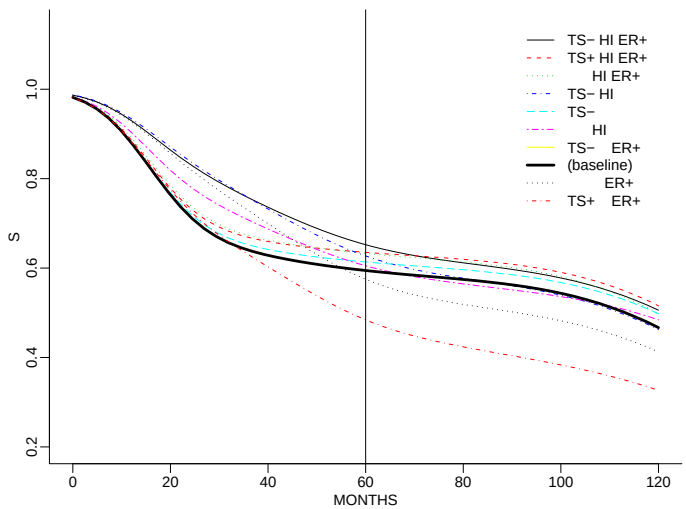
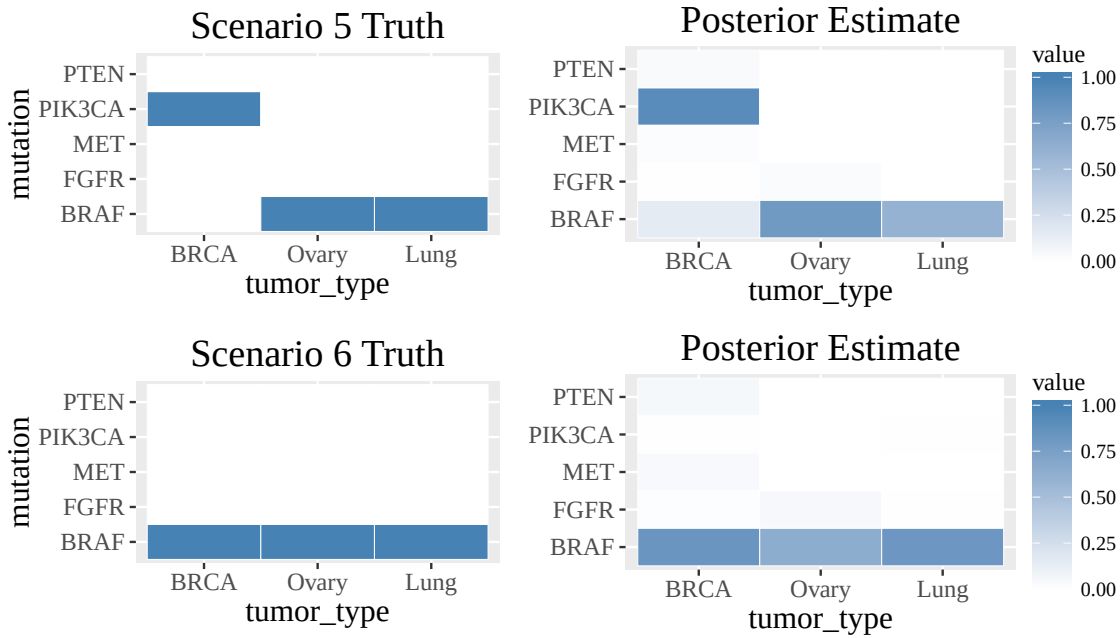
Nonparametric Bayesian Data Analysis for Causal Inference

Part 2 – Regression

Peter Müller, UT Austin



regression



left = truth; right = estimate as $p(a)$ over repeat sim.

Slides: www.math.utexas.edu/users/pmueller/osu.pdf

2. Regression

Regression: $y_i \mid x_i = x \sim F_x(y_i)$.

2. Regression

Regression: $y_i \mid x_i = x \sim F_x(y_i)$.

1. NP on residual: $y_i = f_\theta(x_i) + \epsilon_i$, $\epsilon_i \sim G$ and $G \sim p(G)$.
   Semiparametric Bayes, density estimation for residuals ϵ_i, e.g., PT prior (Hanson & Johnson, 2002 JASA).

2. Regression

Regression: $y_i \mid x_i = x \sim F_x(y_i)$.

1. NP on residual: $y_i = f_\theta(x_i) + \epsilon_i$, $\epsilon_i \sim G$ and $G \sim p(G)$.
   Semiparametric Bayes, density estimation for residuals ϵ_i, e.g., PT prior (Hanson & Johnson, 2002 JASA).
2. Random regression mean function :
   $y_i = f(x_i) + \epsilon_i$ and $f(\cdot) \sim p(f)$
   GP prior, wavelet bases, neural networks, hierarchical mixture of experts, etc.

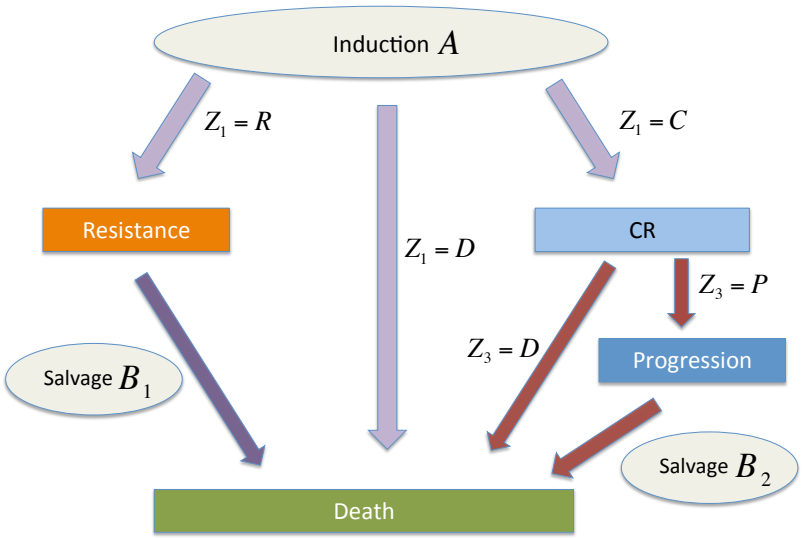
2. Regression

Regression: $y_i \mid x_i = x \sim F_x(y_i)$.

1. NP on residual: $y_i = f_\theta(x_i) + \epsilon_i$, $\epsilon_i \sim G$ and $G \sim p(G)$.
   Semiparametric Bayes, density estimation for residuals ϵ_i, e.g., PT prior (Hanson & Johnson, 2002 JASA).
2. Random regression mean function :
   $y_i = f(x_i) + \epsilon_i$ and $f(\cdot) \sim p(f)$
   GP prior, wavelet bases, neural networks, hierarchical mixture of experts, etc.
3. Fully non-parametric regression:
   $y_i \mid x_i \sim F_{x_i}$, with $\mathcal{F} = \{F_x,\ x \in X\} \sim p(\mathcal{F})$.
   For example, DDP model, dependent PT etc.
   Introduce the DDP next ...

Example 1: Dynamic treatment regimen

Xu et al. (2016 JASA)



Problem: Frontline therapy (A) is randomized, salvage therapy (B) is usually **not randomized**. Adjust for the lack of randomization.

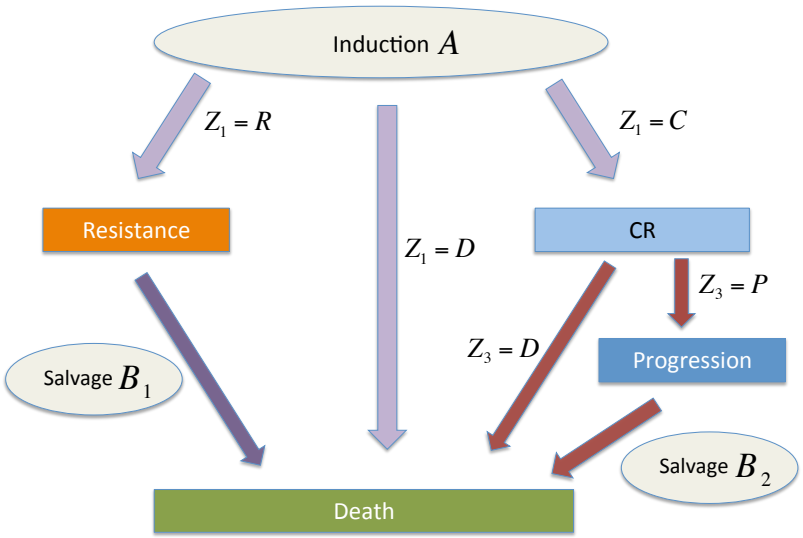
Motivating leukemia trial

Aim: BNP approach to evaluate DTRs, using model-based inference to undo the lack of randomization.

- 4 **induction** trts: FAI, FAI+ATRA, FAI+GCSF, FAI+ATRA+GCSF.
- 2 **salvage** trts: HDAC or not.

Example 1: Dynamic treatment regimen

Xu et al. (2016 JASA)



Problem: Frontline therapy (A) is randomized, salvage therapy (B) is usually **not randomized**. Adjust for the lack of randomization.

Motivating leukemia trial

Aim: BNP approach to evaluate DTRs, using model-based inference to undo the lack of randomization.

- 4 **induction** trts: FAI, FAI+ATRA, FAI+GCSF, FAI+ATRA+GCSF.
- 2 **salvage** trts: HDAC or not.

BNP Model for Evaluating DTRs

Data:

Outcome: $Y^k = \log(T^k) = (\log)\ k^{th}$ transition time (e.g., R $\rightarrow$ D)
Covariates: $\boldsymbol{x}^k$, incl. T^ℓ, $\ell < k$

BNP Model for Evaluating DTRs

Data:

- Outcome: $Y^k = \log(T^k) = (\log)\ k^{th}$ transition time (e.g., R $\rightarrow$ D)
- Covariates: $\boldsymbol{x}^k$, incl. T^ℓ, $\ell < k$

Pars: $\mathcal{F} = \{F^k;\ k = 1, \ldots, K\}$, (unknown) distributions of 7 transition times

Likelihood:

$$\prod_{k=1}^{K} p(Y^k \mid \boldsymbol{x}^k, \mathcal{F}) = \prod_{k=1}^{K} F^k_{\boldsymbol{x}^k}(Y^k)$$

BNP Model for Evaluating DTRs

Data:

Outcome: $Y^k = \log(T^k) = (\log)\ k^{th}$ transition time (e.g., R $\rightarrow$ D)

Covariates: $\boldsymbol{x}^k$, incl. T^ℓ, $\ell < k$

Pars: $\mathcal{F} = \{F^k;\ k = 1, \ldots, K\}$, (unknown) distributions of 7 transition times

Likelihood:

$$\prod_{k=1}^{K} p(Y^k \mid \boldsymbol{x}^k, \mathcal{F}) = \prod_{k=1}^{K} F^k_{\boldsymbol{x}^k}(Y^k)$$

Prior: BNP prior for $\mathcal{F}$

$$\mathcal{F} = \{F^k_x;\ x \in X,\} \sim \text{DDP}, \quad k = 1, \ldots, K$$

with $F^k_x = \sum_{h=0}^{\infty} p^k_h\, N(y;\ \theta^k_{h,x},\ \sigma^k)$.

BNP Model for Evaluating DTRs

Data:

Outcome: $Y^k = \log(T^k) = (\log)\ k^{th}$ transition time (e.g., R $\rightarrow$ D)

Covariates: $\boldsymbol{x}^k$, incl. T^ℓ, $\ell < k$

Pars: $\mathcal{F} = \{F^k;\ k = 1, \ldots, K\}$, (unknown) distributions of 7 transition times

Likelihood:

$$\prod_{k=1}^{K} p(Y^k \mid \boldsymbol{x}^k, \mathcal{F}) = \prod_{k=1}^{K} F^k_{\boldsymbol{x}^k}(Y^k)$$

Prior: BNP prior for $\mathcal{F}$

$$\mathcal{F} = \{F^k_x;\ x \in X,\} \sim \text{DDP}, \quad k = 1, \ldots, K$$

with $F^k_x = \sum_{h=0}^{\infty} p^k_h\, N(y;\ \theta^k_{h,x},\ \sigma^k)$.

GP prior on $\{\theta^k_{h,x}\}_{\boldsymbol{x}}$

Prior: (skip "k" superindex for the moment)

$F_x = \sum_{h=0}^{\infty} p_h \, N(y;\ \theta_{h,x}^k,\ \sigma)$.

Prior: (skip "k" superindex for the moment)

$F_x = \sum_{h=0}^{\infty} p_h \, N(y;\ \theta_{h,x}^k,\ \sigma)$.

