Probabilistic Methods in Linguistics
Lecture 15: Introduction to hierarchical models

Roger Levy

UC San Diego
Department of Linguistics

November 27, 2012
What we’ll start today

- An important class of *hierarchical* probabilistic models
What we’ll start today

- An important class of **hierarchical** probabilistic models
- The class we’ll cover is also called **multi-level** or **mixed-effects** models
What we’ll start today

- An important class of *hierarchical* probabilistic models
- The class we’ll cover is also called *multi-level* or *mixed-effects* models
- Super-important for contemporary linguistic data analysis
What we’ll start today

- An important class of hierarchical probabilistic models
- The class we’ll cover is also called multi-level or mixed-effects models
- Super-important for contemporary linguistic data analysis
- Also a wonderful stepping stone to a large family of richer probabilistic models in linguistics & cognitive science (Ling 252 and other classes)
Motivation

- Observations related to linguistic behavior are clustered at the level of the speaker, and speakers vary from one another in all sorts of attributes.
Motivation

- Observations related to linguistic behavior are clustered at the level of the speaker, and speakers vary from one another in all sorts of attributes.
- Different sentences or even words may have idiosyncratic differences in their ease of understanding or production.
Motivation

- Observations related to linguistic behavior are clustered at the level of the speaker, and speakers vary from one another in all sorts of attributes.
- Different sentences or even words may have idiosyncratic differences in their ease of understanding or production.
- Education-related observations (e.g., vocabulary size) of students have multiple levels of clustering.
Example of clustered data

Empirically observed male adult speaker means for first and second formants of [A] (Peterson and Barney, 1952):
Example of clustered data

- Speakers vary from one another
Example of clustered data

- Speakers vary from one another
- But productions from each speaker vary too
Example of clustered data

- Speakers vary from one another
- But productions from each speaker vary too
- Natural probabilistic model for the $j$-th observation from speaker $i$:

\[
\mu_i \sim \mathcal{N}(\mu, \Sigma_b)
\]

\[
y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y)
\]
Example of clustered data

- Speakers vary from one another
- But productions from each speaker vary too
- Natural probabilistic model for the $j$-th observation from speaker $i$:

\[
\mu_i \sim \mathcal{N}(\mu, \Sigma_b)
\]
\[
y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y)
\]
Example of clustered data

- Speakers vary from one another
- But productions from each speaker vary too
- Natural probabilistic model for the $j$-th observation from speaker $i$:

\[
\mu_i \sim \mathcal{N}(\mu, \Sigma_b) \\
y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y)
\]
Example of clustered data

- We could also write this as

\[ b_i \sim \mathcal{N}(0, \Sigma_b) \]
\[ \mu_i = \mu + b_i \]
\[ y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y) \]

where \( b_i \) is the \( i \)-th speaker’s deviation from the mean of the “hypothetical average speaker’s” mean.
Example of clustered data

- We could also write this as

\[ b_i \sim \mathcal{N}(0, \Sigma_b) \]
\[ \mu_i = \mu + b_i \]
\[ y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y) \]

where \( b_i \) is the \( i \)-th speaker’s deviation from the mean of the “hypothetical average speaker’s” mean

- or as

\[ y_{ij} = \mu + \underbrace{b_i}_{\sim \mathcal{N}(0, \Sigma_b)} + \underbrace{\epsilon_{ij}}_{\sim \mathcal{N}(0, \Sigma_y)} \]
Example of clustered data

We could also write this as

\[ b_i \sim \mathcal{N}(0, \Sigma_b) \]
\[ \mu_i = \mu + b_i \]
\[ y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y) \]

where \( b_i \) is the \( i \)-th speaker’s deviation from the mean of the “hypothetical average speaker’s” mean

or as

\[ y_{ij} = \mu + b_i + \epsilon_{ij} \]
\[ \sim \mathcal{N}(0, \Sigma_b) \]
\[ \sim \mathcal{N}(0, \Sigma_y) \]

