Hierarchical (Multilevel) Modeling

Structure and Inference

Joachim Vandekerckhove

Winter 2025

Hierarchical Modeling

Hierarchical Modeling: Core Idea

A statistical framework for data with **dependencies** from **group structure** or **repeated measures**.

Hierarchical Modeling: Core Idea

A statistical framework for data with **dependencies** from **group structure** or **repeated measures**.

- Examples: Students in classrooms, patients in hospitals, trials within participants, stimuli within conditions.
- Observations within the same group are typically correlated (non-independent).
- Standard methods (like OLS) assume independence, leading to issues.

Hierarchical Modeling: Core Idea

A statistical framework for data with **dependencies** from **group structure** or **repeated measures**.

- Examples: Students in classrooms, patients in hospitals, trials within participants, stimuli within conditions.
- Observations within the same group are typically correlated (non-independent).
- Standard methods (like OLS) assume independence, leading to issues.

Addresses limitations of simpler approaches:

Complete Pooling (Ignore Structure)

Analyze all data together.

- × Underestimates errors.
- × Hides group differences (Ecological Fallacy).
- × Doesn't quantify group variability.

No Pooling (Separate Analyses)

Analyze each group separately.

- × Ignores group similarities.
- × Inefficient.
- × Noisy estimates (esp. small groups).

Hierarchical Modeling: Partial Pooling

Hierarchical models provide a statistically principled compromise: **Partial Pooling**

Hierarchical Modeling: Partial Pooling

Hierarchical models provide a statistically principled compromise: **Partial Pooling**

- Information is adaptively shared across groups.

Hierarchical Modeling: Partial Pooling

Hierarchical models provide a statistically principled compromise: **Partial Pooling**

- Information is adaptively shared across groups.
- "Borrow strength": Groups inform each other.

Hierarchical Modeling: Partial Pooling

Hierarchical models provide a statistically principled compromise: **Partial Pooling**

- Information is adaptively shared across groups.
- "Borrow strength": Groups inform each other.
- Improves individual group estimates (especially for noisy/small groups).

Hierarchical Modeling: Partial Pooling

Hierarchical models provide a statistically principled compromise: **Partial Pooling**

- Information is adaptively shared across groups.
- "Borrow strength": Groups inform each other.
- Improves individual group estimates (especially for noisy/small groups).
- Simultaneously estimates population-level effects and the extent of variation between groups.

Hierarchical Model: Formulation (Varying Intercepts)

Let y_{ij} be outcome for observation i in group j, x_{ij} a predictor.

Hierarchical Model: Formulation (Varying Intercepts)

Let y_{ij} be outcome for observation i in group j, x_{ij} a predictor. **Model Structure:**

- **Level 1 (Within-Group):**

  $$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$

  where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$ (Residual variance).

  Here, the intercept α_j varies by group, but the slope β is fixed (common).

Hierarchical Model: Formulation (Varying Intercepts)

Let y_{ij} be outcome for observation i in group j, x_{ij} a predictor. **Model Structure:**

- **Level 1 (Within-Group):**

  $$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$

  where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$ (Residual variance).

  Here, the intercept α_j varies by group, but the slope β is fixed (common).
- **Level 2 (Between-Group):**

  $$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$$

  The group intercepts (α_j) are drawn from a population distribution.

  - μ_α: Population average intercept.
  - σ_α^2: Variance of intercepts across groups.

Hierarchical Model: Formulation (Varying Intercepts)

Let y_{ij} be outcome for observation i in group j, x_{ij} a predictor. **Model Structure:**

- **Level 1 (Within-Group):**

  $$y_{ij} = \alpha_j + \beta x_{ij} + \epsilon_{ij}$$

  where $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$ (Residual variance).

  Here, the intercept α_j varies by group, but the slope β is fixed (common).
- **Level 2 (Between-Group):**

  $$\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$$

  The group intercepts (α_j) are drawn from a population distribution.

  - μ_α: Population average intercept.
  - σ_α^2: Variance of intercepts across groups.

This structure explicitly models the dependency within groups.

