Covariate-Dependent Bayesian Clustering

Peter Müller, M.D. Anderson Cancer Center
Fernando Quintana, Pontifica Universidad Catolica de Chile (PUCC)
Gary Rosner, MDACC

Outline

1. Intro
   - Example
   - Notation
   - Random partition models w/o covariates
2. A covariate dependent PPM
   - Model
   - Similarity function
3. Posterior inference
4. A mixed model with covariate-dependent clustering
5. A survival time model with covariate-dependent clustering

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

- Match her with already observed patients, using clustering

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

- Match her with already observed patients, using clustering
- predict a similar response

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

- Match her with already observed patients, using clustering
- predict a similar response

Clustering: Cluster patients into subsets of similar responses (easy, but useless for prediction).

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

- Match her with already observed patients, using clustering
- predict a similar response

Clustering: Cluster patients into subsets of similar responses (easy, but useless for prediction).

Two patients with similar covariates (dose) should be a priori more likely to co-cluster (not so easy).

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

- Match her with already observed patients, using clustering
- predict a similar response

Clustering: Cluster patients into subsets of similar responses (easy, but useless for prediction).
Two patients with similar covariates (dose) should be a priori more likely to co-cluster (not so easy).

Note: For predictive inference we need a prob model on the clustering.

Intro – Motivating Example

Study: chemo-immunotherapy for ovarian cancer patients

Data: Monocyte counts over time (y_{ij}); patient covariates (x_i), including treatment dose, $i = 1, \ldots, n$.

Prediction: Want to formalize prediction for next patient $n + 1$:

- Match her with already observed patients, using clustering
- predict a similar response

Clustering: Cluster patients into subsets of similar responses (easy, but useless for prediction).
Two patients with similar covariates (dose) should be a priori more likely to co-cluster (not so easy).

Note: For predictive inference we need a prob model on the clustering.
Deterministic heuristic clustering algorithms will not do.

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \ldots S_k$,
clusters (partitioning subsets) $S_i \subset S$,

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \ldots S_k$,
clusters (partitioning subsets) $S_i \subset S$,
Re-parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \ldots S_k$,
clusters (partitioning subsets) $S_i \subset S$,
Re-parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$

Random Partition Models – Data

Responses $y^n = (y_1, \ldots, y_n)$;

Random Partition Models – Data

Responses $y^n = (y_1, \ldots, y_n)$;
Covariates $x^n = (x_1, \ldots, x_n)$

Random Partition Models – Data

Responses $y^n = (y_1, \ldots, y_n)$;
Covariates $x^n = (x_1, \ldots, x_n)$
y and x by cluster: : $x_j^* = \{x_i;\ i \in S_j\}$ and $y_j^* = \ldots$

Random Partition Models w/o Covariates

Sampling model: conditional on partition ρ_n, assume exchangeability,

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

with cluster-specific parameters θ_j

Random Partition Models w/o Covariates

Sampling model: conditional on partition ρ_n, assume exchangeability,

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

with cluster-specific parameters θ_j

Prior $p(\theta_j)$: conjugate ...

Random Partition Models w/o Covariates

Sampling model: conditional on partition ρ_n, assume exchangeability,

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

with cluster-specific parameters θ_j

Prior $p(\theta_j)$: conjugate ...

Prior $p(\rho)$: PPM, SSM, model-based clustering etc. $\longrightarrow$ next slide

Random Partition Models

Product partition model (PPM): Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA), Crowley (1997 JASA), Quintana (2006 JSPI)

cohesion functions $c(S_j)$ define similarity of a cluster,

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

together with the sampling model (*)

Species sampling model (SSM): Pitman (1996), Ishwaran and James (2003 Stat Sinica)

$p(\rho)$ depends on S indirectly through $|S_j|$:

$$p(\rho_n) = p(|S_1|, \ldots, |S_k|).$$

Species sampling model (SSM): Pitman (1996), Ishwaran and James (2003 Stat Sinica)

$p(\rho)$ depends on S indirectly through $|S_j|$:

$$p(\rho_n) = p(|S_1|, \ldots, |S_k|).$$

Alternative characterization by predictive prob function (PPF):

$$p(s_{i+1} = j \mid s_1, \ldots, s_i)$$

Species sampling model (SSM): Pitman (1996), Ishwaran and James (2003 Stat Sinica)

$p(\rho)$ depends on S indirectly through $|S_j|$:

$$p(\rho_n) = p(|S_1|, \ldots, |S_k|).$$

Alternative characterization by predictive prob function (PPF):

$$p(s_{i+1} = j \mid s_1, \ldots, s_i)$$

careful! PPF is not arbitrary, subject to constraints (EEPF).

Species sampling model (SSM): Pitman (1996), Ishwaran and James (2003 Stat Sinica)

$p(\rho)$ depends on S indirectly through $|S_j|$:

$$p(\rho_n) = p(|S_1|, \ldots, |S_k|).$$

Alternative characterization by predictive prob function (PPF):

$$p(s_{i+1} = j \mid s_1, \ldots, s_i)$$

careful! PPF is not arbitrary, subject to constraints (EEPF). Otherwise the implied joint might not be exchangeable.

Random Partition Models (ctd.)

Model based clustering: Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

implicitely define $p(\rho)$ by

$$p(y_i \mid k, \theta_1, \ldots, \theta_k, \pi_{k1}, \ldots, \pi_{kk}) = \sum_{j=1}^{k} \pi_{kj} f_j(y_i \mid \theta_j),$$

Random Partition Models (ctd.)

Model based clustering: Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

implicitely define $p(\rho)$ by

$$p(y_i \mid k, \theta_1, \ldots, \theta_k, \pi_{k1}, \ldots, \pi_{kk}) = \sum_{j=1}^{k} \pi_{kj} f_j(y_i \mid \theta_j),$$

and equivalent hierarchical model with latent s_i:

$$p(y_i \mid s_i = j, \theta, k) = f_j(y_i \mid \theta_j)$$
$$Pr(s_i = j) = \pi_j$$

Polya urn: predictive rule; let $k_i =$ no. clusters among $\{1, \ldots, i\}$.

$$p(s_{i+1} \mid s_1, \ldots, i) = \begin{cases} s_h & \text{with prob } 1/(M+i) \\ k_i + 1 & \text{with prob } M/(M+i) \end{cases}$$

Polya urn: predictive rule; let $k_i =$ no. clusters among $\{1, \ldots, i\}$.

$$p(s_{i+1} \mid s_1, \ldots, i) = \begin{cases} s_h & \text{with prob } 1/(M+i) \\ k_i + 1 & \text{with prob } M/(M+i) \end{cases}$$

