Covariate-Dependent Bayesian Clustering

Peter Müller & Gary Rosner, M.D. Anderson Cancer Center
Fernando A. Quintana, Pontificia Universidad Católica de Chile

Construction and Properties of Bayesian Nonparametric Regression Models
Cambridge, UK

August 10, 2007

Outline

1. Introduction
   - Two Examples:
     - chemo-immunotherapy for ovarian cancer patients
     - survival time for breast cancer patients
   - Notation
   - Random partition models w/o covariates
2. A covariate dependent PPM
   - Model
   - Similarity function
3. Posterior inference
4. Data Illustrations
5. Discussion

Introduction – Motivating Example 1

Study: chemo-immunotherapy for ovarian cancer patients (Wang et al. 2005)

Data: $n = 47$ patients subject to variable doses of carboplatinum (chemotherapy agent) plus equal doses of γ-interferon and a colony stimulating factor.

Response: Monocyte (white cells that protect against blood-borne pathogens; have quick action) counts over time (y_{ij}).

Covariates: treatment dose (x_i).

Introduction – Motivating Example 1

Study: chemo-immunotherapy for ovarian cancer patients (Wang et al. 2005)

Data: $n = 47$ patients subject to variable doses of carboplatinum (chemotherapy agent) plus equal doses of γ-interferon and a colony stimulating factor.

Response: Monocyte (white cells that protect against blood-borne pathogens; have quick action) counts over time (y_{ij}).

Covariates: treatment dose (x_i).

Introduction – Motivating Example 1

Study: chemo-immunotherapy for ovarian cancer patients (Wang et al. 2005)

Data: $n = 47$ patients subject to variable doses of carboplatinum (chemotherapy agent) plus equal doses of γ-interferon and a colony stimulating factor.

Response: Monocyte (white cells that protect against blood-borne pathogens; have quick action) counts over time (y_{ij}).

Covariates: treatment dose (x_i).

Introduction – Motivating Example 1

Study: chemo-immunotherapy for ovarian cancer patients (Wang et al. 2005)

Data: $n = 47$ patients subject to variable doses of carboplatinum (chemotherapy agent) plus equal doses of γ-interferon and a colony stimulating factor.

Response: Monocyte (white cells that protect against blood-borne pathogens; have quick action) counts over time (y_{ij}).

Covariates: treatment dose (x_i).

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

Introduction – Motivating Example 2

Study: Chemotherapy for breast cancer patients. (same as in De Iorio et al. 2007)

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

Data: $n = 765$ patients, randomized to A or B.

Response: disease-free survival, i.e. time until death from any cause, relapse, or diagnosis with a second malignancy (y_i).

Covariates: Covariate vector of dimension 6 and of mixed types:

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

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

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

- Match her with already observed patients
- predict a similar response

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

Two patients with similar covariate values (e.g. dose, treatment, tumor size, etc.) should be a priori more likely to co-cluster (not so easy).

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

Deterministic heuristic clustering algorithms will not do.

Random Partition Models – Notation

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

Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,

clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.

Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$

Random partition: $p(\rho_n)$

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 $y_j^* = \{y_i;\ i \in S_j\}$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,
clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.
Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$
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 $y_j^* = \{y_i;\ i \in S_j\}$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,
clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.
Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$
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 $y_j^* = \{y_i;\ i \in S_j\}$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,
clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.
Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$
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 $y_j^* = \{y_i;\ i \in S_j\}$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,
clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.
Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$
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 $y_j^* = \{y_i;\ i \in S_j\}$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,
clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.
Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$
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 $y_j^* = \{y_i;\ i \in S_j\}$

Random Partition Models – Notation

Notation: units (patients) $i \in S = \{1, \ldots, n\}$,
Partition $\rho_n = \{S_1, \ldots, S_k\}$, with $S = S_1 \cup S_2 \cup \ldots \cup S_k$,
clusters (partitioning subsets) $S_i \subset S$, $S_i \neq \emptyset$.
Alternative parametrization: $\rho_n \leftrightarrow (s, k)$ with $s_i = j$ iff $i \in S_j$
Random partition: $p(\rho_n)$
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 $y_j^* = \{y_i;\ i \in S_j\}$

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 ... (just by simplicity)

Prior $p(\rho)$: product distribution

SSM (DP, Pitman-Yor, stick-breaking, normalized inverse-gaussian, Poisson-Kingman, ...)

model-based clustering, e.g. Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

Sometimes they are easier to describe in predictive terms.

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 ... (just by simplicity)

Prior $p(\rho)$: product distribution

SSM (DP, Pitman-Yor, stick-breaking, normalized inverse-gaussian, Poisson-Kingman, ...)

model-based clustering, e.g. Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

Sometimes they are easier to describe in predictive terms.

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 ... (just by simplicity)

Prior $p(\rho)$: product distribution

SSM (DP, Pitman-Yor, stick-breaking, normalized inverse-gaussian, Poisson-Kingman, ...)

model-based clustering, e.g. Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

Sometimes they are easier to describe in predictive terms.

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 ... (just by simplicity)

Prior $p(\rho)$: product distribution

SSM (DP, Pitman-Yor, stick-breaking, normalized inverse-gaussian, Poisson-Kingman, ...)

model-based clustering, e.g. Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

Sometimes they are easier to describe in predictive terms.

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 ... (just by simplicity)

Prior $p(\rho)$: product distribution

SSM (DP, Pitman-Yor, stick-breaking, normalized inverse-gaussian, Poisson-Kingman, ...)

model-based clustering, e.g. Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

Sometimes they are easier to describe in predictive terms.

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 ... (just by simplicity)

Prior $p(\rho)$: product distribution

SSM (DP, Pitman-Yor, stick-breaking, normalized inverse-gaussian, Poisson-Kingman, ...)

model-based clustering, e.g. Fraley and Raftery (2002 JASA), Richardson and Green (1997 JRSSB)

Sometimes they are easier to describe in predictive terms.

Random Partition Models

Parametric product partition model (PPM): Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA), Crowley (1997 JASA),...

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)

Polya urn: a PPM with $c(S_i) = M \times (|S_i| - 1)!$

also a SSM with prior predictive probabilities

$$p(s_{n+1} = j \mid s_1, \ldots, s_n) = \frac{1}{M+n} \begin{cases} |S_j| & \text{if } 1 \leq j \leq k \\ M & \text{if } j = k+1 \end{cases}$$

Random Partition Models

Parametric product partition model (PPM): Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA), Crowley (1997 JASA),...

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)

Polya urn: a PPM with $c(S_i) = M \times (|S_i| - 1)!$

also a SSM with prior predictive probabilities

$$p(s_{n+1} = j \mid s_1, \ldots, s_n) = \frac{1}{M+n} \begin{cases} |S_j| & \text{if } 1 \leq j \leq k \\ M & \text{if } j = k+1 \end{cases}$$

Random Partition Models

Parametric product partition model (PPM): Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA), Crowley (1997 JASA),...

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)

Polya urn: a PPM with $c(S_i) = M \times (|S_i| - 1)!$

also a SSM with prior predictive probabilities

$$p(s_{n+1} = j \mid s_1, \ldots, s_n) = \frac{1}{M+n} \begin{cases} |S_j| & \text{if } 1 \leq j \leq k \\ M & \text{if } j = k+1 \end{cases}$$

Random Partition Models

Parametric product partition model (PPM): Hartigan (1990 Comm Stat), Barry and Hartigan (1993 JASA), Crowley (1997 JASA),...

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)

Polya urn: a PPM with $c(S_i) = M \times (|S_i| - 1)!$

also a SSM with prior predictive probabilities

$$p(s_{n+1} = j \mid s_1, \ldots, s_n) = \frac{1}{M+n} \begin{cases} |S_j| & \text{if } 1 \leq j \leq k \\ M & \text{if } j = k+1 \end{cases}$$

Covariate-dependent PPM: how to define it?

Similarity function: define $g(x_j^\star) > 0$ to characterize the similarity of $x_j^* = \{x_i;\ i \in S_j\}$ 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 normalization constant $g_n(x^n) = \sum_\rho \prod_{j=1}^k g(x_j^\star) c(S_j)$

Natural choice:

define $g(x_j^\star)$ as an (auxiliary) exchangeable probability 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: how to define it?

