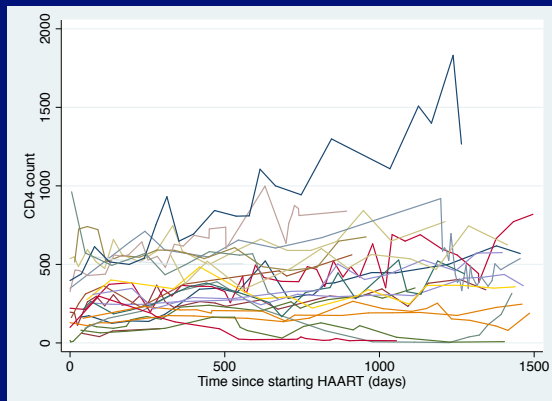


Longitudinal Data

Fall 2015



Chapter 9

Latent Trajectory Models

Instructors

Nick Jewell (jewell@berkeley.edu)



GSI

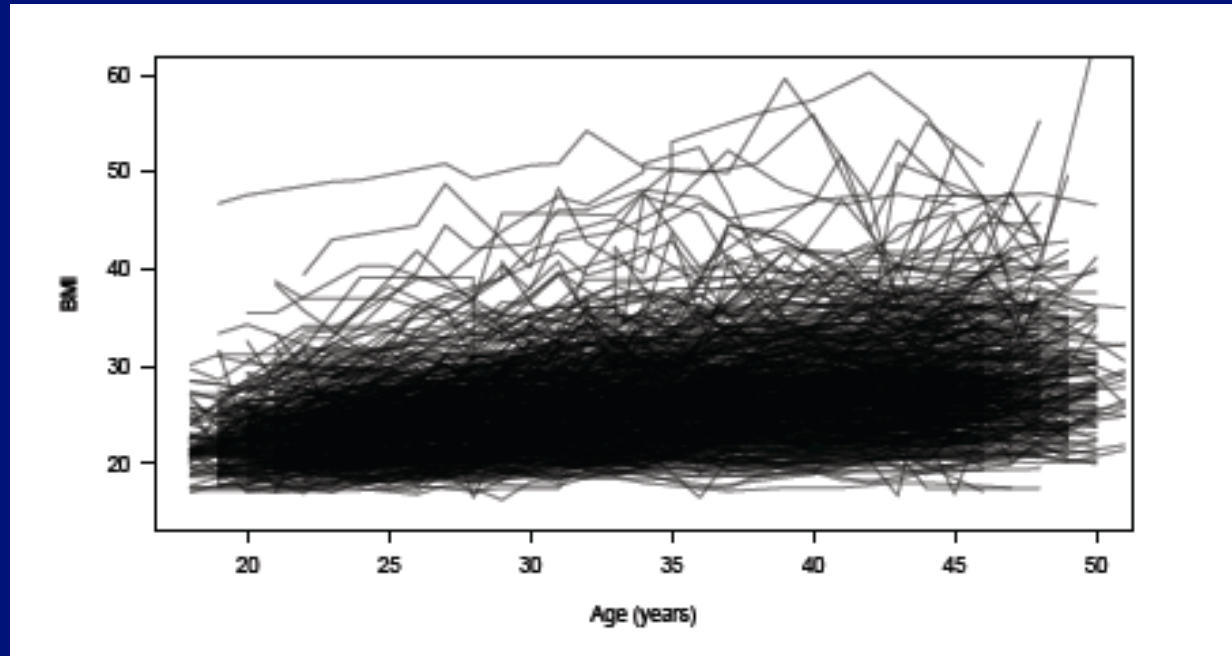
Robin Mejia (mejia@nasw.org)

“Understanding our world requires conceptualizing the similarities and differences between the entities that compose it”

Robert Tryon and Daniel Bailey, 1970

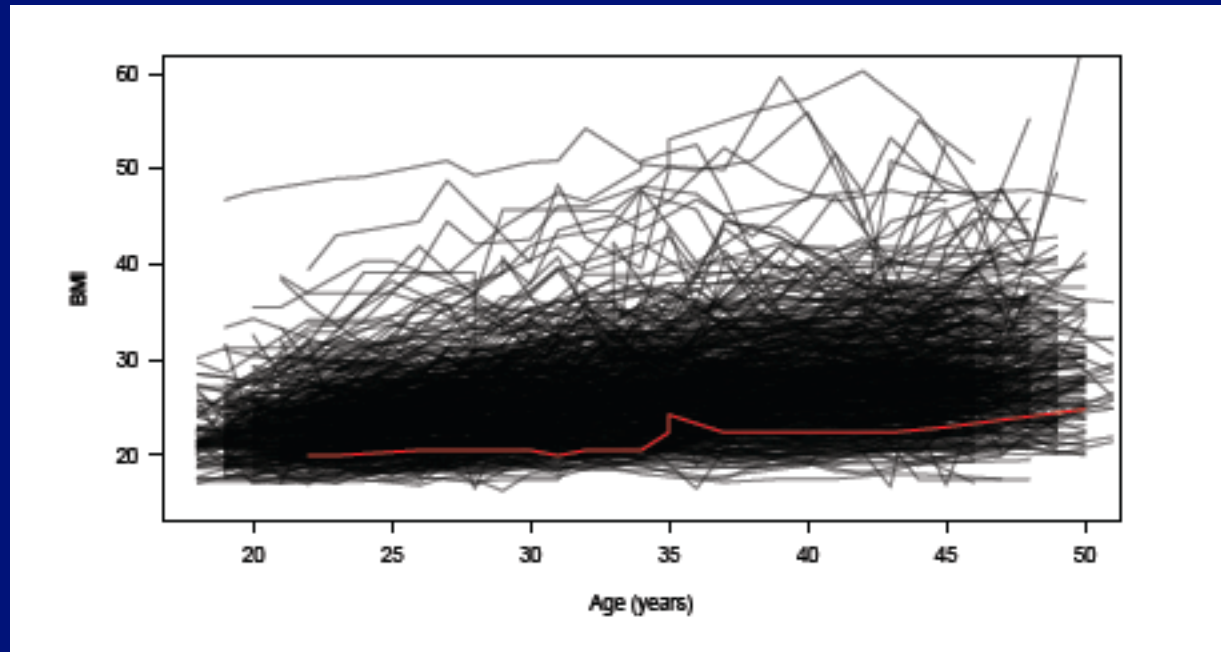
Mixed effects models fail to model inter-individual variability that is not explained by known and observed covariates (and can struggle even with known covariates)

How does BMI change with age?



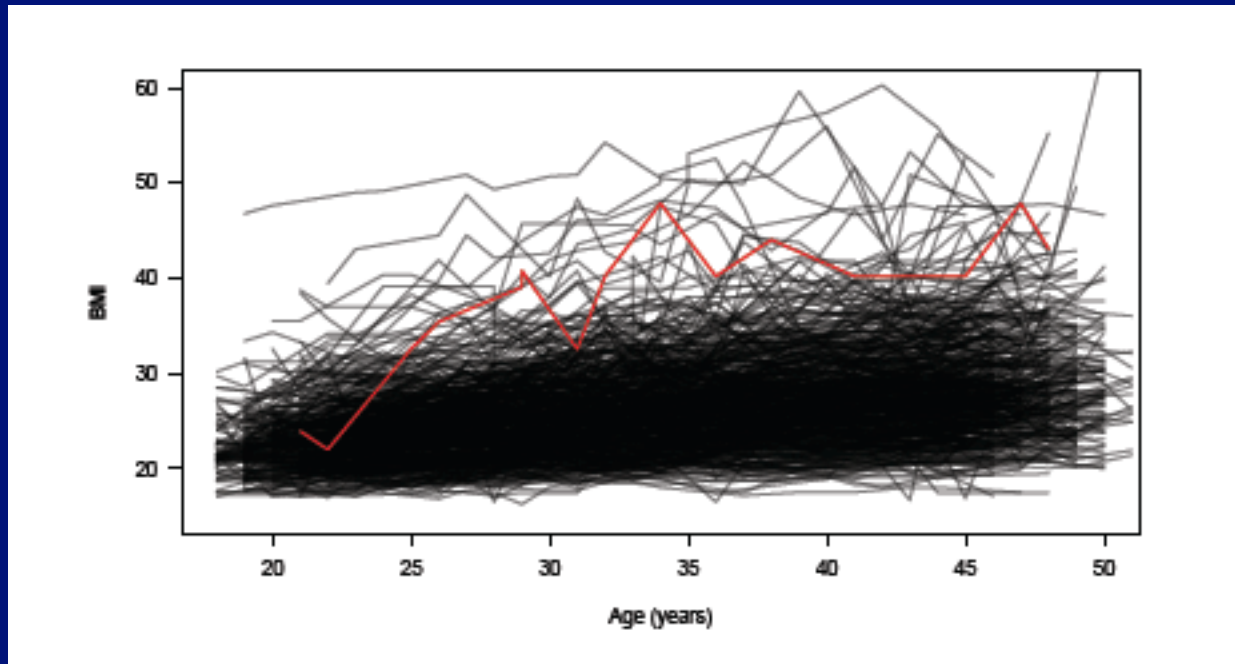
National Longitudinal Study of Youth (NLSY) from 1979 - 2008.

How does BMI change with age?



National Longitudinal Study of Youth (NLSY) from 1979 - 2008.

How does BMI change with age?



National Longitudinal Study of Youth (NLSY) from 1979 - 2008.

Finite Mixture Models (Approach 1)

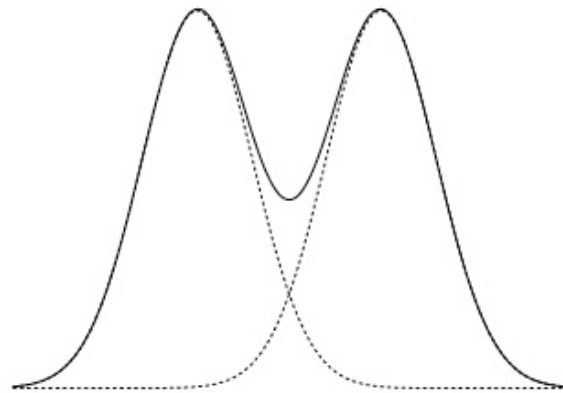


Figure 8.1: Gaussian mixture model involving two subgroups with different means and equal spread.

Solid line is what you see—can you infer the dotted line information?

(That is, the mean (and ultimately regression model) for the two different subgroups, what fraction of the population belong to each subgroup, and description of the subgroups in terms of known covariates.)

Finite Mixture Models (Approach 1)


$$f(\mathbf{Y}_i) = \sum_{k=1}^K \pi_k f_k(\mathbf{Y}_i)$$

$$\pi_k \leq 1 \quad \sum_{k=1}^K \pi_k = 1$$

Typical Longitudinal Analysis

- Use Generalized Estimating Equations (GEE) to estimate the mean outcome, and how it changes over time, adjusting for covariates
 - regression parameter estimation is consistent despite potential covariance misspecification
 - efficiency can be gained through use of a more appropriate working correlation structure
 - robust (sandwich) standard error estimators available
- But, with a heterogeneous population,
 - BMI does not change much for some people as they age
 - BMI changes considerably for some people as they age
- We don't wish to average out these separate trajectories by modeling the mean over time

Finite Mixture Models

- **Data for n individuals:** $\mathbf{y}_i = (y_{i1}, \dots, y_{im_i})$ measured at times $t_i = (t_{i1}, \dots, t_{im_i})$
- **We assume K latent trajectories in the population that are distributed with frequencies:** π_1, \dots, π_K where $\pi_k > 0$ and $\sum_{k=1}^K \pi_k = 1$.
 $f(\mathbf{y}|\mathbf{t}, \theta) = \pi_1 \mathbf{f}(\mathbf{y}|\mathbf{t}, \beta_1, \Sigma_1) + \dots + \pi_K \mathbf{f}(\mathbf{y}|\mathbf{t}, \beta_K, \Sigma_K)$ 
- **The (conditional) mixture density is** $f(\mathbf{y}|\mathbf{t}, \beta_k, \Sigma_k)$, **a multivariate Gaussian with mean** μ_k **and covariance** Σ_k .
- **In most trajectory software, (conditional) independence is assumed as a working correlations structure:** $(\Sigma_k = \sigma_k^2 I)$.



$$\theta = (\pi_1, \dots, \pi_K; \beta_1, \dots, \beta_K; \Sigma_1, \dots, \Sigma_K)$$

How does this relate to our discussion of mixed models thus far?

- Before, we discussed random effects (latent variables) that normal distributions, e.g., $\beta_{0i} \sim N(0, \sigma^2)$.
- This model has discrete latent variables, say $C=1, \dots, K$ with simple multinomial distribution defined by probabilities of being in each class: $P(C=k) = \pi_k$.
- So, another generalization of the mixed model that is convenient for modeling the data as derived from a random set of distinct groups (with distinct trajectories).
- These latent class models are convenient for clustering problems, and so are the basis of various model-based clustering routines.


