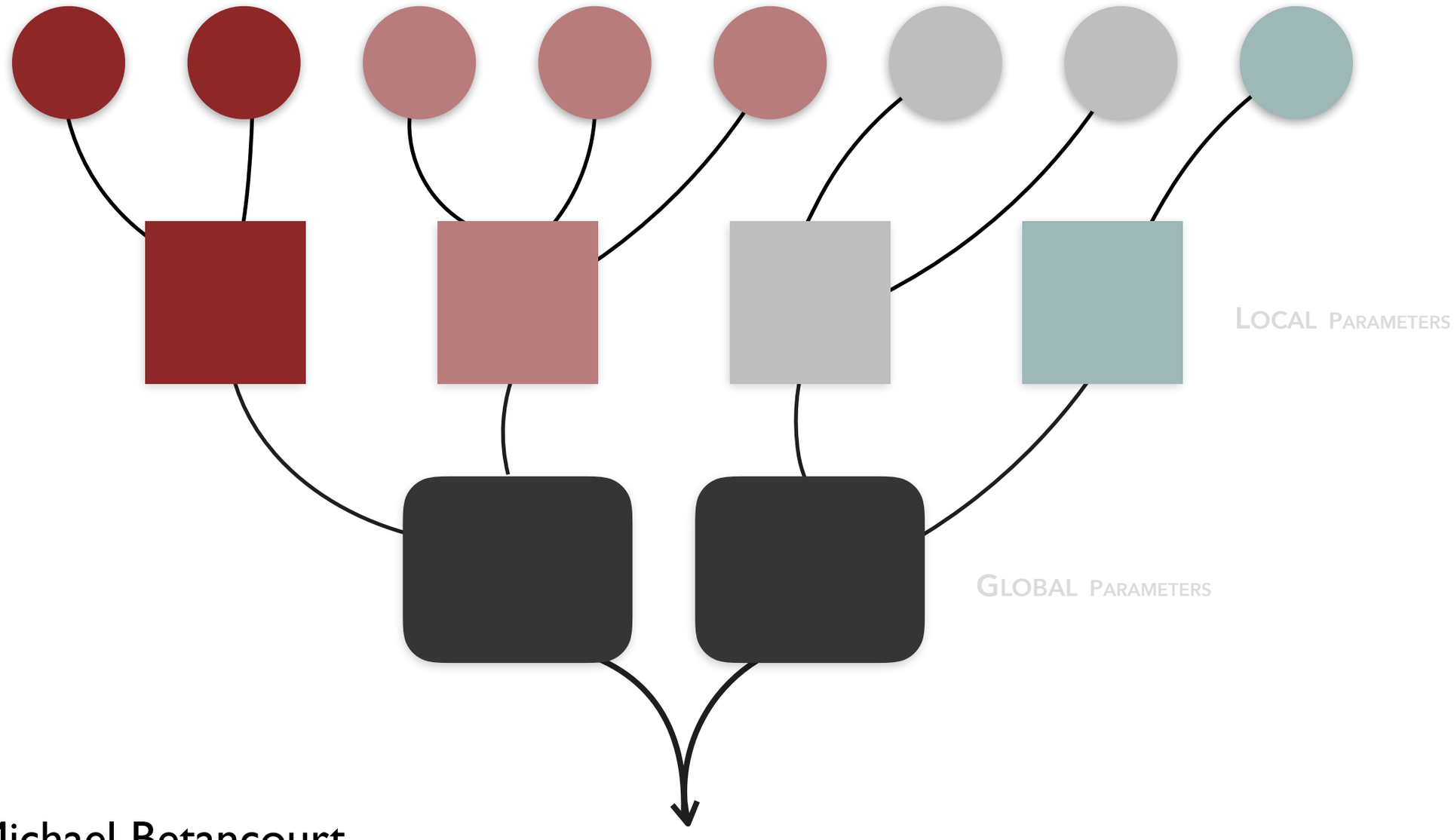
Hierarchical and Multilevel Models

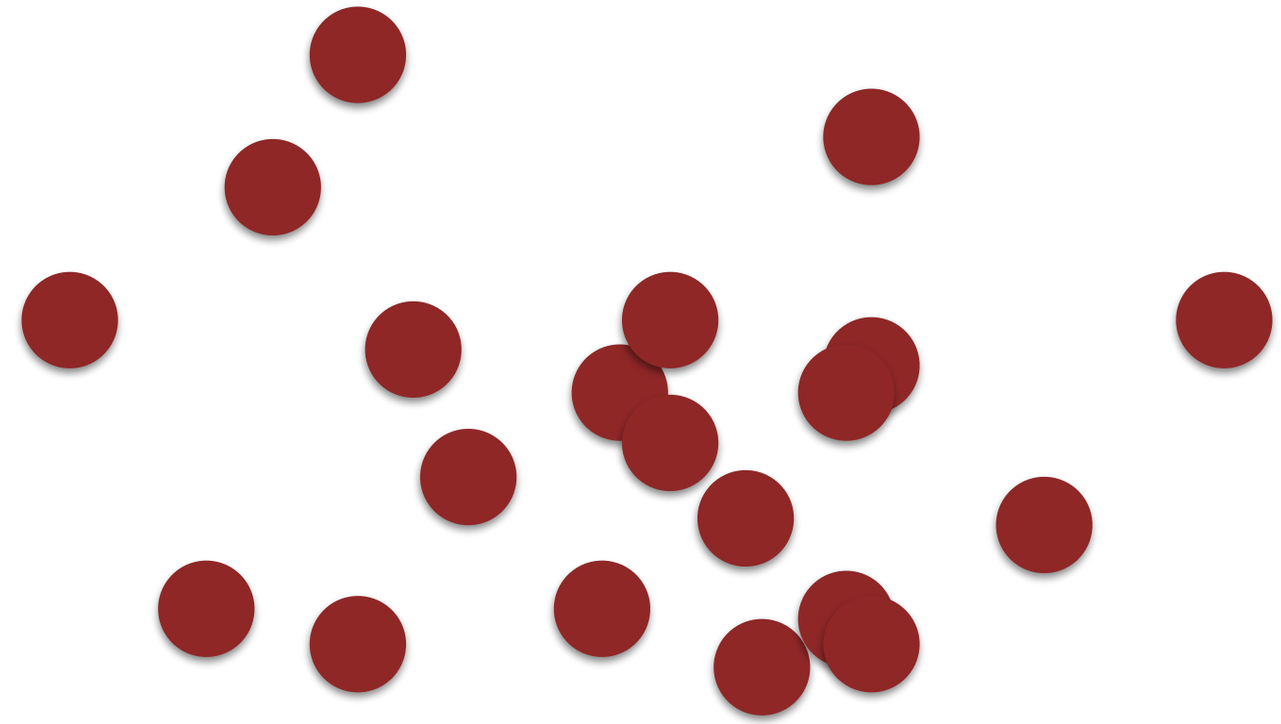


Thanks to Michael Betancourt

A common problem in applied statistics
is modeling individuals of a *population*.

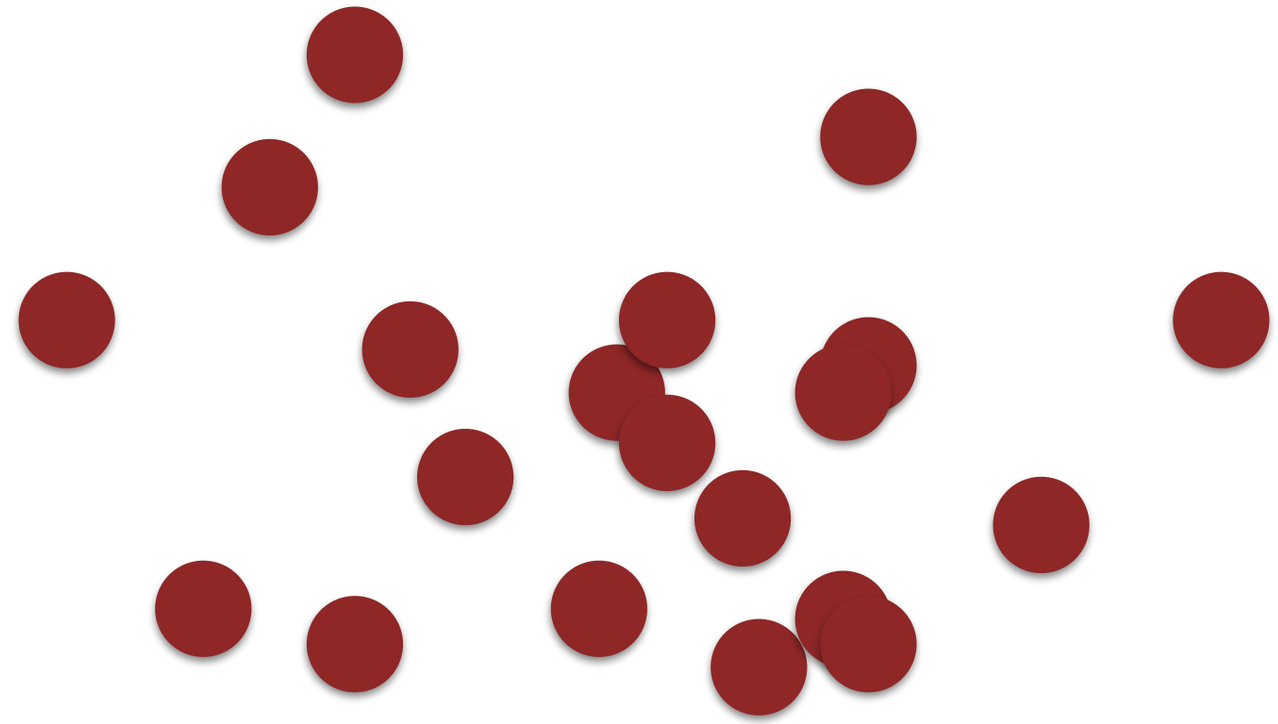
$$y \sim \mathcal{N}(\mathbf{X}^T \boldsymbol{\beta} + \alpha, \sigma)$$

A common problem in applied statistics is modeling individuals of a *population*.



$$y \sim \mathcal{N}(\mathbf{X}^T \boldsymbol{\beta} + \alpha, \sigma)$$

If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias.



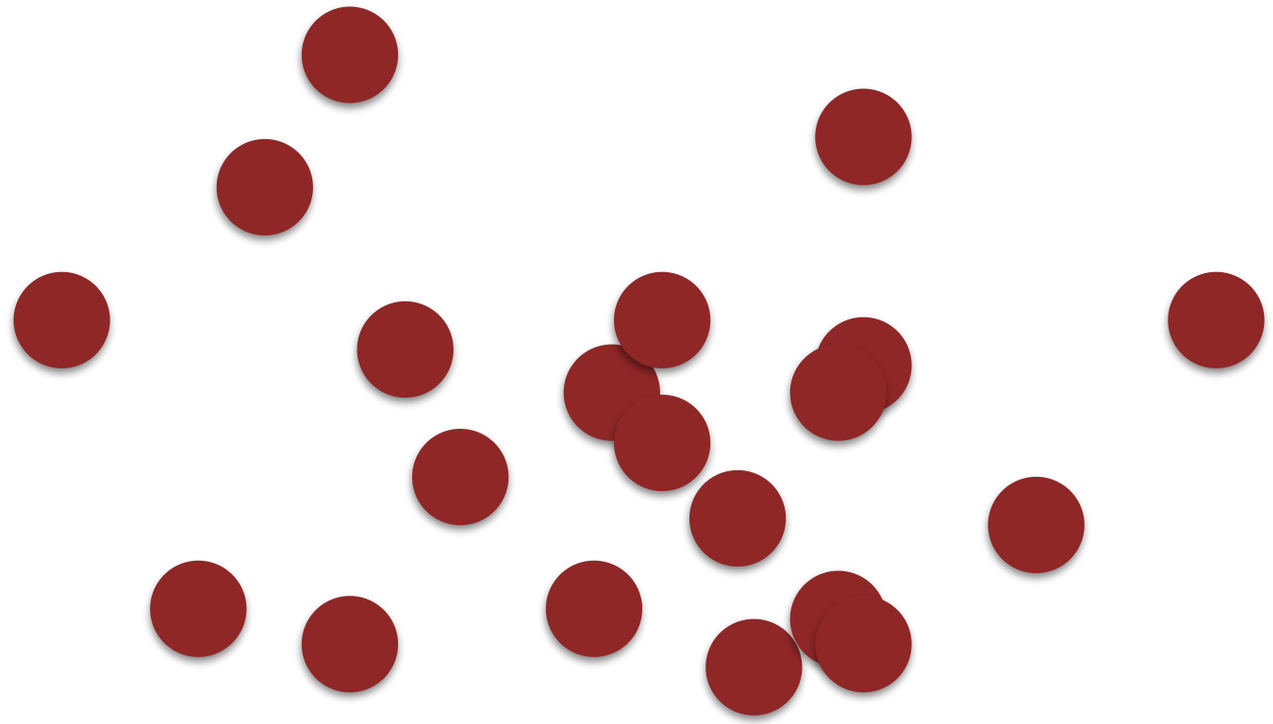
If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias.

$$y_n \sim \mathcal{N}\left(\mathbf{X}^T \boldsymbol{\beta} + \alpha, \sigma\right)$$

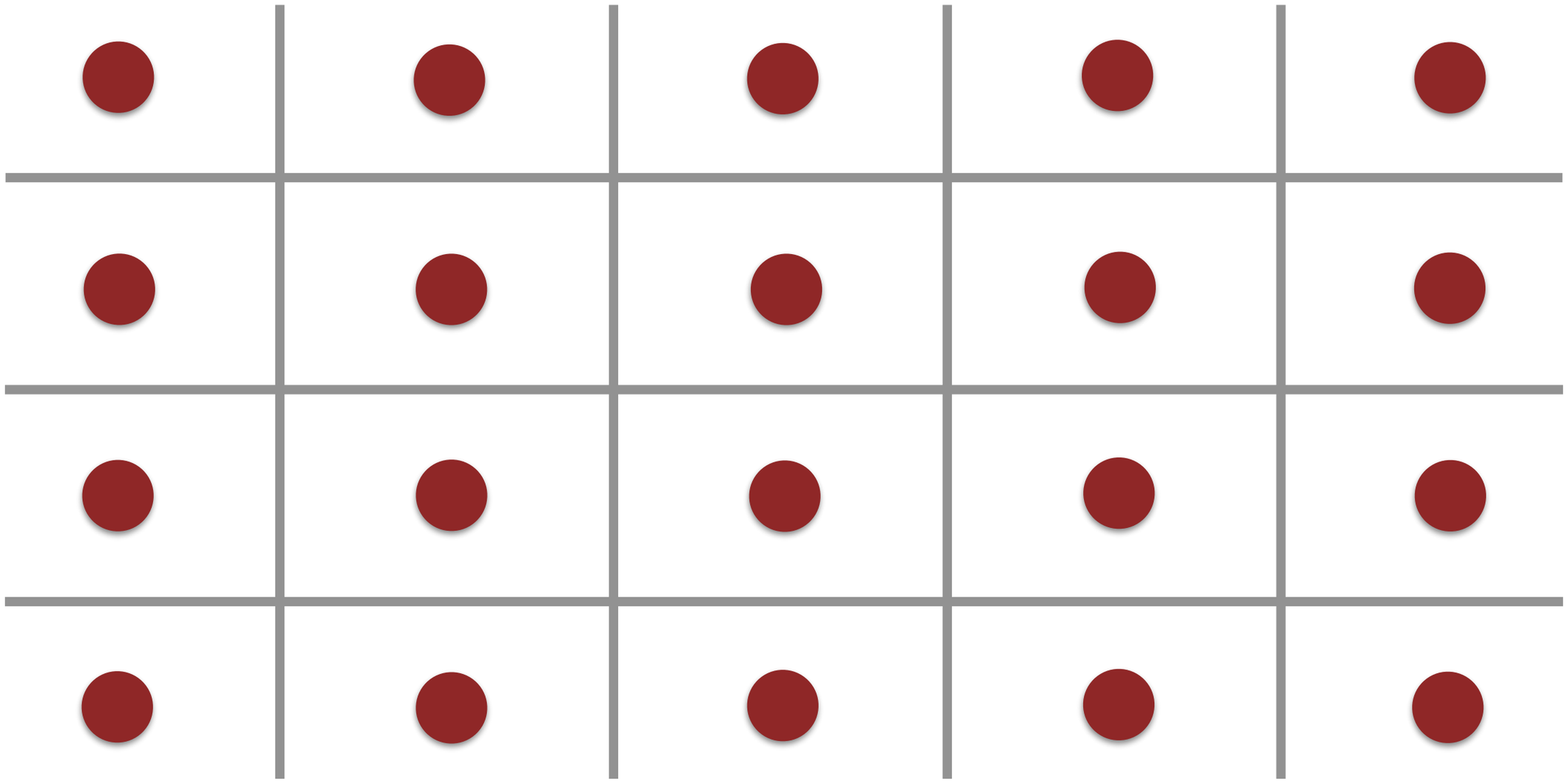
Complete pooling model

```
data {
  int N;
  int K;
  int<lower=0, upper=1> y[N];
  vector[N] x;
}

parameters {
  real beta;
  real alpha;
}

model {
   beta ~ normal(0, 10);
   alpha ~ normal(0, 10);
   y ~ bernoulli_logit(x * beta + alpha);
}
```

Modeling every individual separately avoids any bias, but then the data becomes very sparse and inferences weak.



Modeling every individual separately avoids any bias, but then the data becomes very sparse and inferences weak.



$$y_n \sim \mathcal{N}\left(\mathbf{X}^T \boldsymbol{\beta}_n + \alpha_n, \sigma\right)$$

No pooling model

```
data {
  int N;
  int K;
  int<lower=0, upper=1> y[N];
  vector[N] x;
}

parameters {
  // different slope and intercept for each indiv
  vector[N] beta;
  vector[N] alpha;
}

model {
   beta ~ normal(0, 10);
   alpha ~ normal(0, 10);
   y ~ bernoulli_logit(x .* beta + alpha);
}
```

A compromise between complete pooling and no pooling that could balance bias and variance would be ideal.

A compromise between complete pooling and no pooling that could balance bias and variance would be ideal.

- **Simultaneously** estimate individual (local) parameters and population (global) parameters

A compromise between complete pooling and no pooling that could balance bias and variance would be ideal.

