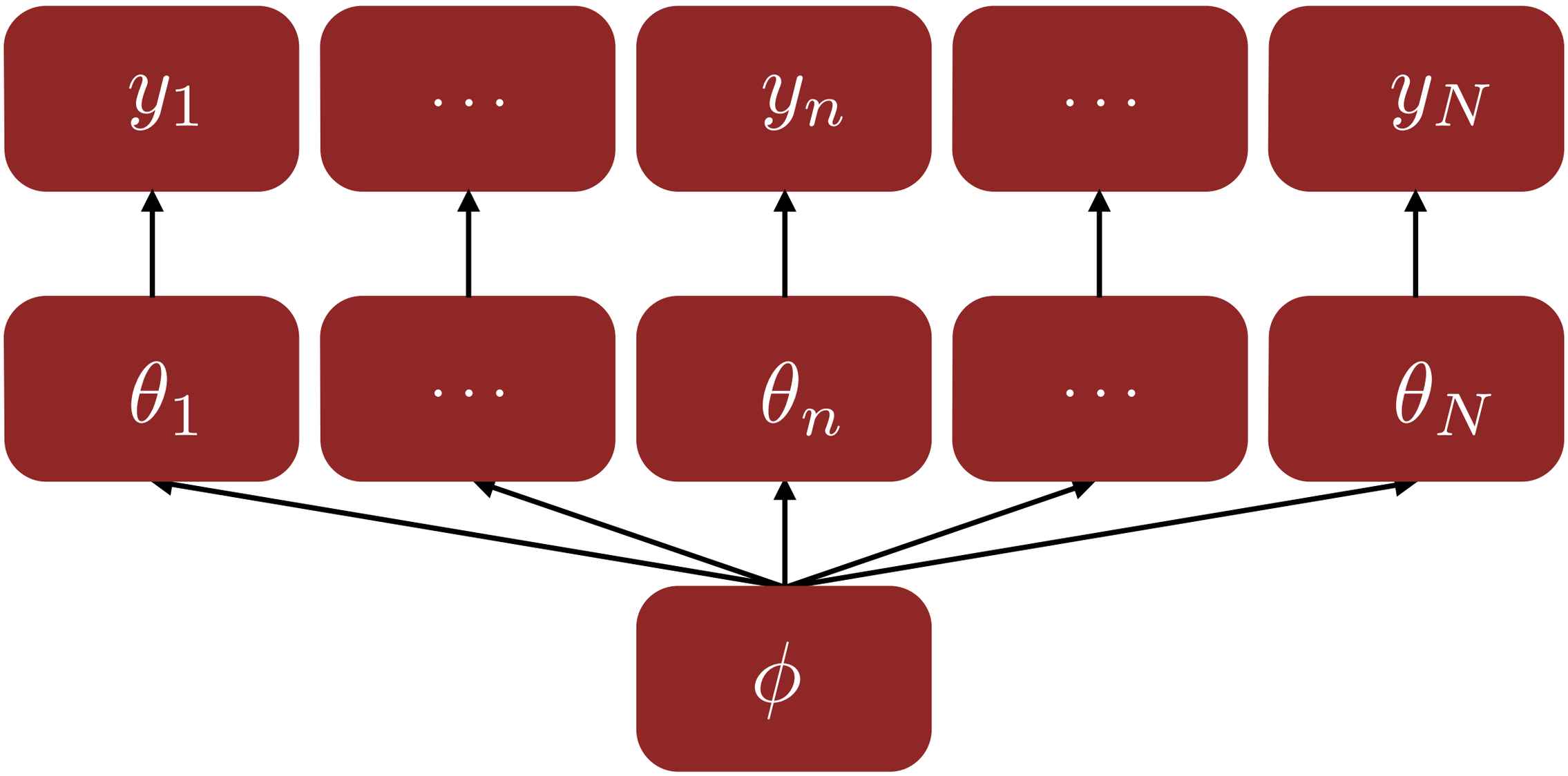
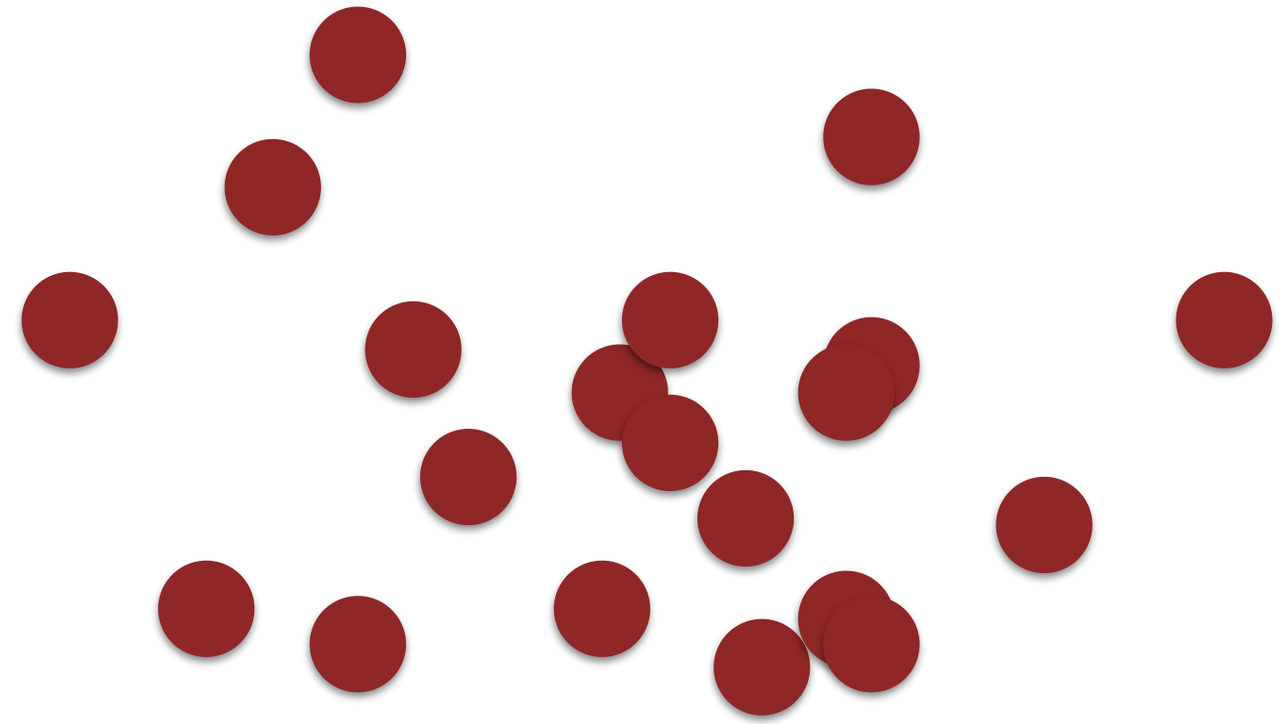
Hierarchical Models



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

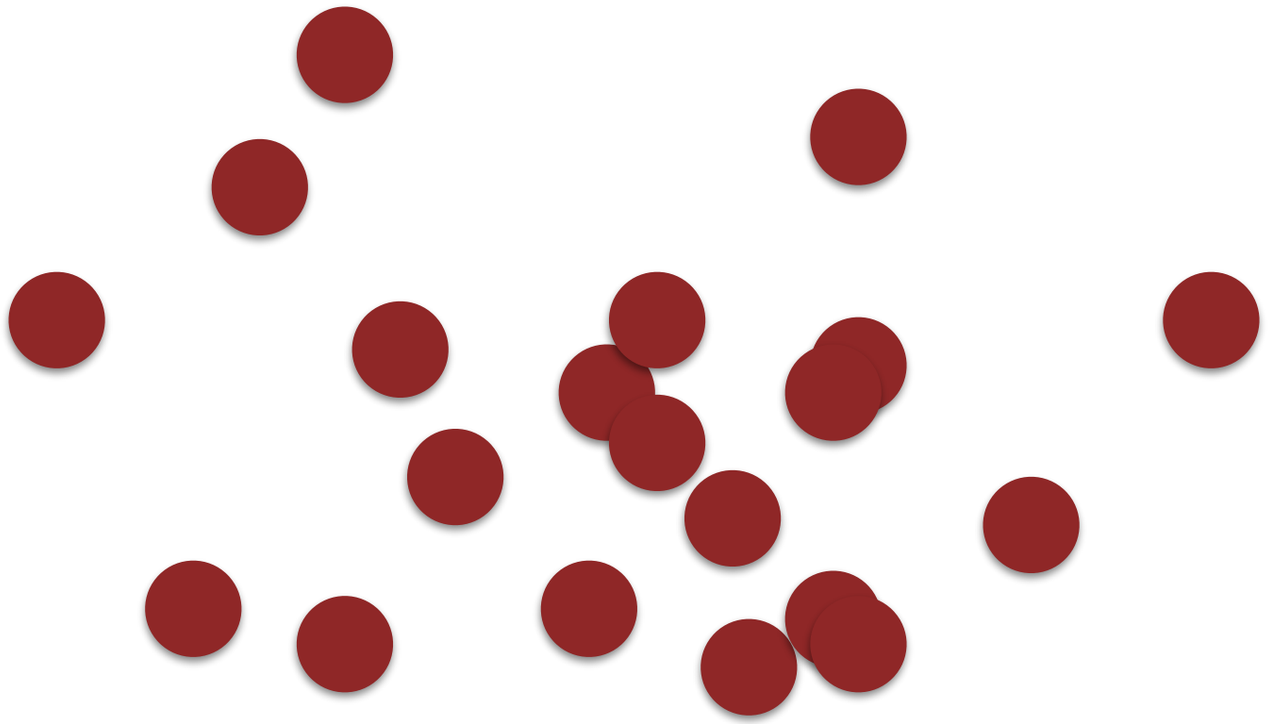
$$y \sim \mathcal{N}(\mathbf{X}^T \boldsymbol{\beta} + \alpha, \sigma)$$

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



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

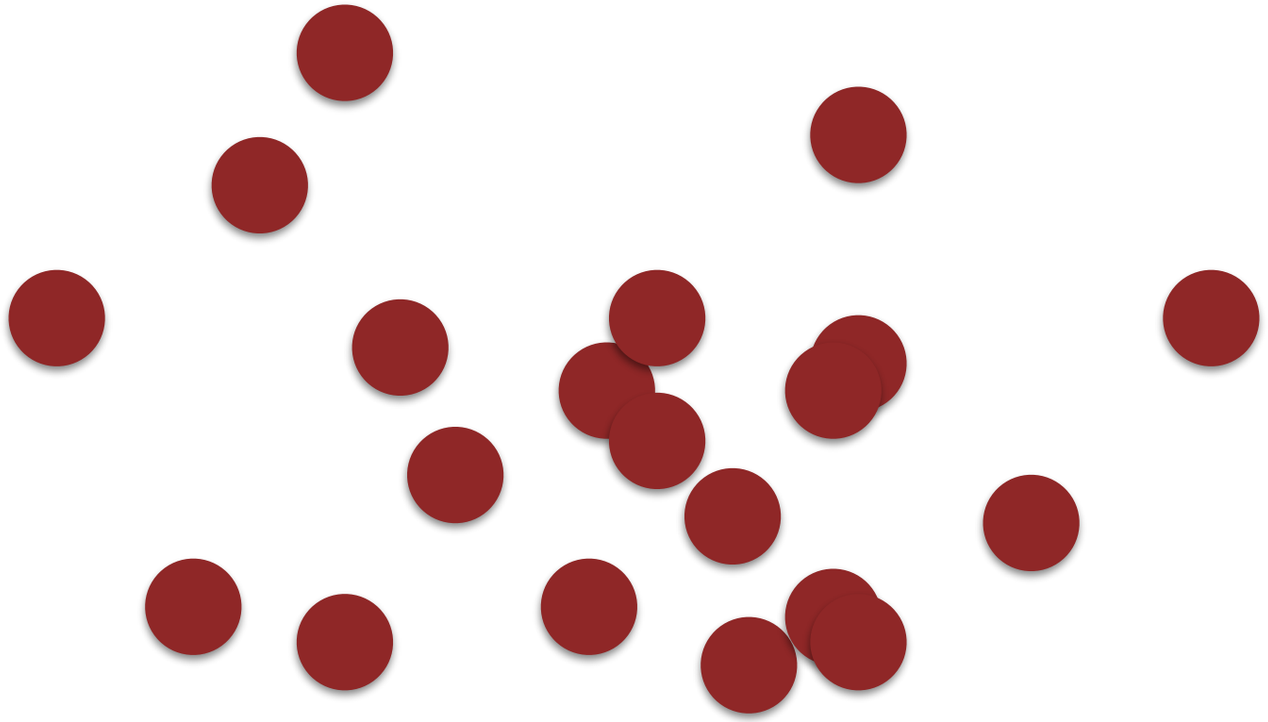
If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias.



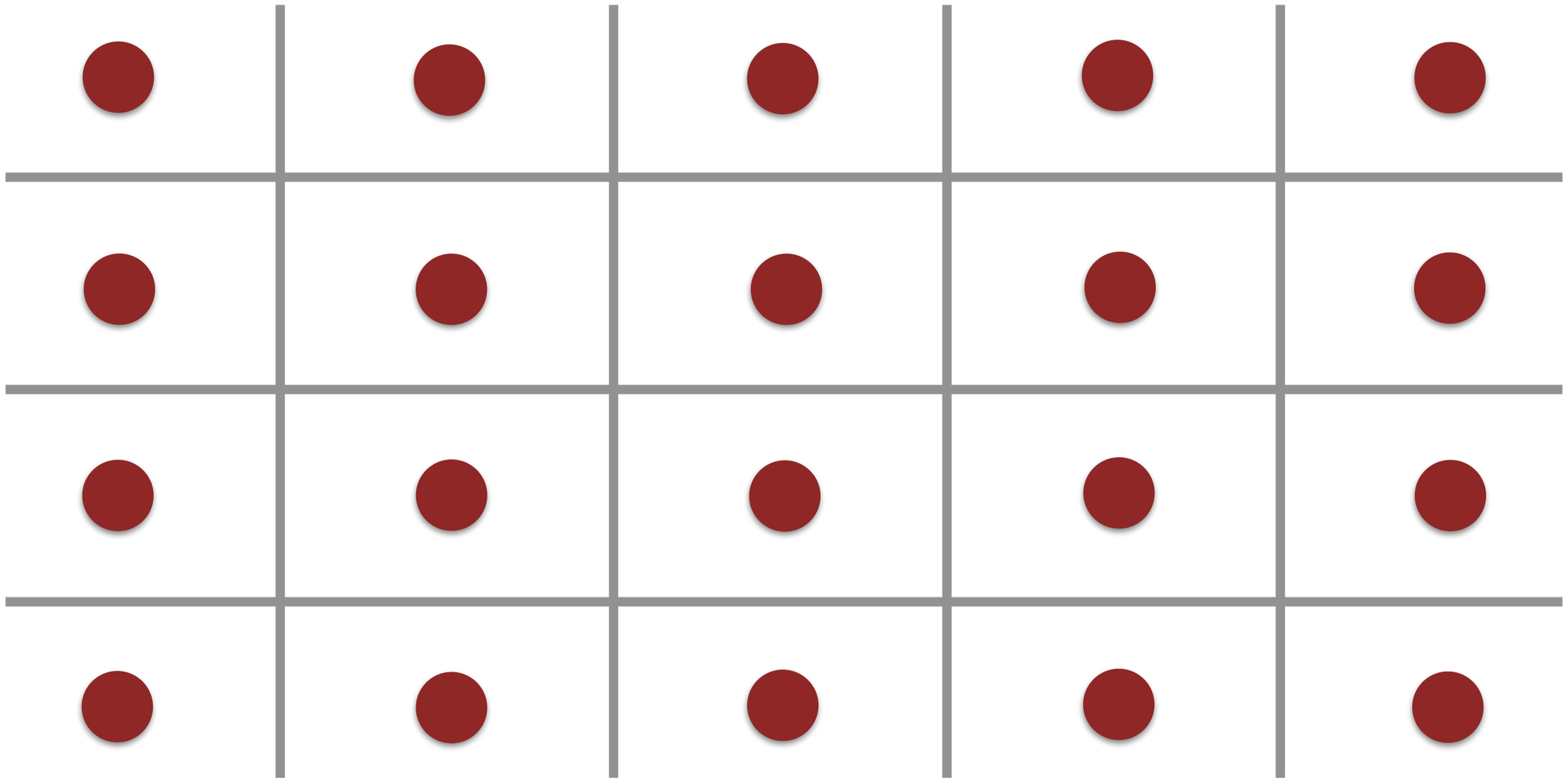
If we assume that every individual is equivalent then we can pool the data, but only at the expense of bias.