Finite Mixture Models

- The mean vector μ_k is related to the observation times as follows:
 - Linear: $(\mu_k)_j = \beta_0 + \beta_1 t_{ij}$
 - Quadratic: $(\mu_k)_j = \beta_0 + \beta_1 t_{ij} + \beta_2 t_{ij}^2$
 - Splines in observation times

where the regression model (and coefficients) are assumed the same for each cluster, and t_{ij} is the j^{th} observation for the i^{th} individual where $1 \leq j \leq m_i$

Finite Mixture Models



- **Group membership:** $\pi_k = \frac{\exp(\gamma_k z)}{\sum_{j=1}^K \exp(\gamma_j z)}$ 

Z is set of same or different covariates

This expands θ to include the γ s also

Estimation for Mixture Models

- Maximum likelihood estimation for θ via the EM algorithm



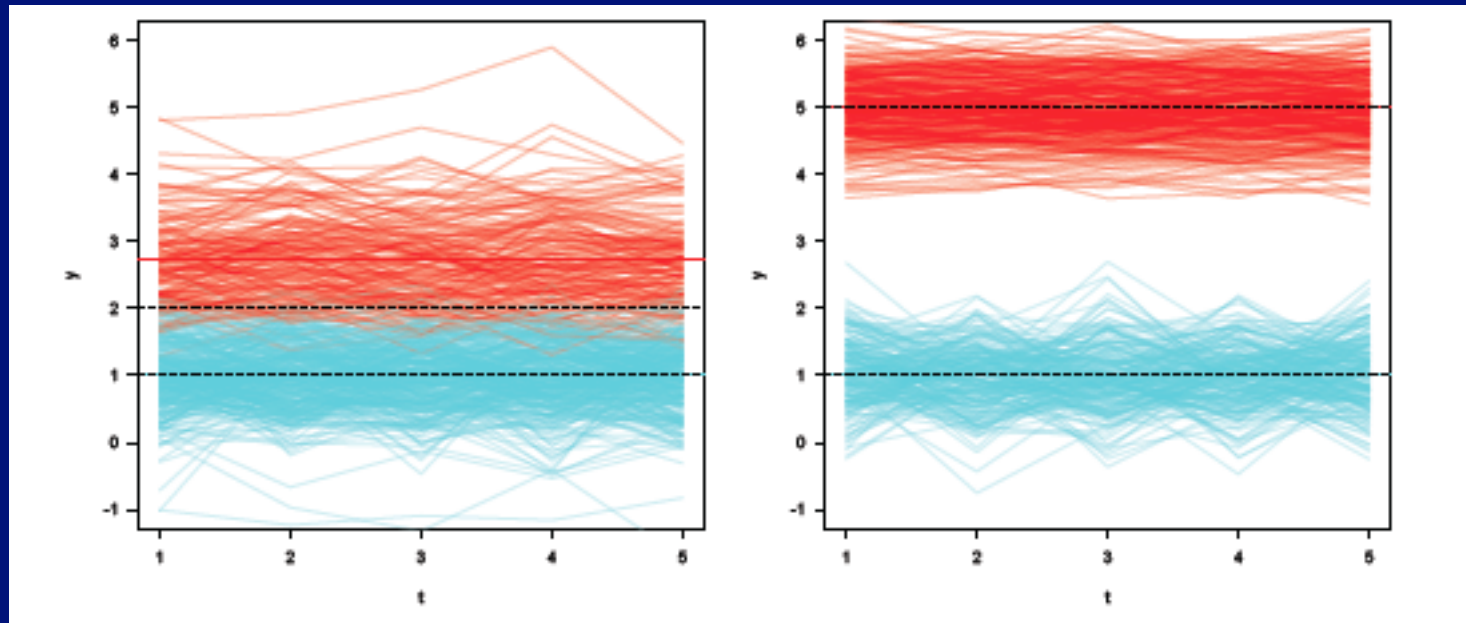
- K is pre-specified; can be chosen using the BIC
- Parameter estimators are not consistent under covariance misspecification (White, 1982; Heggeseeth and Jewell, 2013).
- Robust (sandwich) standard error estimators are available.
- How bad can the bias in regression estimators be? What influences its size?



Mispecified Covariance Structure

Bias and Separation of Trajectories

- Separated components lead to little bias even when you wrongly assume independence.



Black dashed -- true means, Solid lines – estimated means

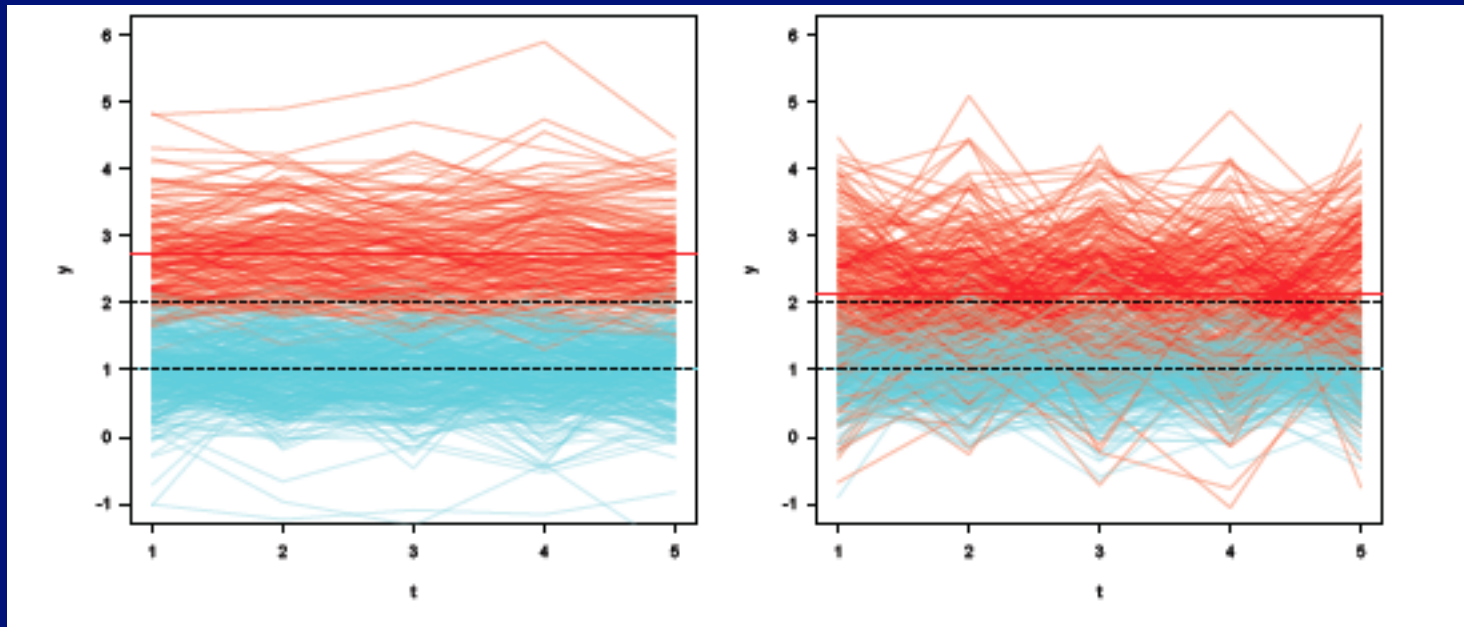
$$\hat{SE}_I(\beta_{01}) = 0.02, \hat{SE}_R(\beta_{01}) = 0.06$$

$$\hat{SE}_I(\beta_{01}) = 0.01, \hat{SE}_R(\beta_{01}) = 0.01^{14}$$

Mispecified Covariance Structure

Bias and Level of Dependence

- Components with little dependence lead to small bias even when you wrongly assume independence.

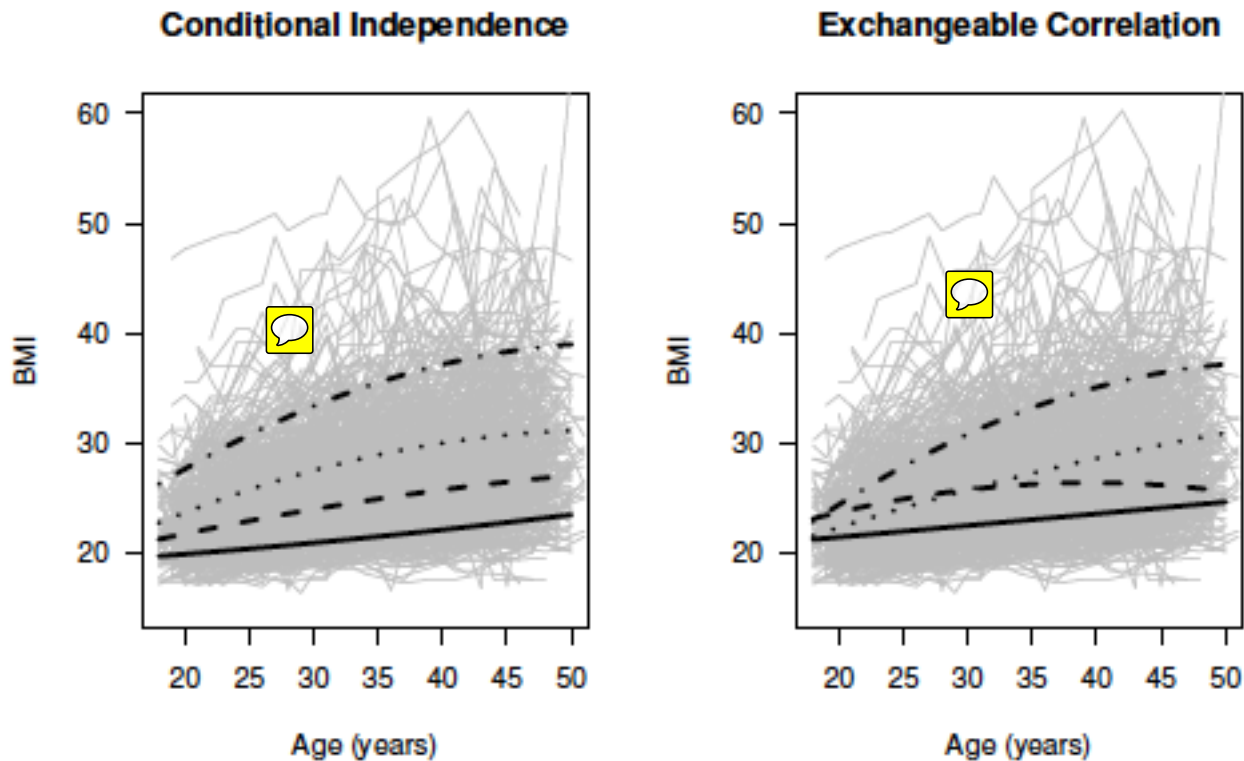


Black dashed -- true means, Solid lines – estimated means

$$\hat{SE}_I(\beta_{01}) = 0.02, \hat{SE}_R(\beta_{01}) = 0.06$$

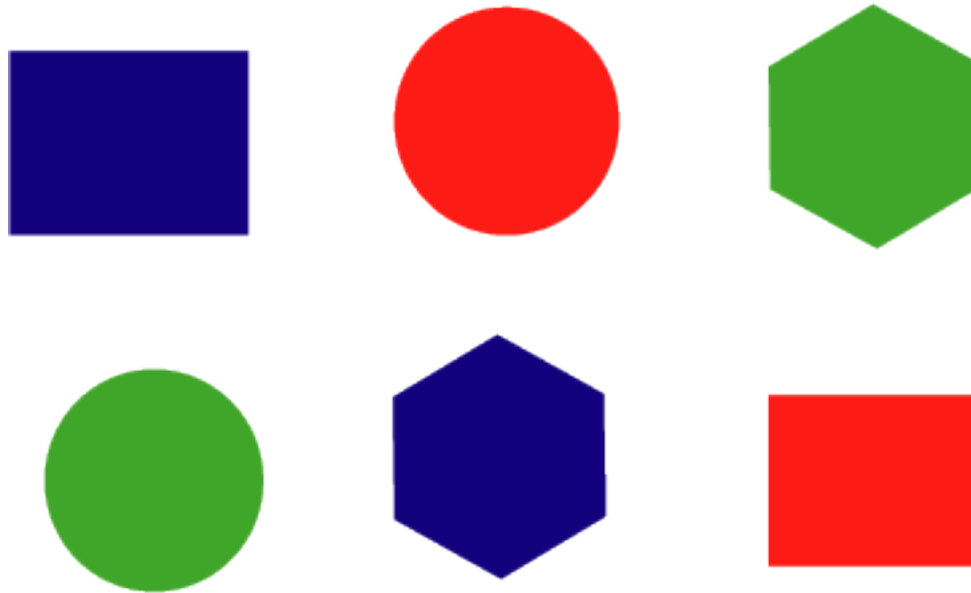
$$\hat{SE}_I(\beta_{01}) = 0.03, \hat{SE}_R(\beta_{01}) = 0.04$$

NLSY Data Analysis



- Covariance makes a difference to the trajectories
- hard to estimate bias from misspecified covariance

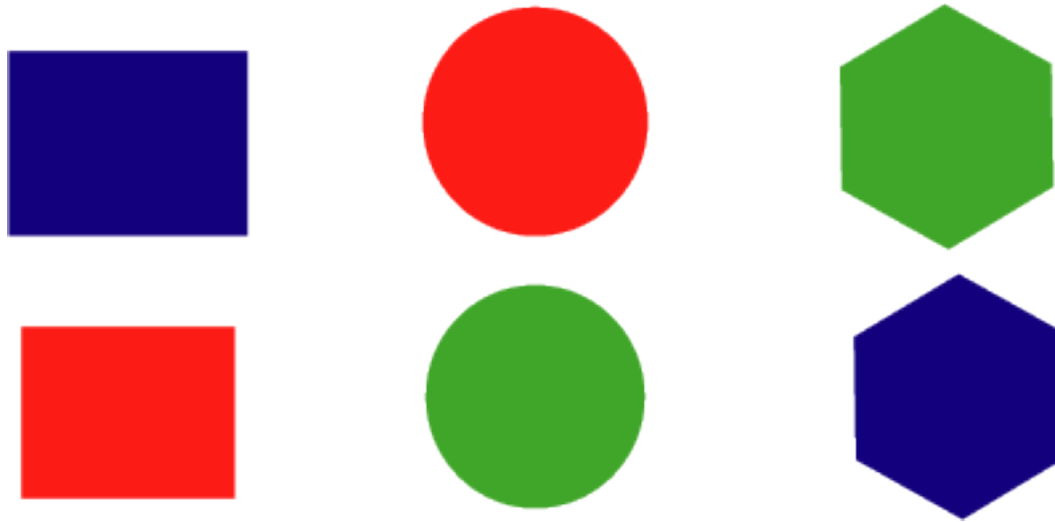
How Do We Group These Blocks?