Hierarchical Model: Formulation (Varying Slopes)

We can also allow the slope (β_j) to vary across groups.

Hierarchical Model: Formulation (Varying Slopes)

We can also allow the slope (β_j) to vary across groups. **Model Structure:**

- **Level 1 (Within-Group):**

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$$

Hierarchical Model: Formulation (Varying Slopes)

We can also allow the slope (β_j) to vary across groups. **Model Structure:**

- **Level 1 (Within-Group):**

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma_y^2)$$

- **Level 2 (Between-Group):**

  Intercepts α_j and slopes β_j are drawn from a population distribution, often modeled as multivariate normal to capture potential correlation (ρ).

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathcal{MVN}\left(\begin{pmatrix} \mu_\alpha \\ \mu_\beta \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right)$$

  - μ_α, μ_β: Population average intercept and slope.
  - $\sigma_\alpha^2, \sigma_\beta^2$: Variance of intercepts and slopes across groups.
  - ρ: Correlation between intercepts and slopes across groups.

Hierarchical Modeling: Terminology

Fixed Effects Parameters assumed constant across all groups (e.g., common slope β) OR population-level means (e.g., μ_α, μ_β). Represent average effects.

Hierarchical Modeling: Terminology

Fixed Effects Parameters assumed constant across all groups (e.g., common slope β) OR population-level means (e.g., μ_α, μ_β). Represent average effects.

Random Effects Parameters allowed to vary across groups (e.g., group intercepts α_j, group slopes β_j). Often conceptualized as group-specific deviations from the fixed effect mean (e.g., $u_{0j} = \alpha_j - \mu_\alpha$). Capture heterogeneity.

Hierarchical Modeling: Terminology

Fixed Effects Parameters assumed constant across all groups (e.g., common slope β) OR population-level means (e.g., μ_α, μ_β). Represent average effects.

Random Effects Parameters allowed to vary across groups (e.g., group intercepts α_j, group slopes β_j). Often conceptualized as group-specific deviations from the fixed effect mean (e.g., $u_{0j} = \alpha_j - \mu_\alpha$). Capture heterogeneity.

Variance Components Parameters characterizing variability at different levels:

- Level 1: Residual variance σ_y^2.
- Level 2: Random effect variances $(\sigma_\alpha^2, \sigma_\beta^2)$ and their correlation/covariance ($\mathbf{\Sigma}$). Quantify the magnitude of group

Hierarchical Modeling: Partial Pooling / Shrinkage

The estimate for a group-specific parameter (e.g., $\hat{\alpha}_j$) balances two sources of information:

Hierarchical Modeling: Partial Pooling / Shrinkage

The estimate for a group-specific parameter (e.g., $\hat{\alpha}_j$) balances two sources of information:

1. Information from the groupś own data (the no-pooling estimate, $\hat{\alpha}_{j,\text{no pool}}$).
2. Information from the overall population distribution (the population mean estimate, $\hat{\mu}_\alpha$).

Hierarchical Modeling: Partial Pooling / Shrinkage

The estimate for a group-specific parameter (e.g., $\hat{\alpha}_j$) balances two sources of information:

1. Information from the groupś own data (the no-pooling estimate, $\hat{\alpha}_{j,\text{no pool}}$).
2. Information from the overall population distribution (the population mean estimate, $\hat{\mu}_\alpha$).

$$\hat{\alpha}_j \approx w_j \hat{\alpha}_{j,\text{no pool}} + (1 - w_j)\hat{\mu}_\alpha$$

Hierarchical Modeling: Partial Pooling / Shrinkage

The estimate for a group-specific parameter (e.g., $\hat{\alpha}_j$) balances two sources of information:

1. Information from the groupś own data (the no-pooling estimate, $\hat{\alpha}_{j,\text{no pool}}$).
2. Information from the overall population distribution (the population mean estimate, $\hat{\mu}_\alpha$).

