A Bayesian Nonparametric Model for Classification of Longitudinal Profiles

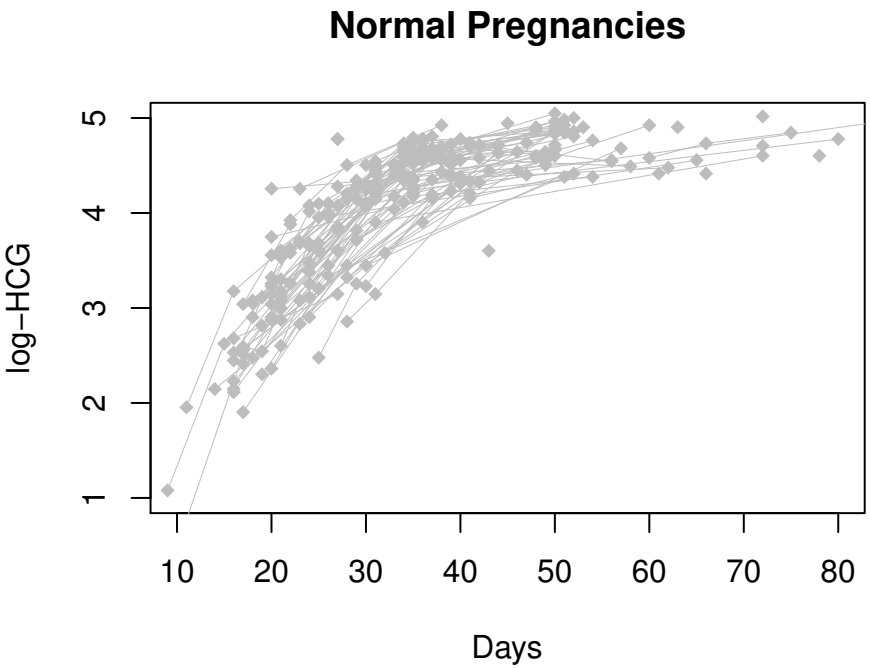
Claudio Fuentes
Department of Statistics
Oregon State University

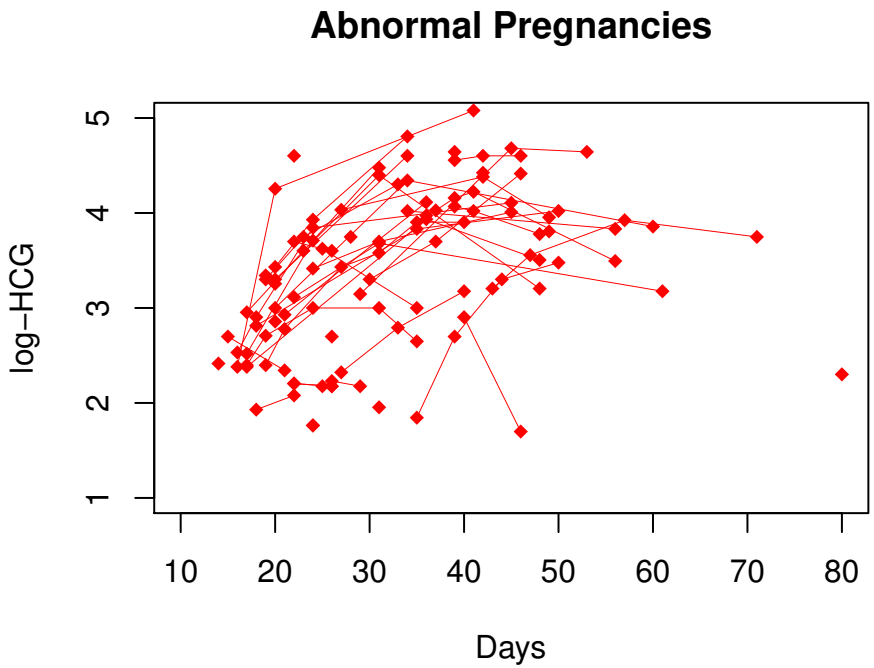
(Joint work with Jeremy Gaskins and Rolando de la Cruz)

Workshop on Machine and Statistical Learning with Applications
Santiago - 2023

The Motivating Question

- Women undergoing assisted reproductive therapy (ART) in order to become pregnant are at heightened risk of early pregnancy loss.
- The concentration of the hormone β-HCG is associated with the growth of the fetus in early pregnancy and may be predictive of abnormal pregnancy (loss of fetus or complications leading to nonterminal delivery).