Notes

- This is the clustering model implied by random sampling from an unknown discrete G with DP prior:

$$y_i \sim G \text{ and } G \sim DP(G^\star, M)$$

Polya urn: predictive rule; let k_i = no. clusters among $\{1, \ldots, i\}$.

$$p(s_{i+1} \mid s_1, \ldots, i) = \begin{cases} s_h & \text{with prob } 1/(M+i) \\ k_i + 1 & \text{with prob } M/(M+i) \end{cases}$$

Notes

- This is the clustering model implied by random sampling from an unknown discrete G with DP prior:

$$y_i \sim G \text{ and } G \sim DP(G^\star, M)$$

- The P.U. is a PPM with $c(|S_i|) = \ldots$

Polya urn: predictive rule; let k_i = no. clusters among $\{1, \ldots, i\}$.

$$p(s_{i+1} \mid s_1, \ldots, i) = \begin{cases} s_h & \text{with prob } 1/(M+i) \\ k_i + 1 & \text{with prob } M/(M+i) \end{cases}$$

Notes

- This is the clustering model implied by random sampling from an unknown discrete G with DP prior:

$$y_i \sim G \text{ and } G \sim DP(G^\star, M)$$

- The P.U. is a PPM with $c(|S_i|) = \ldots$
- The P.U. is a SSM with ....

Polya urn: predictive rule; let k_i = no. clusters among $\{1, \ldots, i\}$.

$$p(s_{i+1} \mid s_1, \ldots, i) = \begin{cases} s_h & \text{with prob } 1/(M+i) \\ k_i + 1 & \text{with prob } M/(M+i) \end{cases}$$

Notes

- This is the clustering model implied by random sampling from an unknown discrete G with DP prior:

$$y_i \sim G \text{ and } G \sim DP(G^\star, M)$$

- The P.U. is a PPM with $c(|S_i|) = \ldots$
- The P.U. is a SSM with ....
- It is a special (limiting) case of model-based clustering.

Covariate-dependent PPM

Similarity function: 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

Similarity function: 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:

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

Covariate-dependent PPM

Similarity function: 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:

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

with norm. const. $g_n(x^n) = \sum_\rho \prod_{j=1}^k g(x_j^\star) c(S_j)$

Covariate-dependent PPM

Similarity function: 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:

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

with norm. const. $g_n(x^n) = \sum_\rho \prod_{j=1}^k g(x_j^\star) c(S_j)$

Natural choice:

define $g(x_j^\star)$ as (auxiliary) exchangeable prob. model $q(\cdot)$:

$$g(x_j^\star) \equiv \int \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)\, d\xi_j$$

Covariate-dependent PPM: Desiderata

Symmetry: $p(\rho_n \mid x^\star)$ is invariant w.r.t. permutations of the indices $i = 1, \ldots, n$, i.e., $p(\rho_n \mid x^\star)$ does not depend on the order of experiments.

Covariate-dependent PPM: Desiderata

Symmetry: $p(\rho_n \mid x^\star)$ is invariant w.r.t. permutations of the indices $i = 1, \ldots, n$, i.e., $p(\rho_n \mid x^\star)$ does not depend on the order of experiments.

Averaging: $g(x^\star) = \int g(x^\star, x) dx$

Covariate-dependent PPM: Desiderata

Symmetry: $p(\rho_n \mid x^\star)$ is invariant w.r.t. permutations of the indices $i = 1, \ldots, n$, i.e., $p(\rho_n \mid x^\star)$ does not depend on the order of experiments.

Averaging: $g(x^\star) = \int g(x^\star, x)dx$

$\Rightarrow$ Similarity function must be of form

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

Covariate-dependent PPM: Desiderata

Symmetry: $p(\rho_n \mid x^\star)$ is invariant w.r.t. permutations of the indices $i = 1, \ldots, n$, i.e., $p(\rho_n \mid x^\star)$ does not depend on the order of experiments.

Averaging: $g(x^\star) = \int g(x^\star, x)dx$

$\Rightarrow$ Similarity function must be of form

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

$\Rightarrow$ Coherence across n:

$$p(\rho_n \mid x^n) = \sum_{s_{n+1}} \int p(\rho_{n+1} \mid x^n, x_{n+1})$$

$$q(x_{n+1} \mid x^n)\, dx_{n+1}$$

Covariate-dependent PPM: Desiderata

Symmetry: $p(\rho_n \mid x^\star)$ is invariant w.r.t. permutations of the indices $i = 1, \ldots, n$, i.e., $p(\rho_n \mid x^\star)$ does not depend on the order of experiments.

Averaging: $g(x^\star) = \int g(x^\star, x)dx$

$\Rightarrow$ Similarity function must be of form

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

$\Rightarrow$ Coherence across n:

$$p(\rho_n \mid x^n) = \sum_{s_{n+1}} \int p(\rho_{n+1} \mid x^n, x_{n+1}) \; q(x_{n+1} \mid x^n)\, dx_{n+1}$$

for $q(x_{n+1} \mid x^n) = g_{n+1}(x^{n+1})/g_n(x^n)$.

Covariate-dependent PPM (ctd.)

Similarity function: $g(x_j^\star) = \int \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j) \mathrm{d}\xi_j$

defaults for common data formats

Covariate-dependent PPM (ctd.)

Similarity function: $g(x_j^\star) = \int \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j) \mathrm{d}\xi_j$

defaults for common data formats

Continuous covariates: use

$q(x_i \mid \xi_i) = N(\xi_j, V)$ and $q(\xi_j) = N(\ldots)$.

Covariate-dependent PPM (ctd.)

Similarity function: $g(x_j^\star) = \int \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j) \mathrm{d}\xi_j$
defaults for common data formats

Continuous covariates: use
$q(x_i \mid \xi_i) = N(\xi_j, V)$ and $q(\xi_j) = N(\ldots)$.
Categorical: Multinomial and Dirichlet
Ordinal: multinomial probit model with fixed cutoffs and
(conditionally conjugate) normal prior.
Counts: $q(x_i \mid \xi_j) = Poi(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$,

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k;

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$.

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$. If all is done carefully, reinvent the proposed PPM

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$. If all is done carefully, reinvent the proposed PPM

Modify PPF: modify predictive probability function for the desired regression

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$. If all is done carefully, reinvent the proposed PPM

Modify PPF: modify predictive probability function for the desired regression

$$p(s_{i+1} \mid s_1, \ldots, s_i) = \begin{cases} j & \text{with prob } \propto 1/(M+i)\ q(x_j^\star, x_{i+1}) \\ k_i + 1 & \text{with prob } \propto M/(M+i)\ q(x_{i+1}) \end{cases}$$

for some "similarity" q

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$. If all is done carefully, reinvent the proposed PPM

Modify PPF: modify predictive probability function for the desired regression

$$p(s_{i+1} \mid s_1, \ldots, s_i) = \begin{cases} j & \text{with prob } \propto 1/(M+i)\ q(x_j^\star, x_{i+1}) \\ k_i + 1 & \text{with prob } \propto M/(M+i)\ q(x_{i+1}) \end{cases}$$

for some "similarity" q

- verboten!