Similarity function: define $g(x_j^\star) > 0$ to characterize the similarity of $x_j^* = \{x_i;\ i \in S_j\}$ 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 normalization constant $g_n(x^n) = \sum_\rho \prod_{j=1}^k g(x_j^\star) c(S_j)$

Natural choice:

define $g(x_j^\star)$ as an (auxiliary) exchangeable probability 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: how to define it?

Similarity function: define $g(x_j^\star) > 0$ to characterize the similarity of $x_j^* = \{x_i;\ i \in S_j\}$ 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 normalization constant $g_n(x^n) = \sum_\rho \prod_{j=1}^k g(x_j^\star) c(S_j)$

Natural choice:

define $g(x_j^\star)$ as an (auxiliary) exchangeable probability 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 (ctd.)

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

Coherence across n: desirable (but not critical),

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

for some probability model $q(x_{n+1} \mid x^n)$.
e.g., for $q(x_{n+1} \mid x^n) = g_{n+1}(x^{n+1})/g_n(x^n)$.

Covariate-dependent PPM (ctd.)

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

Coherence across n: desirable (but not critical),

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

for some probability model $q(x_{n+1} \mid x^n)$.

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

Covariate-dependent PPM (ctd.)

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

Coherence across n: desirable (but not critical),

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

for some probability model $q(x_{n+1} \mid x^n)$.
e.g., for $q(x_{n+1} \mid x^n) = g_{n+1}(x^{n+1})/g_n(x^n)$.

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Covariate-dependent PPM (ctd.)

Similarity function $g(x_j^\star)$: computationally efficient posterior simulations with

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

where $q(x_i \mid \xi)$ and $q(\xi)$ are probability models.

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

Categorical: Multinomial and Dirichlet

Ordinal: multinomial probit model with fixed cutoffs and (conditionally conjugate) normal prior.

Counts: $q(x_i \mid \xi_j) = Poisson(\xi_j)$ and $q(\xi_j) = Ga(a, b)$.

In principle we assume independence across different types of covariates but can also consider dependence ("correlation").

Posterior Inference – w/o Covariates

- Lau and Green (2007) 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, Lau and Green, 2007).
- Alternatively Bayesian hierarchical clustering (Heard et al. 2006 JASA). Iteratively combine clusters to maximize posterior probabilities.
- Least squares clustering (Dahl, 2006).

Posterior Inference – w/o Covariates

- Lau and Green (2007) 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, Lau and Green, 2007).
- Alternatively Bayesian hierarchical clustering (Heard et al. 2006 JASA). Iteratively combine clusters to maximize posterior probabilities.
- Least squares clustering (Dahl, 2006).

Posterior Inference – w/o Covariates

- Lau and Green (2007) 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, Lau and Green, 2007).
- Alternatively Bayesian hierarchical clustering (Heard et al. 2006 JASA). Iteratively combine clusters to maximize posterior probabilities.
- Least squares clustering (Dahl, 2006).

Posterior Inference – w/o Covariates

- Lau and Green (2007) 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, Lau and Green, 2007).
- Alternatively Bayesian hierarchical clustering (Heard et al. 2006 JASA). Iteratively combine clusters to maximize posterior probabilities.
- Least squares clustering (Dahl, 2006).

Posterior Inference – w/o Covariates

- Lau and Green (2007) 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, Lau and Green, 2007).
- Alternatively Bayesian hierarchical clustering (Heard et al. 2006 JASA). Iteratively combine clusters to maximize posterior probabilities.
- Least squares clustering (Dahl, 2006).

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
- Similar approach proposed in Neal (2007).
- Alternative constructions: DDPs, hybrid DPs, order-restricted DPs, DP regression smoother, mixtures of experts, monotonic nonparametric regression, matrix stick-breaking priors, Normalized kernel-weighted random measures,...

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
- Similar approach proposed in Neal (2007).
- Alternative constructions: DDPs, hybrid DPs, order-restricted DPs, DP regression smoother, mixtures of experts, monotonic nonparametric regression, matrix stick-breaking priors, Normalized kernel-weighted random measures,...

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
- Similar approach proposed in Neal (2007).
- Alternative constructions: DDPs, hybrid DPs, order-restricted DPs, DP regression smoother, mixtures of experts, monotonic nonparametric regression, matrix stick-breaking priors, Normalized kernel-weighted random measures,...

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
- Similar approach proposed in Neal (2007).
- Alternative constructions: DDPs, hybrid DPs, order-restricted DPs, DP regression smoother, mixtures of experts, monotonic nonparametric regression, matrix stick-breaking priors, Normalized kernel-weighted random measures,...

Posterior Inference w. Covariates (ctd.)

Model augmentation: augment original model to include $q(\cdot)$ as an auxiliary model on x and ξ,

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

$$(x^n, \xi \mid \rho) \sim \prod_{j=1}^{k} q(x_j^\star \mid \xi_j) \cdot q(\xi_j)$$

and change the prior on ρ

$$q(\rho_n) \propto \prod c(S_j)$$

$p(\rho^n \mid y^n, x^n)$ is exactly as in the original model,
assuming independence of θ_j and ξ_j, *and* no hyperparameters in $q(\xi_j)$.
With hyperparameters the missing $g_n(x^n)$ matters.

Posterior Inference w. Covariates (ctd.)

Model augmentation: augment original model to include $q(\cdot)$ as an auxiliary model on x and ξ,

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

$$(x^n, \xi \mid \rho) \sim \prod_{j=1}^{k} q(x_j^\star \mid \xi_j) \cdot q(\xi_j)$$

and change the prior on ρ

$$q(\rho_n) \propto \prod c(S_j)$$

$p(\rho^n \mid y^n, x^n)$ is exactly as in the original model, assuming independence of θ_j and ξ_j, *and* no hyperparameters in $q(\xi_j)$.
With hyperparameters the missing $g_n(x^n)$ matters.

Posterior Inference w. Covariates (ctd.)

Model augmentation: augment original model to include $q(\cdot)$ as an auxiliary model on x and ξ,

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

$$(x^n, \xi \mid \rho) \sim \prod_{j=1}^{k} q(x_j^\star \mid \xi_j) \cdot q(\xi_j)$$

and change the prior on ρ

$$q(\rho_n) \propto \prod c(S_j)$$

$p(\rho^n \mid y^n, x^n)$ is exactly as in the original model,
assuming independence of θ_j and ξ_j, *and* no hyperparameters in $q(\xi_j)$.
With hyperparameters the missing $g_n(x^n)$ matters.

Posterior Inference w. Covariates (ctd.)

Model augmentation: augment original model to include $q(\cdot)$ as an auxiliary model on x and ξ,

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

$$(x^n, \xi \mid \rho) \sim \prod_{j=1}^{k} q(x_j^\star \mid \xi_j) \cdot q(\xi_j)$$

and change the prior on ρ

$$q(\rho_n) \propto \prod c(S_j)$$

$p(\rho^n \mid y^n, x^n)$ is exactly as in the original model,
assuming independence of θ_j and ξ_j, *and* no hyperparameters in $q(\xi_j)$.
With hyperparameters the missing $g_n(x^n)$ matters.

Posterior Inference w. Covariates (ctd.)

Model augmentation: augment original model to include $q(\cdot)$ as an auxiliary model on x and ξ,

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

$$(x^n, \xi \mid \rho) \sim \prod_{j=1}^{k} q(x_j^\star \mid \xi_j) \cdot q(\xi_j)$$

and change the prior on ρ

$$q(\rho_n) \propto \prod c(S_j)$$

$p(\rho^n \mid y^n, x^n)$ is exactly as in the original model,
assuming independence of θ_j and ξ_j, *and* no hyperparameters in $q(\xi_j)$.
With hyperparameters the missing $g_n(x^n)$ matters.

Posterior Inference w. Covariates (ctd.)