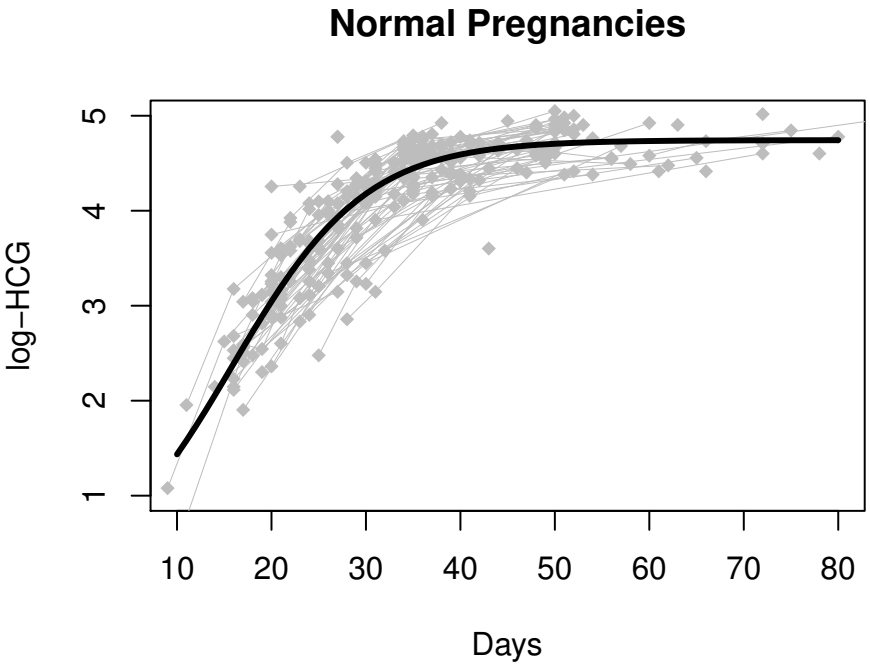


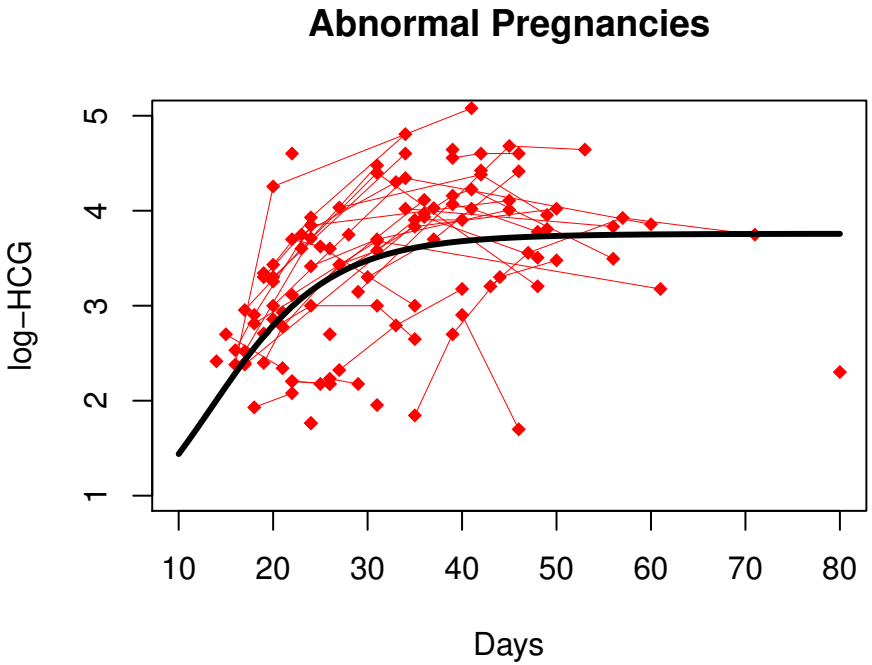
The Motivating Question

- If we know (a part of) a patient's β-HCG trajectory, can we use this information to predict whether she will go on to have a successful pregnancy?
- Let $D = 1$ denote disease (abnormal pregnancy) and let $\mathbf{Y}$ denote the vector of HCG-values for a given patient.
- Create a model for the longitudinal HCG trajectory for the normal pregnancy patients $f(\mathbf{Y}|D = 0)$ and a model for the longitudinal HCG trajectory for the abnormal pregnancy patients $f(\mathbf{Y}|D = 1)$.
- Use Bayes' Theorem to get a probability of abnormal pregnancy, given the HCG trajectory.

$$\Pr(D = 1|\mathbf{Y}) = \frac{f(\mathbf{Y}|D = 1)\Pr(D = 1)}{f(\mathbf{Y}|D = 1)\Pr(D = 1) + f(\mathbf{Y}|D = 0)\Pr(D = 0)}.$$

The Motivating Question





A couple of concerns:

- The abnormal pregnancy model does not appear to fit very well.
- There may be different ways for the pregnancy to fail and a model that tries to explain all of these in the same way will not work well.
- We need a model that allows multiple types of trajectories, but we do not necessarily know how many there should be.

Some Notation

- N = number of patients
  - In pregnancy data: $N = 173$
- D_i = disease status (1 for disease, 0 for healthy)
  - 49 (28.3%) abnormal pregnancies
- t_{ij} = the jth measurement occasion for patient i
- n_i = number of measurement occasions for patient i
  - Measurement times are not aligned and n_i ranges from 1 to 6, with 30% having only 1 measurement.
- Y_{ij} = longitudinal biomarker measurement for patient i at the jth observation
  - $Y_{ij} = \log_{10}(\beta\text{-HCG})$
- $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{i,n_i})$ = vector of biomarkers for patient i

The 2-Component Model

First, let's formalize the 2-component model that we compare our approach to (Marshall and Baron, 2000).

$$
\begin{aligned}
D_i &\sim \text{Bern}(\phi) \\
Y_{ij} &= f(t_{ij}; \boldsymbol{\theta}_{D_i}) + \gamma_i + \epsilon_{ij} \\
f(t; \boldsymbol{\theta}) &= \frac{\theta_1}{1 + \exp\{-\theta_2 t - \theta_3\}} \\
\gamma_i &\sim \text{N}(0, \gamma^2) \\
\boldsymbol{\epsilon}_i &\sim \text{MVN}_{n_i}\left(\mathbf{0}_{n_i}, \sigma^2 \mathbf{R}_i(\rho)\right)
\end{aligned}
$$

This model has two components:

- Healthy: probability $1 - \phi$ and parameter $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \theta_{03})$
- Disease: probability ϕ and parameter $\boldsymbol{\theta}_1 = (\theta_{11}, \theta_{12}, \theta_{13})$

2-Component Model

A few comments:

- The sigmoid function has been shown to reasonably model HCG in previous analyses and represents biologically plausible behavior

$$f(t;\boldsymbol{\theta}) = \frac{\theta_1}{1+\exp\{-\theta_2 t - \theta_3\}}.$$

- θ_1 represents the maximum height (plateou value) of the HCG curve.
- If $\theta_2 > 0$, then $f(t;\boldsymbol{\theta})$ is increasing (toward θ_1) with large values of θ_2 providing very steep increases.
- θ_3 is related to when the steep increase in HCG begins.
- Nothing requires that we use this sigmoid function. Other parametric models (polynomial, spline, etc.) could equivalently be used in place of the $f(t;\boldsymbol{\theta})$ function.

2-Component Model

A few comments:

- Dependence between measurements Y_{ij} and $Y_{ij'}$ is determined by the random effect γ_i and the autoregressive correlation matrix $\mathbf{R}_i$.

$$\text{corr}(Y_{ij}, Y_{ij'}) = \frac{\sigma^2 \rho^{|t_{ij} - t_{ij'}|} + \gamma^2}{\sigma^2 + \gamma^2}$$

  As $|t_{ij} - t_{ij'}| \to 0$, $\text{corr}(Y_{ij}, Y_{ij'}) \to 1$.
  As $|t_{ij} - t_{ij'}| \to \infty$, $\text{corr}(Y_{ij}, Y_{ij'}) \to \gamma^2/[\sigma^2 + \gamma^2]$.

- Using only component to define the dependence is more common but less realistic.

- As we showed earlier, the two-component does not fit well, so we will need to consider a more flexible model using **Bayesian Nonparametrics**.

Bayesian Nonparametrics

Each observation i depends on some random parameter θ_i.

$$Y_i \sim \text{indep } p(y; \theta_i, \boldsymbol{\phi})$$

Typically, we choose a **parametric** distribution for the θ_is, such as a normal distribution. But we may think this is restrictive or inappropriate.

Instead, we will let the distribution for θ be a random variable:

$$\theta_i \sim \text{indep } F, \qquad F \sim \mathcal{P},$$

where $\mathcal{P}$ is a distribution on the set of distributions on Θ.

The choice of $\mathcal{P}$ that we make is the **Dirichlet Process**.

The Dirichlet Process

Definition

Let G be a probability distribution on Θ and $\alpha > 0$. Then F has the **Dirichlet Process** distribution with parameter αG if for every finite partition $A_1, \ldots, A_k$ of Θ,

$$(F\{A_1\}, \ldots, F\{A_k\}) \sim \text{Dir}\left(\alpha G\{A_1\}, \ldots, \alpha G\{A_k\}\right).$$

We denote this as $F \sim \text{DP}(\alpha G)$.

See Hjort et al. (2010).

The Dirichlet Process

A more constructive definition:

$$\begin{aligned}
\theta_1, \theta_2, \theta_3, \ldots &\sim \text{iid } G(\cdot) \\
V_1, V_2, V_3, \ldots &\sim \text{iid Beta}(1, \alpha) \\
\psi_h &= V_h \prod_{l=1}^{h-1} (1 - V_l) \\
F(\cdot) &= \sum_{h=1}^{\infty} \psi_h \delta_{\theta_h}(\cdot)
\end{aligned}$$

Then, $F \sim \text{DP}(\alpha G)$.