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$. If all is done carefully, reinvent the proposed PPM

Modify PPF: modify predictive probability function for the desired regression

$$p(s_{i+1} \mid s_1, \ldots, s_i) = \begin{cases} j & \text{with prob } \propto 1/(M+i)\ q(x_j^\star, x_{i+1}) \\ k_i + 1 & \text{with prob } \propto M/(M+i)\ q(x_{i+1}) \end{cases}$$

for some "similarity" q

- verboten! Not clear that $\exists$ joint prob model $p(s)$

Alternative Constructions

Augmented REsponse Vector: replace y_i by $\tilde{y} = (y_i, x_i)$, and use posterior predictive $p(y_{n+1} \mid x_{n+1}, y^n, x^n)$.

- wrong likelihood with extra factor $p(x_i \mid \theta_{s_i})$
- nothing easier

Mixture model with covariate-dependent weights: covariate dependent weights $\pi_j = \pi_j(x_i)$.

- parameters ξ_j for $\pi_j(x_i;\ \xi_j)$
- fixed size k; could be fixed by considering $k \to \infty$. If all is done carefully, reinvent the proposed PPM

Modify PPF: modify predictive probability function for the desired regression

$$p(s_{i+1} \mid s_1, \ldots, s_i) = \begin{cases} j & \text{with prob } \propto 1/(M+i)\ q(x_j^\star, x_{i+1}) \\ k_i + 1 & \text{with prob } \propto M/(M+i)\ q(x_{i+1}) \end{cases}$$

for some "similarity" q

- verboten! Not clear that $\exists$ joint prob model $p(s)$
- but still works very fine

Posterior Inference – w/o Covariates

Lau and Green (2007 JCGS)

- stochastic search by MCMC; useful for predictive inference

Posterior Inference – w/o Covariates

Lau and Green (2007 JCGS)

- stochastic search by MCMC; useful for predictive inference
- The MAP model is not necessarily representative

Posterior Inference – w/o Covariates

Lau and Green (2007 JCGS)

- stochastic search by MCMC; useful for predictive inference
- The MAP model is not necessarily representative
- Inference with a loss function (e.g., Quintana and Iglesias, 2003 JRSSB).

Posterior Inference – w/o Covariates

Lau and Green (2007 JCGS)

- stochastic search by MCMC; useful for predictive inference
- The MAP model is not necessarily representative
- Inference with a loss function (e.g., Quintana and Iglesias, 2003 JRSSB).
- Alternatively Bayesian hierarchical clustering (Heard et al. 2005 JASA). Iteratively combine clusters to max post prob.

Posterior Inference w. Covariates

- After a model augmentation, computation reduces to a model w/o covariates.

Posterior Inference w. Covariates

- After a model augmentation, computation reduces to a model w/o covariates.
- In words, simply change the interpetation of $g(x_j^*)$ to consider x_i as r.v.'s

Model augmentation:

- augment original model by defining x^n as r.v. with

$$q(x^n, \xi \mid \rho) = \prod_j \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)$$

Model augmentation:

- augment original model by defining x^n as r.v. with

$$q(x^n, \xi \mid \rho) = \prod_j \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)$$

- change $p(\rho_n)$ to PPM w/o covariates: $q(\rho_n) = \prod c(S_j)$

Model augmentation:

- augment original model by defining x^n as r.v. with

$$q(x^n, \xi \mid \rho) = \prod_j \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)$$

- change $p(\rho_n)$ to PPM w/o covariates: $q(\rho_n) = \prod c(S_j)$
- keep $p(y \mid \rho_n, \theta)$ and $p(\theta)$ unchanged.

Model augmentation:

- augment original model by defining x^n as r.v. with

$$q(x^n, \xi \mid \rho) = \prod_j \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)$$

- change $p(\rho_n)$ to PPM w/o covariates: $q(\rho_n) = \prod c(S_j)$
- keep $p(y \mid \rho_n, \theta)$ and $p(\theta)$ unchanged.

$$q(y^n, x^n, \theta, \xi, \rho_n) =$$

$$\prod_j \prod_{i \in S_j} \underbrace{p(y_i \mid \theta_j, x_i)\, q(x_i \mid \xi_j)}_{\text{likelihood for } (y_i, x_i)} \underbrace{p(\theta_j)\, q(\xi_j)}_{\text{indep}} \underbrace{c(S_j)}_{\text{PPM}}$$

Model augmentation:

- augment original model by defining x^n as r.v. with

$$q(x^n, \xi \mid \rho) = \prod_j \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)$$

- change $p(\rho_n)$ to PPM w/o covariates: $q(\rho_n) = \prod c(S_j)$
- keep $p(y \mid \rho_n, \theta)$ and $p(\theta)$ unchanged.

$$q(y^n, x^n, \theta, \xi, \rho_n) =$$

$$\prod_j \prod_{i \in S_j} \underbrace{p(y_i \mid \theta_j, x_i)\, q(x_i \mid \xi_j)}_{\text{likelihood for } (y_i, x_i)} \underbrace{p(\theta_j)\, q(\xi_j)}_{\text{indep}} \underbrace{c(S_j)}_{\text{PPM}}$$

$q(\rho_n \mid x^n, y^n) \propto p(\rho_n \mid x^n, y^n)$, under the proposed model
$\Rightarrow$ straightforward posterior inference

Model augmentation:

- augment original model by defining x^n as r.v. with

$$q(x^n, \xi \mid \rho) = \prod_j \prod_{i \in S_j} q(x_i \mid \xi_j)\, q(\xi_j)$$

- change $p(\rho_n)$ to PPM w/o covariates: $q(\rho_n) = \prod c(S_j)$
- keep $p(y \mid \rho_n, \theta)$ and $p(\theta)$ unchanged.

$$q(y^n, x^n, \theta, \xi, \rho_n) = \prod_j \prod_{i \in S_j} \underbrace{p(y_i \mid \theta_j, x_i)\, q(x_i \mid \xi_j)}_{\text{likelihood for } (y_i, x_i)} \underbrace{p(\theta_j)\, q(\xi_j)}_{\text{indep}} \underbrace{c(S_j)}_{\text{PPM}}$$

$q(\rho_n \mid x^n, y^n) \propto p(\rho_n \mid x^n, y^n)$, under the proposed model
$\Rightarrow$ straightforward posterior inference

Opposite is not true! A model with combined response is not necessarily interpretable as covariate-dependent PPM.