- stick-breaking prior on p_h
- GP prior on the functions $\{\theta_{h,x}\}_{\boldsymbol{x}}$, dependent across x, independent across h

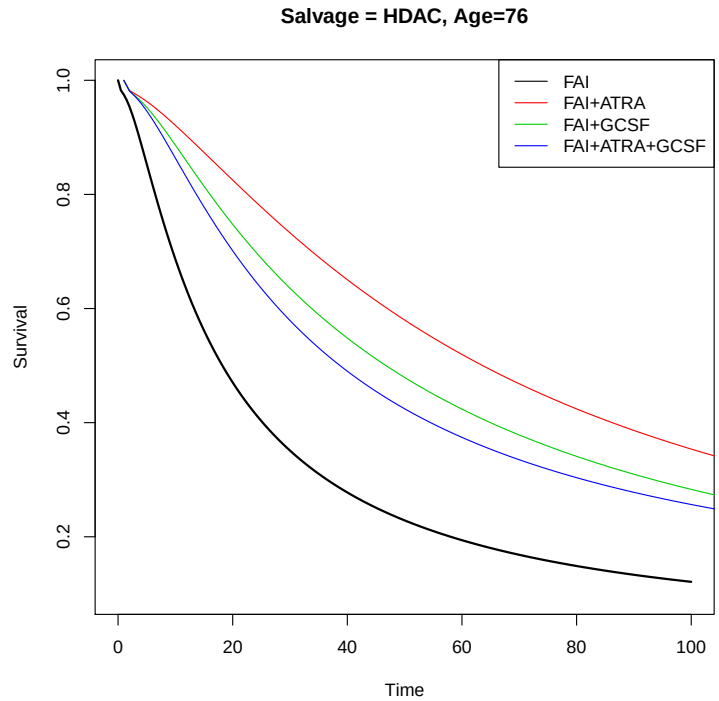
Results: Survival regression and optimal policy

Survival regression: for each $T_{\boldsymbol{x}}^k$, using DDP mix of normal

Results: Survival regression and optimal policy

Survival regression: for each $T_{\boldsymbol{x}}^k$, using DDP mix of normal

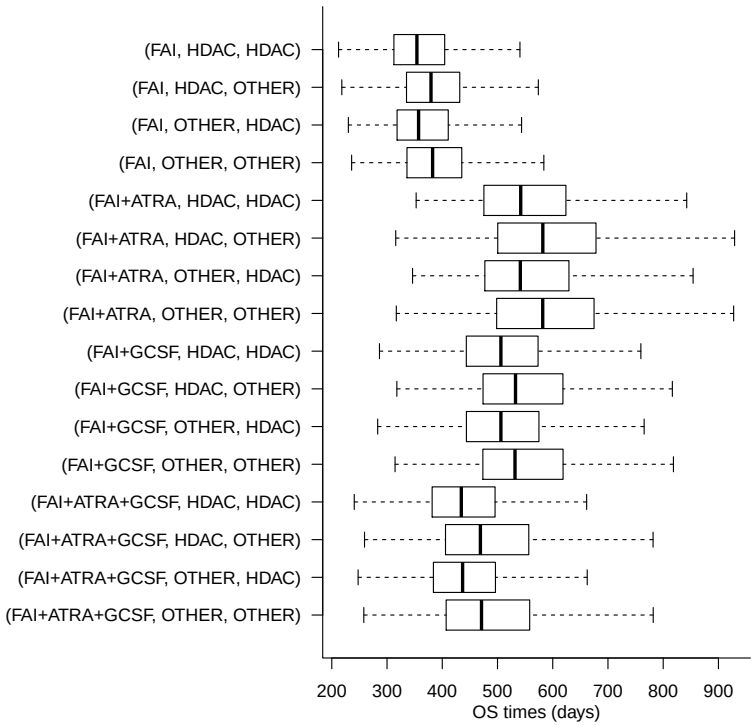
Prior support: full prior support; BNP is *always right*; this mitigates concerns about extrapolation.



survival regr for T^{PD}

Comparing policies

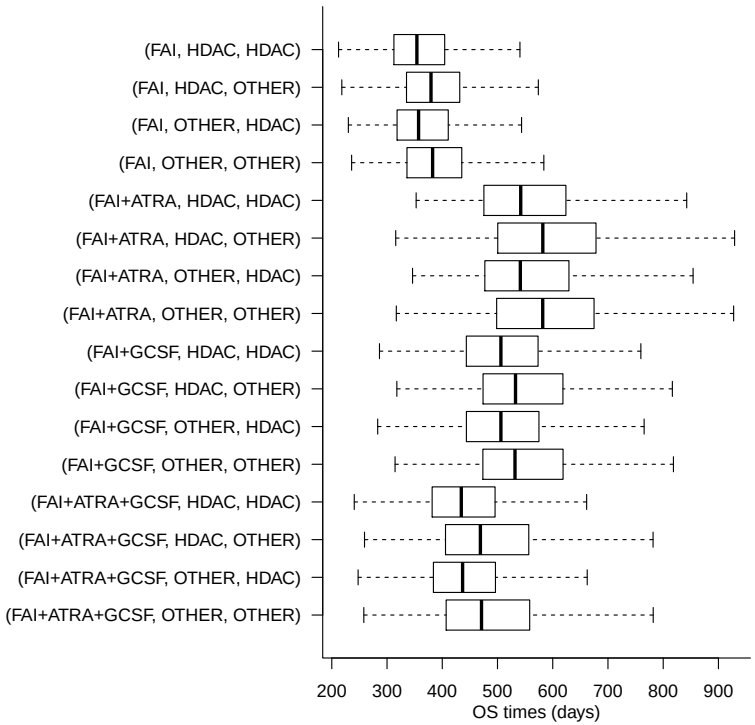
Overall survival for alternative policies (A, B_1, B_2).



Potential outcomes: evaluate mean OS for possible treatment policies

Comparing policies

Overall survival for alternative policies (A, B_1, B_2).

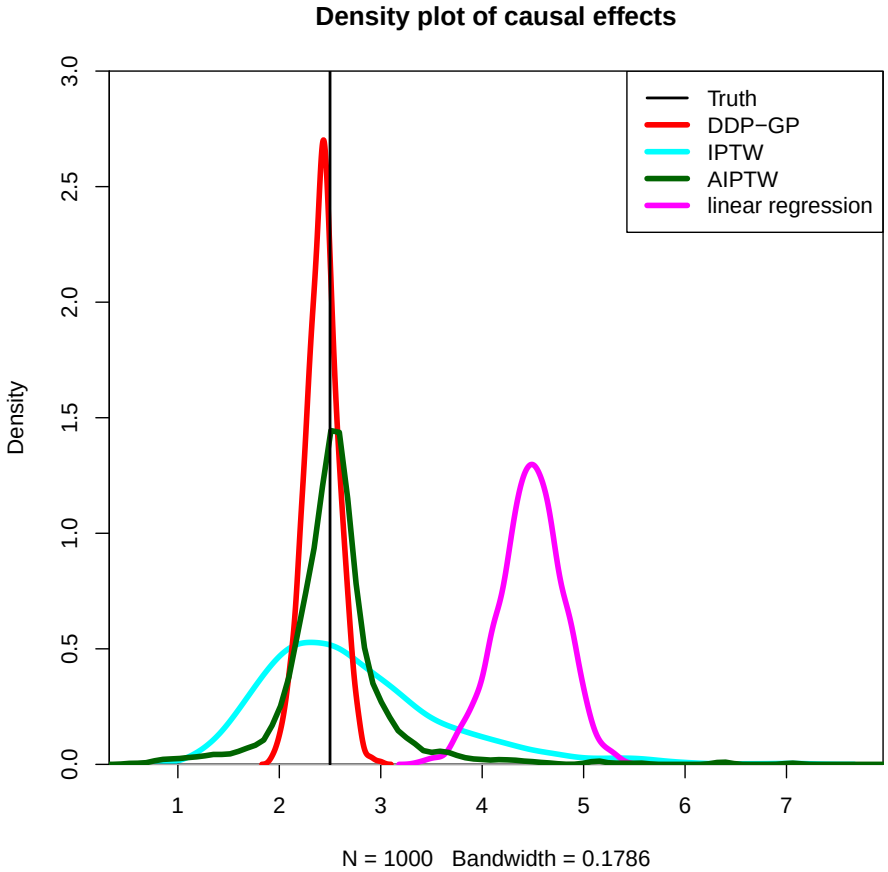


Potential outcomes: evaluate mean OS for possible treatment policies

Optimal policy: compare by mean OS

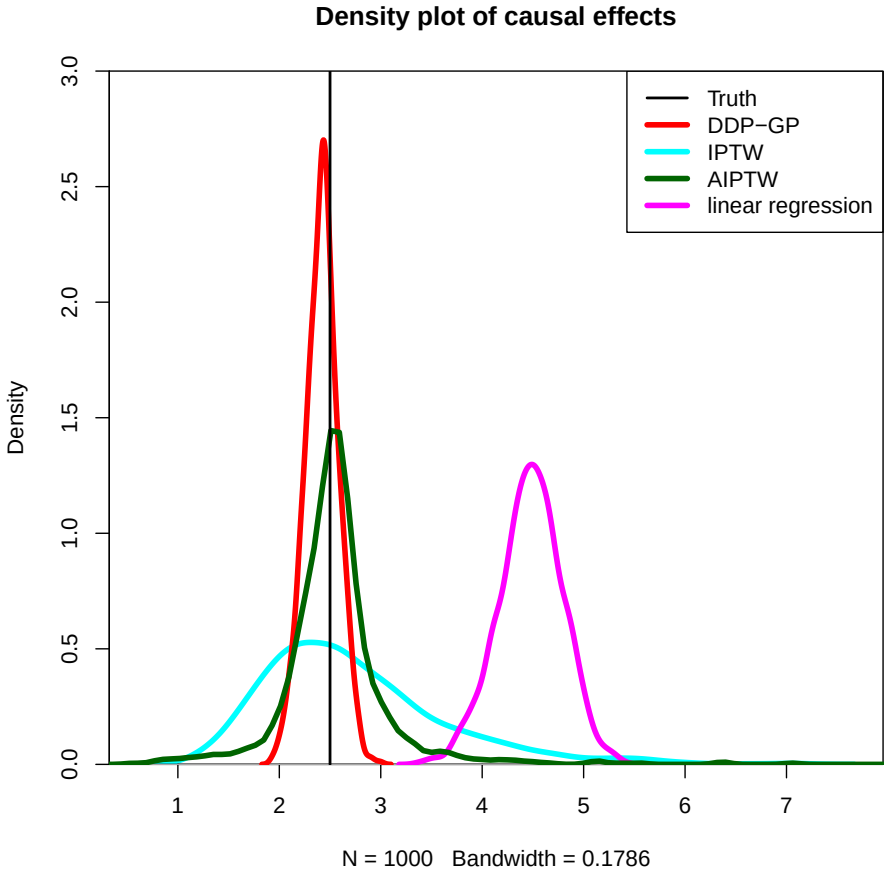
Comparison with double robust methods

Two simulations to compare with inverse prob weighting, using correct model (left) and mis-specified model (right)

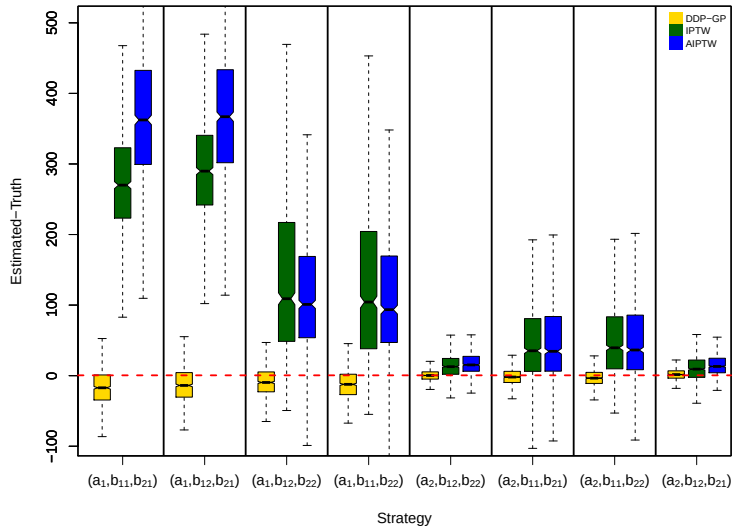


Comparison with double robust methods

Two simulations to compare with inverse prob weighting, using correct model (left) and mis-specified model (right)



single event time
(correct model)



DTR, with both
models wrong

Example 2: Semicompeting risks

Xu, Scharfstein, M and Daniels (2019, arXiv). Another application of (almost) the same model for pairs of event times.

Event times: progression P_j & overall survival D_j
under control ($j = 0$) and treatment ($j = 1$).

Example 2: Semicompeting risks

Xu, Scharfstein, M and Daniels (2019, arXiv). Another application of (almost) the same model for pairs of event times.

Event times: progression P_j & overall survival D_j
under control ($j = 0$) and treatment ($j = 1$).

Censoring: D_j censors P_j;
and independent censoring C_j

Example 2: Semicompeting risks

Xu, Scharfstein, M and Daniels (2019, arXiv). Another application of (almost) the same model for pairs of event times.

Event times: progression P_j & overall survival D_j
under control ($j = 0$) and treatment ($j = 1$).

Censoring: D_j censors P_j;
and independent censoring C_j

Inference: compare P_j adjusting for D_j

Inference target: conditional odds

$$\tau_{\boldsymbol{x}}(u) = \frac{p_{\boldsymbol{x}}(P_1 < u \mid D_0 > u, D_1 > u)}{p_{\boldsymbol{x}}(P_0 < u \mid D_0 > u, D_1 > u)}$$

Joint distribution for P_j, D_j

Identifiability: Let

$$G_j = p(D_j)$$

Joint distribution for P_j, D_j

Identifiability: Let

$$G_j = p(D_j)$$

and

$$V_j(s \mid t) = p(P_j \leq s, P_j < D_j \mid D_j = t).$$

$s < t$ (for the moment, ignoring regression on "$\boldsymbol{x}$").
Under random censoring G_j and V_j are identifiable –
just use the corresponding sample statistics.