We call this the **stick-breaking representation** due to the way we form the ψ_h's. (Sethuraman, 1994)

The Dirichlet Process

Some important properties of the Dirichlet Process:

Let $A \subset \Theta$, $F \sim \text{DP}(\alpha G)$, and $\theta_1, \ldots, \theta_n \sim$ iid F. Then:

- $\text{E}\left(F\{A\}\right) = G\{A\}$
- $\text{Var}\left(F\{A\}\right) = \frac{1}{1+\alpha} G\{A\}\left(1 - G\{A\}\right)$
- F is discrete with probability 1.
- The DP promotes clustering: $\Pr(\theta_i = \theta_j) = 1/(1+\alpha)$.
- The DP is a conjugate distribution.
  Let $\mathbb{F}_n$ be the empirical distribution function. The posterior distribution of F is DP with parameter $\alpha G + n\mathbb{F}_n$.
- The DP is "easy" to use in an MCMC algorithm.

$$p\left(\theta_i | \boldsymbol{\theta}_{(-i)}, y\right) \propto p(y_i; \theta_i, \boldsymbol{\phi}) \left[\sum_{i' \neq i} \frac{1}{\alpha + n - 1} \delta_{\theta_{i'}}(\theta_i) + \frac{\alpha}{\alpha + n - 1} g(\theta_i)\right].$$

Bayesian Nonparametric Model for Longitudinal Trajectory

We extend the 2-component model to a Bayesian Nonparametric version:

$$
\begin{aligned}
(\boldsymbol{\theta}_i, \phi_i) &\sim \text{DP}\left(\alpha,\ \text{MVN}_3(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}) \otimes \text{Beta}(a, b)\right) \\
D_i &\sim \text{Bern}(\phi_i) \\
Y_{ij} &= f(t_{ij}; \boldsymbol{\theta}_i) + \gamma_i + \epsilon_{ij} \\
\gamma_i &\sim \text{N}(0, \gamma^2) \\
\boldsymbol{\epsilon}_i &\sim \text{MVN}_{n_i}\left(\mathbf{0}_{n_i}, \sigma^2 \mathbf{R}_i(\rho)\right),
\end{aligned}
$$

Now each patient has her own "unique" parameter vector $(\boldsymbol{\theta}_i, \phi_i)$.

Due to the DP choice, there are many ties in the $(\boldsymbol{\theta}_i, \phi_i)$s across patients. In essence, we have a small number of clusters with unique values of these parameters.

Bayesian Nonparametric Model for Longitudinal Trajectory

Let $c_i \in \{1, 2, 3, \ldots\}$ be the cluster for patient i, we can equivalently express the model as:

$$
\begin{aligned}
\Pr(c_i = k) &= \psi_k = V_k \prod_{j=1}^{k-1} (1 - V_j) \\
D_i | c_i &\sim \text{Bern}(\phi^{(c_i)}) \\
Y_{ij} | c_i &= f(t_{ij}; \boldsymbol{\theta}_i^{(c_i)}) + \gamma_i + \epsilon_{ij} \\
\gamma_i &\sim \text{N}(0, \gamma^2) \\
\boldsymbol{\epsilon}_i &\sim \text{MVN}_{n_i}\left(\mathbf{0}_{n_i}, \sigma^2 \mathbf{R}_i\right) \\
\\
\boldsymbol{\theta}^{(k)} &\sim \text{MVN}_3(\boldsymbol{\theta}^\star, \boldsymbol{\Sigma}) \\
\phi^{(k)} &\sim \text{Beta}(a, b) \\
V_k &\sim \text{Beta}(1, \alpha)
\end{aligned}
$$

Bayesian Nonparametric Model for Longitudinal Trajectory

A few comments:

- This model will allow multiple types of clusters. Clusters with $\phi^{(c)}$ near 0 will contain mainly healthy patients and clusters $\phi^{(c)}$ near 1 mainly diseased patients.
- Clusters with $\phi^{(c)}$ in the middle contain both healthy and diseased patients with similar trajectories $f(t; \boldsymbol{\theta}^{(c)})$.
- A computational algorithm will not allow infinitely many clusters. We truncate so that we only have H of the cluster-specific parameters: $\phi^{(c)}, \boldsymbol{\theta}^{(c)}, V_c, \psi_c$. H should be large enough that we frequently have empty clusters.
- Even though clusters have unique heights $\theta_1^{(c)}$, the random intercept γ_i is helpful in avoiding outlier clusters.

Bayesian Nonparametric Model for Longitudinal Trajectory

- If we observe a patient's longitudinal trajectory $\mathbf{y}$, we can predict disease status as follows.

$$\begin{aligned}\Pr(D=1\,|\,\mathbf{y}) &= \mathsf{E}\{I(D=1)\,|\,\mathbf{y}\} = \mathsf{E}\left[\mathsf{E}\left\{I(D=1)\,|\,C=k,\mathbf{y}\right\}\,|\,\mathbf{y}\right] \\ &= \sum_{k=1}^{H}\phi^{(k)}\Pr(C=k\,|\,\mathbf{y}) \\ &= \sum_{k=1}^{H}\phi^{(k)}\frac{\psi_k\ \mathsf{MVN}\left(\mathbf{y};f(\mathbf{t};\boldsymbol{\theta}^{(k)}),\sigma^2\mathbf{R}_i(\rho)+\gamma^2\mathbf{1}\mathbf{1}'\right)}{\sum_{h=1}^{H}\psi_h\ \mathsf{MVN}\left(\mathbf{y};f(\mathbf{t};\boldsymbol{\theta}^{(h)}),\sigma^2\mathbf{R}_i(\rho)+\gamma^2\mathbf{1}\mathbf{1}'\right)}\end{aligned}$$

- These probabilities may or may not be well-calibrated. Regardless, we can treat them as a score with high values indicative of disease.

Bayesian Nonparametric Model for Longitudinal Trajectory

We will also need to specify priors for the remaining distributions:

$$
\begin{aligned}
\alpha &\sim \text{Gamma}(1, 1) \\
\gamma^2 &\sim \text{InvGamma}(0.1, 0.1) \\
\sigma^2 &\sim \text{InvGamma}(0.1, 0.1) \\
\rho &\sim \text{Unif}(0, 1) \\
\boldsymbol{\theta}^\star &\sim \text{MVN}_3(\mathbf{1}_3, 10^2\mathbf{I}_3) \\
\boldsymbol{\Sigma} &\sim \text{InvWish}(5, \mathbf{I}_3) \\
& \quad a = b = 0.5
\end{aligned}
$$

Markov chain Monte Carlo sampling

In order to obtain any posterior inference, we will need a Markov chain Monte Carlo (MCMC) sampling algorithm.

- In every iteration, we cycle through all parameters updating each given the values of all others.
- We sample from the model specification that explicitly includes the cluster indicators C_i.
- We perform conditionally conjugate sampling steps for the following parameters: C_i from multinomial, $\phi^{(c)}$ from beta, $\boldsymbol{\theta}^\star$ from MVN_3, $\boldsymbol{\Sigma}$ from InvWish, V_k from beta, and α from gamma.
- We perform adaptive Metropolis-Hastings sampling for the conjugate parameters: $\boldsymbol{\theta}^{(c)}$, ρ, γ^2, σ^2.