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

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.

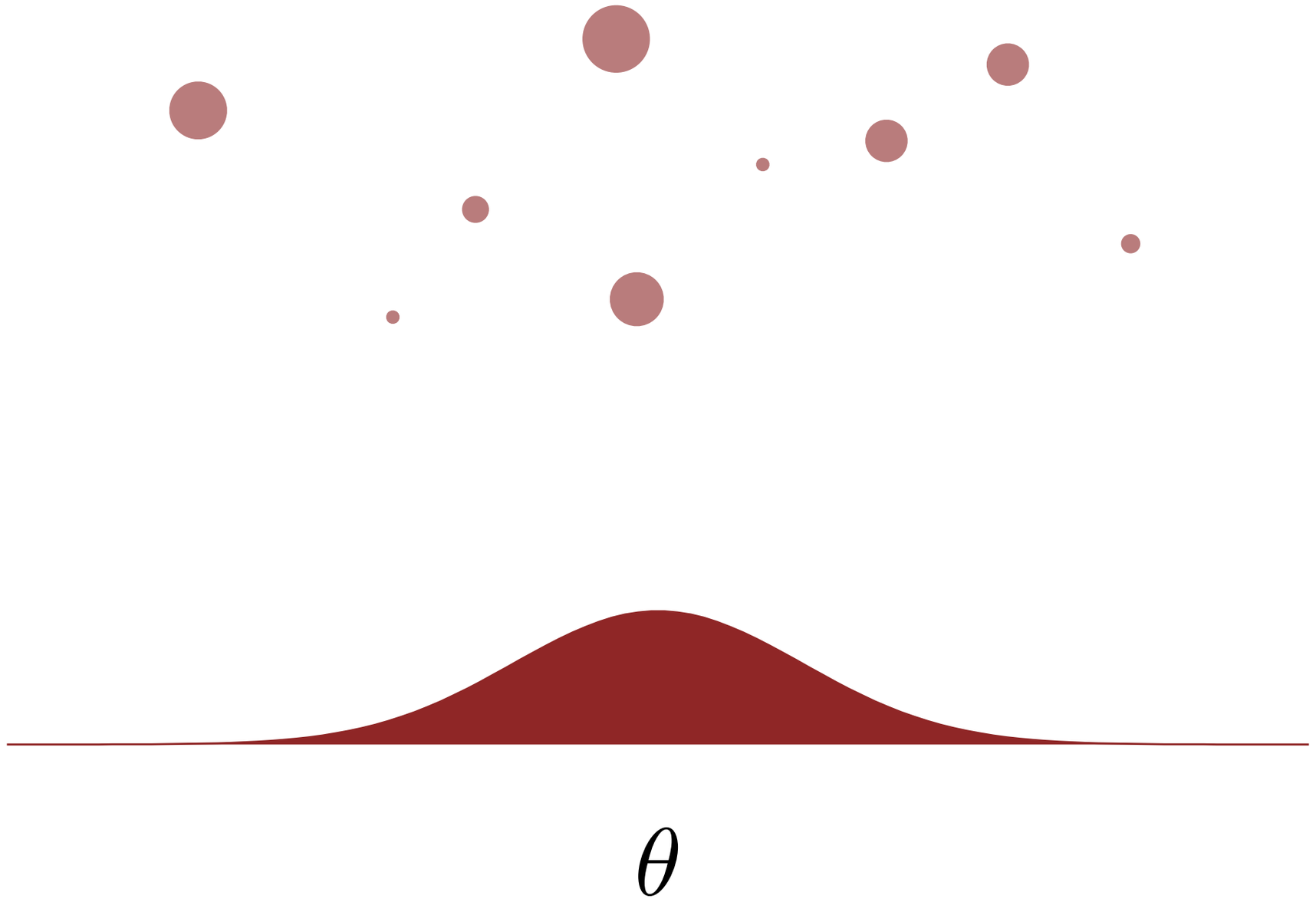


$$y_n \sim \mathcal{N}\left(\mathbf{X}^T \boldsymbol{\beta}_n + \alpha_n, \sigma\right)$$

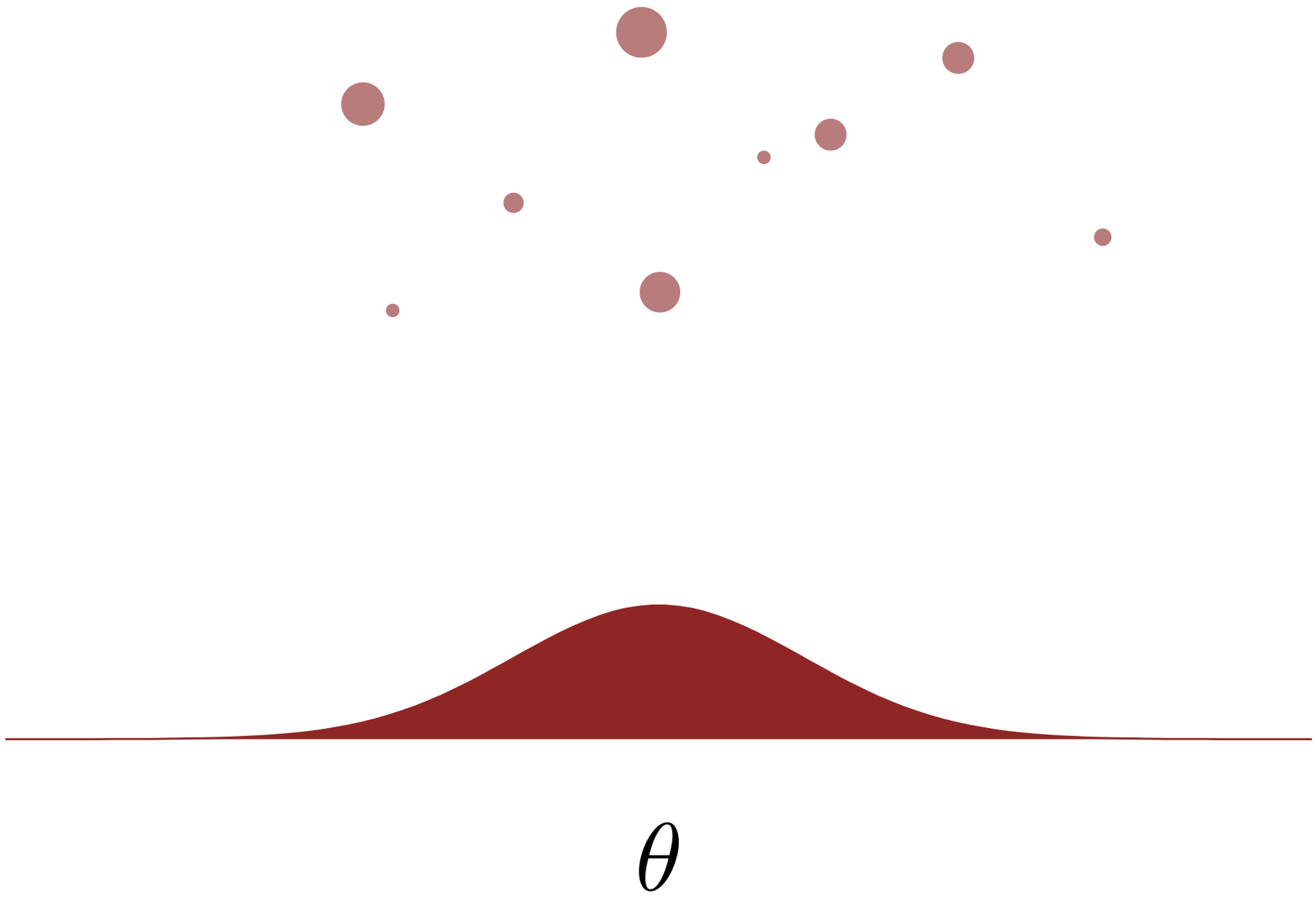
A compromise between complete pooling and no pooling that could balance bias and variance would be ideal.

θ

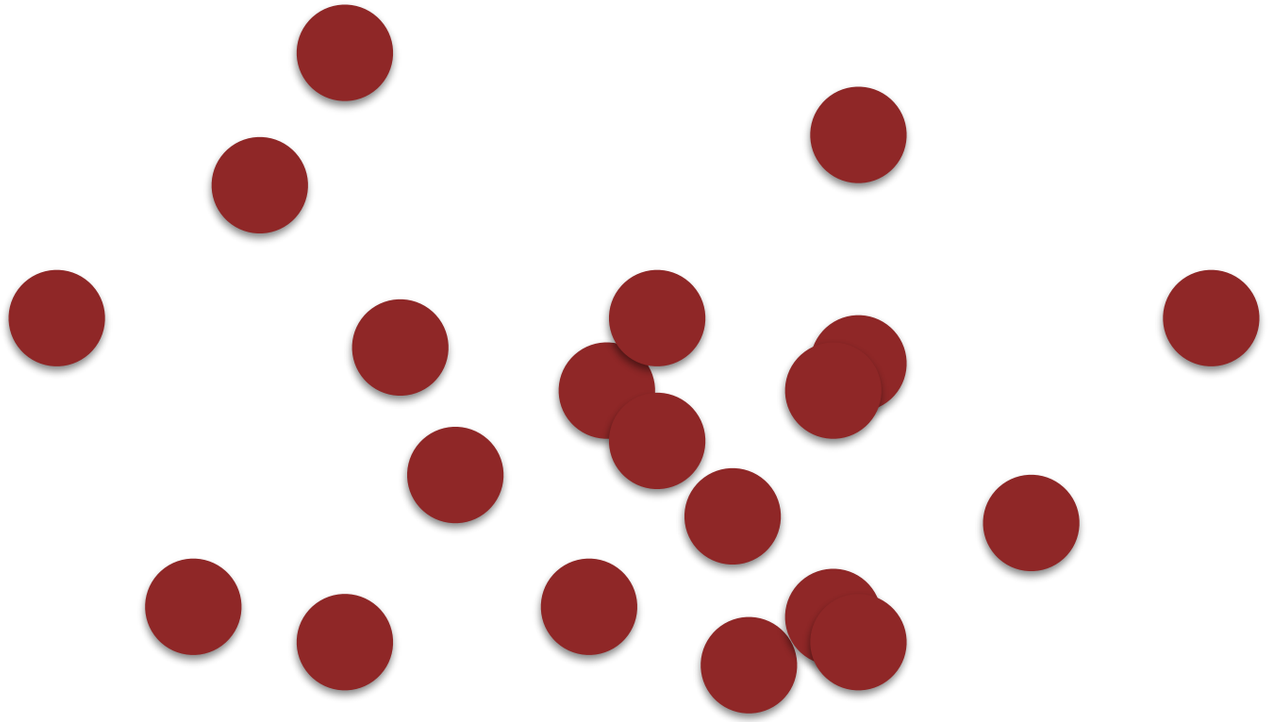
A compromise between complete pooling and no pooling that could balance bias and variance would be ideal.



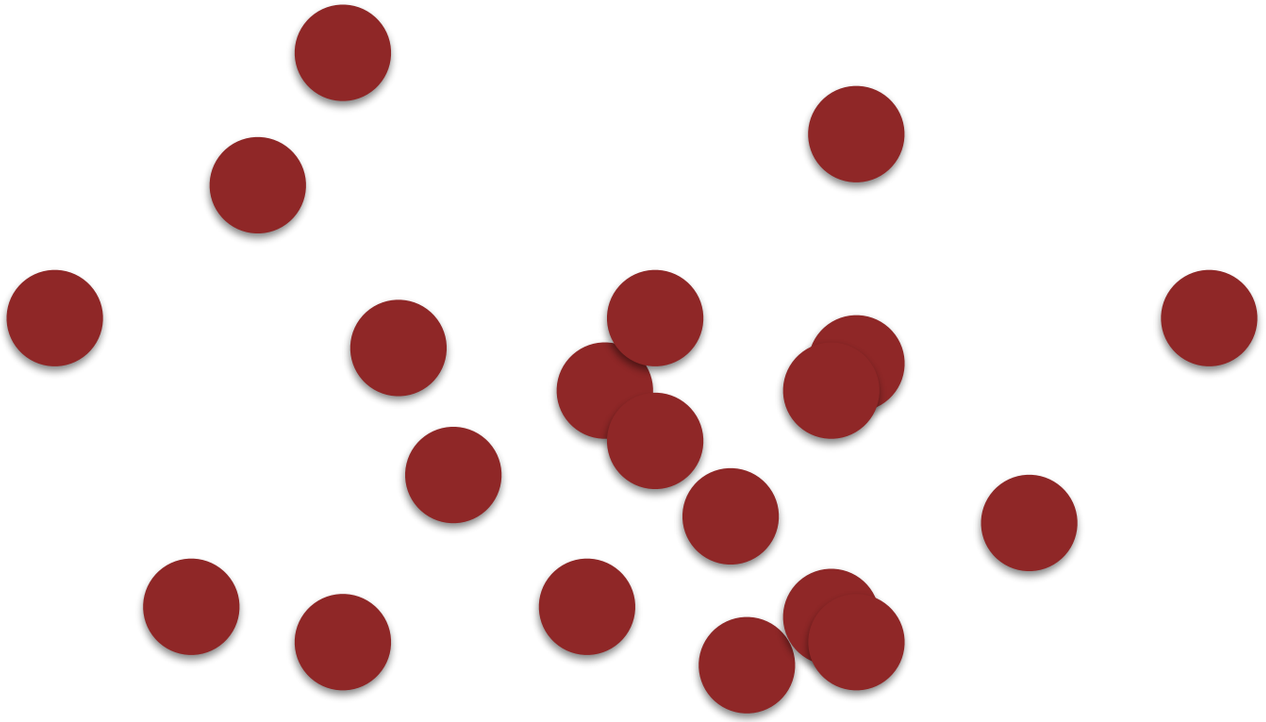
A compromise between complete pooling and no pooling that could balance bias and variance would be ideal.