- **Simultaneously** estimate individual (local) parameters and population (global) parameters

- But can be helpful to imagine it **iteratively**

Start by fitting all individuals separately

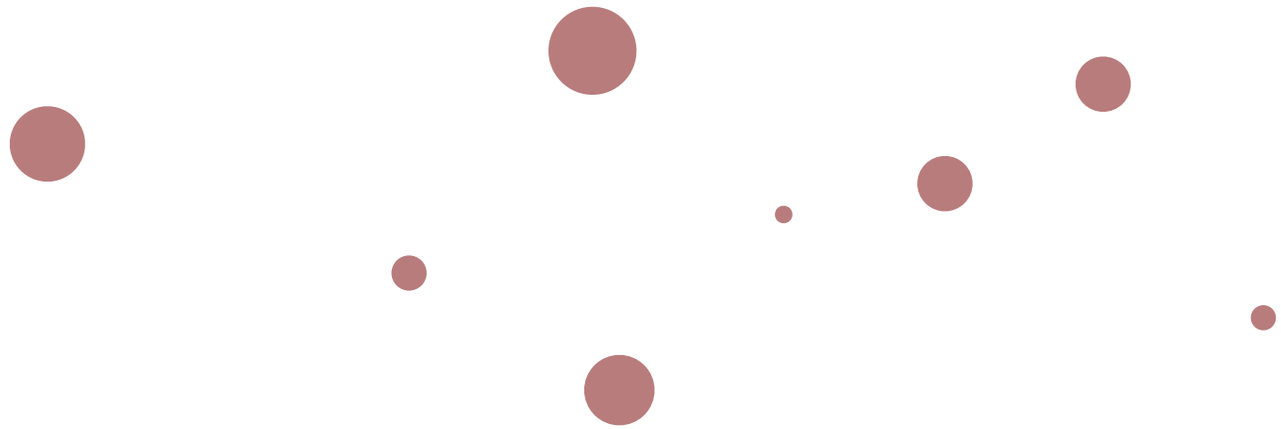
Start by fitting all individuals separately

- Larger circle = larger uncertainty
  - less data

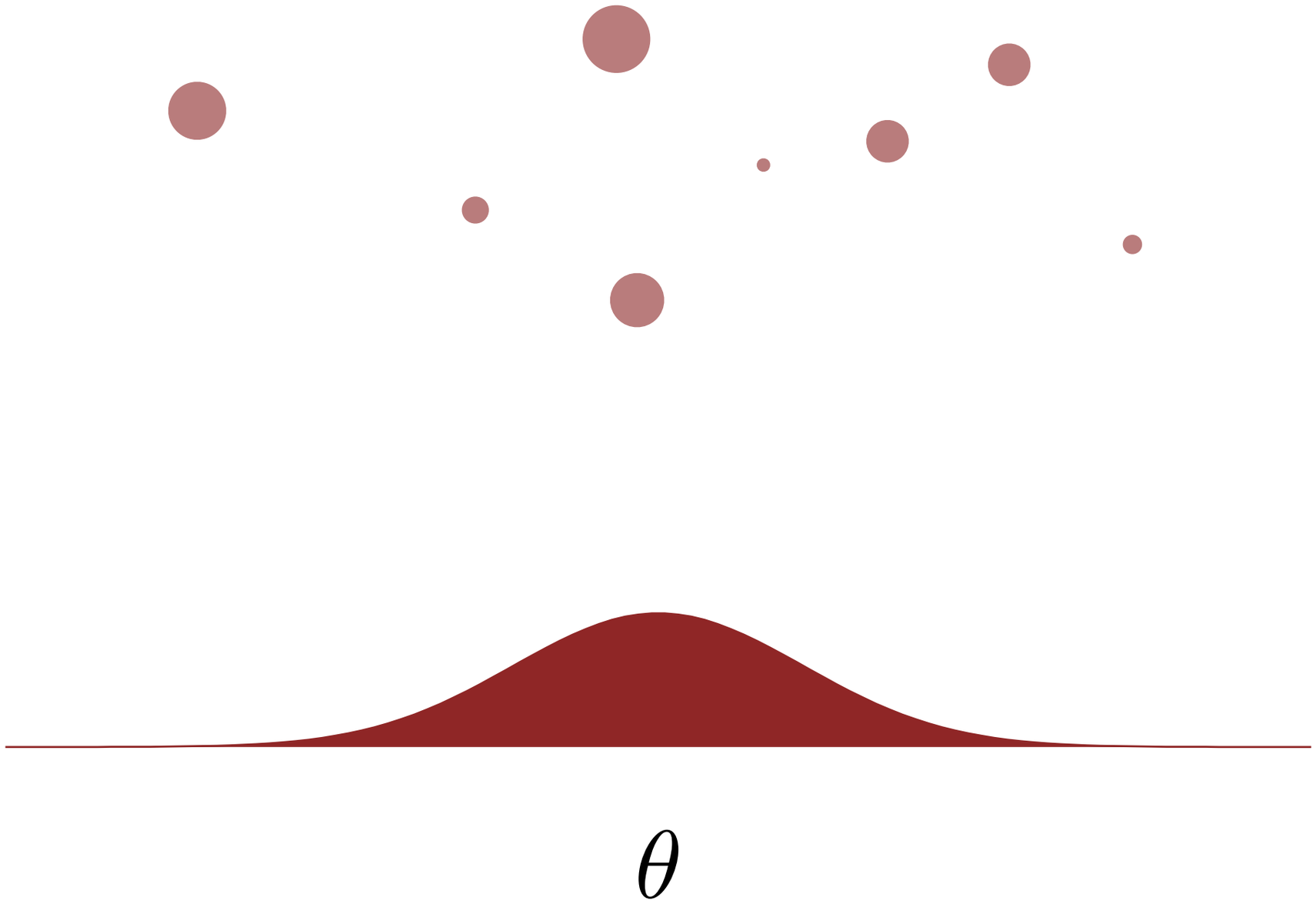
Start by fitting all individuals separately

- Larger circle = larger uncertainty
  - less data
- Smaller circle = smaller uncertainty
  - more data

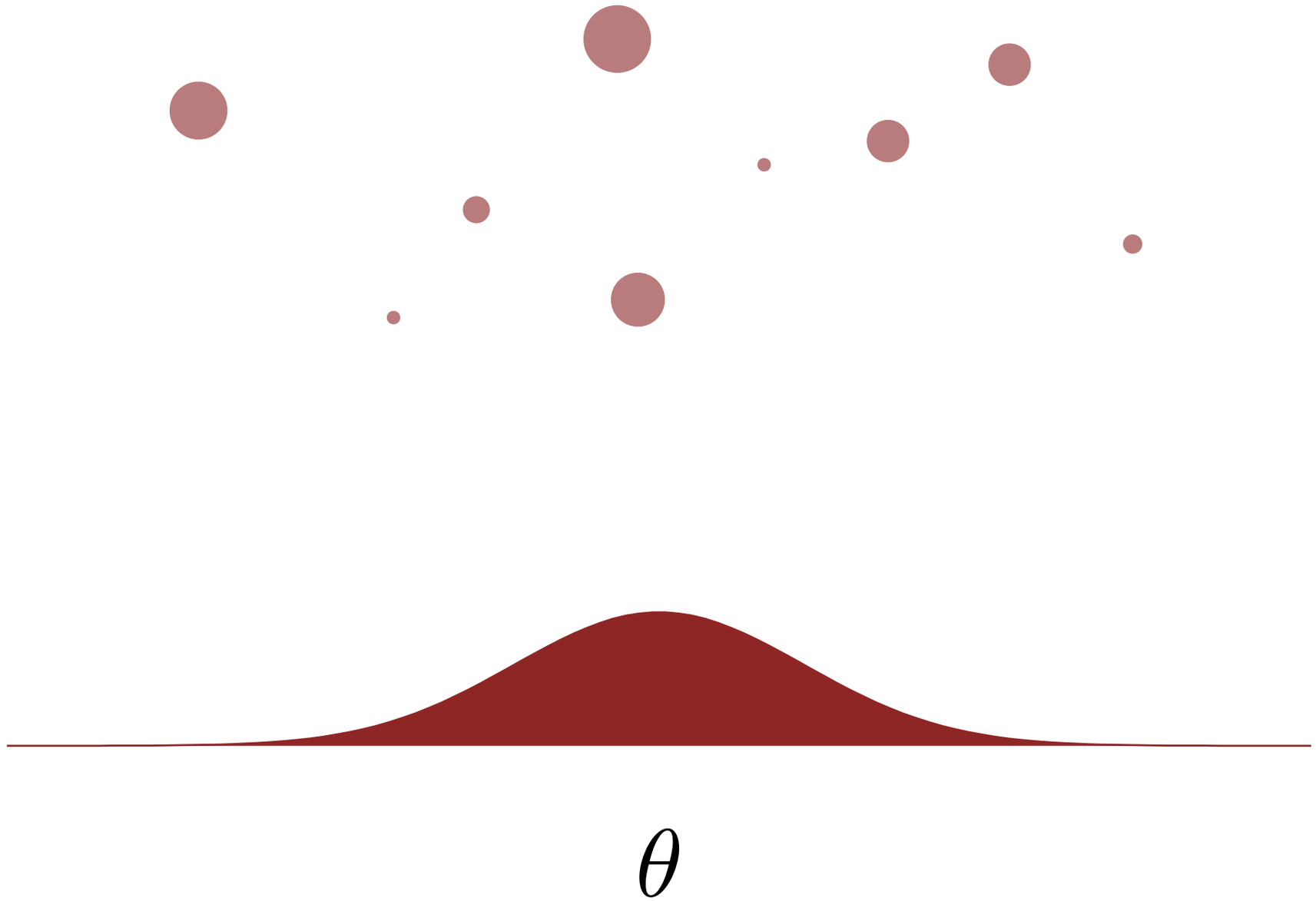
Take individual fits and estimate population distribution



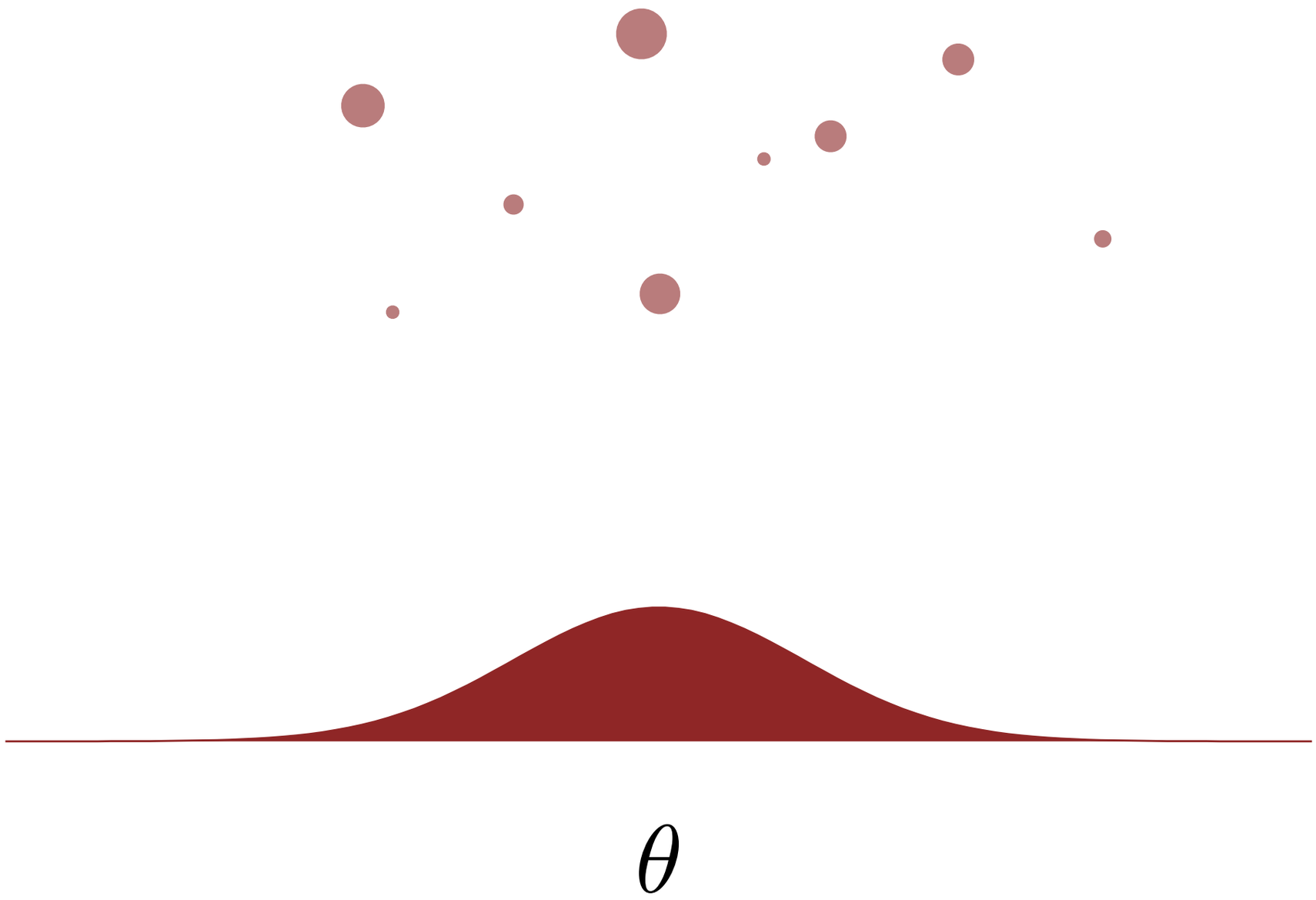
Take individual fits and estimate population distribution



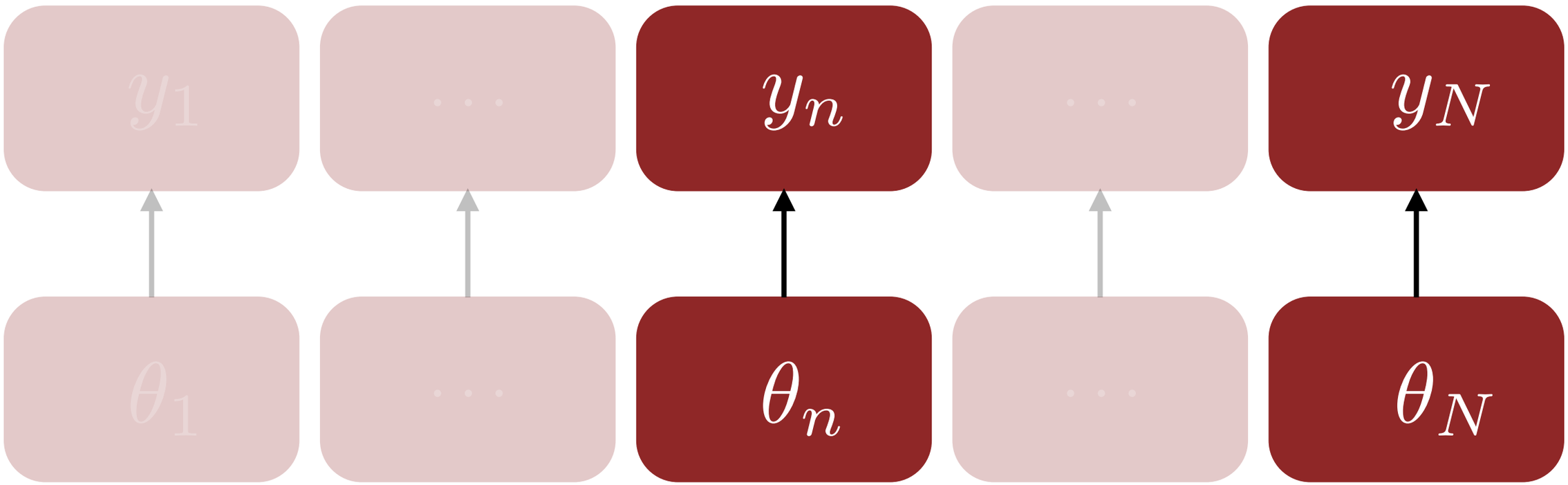
Use population distribution as prior for individuals



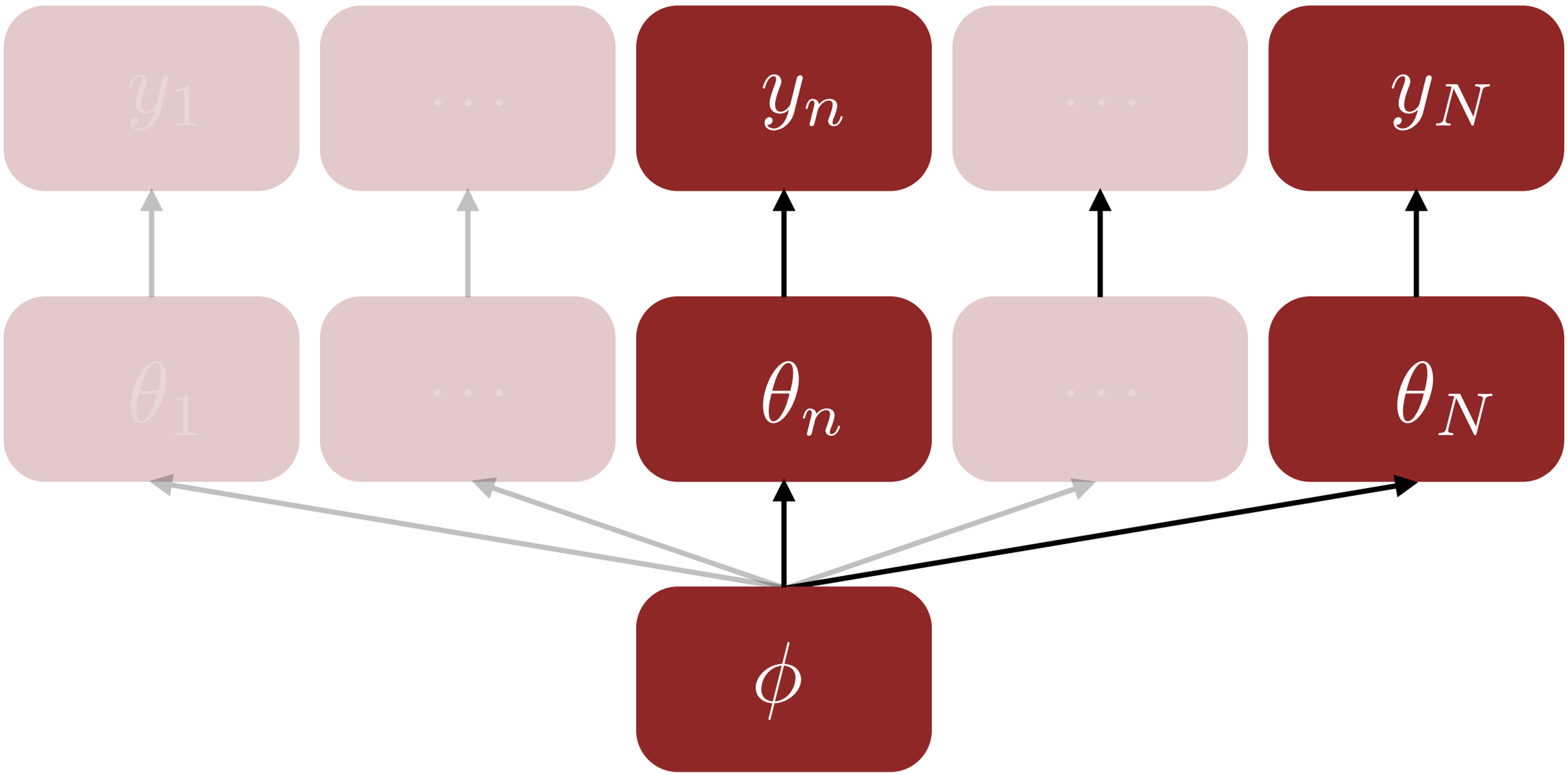
Use population distribution as prior for individuals



The hyperparameters couple all of the groups together, partially pooling the data and balancing bias and variance.



The hyperparameters couple all of the groups together, partially pooling the data and balancing bias and variance.



The most common population model is a Gaussian, where we model the population mean and variance.

$$\prod_{n=1}^{N} p(y_n|\theta_n)\, p(\theta_n|\phi)\, p(\phi)$$

The most common population model is a Gaussian, where we model the population mean and variance.

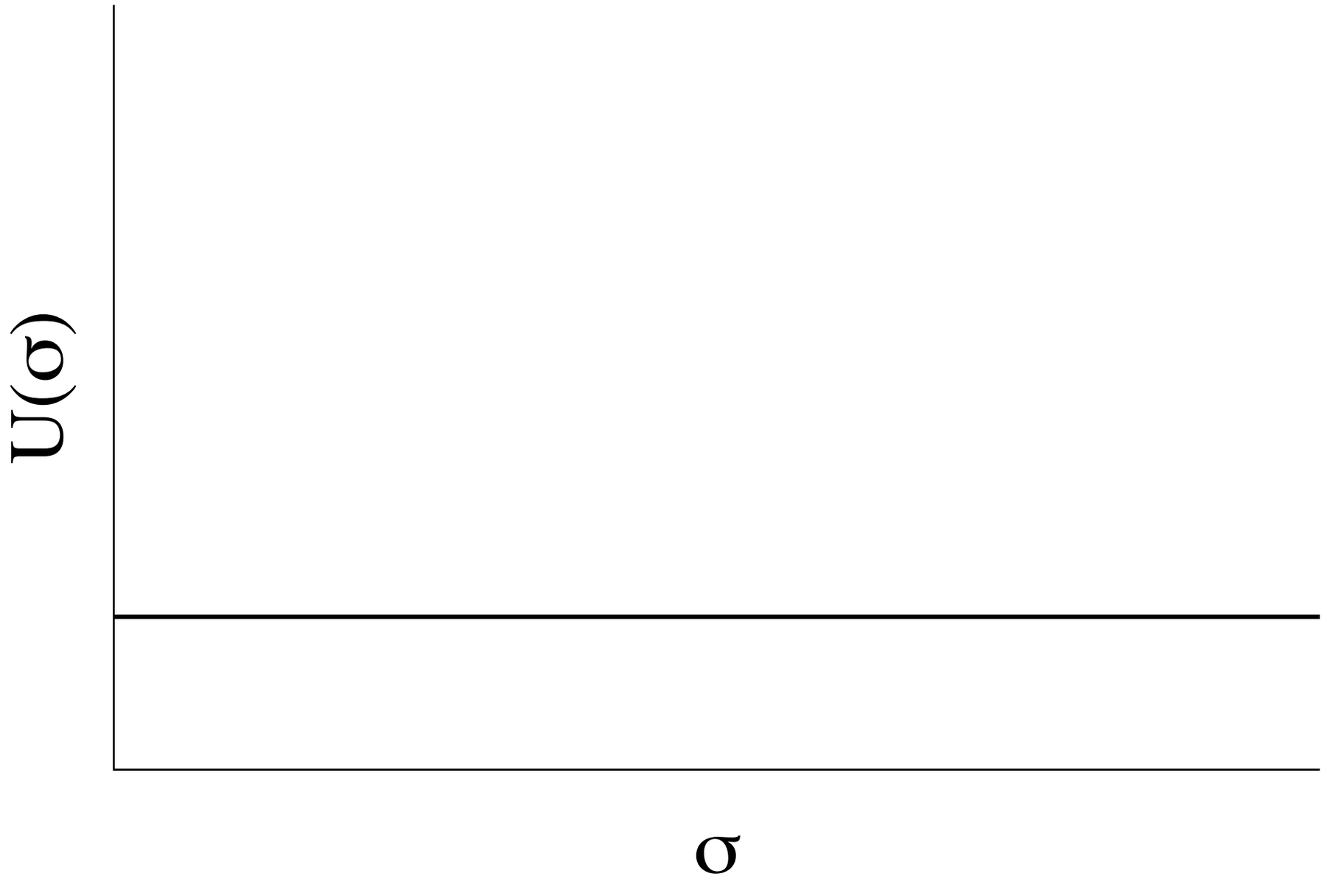
$$\prod_{n=1}^{N} p(y_n|\theta_n)\, p(\theta_n|\phi)\, p(\phi)$$

$$\prod_{n=1}^{N} p(y_n|\theta_n)\, \mathcal{N}(\theta_n|\mu,\sigma)\, p(\mu)\, p(\sigma)$$