Label switching and parameter identifiability

However, there are some challenges to drawing inference in our model.

- **Label Switching**: The likelihood function is invariant to permutations of the labels. If I switch the names of cluster 1 and 2 to clusters 2 and 1, the likelihood function is the same (Stephens, 2000).
- If our MCMC chain is mixing well, we should *expect* to see label switches.
- But can we perform inference? I can't just average over $\phi^{(1)}$ over iterations, because sometimes cluster 1 may correspond to a disease cluster and sometimes to a healthy cluster.
- While we have a total of H potential clusters in each iteration, many of these are empty. The number of non-empty clusters is also changing every iteration.

Label switching and parameter identifiability

An example of label switching:

Obviously, cluster-defined parameters are not identifiable, but these two parameters are estimable:

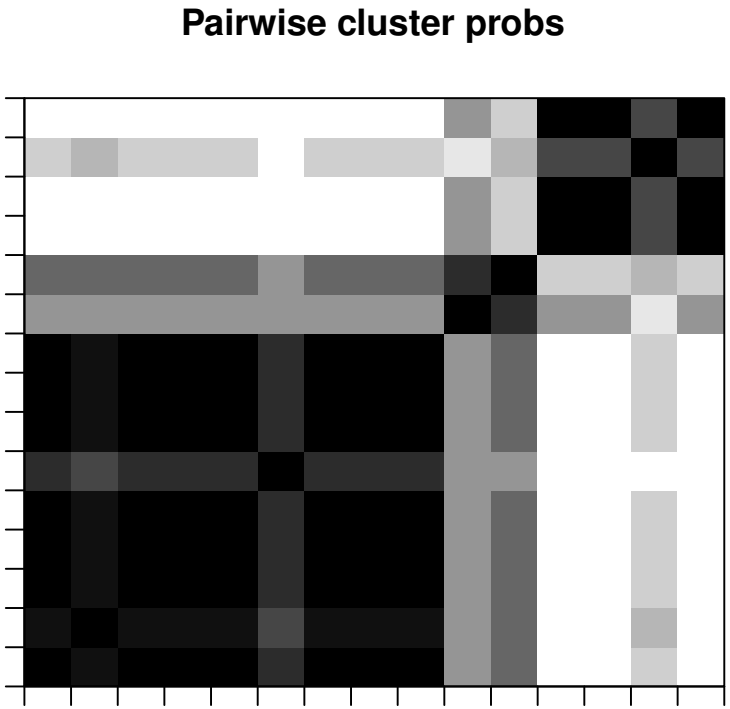
- Prob. two patients are in the same cluster: $\Pr(C_i = C_j \,|\, \mathbf{y}_i, \mathbf{y}_j)$
- Disease predictions:

$$\Pr(D = 1 \,|\, \mathbf{y}) = \sum_{k=1}^{H} \phi^{(k)} \frac{\psi_k \ \text{MVN}\left(\mathbf{y}; f(\mathbf{t}; \boldsymbol{\theta}^{(k)}), \sigma^2 \mathbf{R}_i(\rho) + \gamma^2 \mathbf{1}\mathbf{1}'\right)}{\sum_{h=1}^{H} \psi_h \ \text{MVN}\left(\mathbf{y}; f(\mathbf{t}; \boldsymbol{\theta}^{(h)}), \sigma^2 \mathbf{R}_i(\rho) + \gamma^2 \mathbf{1}\mathbf{1}'\right)}$$

Label switching and parameter identifiability

- If all we need is to make predictions about disease status, then label switching is not an issue.
- But if we want to make conclusions about particular trajectories, we will need posterior samples with identifiable parameters.

- Based on the posterior pairwise probabilities (right), we estimate an **optimal partition**, the mostly like clustering configuration.
- Given the optimal partition, we can rerun MCMC without updating the cluster membership to get a usable posterior sample.



Estimating the optimal partition

Dahl's Method (Dahl, 2006)
Find $\hat{\mathbf{c}} = (\hat{c}_1, \hat{c}_2, \ldots, \hat{c}_n)$ from MCMC sample that minimizes

$$\sum_{i=1}^{N} \sum_{j=1}^{N} \left[I(\hat{c}_i = \hat{c}_j) - \Pr(C_i = C_j \mid \mathbf{y}) \right]^2 .$$

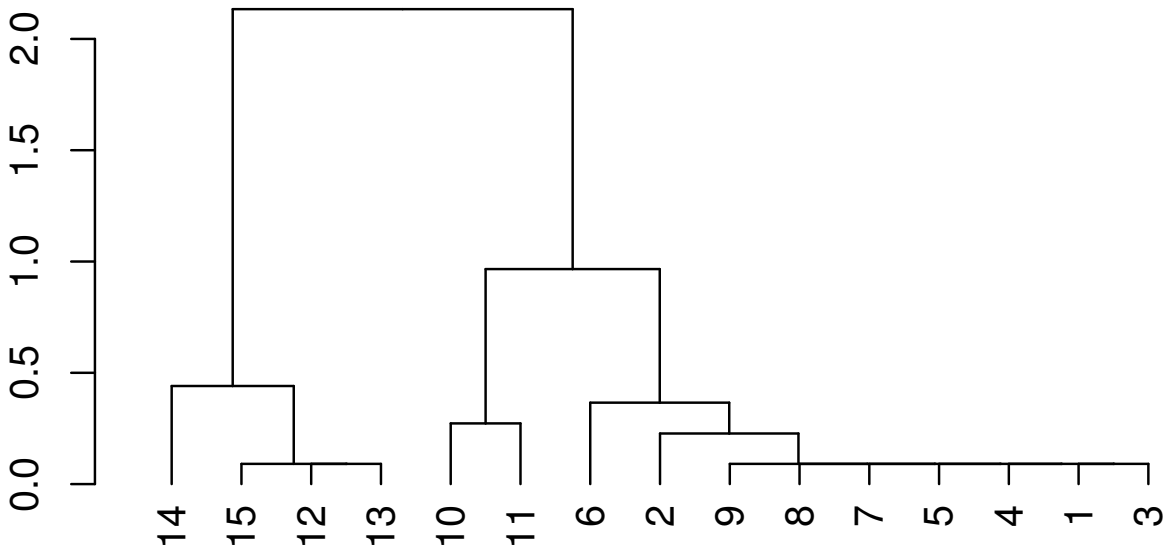
We would estimate the optimal clustering to be

We can then run a new MCMC chain using these labels. The parameters $(\boldsymbol{\theta}^{(1)}, \phi^{(1)})$ correspond to cluster #1 with patients 1–9.

Estimating the optimal partition

Hierarchical Clustering

Using the pairwise clustering probabilities $\Pr(C_i = C_j \mid \mathbf{y}_i, \mathbf{y}_j)$, we can obtain a **dendrogram** describing the clustering relationships.



There are various choices of the linkage criteria that can be used, and we consider the average linkage and Ward's method.

Estimating the optimal partition

Hierarchical Clusterings

This gives us a partition for each number of clusters k, but we still have to choose the value k.