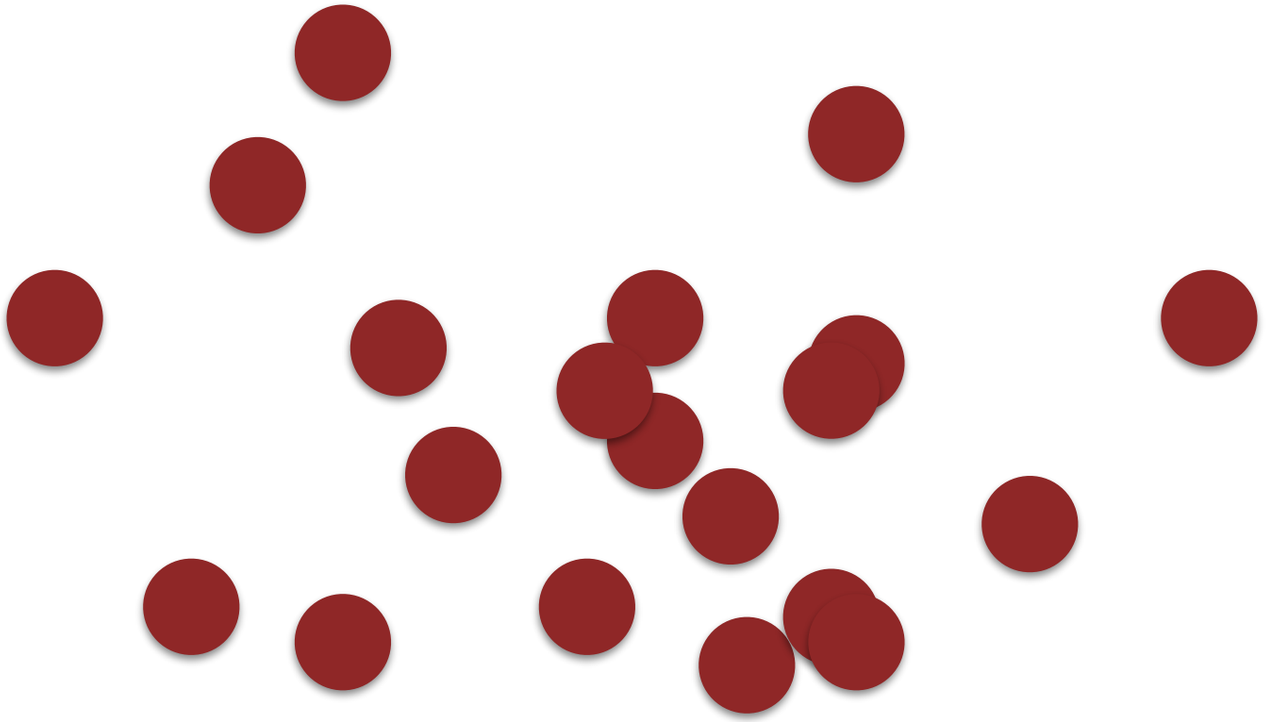
In order to formalize this approach we need to consider *exchangeabilitiy*.



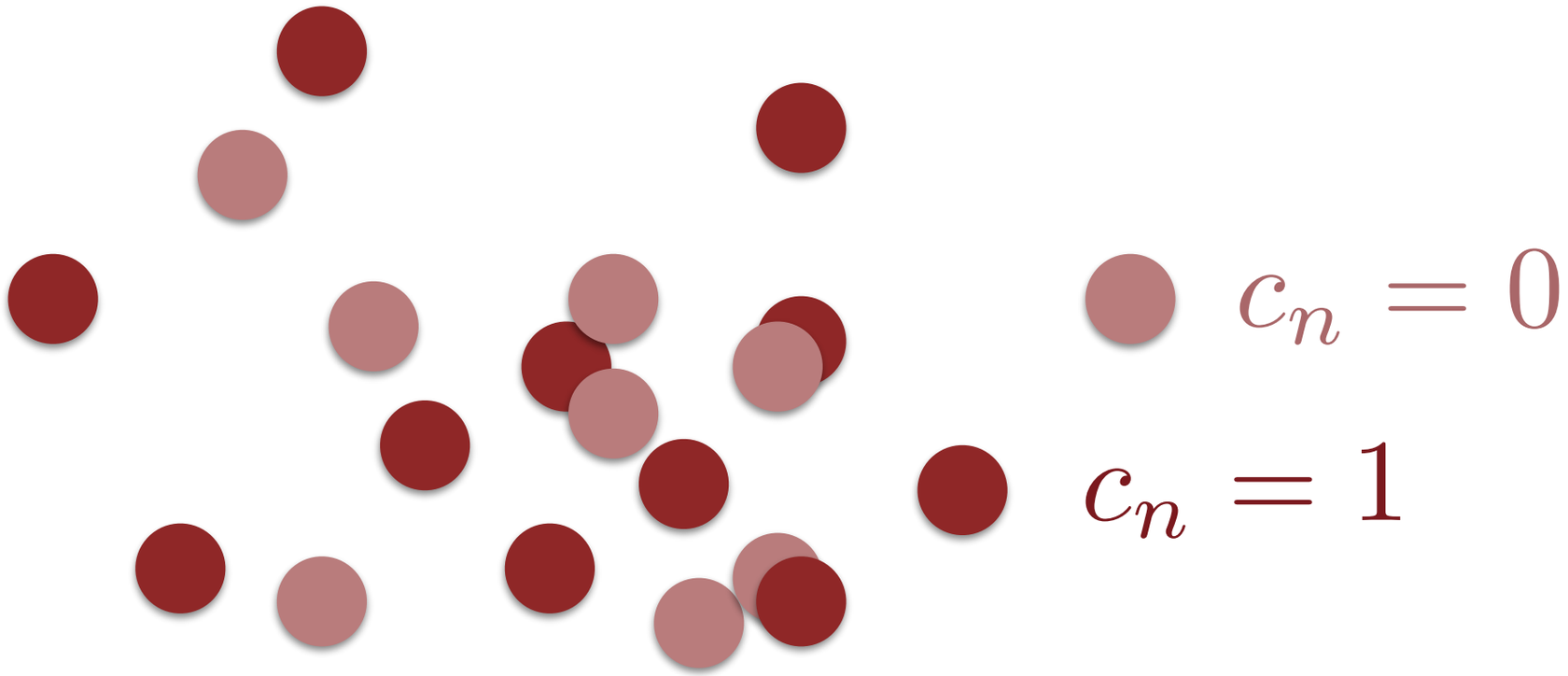
In order to formalize this approach we need to consider *exchangeabilitiy*.



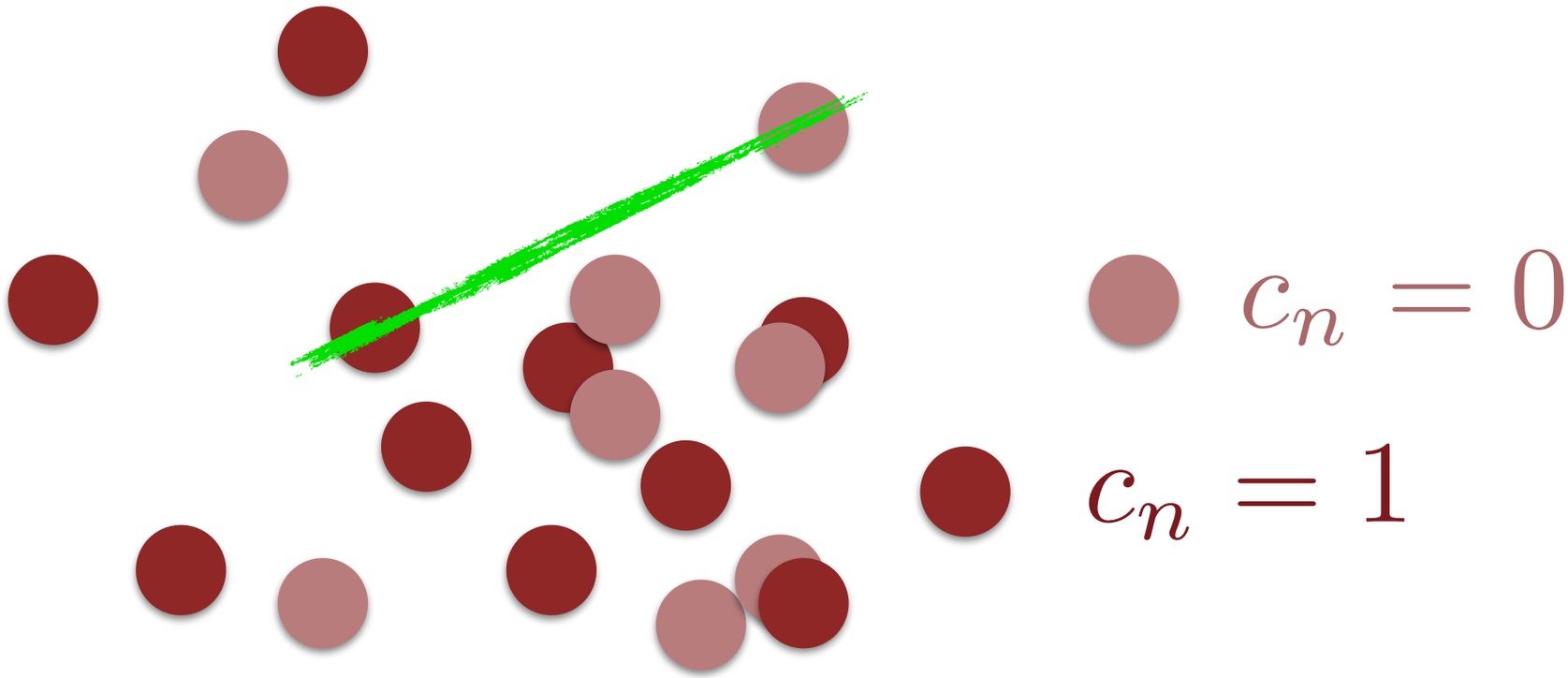
In order to formalize this approach we need to consider *exchangeabilitiy*.



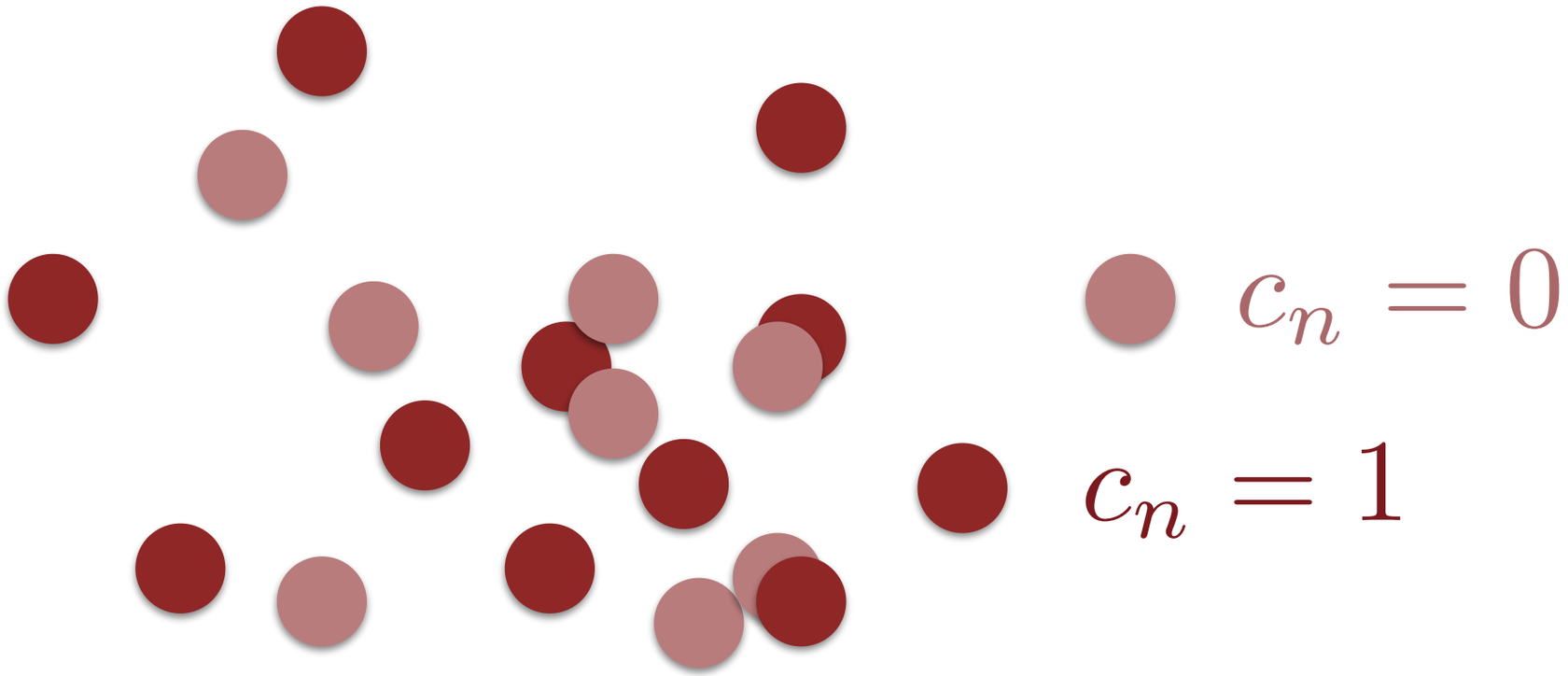
Given labels that discriminate between individuals,
the individuals are no longer exchangeable.