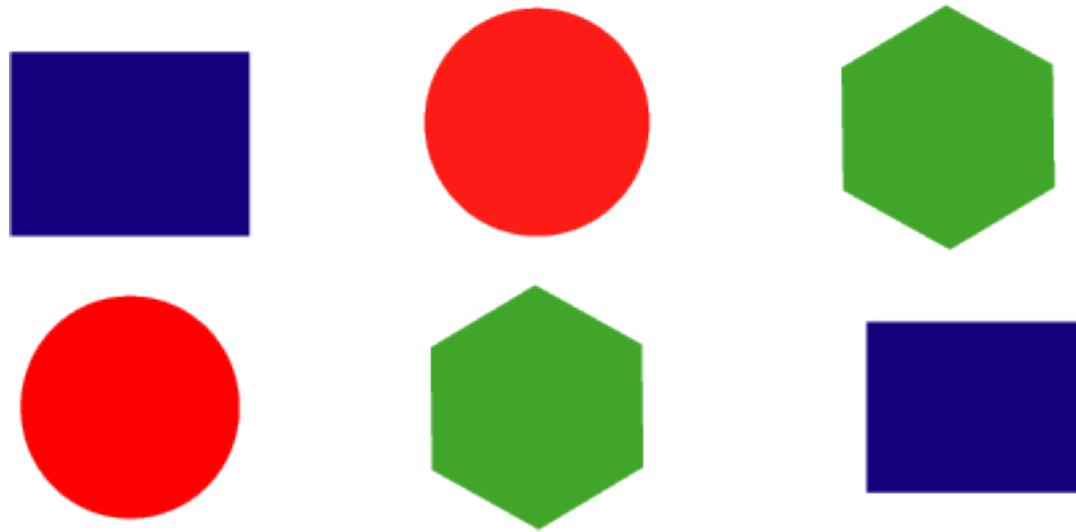
Group by Color



Group by Shape



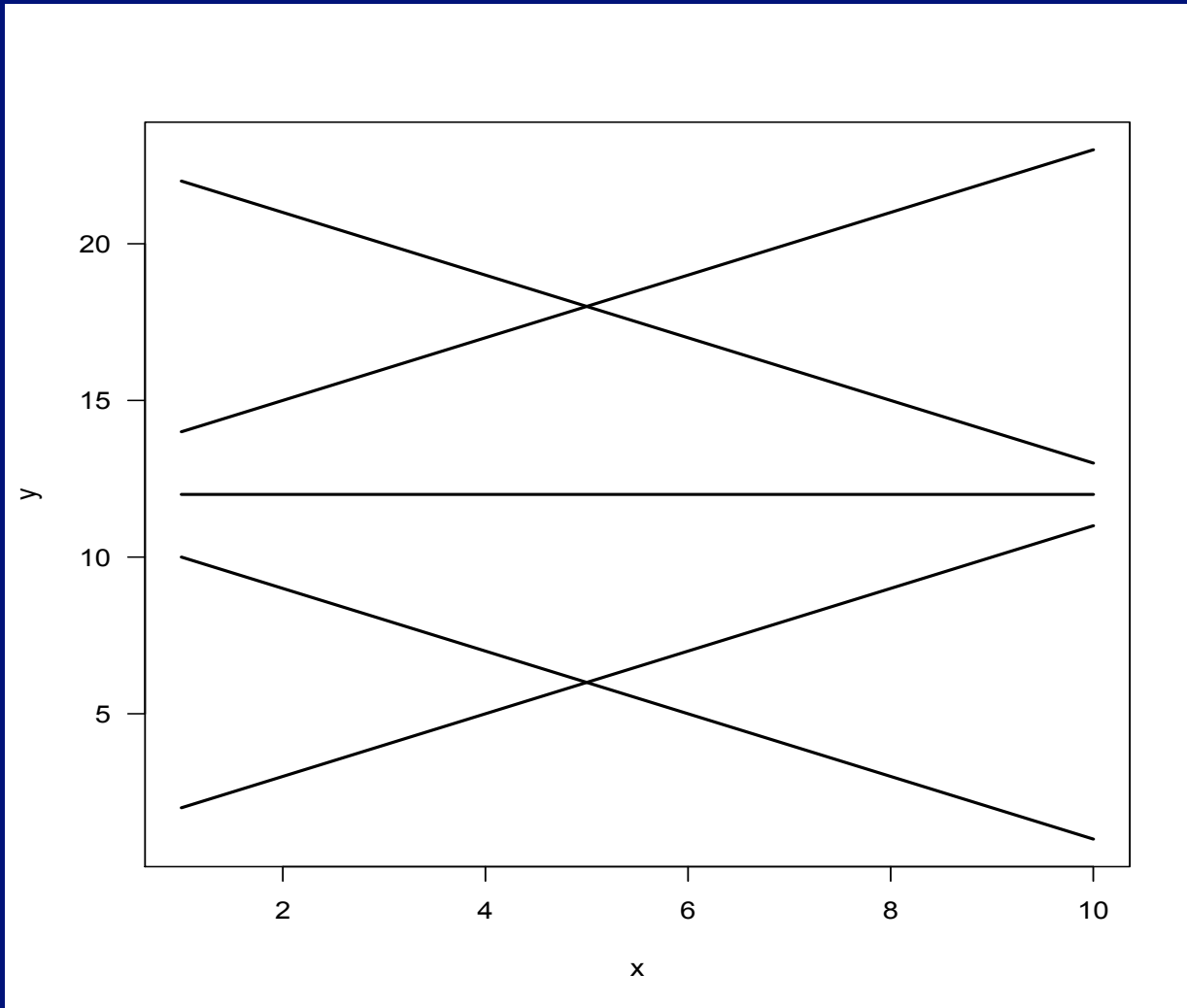
How Do We Group These Blocks?



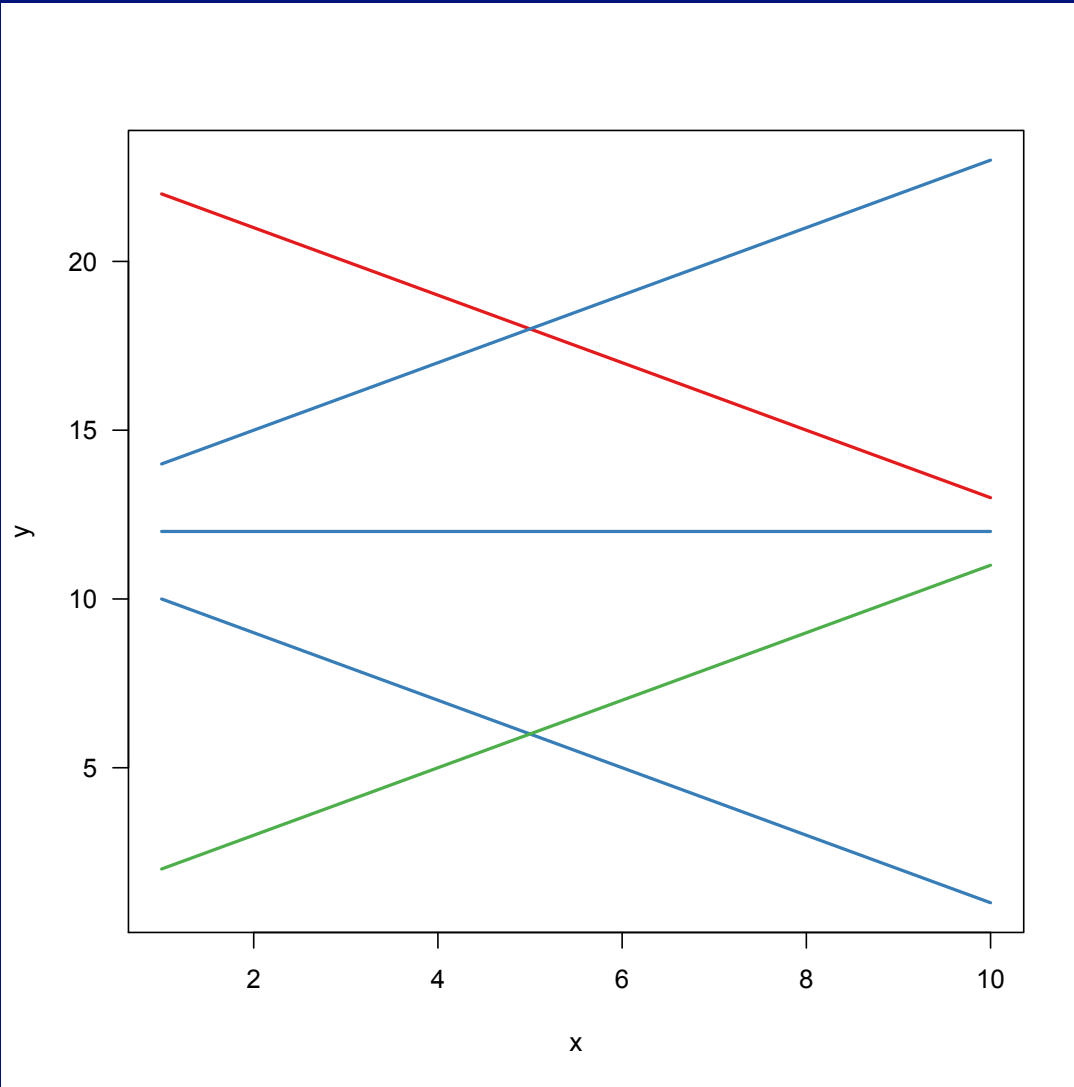
Group by Color or Shape



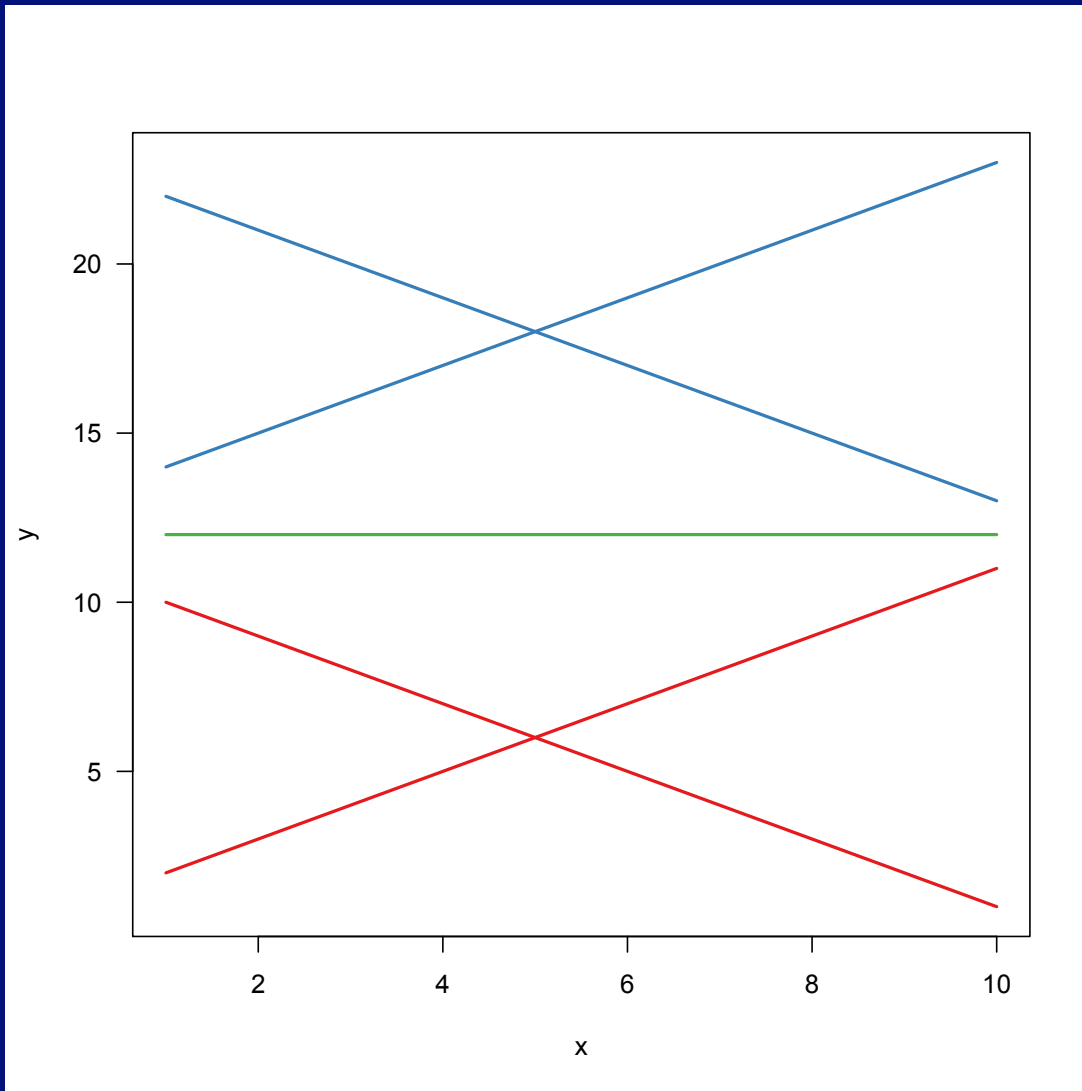
How Do We Group These (Regression) Lines?