- Let $\hat{k}$ be the median number of non-empty clusters from our MCMC sample
- Under average linkage, dendrogram height h represents that for i and j assigned to different clusters

$$\Pr(C_i \neq C_j \mid \mathbf{y}) \geq h,$$

  so we can specify a value of h, such as 0.75 or 0.9.
- Some automated methods designed to choose k based on minimizing some criteria: Gamma measure, Tau measure, Silhouette index (Charrad et al, 2014).

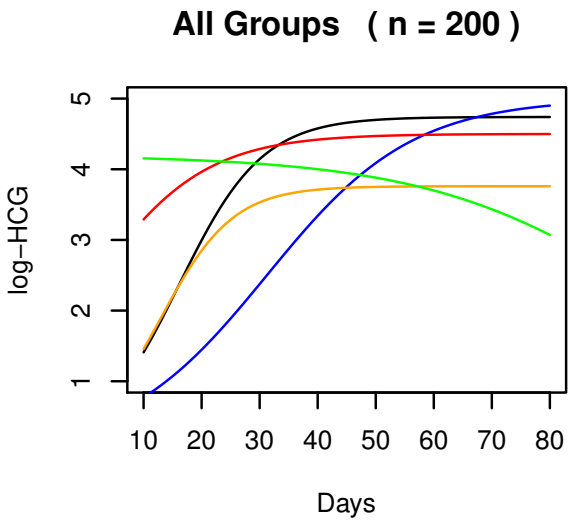
Model Choices

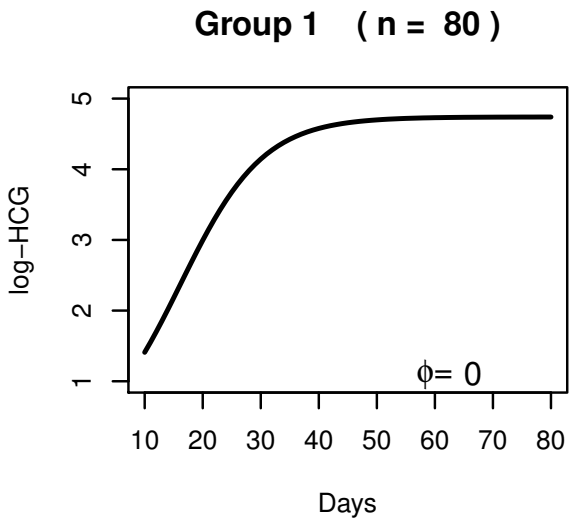
Different Models under consideration:

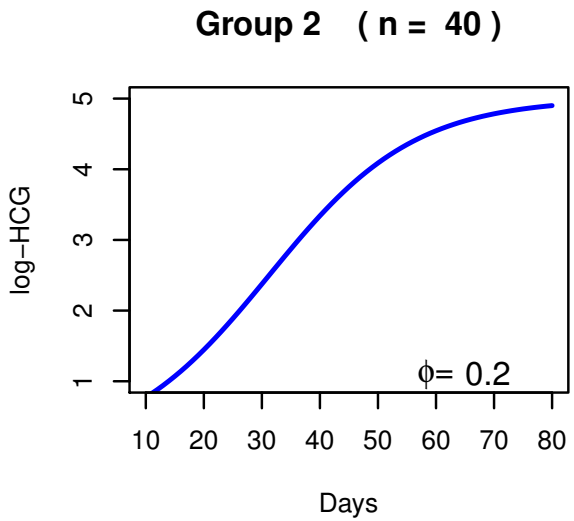
1. Two-component model
2. Bayesian Nonparametric model with label switching
   Disease prediction through Bayesian model averaging (BMA)
3. Bayesian Nonparametric model with 2-stage estimation
   Choose optimal clustering by Dahl's method
4. Bayesian Nonparametric model with 2-stage estimation
   Choose optimal clustering through hierarchical clustering under each method for choosing k

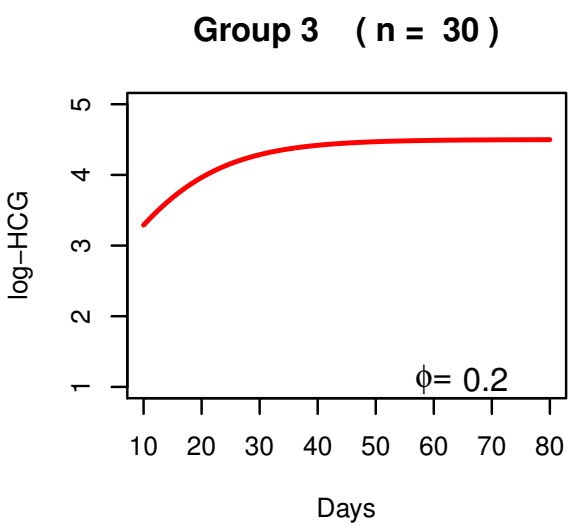
Because the number of clusters and the membership are unknown, the BMA choice most accurately represents what we can learn from the data. Theoretically, predictions under BMA should be best (lowest variance) by Rao-Blackwell considerations.