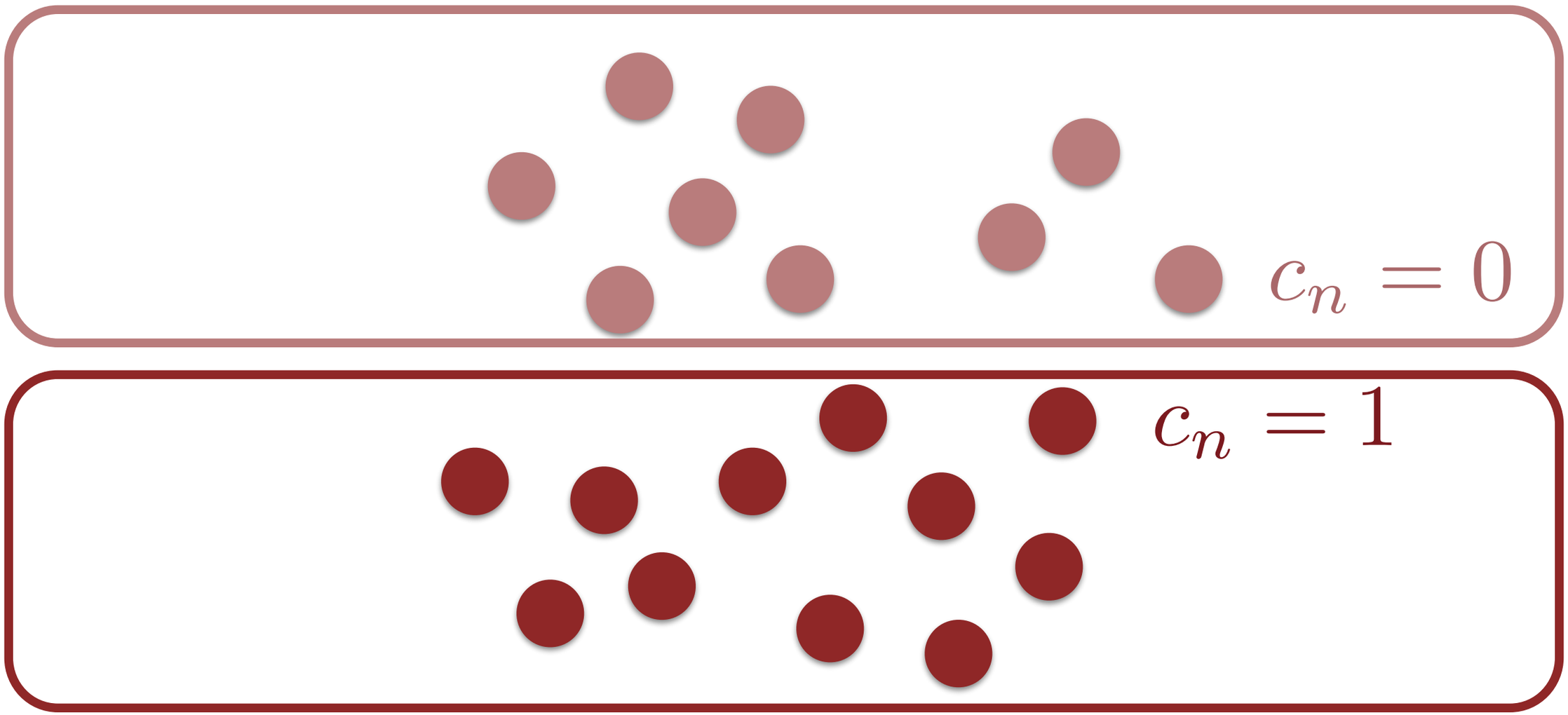
Given labels that discriminate between individuals,
the individuals are no longer exchangeable.



The partitions of the population, however, define exchangeable groups within a metapopulation.



The partitions of the population, however, define exchangeable groups within a metapopulation.



The partitions of the population, however, define exchangeable groups within a metapopulation.

$c_n = 0$

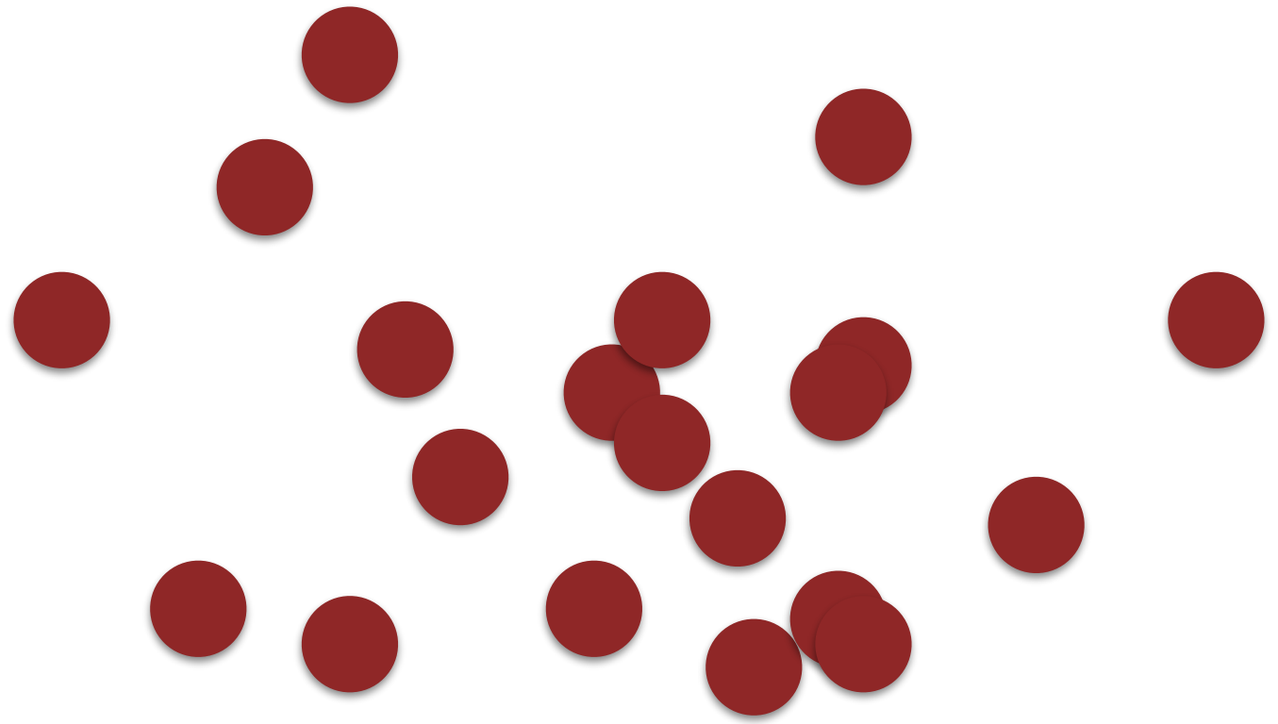
$c_n = 1$

The partitions of the population, however, define exchangeable groups within a metapopulation.

$c_n = 0$

$c_n = 1$

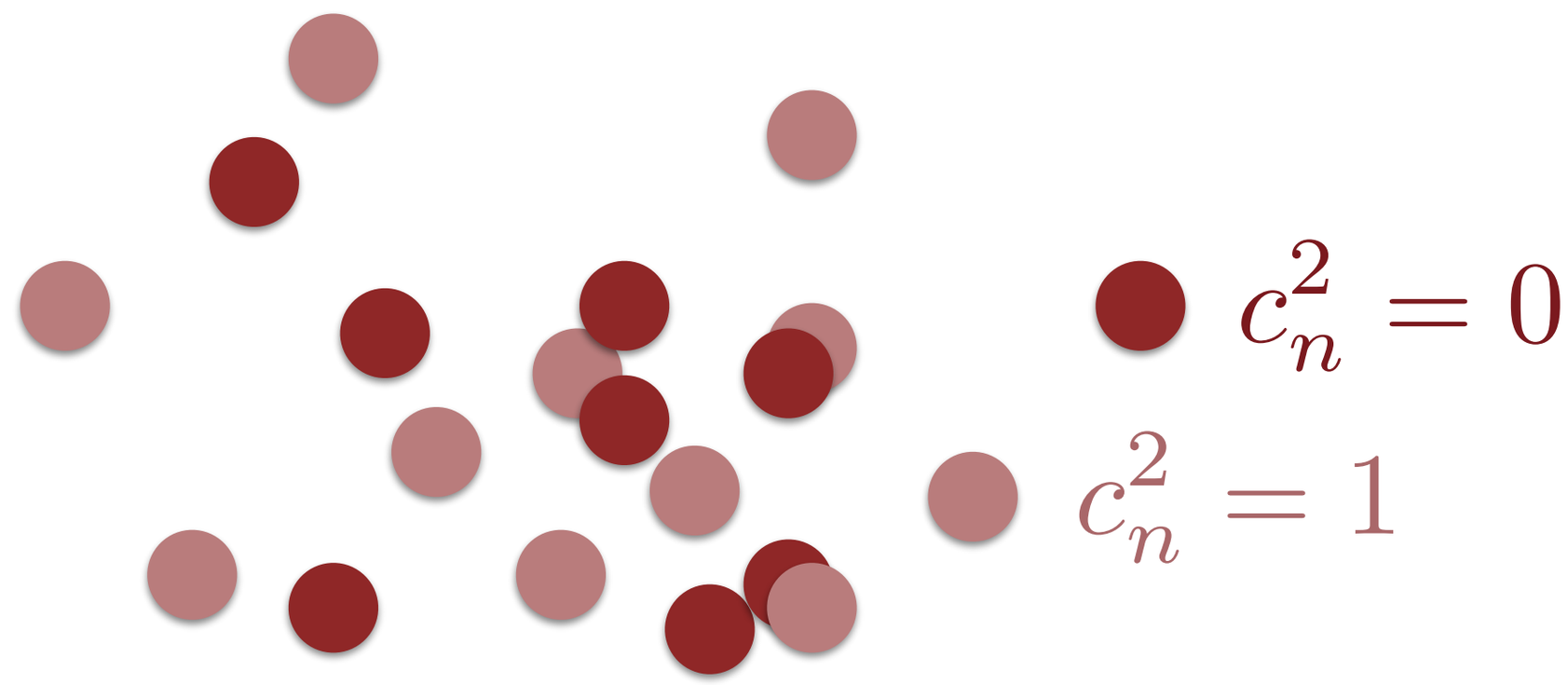
The partitions of the population, however, define exchangeable groups within a metapopulation.



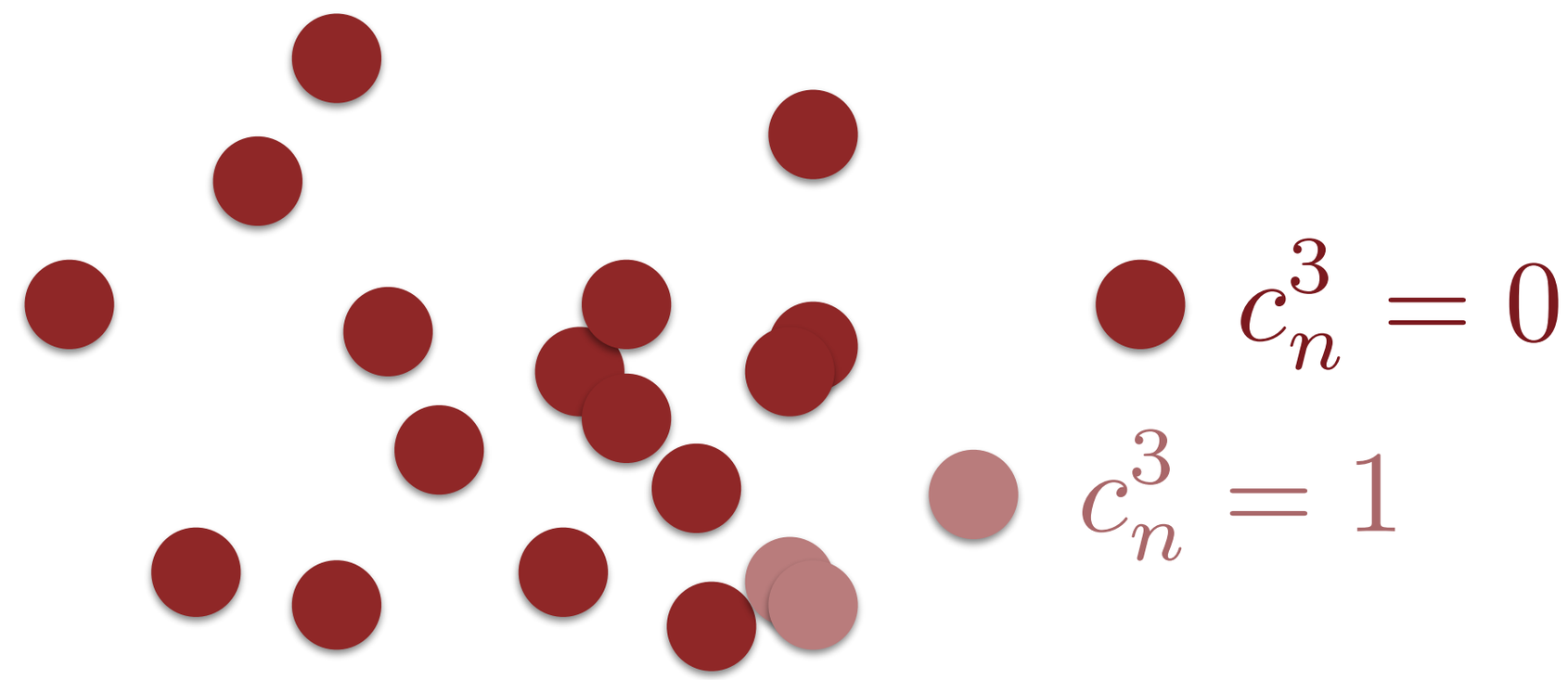
$$\left\{c_n^1, c_n^2, c_n^3\right\} \in \{0, 1\}$$

Every discriminating label defines new, exchangeable groups and new metapopulations.

Every discriminating label defines new, exchangeable groups and new metapopulations.



Every discriminating label defines new, exchangeable groups and new metapopulations.



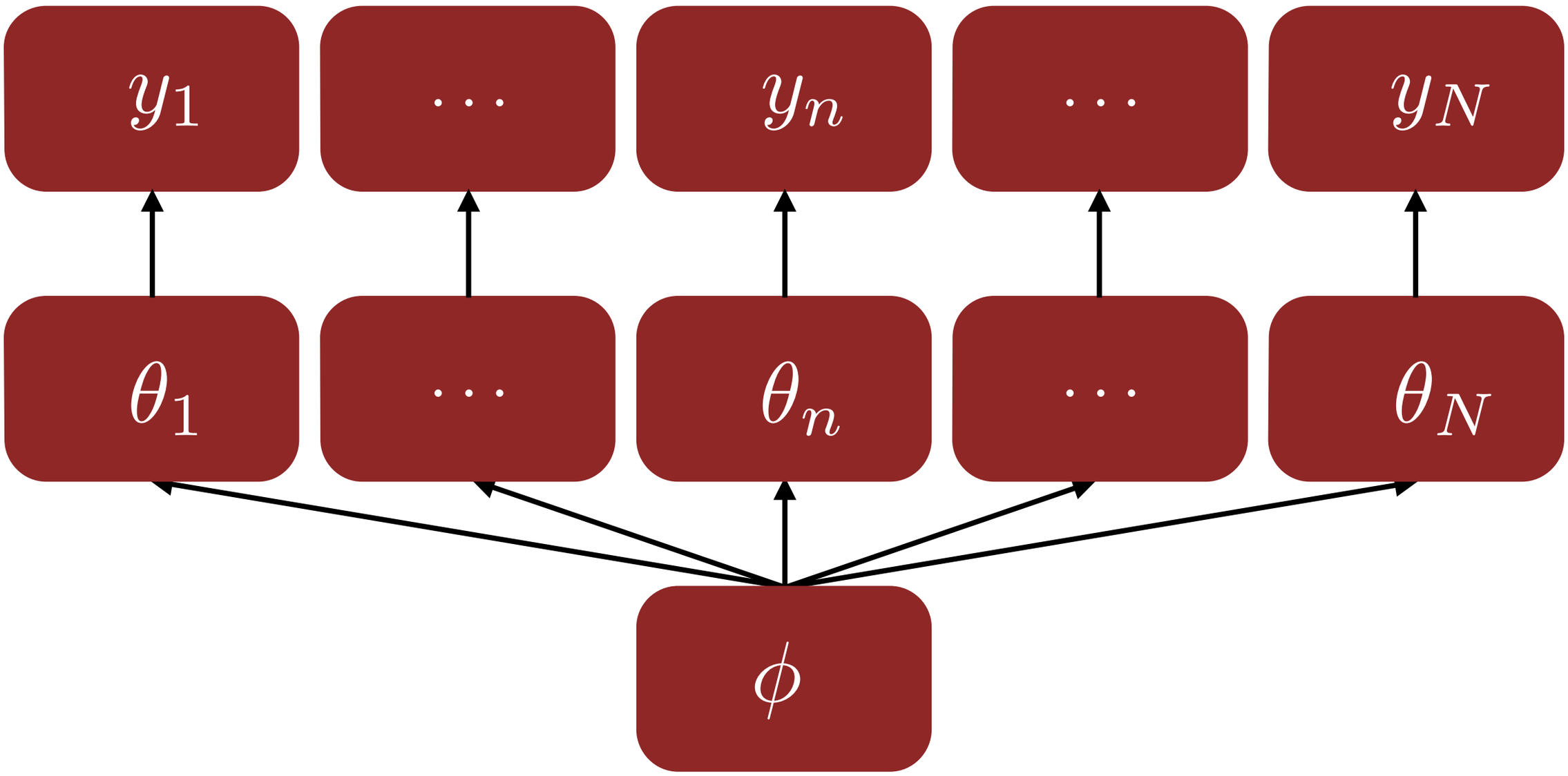
As the population grows, the only prior distribution that respects exchangeability is a *hierarchical* distribution.