$\hat{\alpha}_j \approx w_j \hat{\alpha}_{j,\text{no pool}} + (1 - w_j)\hat{\mu}_\alpha$ The weight w_j (or shrinkage factor) depends on precision:

Hierarchical Modeling: Partial Pooling / Shrinkage

The estimate for a group-specific parameter (e.g., $\hat{\alpha}_j$) balances two sources of information:

1. Information from the groupś own data (the no-pooling estimate, $\hat{\alpha}_{j,\text{no pool}}$).
2. Information from the overall population distribution (the population mean estimate, $\hat{\mu}_\alpha$).

$\hat{\alpha}_j \approx w_j \hat{\alpha}_{j,\text{no pool}} + (1 - w_j)\hat{\mu}_\alpha$ The weight w_j (or shrinkage factor) depends on precision:

$w_j \approx \frac{\text{Precision of Group Estimate}}{\text{Precision of Group Estimate}+\text{Precision of Population Estimate}} = \frac{1/\text{Var}(\hat{\alpha}_{j,\text{no pool}})}{1/\text{Var}(\hat{\alpha}_{j,\text{no pool}})+1/\sigma_\alpha^2}$

Hierarchical Modeling: Partial Pooling / Shrinkage

The estimate for a group-specific parameter (e.g., $\hat{\alpha}_j$) balances two sources of information:

1. Information from the groupś own data (the no-pooling estimate, $\hat{\alpha}_{j,\text{no pool}}$).
2. Information from the overall population distribution (the population mean estimate, $\hat{\mu}_\alpha$).

$\hat{\alpha}_j \approx w_j \hat{\alpha}_{j,\text{no pool}} + (1 - w_j)\hat{\mu}_\alpha$ The weight w_j (or shrinkage factor) depends on precision:

$$w_j \approx \frac{\text{Precision of Group Estimate}}{\text{Precision of Group Estimate}+\text{Precision of Population Estimate}} = \frac{1/\text{Var}(\hat{\alpha}_{j,\text{no pool}})}{1/\text{Var}(\hat{\alpha}_{j,\text{no pool}})+1/\sigma_\alpha^2}$$

- $\text{Var}(\hat{\alpha}_{j,\text{no pool}})$ depends on group size n_j and within-group variance σ_y^2.
- Group estimates are **shrunk** towards the population mean.

Hierarchical Modeling: Adaptive Shrinkage

The amount of shrinkage is adaptive and data-dependent: **More Shrinkage** (towards $\hat{\mu}_\alpha$) when:

- Group has less data / noisy estimate (large $\mathrm{Var}(\hat{\alpha}_{j,\text{no pool}})$).
- Groups are very similar (small between-group variance σ_α^2).

Hierarchical Modeling: Adaptive Shrinkage

The amount of shrinkage is adaptive and data-dependent: **More Shrinkage** (towards $\hat{\mu}_\alpha$) when:

- Group has less data / noisy estimate (large $\text{Var}(\hat{\alpha}_{j,\text{no pool}})$).
- Groups are very similar (small between-group variance σ_α^2).

Less Shrinkage (estimate closer to group's own data) when:

- Group has more data / precise estimate (small $\text{Var}(\hat{\alpha}_{j,\text{no pool}})$).
- Groups are very dissimilar (large between-group variance σ_α^2).

Hierarchical Modeling: Adaptive Shrinkage

The amount of shrinkage is adaptive and data-dependent: **More Shrinkage** (towards $\hat{\mu}_\alpha$) when:

- Group has less data / noisy estimate (large $\text{Var}(\hat{\alpha}_{j,\text{no pool}})$).
- Groups are very similar (small between-group variance σ_α^2).

Less Shrinkage (estimate closer to group's own data) when:

- Group has more data / precise estimate (small $\text{Var}(\hat{\alpha}_{j,\text{no pool}})$).
- Groups are very dissimilar (large between-group variance σ_α^2).

This adaptive regularization prevents overfitting and leads to better out-of-sample predictions compared to no-pooling or complete-pooling models.