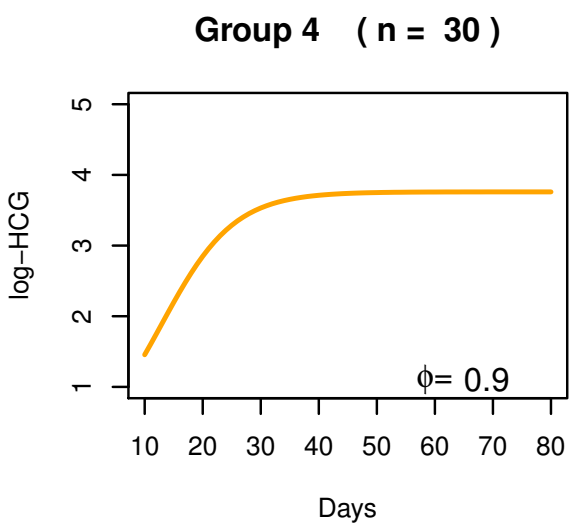
Simulation Study #1

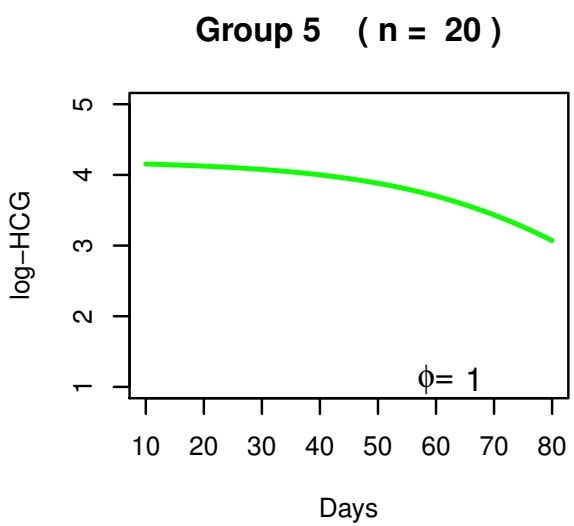




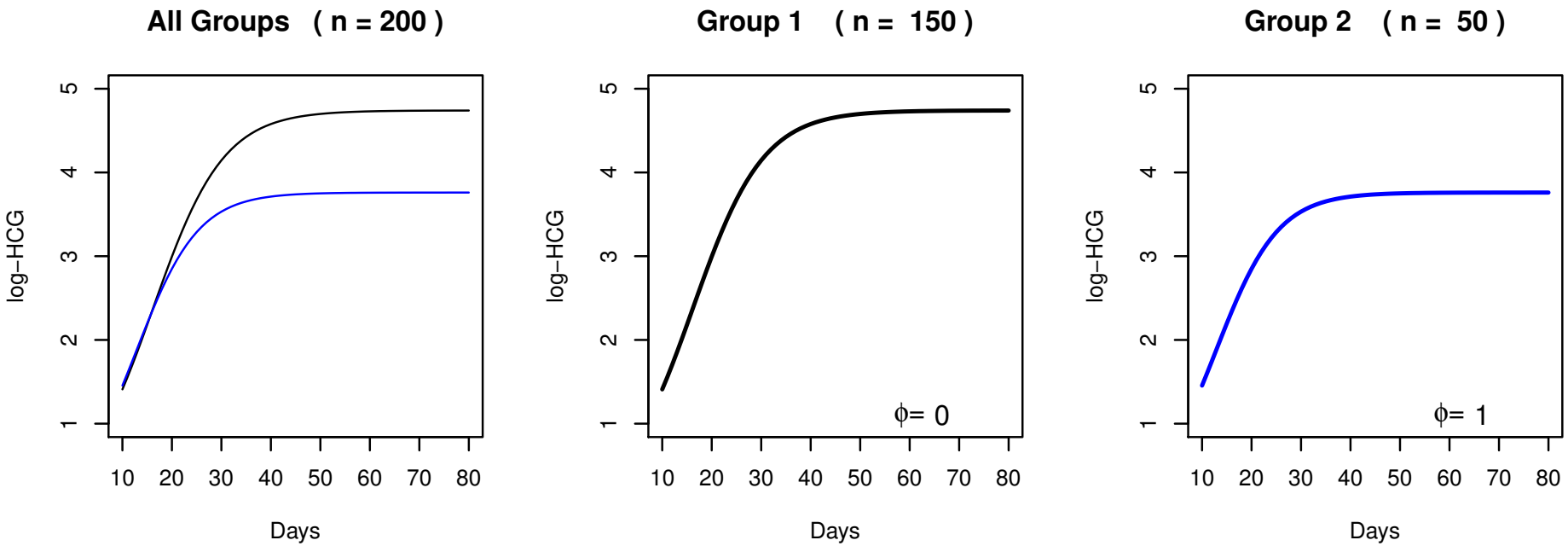








Simulation Study #2



In each case, we generate 200 datasets with 200 observations and apply each of the methods.

We estimate within sample accuracy, as well as out of sample accuracy from an additional set of 25,000 patients.

Simulation Study Results

Model	Clusters	Out of sample % error	Out of sample AUC	Within sample % error	Within sample AUC
Simulation Study #1 (5 components)					
2-component	2	$25.4\%_{(0.9)}$	$0.768_{(.013)}$	$25.0\%_{(2.5)}$	$0.783_{(.013)}$
BMA	$8.1_{(1.1)}$	$21.7\%_{(0.7)}$	$0.823_{(.007)}$	$20.1\%_{(3.0)}$	$0.848_{(.029)}$
Dahl	$9.0_{(2.9)}$	$22.6\%_{(0.9)}$	$0.813_{(.009)}$	$20.2\%_{(2.9)}$	$0.844_{(.030)}$
Avg($h = .75$)	$5.0_{(1.1)}$	$22.8\%_{(1.5)}$	$0.809_{(.022)}$	$20.7\%_{(3.3)}$	$0.837_{(.035)}$
Avg(median)	$7.7_{(1.1)}$	$22.9\%_{(1.8)}$	$0.806_{(.033)}$	$20.6\%_{(3.3)}$	$0.837_{(.043)}$
Avg(Silhouette)	$4.2_{(0.9)}$	$23.0\%_{(2.2)}$	$0.806_{(.018)}$	$21.2\%_{(3.2)}$	$0.831_{(.036)}$
⋮					
Simulation Study #2 (2 components)					
2-component	2	$18.2\%_{(1.0)}$	$0.853_{(.007)}$	$16.0\%_{(2.1)}$	$0.877_{(.027)}$
BMA	$3.0_{(0.6)}$	$18.0\%_{(1.0)}$	$0.854_{(.008)}$	$16.0\%_{(2.0)}$	$0.875_{(.028)}$
Dahl	$2.5_{(0.8)}$	$18.0\%_{(1.0)}$	$0.856_{(.008)}$	$16.1\%_{(2.2)}$	$0.875_{(.027)}$
Avg($h = .75$)	$2.1_{(0.2)}$	$18.0\%_{(1.1)}$	$0.857_{(.009)}$	$16.1\%_{(2.0)}$	$0.874_{(.028)}$
Avg(median)	$2.7_{(0.7)}$	$18.2\%_{(1.6)}$	$0.851_{(.032)}$	$16.3\%_{(2.7)}$	$0.871_{(.045)}$
Avg(Silhouette)	$2.1_{(0.4)}$	$18.0\%_{(1.1)}$	$0.856_{(.009)}$	$16.1\%_{(2.2)}$	$0.874_{(.029)}$
⋮					

Simulation Study Results

Comments:

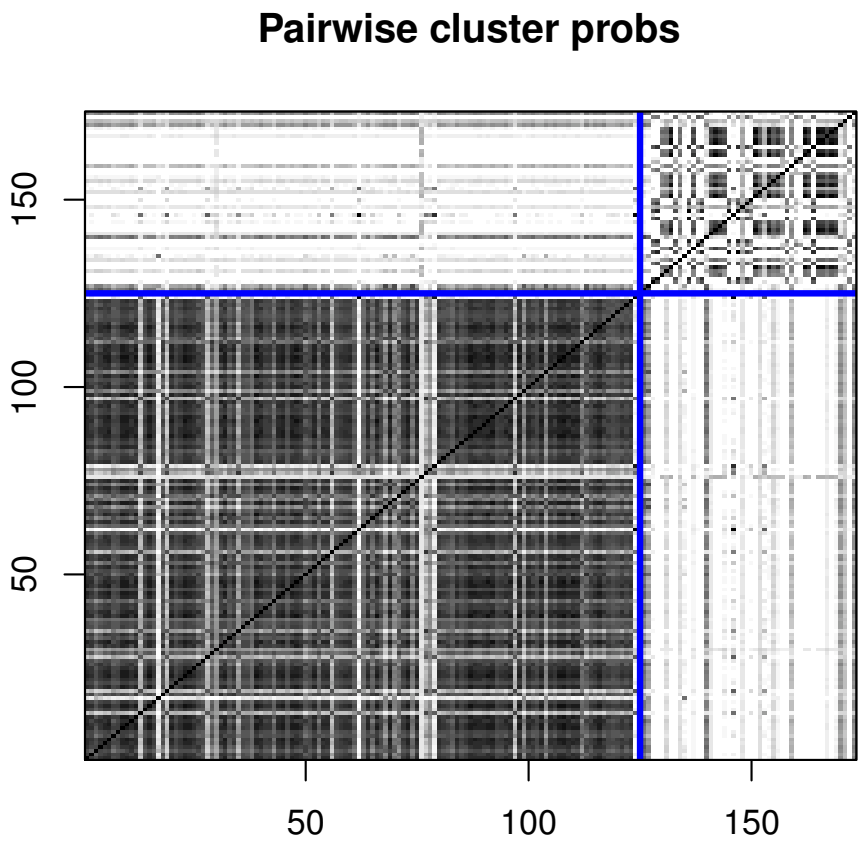
- If the two-component model is not correct, it produces substantially worse predictions; when two-component model is correct, BMA and two-stage BNP estimators do just as well.
- BMA beats the two-stage BNP predictions but not by too much. The minor loss in prediction accuracy may be justified by more clear interpretations.
- The Dahl method is among the best two-stage methods in prediction but tends to have more clusters. Estimating k from the median of the MCMC chain also produces many clusters with few observations.