Posterior predictive: Let $\tilde{y} = y_{n+1}$ denote a future response for a patient with covariates $\tilde{x} = x_{n+1}$. We need

$$p(\tilde{y} \mid \tilde{x}, x^n, y^n) = \int p(\tilde{y} \mid \tilde{x}, \rho_{n+1}, x^n, y^n)\, dp(\rho_{n+1} \mid \tilde{x}, x^n, y^n).$$

MCMC evaluation: For an MCMC implementation we need to reduce it to an integral w.r.t. $p(\rho_n \mid x^n, y^n)$.

Posterior Inference w. Covariates (ctd.)

Posterior predictive: Let $\tilde{y} = y_{n+1}$ denote a future response for a patient with covariates $\tilde{x} = x_{n+1}$. We need

$$p(\tilde{y} \mid \tilde{x}, x^n, y^n) = \int p(\tilde{y} \mid \tilde{x}, \rho_{n+1}, x^n, y^n)\, dp(\rho_{n+1} \mid \tilde{x}, x^n, y^n).$$

MCMC evaluation: For an MCMC implementation we need to reduce it to an integral w.r.t. $p(\rho_n \mid x^n, y^n)$.

Posterior Inference w. Covariates (ctd.)

Posterior predictive: Let $\tilde{y} = y_{n+1}$ denote a future response for a patient with covariates $\tilde{x} = x_{n+1}$. We need

$$p(\tilde{y} \mid \tilde{x}, x^n, y^n) = \int p(\tilde{y} \mid \tilde{x}, \rho_{n+1}, x^n, y^n)\, dp(\rho_{n+1} \mid \tilde{x}, x^n, y^n).$$

MCMC evaluation: For an MCMC implementation we need to reduce it to an integral w.r.t. $p(\rho_n \mid x^n, y^n)$.

Posterior Inference w. Covariates (ctd.)

Let $\ell = s_{n+1}$ and $p(\tilde{y} \mid y^\star_\ell) = \int p(\tilde{y} \mid \theta_\ell) dp(\theta_\ell \mid y^\star_\ell)$.

$$p(\tilde{y} \mid \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} \mid y^\star_\ell, \rho_n) \underbrace{\frac{g(\tilde{x}^\star_\ell)}{g(x^\star_\ell)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}}_{w_\ell} p(\rho_n \mid y^n, x^n) d\rho_n$$

In words

- Use posterior samples ρ_n for size n sample;
- set s_{n+1} with $Pr(s_{n+1} = \ell) \propto w_\ell$
- evaluate predictive for $\tilde{y}$, assuming cluster ℓ.
- average over posterior samples

Posterior Inference w. Covariates (ctd.)

Let $\ell = s_{n+1}$ and $p(\tilde{y} \mid y^\star_\ell) = \int p(\tilde{y} \mid \theta_\ell) dp(\theta_\ell \mid y^\star_\ell)$.

$$p(\tilde{y} \mid \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} \mid y^\star_\ell, \rho_n) \underbrace{\frac{g(\tilde{x}^\star_\ell)}{g(x^\star_\ell)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}}_{w_\ell} p(\rho_n \mid y^n, x^n) d\rho_n$$

In words

- Use posterior samples ρ_n for size n sample;
- set s_{n+1} with $Pr(s_{n+1} = \ell) \propto w_\ell$
- evaluate predictive for $\tilde{y}$, assuming cluster ℓ.
- average over posterior samples

Posterior Inference w. Covariates (ctd.)

Let $\ell = s_{n+1}$ and $p(\tilde{y} \mid y^\star_\ell) = \int p(\tilde{y} \mid \theta_\ell) dp(\theta_\ell \mid y^\star_\ell)$.

$$p(\tilde{y} \mid \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} \mid y^\star_\ell, \rho_n) \underbrace{\frac{g(\tilde{x}^\star_\ell)}{g(x^\star_\ell)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}}_{w_\ell} p(\rho_n \mid y^n, x^n) d\rho_n$$

In words

- Use posterior samples ρ_n for size n sample;
- set s_{n+1} with $Pr(s_{n+1} = \ell) \propto w_\ell$
- evaluate predictive for $\tilde{y}$, assuming cluster ℓ.
- average over posterior samples

Posterior Inference w. Covariates (ctd.)

Let $\ell = s_{n+1}$ and $p(\tilde{y} \mid y^\star_\ell) = \int p(\tilde{y} \mid \theta_\ell) dp(\theta_\ell \mid y^\star_\ell)$.

$$p(\tilde{y} \mid \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} \mid y^\star_\ell, \rho_n) \underbrace{\frac{g(\tilde{x}^\star_\ell)}{g(x^\star_\ell)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}}_{w_\ell} p(\rho_n \mid y^n, x^n) d\rho_n$$

In words

- Use posterior samples ρ_n for size n sample;
- set s_{n+1} with $Pr(s_{n+1} = \ell) \propto w_\ell$
- evaluate predictive for $\tilde{y}$, assuming cluster ℓ.
- average over posterior samples

Posterior Inference w. Covariates (ctd.)

Let $\ell = s_{n+1}$ and $p(\tilde{y} \mid y^\star_\ell) = \int p(\tilde{y} \mid \theta_\ell) dp(\theta_\ell \mid y^\star_\ell)$.

$$p(\tilde{y} \mid \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} \mid y^\star_\ell, \rho_n) \underbrace{\frac{g(\tilde{x}^\star_\ell)}{g(x^\star_\ell)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}}_{w_\ell} p(\rho_n \mid y^n, x^n) d\rho_n$$

In words

- Use posterior samples ρ_n for size n sample;
- set s_{n+1} with $Pr(s_{n+1} = \ell) \propto w_\ell$
- evaluate predictive for $\tilde{y}$, assuming cluster ℓ.
- average over posterior samples

Posterior Inference w. Covariates (ctd.)

Let $\ell = s_{n+1}$ and $p(\tilde{y} \mid y^\star_\ell) = \int p(\tilde{y} \mid \theta_\ell) dp(\theta_\ell \mid y^\star_\ell)$.

$$p(\tilde{y} \mid \tilde{x}, y^n, x^n) \propto \int \sum_{\ell=1}^{k_n+1} p(\tilde{y} \mid y^\star_\ell, \rho_n) \underbrace{\frac{g(\tilde{x}^\star_\ell)}{g(x^\star_\ell)} \frac{c(S_\ell \cup \{n+1\})}{c(S_\ell)}}_{w_\ell} p(\rho_n \mid y^n, x^n) d\rho_n$$

In words

- Use posterior samples ρ_n for size n sample;
- set s_{n+1} with $Pr(s_{n+1} = \ell) \propto w_\ell$
- evaluate predictive for $\tilde{y}$, assuming cluster ℓ.
- average over posterior samples

Example: Ovarian Cancer

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

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

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.

Example: Ovarian Cancer

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

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

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.

Example: Ovarian Cancer

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

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

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.

Example: Ovarian Cancer

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

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

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.

Example: Ovarian Cancer

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

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

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.

Example: Ovarian Cancer

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

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

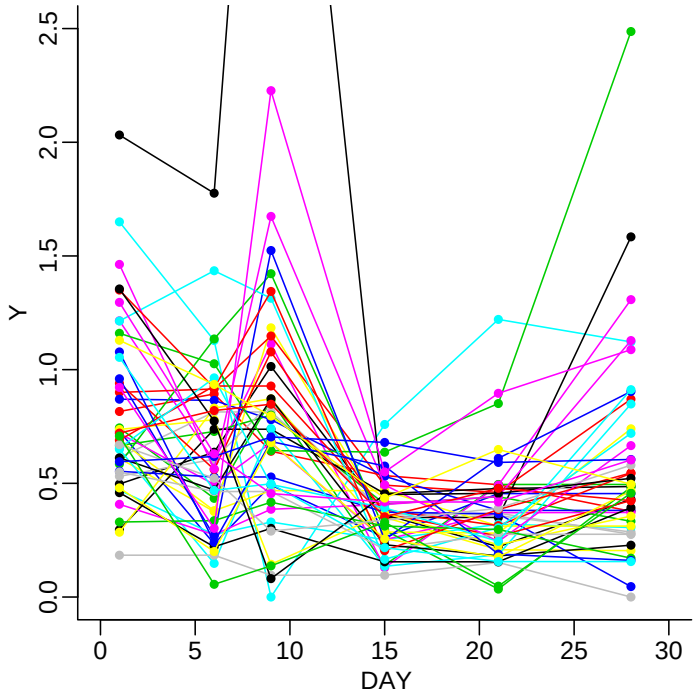
Model: PPM model with additional covariate dependence.