Joint distribution for P_j, D_j

Identifiability: Let

$$G_j = p(D_j)$$

and

$$V_j(s \mid t) = p(P_j \leq s, P_j < D_j \mid D_j = t).$$

$s < t$ (for the moment, ignoring regression on "$\boldsymbol{x}$").
Under random censoring G_j and V_j are identifiable – just use the corresponding sample statistics.

Bivariate sub-distribution: together G_j & V_j define

$$\widetilde{F}_1(s,t) = p(P_1 \leq s, D_1 \leq t, P_1 \leq D_1)$$

$s \leq t$, and same for $\widetilde{F}_0$.

Joint distribution for P_j, D_j

Identifiability: Let

$$G_j = p(D_j)$$

and

$$V_j(s \mid t) = p(P_j \leq s, P_j < D_j \mid D_j = t).$$

$s < t$ (for the moment, ignoring regression on "$\boldsymbol{x}$").
Under random censoring G_j and V_j are identifiable – just use the corresponding sample statistics.

Bivariate sub-distribution: together G_j & V_j define

$$\widetilde{F}_1(s,t) = p(P_1 \leq s, D_1 \leq t, P_1 \leq D_1)$$

$s \leq t$, and same for $\widetilde{F}_0$.

Random prob measures, $F_1(s,t)$ & $F_0(s,t)$ imply $\widetilde{F}_1$ & $\widetilde{F}_0$.
DDP mix of normals, as before

Copula $G(D_0, D_1)$

Copula: Link F_0 and F_1 with a normal copula.

Φ = standard normal c.d.f and

$\Phi_{2,\rho}$ = bivariate normal with correlation ρ.

Copula $G(D_0, D_1)$

Copula: Link F_0 and F_1 with a normal copula.

Φ = standard normal c.d.f and

$\Phi_{2,\rho}$ = bivariate normal with correlation ρ.

$$G(D_0, D_1;\ \rho) = \Phi_{2,\rho}\left[\Phi^{-1}\{G_0(D_0)\}, \Phi^{-1}\{G_1(D_1)\}\right]$$

Copula $G(D_0, D_1)$

Copula: Link F_0 and F_1 with a normal copula.

Φ = standard normal c.d.f and

$\Phi_{2,\rho}$ = bivariate normal with correlation ρ.

$$G(D_0, D_1;\ \rho) = \Phi_{2,\rho}\left[\Phi^{-1}\{G_0(D_0)\}, \Phi^{-1}\{G_1(D_1)\}\right]$$

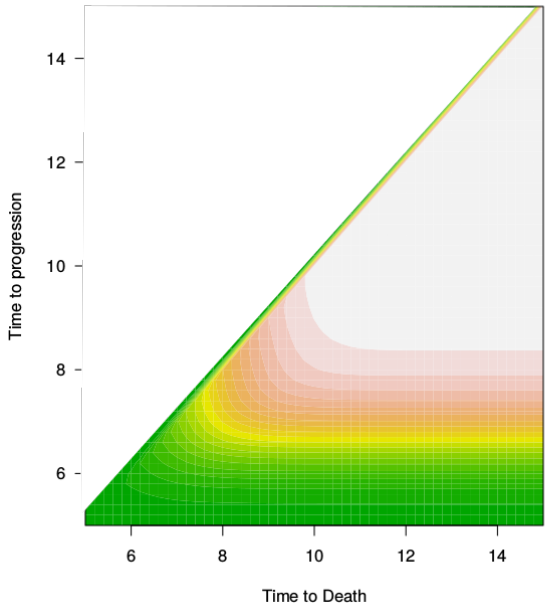
ρ is not identifiable – choice of ρ is an assumption.

Odds of progression

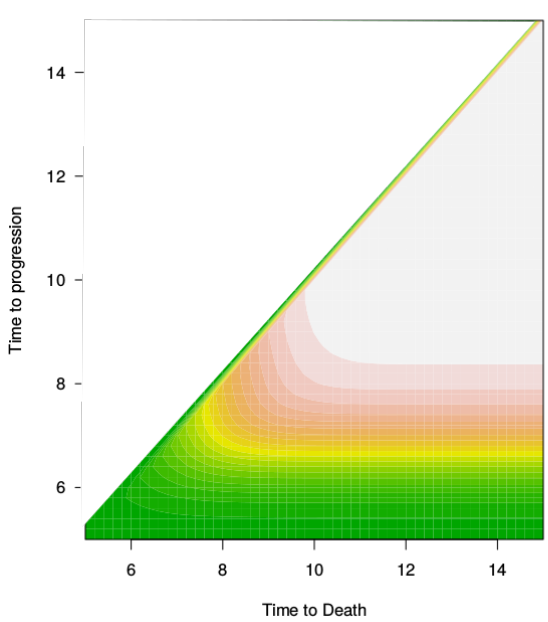
Then

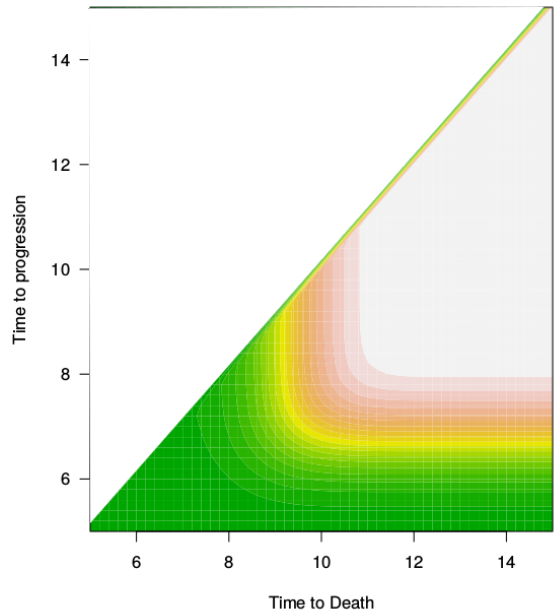
$$\tau_{\boldsymbol{x}}(u) = \frac{\int_{P_1<u} \int_{D_0 \geq u} \int_{D_1 \geq u} dV_1(P_1 \mid D_1, \boldsymbol{x})\, dG_{\boldsymbol{x}}(D_0, D_1)}{\int_{P_0<u} \int_{D_0 \geq u} \int_{D_1 \geq u} dV_0(P_0 \mid D_1), \boldsymbol{x}\, dG_{\boldsymbol{x}}(D_0, D_1)}$$

Results – Brain tumor study

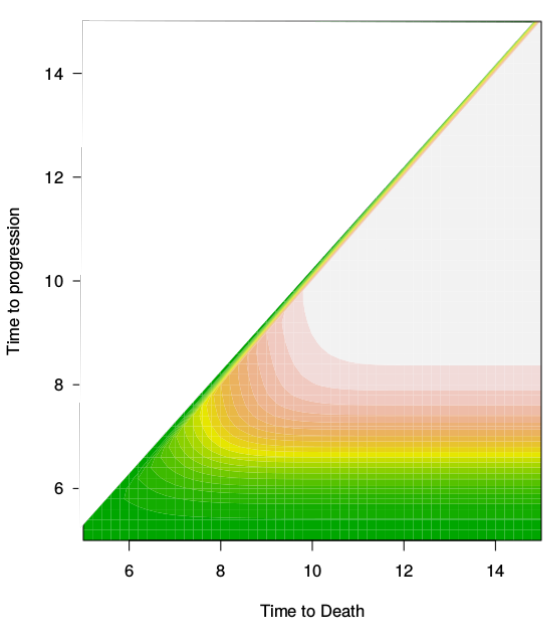


Results – Brain tumor study

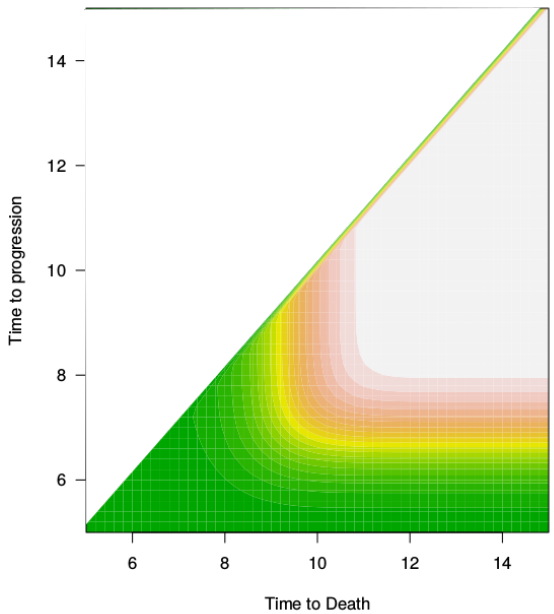




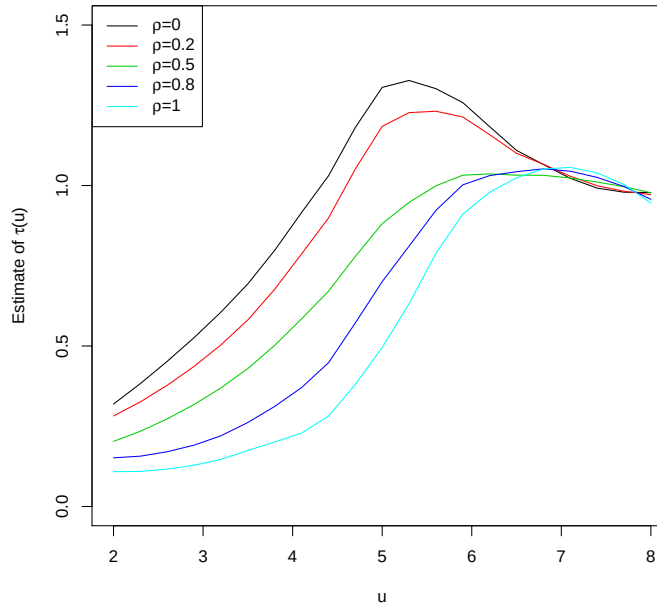
Results – Brain tumor study



$p(D_0, P_0)$



$p(D_1, P_1)$



$\tau(u)$ (on *log* scale!)

BNP regression by covariate-dependent partitions

Define BNP regression by

1. random partition, indexed by covariates;
2. cluster-specific sampling model.

$\rightarrow$ next topic..

3. Classification

Categorical x_i: different subpopulations of interest

Aim: classify a new patient as $x_{n+1} = x \in \{0, 1\}$

Model:

$$y_i \mid x_i = 1 \sim F_x \text{ and } \{F_x;\ x = 0, 1\} \sim \text{DDP}$$

as before (GP simplifies to bivariate normal for $x \in \{0, 1\}$), but ...

3. Classification

Categorical x_i: different subpopulations of interest

Aim: classify a new patient as $x_{n+1} = x \in \{0, 1\}$

Model:

$$y_i \mid x_i = 1 \sim F_x \text{ and } \{F_x;\ x = 0, 1\} \sim \text{DDP}$$

as before (GP simplifies to bivariate normal for $x \in \{0, 1\}$), but ...

Simple augmentation: with

$$p(x_i = 1) = \pi$$

allows the desired ...

3. Classification

Categorical x_i: different subpopulations of interest

Aim: classify a new patient as $x_{n+1} = x \in \{0, 1\}$

Model:

$$y_i \mid x_i = 1 \sim F_x \text{ and } \{F_x;\ x = 0, 1\} \sim \text{DDP}$$

as before (GP simplifies to bivariate normal for $x \in \{0, 1\}$), but ...

Simple augmentation: with

$$p(x_i = 1) = \pi$$

allows the desired ...