Hierarchical Modeling: Bayesian Approach

Bayesian methods provide a natural framework:

Hierarchical Modeling: Bayesian Approach

Bayesian methods provide a natural framework:

- **Full Probability Model:** Specify the entire structure:
  - Level 1 Likelihood: $P(\text{Data}|\text{Level 1 Params})$
  - Level 2 Priors: $P(\text{Level 1 Params}|\text{Level 2 Hyperparams})$
  - Hyperpriors: $P(\text{Level 2 Hyperparams})$

Hierarchical Modeling: Bayesian Approach

Bayesian methods provide a natural framework:

- **Full Probability Model:** Specify the entire structure:
  - Level 1 Likelihood: $P(\text{Data}|\text{Level 1 Params})$
  - Level 2 Priors: $P(\text{Level 1 Params}|\text{Level 2 Hyperparams})$
  - Hyperpriors: $P(\text{Level 2 Hyperparams})$
- **Coherent Uncertainty:** Get full posterior distributions for all parameters (fixed effects, random effects, variance components), naturally propagating uncertainty.

Hierarchical Modeling: Bayesian Approach

Bayesian methods provide a natural framework:

- **Full Probability Model:** Specify the entire structure:
  - Level 1 Likelihood: $P(\text{Data}|\text{Level 1 Params})$
  - Level 2 Priors: $P(\text{Level 1 Params}|\text{Level 2 Hyperparams})$
  - Hyperpriors: $P(\text{Level 2 Hyperparams})$
- **Coherent Uncertainty:** Get full posterior distributions for all parameters (fixed effects, random effects, variance components), naturally propagating uncertainty.
- **Computation:** Modern MCMC (e.g., HMC/NUTS in Stan, PyMC) handles complex posteriors effectively.

Hierarchical Modeling: Bayesian Approach

Bayesian methods provide a natural framework:

- **Full Probability Model:** Specify the entire structure:
  - Level 1 Likelihood: $P(\text{Data}|\text{Level 1 Params})$
  - Level 2 Priors: $P(\text{Level 1 Params}|\text{Level 2 Hyperparams})$
  - Hyperpriors: $P(\text{Level 2 Hyperparams})$
- **Coherent Uncertainty:** Get full posterior distributions for all parameters (fixed effects, random effects, variance components), naturally propagating uncertainty.
- **Computation:** Modern MCMC (e.g., HMC/NUTS in Stan, PyMC) handles complex posteriors effectively.
- **Prior Specification:** Requires care!
  - Fixed Effects / Means (μś): Often weakly informative (e.g., wide Normal).
  - Variance Components (σś): Crucial! Use weakly informative priors concentrated away from zero (e.g., Half-Normal, Half-Cauchy) to avoid issues.
  - Correlations (ρś): LKJ priors are common for correlation matrices.

Fitting is just the start! Rigorous checking is essential:

Hierarchical Modeling: Checking & Interpretation

Fitting is just the start! Rigorous checking is essential:

- **MCMC Convergence:** Check $\hat{R}$ (should be ≈ 1.0), Effective Sample Size (ESS), trace plots.

Hierarchical Modeling: Checking & Interpretation

Fitting is just the start! Rigorous checking is essential:

- **MCMC Convergence:** Check $\hat{R}$ (should be ≈ 1.0), Effective Sample Size (ESS), trace plots.
- **Prior Sensitivity Analysis:** Do results change much with different reasonable priors?

Hierarchical Modeling: Checking & Interpretation

Fitting is just the start! Rigorous checking is essential:

- **MCMC Convergence:** Check $\hat{R}$ (should be ≈ 1.0), Effective Sample Size (ESS), trace plots.
- **Prior Sensitivity Analysis:** Do results change much with different reasonable priors?
- **Posterior Predictive Checks (PPCs):**
  - Simulate data from the fitted model.
  - Compare simulated data distributions to observed data. Mismatches indicate model problems.