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

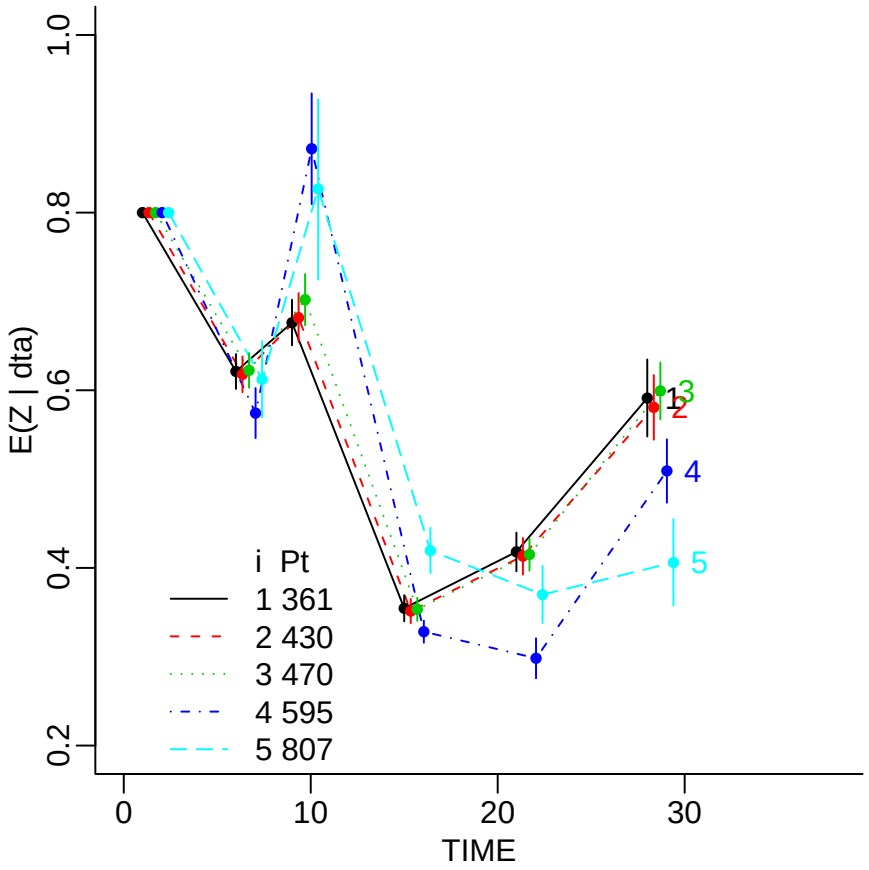
Random partition: $p(\rho_n \mid x^n) \propto \prod c(S_j) g(x_j^\star)$ and conjugate prior $p(\theta_j)$.