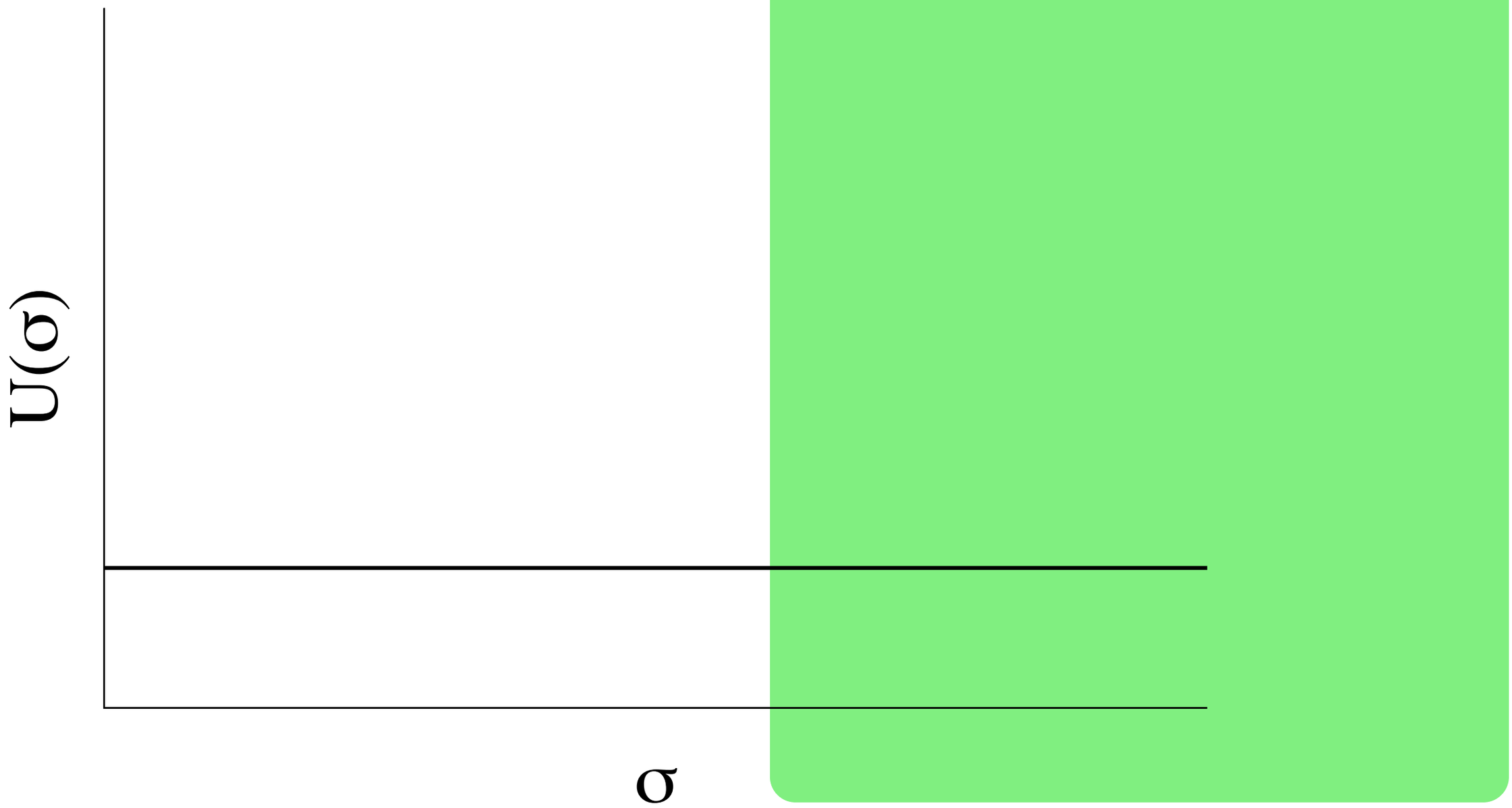
Informative prior distributions on the population hyperparameters are incredibly important.

$$p(\mu)\, p(\sigma)$$

A uniform prior on the population deviance places too much probability at high values and no pooling.

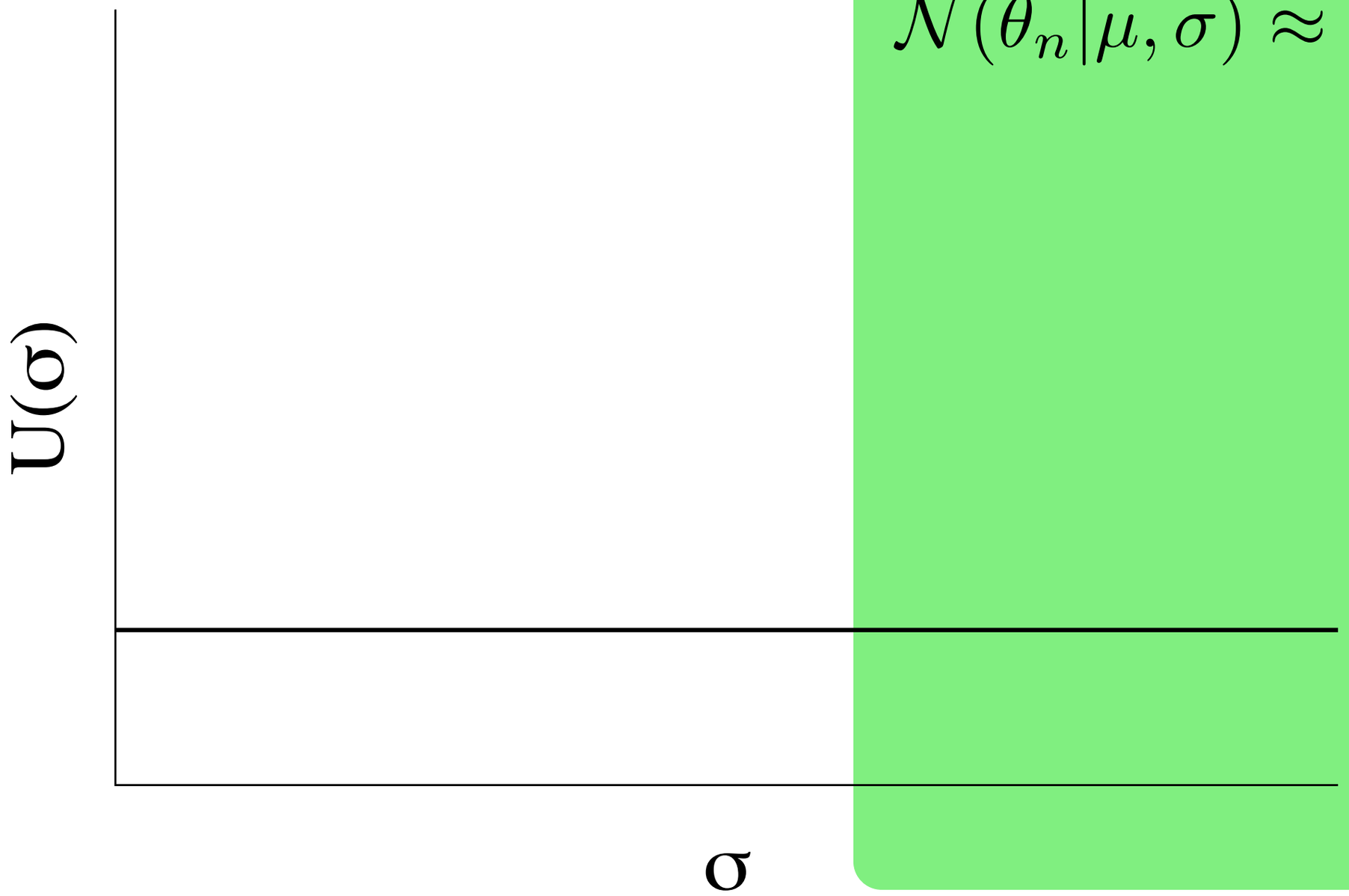


A uniform prior on the population deviance places too much probability at high values and no pooling.

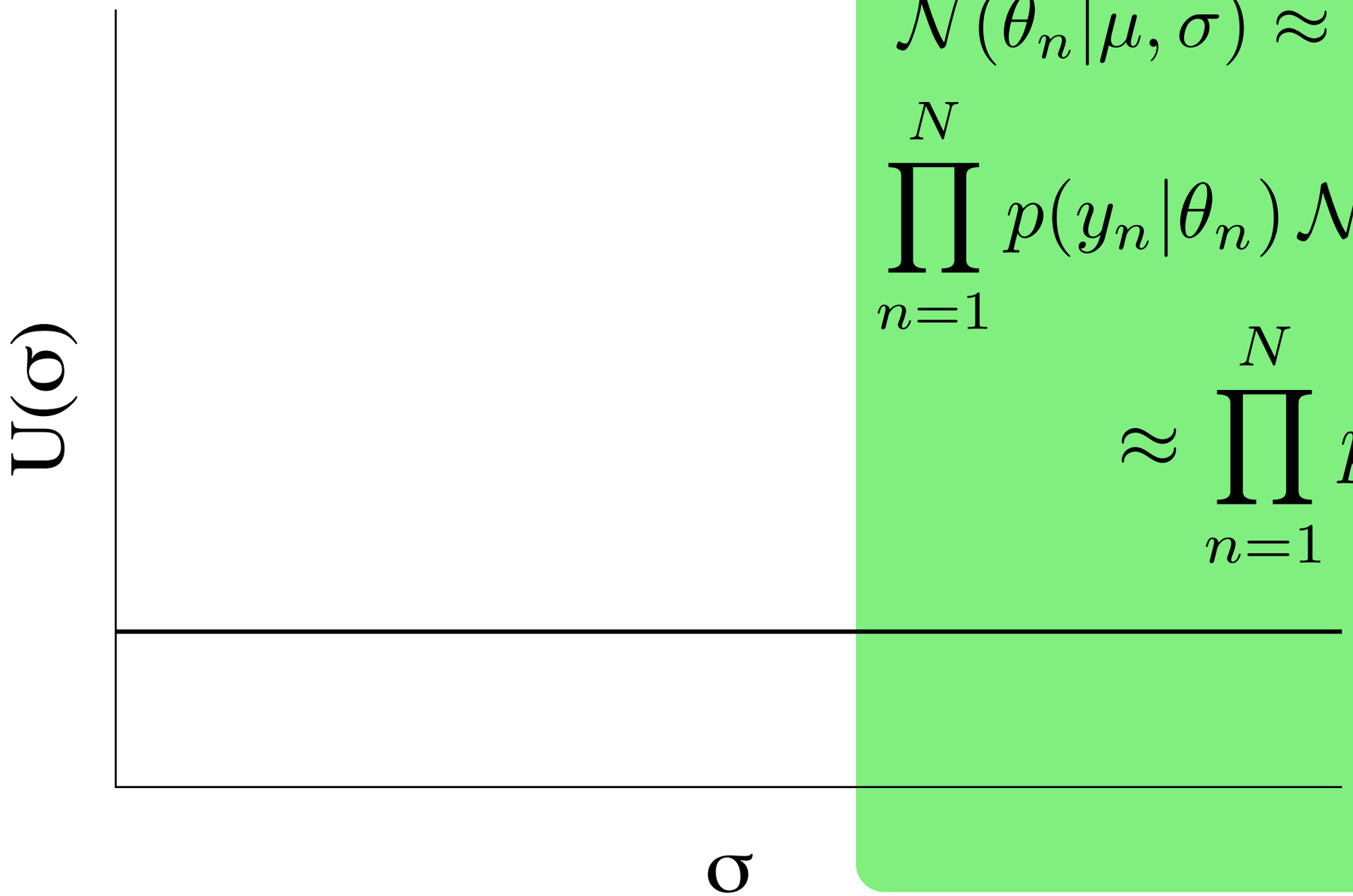


A uniform prior on the population deviance places too much probability at high values and no pooling.

$$\mathcal{N}(\theta_n|\mu,\sigma) \approx U(\theta_n)$$



A uniform prior on the population deviance places too much probability at high values and no pooling.

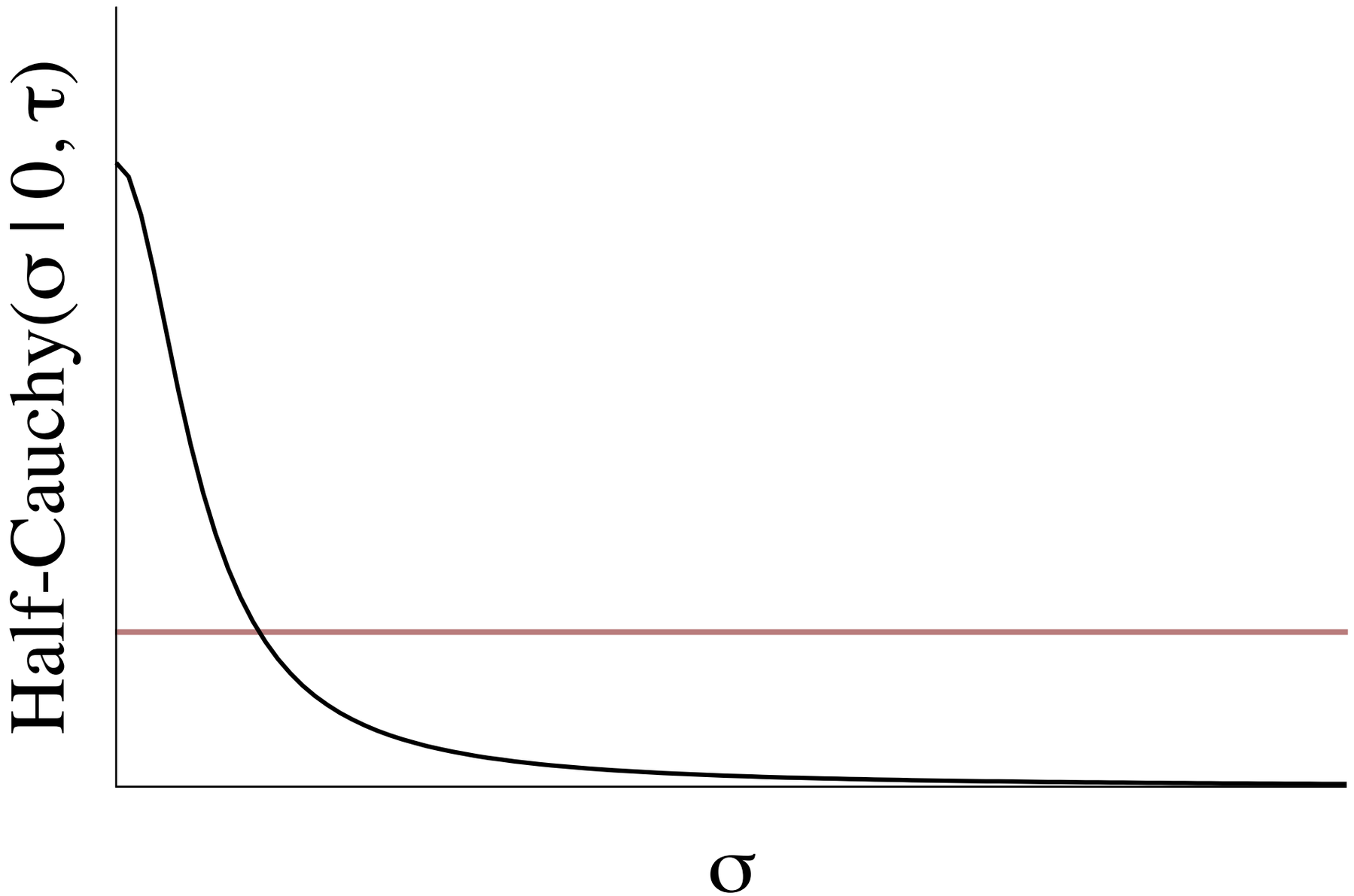


$$\mathcal{N}(\theta_n|\mu,\sigma) \approx U(\theta_n)$$

$$\prod_{n=1}^{N} p(y_n|\theta_n)\,\mathcal{N}(\theta_n|\mu,\sigma)$$

$$\approx \prod_{n=1}^{N} p(y_n|\theta_n)$$

For the model to be able to partially pool we need a weakly informative prior that concentrates around zero.



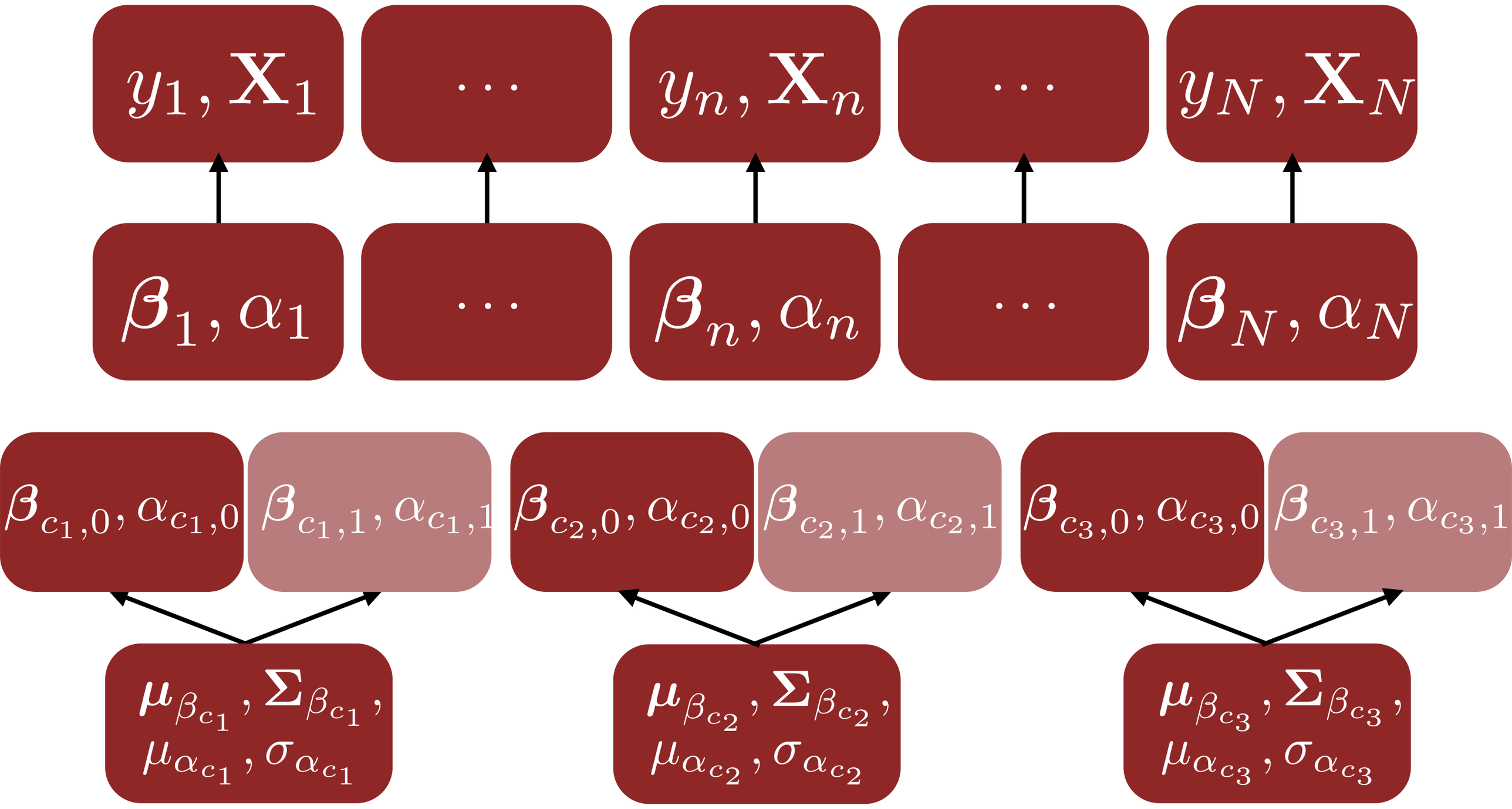
In a *multilevel* model the linear parameters vary across multiple hierarchies.