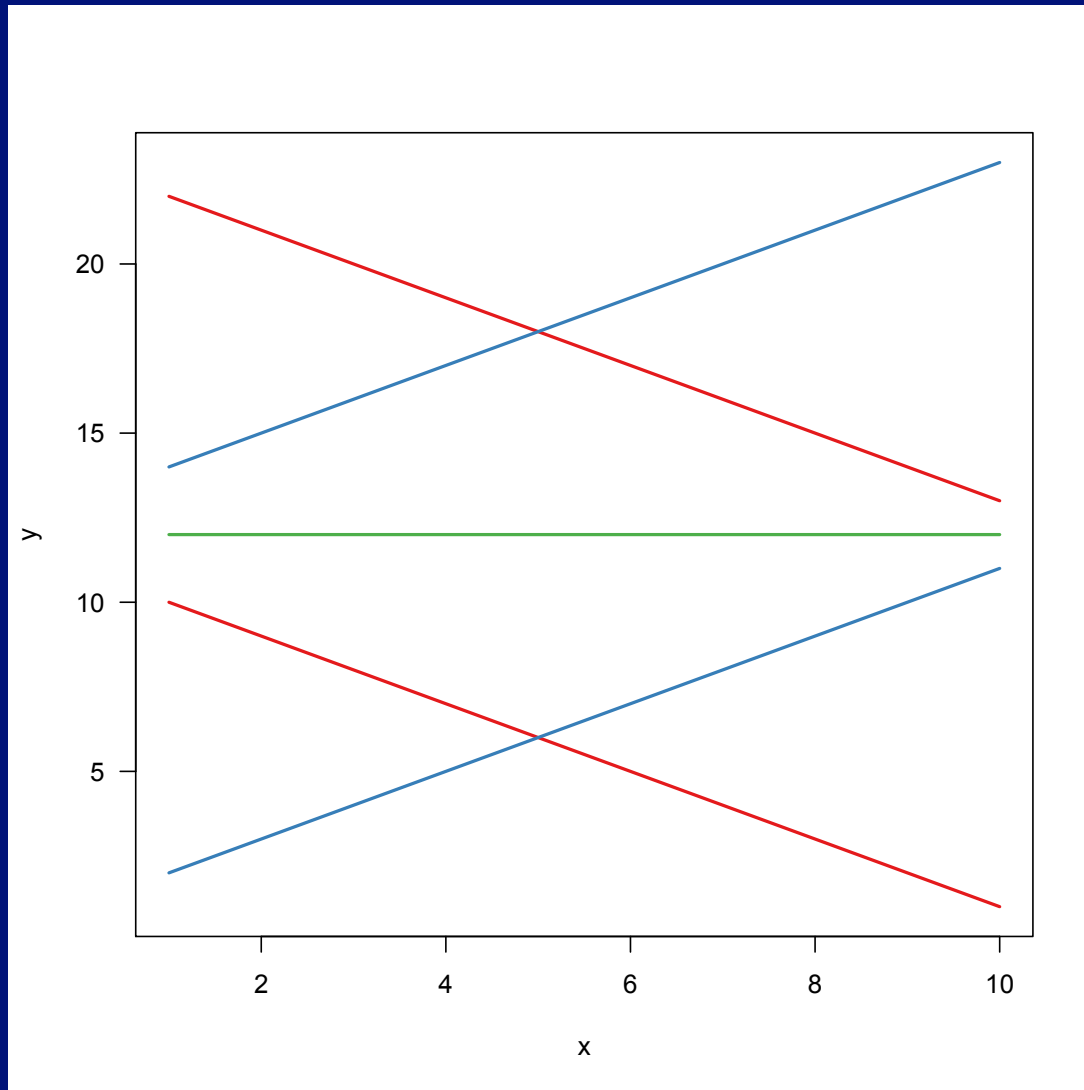
Group by Intercept



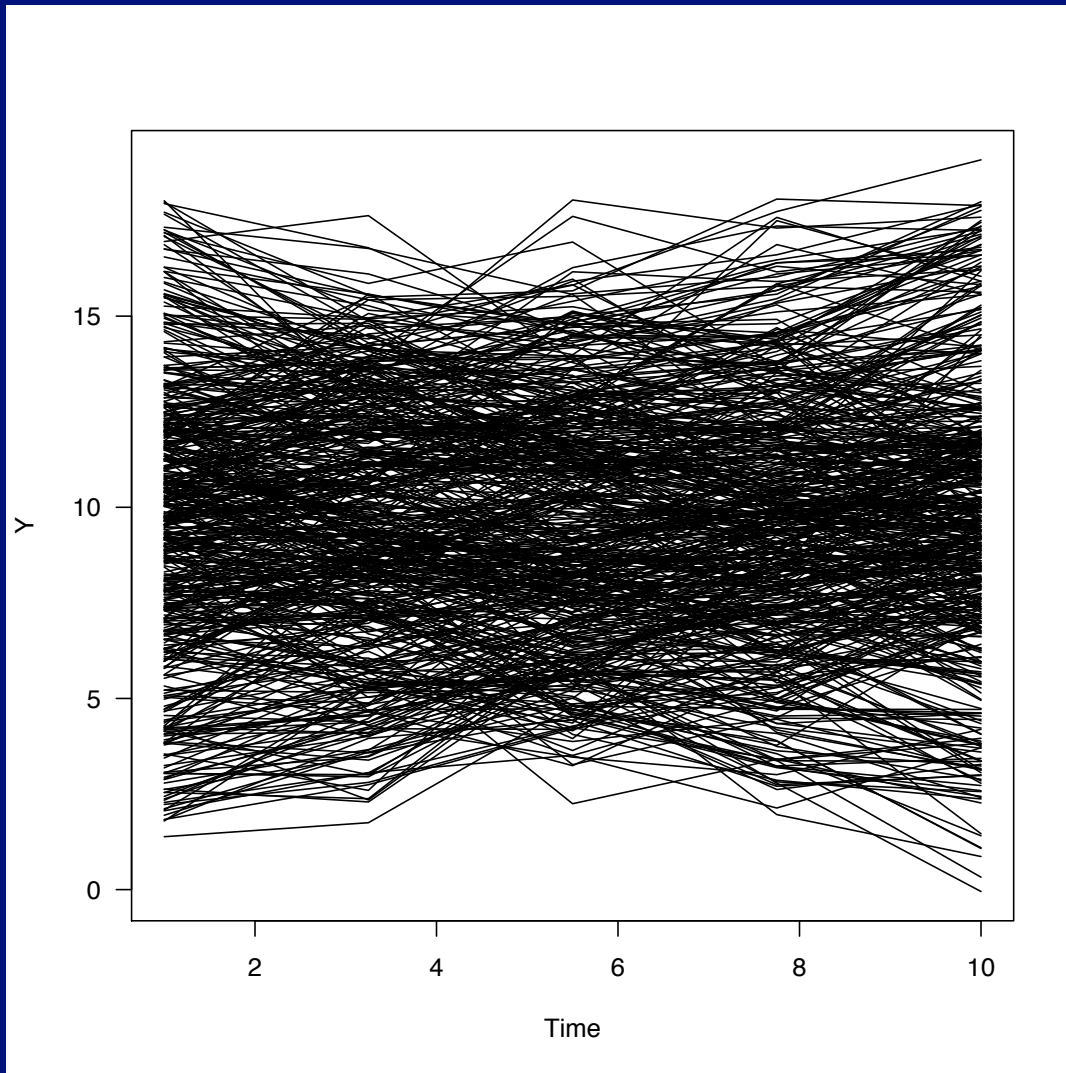
Group by Level



Group by Shape (Slope)

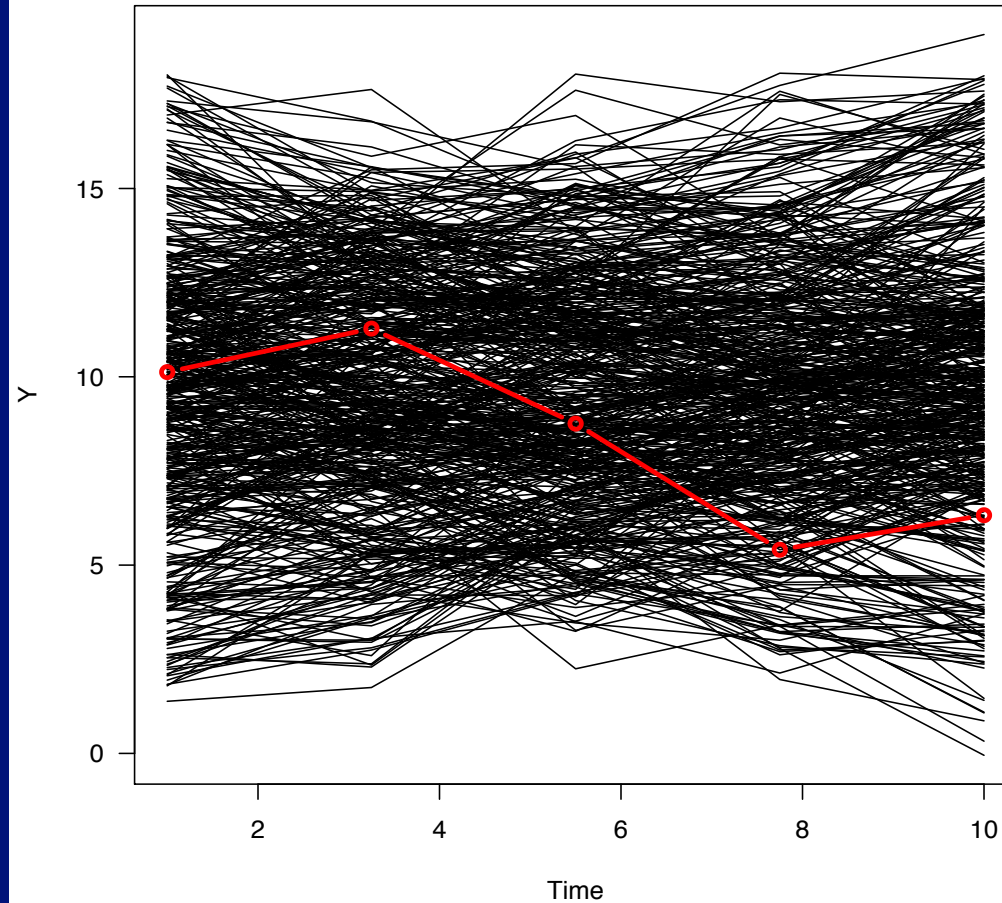


Simulated Data



How could we group these individuals?

Simulated Data

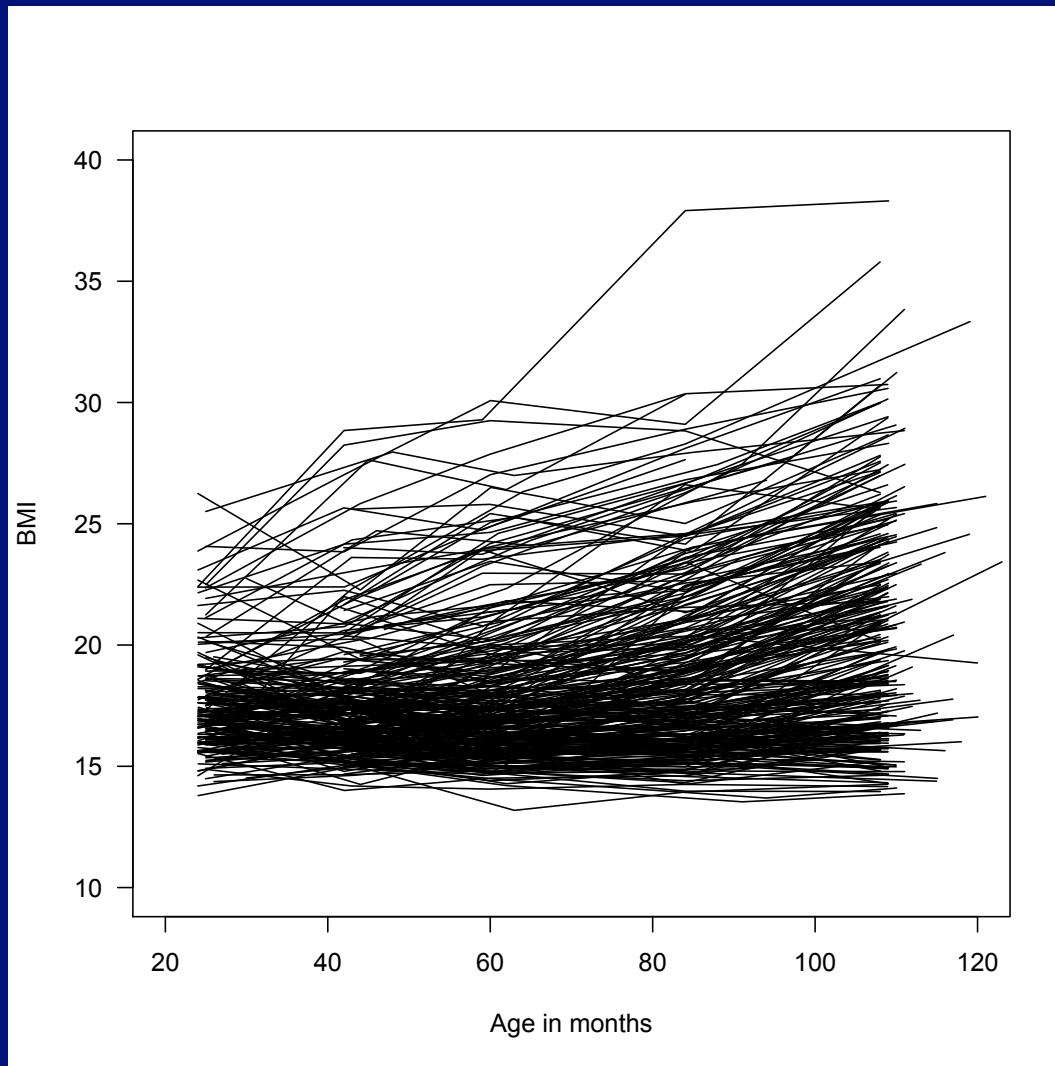


How could we group these individuals?