Classification: $p(x_{n+1} = 1 \mid data)$ – that's all! (de la Cruz et al., 2007 ApplStat)

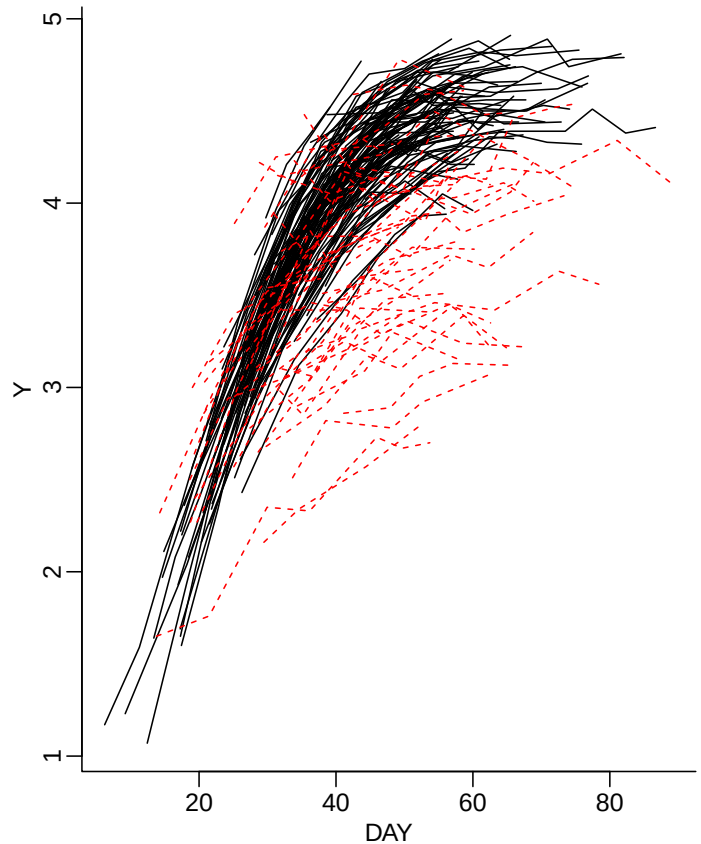
Example 3: Pregnancy classification

De la Cruz-Mesia et al. (2007, ApplStat)

Data: hormone data y_{ij} for $n = 173$ pregnant women, repeat mmt at times t_{ij}, $j = 1, \ldots, n_i$

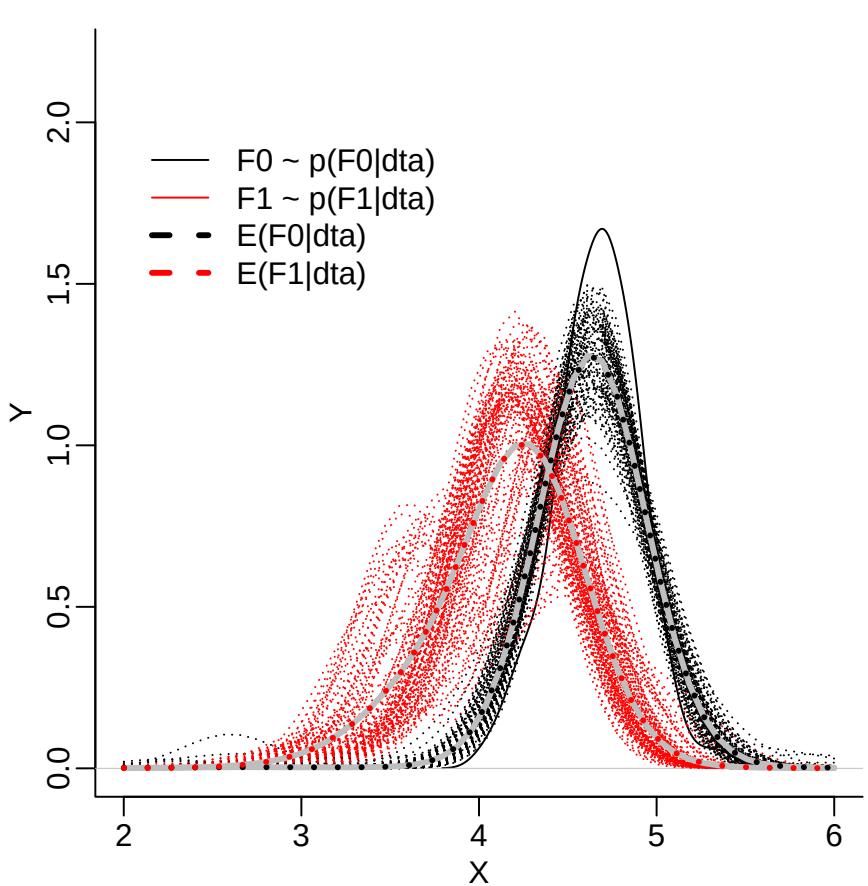
Subpopulations: $x_i = 0$, normal pregnancies, $n_0 = 124$ women

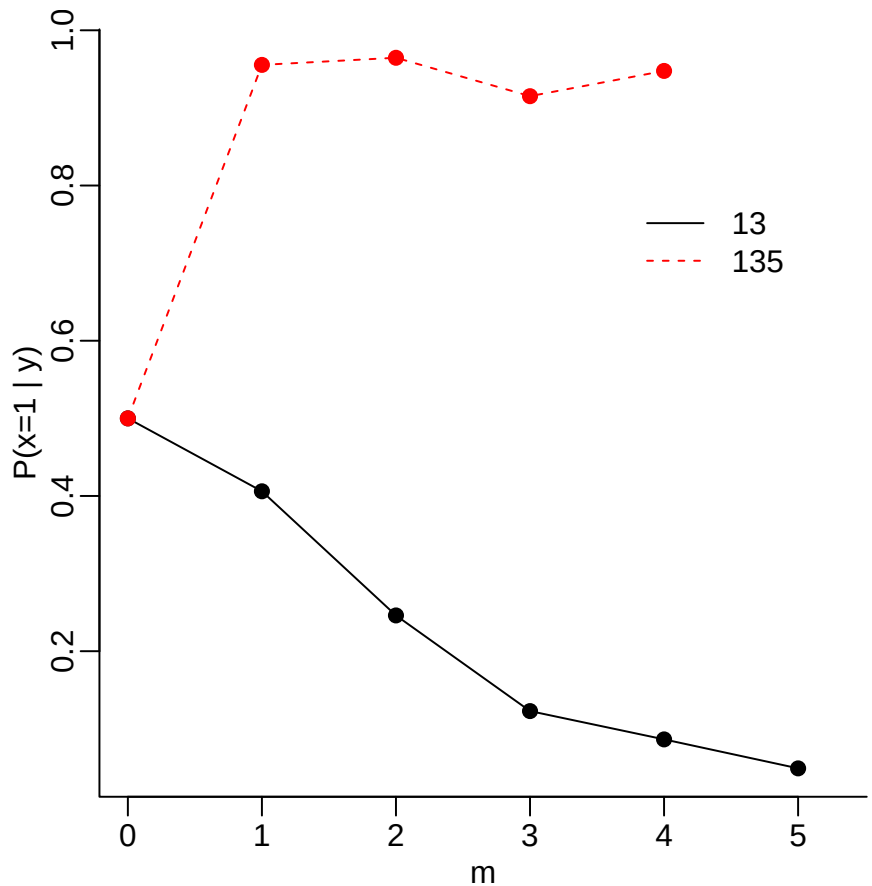
$x_i = 1$, abnormal, $n_1 = 49$



Sampling model: $y_{ij} \mid x_i = x, \ldots \sim N(m_{ij}, \sigma_x^2)$

with $m_{ij} = \theta_i / \left\{1 + e^{-(t_{ij} - \beta_{1x})/\beta_{2x}}\right\}$





(a) $E(F_x \mid data)$ (b)$p(x_{n+1} = 1 \mid y_{n+1,1\ldots m}, data)$

Estimated F_x under $x = 0$ (thick black curve) and $x = 1$ (thick red or grey) (panel a), and posterior probability $p(x_{n+1} = 1 \mid y_{n+1,1\ldots m}, data)$

2. Clustering

Recall: DP Mixtures: convolution of discrete $F = \sum p_h \delta_{m_h}$ with (continuous) kernel, e.g., normal

$$G(y) \;=\; \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP}$$

2. Clustering

Recall: DP Mixtures: convolution of discrete $F = \sum p_h \delta_{m_h}$ with (continuous) kernel, e.g., normal

$$\begin{aligned} G(y) &= \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP} \\ &= \sum_{h=1}^{\infty} p_h N(y \mid m_h, \sigma) \end{aligned}$$

continous $G(\cdot)$ (and hyperpar σ^2)

2. Clustering

Recall: DP Mixtures: convolution of discrete $F = \sum p_h \delta_{m_h}$ with (continuous) kernel, e.g., normal

$$\begin{aligned} G(y) &= \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP} \\ &= \sum_{h=1}^{\infty} p_h N(y \mid m_h, \sigma) \end{aligned}$$

continous $G(\cdot)$ (and hyperpar σ^2)

Latent vars: write $\int \ldots dF(\theta)$ as hierarchical model

$$\begin{aligned} y_i \mid \theta_i &\sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n \\ \theta_i \mid F &\sim F \end{aligned}$$

2. Clustering

Recall: DP Mixtures: convolution of discrete $F = \sum p_h \delta_{m_h}$ with (continuous) kernel, e.g., normal

$$\begin{aligned} G(y) &= \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP} \\ &= \sum_{h=1}^{\infty} p_h N(y \mid m_h, \sigma) \end{aligned}$$

continous $G(\cdot)$ (and hyperpar σ^2)

Latent vars: write $\int \ldots dF(\theta)$ as hierarchical model

$$\begin{aligned} y_i \mid \theta_i &\sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n \\ \theta_i \mid F &\sim F \end{aligned}$$

Notation: discrete $F \Rightarrow K \leq n$ unique θ_i's $= \{\phi_1^\star, \ldots, \phi_K^\star\}$.

2. Clustering

Recall: DP Mixtures: convolution of discrete $F = \sum p_h \delta_{m_h}$ with (continuous) kernel, e.g., normal

$$\begin{aligned} G(y) &= \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP} \\ &= \sum_{h=1}^{\infty} p_h N(y \mid m_h, \sigma) \end{aligned}$$

continous $G(\cdot)$ (and hyperpar σ^2)

Latent vars: write $\int \ldots dF(\theta)$ as hierarchical model

$$\begin{aligned} y_i \mid \theta_i &\sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n \\ \theta_i \mid F &\sim F \end{aligned}$$

Notation: discrete $F \Rightarrow K \leq n$ unique θ_i's $= \{\phi_1^\star, \ldots, \phi_K^\star\}$.
Latent indicators: $z_i = j$ iff $\theta_i = \phi_j^\star$ match θ_i with $\phi_j^\star$'s.

2. Clustering

Recall: DP Mixtures: convolution of discrete $F = \sum p_h \delta_{m_h}$ with (continuous) kernel, e.g., normal

$$\begin{aligned} G(y) &= \int N(y \mid \theta, \sigma^2)\, dF(\theta), \quad F \sim \text{DP} \\ &= \sum_{h=1}^{\infty} p_h N(y \mid m_h, \sigma) \end{aligned}$$

continous $G(\cdot)$ (and hyperpar σ^2)

Latent vars: write $\int \ldots dF(\theta)$ as hierarchical model

$$\begin{aligned} y_i \mid \theta_i &\sim N(\theta_i, \sigma^2), \quad i = 1, \ldots, n \\ \theta_i \mid F &\sim F \end{aligned}$$

Notation: discrete $F \Rightarrow K \leq n$ unique θ_i's $= \{\phi_1^\star, \ldots, \phi_K^\star\}$.
Latent indicators: $z_i = j$ iff $\theta_i = \phi_j^\star$ match θ_i with $\phi_j^\star$'s.

Random Partition Models

Product partition model (PPM): cohesion functions $c(S_j)$ define similarity of a cluster,

$$p(\rho_n) \propto \prod_{j=1}^{k} c(S_j).$$

Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA)

Random Partition Models

Product partition model (PPM): cohesion functions $c(S_j)$ define similarity of a cluster,

$$p(\rho_n) \propto \prod_{j=1}^{k} c(S_j).$$

Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA)

Sampling model: conditional on partition ρ_n, assume exchangeability,

$$p(y^n \mid \rho, \phi^\star) = \prod_{j=1}^{k} \left\{ \prod_{i \in S_j} p(y_i \mid \phi_j^\star) \right\} \qquad (*)$$

with cluster-specific parameters $\phi_j^\star$

Random Partition Models

Product partition model (PPM): cohesion functions $c(S_j)$ define similarity of a cluster,

$$p(\rho_n) \propto \prod_{j=1}^{k} c(S_j).$$

Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA)

Sampling model: conditional on partition ρ_n, assume exchangeability,

$$p(y^n \mid \rho, \phi^\star) = \prod_{j=1}^{k} \left\{ \prod_{i \in S_j} p(y_i \mid \phi_j^\star) \right\} \qquad (*)$$

with cluster-specific parameters $\phi_j^\star$

Prior $p(\phi_j^\star)$: conjugate ...

Covariate-dependent PPM (PPMx)

M et al. (2011 JCGS), Quintana et al. (2015 ScandJS)

Random partition: to favor clusters of patients with similar covariates,

Covariate-dependent PPM (PPMx)

M et al. (2011 JCGS), Quintana et al. (2015 ScandJS)

Random partition: to favor clusters of patients with similar covariates, define $g(x_j^\star) > 0$ to characterize the similarity of $\{x_i;\ i \in S_j\}$ with low values for bad clusters:

Covariate-dependent PPM (PPMx)

M et al. (2011 JCGS), Quintana et al. (2015 ScandJS)

Random partition: to favor clusters of patients with similar covariates, define $g(x_j^\star) > 0$ to characterize the similarity of $\{x_i;\ i \in S_j\}$ with low values for bad clusters:

$$p(\rho_n \mid x^n) \propto \prod_{j=1}^{k} g(x_j^\star) \cdot c(S_j)$$

Covariate-dependent PPM (PPMx)

M et al. (2011 JCGS), Quintana et al. (2015 ScandJS)

Random partition: to favor clusters of patients with similar covariates, define $g(x_j^\star) > 0$ to characterize the similarity of $\{x_i;\ i \in S_j\}$ with low values for bad clusters:

$$p(\rho_n \mid x^n) \propto \prod_{j=1}^{k} g(x_j^\star) \cdot c(S_j)$$

Similarity function: easy computation with

$$g(x_j^\star) = \int \prod_{i \in S_j} q(x_i \mid \xi_j^\star)\ q(\xi_j^\star) \mathrm{dd}\xi_j^\star$$

Covariate-dependent PPM (PPMx)

M et al. (2011 JCGS), Quintana et al. (2015 ScandJS)

Random partition: to favor clusters of patients with similar covariates, define $g(x_j^\star) > 0$ to characterize the similarity of $\{x_i;\ i \in S_j\}$ with low values for bad clusters:

$$p(\rho_n \mid x^n) \propto \prod_{j=1}^{k} g(x_j^\star) \cdot c(S_j)$$

Similarity function: easy computation with

$$g(x_j^\star) = \int \prod_{i \in S_j} q(x_i \mid \xi_j^\star)\ q(\xi_j^\star) \mathrm{dd}\xi_j^\star$$

using, e.g., $q(x_i \mid \xi_i) = N(\xi_j^\star, V)$ and $q(\xi_j^\star) = N(\ldots)$ for continuous x_i,

Covariate-dependent PPM (PPMx)

M et al. (2011 JCGS), Quintana et al. (2015 ScandJS)

Random partition: to favor clusters of patients with similar covariates, define $g(x_j^\star) > 0$ to characterize the similarity of $\{x_i;\ i \in S_j\}$ with low values for bad clusters:

$$p(\rho_n \mid x^n) \propto \prod_{j=1}^{k} g(x_j^\star) \cdot c(S_j)$$

Similarity function: easy computation with

$$g(x_j^\star) = \int \prod_{i \in S_j} q(x_i \mid \xi_j^\star)\ q(\xi_j^\star) \mathrm{dd}\xi_j^\star$$

using, e.g., $q(x_i \mid \xi_i) = N(\xi_j^\star, V)$ and $q(\xi_j^\star) = N(\ldots)$ for continuous x_i, and similar conjugate choices for categorical, ordinal and counts.