$$p(y_n|g(X_n^T\boldsymbol{\beta}_n + \alpha_n), \theta)$$

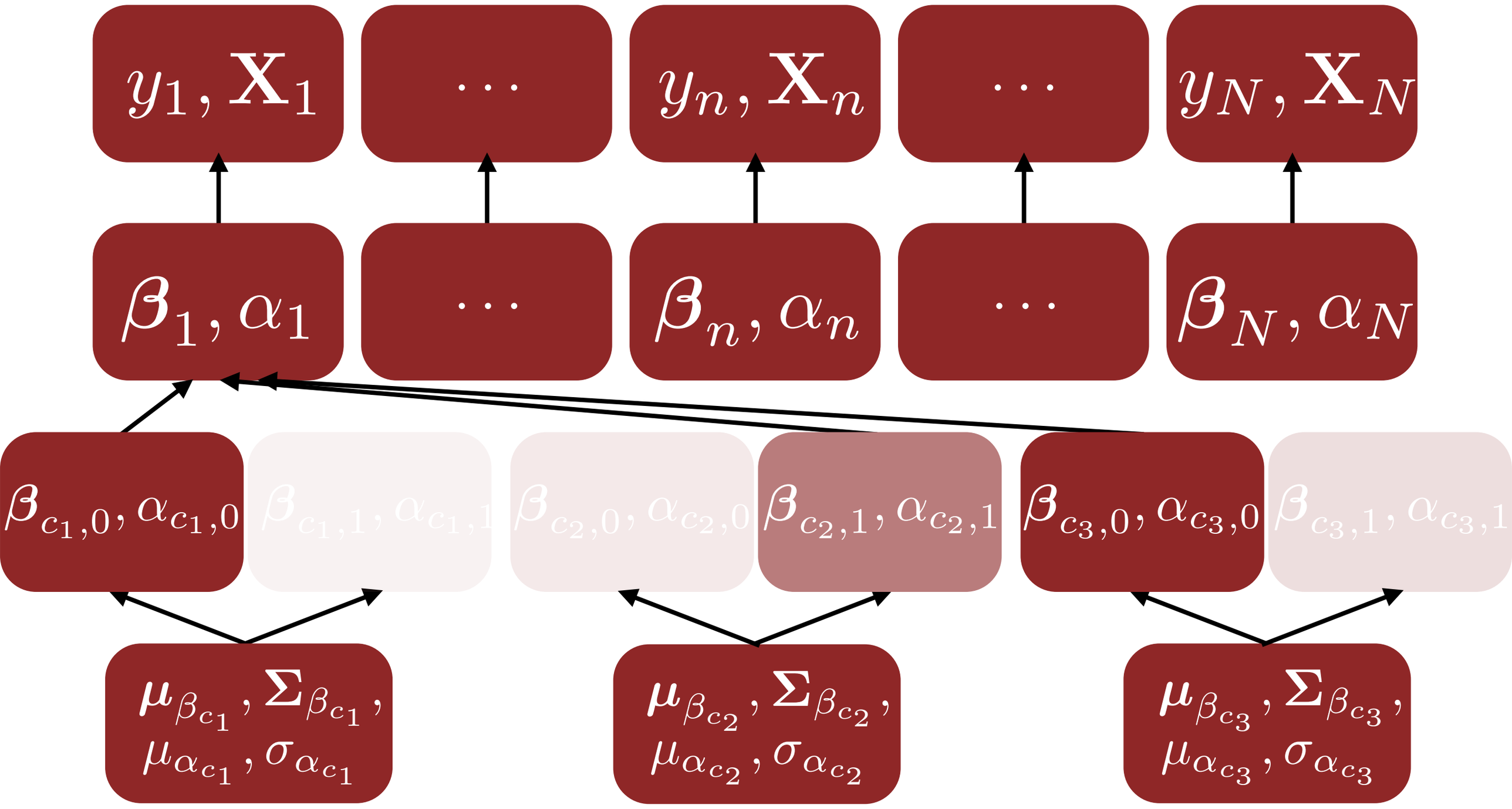
$$\alpha_n = \sum_{k=1}^{N_{\text{groups}}} \alpha_{k,j(n)}$$

$$\alpha_{k,j} \sim \mathcal{N}(\mu_{\alpha_k}, \sigma_{\alpha_k})$$

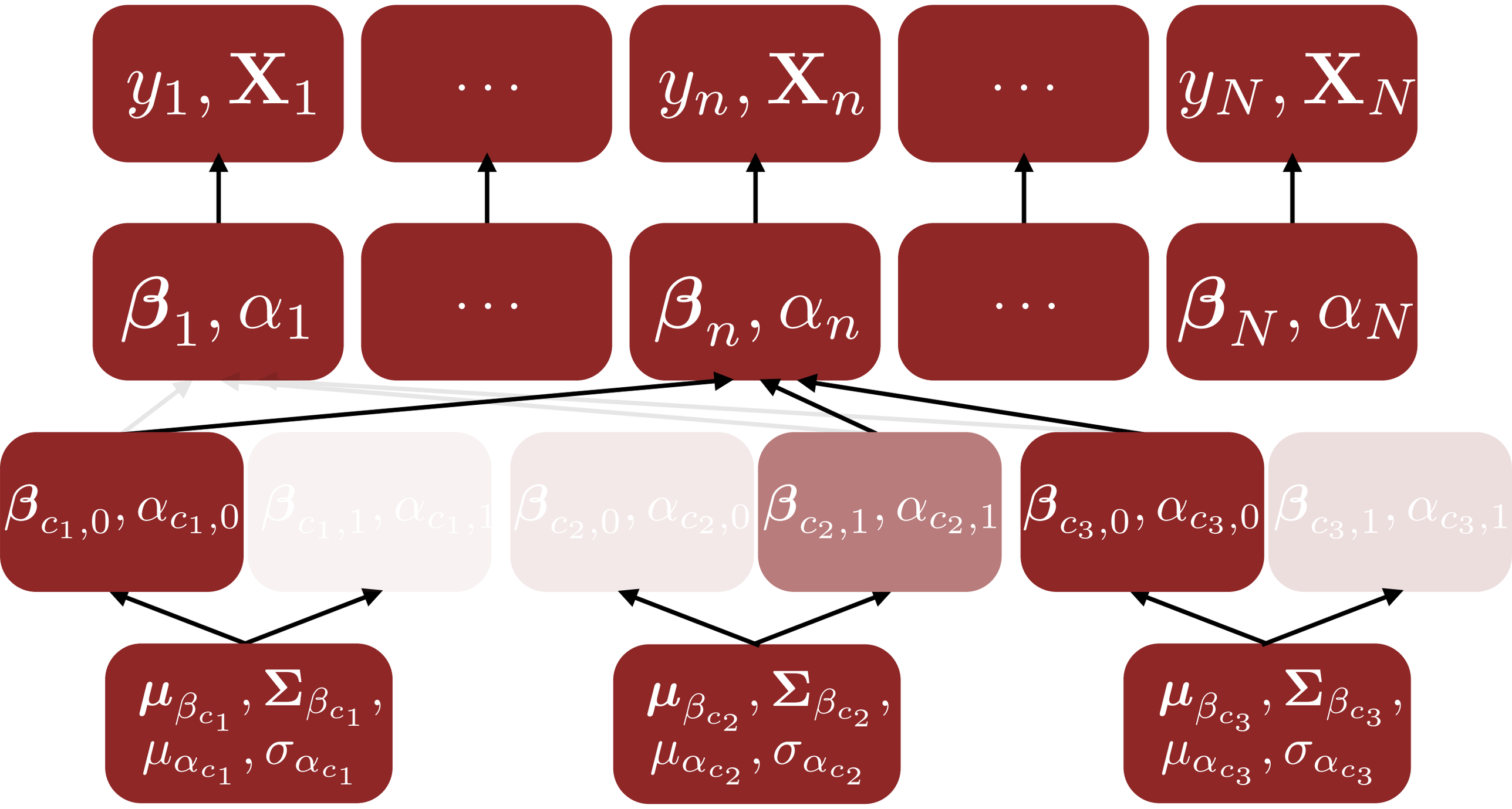
In a *multilevel* model the linear parameters vary across multiple hierarchies.



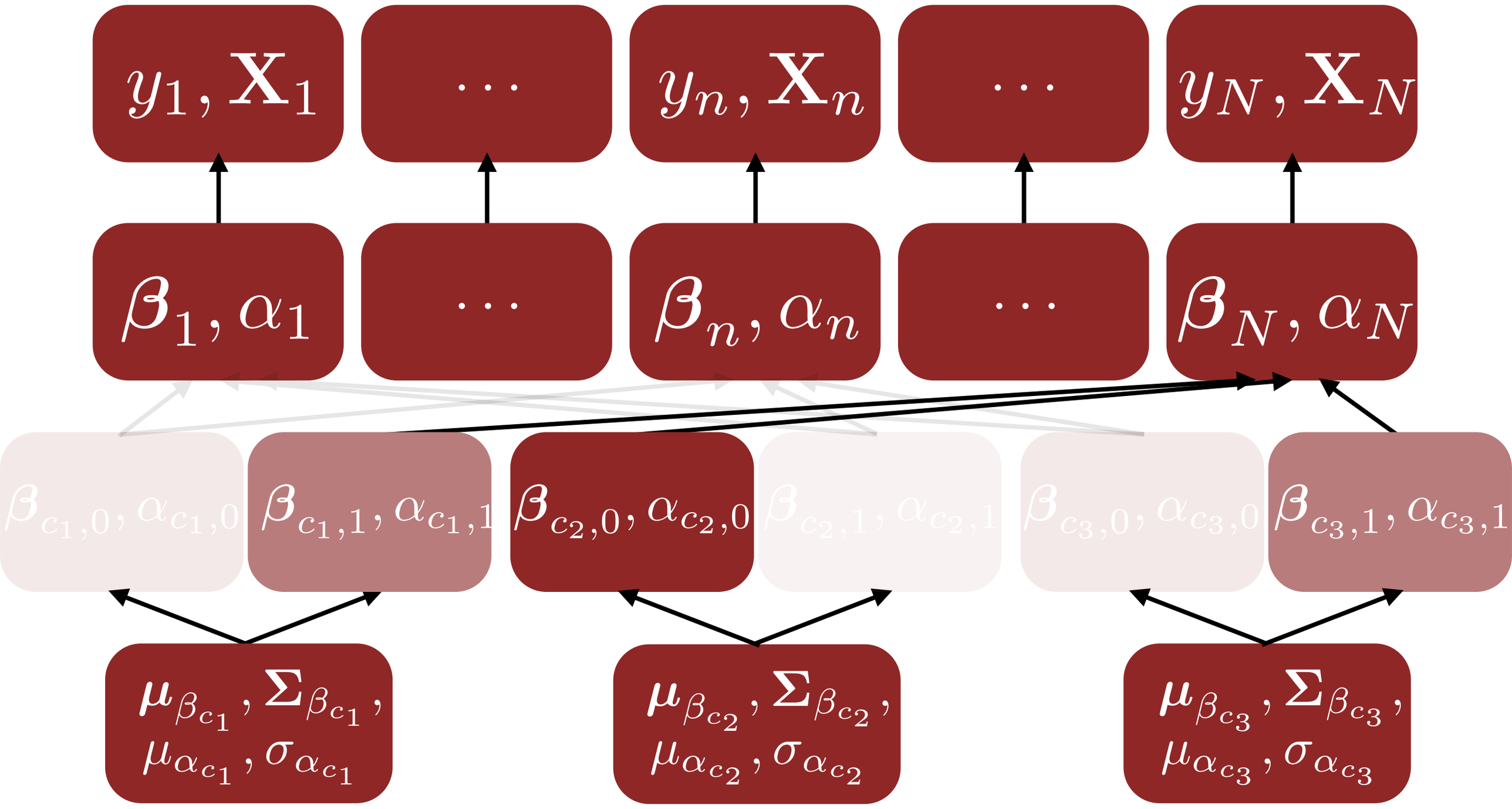
In a *multilevel* model the linear parameters vary across multiple hierarchies.



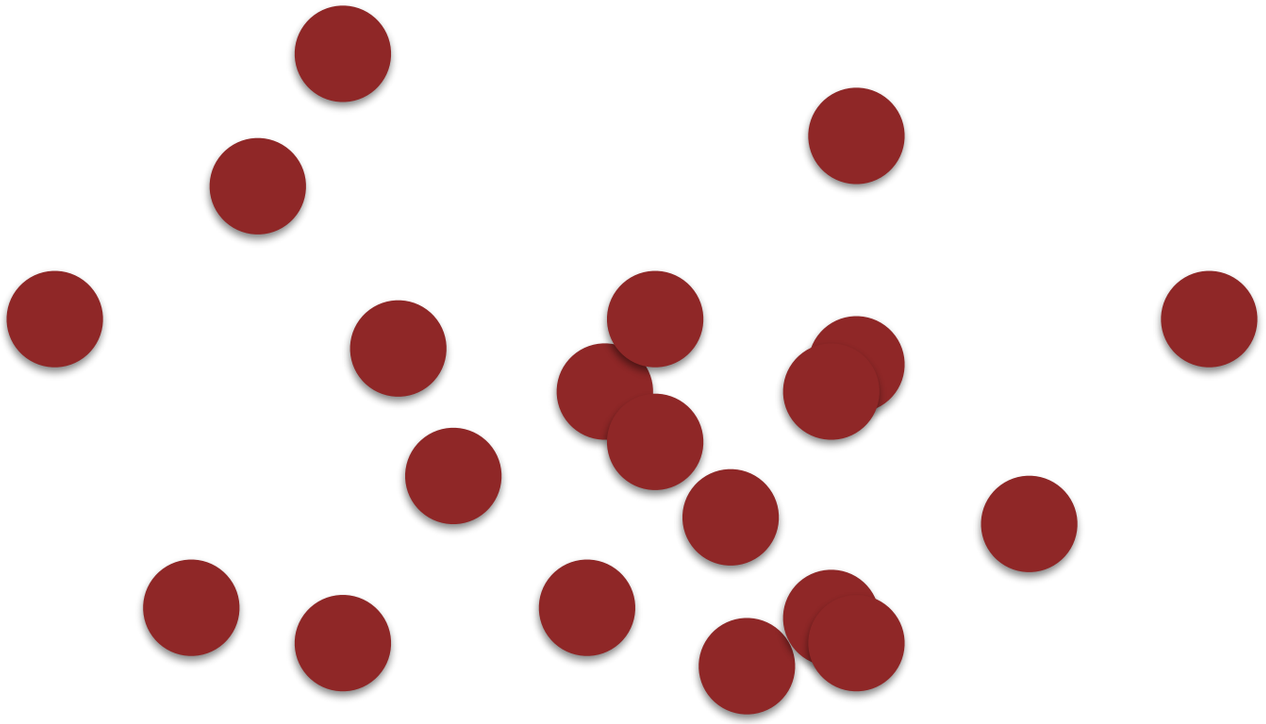
In a *multilevel* model the linear parameters vary across multiple hierarchies.



In a *multilevel* model the linear parameters vary across multiple hierarchies.

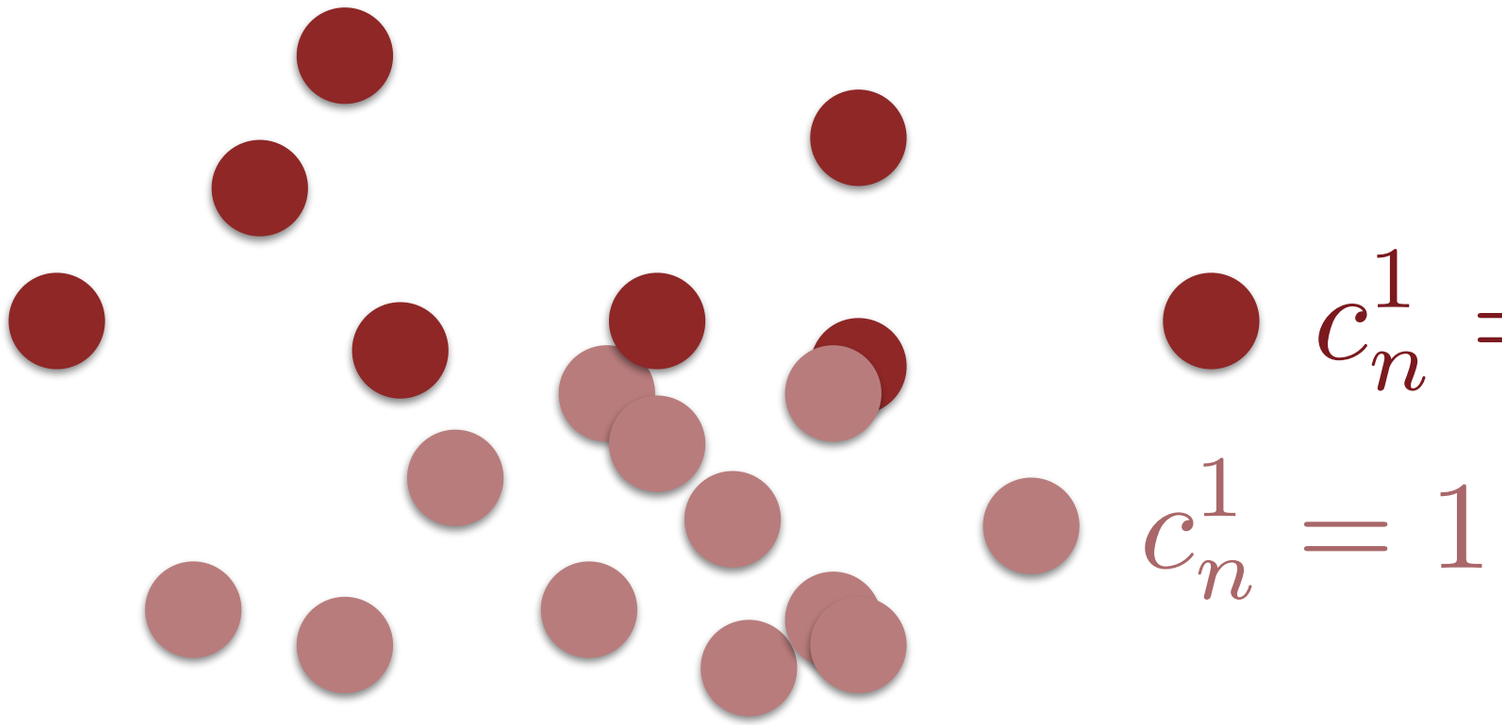


Each level contributes its own slope and/or intercept.



$$\left\{c_n^1, c_n^2, c_n^3\right\} \in \{0, 1\}$$

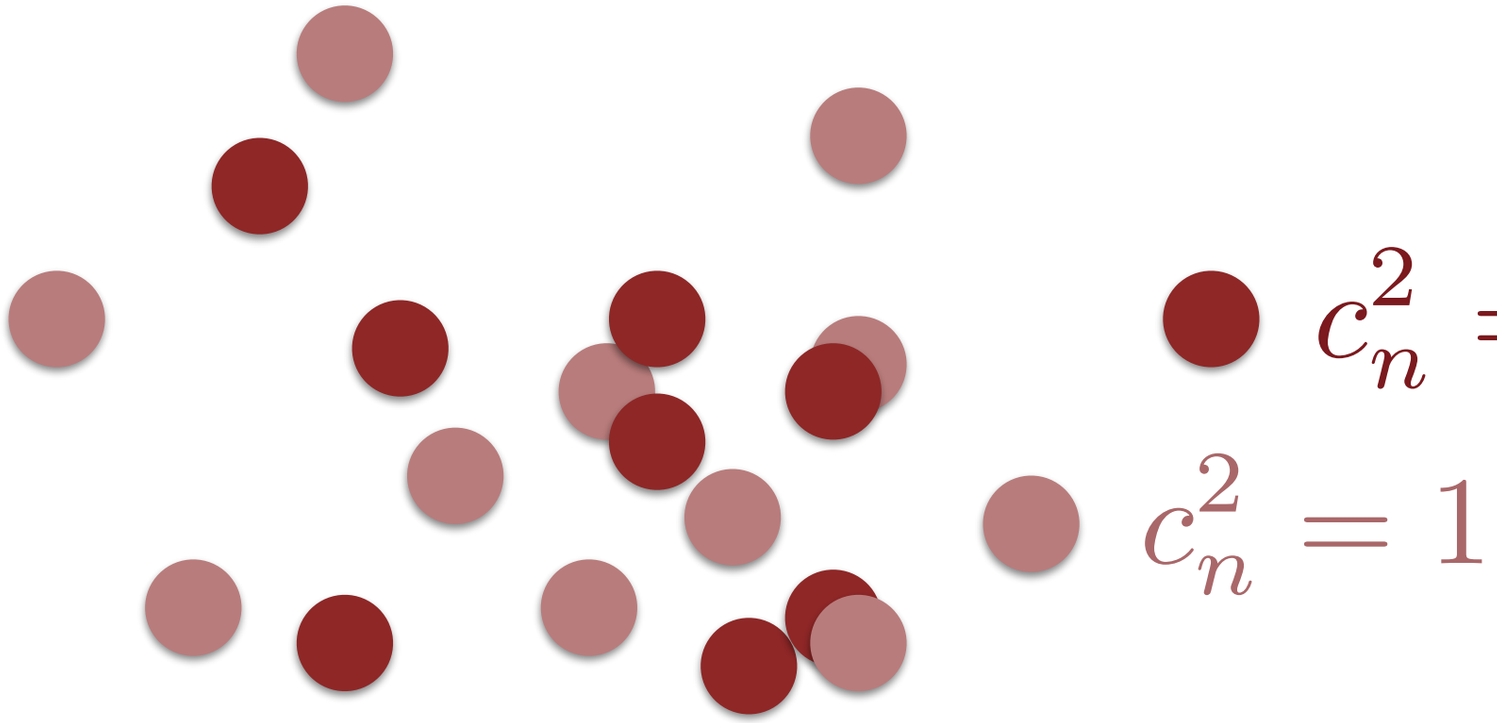
Each level contributes its own slope and/or intercept.



$$\alpha_{c_1,0} \quad \alpha_{c_1,1}$$

$$\alpha_{c_1,j} \sim \mathcal{N}\left(\mu_{\alpha_{c_1}}, \sigma_{\alpha_{c_1}}\right)$$

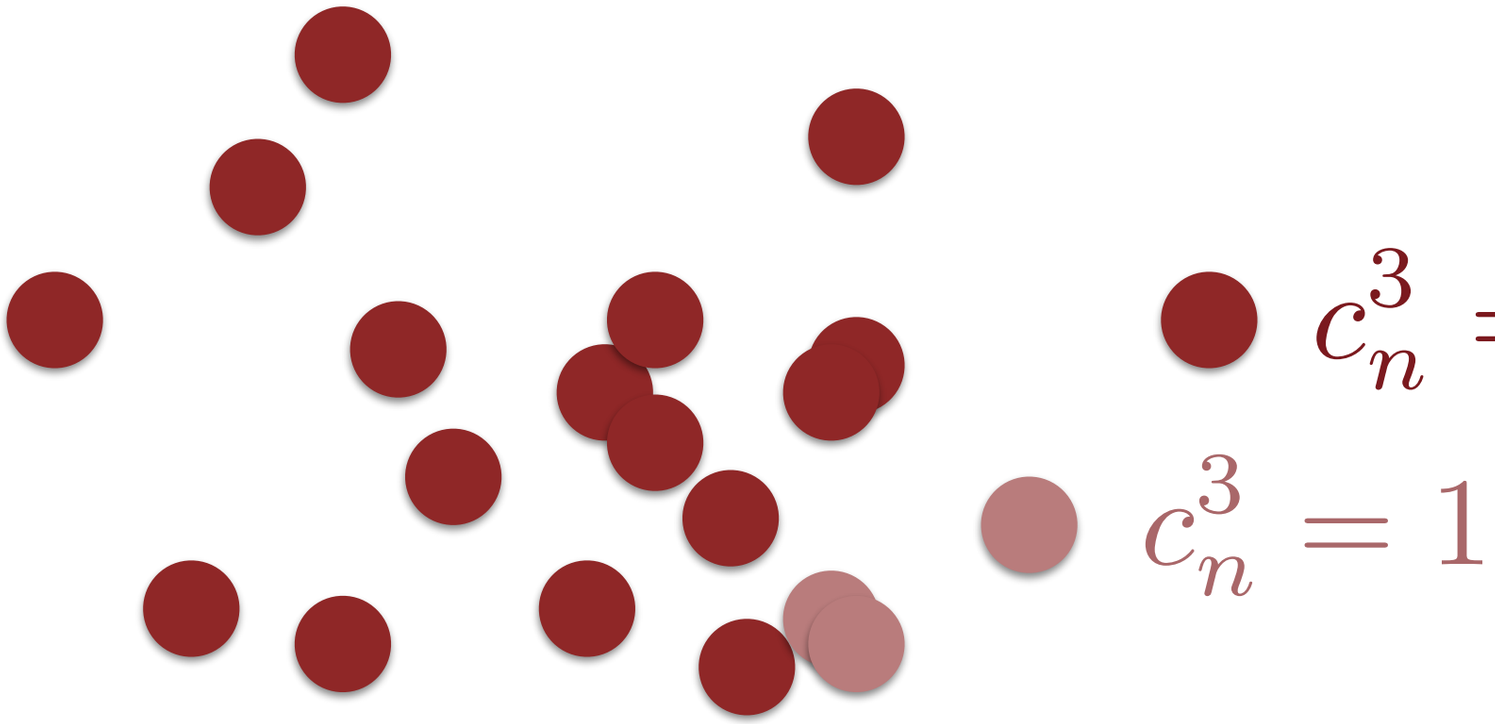
Each level contributes its own slope and/or intercept.



$$\alpha_{c_2,0} \quad \alpha_{c_2,1}$$

$$\alpha_{c_2,j} \sim \mathcal{N}\left(\mu_{\alpha_{c_2}}, \sigma_{\alpha_{c_2}}\right)$$

Each level contributes its own slope and/or intercept.



$$\alpha_{c_3,0} \quad \alpha_{c_3,1}$$

$$\alpha_{c_3,j} \sim \mathcal{N}\left(\mu_{\alpha_{c_3}}, \sigma_{\alpha_{c_3}}\right)$$

Each level contributes its own slope and/or intercept.