Real Longitudinal Data

- **Center for the Health Assessment of Assessment of Mothers and Children of Salinas (CHAMACOS) Study**
 - **In 1999-2000, enrolled 601 pregnant women in agricultural Salinas Valley, CA.**
 - **Mostly Hispanic, agricultural workers.**
 - **Determine if exposure to pesticides and other chemicals impact children's growth patterns (BMI, neurological measures etc_.**
- **First, focus on studying/estimating the growth patterns of children.**
- **Second, determine if early life predictors are related to the patterns**
 - **pesticide/chemical exposure in utero**
 - **ODT, PDT, PDE, BPA (bisphenol A)**

CHAMACOS Data



How could we group these individuals?

Cluster Analyses

- Clustering is the task of assigning a set of objects into groups so that the objects in the same group are more similar to each other than to those in other groups.
- What does it mean for objects to be more similar or more dissimilar?
 - Distance matrix
- Why do we cluster objects?

Standard Clustering Methods

- Partition methods
 - Partition objects into K groups so that an objective function of dissimilarities is minimized or maximized.
 - Example: K-means Algorithm
- Model-based methods
 - Assume a model that includes a grouping structure and estimate parameters.
 - Example: Finite Mixture Models

K-means algorithm

- Input: Data for n individuals in vector form. For individual i , the observed data vector is

$$\mathbf{y}_i = (y_{1i}, \dots, y_{im}).$$

- Measure of Dissimilarity: Squared Euclidean distance. The dissimilarity between the 1st and 2nd individuals is

$$d(\mathbf{y}_1 - \mathbf{y}_2) = \|\mathbf{y}_1 - \mathbf{y}_2\|^2 = (y_{11} - y_{12})^2 + \dots + (y_{im} - y_{2m})^2$$

K-means Algorithm

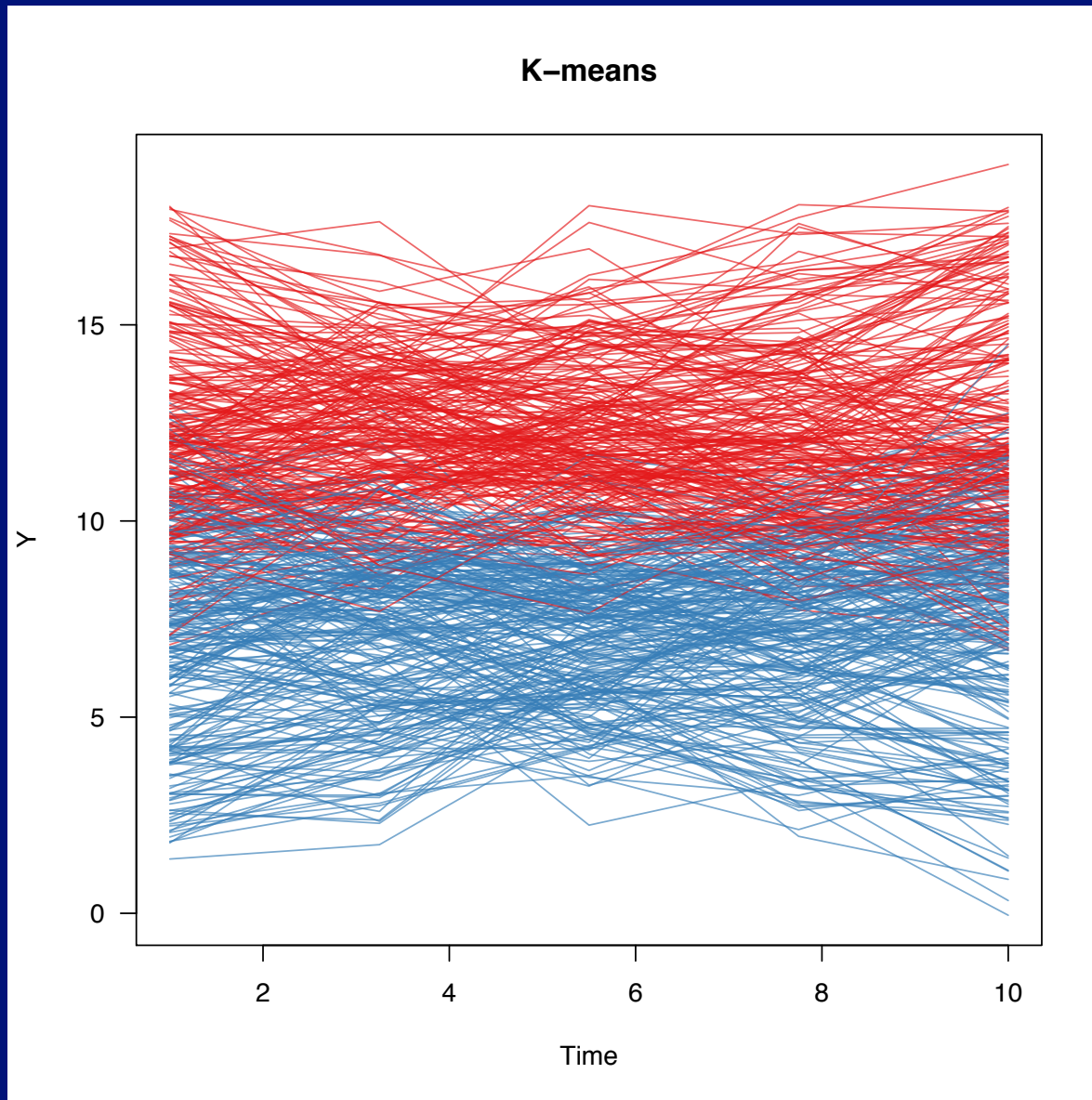
- Goal: Partition individuals into K sets $C = \{C_1, C_2, \dots, C_K\}$ so as to minimize the within-cluster sum of squares

$$\sum_{k=1}^K \sum_{\mathbf{y}_i \in C_k} \|\mathbf{y}_i - \mu_k\|^2$$

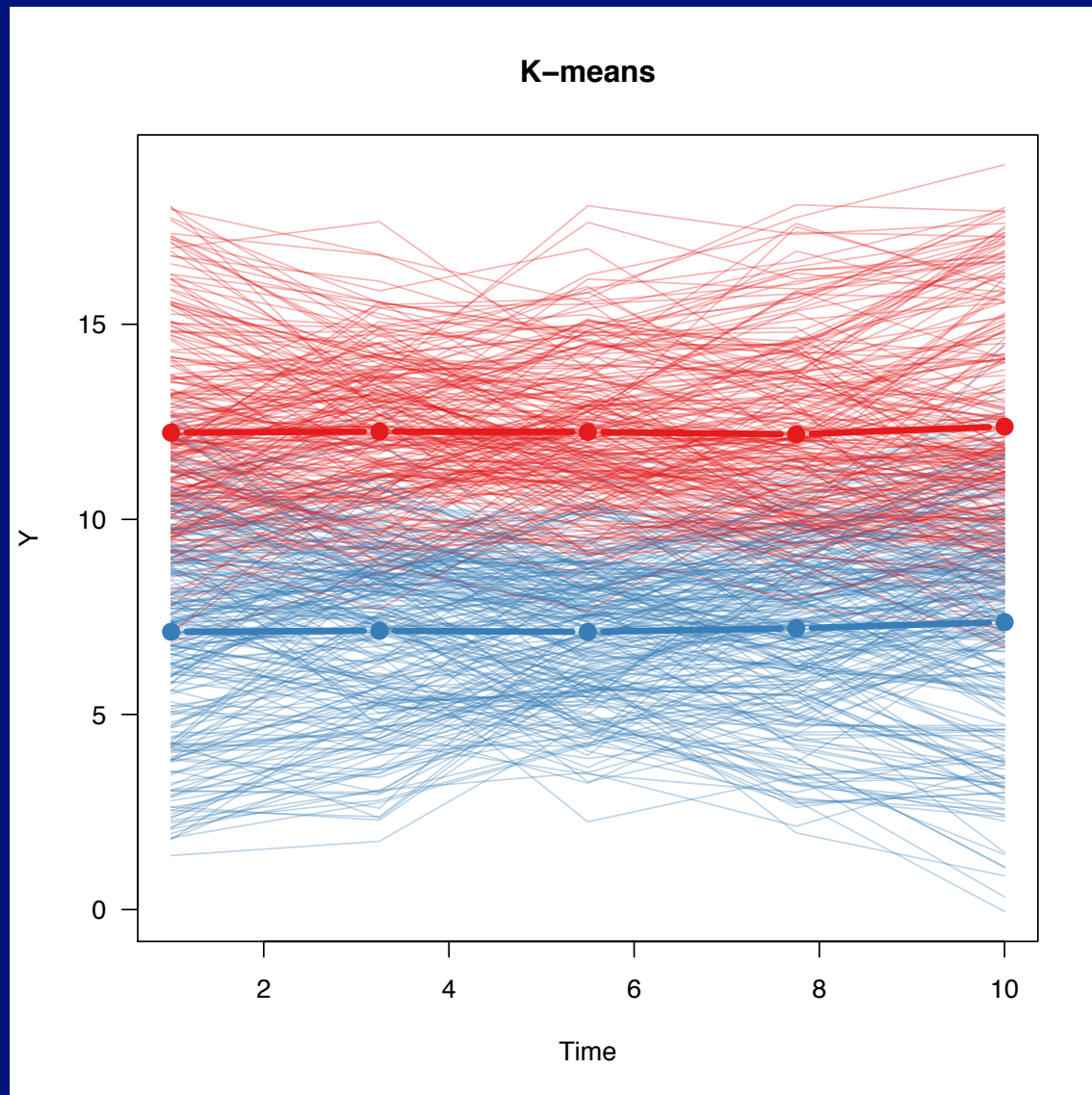
where μ_k is the mean vector of individuals in C_k .

(K must be known before starting K -means. There are many ways to choose K from the data that try to minimize the dissimilarity within each cluster while maximizing the dissimilarity between clusters: for example, the use of *silhouettes*.)