$$\pi(\boldsymbol{\theta}) = \int \mathrm{d}\phi \prod_{n=1}^{N} \pi(\theta_n|\phi)\, \pi(\phi)$$

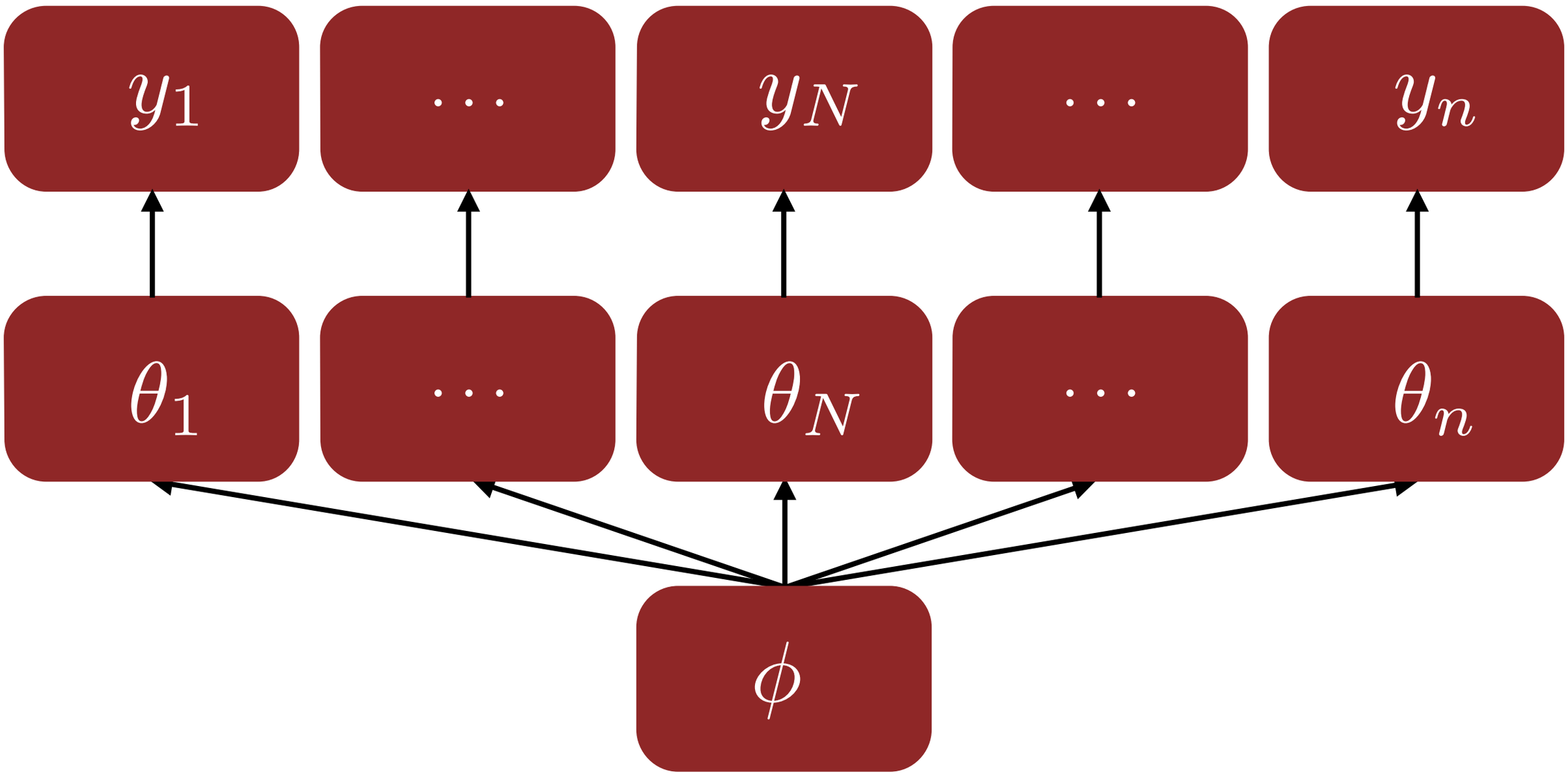
As the population grows, the only prior distribution that respects exchangeability is a *hierarchical* distribution.

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

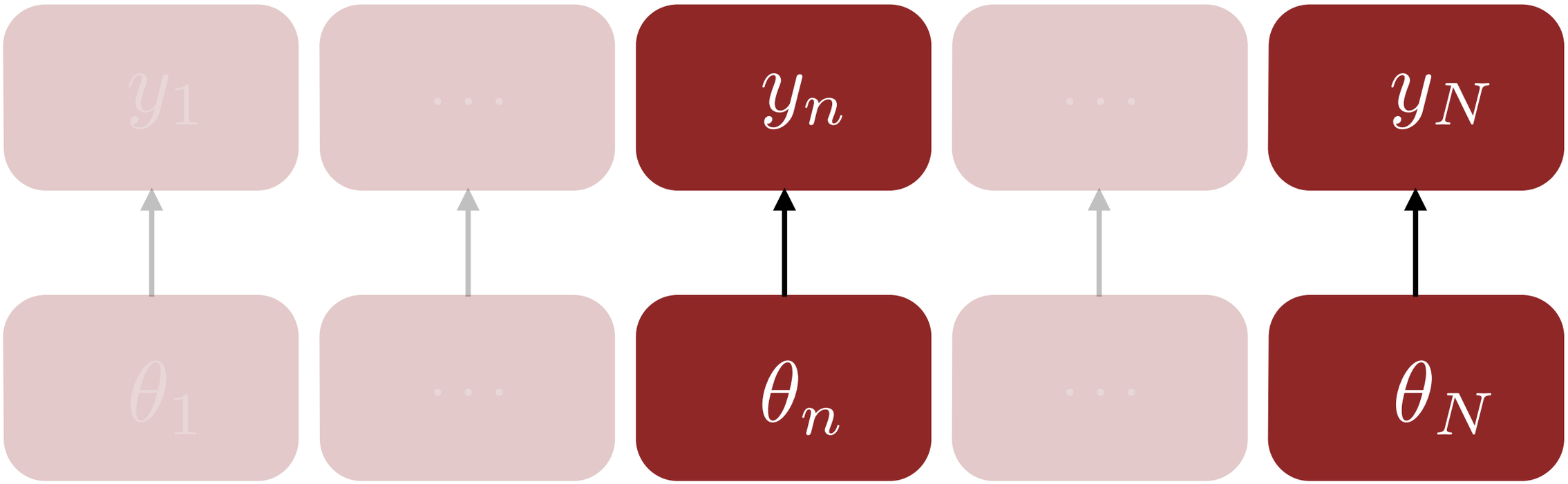
As the population grows, the only distribution that respects exchangeability is a *hierarchical* distribution.



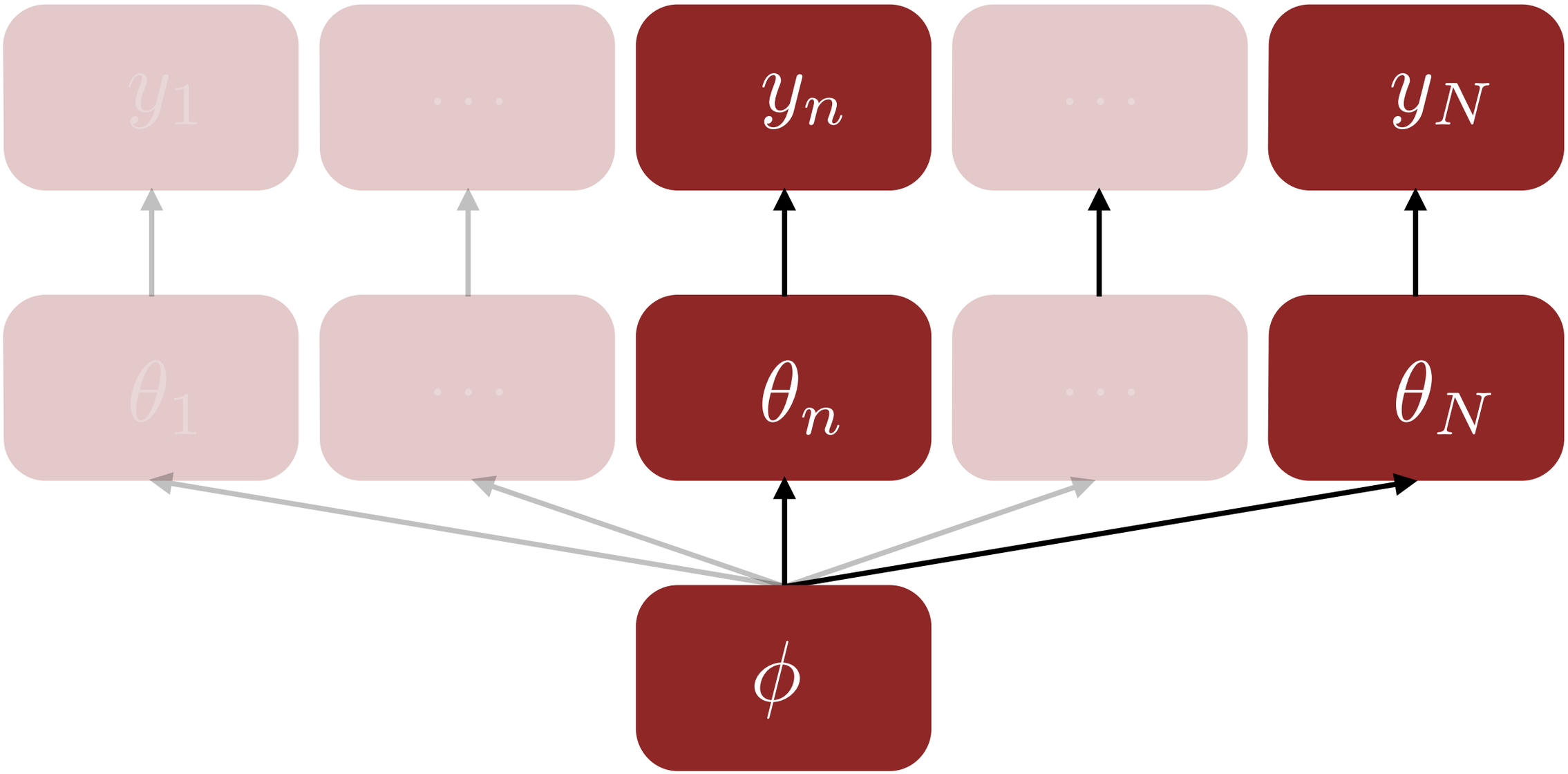
As the population grows, the only distribution that respects exchangeability is a *hierarchical* distribution.



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.

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

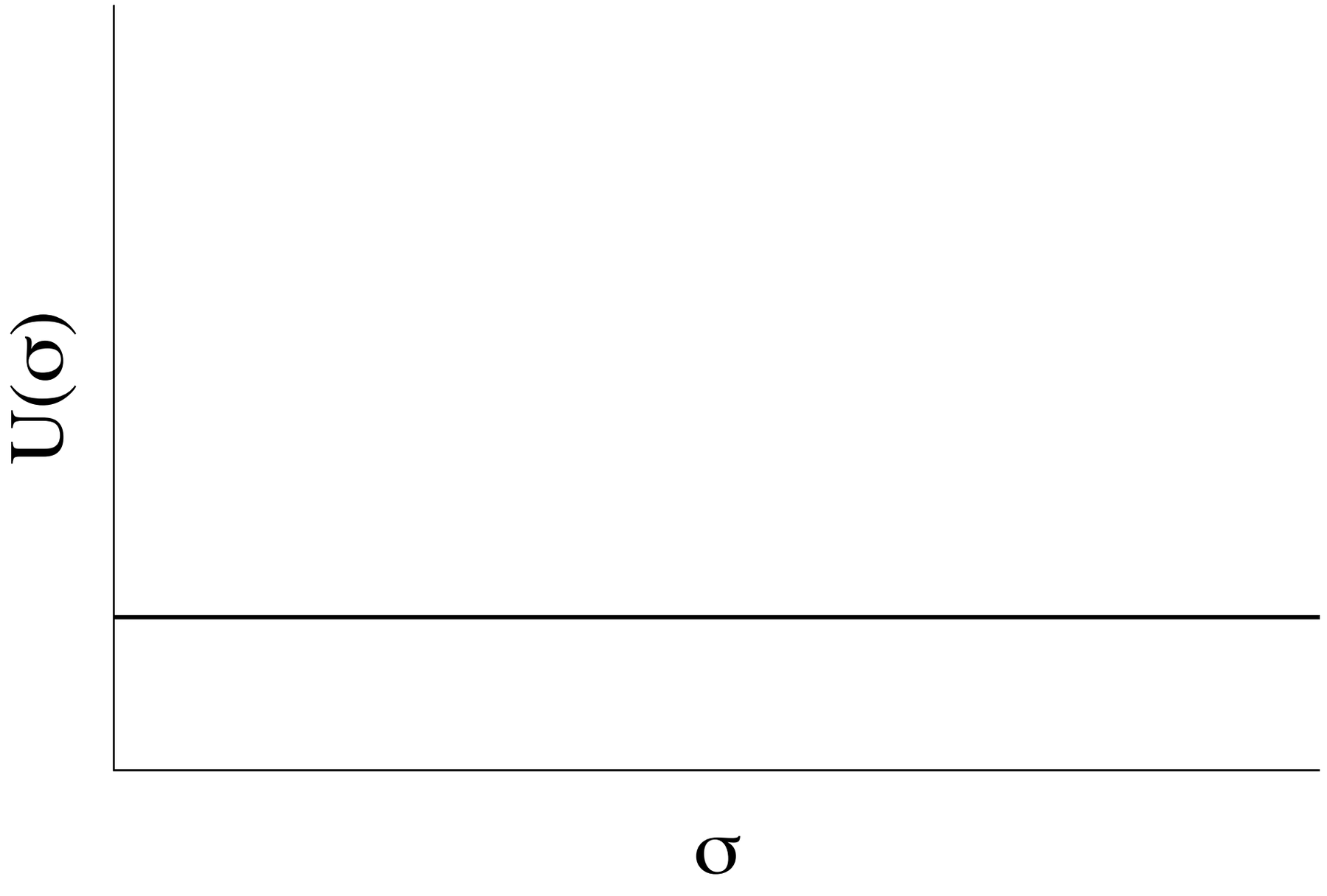
The most common population model is a Gaussian, where we model the population mean and variance.