$c_n^1 = 0$

$c_n^2 = 1$

$c_n^3 = 0$

$\alpha_n =$

Each level contributes its own slope and/or intercept.

$c_n^1 = 0$ $c_n^2 = 1$ $c_n^3 = 0$

$\alpha_n = \quad \alpha_{c_1,0}$

Each level contributes its own slope and/or intercept.

$$c_n^1 = 0 \qquad c_n^2 = 1 \qquad c_n^3 = 0$$

$$\alpha_n = \quad \alpha_{c_1,0} \quad + \quad \alpha_{c_2,1}$$

Each level contributes its own slope and/or intercept.

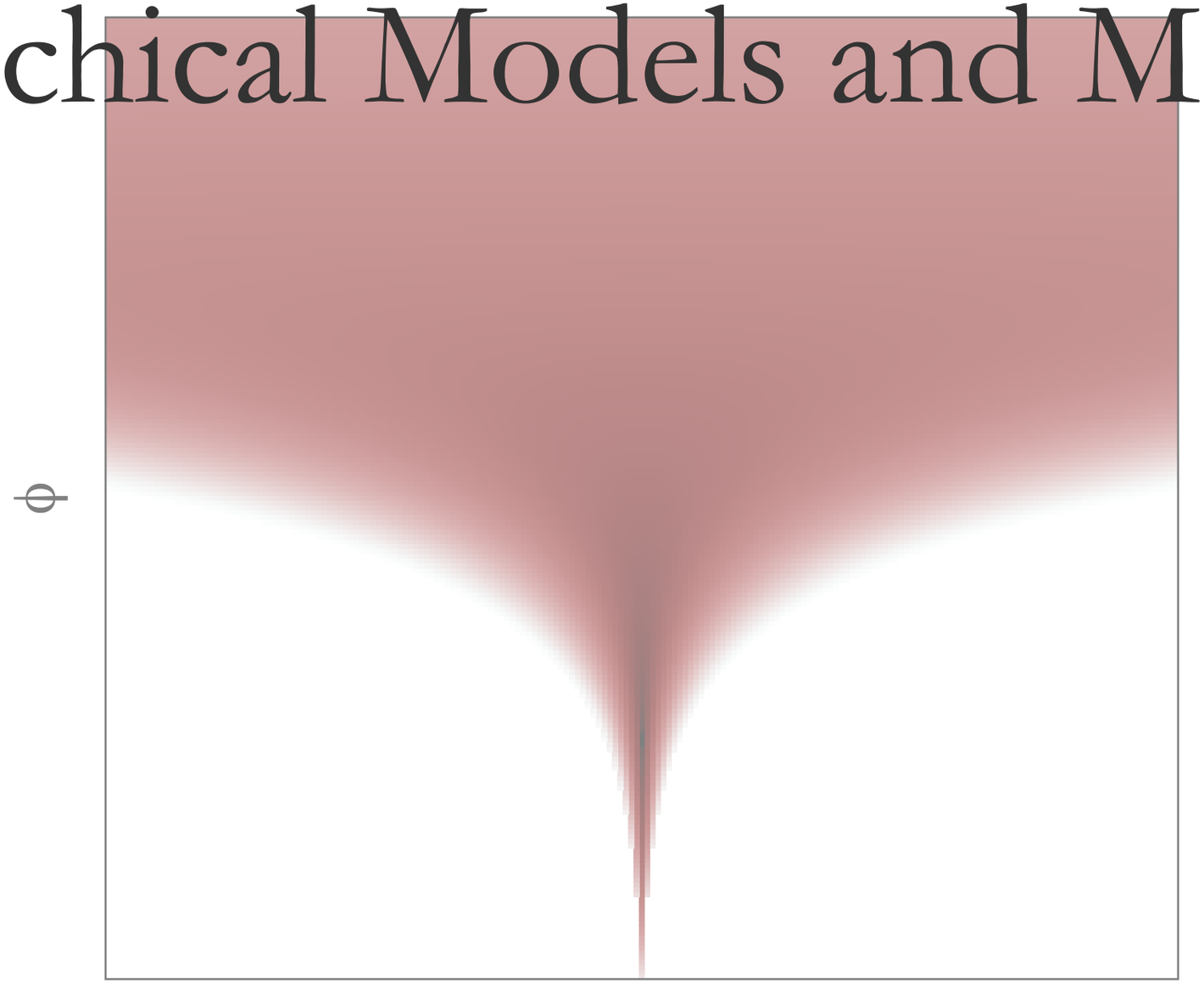
$$c_n^1 = 0 \qquad c_n^2 = 1 \qquad c_n^3 = 0$$

$$\alpha_n = \quad \alpha_{c_1,0} \quad + \quad \alpha_{c_2,1} \quad + \quad \alpha_{c_3,0}$$

Partial pooling model

```
data {
  int N;
  int K;
  int<lower=1> N_groups;
  int<lower=1> group[N];
  int<lower=0, upper=1> y[N];
  vector[N] x;
}

parameters {
  vector[N_groups] alpha; // Intercept for every group
  real beta; // could also let beta vary by group
  real mu_alpha; // Hierarchical mean
  real<lower=0> sigma_alpha; // Hierarchical sd
}

model {
  y ~ bernoulli_logit(x * beta + alpha[group]);

  beta ~ normal(0, 10);
  alpha ~ normal(mu_alpha, sigma_alpha);
  mu_alpha ~ normal(0, 10);
  sigma_alpha ~ cauchy(0, 10);
}
```

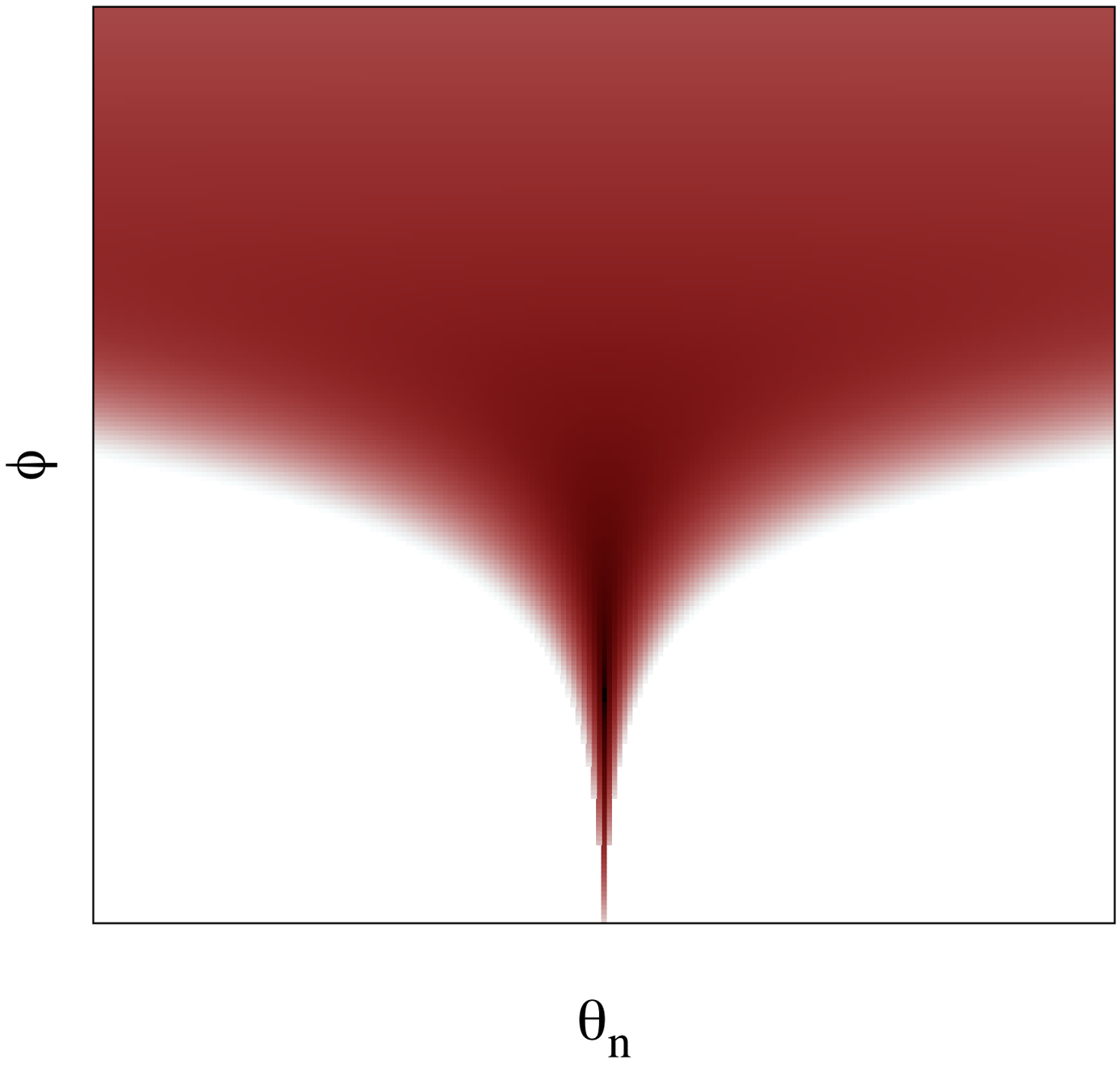
Generalized Linear Hierarchical Models and MCMC



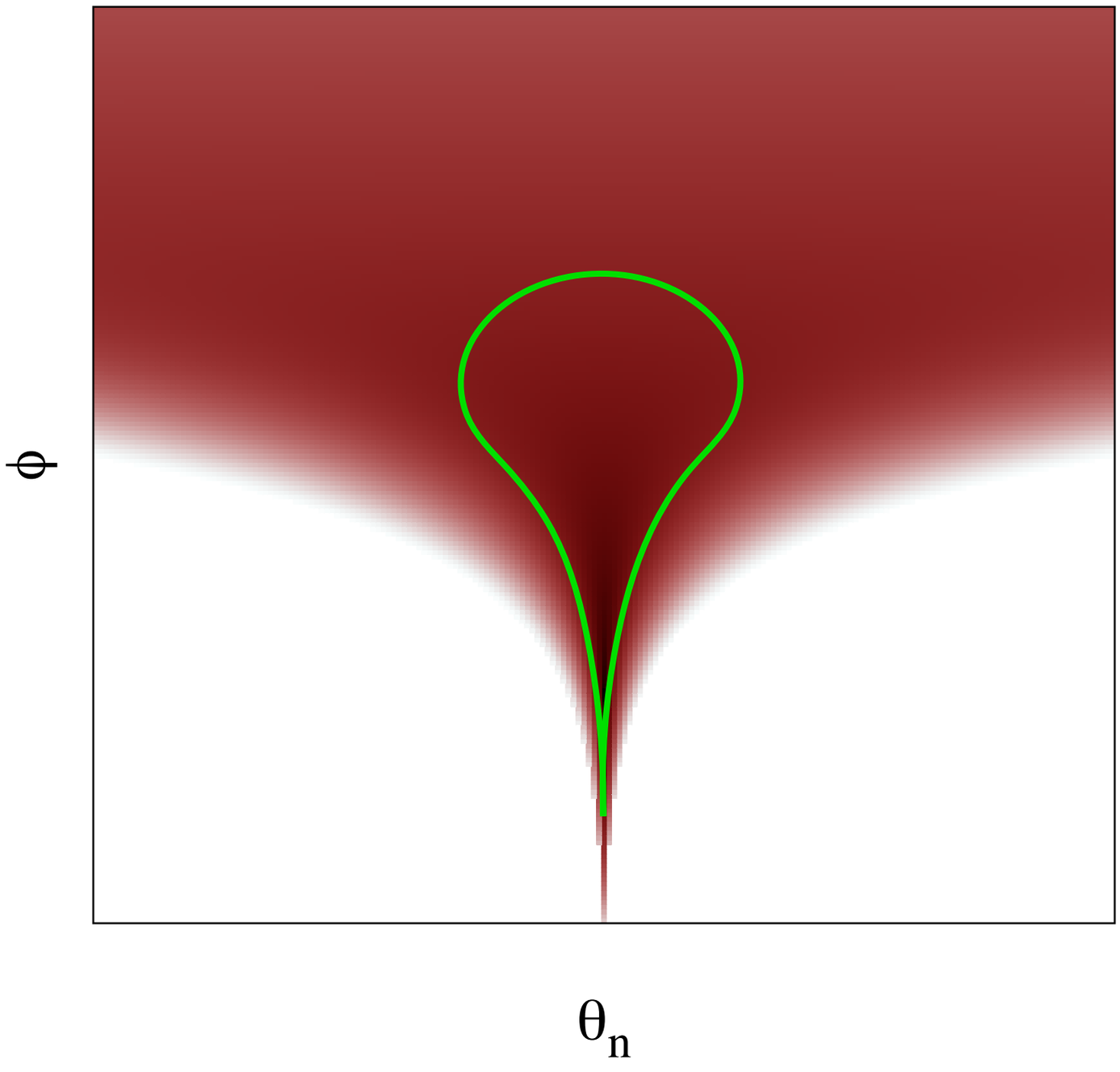
Neal's funnel distribution illustrates some of the problems with sampling from hierarchical distributions.

$$p(\theta|\phi) = \prod_{n=1}^{N} \mathcal{N}\left(\theta_n|0, e^{\frac{\phi}{2}}\right) \mathcal{N}(\phi|0,3)$$

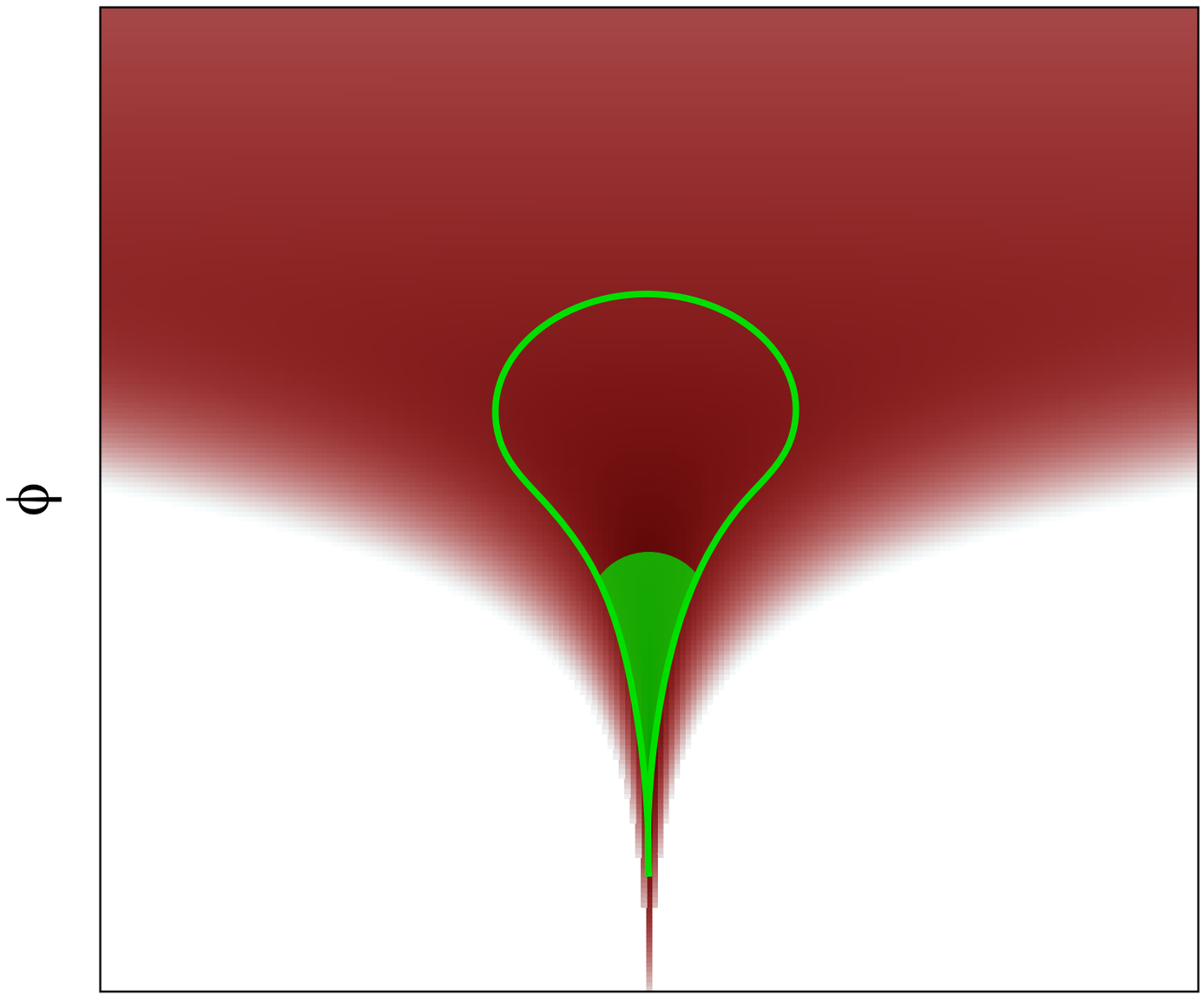
Neal's funnel distribution illustrates some of the problems with sampling from hierarchical priors.