Application to Simulated Data

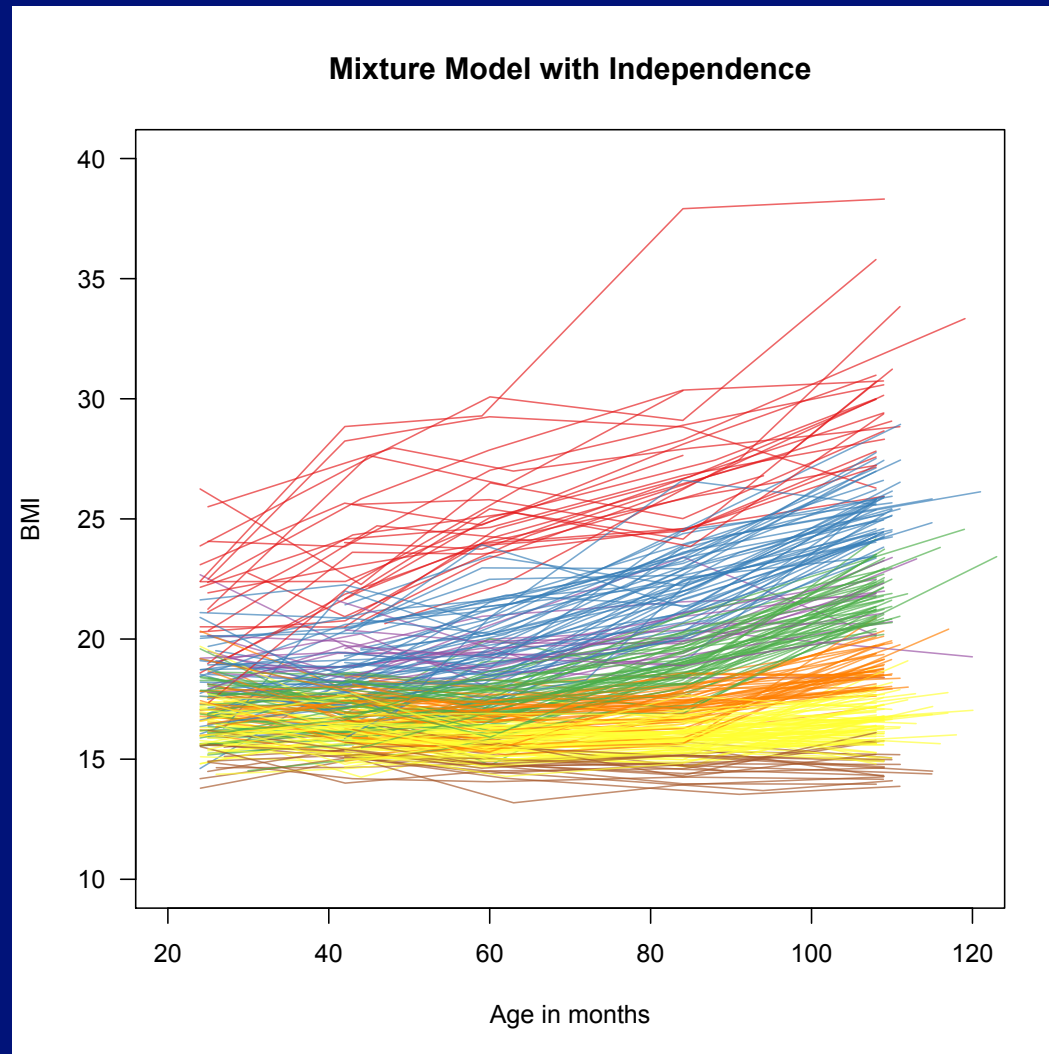


Application to Simulated Data

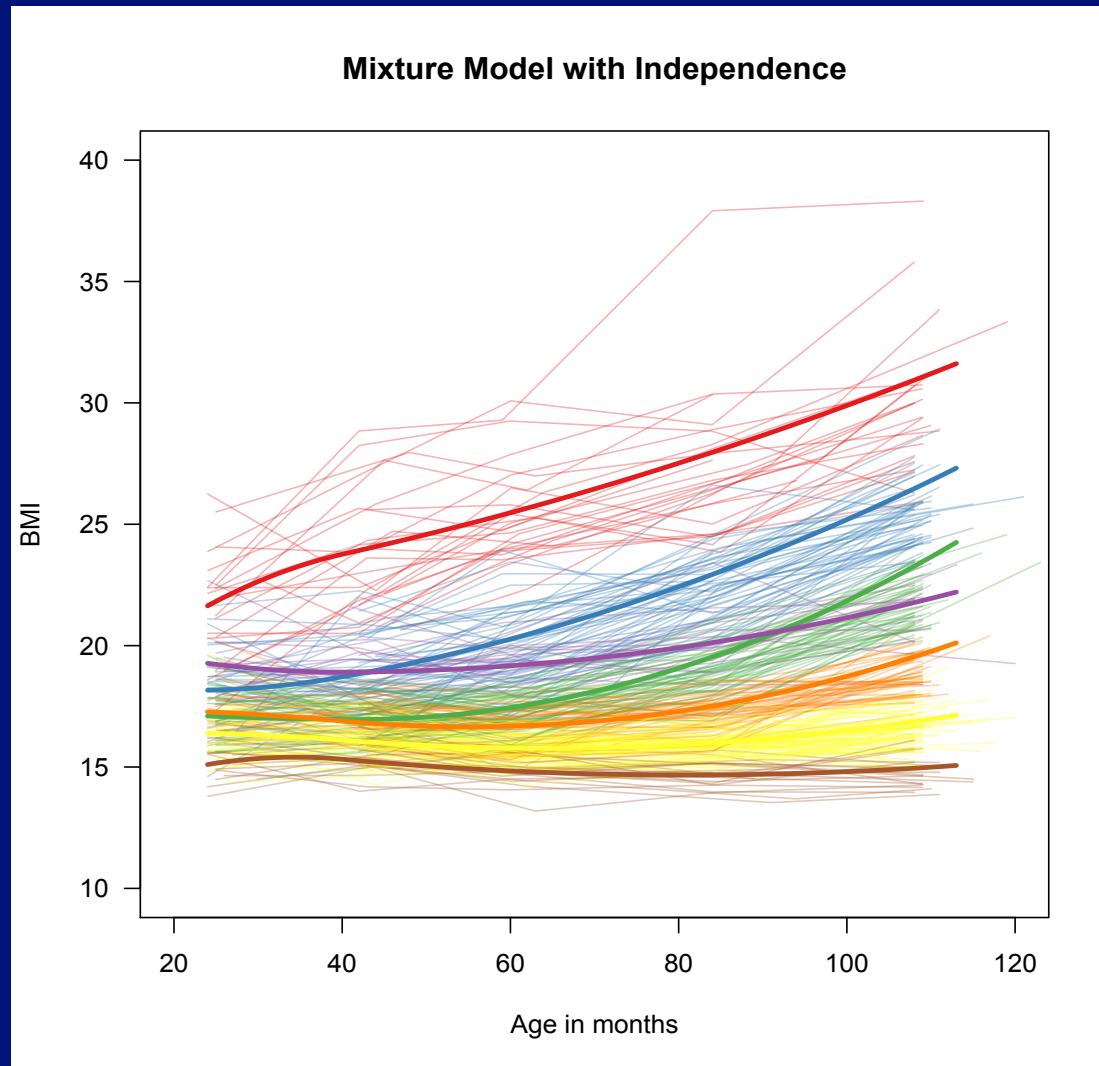


How would you describe—interpret—the group trajectories?

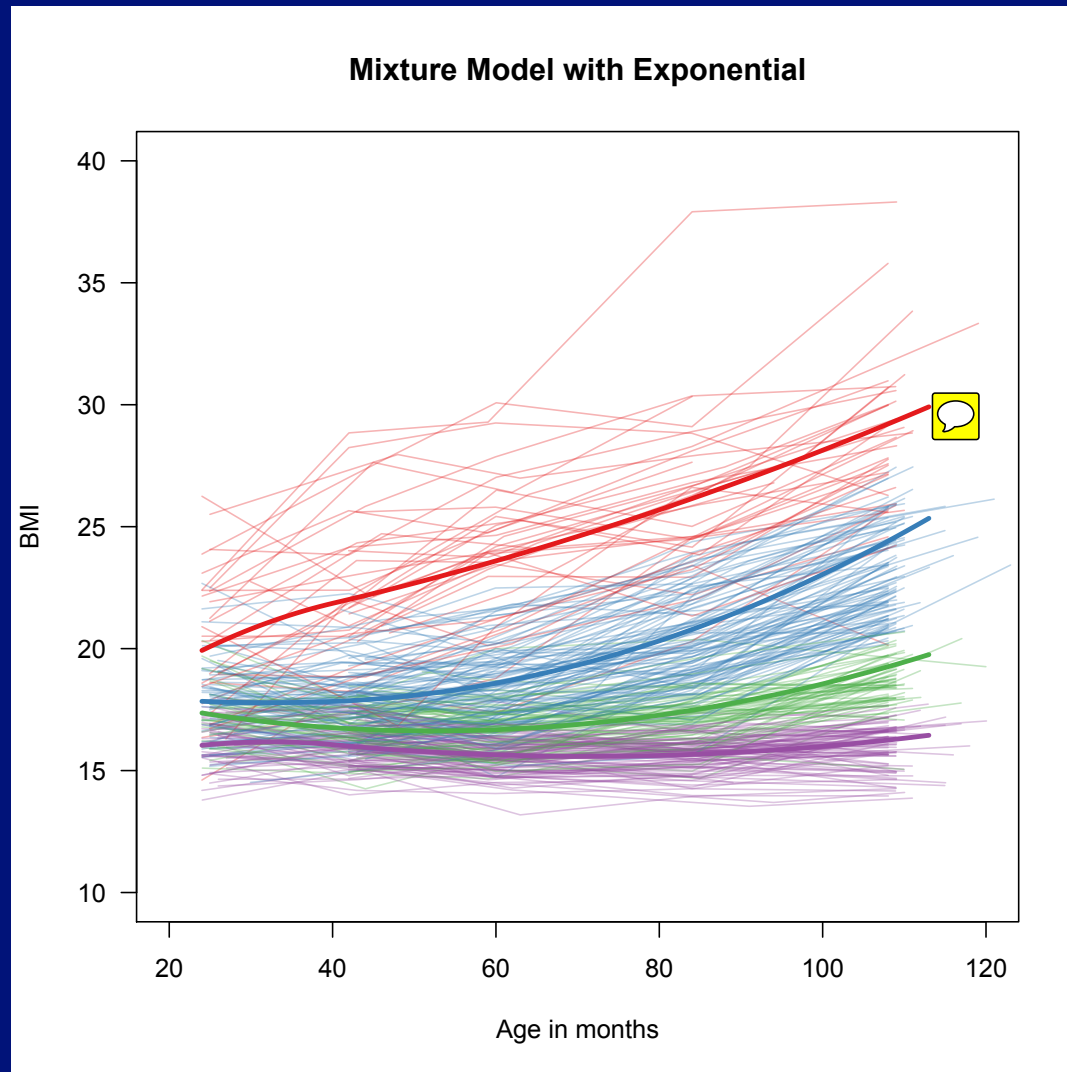
Finite Mixture Model Applied to CHAMACOS Data



Finite Mixture Model Applied to CHAMACOS Data



Finite Mixture Model Applied to CHAMACOS Data



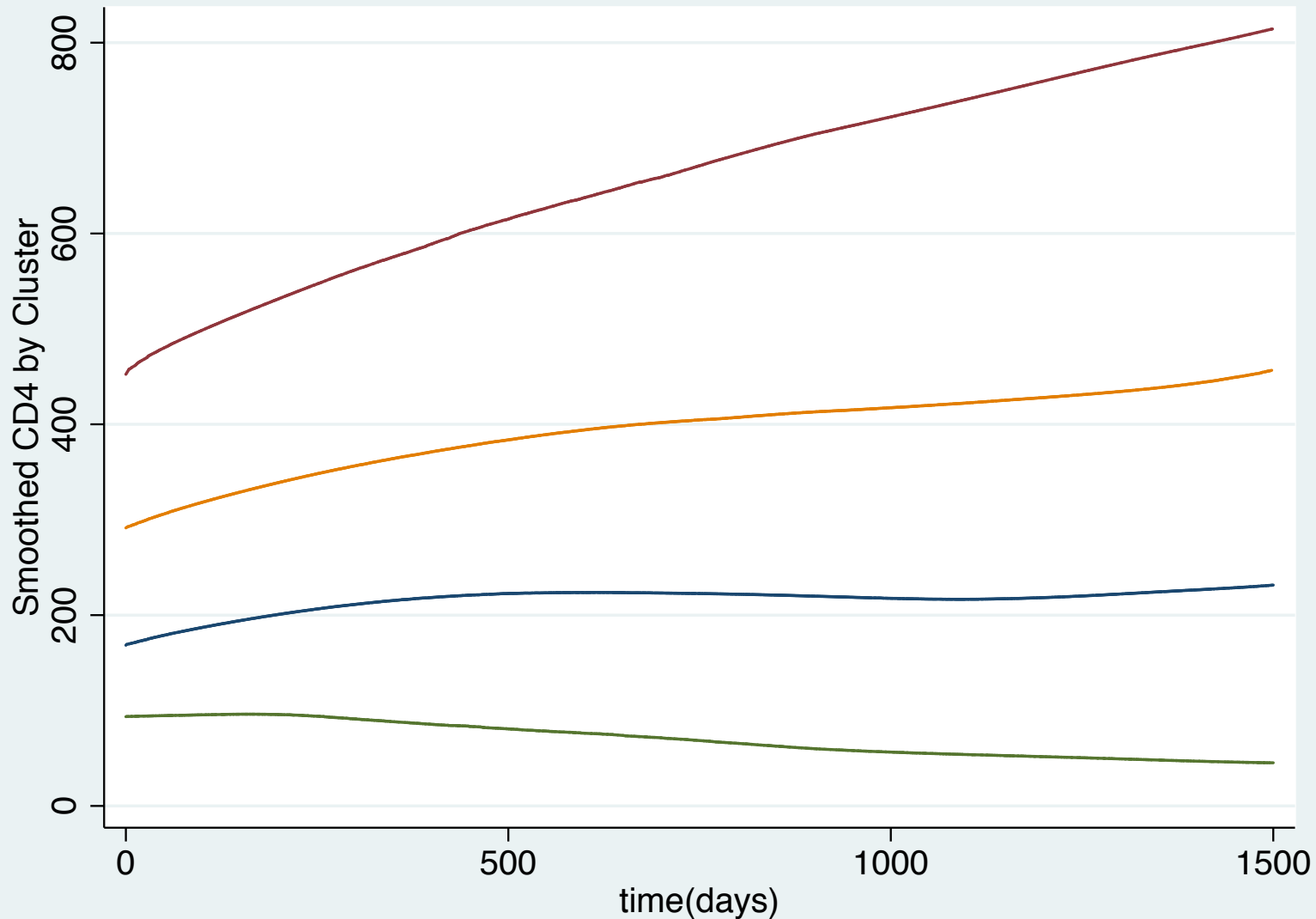
Revisiting Chapter 2: Clustering subjects based on similarity of CD4 versus time

- Treat the CD4 observations at different times as simply different variables one measures on a subject.
- Technique requires all subjects to have the measurements at same times, so we need to interpolate at a fixed number of times for each subject so that everyone has identical times.

Clustering subjects based on similarity of CD4 versus time, cont

- Then, after putting the data in wide format, one simply clusters subjects based on these interpolated times.
- Then, after going back to long format, plot smooths by cluster assignment (see chapter 2 dofile for details).