Example: Mixed Effects Model

Treatment: chemotherapy (carboplatinum) with additional immunotherapy (γ-interferon) and GM-CSF. Immunotherapy boosts monocyte count.

Example: Mixed Effects Model

Treatment: chemotherapy (carboplatinum) with additional immunotherapy (γ-interferon) and GM-CSF. Immunotherapy boosts monocyte count.

Data: $n = 47$ patients, with $n_i = 6$ repeat observations each. $x_i =$ dose carbopt; $y_i = (y_{i1}, \ldots, y_{i6})$.

Example: Mixed Effects Model

Treatment: chemotherapy (carboplatinum) with additional immunotherapy (γ-interferon) and GM-CSF. Immunotherapy boosts monocyte count.

Data: $n = 47$ patients, with $n_i = 6$ repeat observations each. $x_i =$ dose carbopt; $y_i = (y_{i1}, \ldots, y_{i6})$.

Model: PPM model with additional covariate dependence.

Example: Mixed Effects Model

Treatment: chemotherapy (carboplatinum) with additional immunotherapy (γ-interferon) and GM-CSF. Immunotherapy boosts monocyte count.

Data: $n = 47$ patients, with $n_i = 6$ repeat observations each.
$x_i =$ dose carbopt; $y_i = (y_{i1}, \ldots, y_{i6})$.

Model: PPM model with additional covariate dependence.

Sampling model: $p(y_i \mid \theta, \rho_n) = N(\theta_j, \Sigma)$

Example: Mixed Effects Model

Treatment: chemotherapy (carboplatinum) with additional immunotherapy (γ-interferon) and GM-CSF. Immunotherapy boosts monocyte count.

Data: $n = 47$ patients, with $n_i = 6$ repeat observations each.
$x_i =$ dose carbopt; $y_i = (y_{i1}, \ldots, y_{i6})$.

Model: PPM model with additional covariate dependence.

Sampling model: $p(y_i \mid \theta, \rho_n) = N(\theta_j, \Sigma)$

Random partition: $p(\rho_n \mid x^n) \propto \prod c(S_j) g(x_j^\star)$ and conjugate prior $p(\theta_j)$.

Example: Mixed Effects Model

Treatment: chemotherapy (carboplatinum) with additional immunotherapy (γ-interferon) and GM-CSF. Immunotherapy boosts monocyte count.

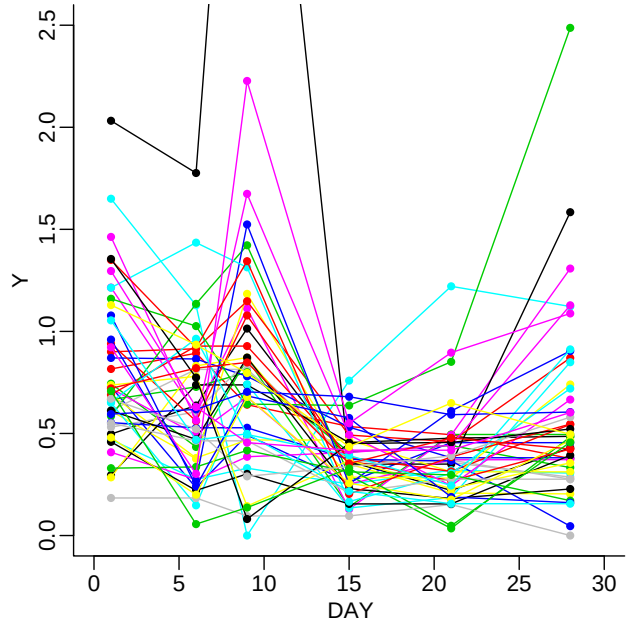
Data: $n = 47$ patients, with $n_i = 6$ repeat observations each.
$x_i =$ dose carbopt; $y_i = (y_{i1}, \ldots, y_{i6})$.

Model: PPM model with additional covariate dependence.

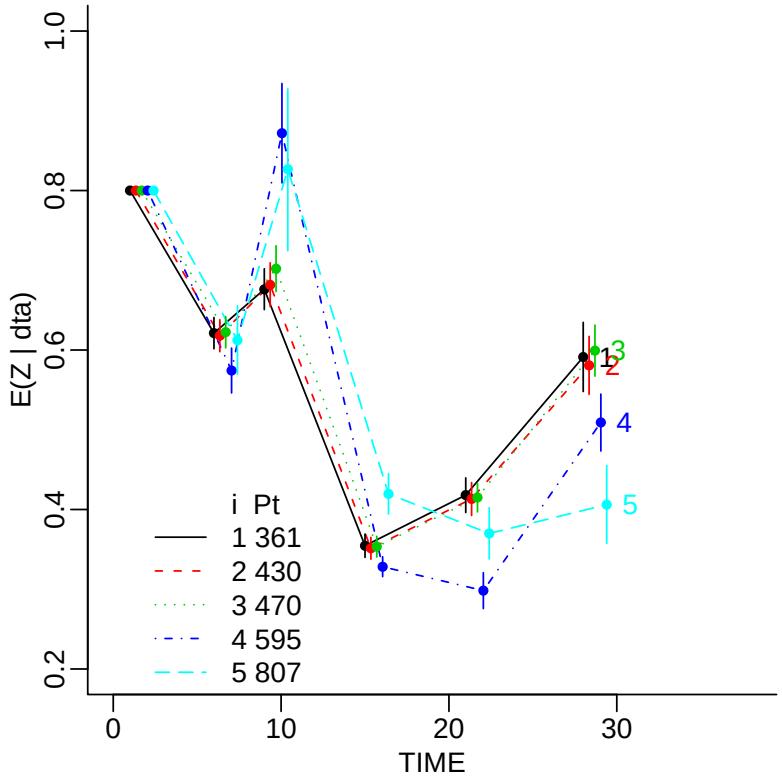
Sampling model: $p(y_i \mid \theta, \rho_n) = N(\theta_j, \Sigma)$

Random partition: $p(\rho_n \mid x^n) \propto \prod c(S_j) g(x_j^\star)$ and conjugate prior $p(\theta_j)$.