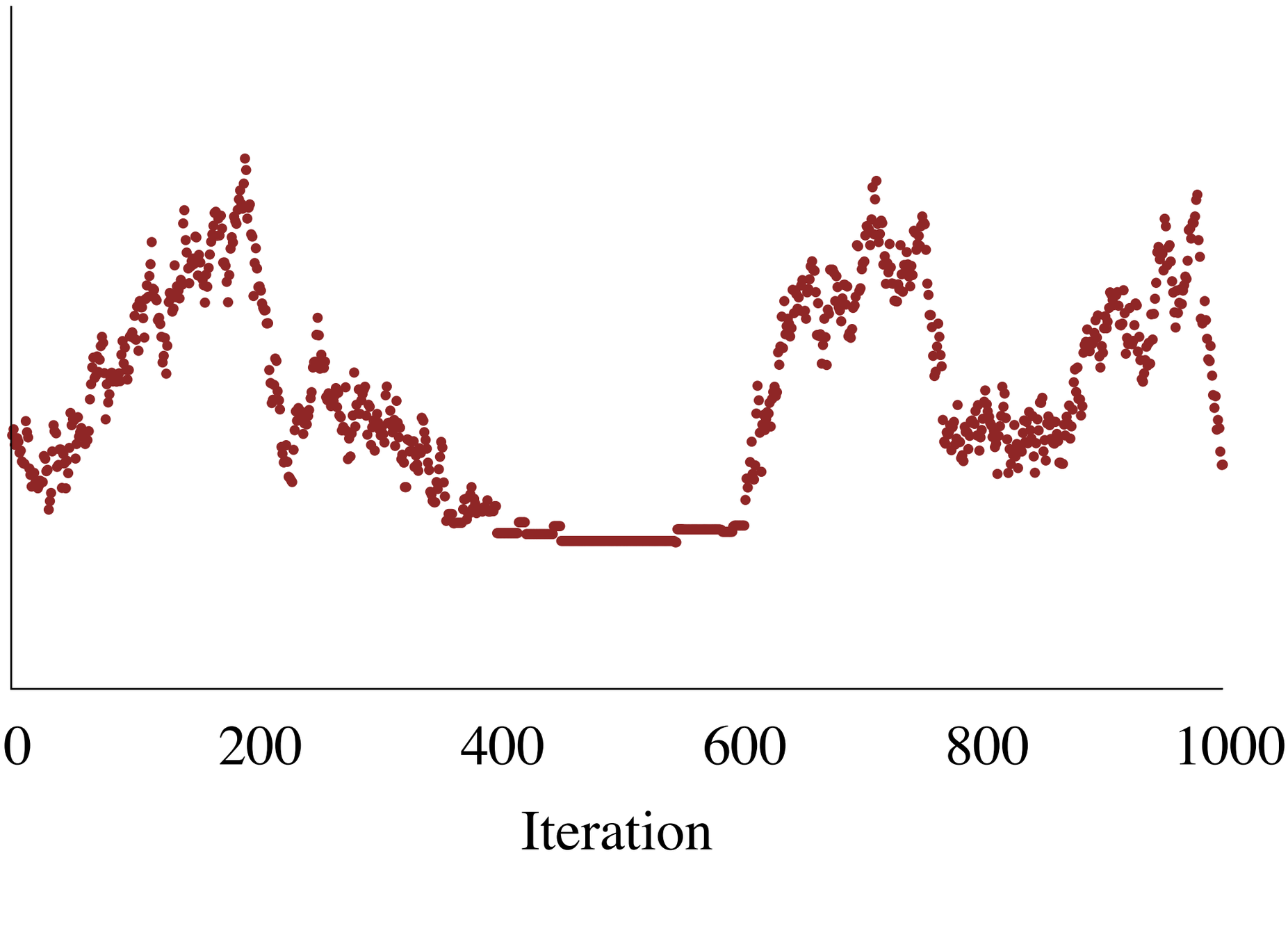
Neal's funnel distribution illustrates some of the problems with sampling from hierarchical priors.



Neal's funnel distribution illustrates some of the problems with sampling from hierarchical priors.



Neal's funnel distribution illustrates some of the problems with sampling from hierarchical priors.



The geometry of the funnel, however, dramatically improves when we sample from auxiliary parameters.

$$p(\theta|\phi) = \prod_{n=1}^{N} \mathcal{N}\left(\theta_n|0, e^{\frac{\phi}{2}}\right) \mathcal{N}(\phi|0, 3)$$

The geometry of the funnel, however, dramatically improves when we sample from auxiliary parameters.

$$p(\theta|\phi) = \prod_{n=1}^{N} \mathcal{N}\left(\theta_n|0, e^{\frac{\phi}{2}}\right) \mathcal{N}(\phi|0, 3)$$

$$\tilde{\theta}_n \sim \mathcal{N}(0, 1)$$

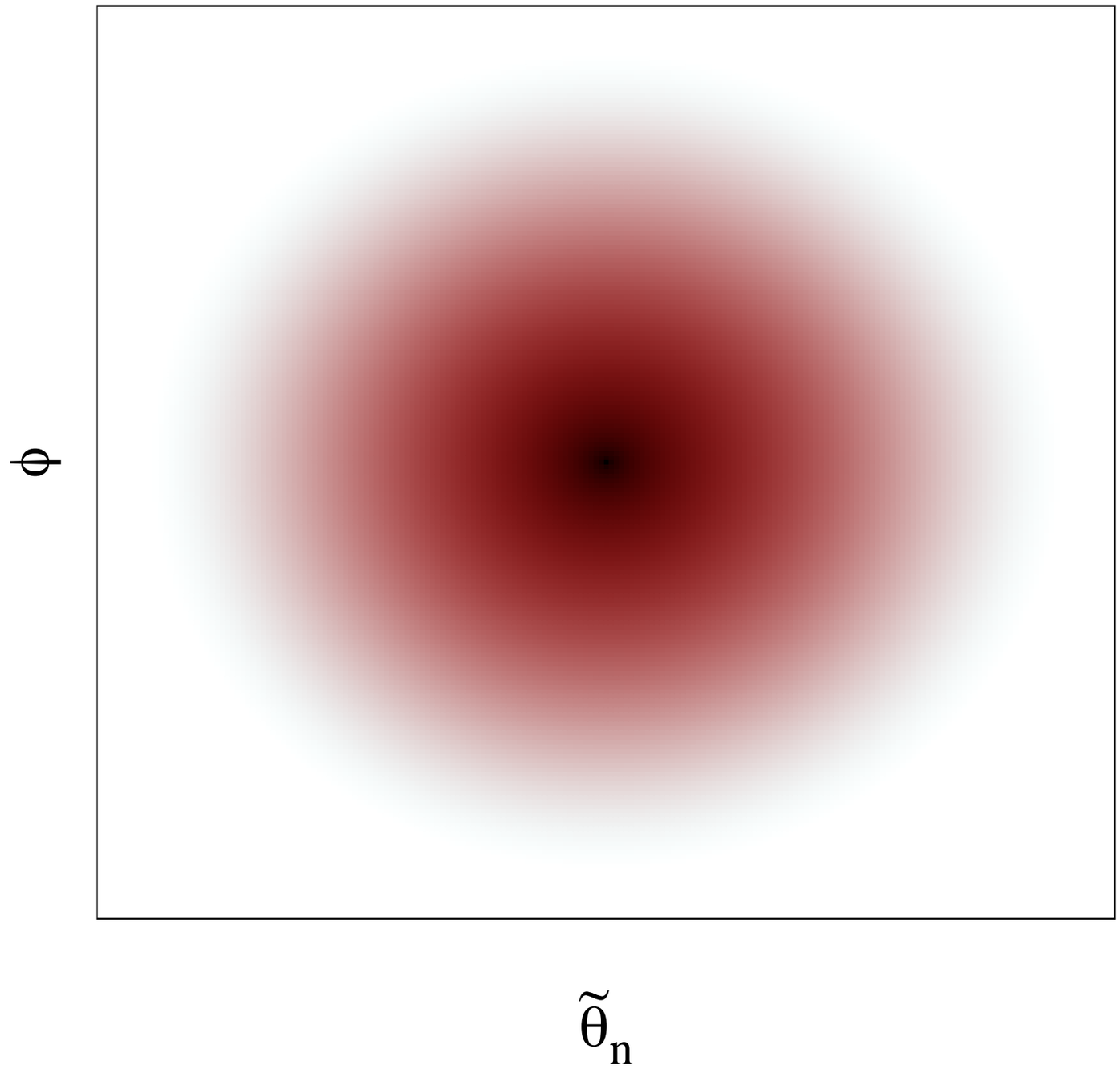
The geometry of the funnel, however, dramatically improves when we sample from auxiliary parameters.

$$p(\theta|\phi) = \prod_{n=1}^{N} \mathcal{N}\left(\theta_n|0, e^{\frac{\phi}{2}}\right) \mathcal{N}(\phi|0,3)$$

$$\tilde{\theta}_n \sim \mathcal{N}(0,1)$$

$$\theta_n = 0 + \tilde{\theta}_n e^{\frac{\phi}{2}}$$

The geometry of the funnel, however, dramatically improves when we sample from auxiliary parameters.



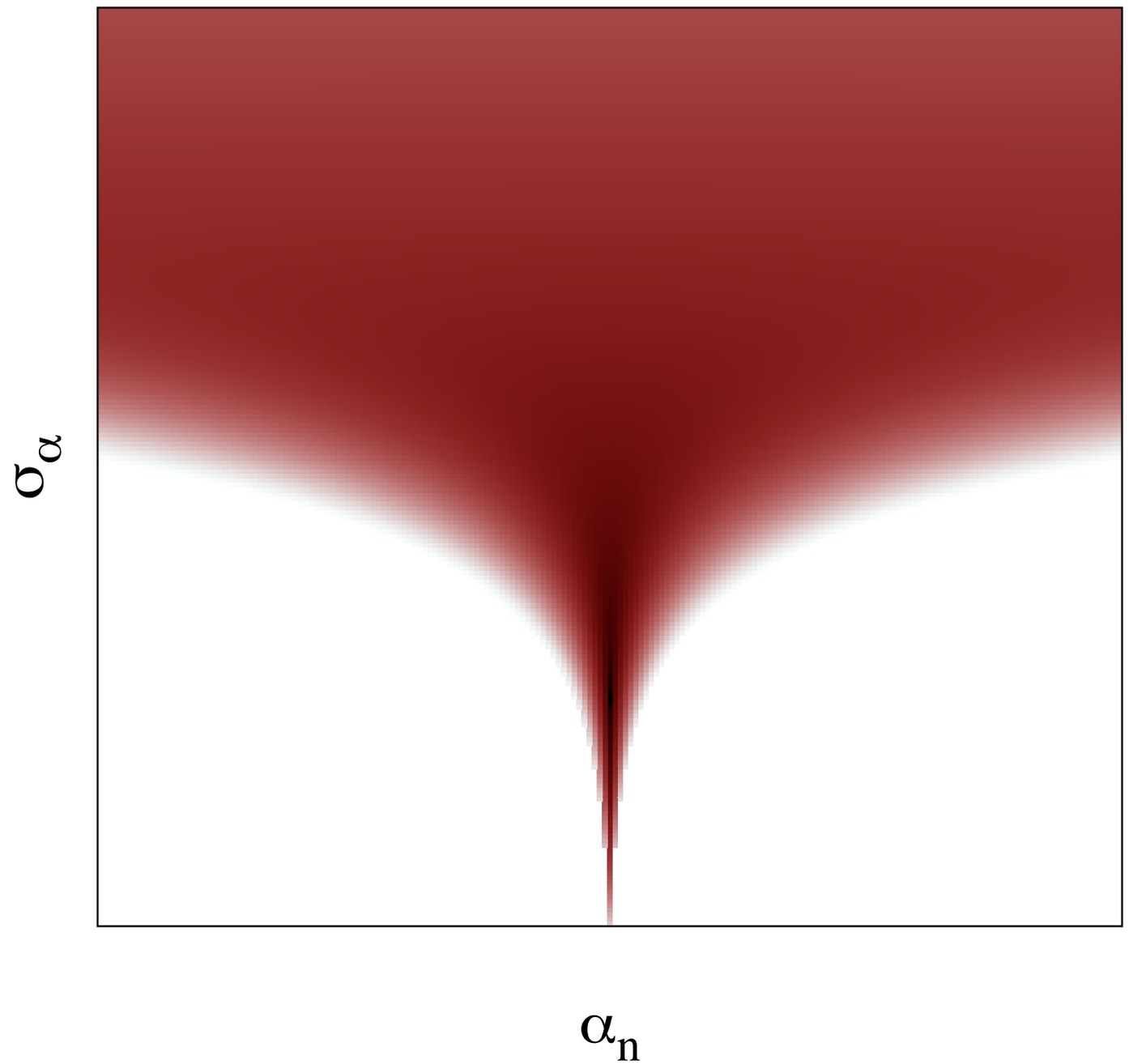
The most natural representation of a general linear hierarchical model is a *centered parameterization*.

$$p(y_n|g(\mathbf{X}_n^T\boldsymbol{\beta}_n + \alpha_n, \theta)$$

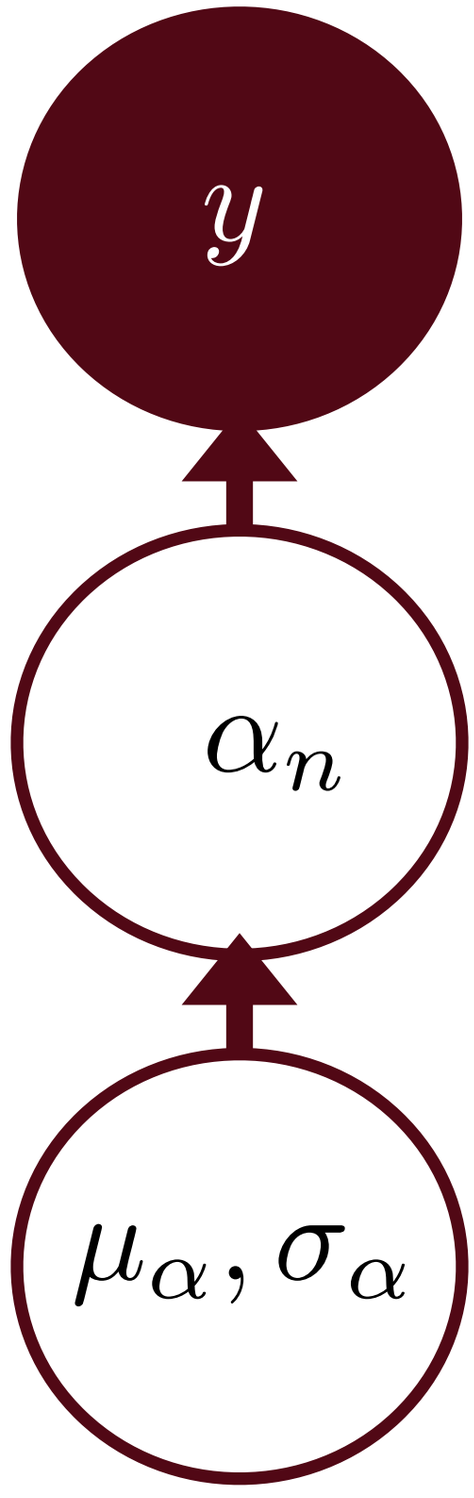
$$\boldsymbol{\beta}_n \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$$

$$\alpha_n \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha)$$

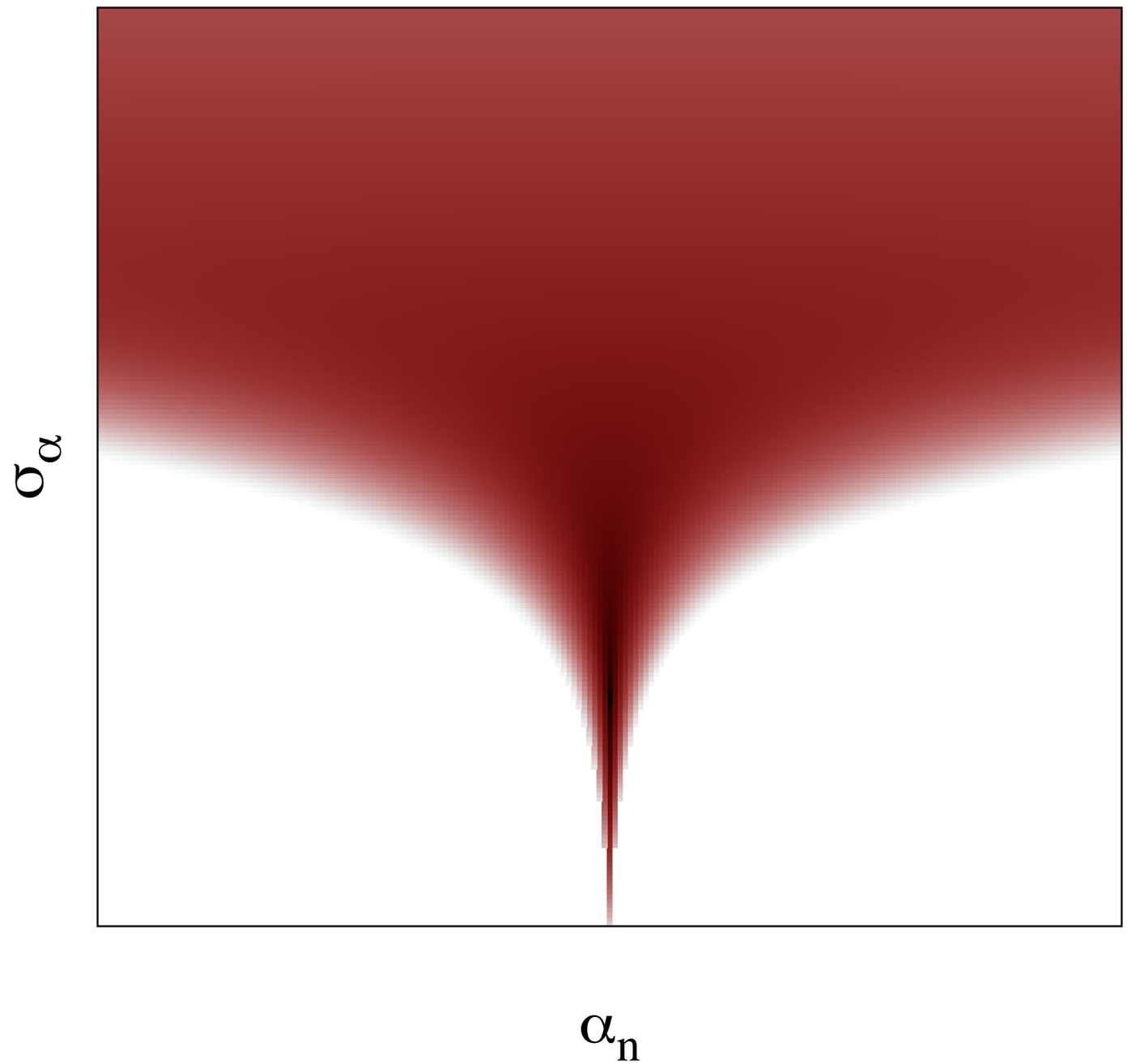
When data are sparse and uninformative, the centered parameterization of the posterior resembles the funnel.



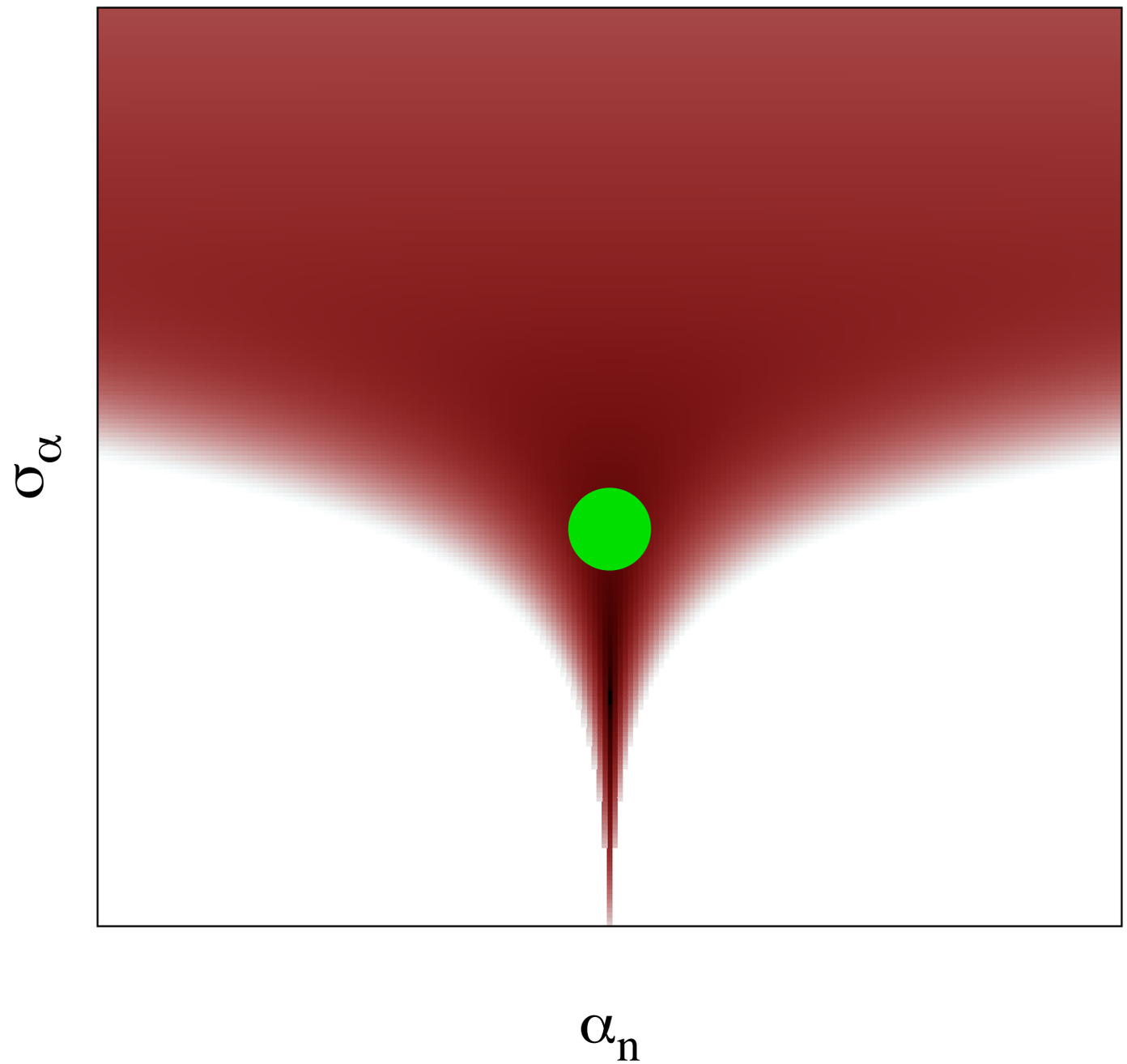
When data are sparse and uninformative, the centered parameterization of the posterior resembles the funnel.