Example

Example 4: Survival regression with PPMx

M, Quintana & Rosner (2011 JCGS) analyze data from a study (CALGB 9082) of breast cancer patients.

Treatment: high dose (A) versus low dose (B) chemotherapy

Example

Example 4: Survival regression with PPMx

M, Quintana & Rosner (2011 JCGS) analyze data from a study (CALGB 9082) of breast cancer patients.

Treatment: high dose (A) versus low dose (B) chemotherapy

Data: 765 patients randomized to A vs. B.

Example

Example 4: Survival regression with PPMx

M, Quintana & Rosner (2011 JCGS) analyze data from a study (CALGB 9082) of breast cancer patients.

Treatment: high dose (A) versus low dose (B) chemotherapy

Data: 765 patients randomized to A vs. B.

Response: time until progression or death

Example

Example 4: Survival regression with PPMx

M, Quintana & Rosner (2011 JCGS) analyze data from a study (CALGB 9082) of breast cancer patients.

Treatment: high dose (A) versus low dose (B) chemotherapy

Data: 765 patients randomized to A vs. B.

Response: time until progression or death

Covariates:
- *Categorical:* dose (A vs. B), menopausal status, estrogen use
- *Continuous:* age, initial tumor size,
- *Count:* number of positive lymph nodes

Example

Example 4: Survival regression with PPMx

M, Quintana & Rosner (2011 JCGS) analyze data from a study (CALGB 9082) of breast cancer patients.

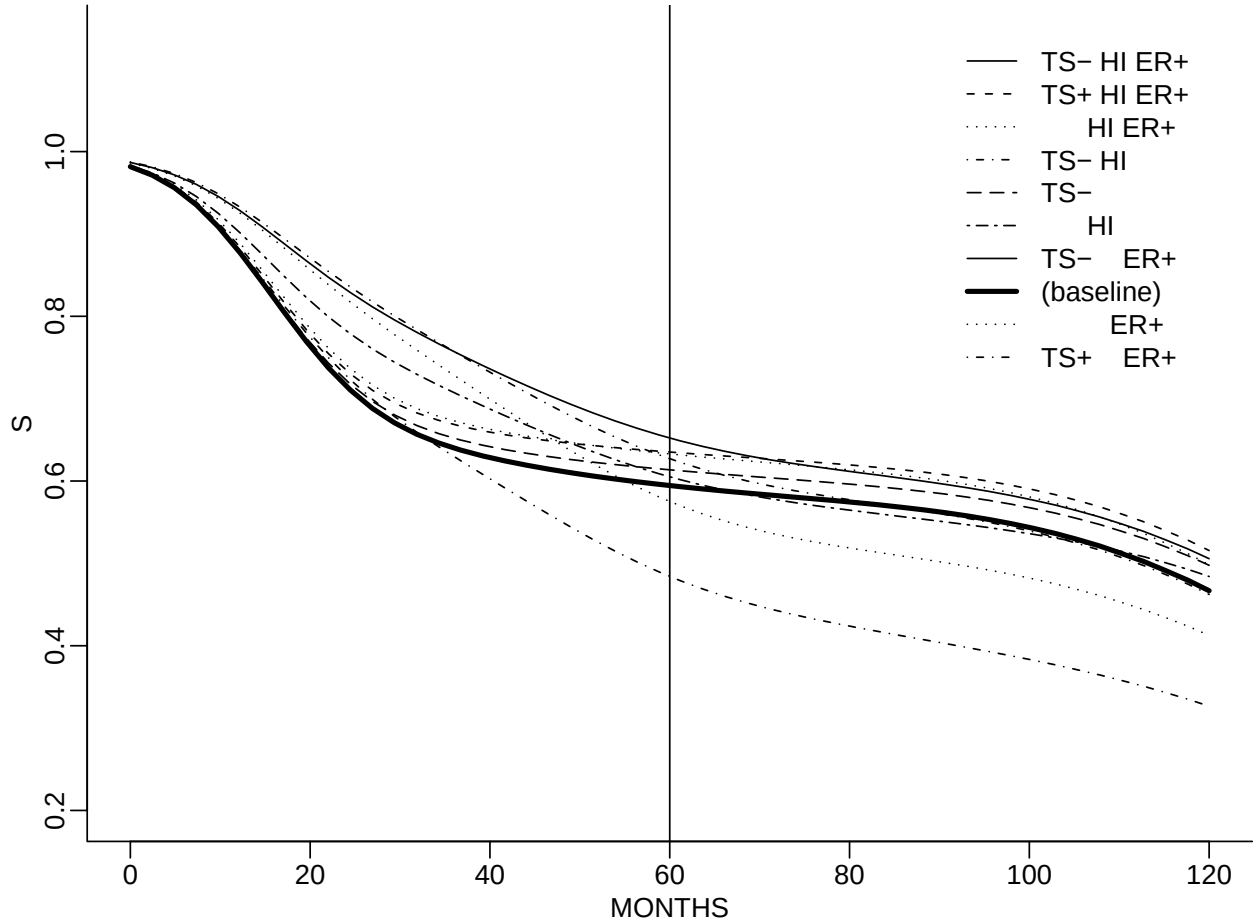
Treatment: high dose (A) versus low dose (B) chemotherapy

Data: 765 patients randomized to A vs. B.

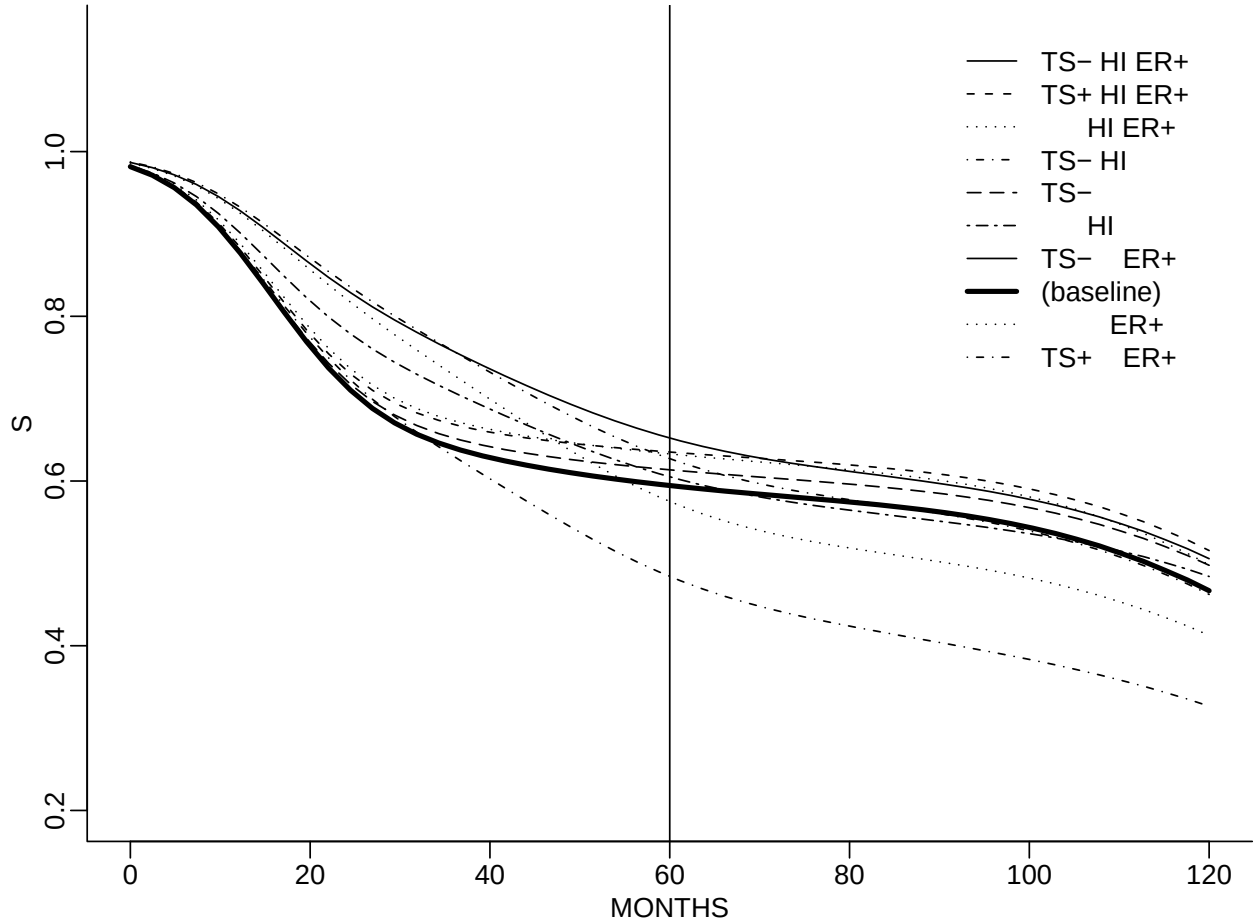
Response: time until progression or death

Covariates:
- *Categorical:* dose (A vs. B), menopausal status, estrogen use
- *Continuous:* age, initial tumor size,
- *Count:* number of positive lymph nodes

Model: PPMx, with cluster-specific normal sampling



$S(t \mid x)$ by covariates



$S(t \mid x)$ by covariates

BNP regression: use the PPMx for BNP regression; allowing regression with variable dimension covariate vector!

Bayesian Subgroup analysis

Subroup analysis problem: inference on exceptions from overall conclusion, typically for a clinical study, for

- a "benefitting population",

vs.

- eligible population of the trial

Bayesian Subgroup analysis

Subroup analysis problem: inference on exceptions from overall conclusion, typically for a clinical study, for

- a "benefitting population",

vs.

- eligible population of the trial

Approaches :

- Treatment/cov interaction: Dixon and Simon (1991 Bmcs), Jones et al. (2011 ClinTrials)

Bayesian Subgroup analysis

Subroup analysis problem: inference on exceptions from overall conclusion, typically for a clinical study, for

- a "benefitting population",

vs.

- eligible population of the trial

Approaches :

- Treatment/cov interaction: Dixon and Simon (1991 Bmcs), Jones et al. (2011 ClinTrials)
- Tree based methods: Foster, Taylor & Ruberg (2011 StatMed)

Bayesian Subgroup analysis

Subroup analysis problem: inference on exceptions from overall conclusion, typically for a clinical study, for

- a "benefitting population",

vs.

- eligible population of the trial

Approaches :

- Treatment/cov interaction: Dixon and Simon (1991 Bmcs), Jones et al. (2011 ClinTrials)
- Tree based methods: Foster, Taylor & Ruberg (2011 StatMed)
- Model selection: Berger, Wang and Shen (2014, J Biopharm Stat), Sivaganesan et al. (2011 StatMed)

Bayesian Subgroup analysis

Subroup analysis problem: inference on exceptions from overall conclusion, typically for a clinical study, for

- a "benefitting population",

vs.

- eligible population of the trial

Approaches :

- Treatment/cov interaction: Dixon and Simon (1991 Bmcs), Jones et al. (2011 ClinTrials)
- Tree based methods: Foster, Taylor & Ruberg (2011 StatMed)
- Model selection: Berger, Wang and Shen (2014, J Biopharm Stat), Sivaganesan et al. (2011 StatMed)
- Decision problem: next slides...

Decision Problem

Data: response y_i, covariates $x_i = (x_{i1}, \ldots, x_{ip})$.

Decision Problem

Data: response y_i, covariates $x_i = (x_{i1}, \ldots, x_{ip})$.

Actions: Report a subgroup of patients who most benefit from the experimental therapy:

$$\boldsymbol{a} = (I, \mathbf{x}^\star),$$

Covariates: $I \subset \{1, \ldots, p\}$

Levels: $\mathbf{x}^\star = (x_j^\star,\ j \in I)$,

(possibly restrict continuous $x_j^\star$ to fixed thresholds)

Decision Problem

Data: response y_i, covariates $x_i = (x_{i1}, \ldots, x_{ip})$.