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

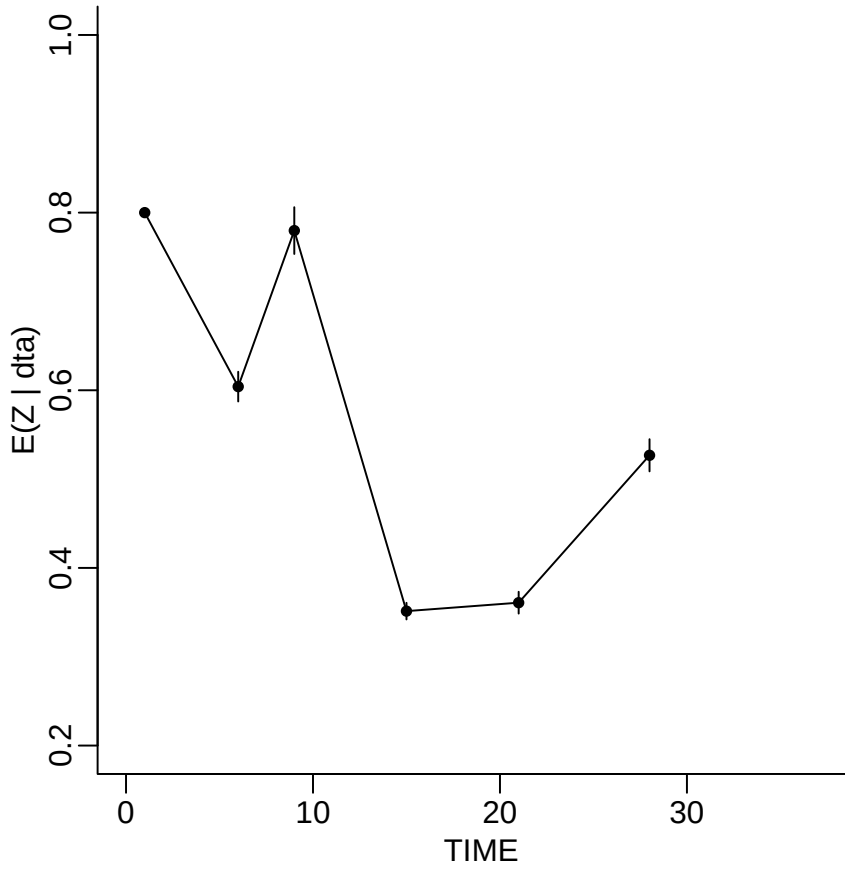
Example: Ovarian Cancer



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



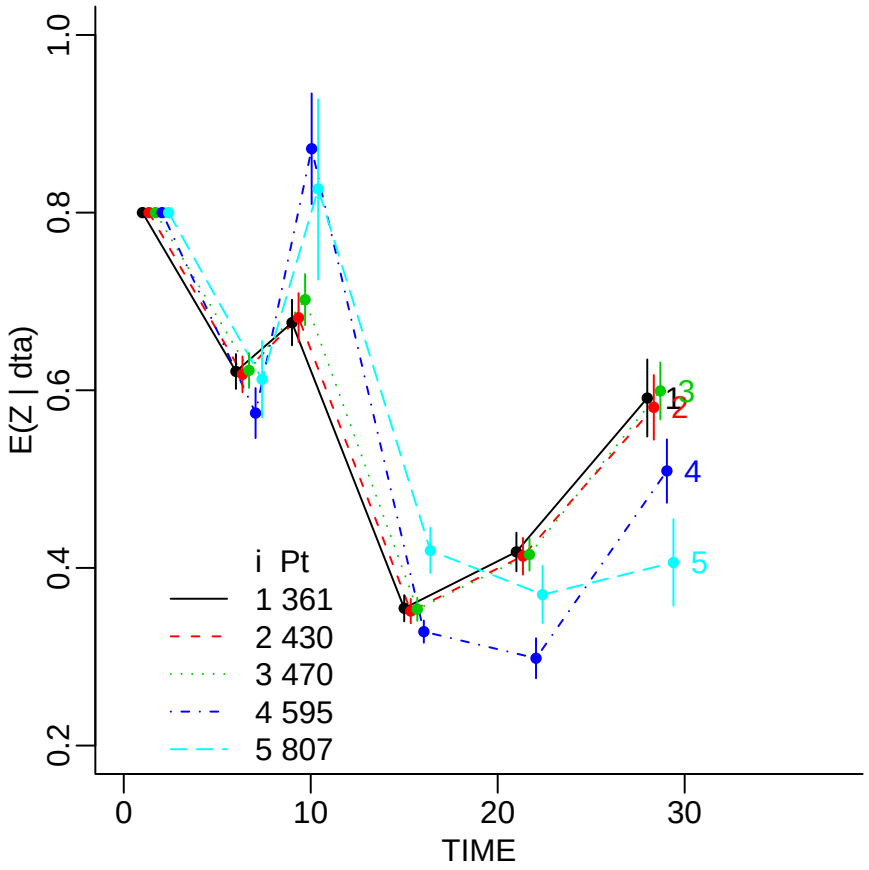
with covariate-dependent clustering

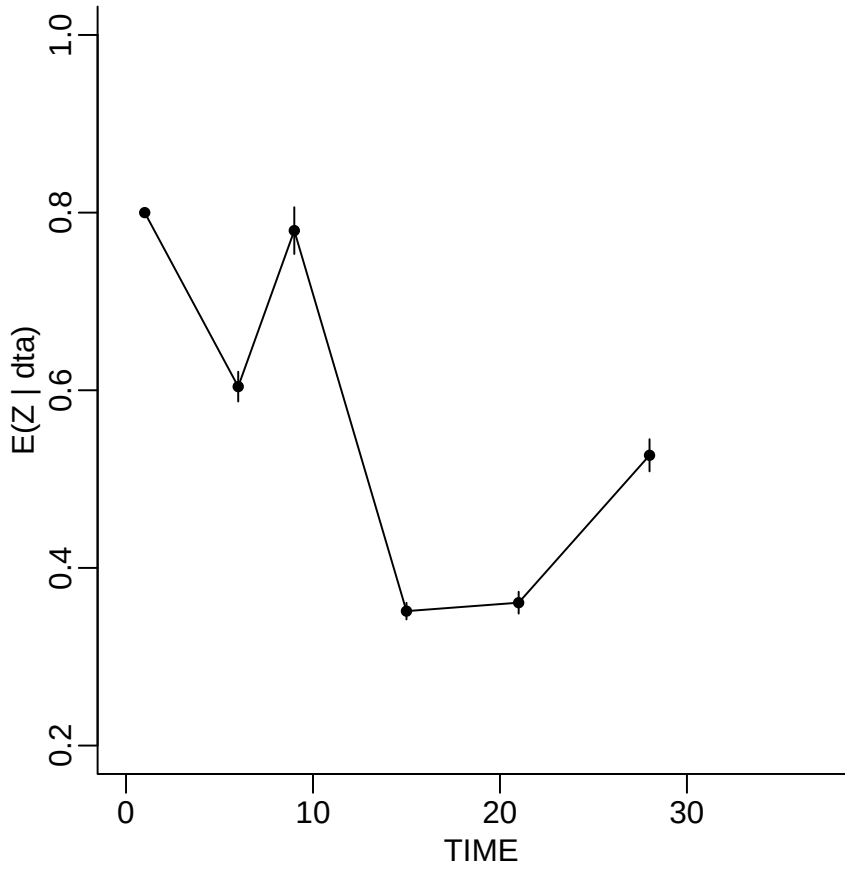


without covariate-dependent clustering

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.



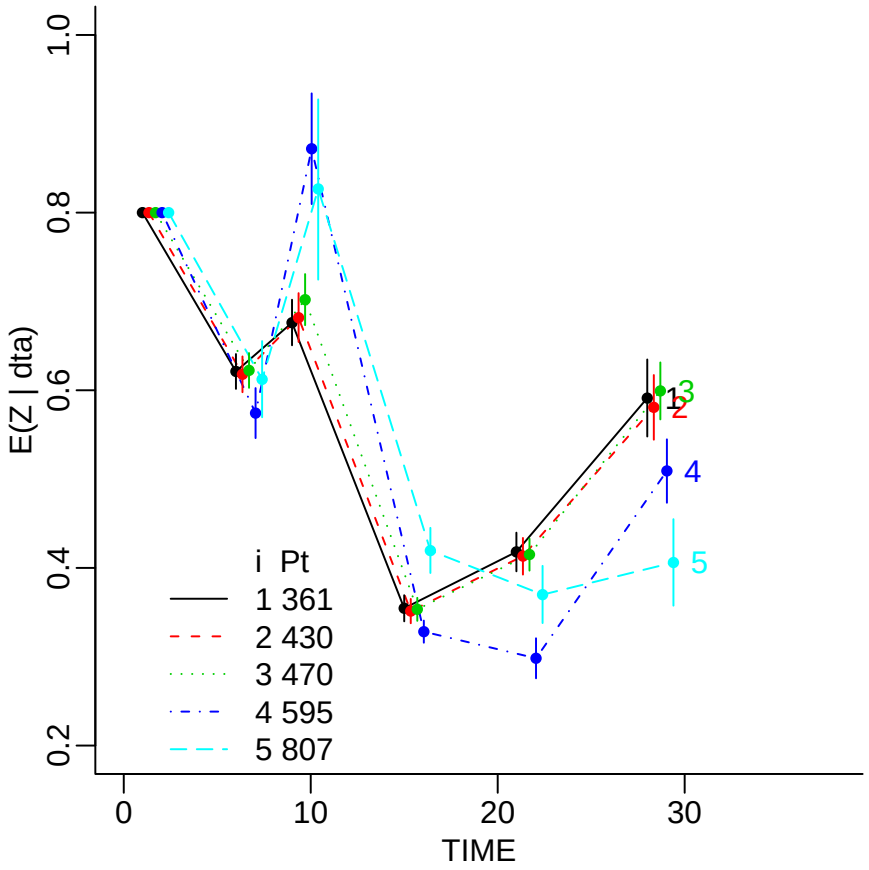


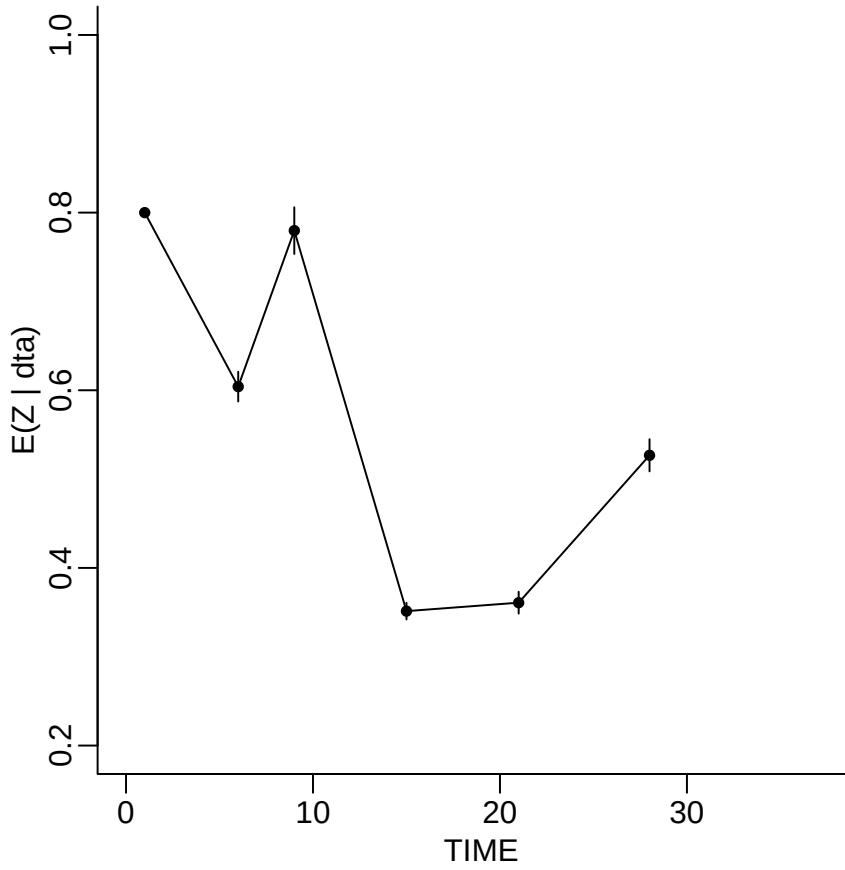
with covariate-dependent clustering

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.





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.

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 posterior inferences that may be quite different between the proposed and auxiliary models. Auxiliary model would then adjust the similarity functions in undesirable ways.

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).