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

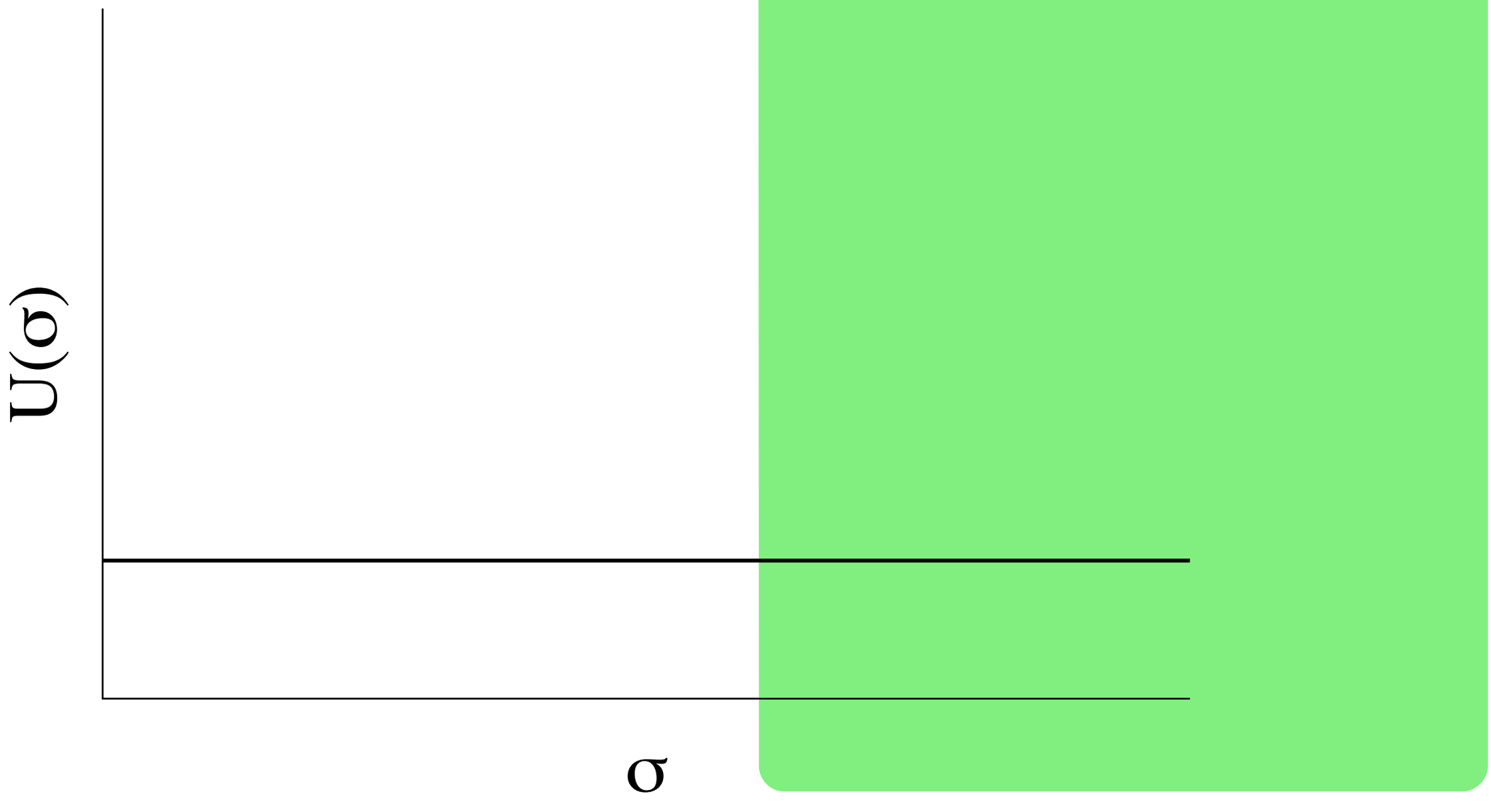
Informative prior distributions on the population hyperparameters are incredibly important.

$$\pi(\mu)\,\pi(\sigma)$$

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

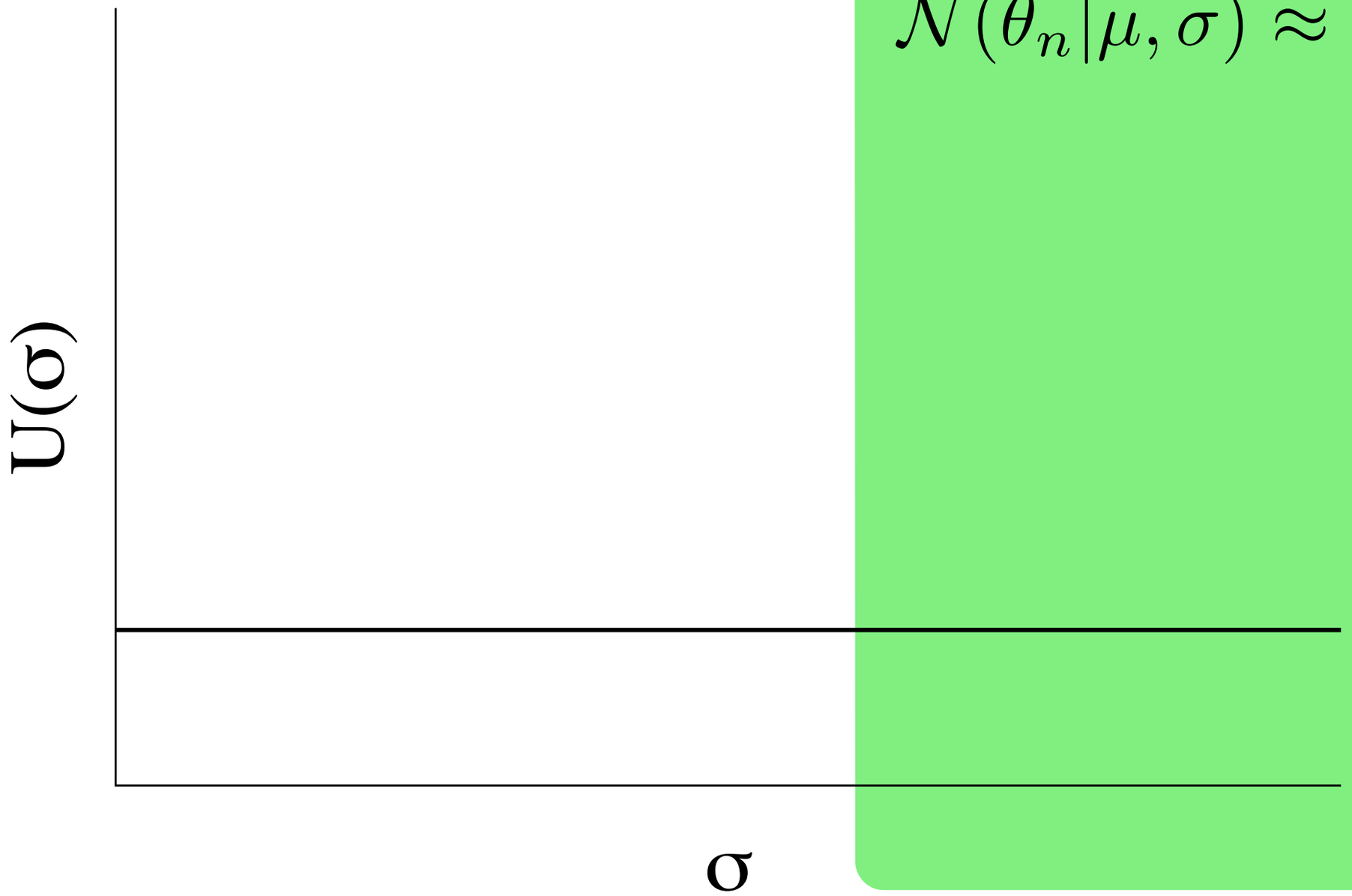


A uniform prior on the population deviance places too much probability at high values and no pooling.

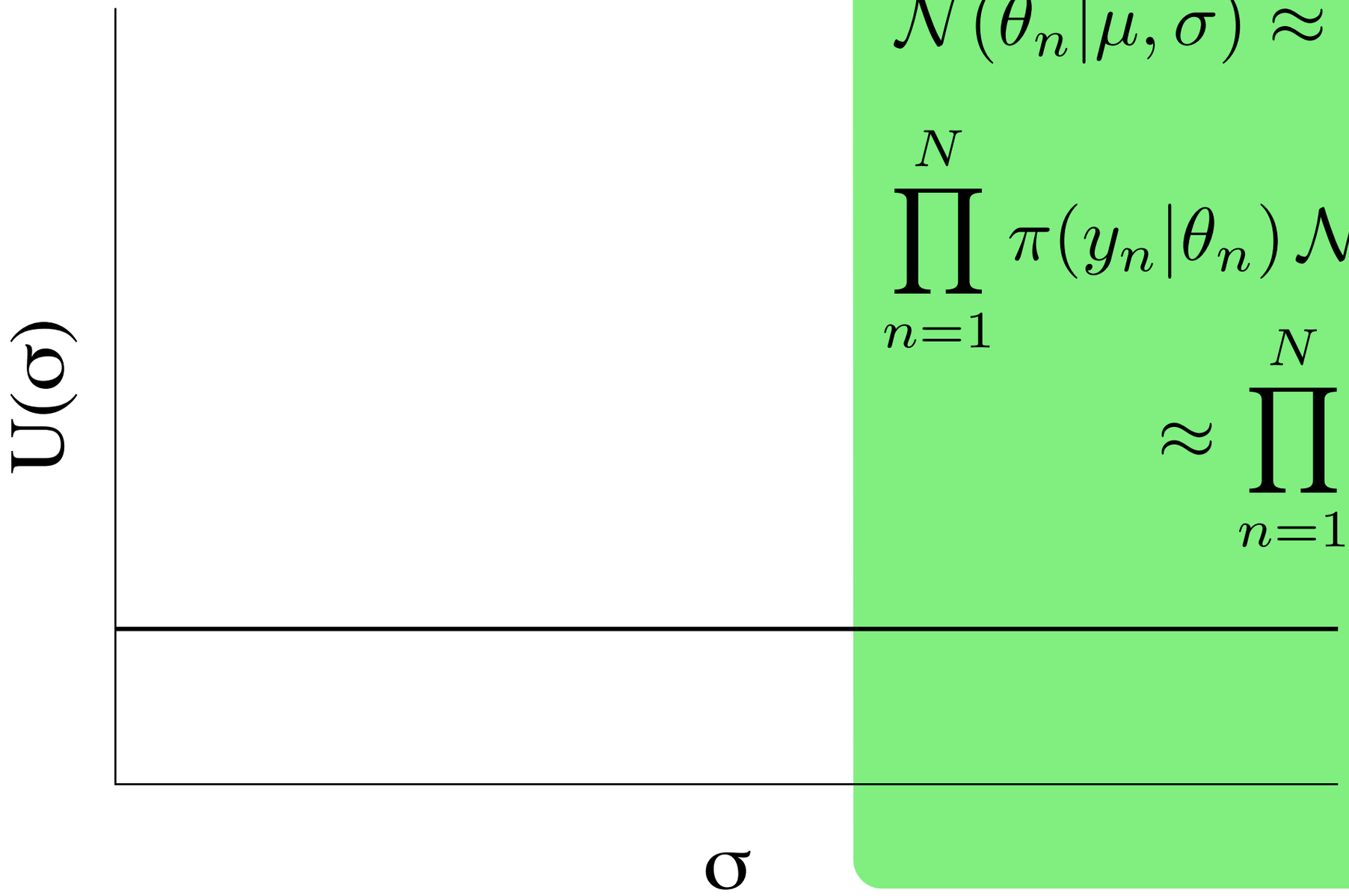


A uniform prior on the population deviance places too much probability at high values and no pooling.