Clustering subjects based on similarity of CD4 versus time



Clustering by Shape

- Interested in *shape* not just level (which appears to dominate clustering techniques)
- Want a method that:
 - Works with irregularly sampled data
 - Includes a way to estimate the relationship between baseline risk factors and group membership
 - Groups individuals according to the outcome pattern over time ignoring the level

Clustering by Shape Options

- Estimate slopes between neighboring observations and cluster on the “derived” observations
- Fit splines for each individual, differentiate, and cluster on coefficients of resulting derivative
- Use partition based cluster methods (like PAM) but use (i) the Pearson coefficient as a distance or dissimilarity measure

$$d_{corr}(\mathbf{x}, \mathbf{y}) = 1 - Corr(\mathbf{x}, \mathbf{y})$$

or the cosine-angle measure of dissimilarity

$$d_{cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{j=1}^m x_j y_j}{(\sum_{j=1}^m x_j^2)(\sum_{j=1}^m y_j^2)}$$

- Vertical shifting individual trajectories

Vertical Shifting

- For each individual, calculate

$$y_i^* = y_i - m_i^{-1} \sum_{j=1}^{m_i} y_{ij}$$

- Each individual now has mean zero and so level is removed from any resulting clustering
- Apply clustering technique to shifted data, e.g. finite mixture model

Correlation Models for Vertical Shifted Data

- Without specifying group, suppose

$$\mathbf{y}_i^* = \lambda_i \mathbf{I}_{m_i} + \mu_i + \epsilon_i, \lambda \sim F_\lambda, \epsilon \sim N(), \Sigma)$$

where \mathbf{I}_{m_i} is an m_i length vector of 1s, and

$\mu_{ij} = \mu_k(t_{ij})$ is the j^{th} element of the vector of mean values for the k^{th} group evaluated at the observation times t_i . Thus,

$$\mathbf{y}_i^* = \mathbf{A}_i \mathbf{y}_i = \mu_i - \bar{\mu}_i + \epsilon_i - \bar{\epsilon}_i$$

Correlation Models for Vertical Shifted Data

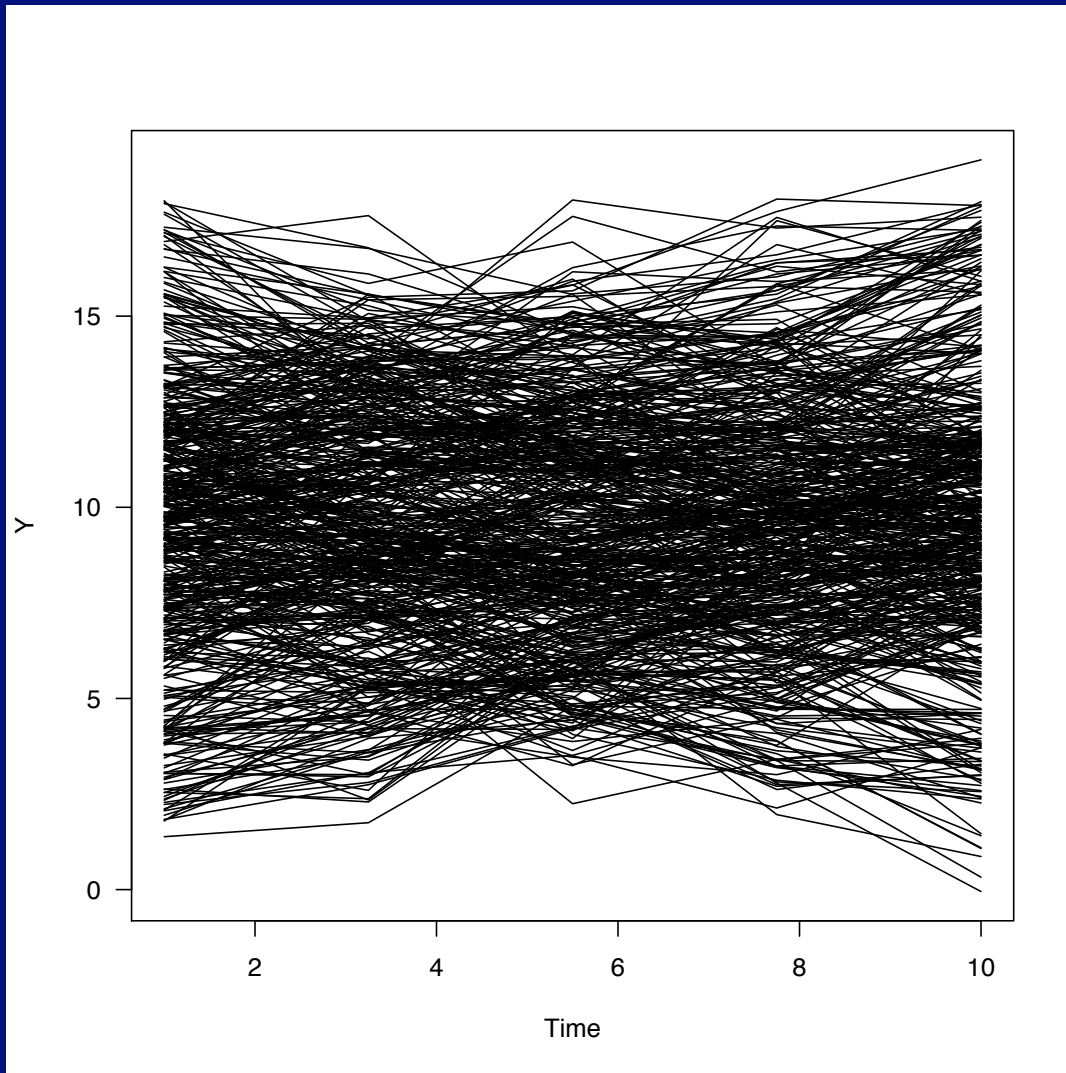


$$Cov(\mathbf{Y}_i^* - \mu_i) = Cov((A - \mathbf{I}_{m_i})\mu_i + A\epsilon)$$

Two components of the covariance

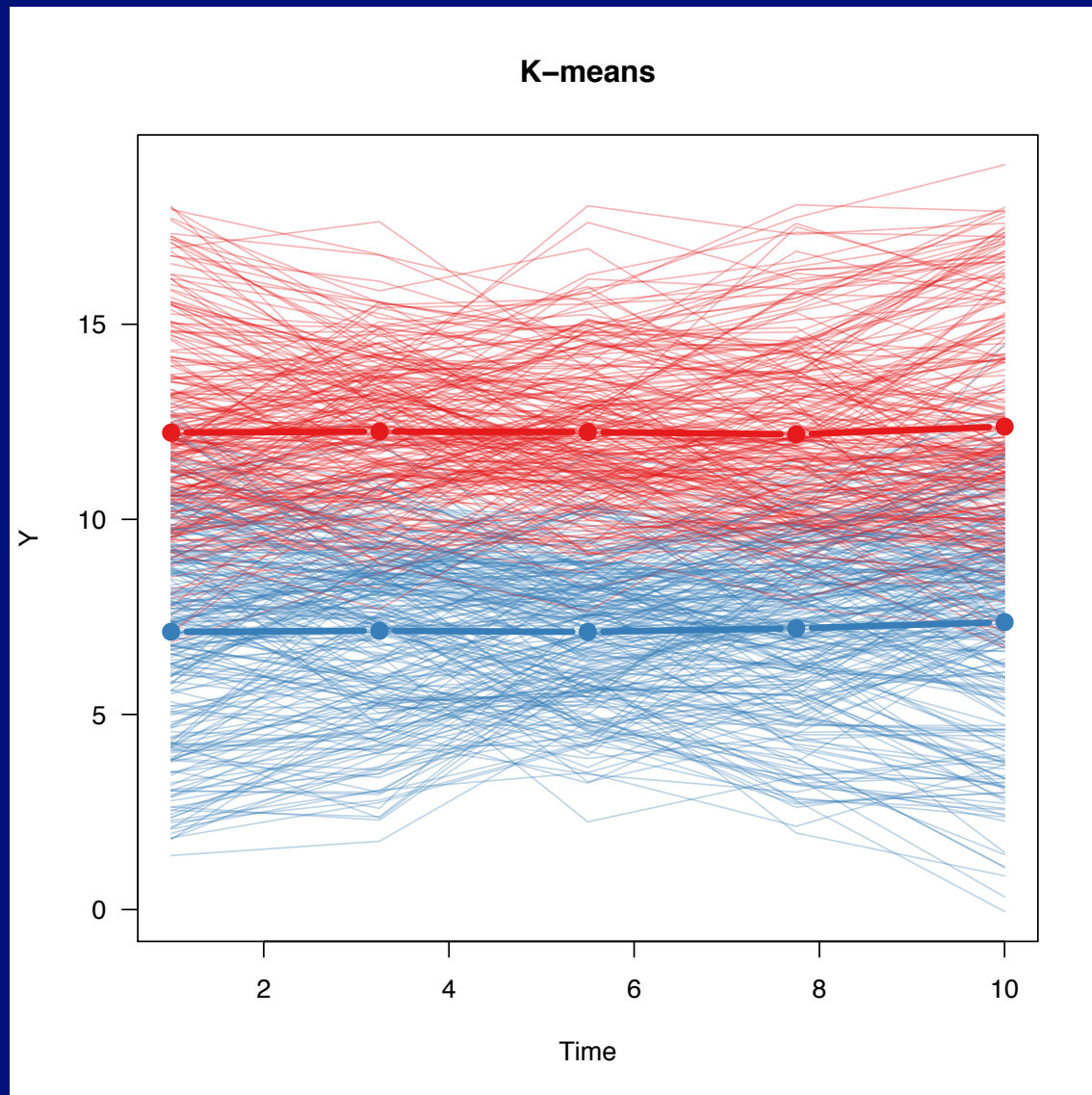
- One induced by the averaging process
- One induced by (random) observation times

Simulated Data



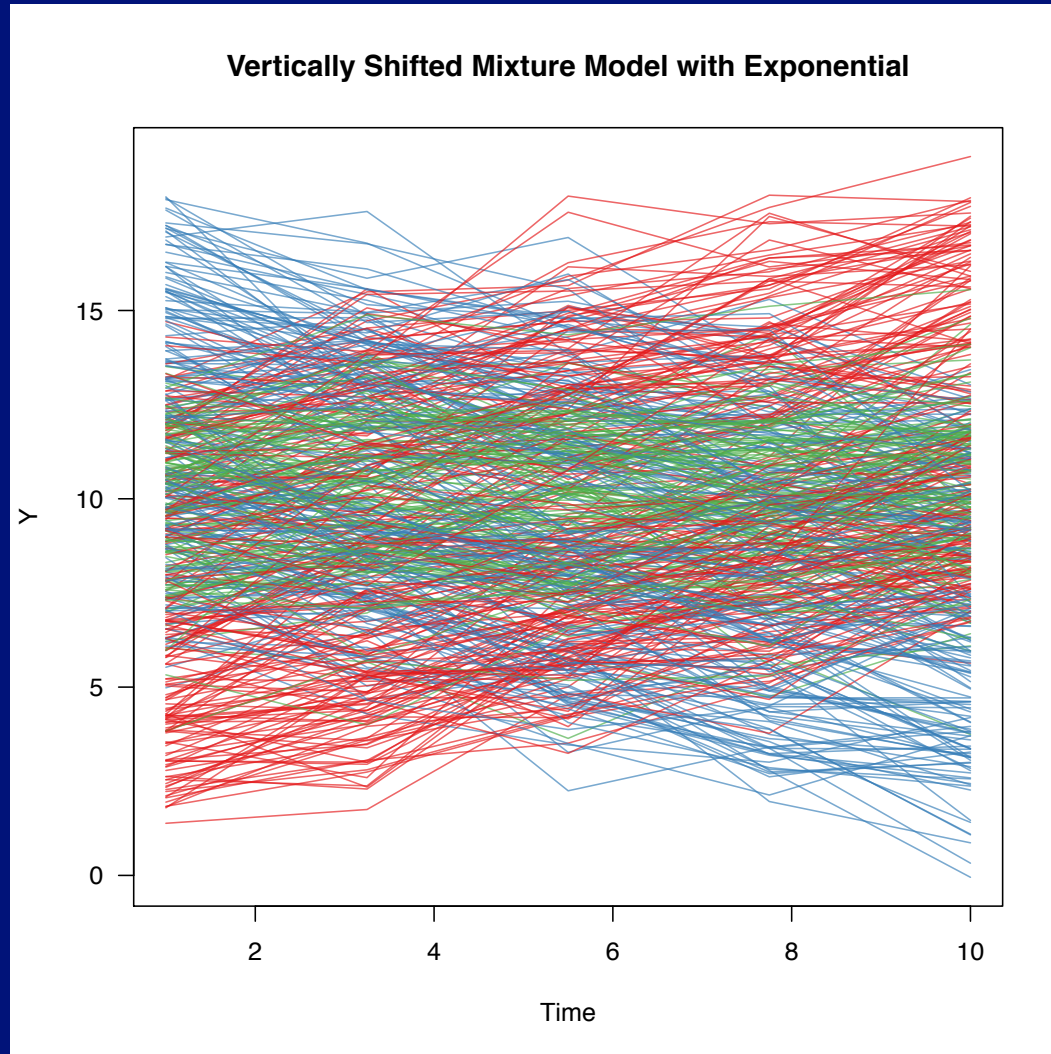
How could we group these individuals?