So the quantities we have to draw inferences about are
Example of clustered data

- We could also write this as

\[ b_i \sim \mathcal{N}(0, \Sigma_b) \]
\[ \mu_i = \mu + b_i \]
\[ y_{ij} \sim \mathcal{N}(\mu_i, \Sigma_y) \]

where \( b_i \) is the \( i \)-th speaker’s deviation from the mean of the “hypothetical average speaker’s” mean

- or as

\[ y_{ij} = \mu + \underbrace{b_i}_{\sim \mathcal{N}(0, \Sigma_b)} + \underbrace{\epsilon_{ij}}_{\sim \mathcal{N}(0, \Sigma_y)} \]

- So the quantities we have to draw inferences about are
Example of clustered data

- We could also write this as

\[
\begin{align*}
b_i &\sim \mathcal{N}(0, \Sigma_b) \\
\mu_i &= \mu + b_i \\
y_{ij} &\sim \mathcal{N}(\mu_i, \Sigma_y)
\end{align*}
\]

where \( b_i \) is the \( i \)-th speaker's deviation from the mean of the “hypothetical average speaker's” mean

- or as

\[
y_{ij} = \mu + \underbrace{b_i}_{\sim \mathcal{N}(0, \Sigma_b)} + \underbrace{\epsilon_{ij}}_{\sim \mathcal{N}(0, \Sigma_y)}
\]

- So the quantities we have to draw inferences about are \( \mu, \{\mu_i\}, \Sigma_b, \Sigma_y \)
Example of clustered data

One approach: estimate \( \{\hat{\mu}_i\} \) as the observed means for each speaker; let everything else follow from that.
Example of clustered data

One approach: estimate $\{\hat{\mu}_i\}$ as the observed means for each speaker; let everything else follow from that.
Example of clustered data

Estimated residuals from the speaker means $\{\hat{\mu}_i\}$:
Example of clustered data

Simulated data for new speakers:
Example of clustered data

- This process looks somewhat reasonable here
Example of clustered data

- This process looks somewhat reasonable here
- But there are many potential issues:
Example of clustered data

- This process looks somewhat reasonable here
- But there are many potential issues:
  - The cluster-level means \( \{\mu_i\} \) are not known with certainty, and our estimates of \( \Sigma_b \) and \( \Sigma_y \) don’t take that into account
Example of clustered data

- This process looks somewhat reasonable here
- But there are many potential issues:
  - The cluster-level means \( \{\mu_i\} \) are not known with certainty, and our estimates of \( \Sigma_b \) and \( \Sigma_y \) don’t take that into account
  - Hence this approach won’t work gracefully with imbalanced data
Example of clustered data

- This process looks somewhat reasonable here
- But there are many potential issues:
  - The cluster-level means \( \{ \mu_i \} \) are not known with certainty, and our estimates of \( \Sigma_b \) and \( \Sigma_y \) don’t take that into account
  - Hence this approach won’t work gracefully with imbalanced data
  - Likewise, it’s not clear how one would extend this approach when \( y \) is categorical
Example of clustered data

- This process looks somewhat reasonable here
- But there are many potential issues:
  - The cluster-level means \( \{\mu_i\} \) are not known with certainty, and our estimates of \( \Sigma_b \) and \( \Sigma_y \) don’t take that into account
  - Hence this approach won’t work gracefully with imbalanced data
  - Likewise, it’s not clear how one would extend this approach when \( y \) is categorical
  - Finally, this looks rather hopeless for crossed clusterings (e.g., both subjects and items)
Hierarchical models

- Hence the goal here will be to explore techniques that allow us to draw inferences \textit{simultaneously} about all these levels of structure.
A graphical-models visualization

The models we’ve looked at prior to today:

\[ \theta \quad y_1 \quad y_2 \quad \cdots \quad y_n \]
Hierarchical models:
Plate notation

Non-hierarchical models:

Non-plate

\[ \theta \]

\[ y_1 \quad y_2 \quad \ldots \quad y_n \]

Plate

\[ \theta \]

\[ y \]

\[ n \]
Plate notation

Hierarchical models:

Non-plate

Plate
Plate notation

Two different versions of plate notation for these hierarchical models:

Plate

Plate v2