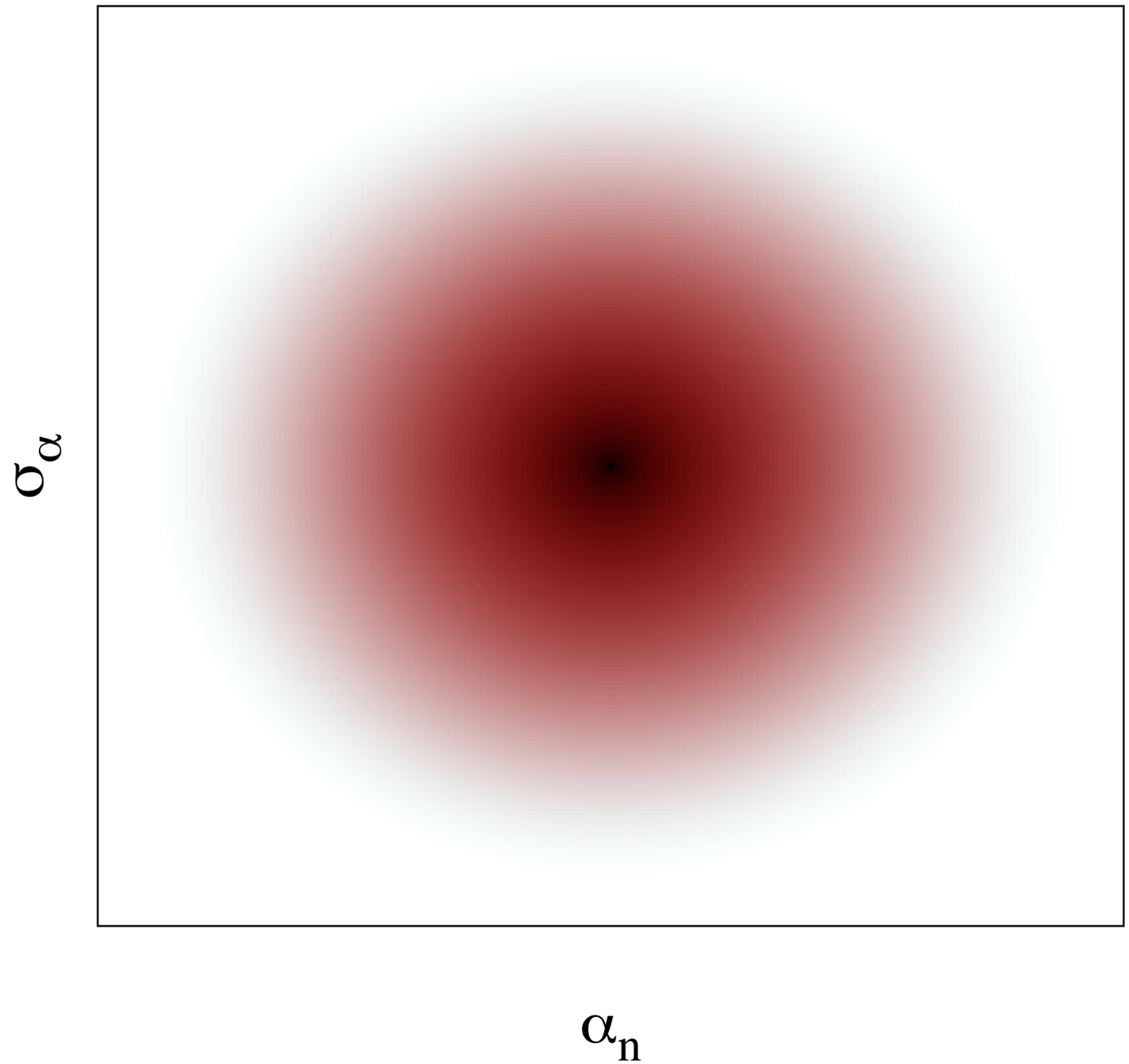
But when the data are more informative
the posterior geometry is much nicer.



But when the data are more informative
the posterior geometry is much nicer.



But when the data are more informative
the posterior geometry is much nicer.



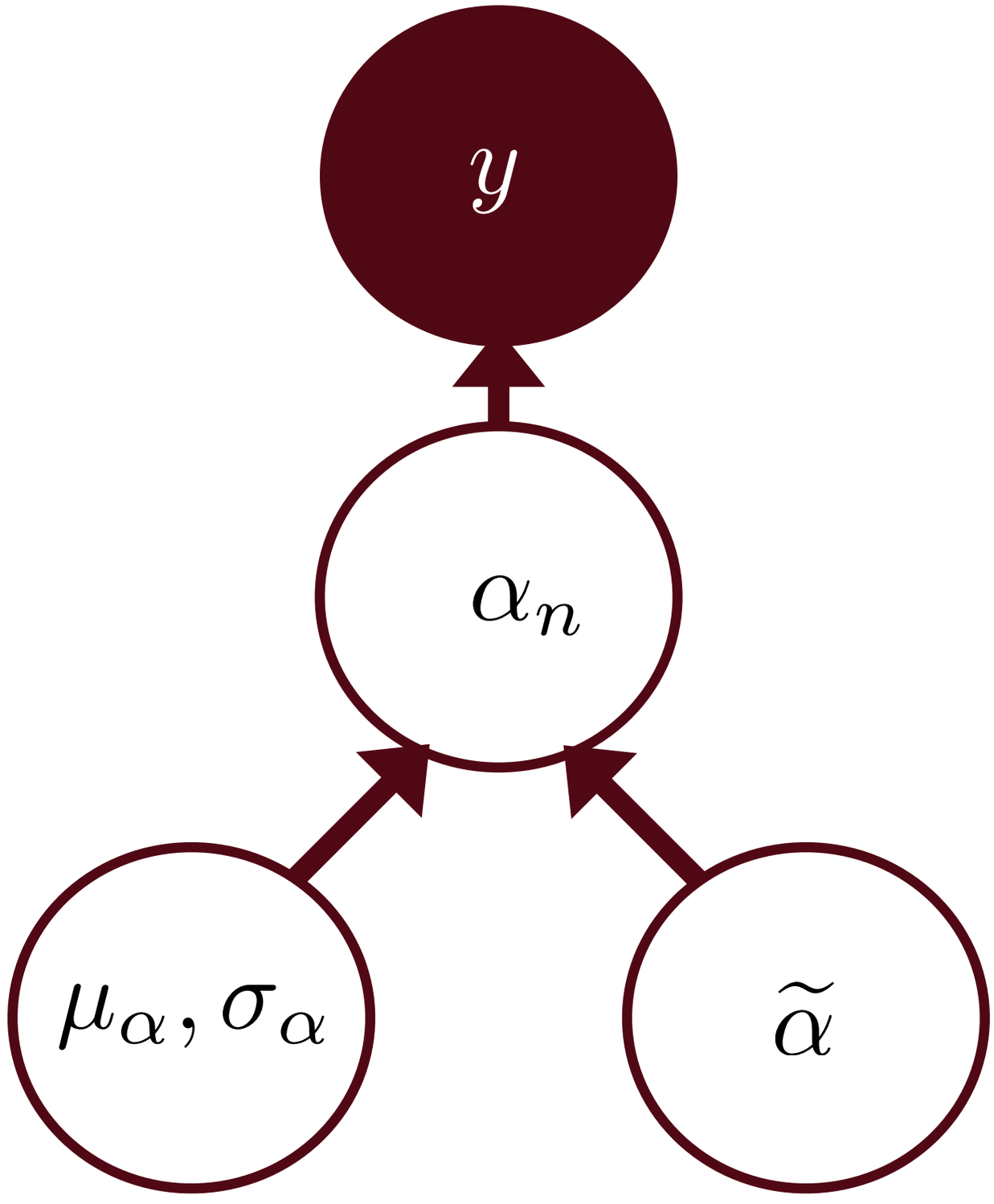
The same general linear hierarchical model can also be implemented with a *non-centered parameterization*.

$$p(y_n|g(\mathbf{X}_n^T\boldsymbol{\beta}_n + \alpha_n, \theta)$$

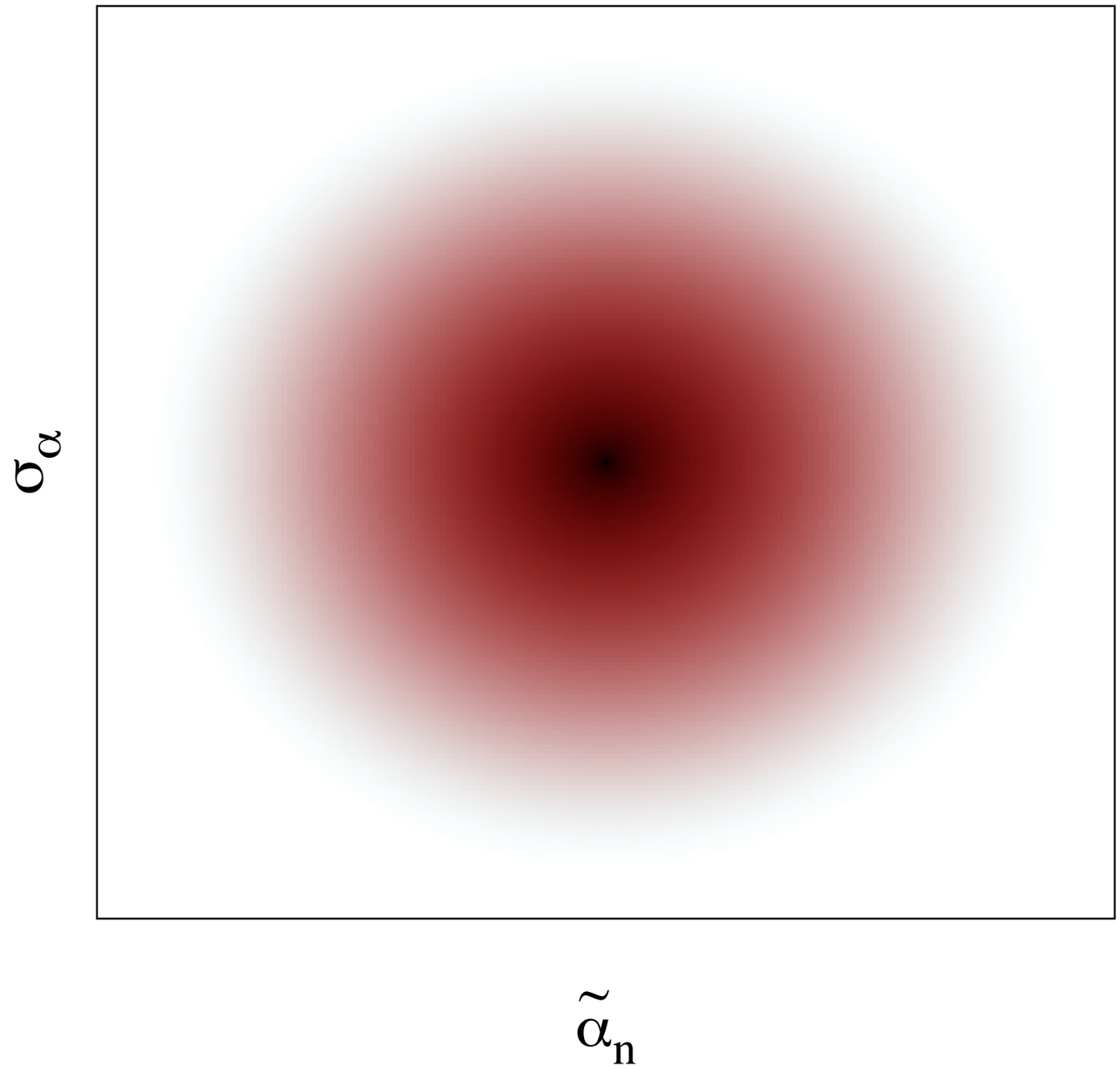
$$\tilde{\boldsymbol{\beta}}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\boldsymbol{\beta}_n = \boldsymbol{\mu}_\beta + \mathbf{L}\tilde{\boldsymbol{\beta}}_n,\ \mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}_\beta$$

$$\tilde{\alpha}_n \sim \mathcal{N}(0, 1)$$
$$\alpha_n = \mu_\alpha + \sigma_\alpha\tilde{\alpha}_n$$

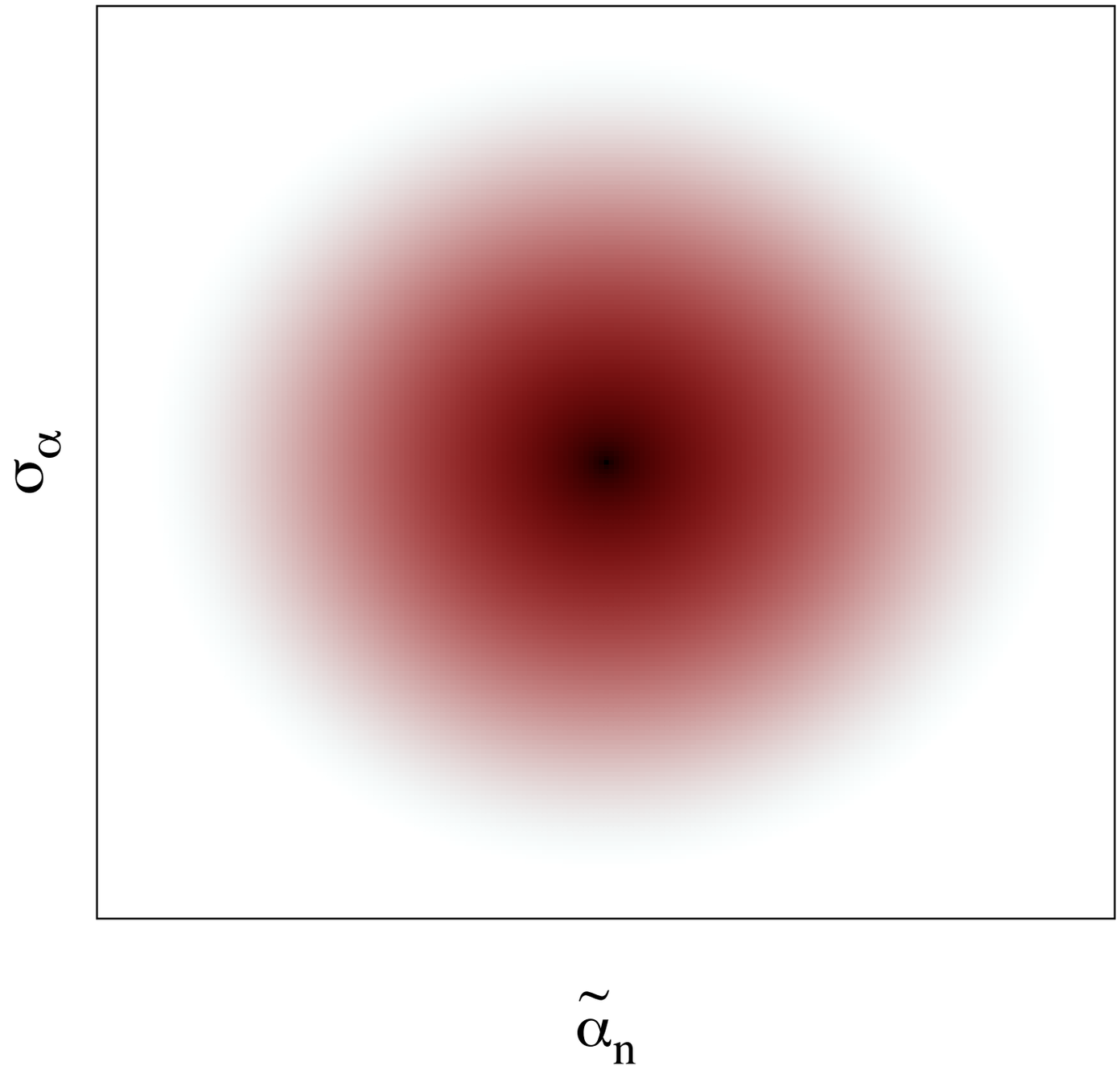
When data are sparse, the non-centered parameterization of the posterior yields a favorable geometry.



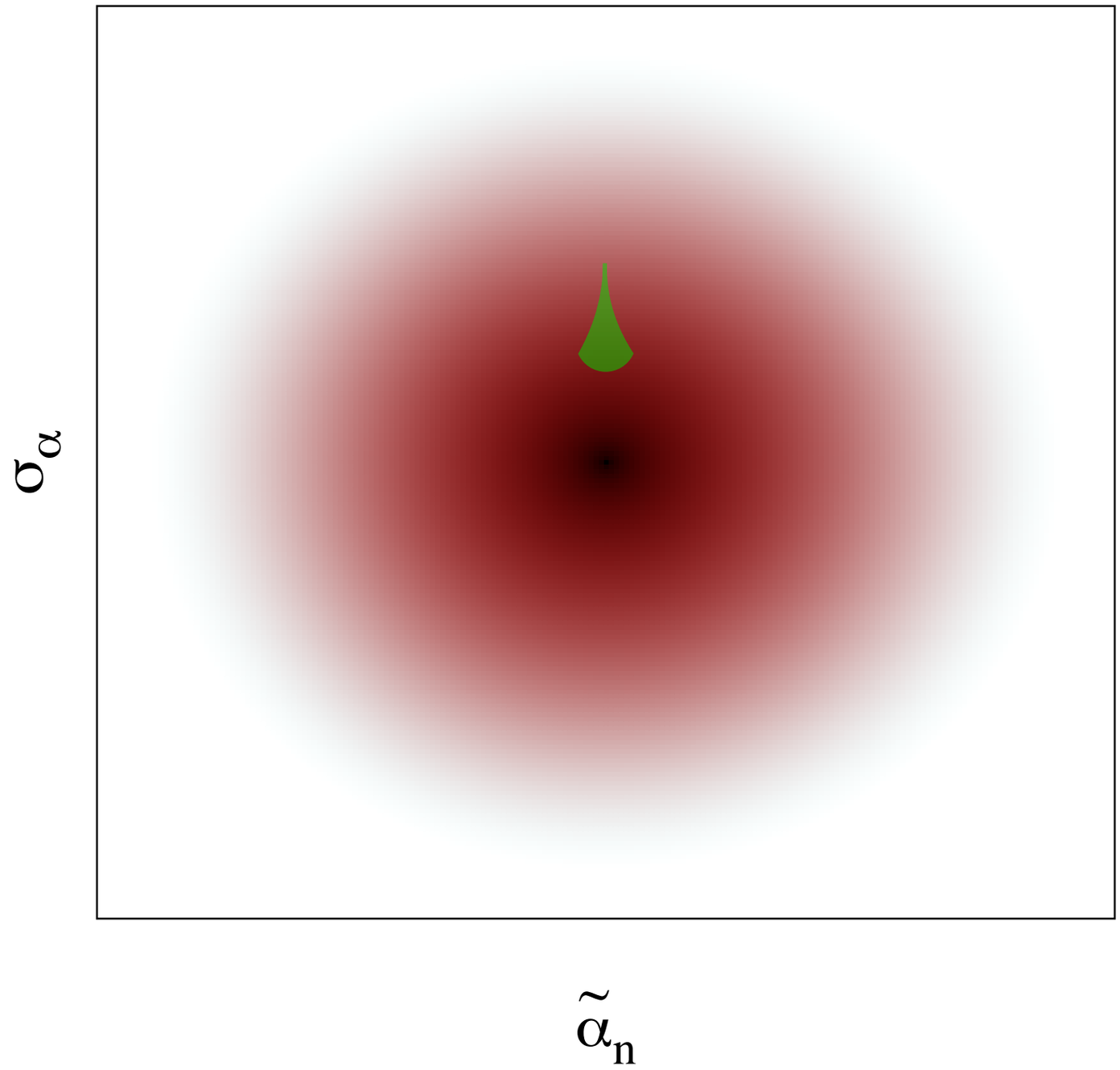
When data are sparse, the non-centered parameterization of the posterior yields a favorable geometry.