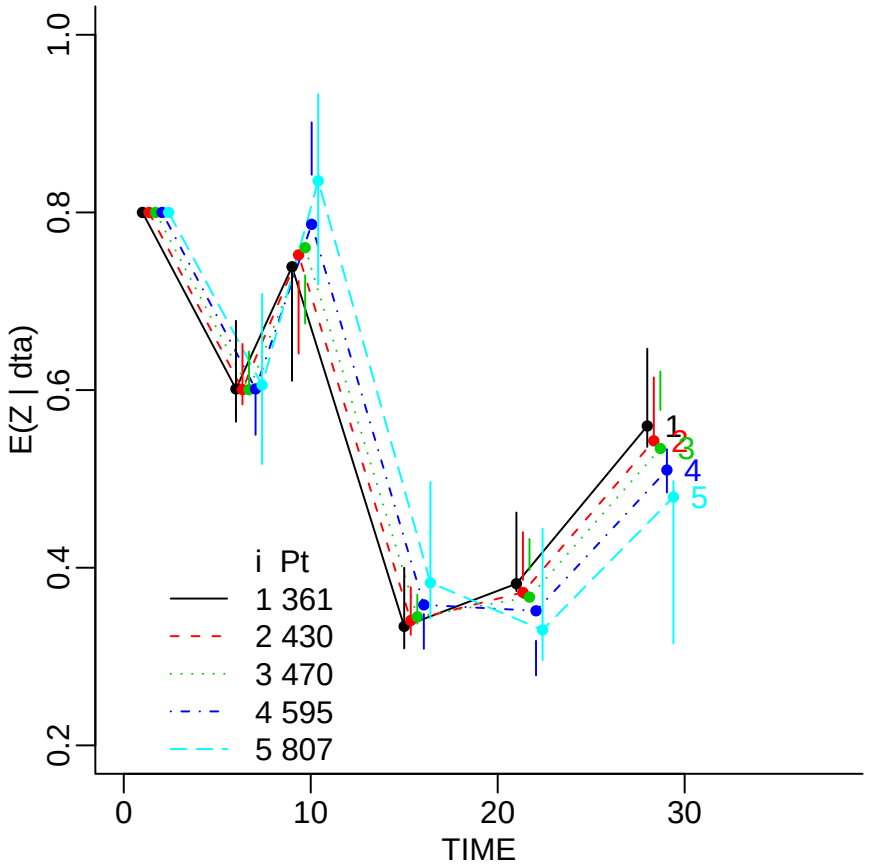
Prediction: becomes inference for a partial response $\tilde{x}$.

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

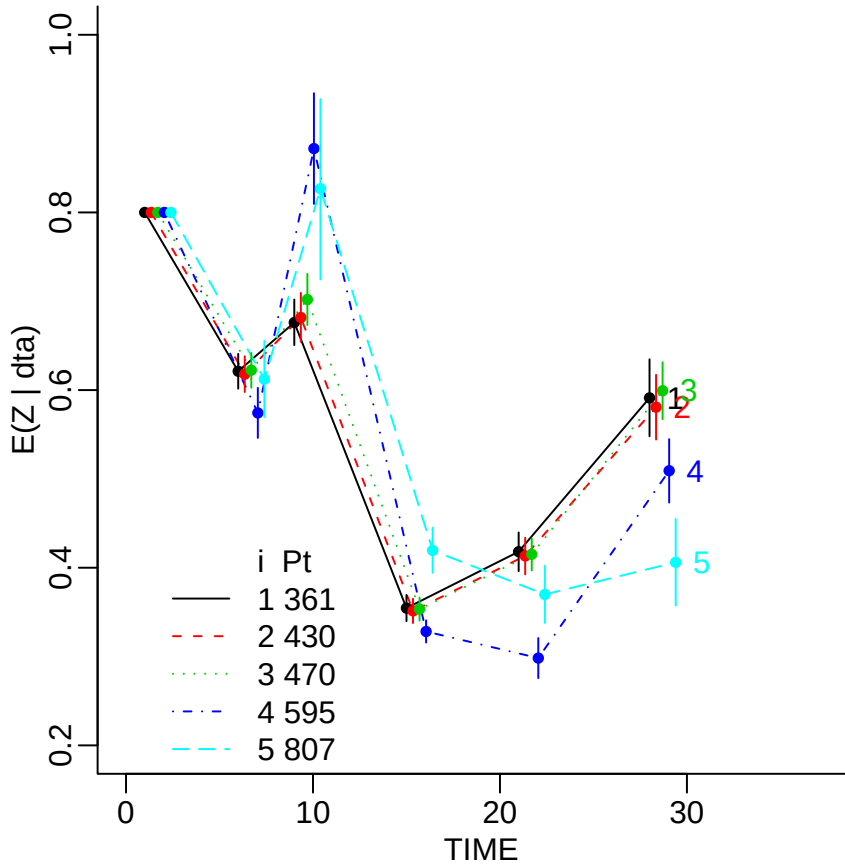
$$(x_i, y_i) \sim N(\theta_i, \Sigma)$$

Including the hyperparameters for x_i leads to posterior inferences that may be quite different between the proposed and auxiliary models. Auxiliary model would then adjust the similarity functions in undesirable ways.

Posterior predictive inference for a future patient by x_i:



augmented response (x_i, y_i)



covariate-dependent clustering

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^*.

We use default choices for auxiliary models.

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^*.

We use default choices for auxiliary models.

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^*.
We use default choices for auxiliary models.

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^*.
We use default choices for auxiliary models.

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

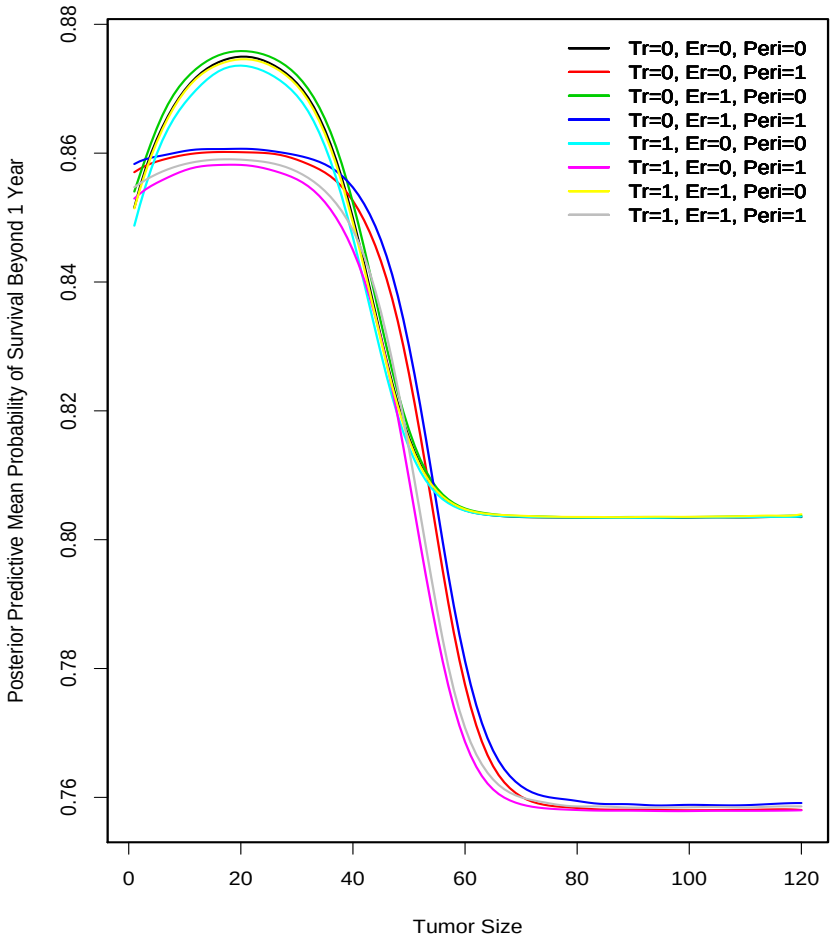
Covariates:

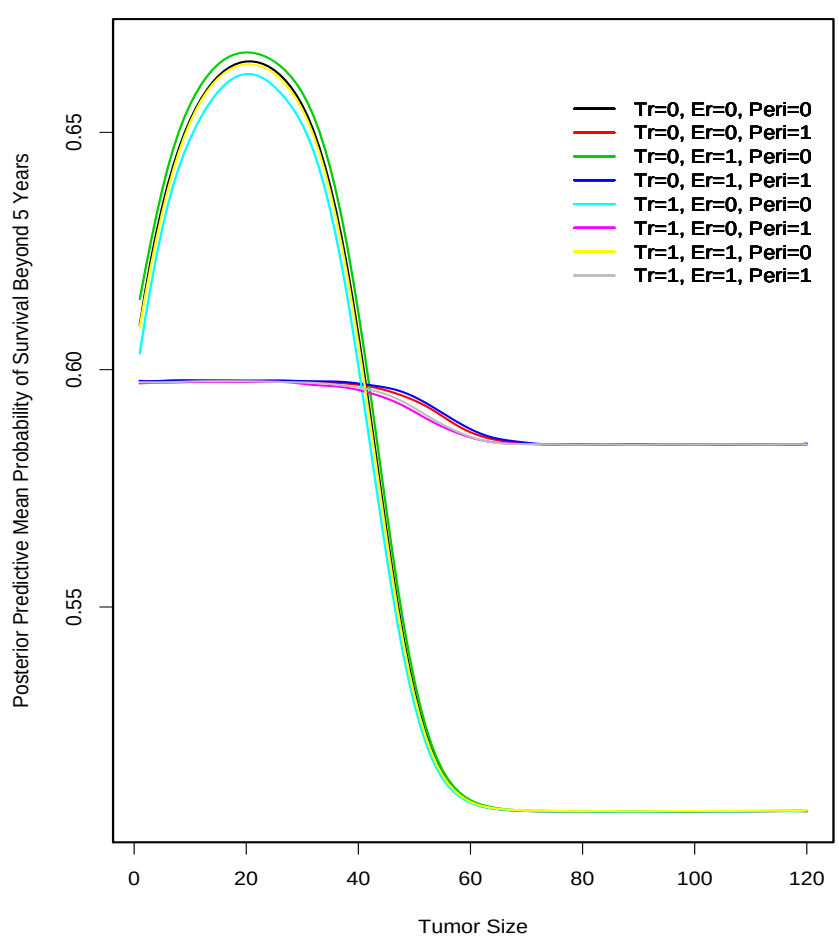
- *Categorical:* dose (A vs. B), menopausal status, estrogen use
- *Continuous:* age, initial tumor size,
- *Count:* number of positive lymph nodes

Model: covariate-dependent PPM.

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

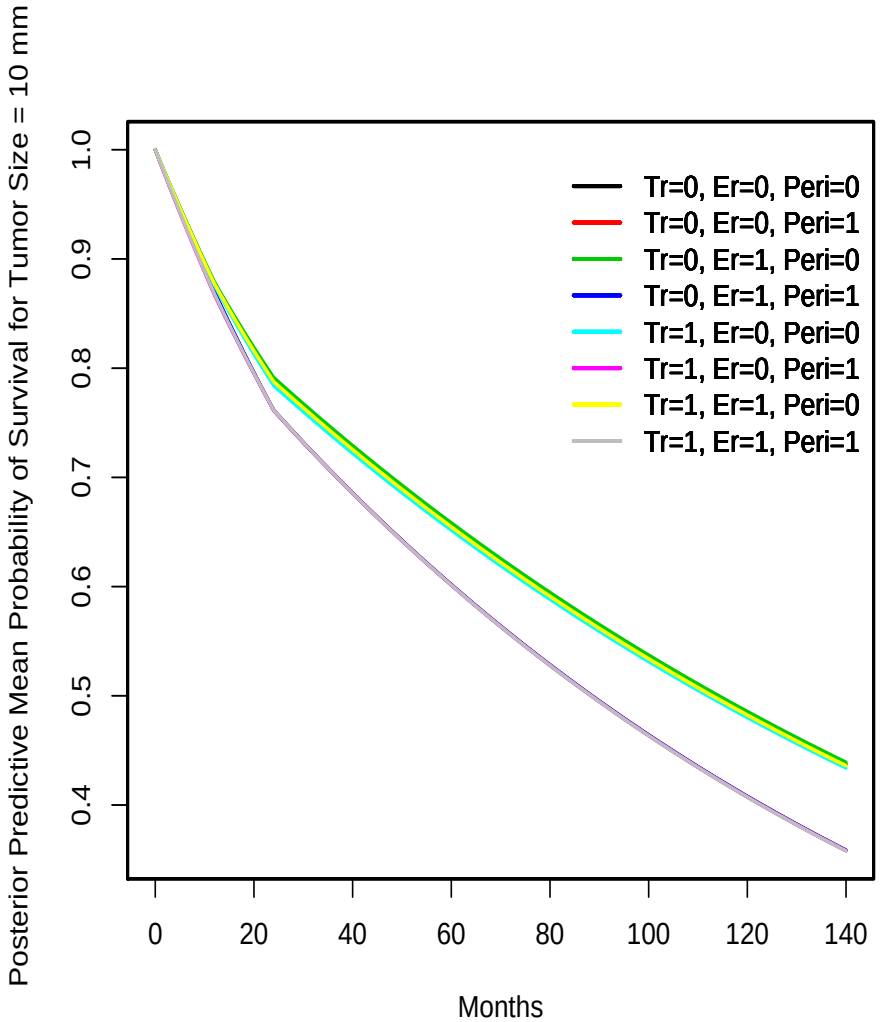
We use default choices for auxiliary models.



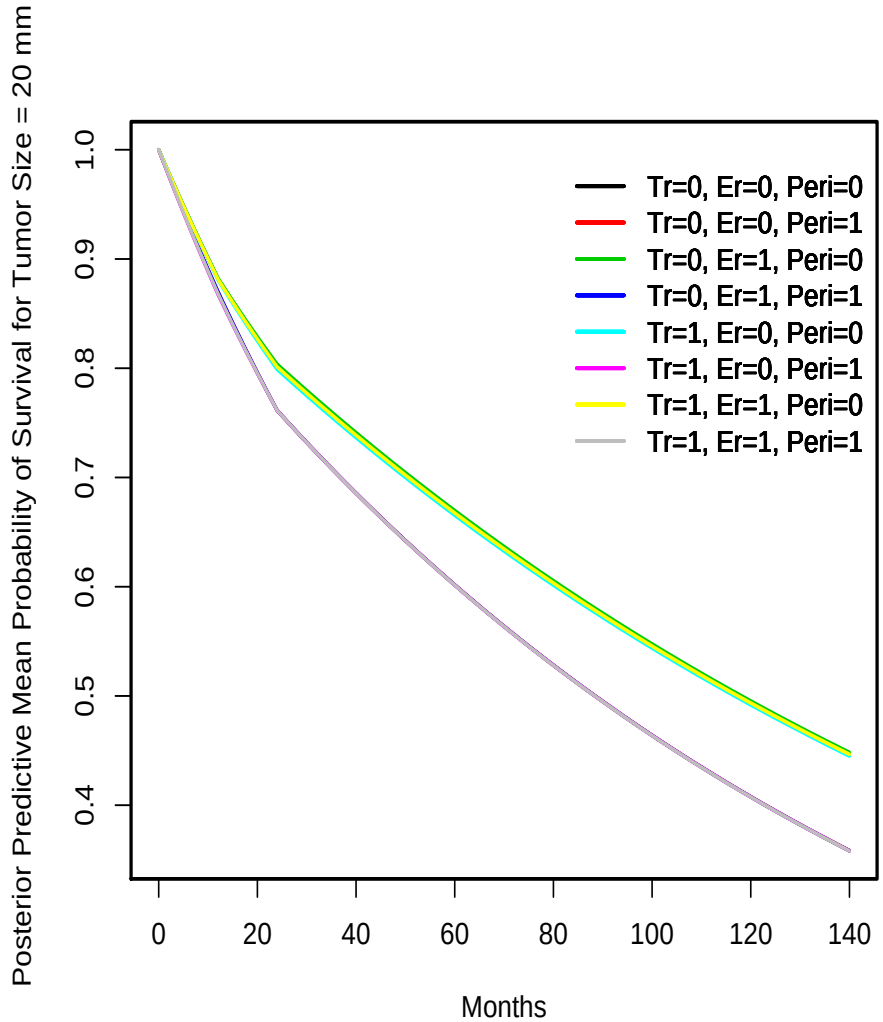


$P(T \geq 1 \mid \tilde{x}, x^n, y^n)$ $P(T \geq 5 \mid \tilde{x}, x^n, y^n)$

Predictions arranged by covariate $\tilde{x}$. Different clustering probabilities lead to differences in prediction, as desired.