We recommend prediction using the BMA estimates and interpretation through two-stage procedure under the average linkage with fixed $h = 0.75$, followed by the Silhouette index (either average or Ward linkage) and Dahl's method.

Data Application: Assisted pregnancy in Chilean women

We now consider the data introduced earlier regarding pregnancy outcome in 173 Chilean women undergoing ART (Marshall and Baron, 2000).

- We run the BNP model MCMC scheme for 80,000 iteration after a burn-in of 2000 and store every 40th value.
- There are an average of 8.6 (95% CI: 6–12) non-empty clusters with $\hat{k} = 8$.



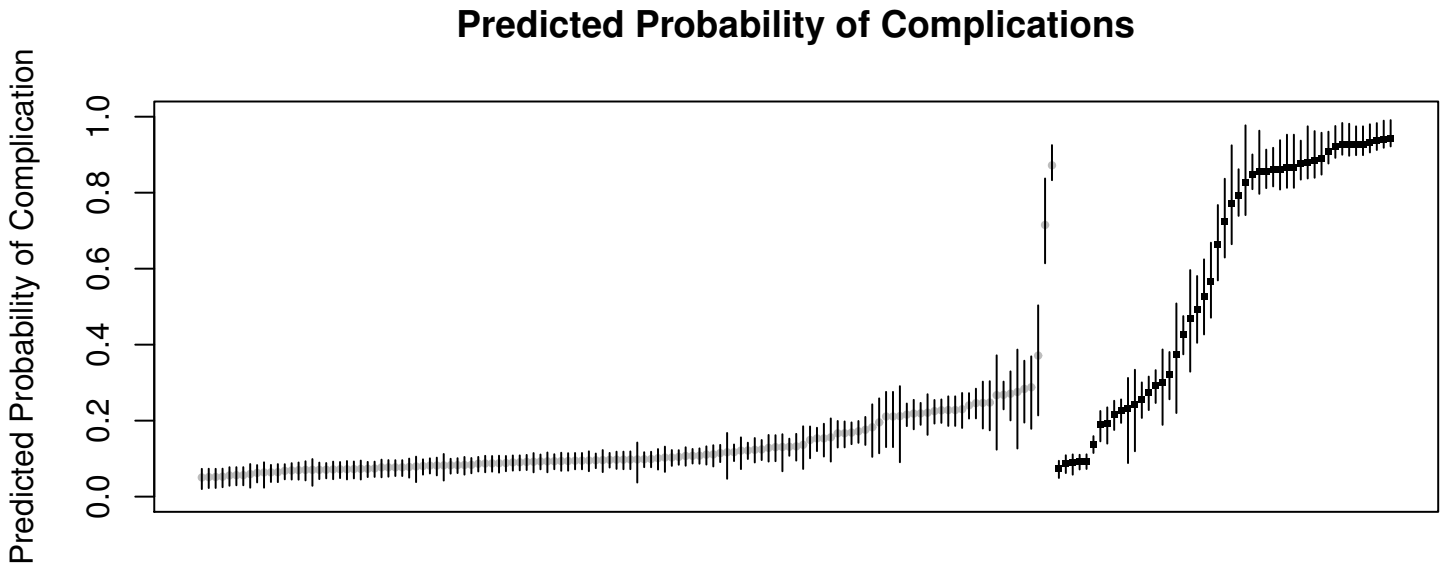
Data Application: Assisted pregnancy in Chilean women

We classify a patient as diseased if $\Pr(D_i = 1 \,|\, \mathbf{y}_i) > 0.5$. We also compute the area under the curve (AUC) from the ROC curve.

We estimate the accuracy using within sample prediction from the full data ($n = 173$) and from a 25-fold cross validation. For CV, we withhold a random 35 patients and predict their disease status after two-stage model fitting with the remaining 138 patients.

		Full data		25-fold CV	
Model	Clusters	% error	AUC	% error	AUC
2-component	2	16.2%	0.863	$19.5\%_{(1.2)}$	$0.865_{(.004)}$
BMA	8.6	13.3%	0.900	$16.2\%_{(1.1)}$	$0.900_{(.003)}$
Dahl	10	12.7%	0.898	$16.8\%_{(1.0)}$	$0.888_{(.004)}$
Avg($h = 0.75$)	5	13.3%	0.883	$17.2\%_{(1.4)}$	$0.881_{(.005)}$
Avg(Silhouette)	5	13.3%	0.880	$16.3\%_{(1.0)}$	$0.890_{(.005)}$
Ward(Silhouette)	3	14.5%	0.889	$17.0\%_{(1.0)}$	$0.884_{(.005)}$

Data Application: Assisted pregnancy in Chilean women

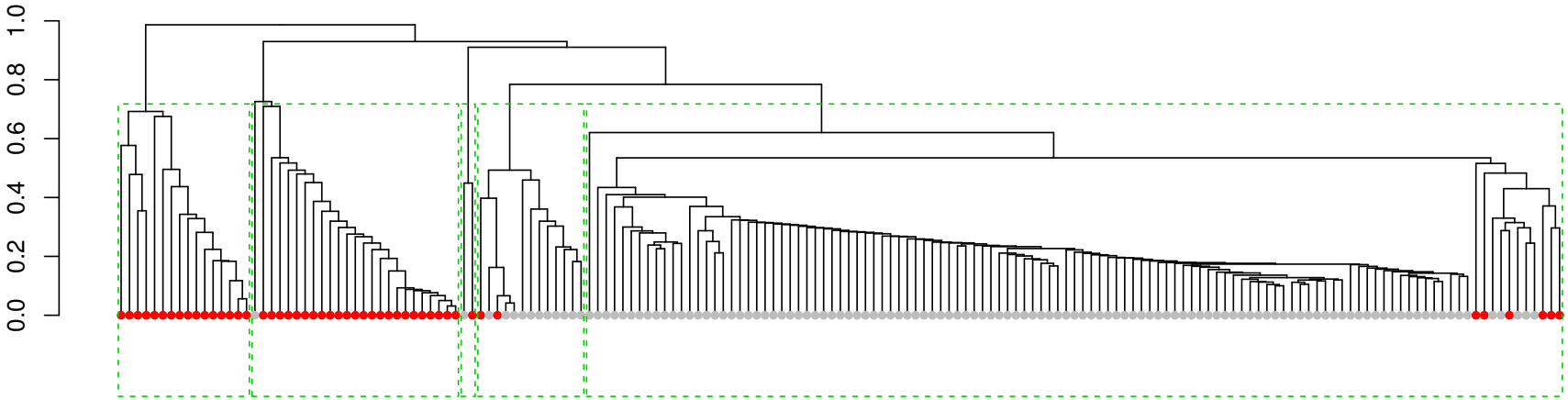


- Classification based on $\Pr(D_i = 1 \,|\, \mathbf{y}_i) > 0.5$ from BMA has 98% specificity but only 57% sensitivity.
- Using the ROC curve to find the best threshold for minimizing Type I and Type II error suggests classifying based on

$$\text{pr}(D_i = 1 \,|\, \mathbf{y}_i) > 0.23,$$

which has 90% specificity and 80% sensitivity.