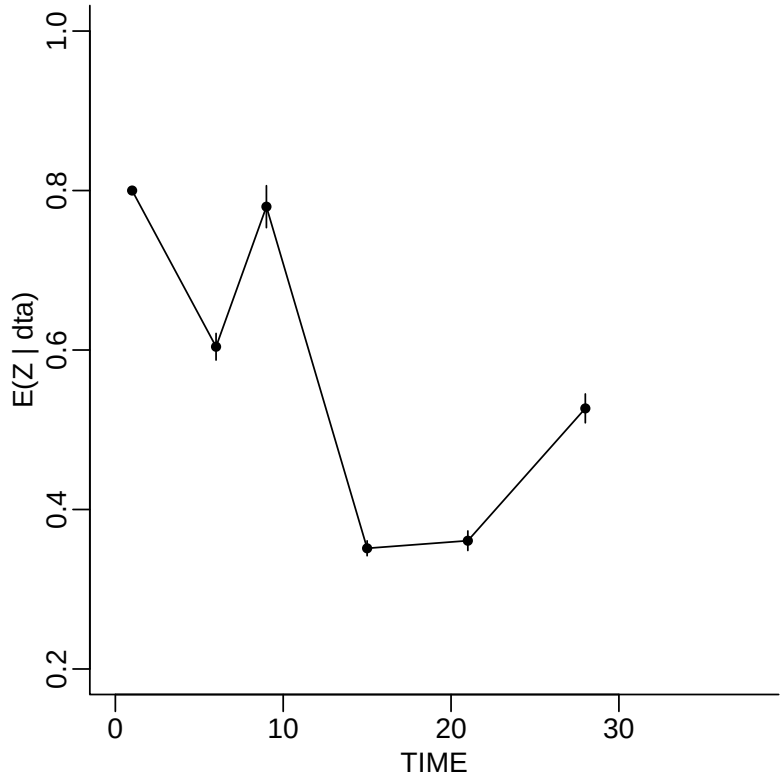
Similarity: $g(x_j^\star) =$ mvn model.



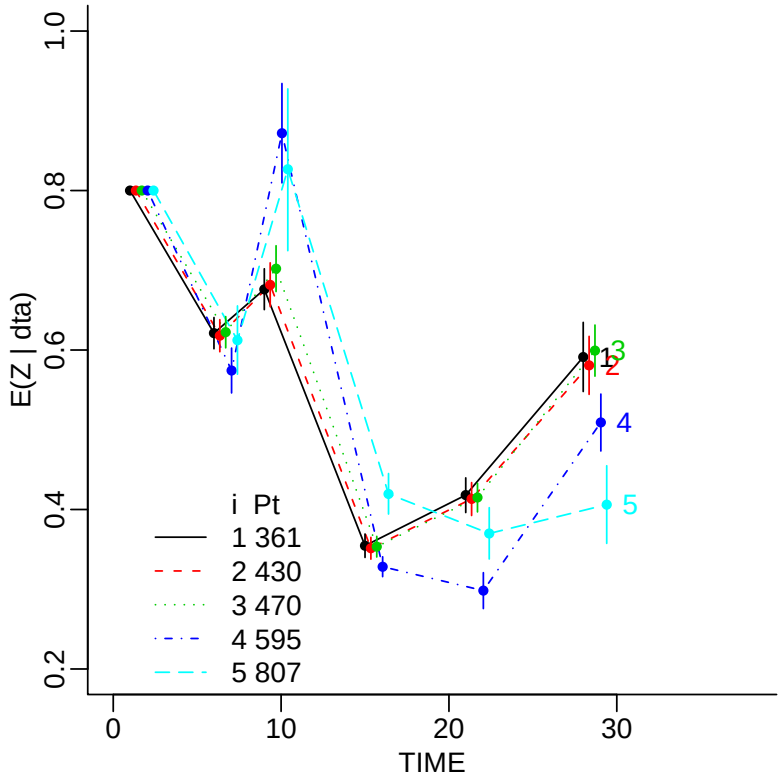
Data. Monocyte count versus day of the first cycle chemotherapy for $n = 47$ patients.