Actions: Report a subgroup of patients who most benefit from the experimental therapy:

$$\boldsymbol{a} = (I, \mathbf{x}^\star),$$

Covariates: $I \subset \{1, \ldots, p\}$

Levels: $\mathbf{x}^\star = (x_j^\star,\ j \in I)$,

(possibly restrict continuous $x_j^\star$ to fixed thresholds)

Decision problem: separate inference (predicting y_{n+1}), with flexible model

Decision Problem

Data: response y_i, covariates $x_i = (x_{i1}, \ldots, x_{ip})$.

Actions: Report a subgroup of patients who most benefit from the experimental therapy:

$$\boldsymbol{a} = (I, \mathbf{x}^\star),$$

Covariates: $I \subset \{1, \ldots, p\}$

Levels: $\mathbf{x}^\star = (x_j^\star,\ j \in I)$,

(possibly restrict continuous $x_j^\star$ to fixed thresholds)

Decision problem: separate inference (predicting y_{n+1}), with flexible model

vs.

decision (report subpopulation), parsimoniously

Decision Problem

Data: response y_i, covariates $x_i = (x_{i1}, \ldots, x_{ip})$.

Actions: Report a subgroup of patients who most benefit from the experimental therapy:

$$\boldsymbol{a} = (I, \mathbf{x}^\star),$$

Covariates: $I \subset \{1, \ldots, p\}$

Levels: $\mathbf{x}^\star = (x_j^\star,\ j \in I)$,

(possibly restrict continuous $x_j^\star$ to fixed thresholds)

Decision problem: separate inference (predicting y_{n+1}), with flexible model

vs.

decision (report subpopulation), parsimoniously

- no need for multiplicity control
- arbitrary prob model
- disentagle stat significance vs. clinical relevance
- allow for variable # covs.

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

$$u(a, \theta) = (\mathrm{LR}(a, \theta) - \beta) \cdot \frac{n(\boldsymbol{a})^{\alpha}}{(|I| + 1)^{\gamma}} \tag{1}$$

where θ are parameters that index the sampling model.

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

$$u(a, \theta) = (\mathrm{LR}(a, \theta) - \beta) \cdot \frac{n(\boldsymbol{a})^{\alpha}}{(|I| + 1)^{\gamma}} \tag{1}$$

  where θ are parameters that index the sampling model.
- *Continous outcome:* e.g., % tumor shrinkage, this could be based on predictive average treatment effect (PATE),

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

$$u(a, \theta) = (\mathrm{LR}(a, \theta) - \beta) \cdot \frac{n(\boldsymbol{a})^{\alpha}}{(|I| + 1)^{\gamma}} \tag{1}$$

  where θ are parameters that index the sampling model.
- *Continous outcome:* e.g., % tumor shrinkage, this could be based on predictive average treatment effect (PATE), averaged over x_i and already averaged w.r.t. $p(\theta \mid data)$.

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

$$u(a, \theta) = (\mathrm{LR}(a, \theta) - \beta) \cdot \frac{n(\boldsymbol{a})^{\alpha}}{(|I| + 1)^{\gamma}} \tag{1}$$

  where θ are parameters that index the sampling model.
- *Continous outcome:* e.g., % tumor shrinkage, this could be based on predictive average treatment effect (PATE), averaged over x_i and already averaged w.r.t. $p(\theta \mid data)$.

$$U(a) = \begin{cases} \{\mathrm{PATE}_{SS}(a) - \beta\} \cdot \frac{|n(a)+1|^{\alpha}}{(|I|+1)^{\gamma}} & \text{if } a \neq H_0 \\ u_0 & \text{if } a = H_0, \end{cases}$$

  where H_0, H_1 are special actions,

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

$$u(a, \theta) = (\mathrm{LR}(a, \theta) - \beta) \cdot \frac{n(\boldsymbol{a})^{\alpha}}{(|I| + 1)^{\gamma}} \tag{1}$$

  where θ are parameters that index the sampling model.
- *Continous outcome:* e.g., % tumor shrinkage, this could be based on predictive average treatment effect (PATE), averaged over x_i and already averaged w.r.t. $p(\theta \mid data)$.

$$U(a) = \begin{cases} \{\mathrm{PATE}_{SS}(a) - \beta\} \cdot \frac{|n(a)+1|^{\alpha}}{(|I|+1)^{\gamma}} & \text{if } a \neq H_0 \\ u_0 & \text{if } a = H_0, \end{cases}$$

  where H_0, H_1 are special actions,

with $\beta > 0$ a fixed clinically decided threshold and $n(\boldsymbol{a})$ is the size of the subpopulation.

θ indexes the sampling model

Utility: we favor a subpopulation with difference (relative to the overall population) in trt effect, large size and parsimonious description with few covariates.

- *Event time:* e.g., for an y_i = PFS (event time), this could be based on log hazard ratio

$$u(a, \theta) = (\mathrm{LR}(a, \theta) - \beta) \cdot \frac{n(\boldsymbol{a})^{\alpha}}{(|I| + 1)^{\gamma}} \tag{1}$$

  where θ are parameters that index the sampling model.
- *Continous outcome:* e.g., % tumor shrinkage, this could be based on predictive average treatment effect (PATE), averaged over x_i and already averaged w.r.t. $p(\theta \mid data)$.

$$U(a) = \begin{cases} \{\mathrm{PATE}_{SS}(a) - \beta\} \cdot \frac{|n(a)+1|^{\alpha}}{(|I|+1)^{\gamma}} & \text{if } a \neq H_0 \\ u_0 & \text{if } a = H_0, \end{cases}$$

  where H_0, H_1 are special actions,

with $\beta > 0$ a fixed clinically decided threshold and $n(\boldsymbol{a})$ is the size of the subpopulation.

θ indexes the sampling model (any model for $p(y \mid x, \theta)$)

Bayes rule: Report $a^\star = \arg\max_a \int u(a, \theta)\, dp(\theta \mid data)$

Bayes rule: Report $a^\star = \arg\max_a \int u(a,\theta)\, dp(\theta \mid data)$

Alternative utility: Foster, Taylor & Ruberg (2011, StatMed) use

$$Q(A) = \text{enhanced treatment effect } - \text{average trt effect}$$

and sensitivity and specificity to evaluate a reported subpopulation A.

Bayes rule: Report $a^\star = \arg\max_a \int u(a,\theta)\, dp(\theta \mid data)$

Alternative utility: Foster, Taylor & Ruberg (2011, StatMed) use

$$Q(A) = \text{enhanced treatment effect } - \text{average trt effect}$$

and sensitivity and specificity to evaluate a reported subpopulation A.

Model: Decicsion problem and solution meaningful for any model.

3. Probability Model

Flexible BNP model. The BNP model "is always right."

- *Event time:* for example, PPMx for the event time
- *Continuous outcome:* e.g., DDP, BART

Example 5: Phase III Study of NSCL Patients

Morita & M, 2017 Bmcs

Patients: advanced non-small cell lung cancer, $n = 267$

Example 5: Phase III Study of NSCL Patients

Morita & M, 2017 Bmcs

Patients: advanced non-small cell lung cancer, $n = 267$

Treatment: carboplatin (N) ($n_0 = 130$) vs. paclitaxel + carboplatin (C) ($n_1 = 137$).

Example 5: Phase III Study of NSCL Patients

Morita & M, 2017 Bmcs

Patients: advanced non-small cell lung cancer, $n = 267$

Treatment: carboplatin (N) ($n_0 = 130$) vs. paclitaxel + carboplatin (C) ($n_1 = 137$).

Baseline covariates: pharmacologically relevant gene expressions, including 16 mRNA (mR1 - mR16) and 1 protein (Pn1) expressiaon levels ($p = 17$).

Example 5: Phase III Study of NSCL Patients

Morita & M, 2017 Bmcs

Patients: advanced non-small cell lung cancer, $n = 267$

Treatment: carboplatin (N) ($n_0 = 130$) vs. paclitaxel + carboplatin (C) ($n_1 = 137$).

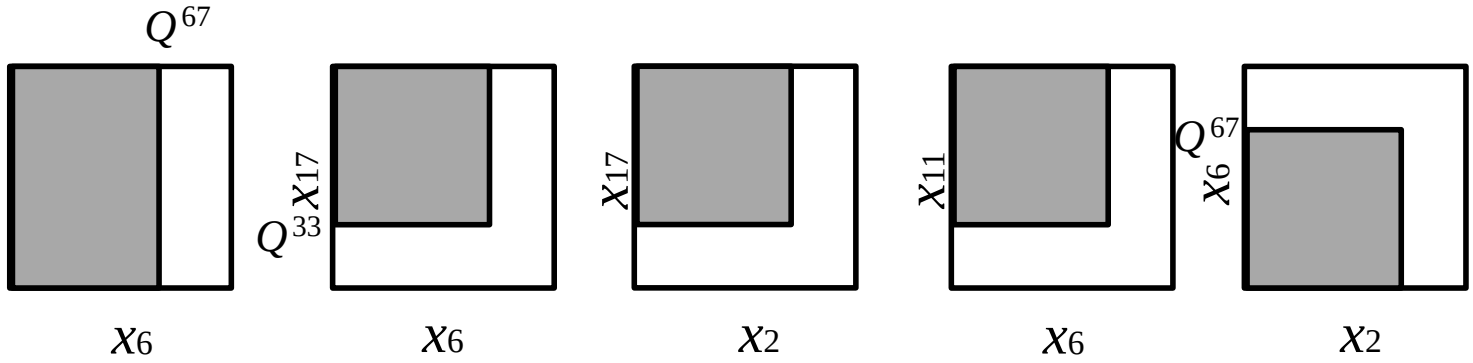
Baseline covariates: pharmacologically relevant gene expressions, including 16 mRNA (mR1 - mR16) and 1 protein (Pn1) expressiaon levels ($p = 17$).

Outcome: y_i =maxTS% (max tumor size shrinkage from baseline)

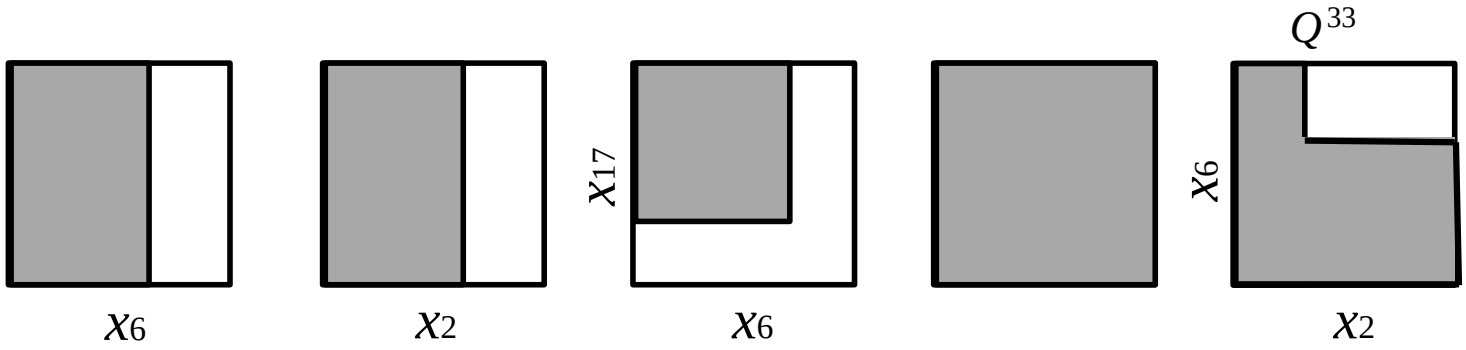
Results

Implement subgroup analysis for the phase III NSCL trial, restricting subgroups to $|I| \leq 2$ covariates.

$(\phi, \zeta) = (0.25, 0.15)$



$(\phi, \zeta) = (0.35, 0.25)$



Example 6: A basket trial design for targeted therapies

Xu et al. (2018 Biometrical J)

Subgroup analysis with a purpose.

IMPACT II: patients across different cancers. Based on molecular alterations patients are eligible for certain targeted therapies (TT)

Subgroup analysis: find subgroup of tumor/mutation pairs who most benefit from TT

Selecting the subpopulations

- Based on a flexible probability model: PPMx

Selecting the subpopulations

- Based on a flexible probability model: PPMx
- Utility function: $u(a, \ldots)$ (1) for event time, PFS

Selecting the subpopulations

- Based on a flexible probability model: PPMx
- Utility function: $u(a, \ldots)$ (1) for event time, PFS
- Report the subpopoulations with largest expected utility
- Adaptive treatment allocation

Simulation

6 scenarios: overall treatment effect (trt);
interaction $z \times$ mutation $\times$ tumor,

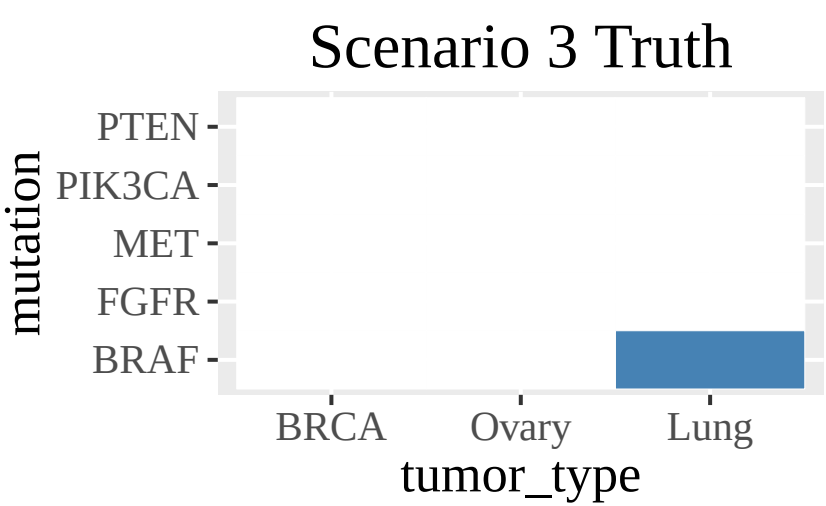
$z \in \{0, 1\}$, mutation $\in \{$BRAF, PIK3CA, PTEN$\}$,
tumors $\in \{$BRCA, Lung, Ovary$\}$.