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



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

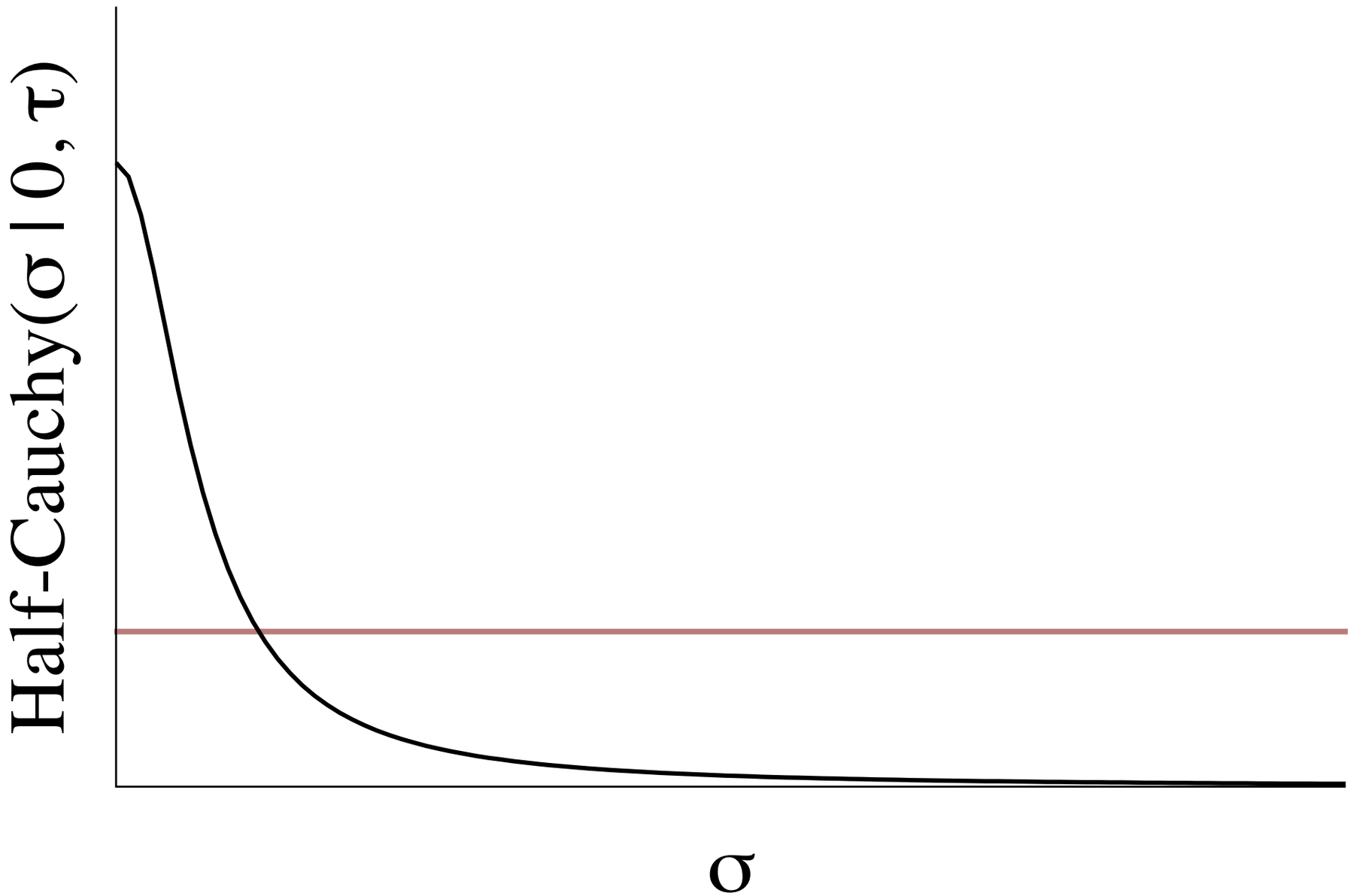


$$\mathcal{N}(\theta_n|\mu,\sigma) \approx U(\theta_n)$$

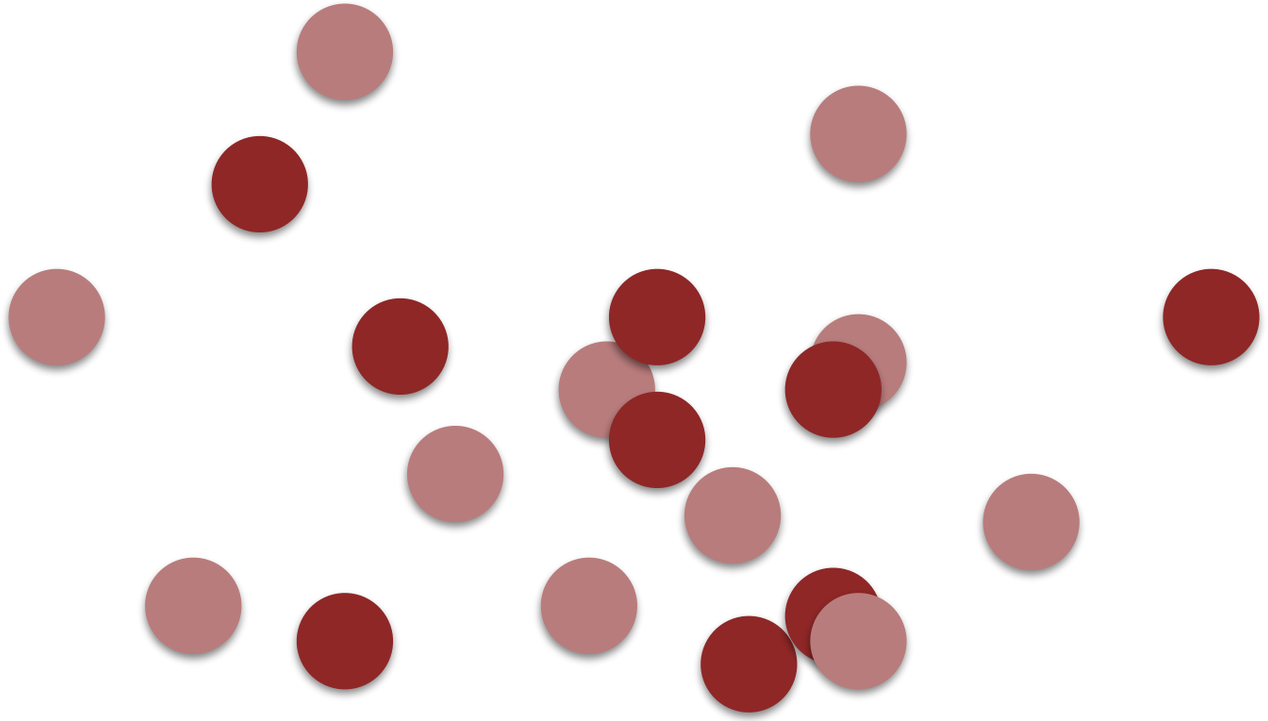
$$\prod_{n=1}^{N} \pi(y_n|\theta_n)\,\mathcal{N}(\theta_n|\mu,\sigma)$$

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

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



General Linear Hierarchical Models



General linear models are naturally suited to exchangeability and hierarchical models.

$$\pi\left(y|g\left(\mathbf{X}^T\boldsymbol{\beta}+\alpha\right),\theta\right)$$

General linear models are naturally suited to exchangeability and hierarchical models.

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

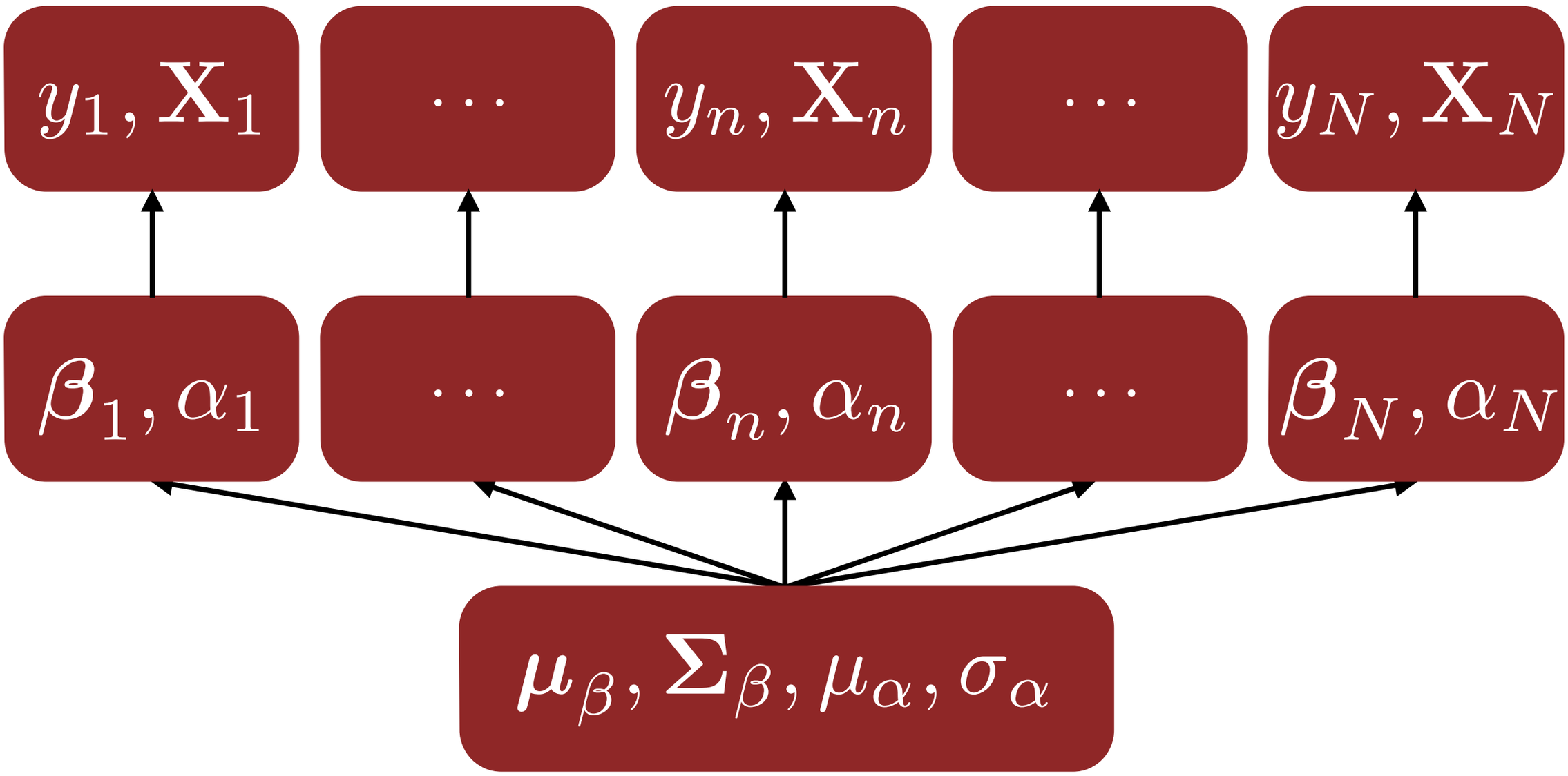
General linear models are naturally suited to exchangeability and hierarchical models.

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

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

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

General linear models are naturally suited to exchangeability and hierarchical models.



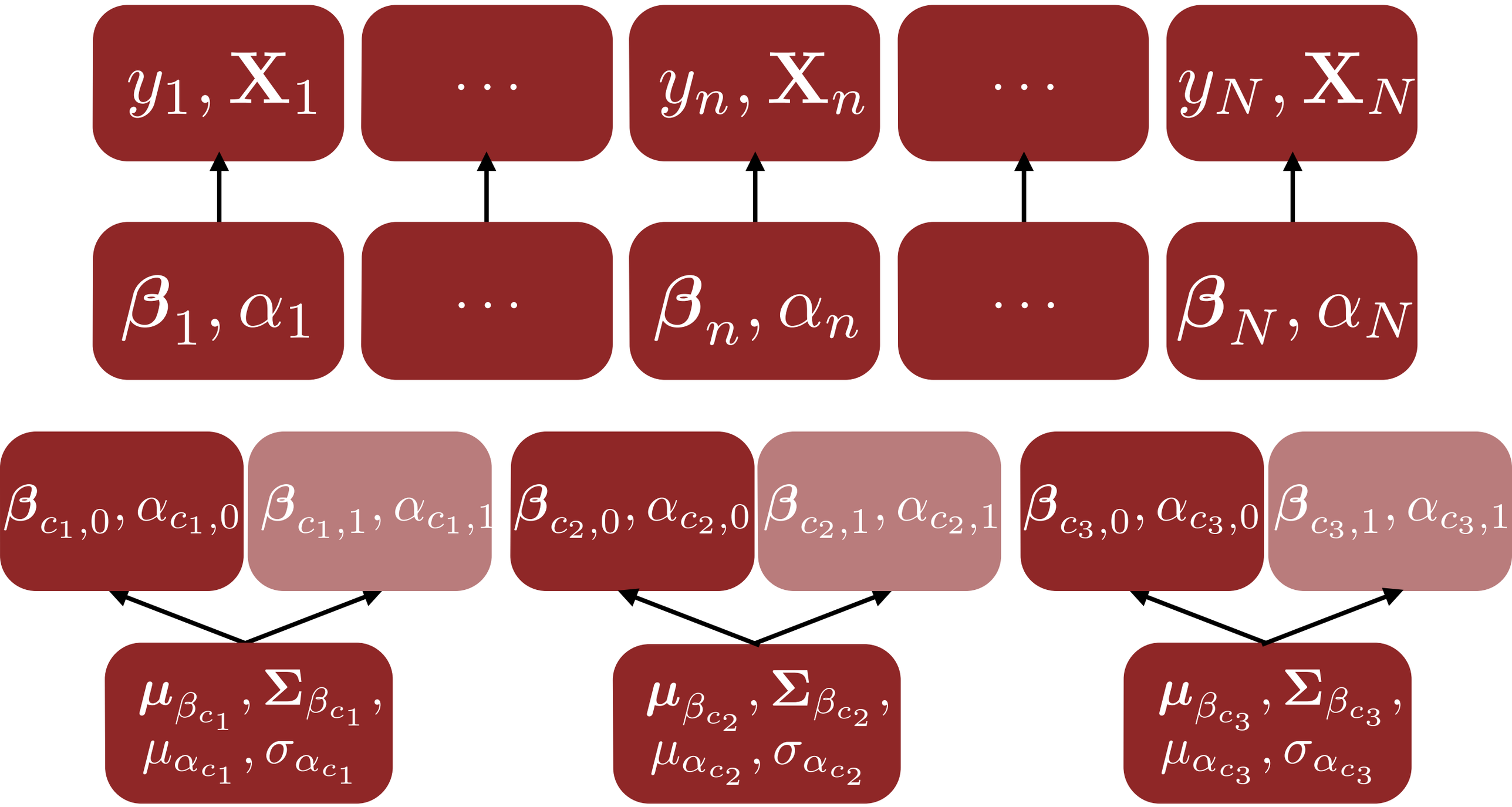
In a *multilevel* model the linear parameters vary across multiple hierarchies.

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

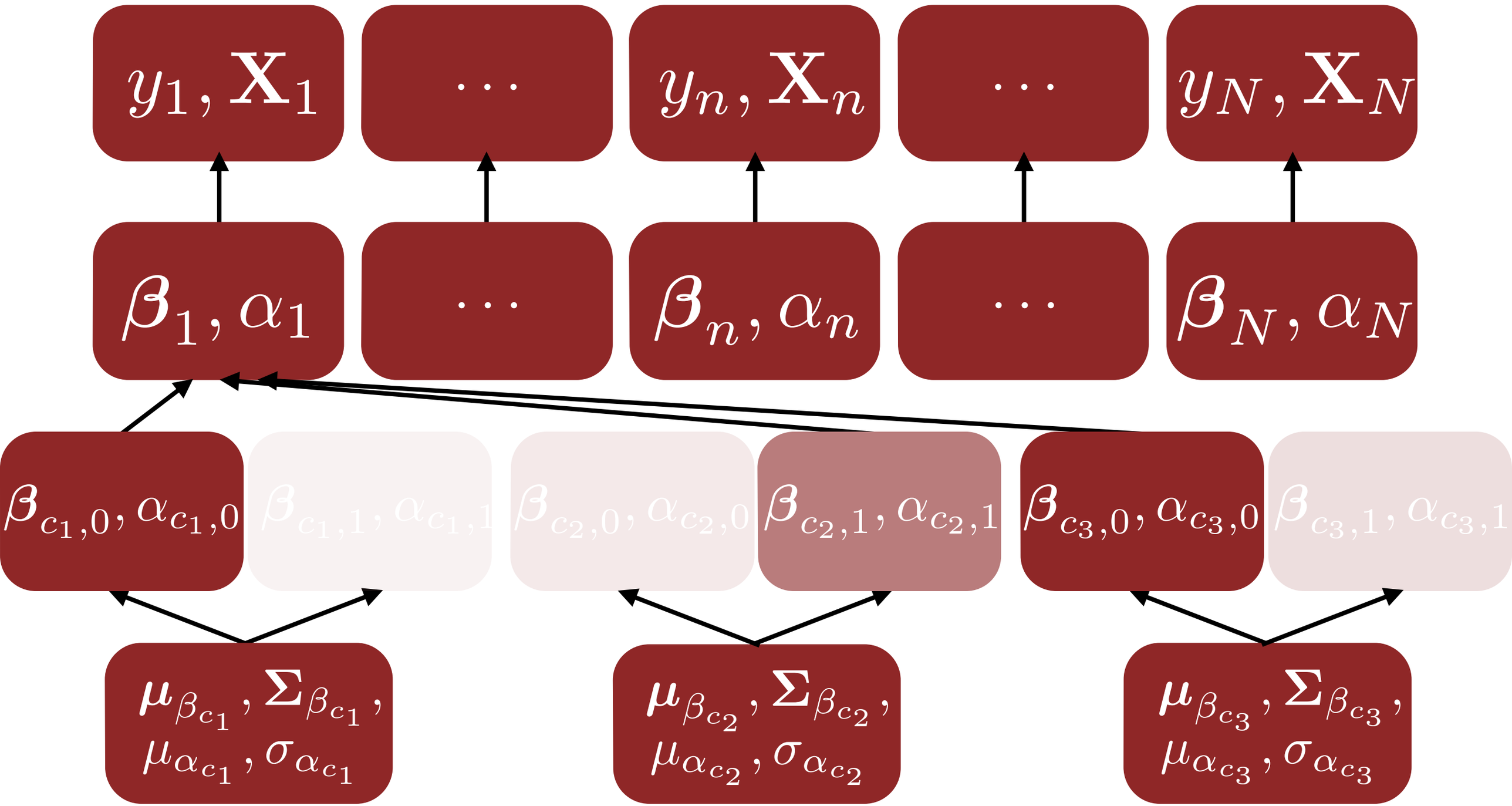
$$\alpha_n = \sum_{k=1}^{N_{\text{groups}}} \alpha_{k,j(n)}$$