Hierarchical Modeling: Checking & Interpretation

Fitting is just the start! Rigorous checking is essential:

- **MCMC Convergence:** Check $\hat{R}$ (should be ≈ 1.0), Effective Sample Size (ESS), trace plots.
- **Prior Sensitivity Analysis:** Do results change much with different reasonable priors?
- **Posterior Predictive Checks (PPCs):**
  - Simulate data from the fitted model.
  - Compare simulated data distributions to observed data. Mismatches indicate model problems.
- **Model Comparison:** Use LOO-CV (via PSIS-LOO) or WAIC to compare models (e.g., different random effects structures). Estimates out-of-sample prediction accuracy.

Hierarchical Modeling: Checking & Interpretation

Fitting is just the start! Rigorous checking is essential:

- **MCMC Convergence:** Check $\hat{R}$ (should be ≈ 1.0), Effective Sample Size (ESS), trace plots.
- **Prior Sensitivity Analysis:** Do results change much with different reasonable priors?
- **Posterior Predictive Checks (PPCs):**
  - Simulate data from the fitted model.
  - Compare simulated data distributions to observed data. Mismatches indicate model problems.
- **Model Comparison:** Use LOO-CV (via PSIS-LOO) or WAIC to compare models (e.g., different random effects structures). Estimates out-of-sample prediction accuracy.

Interpretation: Focus on population parameters (μś, fixed βś), magnitude of variation (σś), and potentially shrunken group estimates (α_jś, β_jś), always with uncertainty

Hierarchical Modeling: Advanced Topics

Briefly:

- **Crossed vs. Nested Random Effects:**
  - Nested: Students in classrooms, classrooms in schools.
  - Crossed: Participants respond to multiple stimuli (random effects for participant AND stimulus, not nested).
  - Models can handle both.

Hierarchical Modeling: Advanced Topics

Briefly:

- **Crossed vs. Nested Random Effects:**
    - Nested: Students in classrooms, classrooms in schools.
    - Crossed: Participants respond to multiple stimuli (random effects for participant AND stimulus, not nested).
    - Models can handle both.
- **Generalized Linear Hierarchical Models (GLMMs):**
    - For non-Gaussian outcomes (binary, count data).
    - Uses appropriate likelihoods (Binomial, Poisson) and link functions (logit, log).

Hierarchical Modeling: Advanced Topics

Briefly:

- **Crossed vs. Nested Random Effects:**
  - Nested: Students in classrooms, classrooms in schools.
  - Crossed: Participants respond to multiple stimuli (random effects for participant AND stimulus, not nested).
  - Models can handle both.
- **Generalized Linear Hierarchical Models (GLMMs):**
  - For non-Gaussian outcomes (binary, count data).
  - Uses appropriate likelihoods (Binomial, Poisson) and link functions (logit, log).
- **Non-Centered Parameterization (Reparameterization):**
  - Crucial for MCMC efficiency, especially with small group variances or sparse data.
  - Instead of $\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$...
  - ... use $z_j \sim \mathcal{N}(0, 1)$ and define $\alpha_j = \mu_\alpha + z_j \times \sigma_\alpha$.
  - Reduces posterior correlations between μ_α and σ_α, avoids "funnels".

Hierarchical Modeling: Advanced Topics

Briefly:

- **Crossed vs. Nested Random Effects:**
  - Nested: Students in classrooms, classrooms in schools.
  - Crossed: Participants respond to multiple stimuli (random effects for participant AND stimulus, not nested).
  - Models can handle both.
- **Generalized Linear Hierarchical Models (GLMMs):**
  - For non-Gaussian outcomes (binary, count data).
  - Uses appropriate likelihoods (Binomial, Poisson) and link functions (logit, log).
- **Non-Centered Parameterization (Reparameterization):**
  - Crucial for MCMC efficiency, especially with small group variances or sparse data.
  - Instead of $\alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$...
  - ... use $z_j \sim \mathcal{N}(0, 1)$ and define $\alpha_j = \mu_\alpha + z_j \times \sigma_\alpha$.
  - Reduces posterior correlations between μ_α and σ_α, avoids "funnels".
- **Multilevel Models for Longitudinal Data:** Analyzing change over time, often

Hierarchical (Multilevel) Modeling

Structure and Inference

Joachim Vandekerckhove

Winter 2025