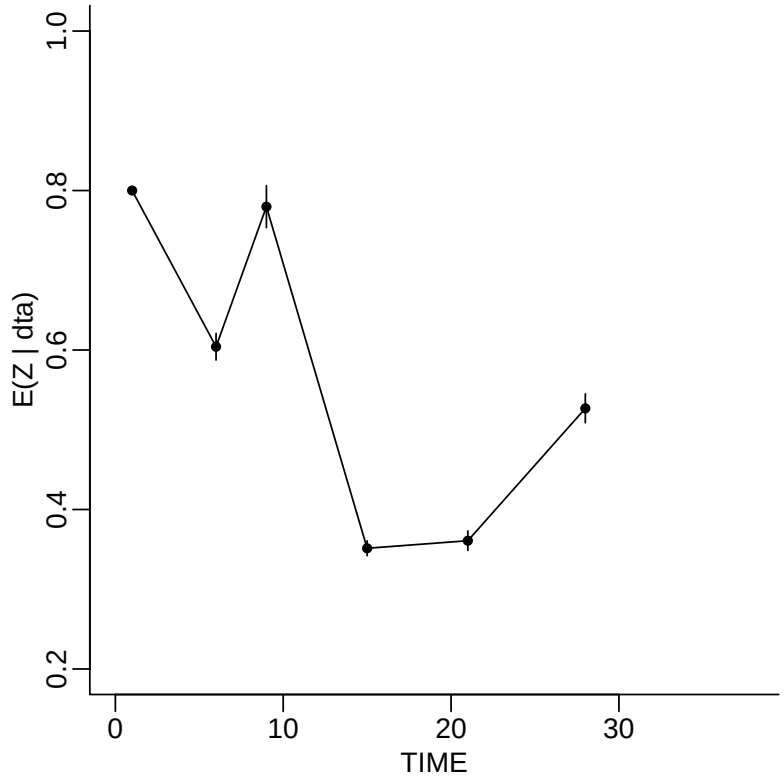


with covariate-dependent clustering



without covariate-dependent clustering

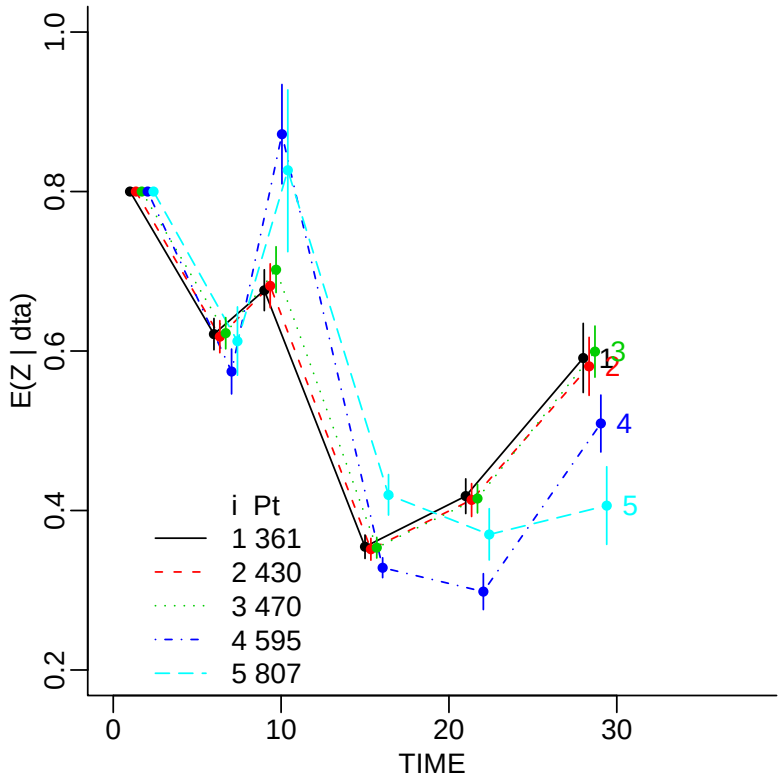


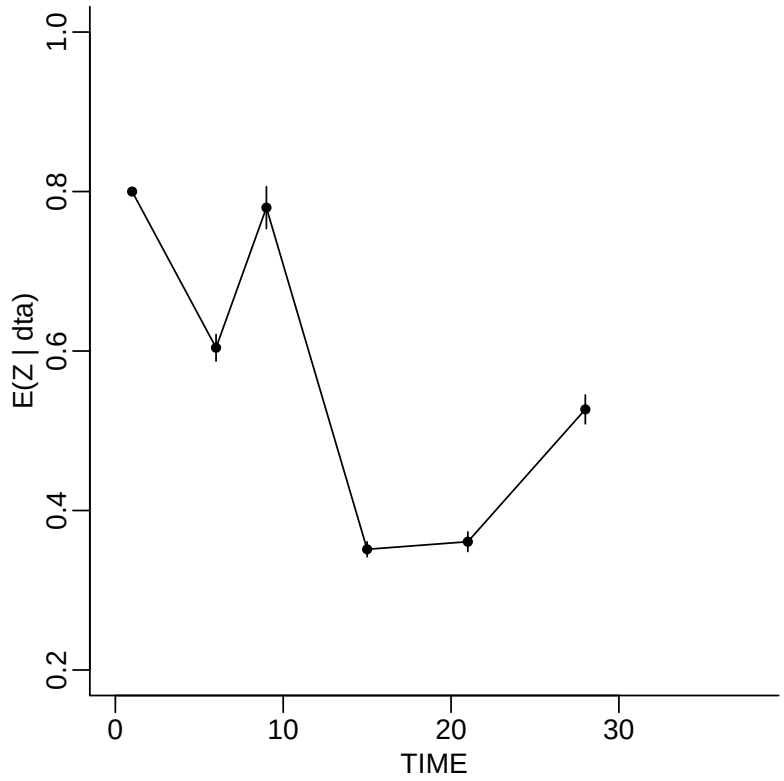


with covariate-dependent clustering

without covariate-dependent clustering

Prediction for $\tilde{y}$ arranged by $\tilde{x}$ (left panel) from lowest ("1") to highest ("5") level of carboplatinum.





with covariate-dependent clustering

without covariate-dependent clustering

Prediction for $\tilde{y}$ arranged by $\tilde{x}$ (left panel) from lowest ("1") to highest ("5") level of carboplatinum.

Without covariates (right panel) prediction is identical for all patients.

Example: Survival Time Model with Clustering

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

Example: Survival Time Model with Clustering

Treatment: high dose (A) versus low dose (B) chemotherapy
Data: 765 patients randomized to A vs. B.

Example: Survival Time Model with Clustering

Treatment: high dose (A) versus low dose (B) chemotherapy
Data: 765 patients randomized to A vs. B.
Response: time until progression or death

Example: Survival Time Model with Clustering

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: Survival Time Model with Clustering

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

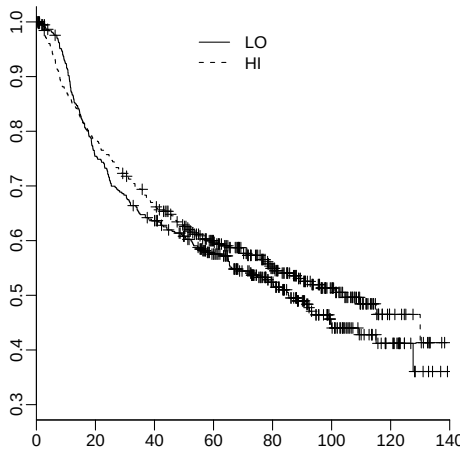
Data: 765 patients randomized to A vs. B.

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

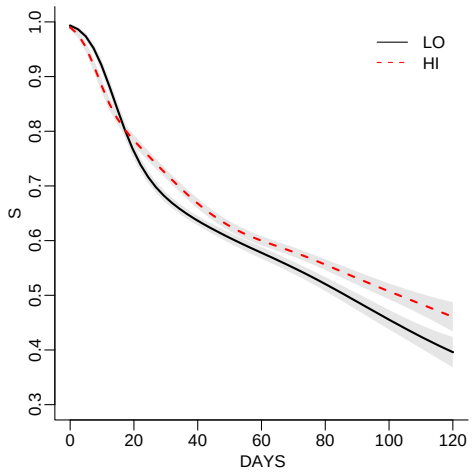
Model: covariate-dependent PPM.

The sampling model is a piecewise exponential model with cluster-specific parameters θ_j^*.