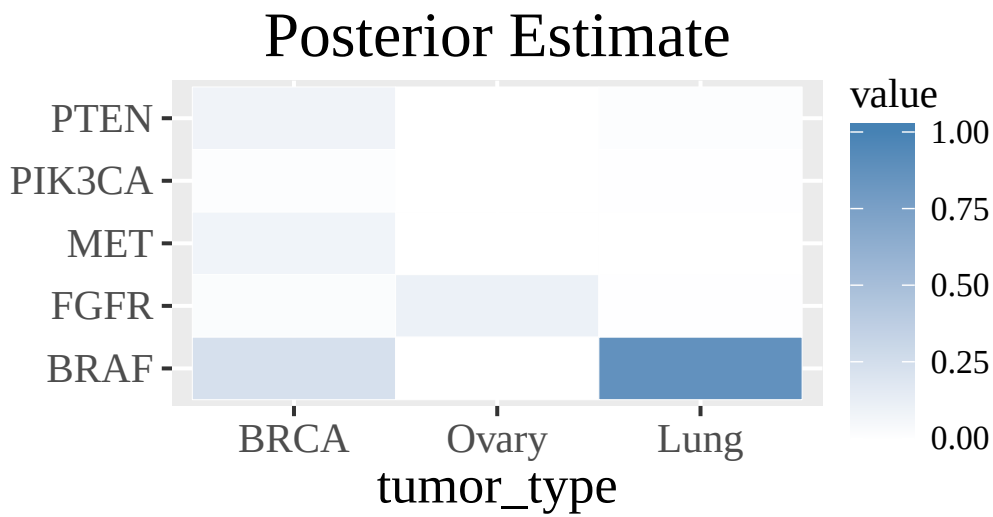
Simulation

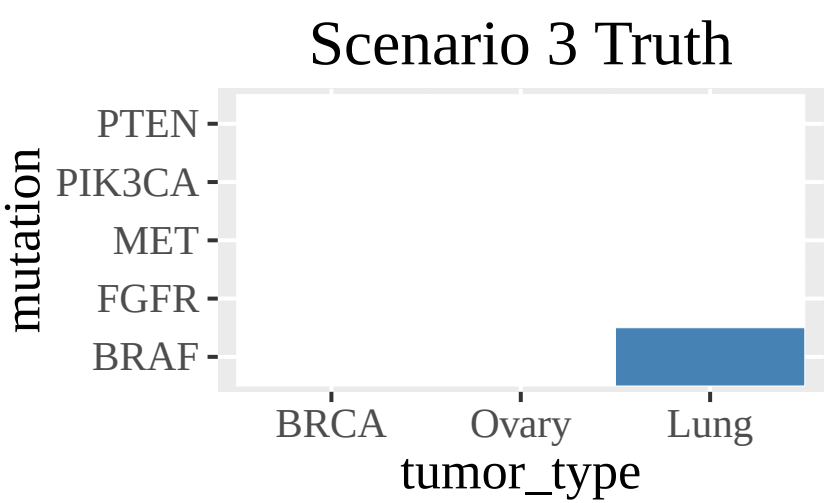
6 scenarios: overall treatment effect (trt);
interaction $z \times$ mutation $\times$ tumor,

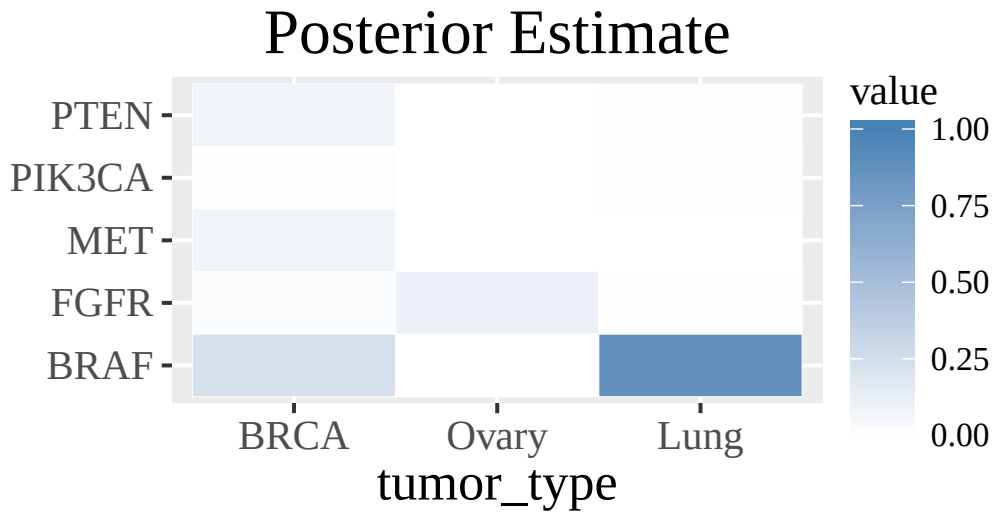
$z \in \{0, 1\}$, mutation $\in \{$BRAF, PIK3CA, PTEN$\}$,
tumors $\in \{$BRCA, Lung, Ovary$\}$.

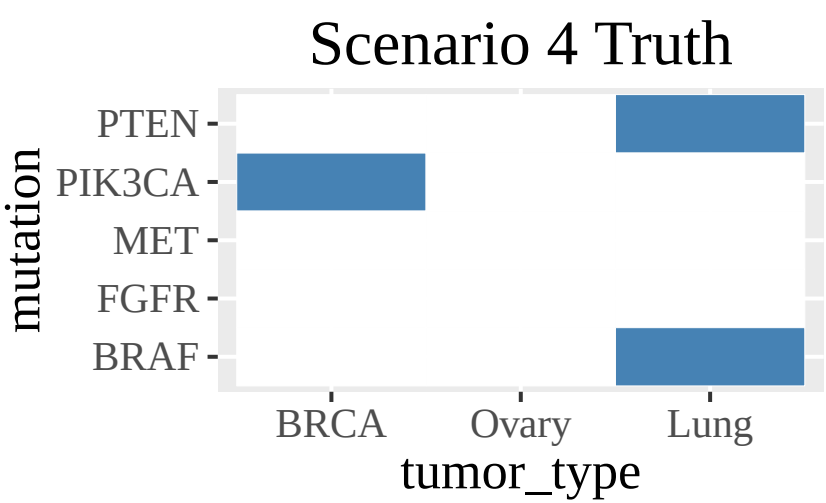
	trt	Interactions (coefficient)
H_0	0	none
H_1	0.4	none
3	0	BRAF*Lung*z (0.4)
4	0	PIK3CA*BRCA*z (0.3), BRAF*Lung*z (0.3) PTEN*Lung*z(0.4)
5	0	PIK3CA*BRCA*z (0.3), BRAF*Ovary*z (0.4) BRAF*Lung*z(0.3)
6	0	BRAF*BRCA(0.4), BRAF*Ovary*z (0.3), BRAF*Lung*z(0.4)

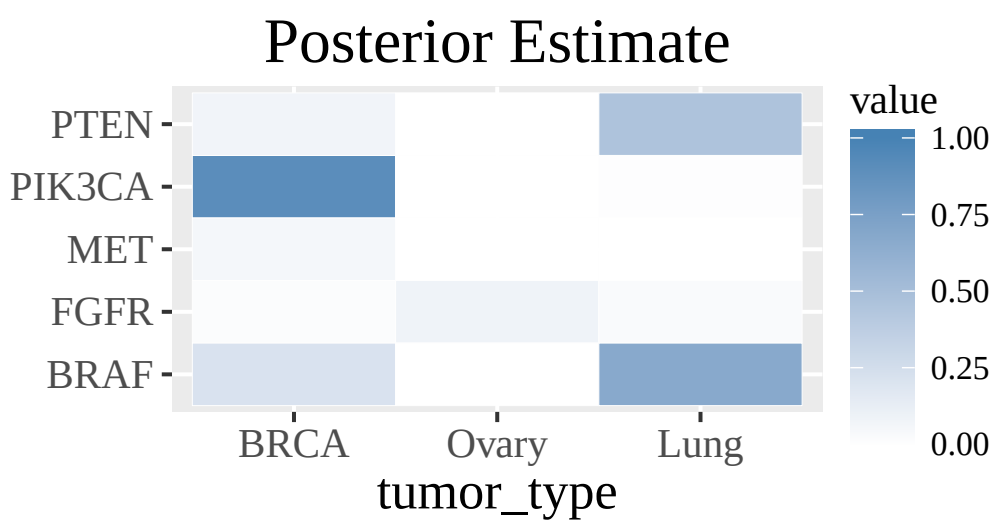

Scenario 3 Truth
mutation
PTEN
PIK3CA
MET
FGFR
BRAF
BRCA
Ovary
Lung
tumor_type


Posterior Estimate
PTEN
PIK3CA
MET
FGFR
BRAF
BRCA
Ovary
Lung
tumor_type
value
1.00
0.75
0.50
0.25
0.00