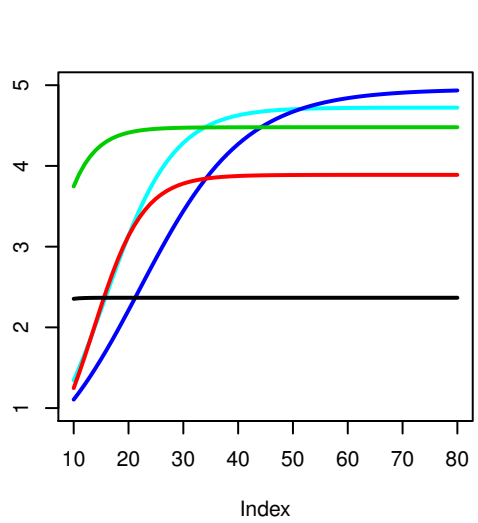
Data Application: Assisted pregnancy in Chilean women

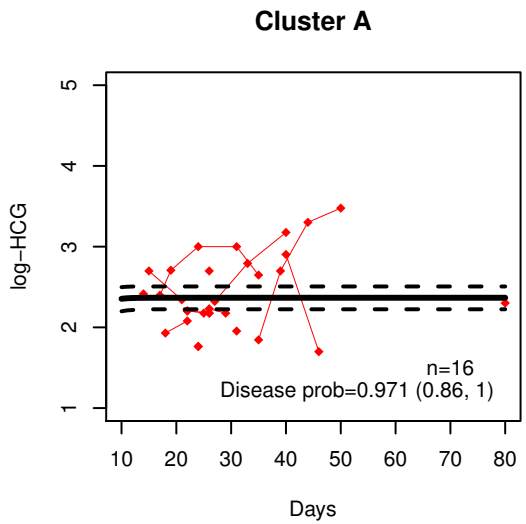


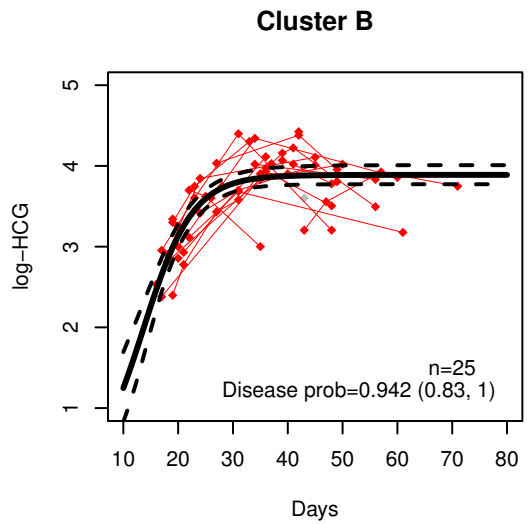
- Optimal partition estimated from the Sihouette index using the average linkage.
- The BNP model seems to do a good job of distinguishing the healthy from the disease patients.
- We can apply the stage 2 approach and refit the MCMC chain to get interpretable results based on this optimal partition.

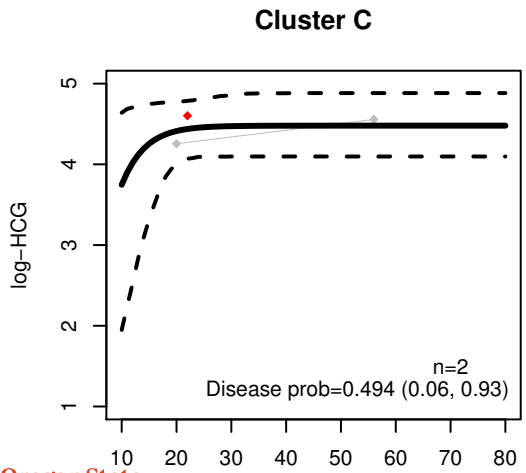
Data Application: Assisted pregnancy in Chilean women

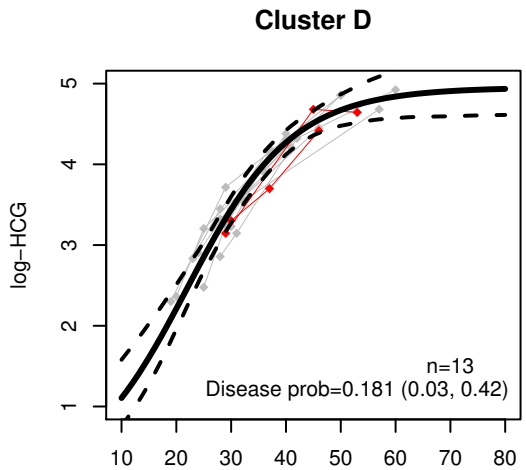
We consider the interpretable conclusions based on the Stage II model fits from the Silhouette index with average linkage.

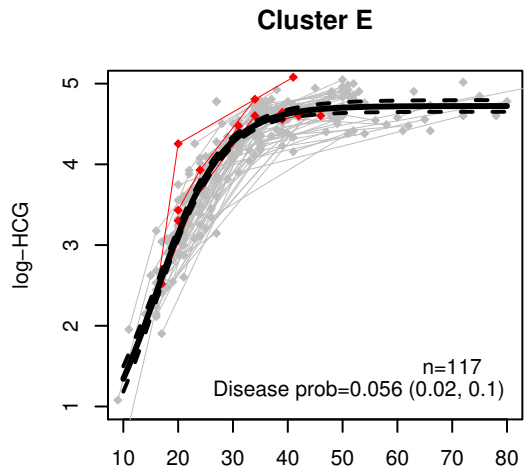












Conclusions

- We proposed a Bayesian nonparametric disease classification model and developed a computationally efficient MCMC scheme for posterior inference.
- We explored and compared methods that can be used to describe cluster-specific parameters in a systematic way.
- In predicting pregnancy outcomes for Chilean ART patients, our model has better predictive performance with 83.8% accuracy, relative to the 2-component model at 80.5% accuracy.
- The loss of information regarding the uncertainty in cluster membership and number does not have a significant impact on prediction and leads to interpretable results.
- A simulation study further confirms these, even when the true model is 2-component.
- If you are interested please check our paper:
  Gaskins, J., Fuentes, C., De La Cruz, R. (2023). "A Bayesian nonparametric model for classification of longitudinal profiles". Biostatistics 24 (1), 209-225.

fuentesc@oregonstate.edu

Thank You!!