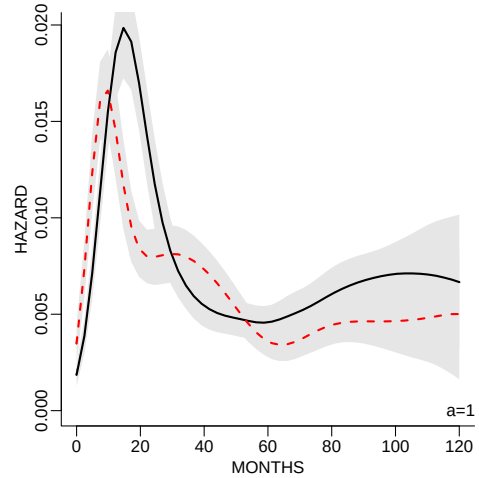


data (KM)

data

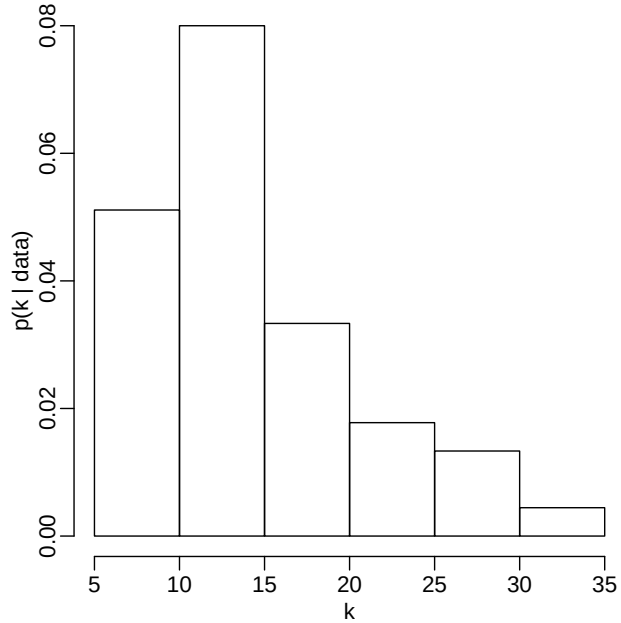


$S(t) \equiv$
$p(y_{n+1} \geq t \mid data)$
estimated survival

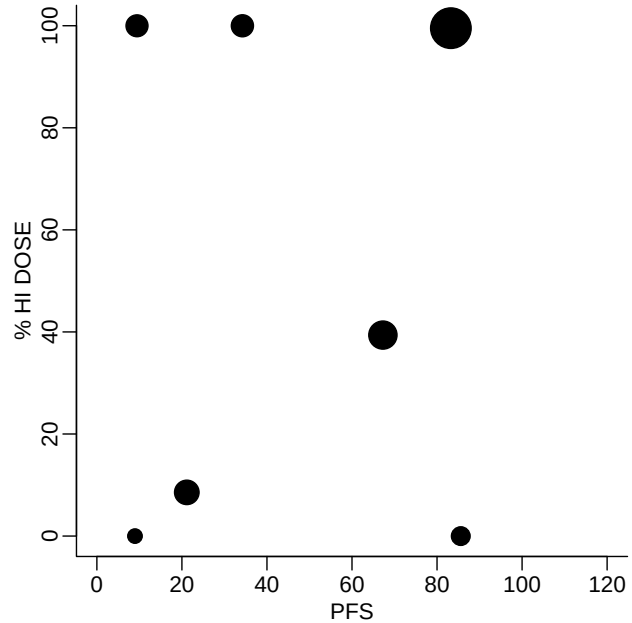


hazard $h(t)$

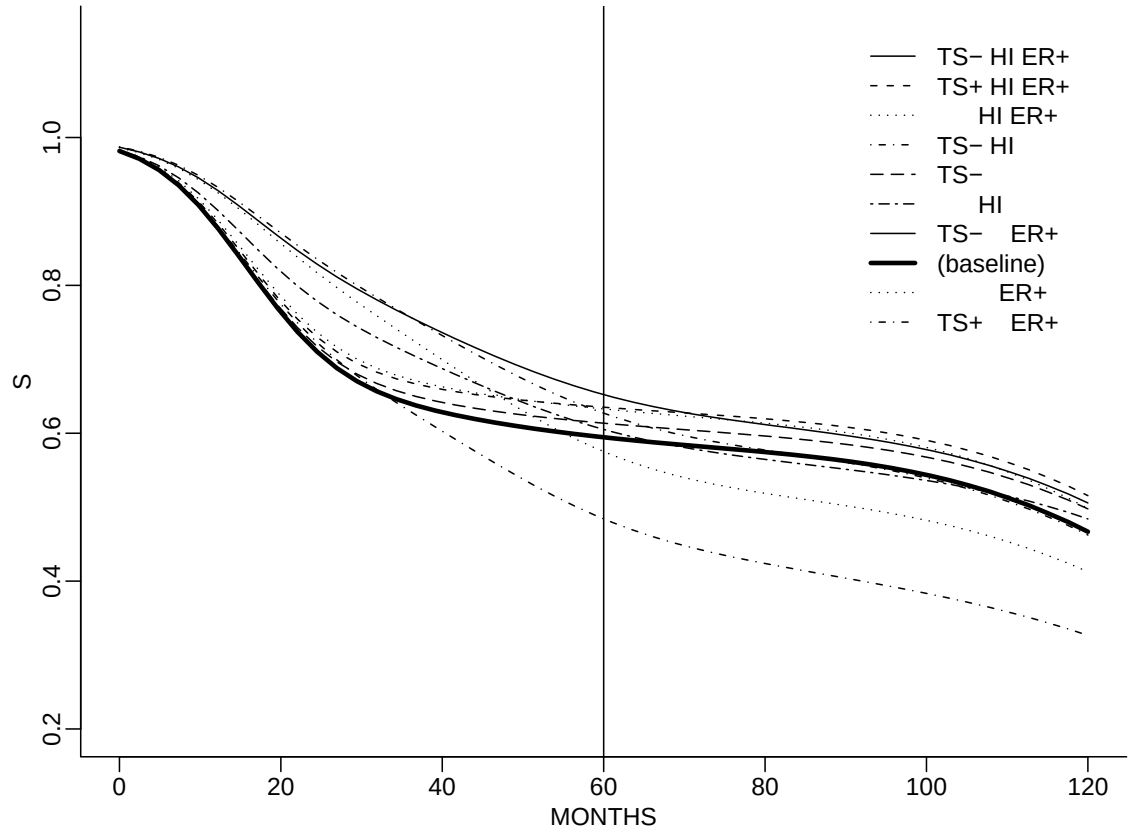
hazard



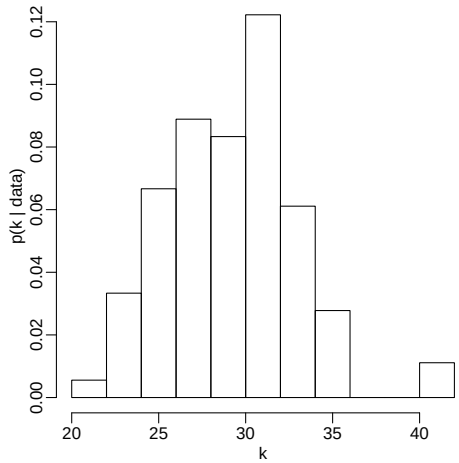
$p(k \mid data)$
n clusters

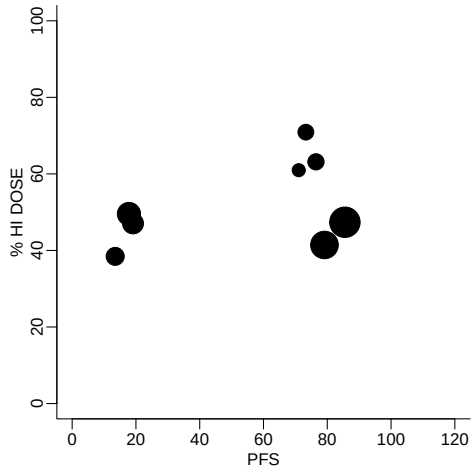


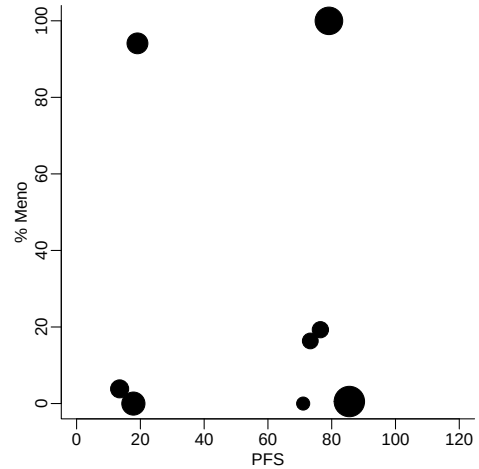
Proportion high dose and mean PFS



$S(t \mid x)$ by covariates







$p(k \mid data)$
n clusters

PFS & % HI

PFS & % Meno

avg covariates by cluster

Summary

Covariate dependent clustering model: proceed as if x were observed, specifying a prob model $q(x_i \mid \xi_i)$ that defines similarity.

Posterior inference: straightforward, includes reweighting of posterior draws for post predictive inference.

Inference summaries: Predictive, Bayes rules for specific loss (Lau and Green).