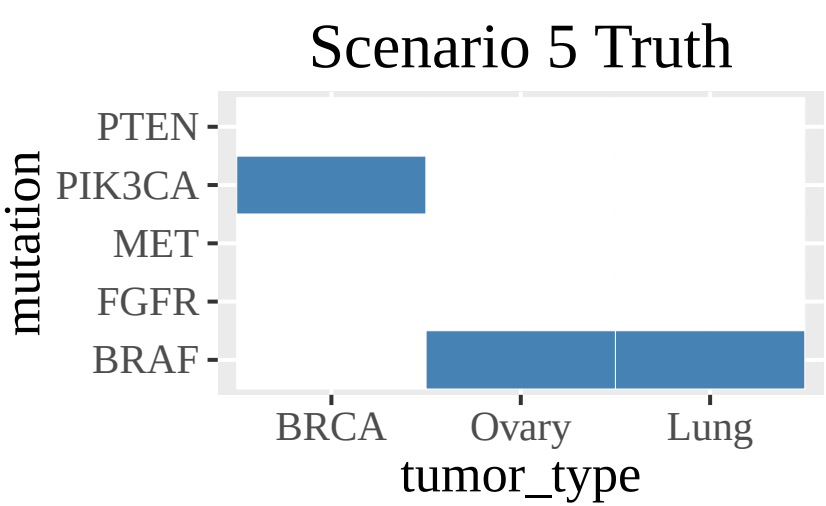


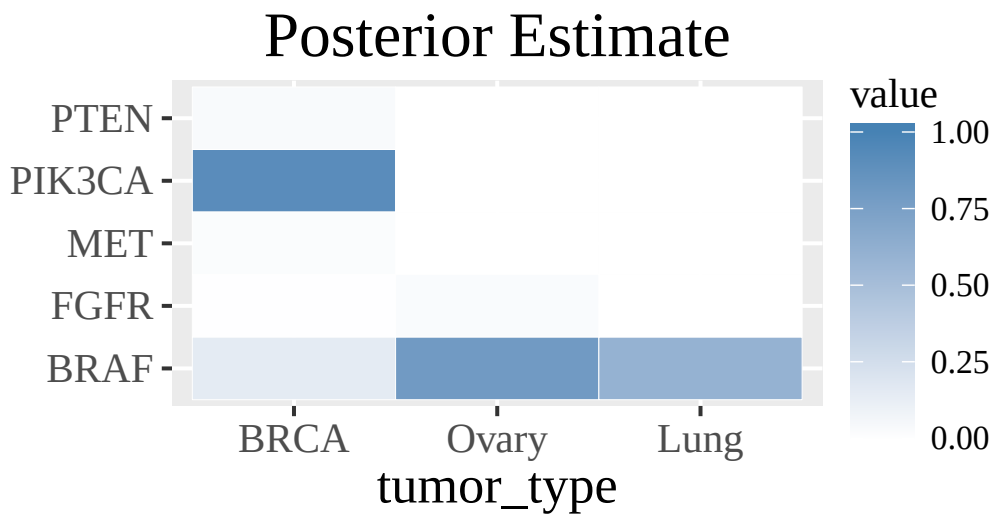


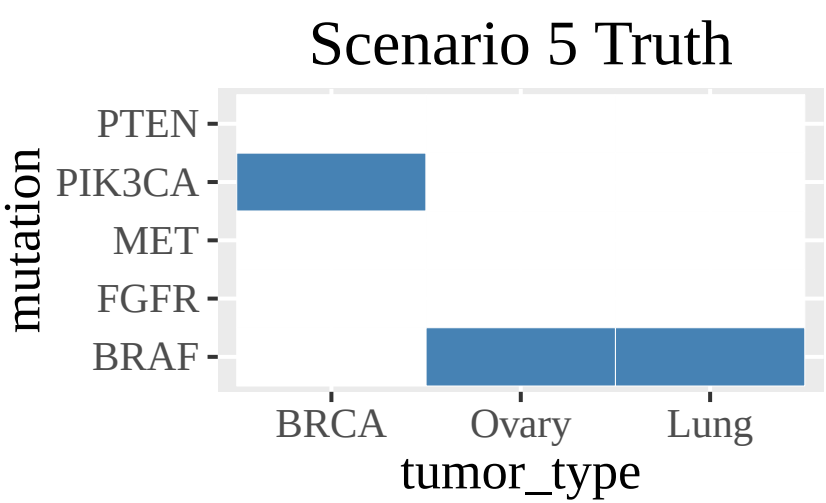


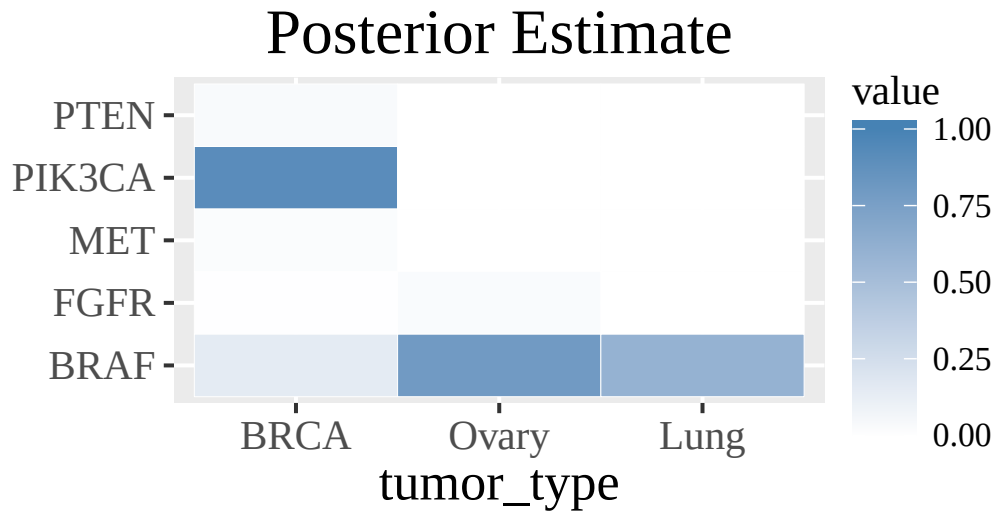


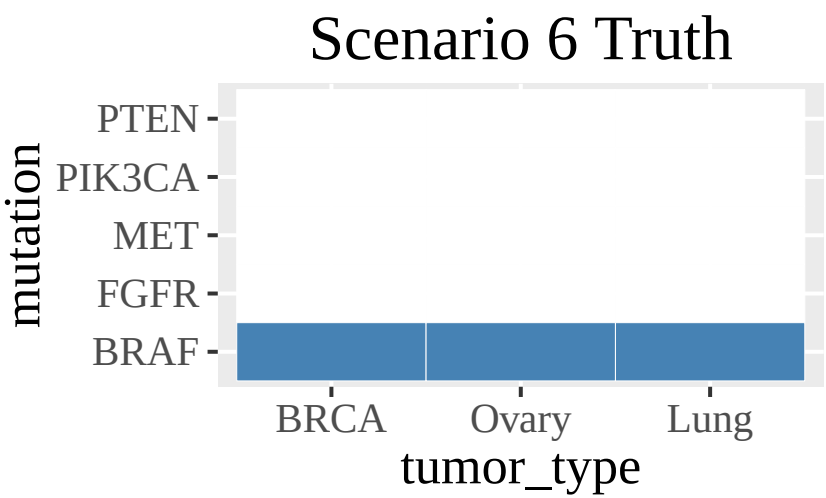
left = truth; right = estimate as $p(a)$ over repeat sim.

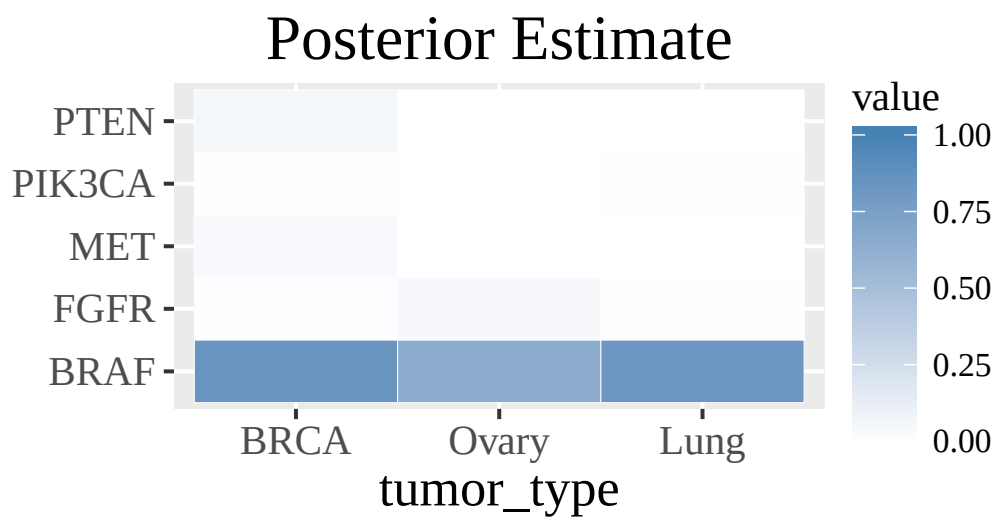

Scenario 5 Truth
mutation
PTEN
PIK3CA
MET
FGFR
BRAF
BRCA
Ovary
Lung
tumor_type


Posterior Estimate
PTEN
PIK3CA
MET
FGFR
BRAF
BRCA
Ovary
Lung
tumor_type
value
1.00
0.75
0.50
0.25
0.00









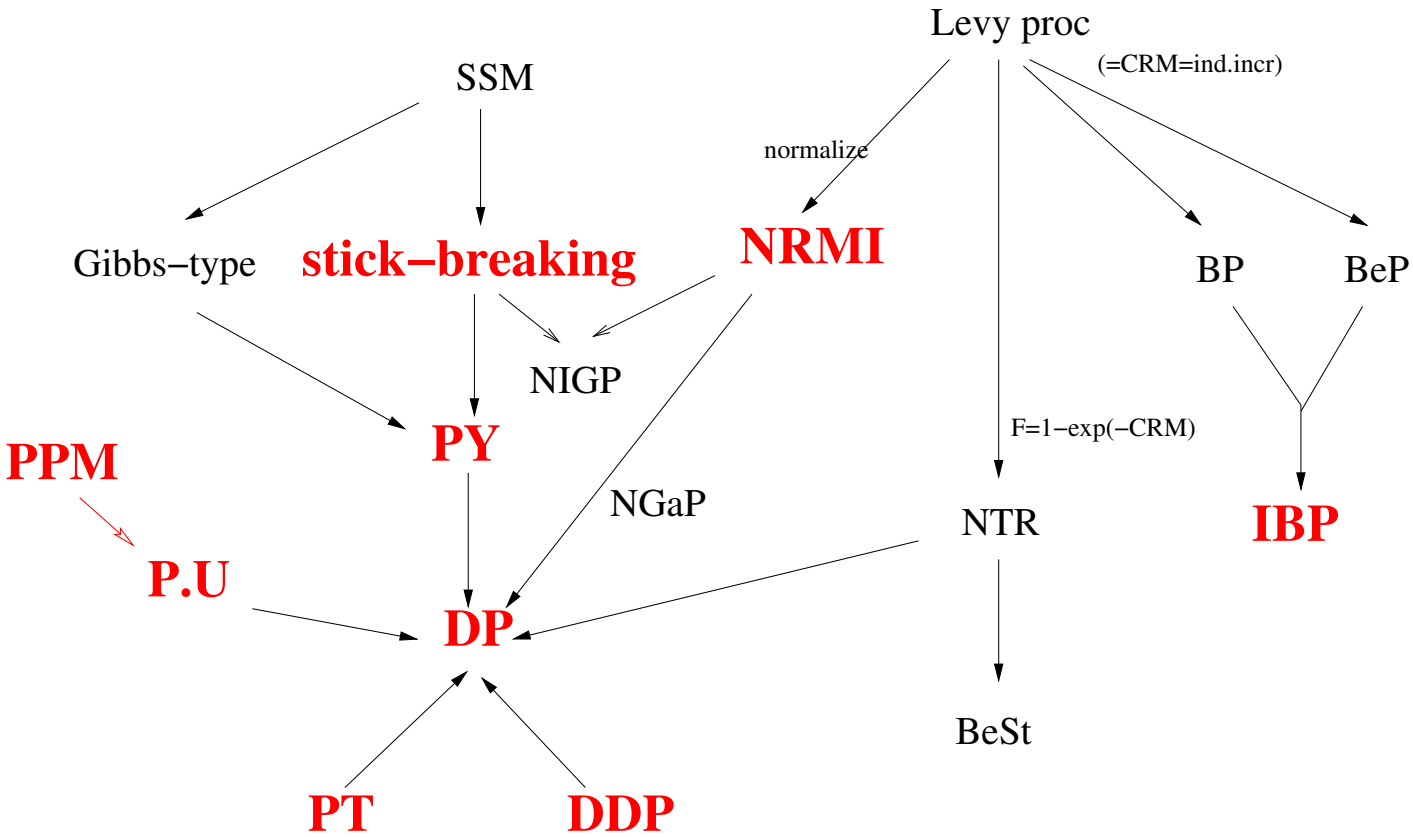
left = truth; right = estimate as $p(a)$ over repeat sim.

Summary

- **Definition:** BNP = prob models for infinite dim parameters.

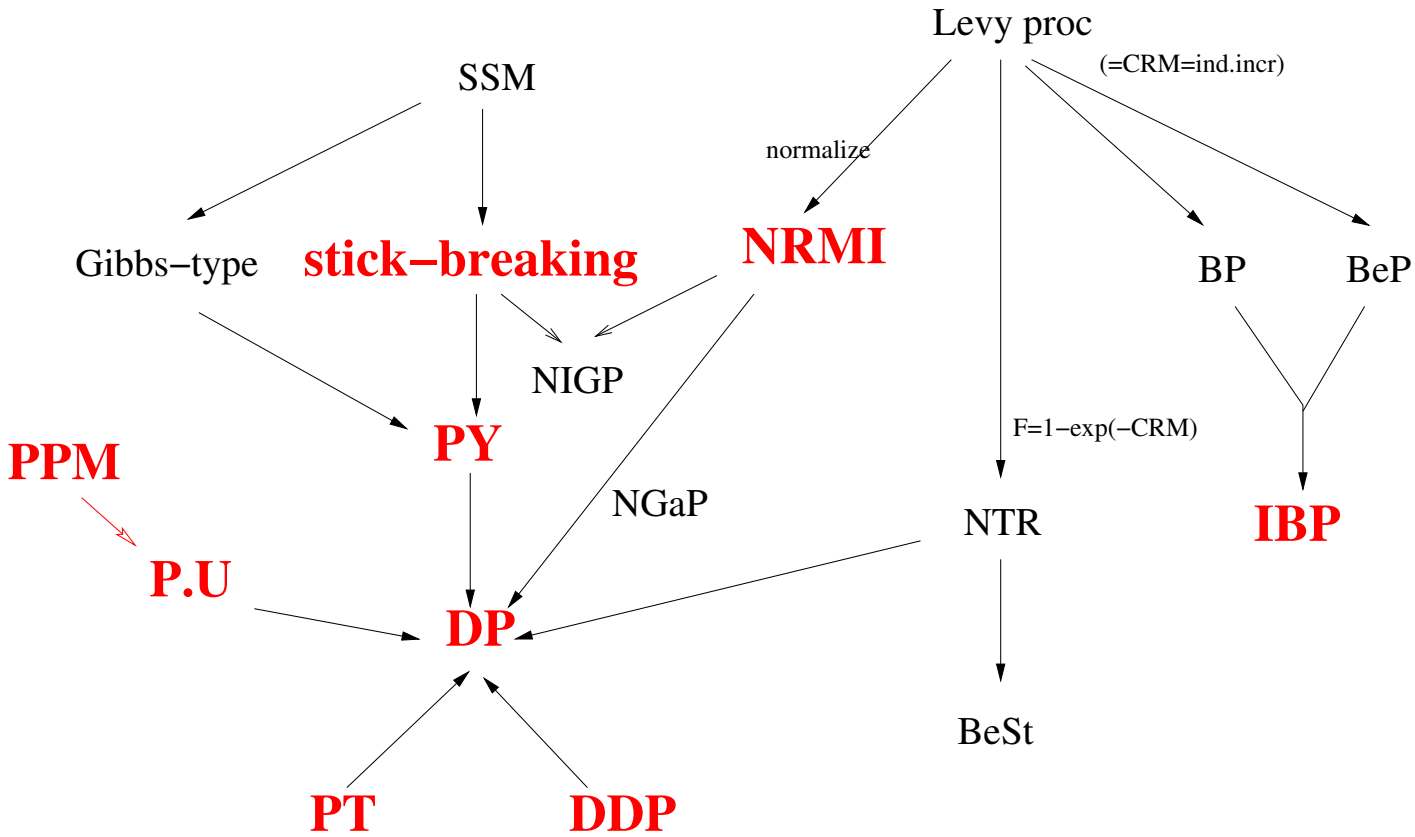
Summary

- **Definition:** BNP = prob models for infinite dim parameters.



Summary

- **Definition:** BNP = prob models for infinite dim parameters.



- Flexible models for full probabilistic description of all uncertainties
- Computation intensive; nonsense in – rubbish out :-)