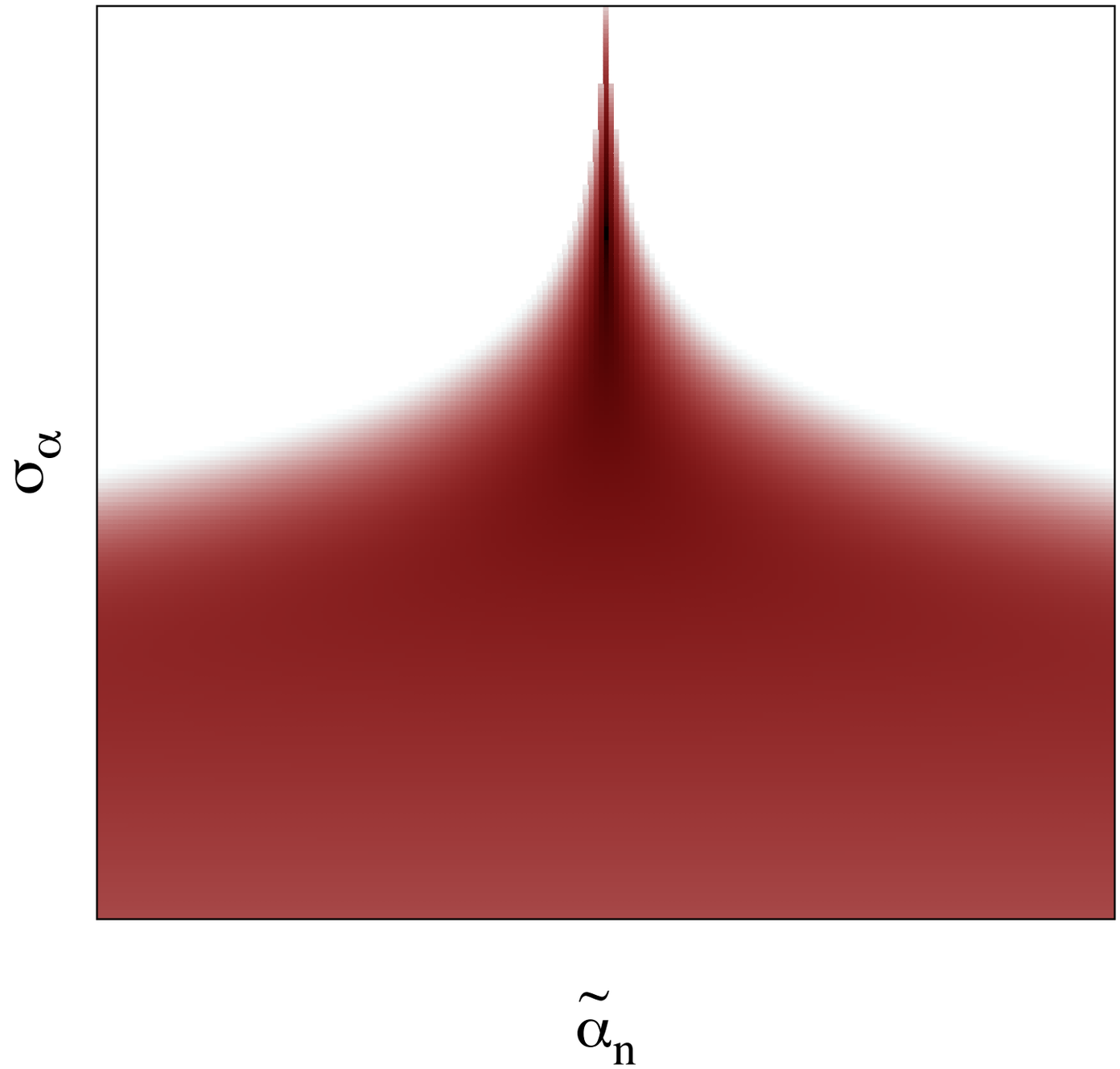
The constraints imposed by the data, however, are more complex, yielding pathological posteriors with more data.



The constraints imposed by the data, however, are more complex, yielding pathological posteriors with more data.



The constraints imposed by the data, however, are more complex, yielding pathological posteriors with more data.



Consequently the two parameterizations are each advantageous in different circumstances.

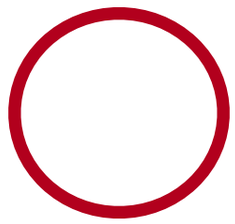
Centered Parameterization

Consequently the two parameterizations are each advantageous in different circumstances.

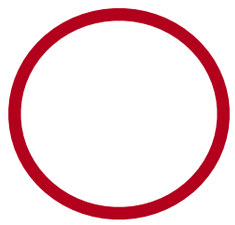
Hierarchical
Geometry

Centered
Parameterization

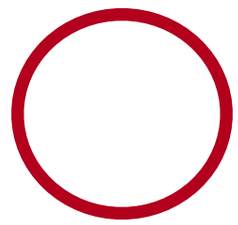
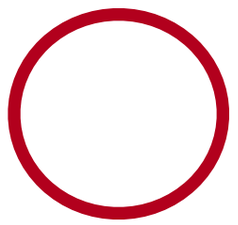
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry
Centered Parameterization		

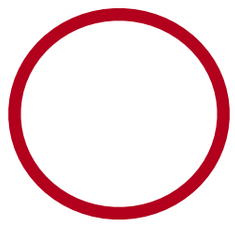
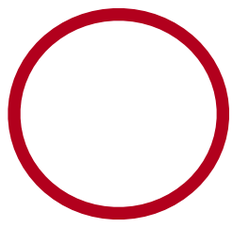
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data
Centered Parameterization			

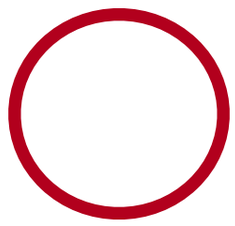
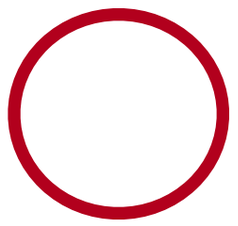
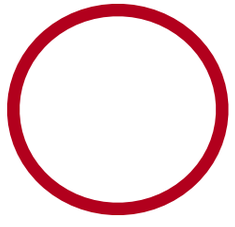
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				

Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

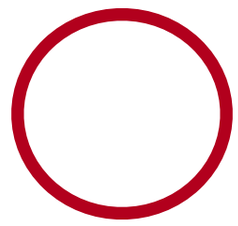
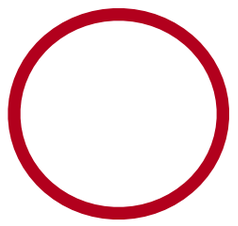
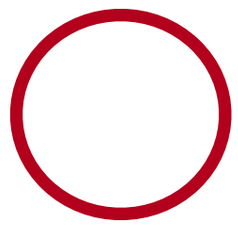
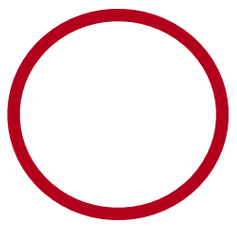
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

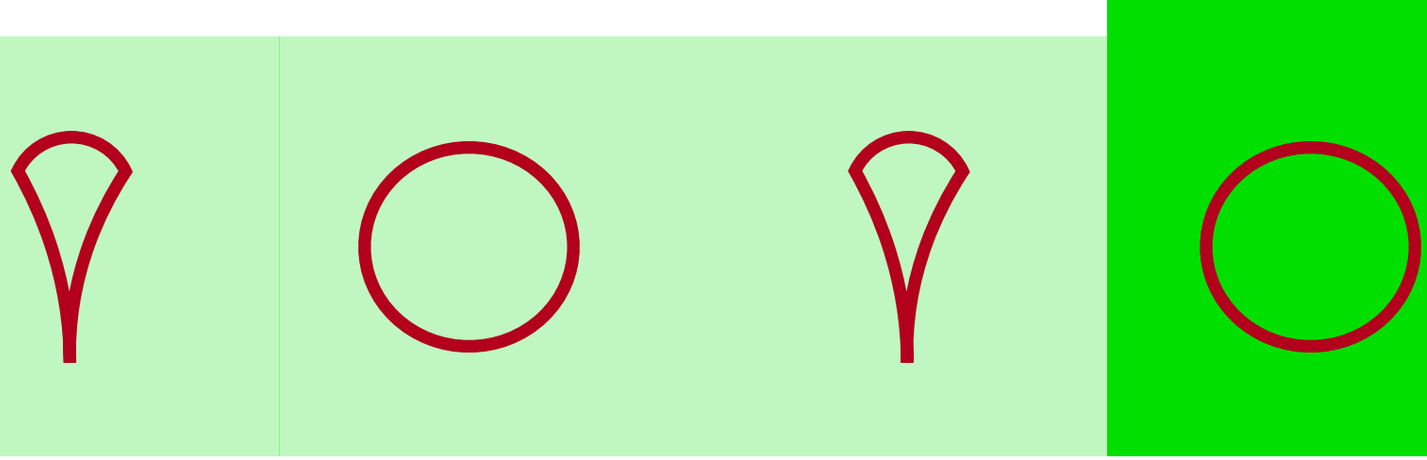
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

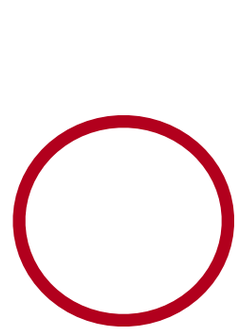
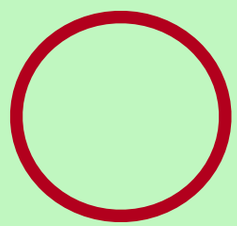
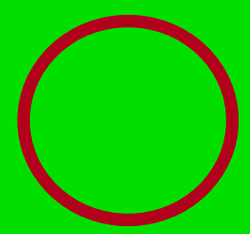
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

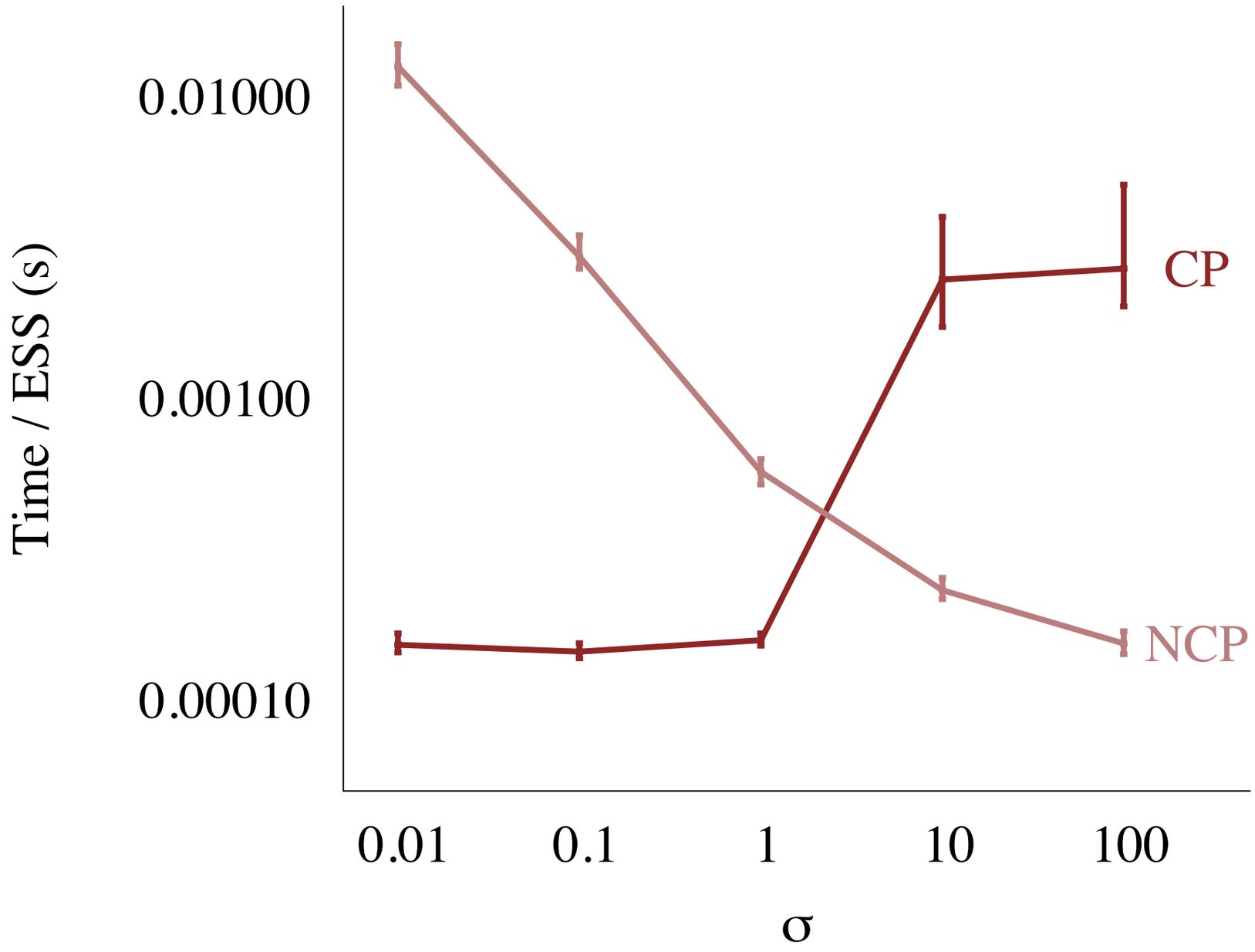
Consequently the two parameterizations are each advantageous in different circumstances.

Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization			
Non-Centered Parameterization			

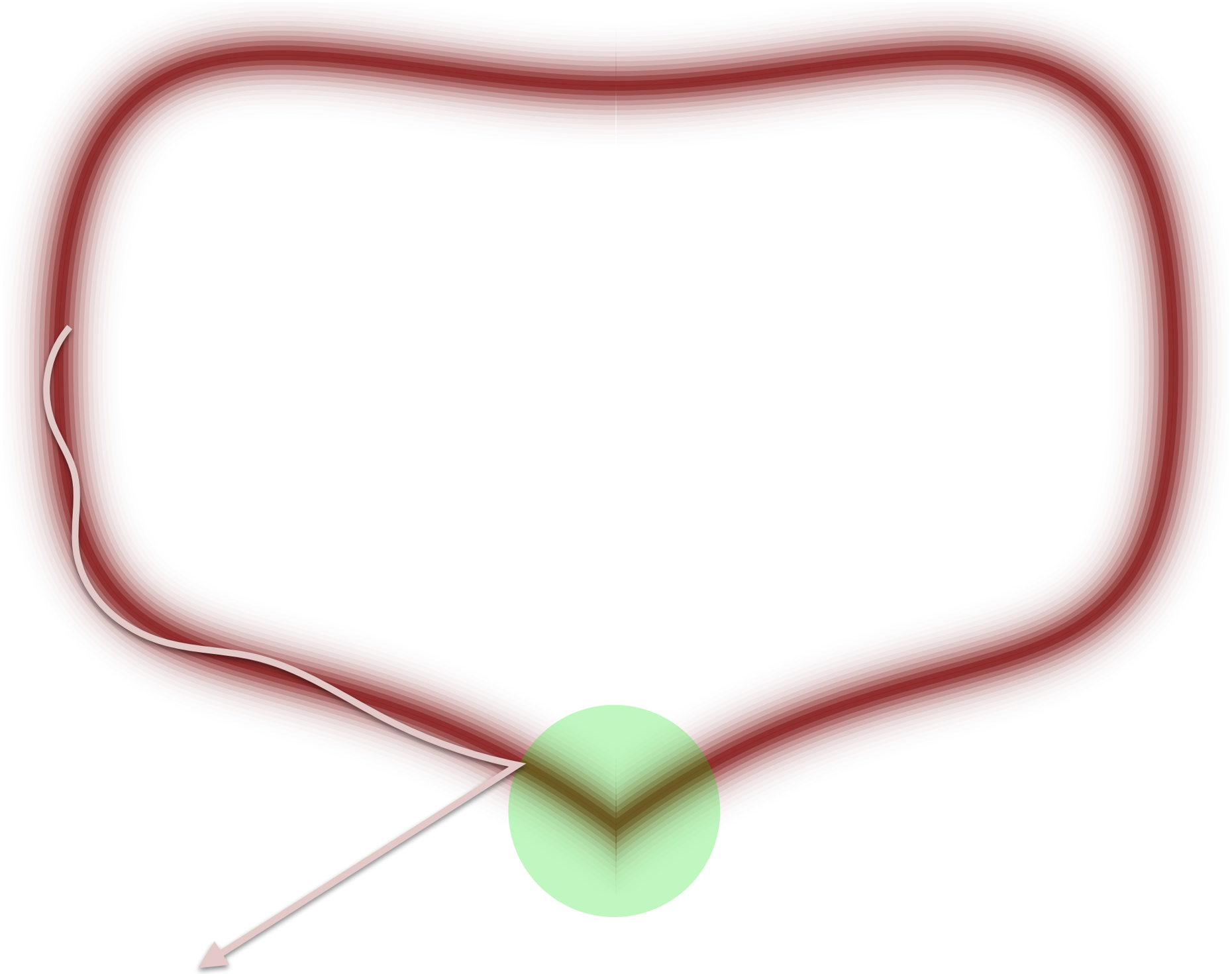
Consequently the two parameterizations are each advantageous in different circumstances.

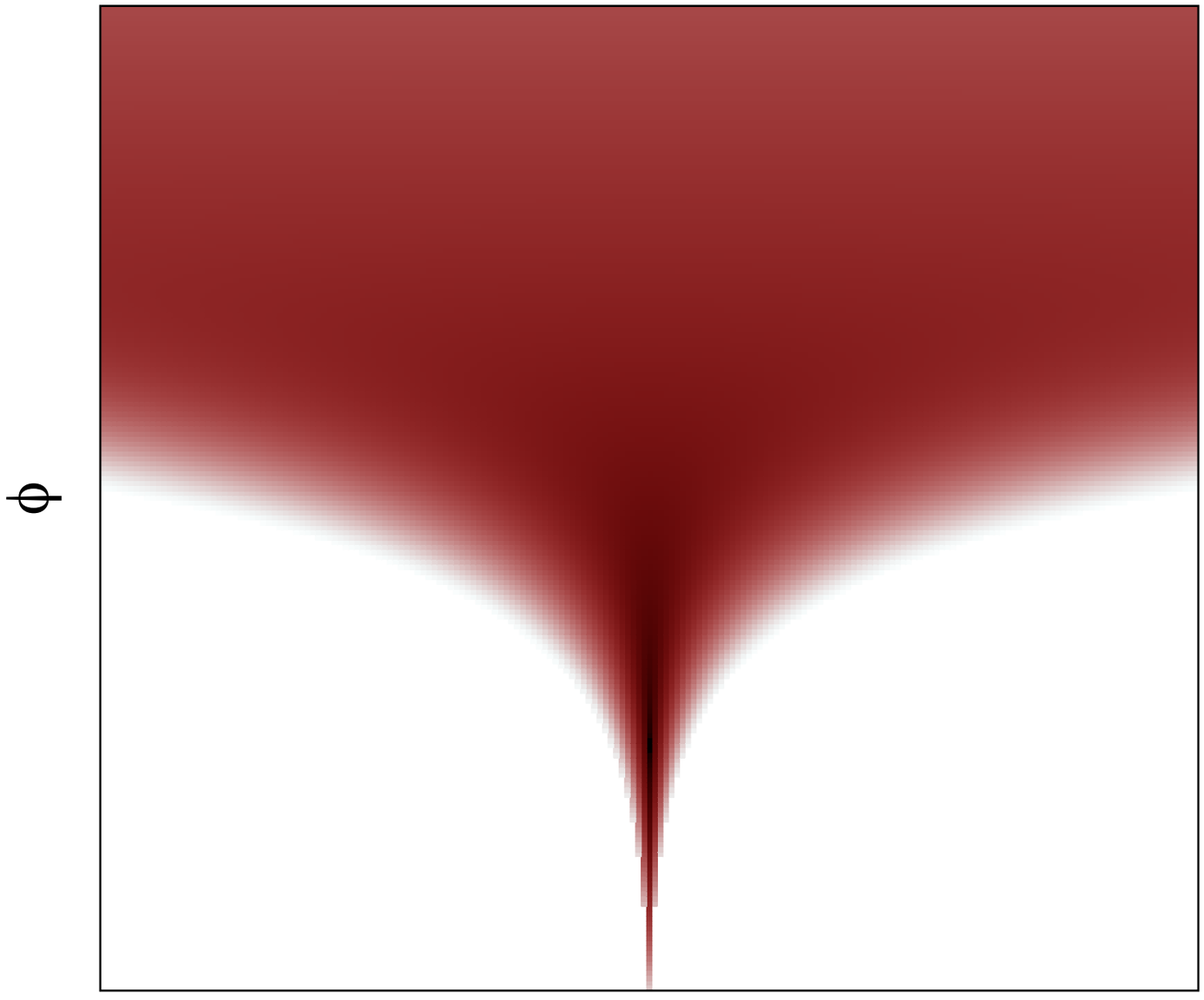
	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

Consequently the two parameterizations are each advantageous in different circumstances.



In practice, careful consideration of MCMC diagnostics help to identify poorly chosen parameterizations.





θ_n

Prediction in Hierarchical Models

In hierarchical models the separation between prior and likelihood is ambiguous.

$$\int \mathrm{d}\phi \prod_{n=1}^{N} p(y_n|\theta_n)\, p(\theta_n|\phi)\, p(\phi)$$

In hierarchical models the separation between prior and likelihood is ambiguous.

$$\underbrace{\prod_{n=1}^{N} p(y_n|\theta_n)}_{p(\mathbf{y}|\boldsymbol{\theta})} \underbrace{\int \mathrm{d}\phi \prod_{n=1}^{N} p(\theta_n|\phi)\, p(\phi)}_{p(\boldsymbol{\theta})}$$

In hierarchical models the separation between prior and likelihood is ambiguous.

$$\int \mathrm{d}\phi \prod_{n=1}^{N} \underbrace{p(y_n|\theta_n)\, p(\theta_n|\phi)}_{p(y_n,\theta_n|\phi)}\, \underbrace{p(\phi)}_{p(\phi)}$$

Although this doesn't affect the posterior distribution,
it does affect posterior predictive distributions.

Interpretation	**Posterior Predictive Sampling Procedure**
Population as *parameter*	$y_n \sim p(y_n \mid \theta_n)$
Population as *missing data*	$\theta_n \sim p(\theta_n \mid \phi)$ $y_n \sim p(y_n \mid \theta_n)$