label switching problem: Inference on a specific terms of the mixture requires resolution of the label switching problem.

Subspace clustering: Hoff (2004 Bayesian Anal) proposes methods that simultanuously select the variables and carries out the clustering.

Covariate-based Clustering vs. Joint Likelihood on (y_i, x_i)

Joint Sampling of (x, y): Include x_i as part of an augmented response vector (x_i, y_i).

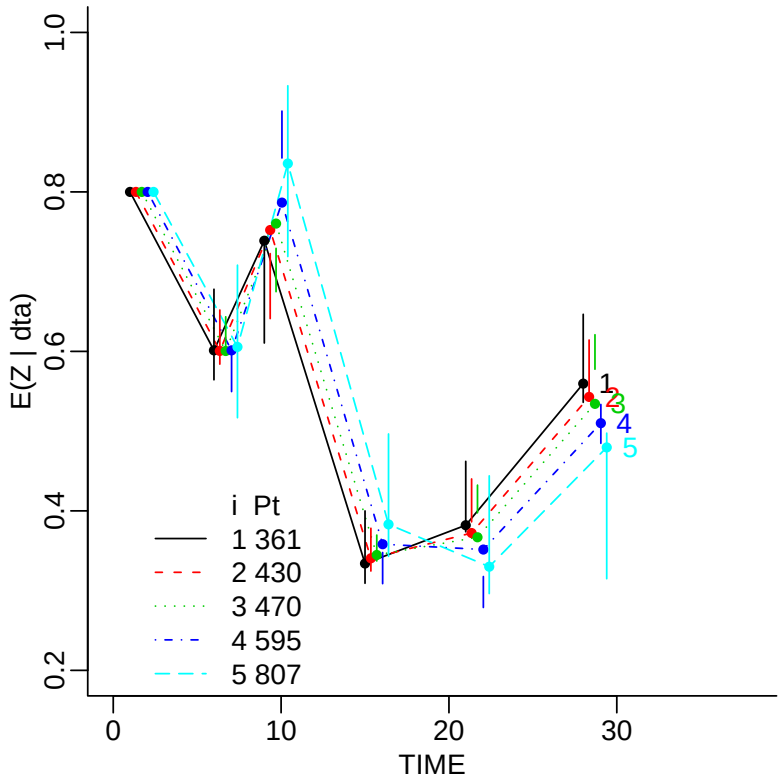
Prediction: becomes inference for a partial response $\tilde{x}$.

Difference: For example 1, e.g., the sampling model would become

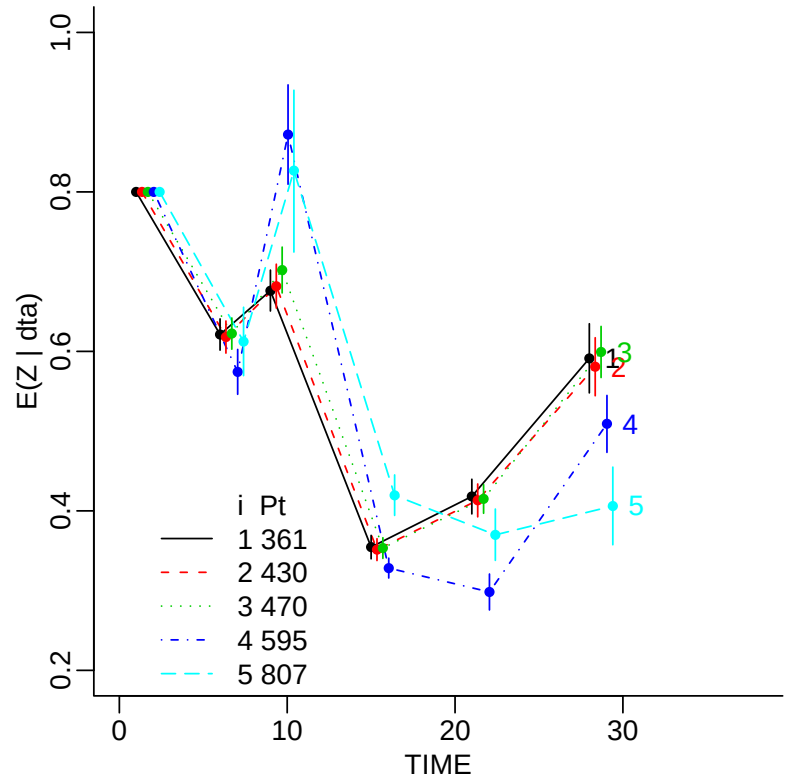
$$(x_i, y_i) \sim N(\theta_i, \Sigma)$$

Including the hyperparameters for x_i leads to a different model that adjusts the similarity functions in undesireable ways.

Posterior predictive inference for a future patient by x_i:



augmented response (x_i, y_i)



covariate-dependent clustering