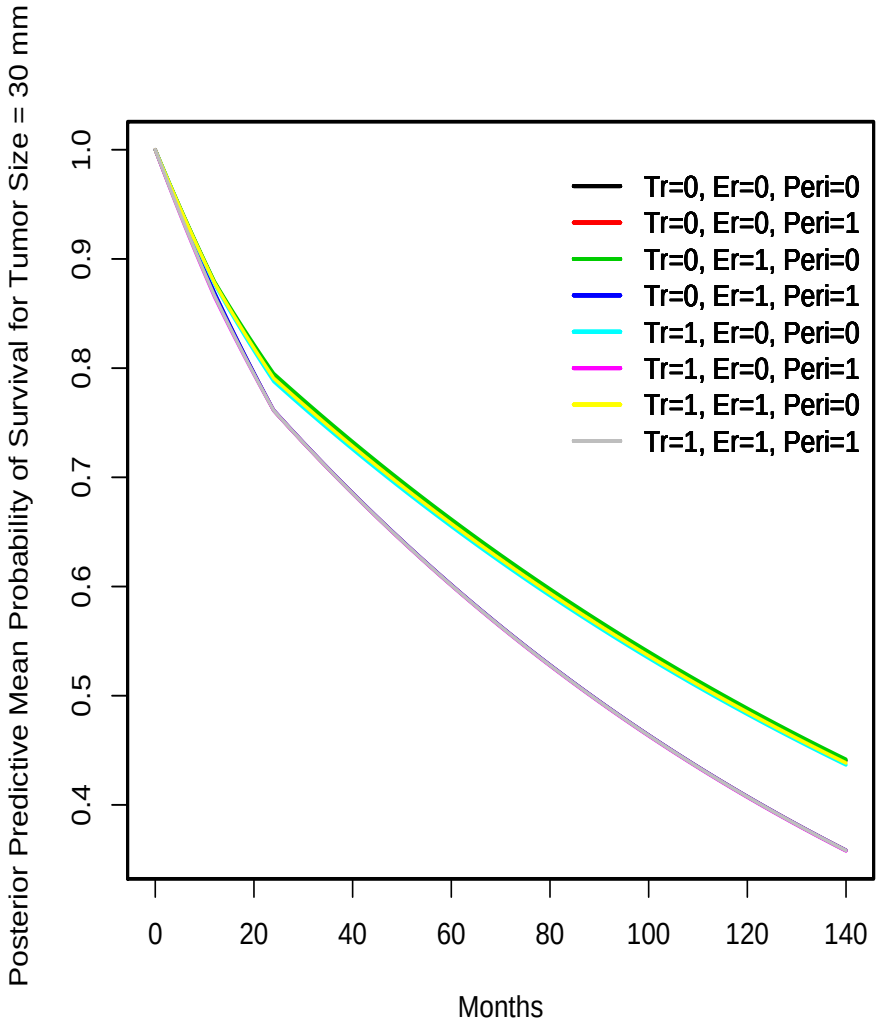


10 mm

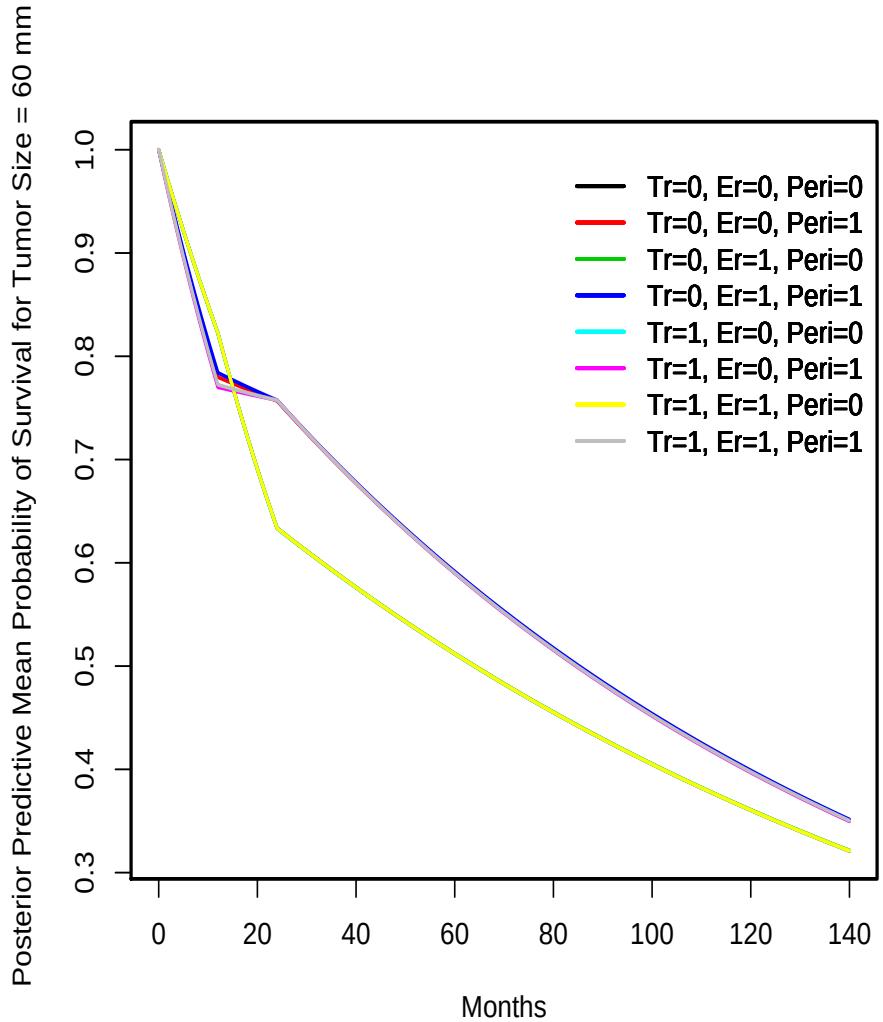


20 mm

Predicted survival curves arranged by covariate $\tilde{x}$.



30 mm



60 mm

Predicted survival curves arranged by covariate $\tilde{x}$.

Discussion

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 posterior 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.

Coherence: Would love to achieve easy coherence across sample size (?).

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

Discussion

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 posterior 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.

Coherence: Would love to achieve easy coherence across sample size (?).

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

Discussion

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 posterior 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.

Coherence: Would love to achieve easy coherence across sample size (?).

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

Discussion

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 posterior 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.

Coherence: Would love to achieve easy coherence across sample size (?).

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

Discussion

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 posterior 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.

Coherence: Would love to achieve easy coherence across sample size (?).

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

Discussion

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 posterior 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.

Coherence: Would love to achieve easy coherence across sample size (?).

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

Other Issues

Penalty on number of clusters: Proposed model may induce a penalty on the action of cohesion functions (i.e. on the probability of partitions).

Automatic choices of similarity function: Need some care when using our proposed choices. May arrange things so that, e.g. with $x_1 = \cdots = x_n$ model reduces to $p(\rho) \propto \prod_{i=1}^{k} c(S_i)$.

Other Issues

Penalty on number of clusters: Proposed model may induce a penalty on the action of cohesion functions (i.e. on the probability of partitions).

Automatic choices of similarity function: Need some care when using our proposed choices. May arrange things so that, e.g. with $x_1 = \cdots = x_n$ model reduces to $p(\rho) \propto \prod_{i=1}^{k} c(S_i)$.