Application to Simulated Data

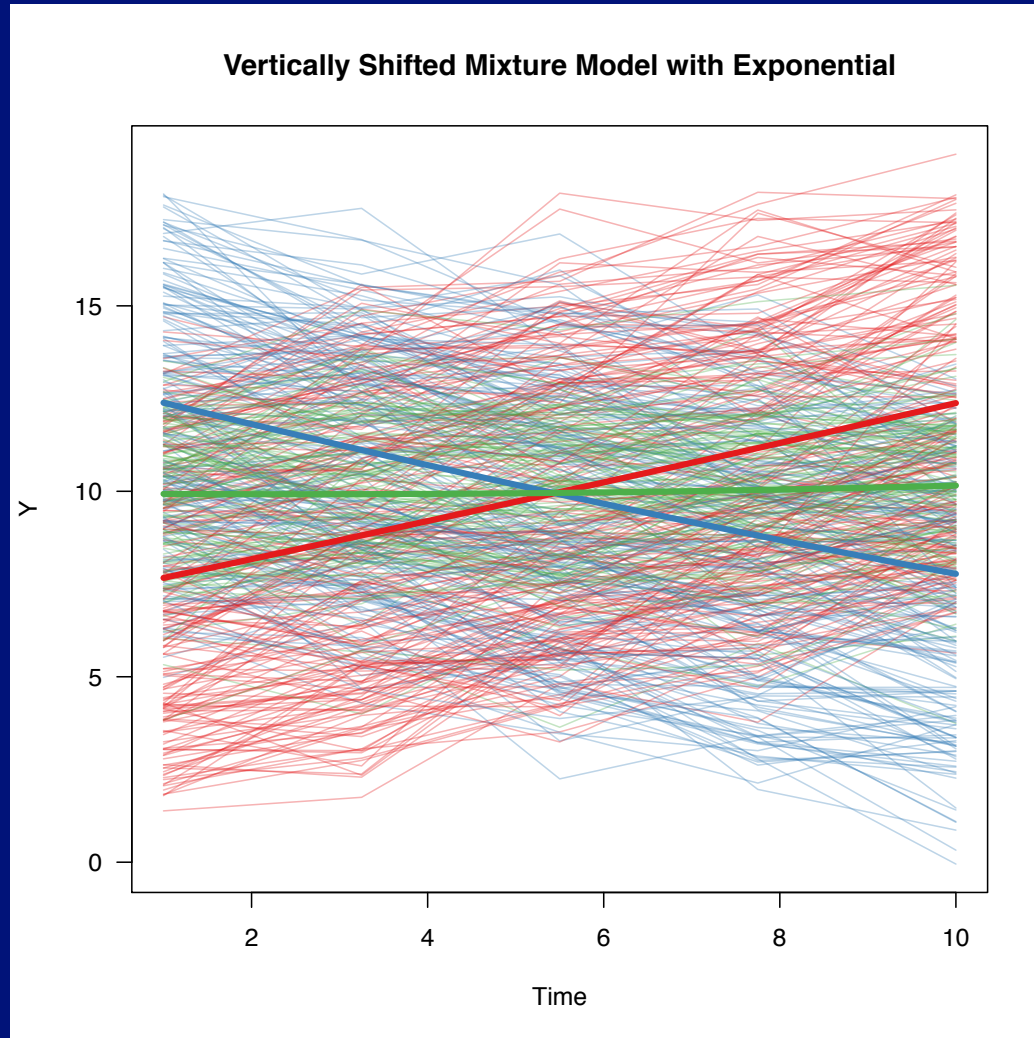


How would you describe—interpret—the group trajectories?

Vertical Shifting Applied to Simulated Data



Vertical Shifting Applied to Simulated Data



Vertical Shifting with CHAMACOS

- Two-part models
 - First, use standard regression models to relate baseline predictors to BMI
 - Then, use vertically shifted shape clustering with (same or different baseline predictors for shape groups)
- For BMI in the CHAMACOS data
 - Works with irregularly sampled data
 - Includes a way to estimate the relationship between baseline risk factors and group membership
 - Groups individuals according to the outcome pattern over time ignoring the level

Regression for BMI Levels

Estimated change in mean BMI from a ten-fold increase in prenatal maternal DDT serum concentration (ng/g of lipid) based on a linear model at ages 2, 3.5, 5, 7, and 9 years, unadjusted and adjusted for baseline risk factors^a.

| | Boy | | | | | Girl | | | | |
|------------------|-------|-------------------|-------------------|-------|-------------------|-------|---------|-------|-------|--------------------|
| | 2 yr. | 3.5 yr. | 5 yr. | 7 yr. | 9 yr. | 2 yr. | 3.5 yr. | 5 yr. | 7 yr. | 9 yr. |
| <i>o,p'</i> -DDT | | | | | | | | | | |
| Unadjusted | 0.35 | 0.77 ⁺ | 0.93 ⁺ | 1.04 | 1.60 [*] | 0 | -0.08 | -0.43 | -0.4 | -0.71 |
| Adjusted | 0.24 | 0.57 | 0.69 | 0.59 | 1.23 ⁺ | -0.04 | -0.22 | -0.64 | -0.72 | -1.10 ⁺ |
| <i>p,p'</i> -DDT | | | | | | | | | | |
| Unadjusted | 0.21 | 0.49 | 0.59 | 0.55 | 1.04 ⁺ | -0.1 | -0.3 | -0.55 | -0.46 | -0.88 |
| Adjusted | 0.14 | 0.35 | 0.41 | 0.20 | 0.70 | -0.08 | -0.33 | -0.65 | -0.56 | -0.94 |
| <i>p,p'</i> -DDE | | | | | | | | | | |
| Unadjusted | 0.14 | 0.46 | 0.5 | 0.3 | 0.82 | -0.14 | -0.41 | -0.72 | -0.58 | -1.3 |
| Adjusted | 0.14 | 0.35 | 0.41 | 0.20 | 0.70 | -0.08 | -0.33 | -0.65 | -0.56 | -0.94 |

Abbreviations: DDE, dichlorodiphenyldichloroethylene; DDT, dichlorodiphenyltrichloroethane.

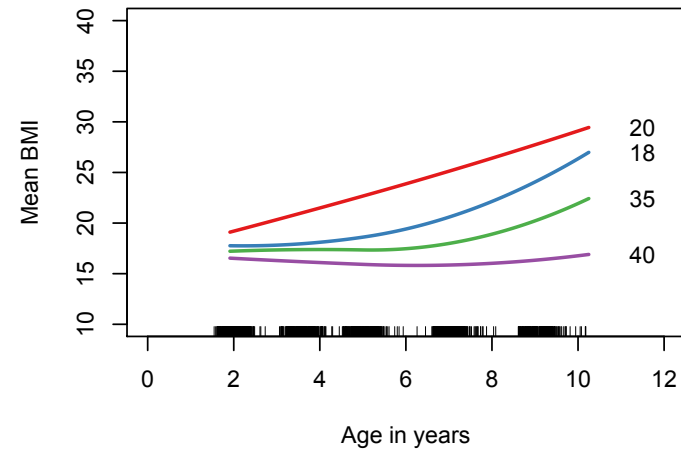
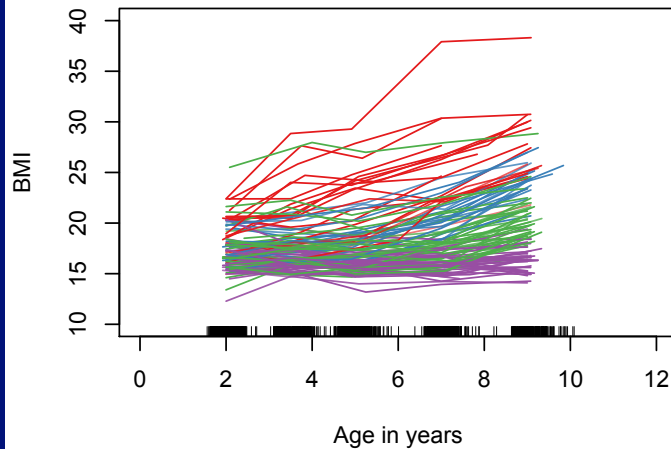
⁺ $P < 0.1$; ^{*} $P < 0.05$

^a Adjusted for maternal pre-pregnancy BMI, number of years in the USA, duration of breastfeeding and birth weight.

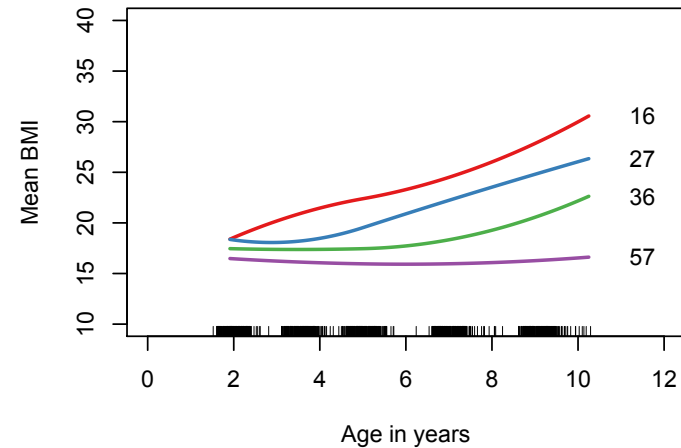
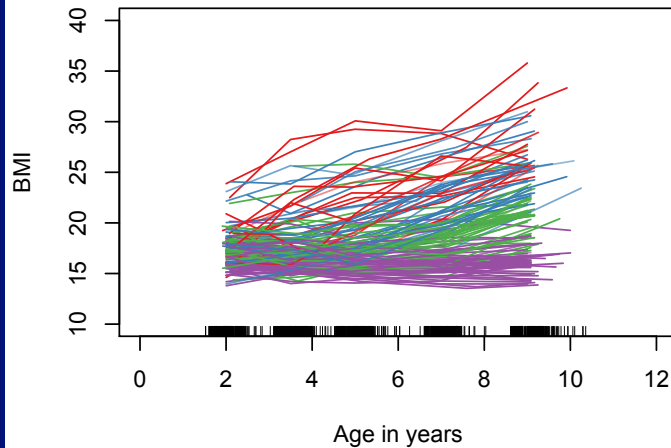
- Can also do regression with average level (over all ages)

Vertical Shifting with CHAMACOS

ODT : Boy



ODT : Girl



Vertical Shifting with CHAMACOS



| | Estimate | 95% CI | $P > z $ |
|--|----------|------------------|-----------|
| Red v. Purple : bmicat_preOverweight | 2.494 | (0.506, 12.294) | 0.261 |
| Red v. Purple : bmicat_preObese | 34.068 | (4.945, 234.699) | 0 |
| Red v. Purple : lng_lod2_ODT_pg | 1.743 | (0.343, 8.857) | 0.503 |
| Red v. Purple : yrsusa | 0.97 | (0.887, 1.062) | 0.514 |
| Blue v. Purple : bmicat_preOverweight | 2.4 | (0.408, 14.129) | 0.333 |
| Blue v. Purple : bmicat_preObese | 8.636 | (0.97, 76.902) | 0.053 |
| Blue v. Purple : lng_lod2_ODT_pg | 6.679 | (1.067, 41.812) | 0.042 |
| Blue v. Purple : yrsusa | 1.048 | (0.955, 1.15) | 0.324 |
| Green v. Purple : bmicat_preOverweight | 1.45 | (0.428, 4.917) | 0.551 |
| Green v. Purple : bmicat_preObese | 15.561 | (2.306, 105.031) | 0.005 |
| Green v. Purple : lng_lod2_ODT_pg | 4.028 | (0.598, 27.113) | 0.152 |
| Green v. Purple : yrsusa | 0.953 | (0.866, 1.048) | 0.322 |

Table 1: Estimated odds ratios for Vertically Shifted Model with ODT for Boy

| | Estimate | 95% CI | $P > z $ |
|--|----------|------------------|-----------|
| Red v. Purple : bmicat_preOverweight | 6.258 | (0.193, 203.281) | 0.302 |
| Red v. Purple : bmicat_preObese | 64.22 | (2.1, 1963.475) | 0.017 |
| Red v. Purple : lng_lod2_ODT_pg | 0.358 | (0.07, 1.846) | 0.22 |
| Red v. Purple : yrsusa | 0.955 | (0.785, 1.162) | 0.646 |
| Blue v. Purple : bmicat_preOverweight | 1.785 | (0.464, 6.86) | 0.399 |
| Blue v. Purple : bmicat_preObese | 7.121 | (1.069, 47.437) | 0.042 |
| Blue v. Purple : lng_lod2_ODT_pg | 0.846 | (0.276, 2.599) | 0.771 |
| Blue v. Purple : yrsusa | 1.011 | (0.922, 1.109) | 0.815 |
| Green v. Purple : bmicat_preOverweight | 0.452 | (0.15, 1.36) | 0.158 |
| Green v. Purple : bmicat_preObese | 4.335 | (0.998, 18.837) | 0.05 |
| Green v. Purple : lng_lod2_ODT_pg | 0.637 | (0.329, 1.233) | 0.181 |
| Green v. Purple : yrsusa | 0.958 | (0.885, 1.038) | 0.295 |

Table 2: Estimated odds ratios for Vertically Shifted Model with ODT for Girl

Vertical Shifting with CHAMACOS

Red (Group 1), Blue (Group 2), Green (Group 3), Purple (Group 4)

Estimated relative risk ratios comparing each group to the referent Group 4 for ten-fold increase in maternal serum concentrations (ng/g of lipid) of *o,p'*-DDT, *p,p'*-DDT and *p,p'*-DDE with and without adjusting for baseline risk factors.

| | Boy | | | | Girl | | | |
|------------------|---------|--------------------|--------------------|---------|--------------------|---------|--------------------|---------|
| | Group 1 | Group 2 | Group 3 | Group 4 | Group 1 | Group 2 | Group 3 | Group 4 |
| <i>o,p'</i> -DDT | | | | | | | | |
| Unadjusted | 2.452 | 7.865* | 5.279* | ref | 0.465 | 0.919 | 0.754 | ref |
| Adjusted | 1.507 | 5.197 [†] | 3.054 | -- | 0.135* | 0.900 | 0.498 ⁺ | -- |
| <i>p,p'</i> -DDT | | | | | | | | |
| Unadjusted | 1.716 | 3.866* | 3.135 ⁺ | ref | 0.466 | 0.922 | 0.814 | ref |
| Adjusted | 1.185 | 2.882 [†] | 2.063 | -- | 0.215 ⁺ | 1.049 | 0.631 [†] | -- |
| <i>p,p'</i> -DDE | | | | | | | | |
| Unadjusted | 1.293 | 3.594* | 2.553 | ref | 0.321 | 0.778 | 0.818 | ref |
| Adjusted | 0.958 | 2.739 [†] | 1.865 | -- | 0.209 | 0.900 | 0.724 | -- |

Abbreviations: DDE, dichlorodiphenyldichloroethylene; DDT, dichlorodiphenyltrichloroethane.

[†] $P < 0.2$; ⁺ $P < 0.1$; * $P < 0.05$

^a Adjusted for maternal pre-pregnancy BMI, number of years in the USA, duration of breastfeeding and birth weight.