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

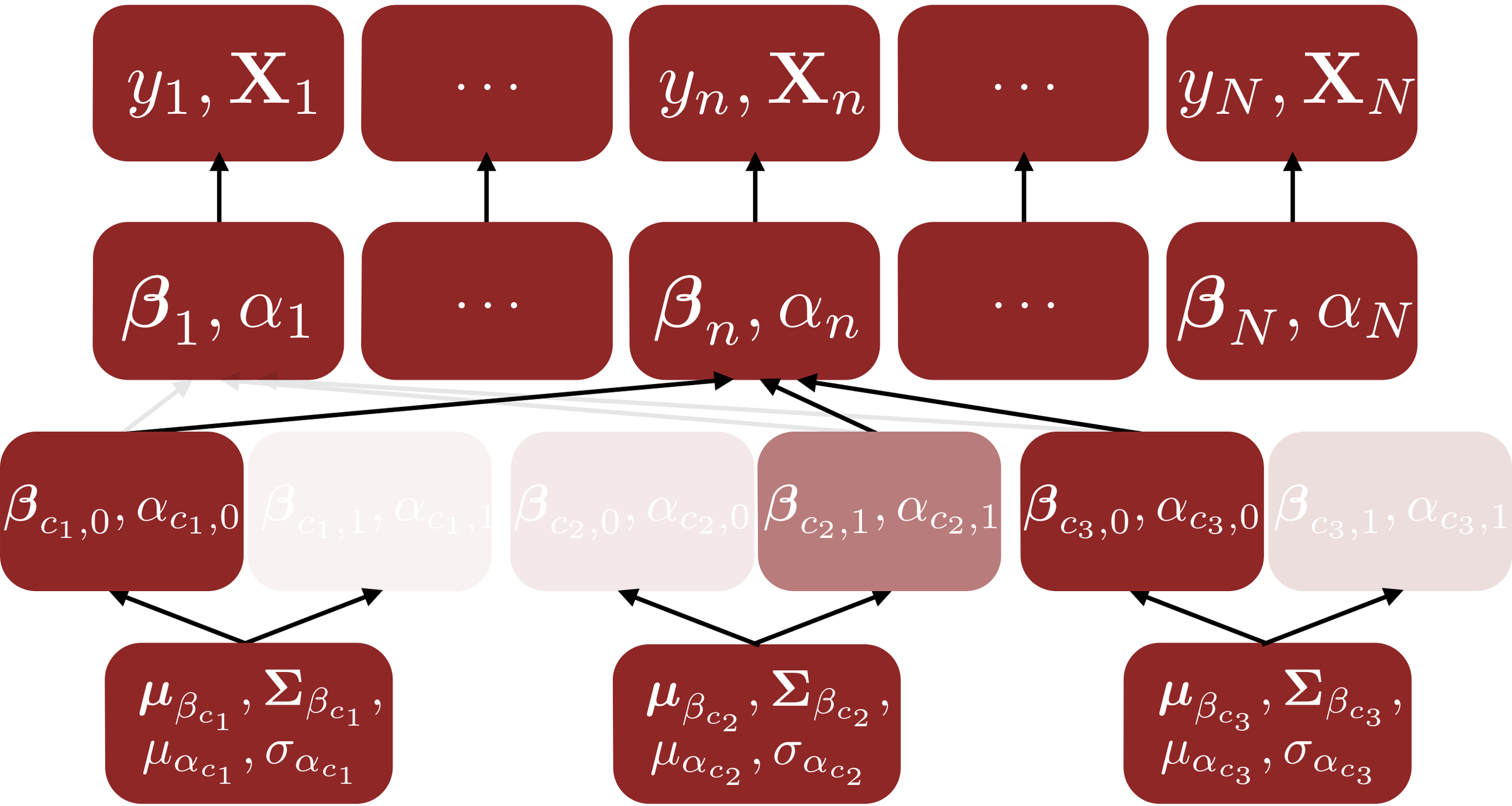
In a *multilevel* model the linear parameters vary across multiple hierarchies.



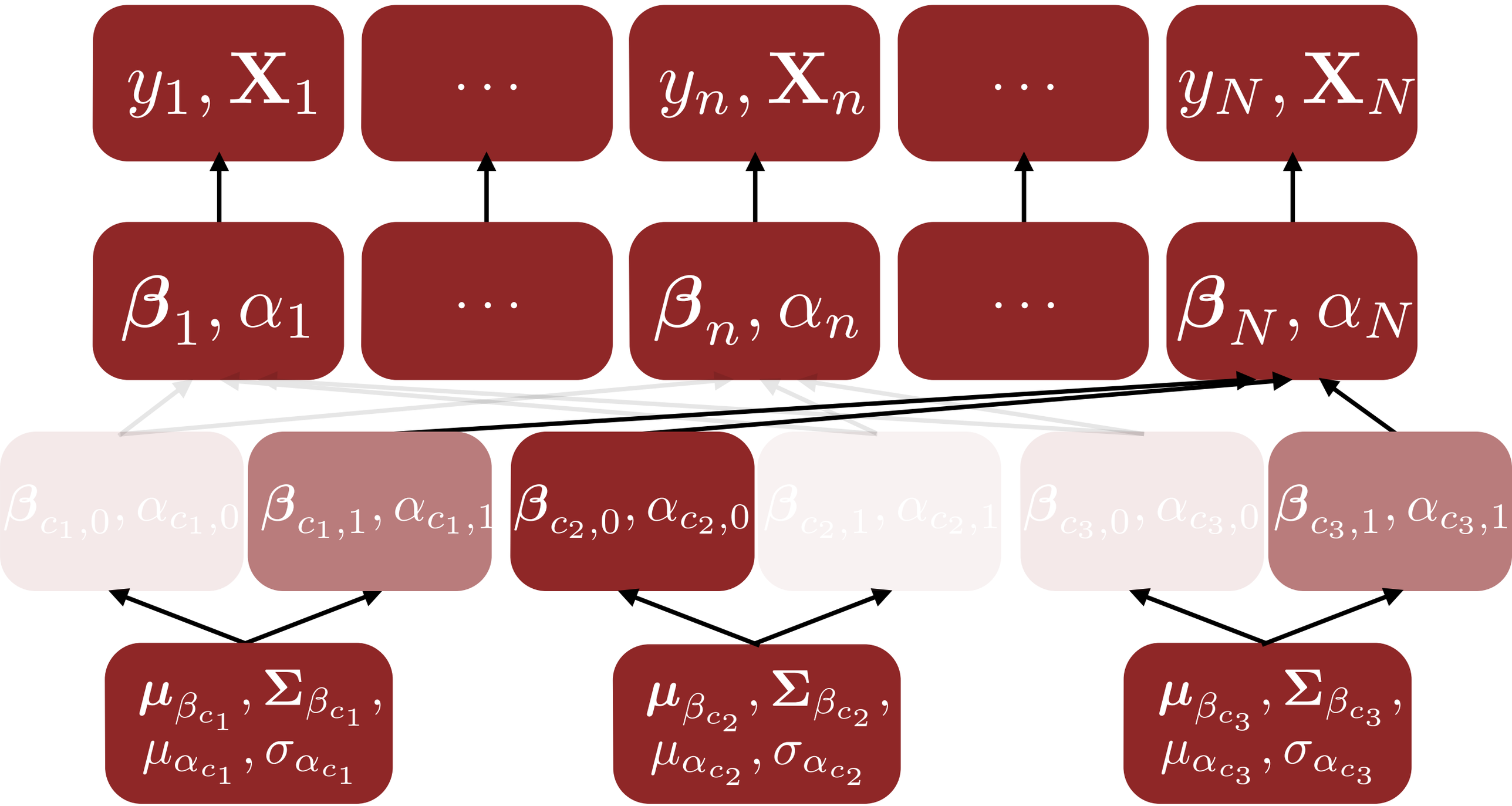
In a *multilevel* model the linear parameters vary across multiple hierarchies.



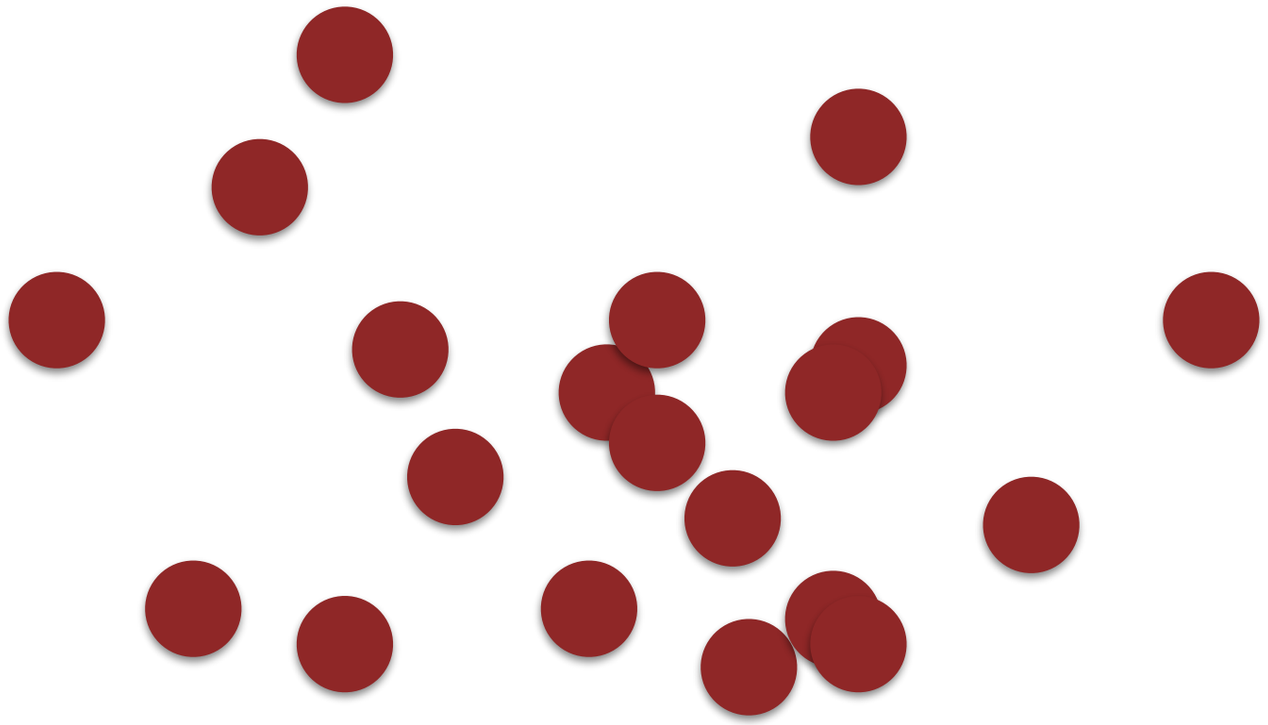
In a *multilevel* model the linear parameters vary across multiple hierarchies.



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

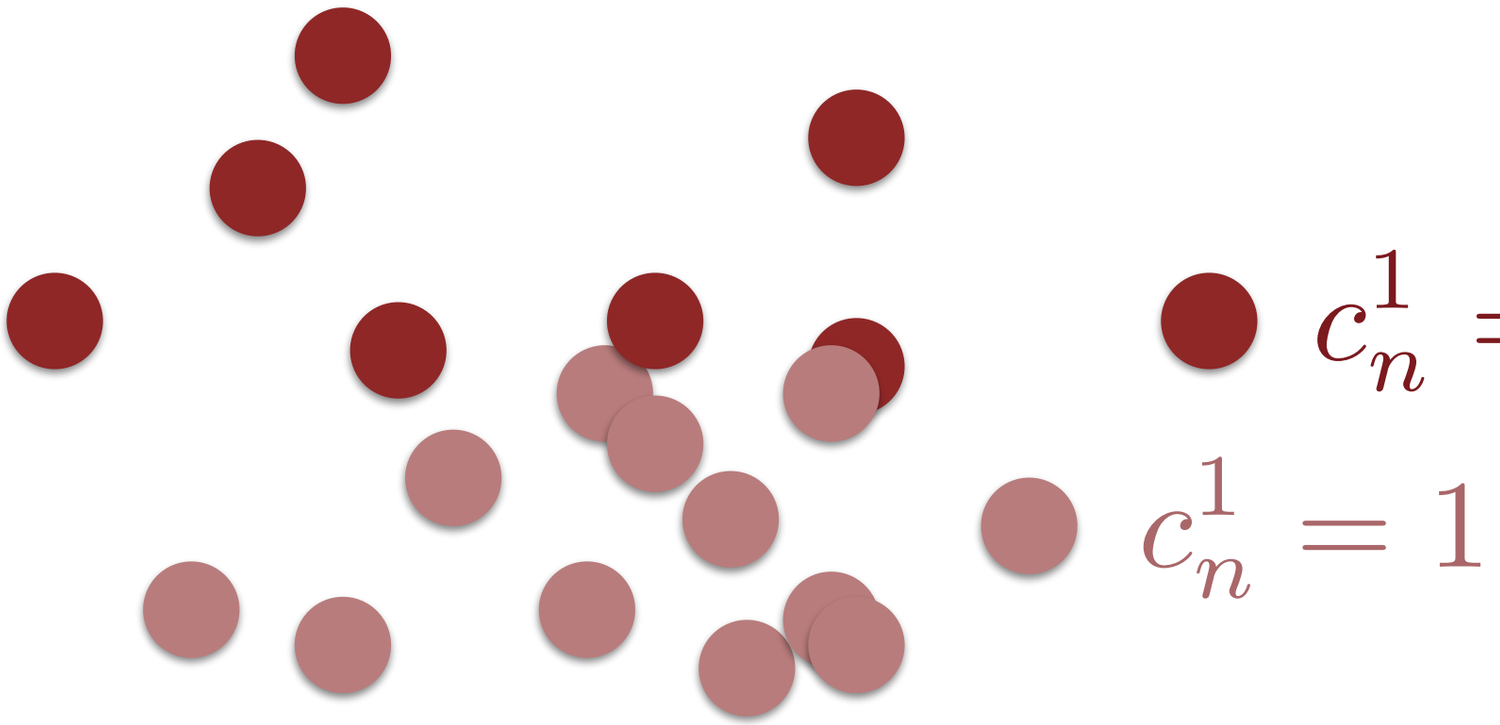


Each level contributes its own slope and intercept.



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

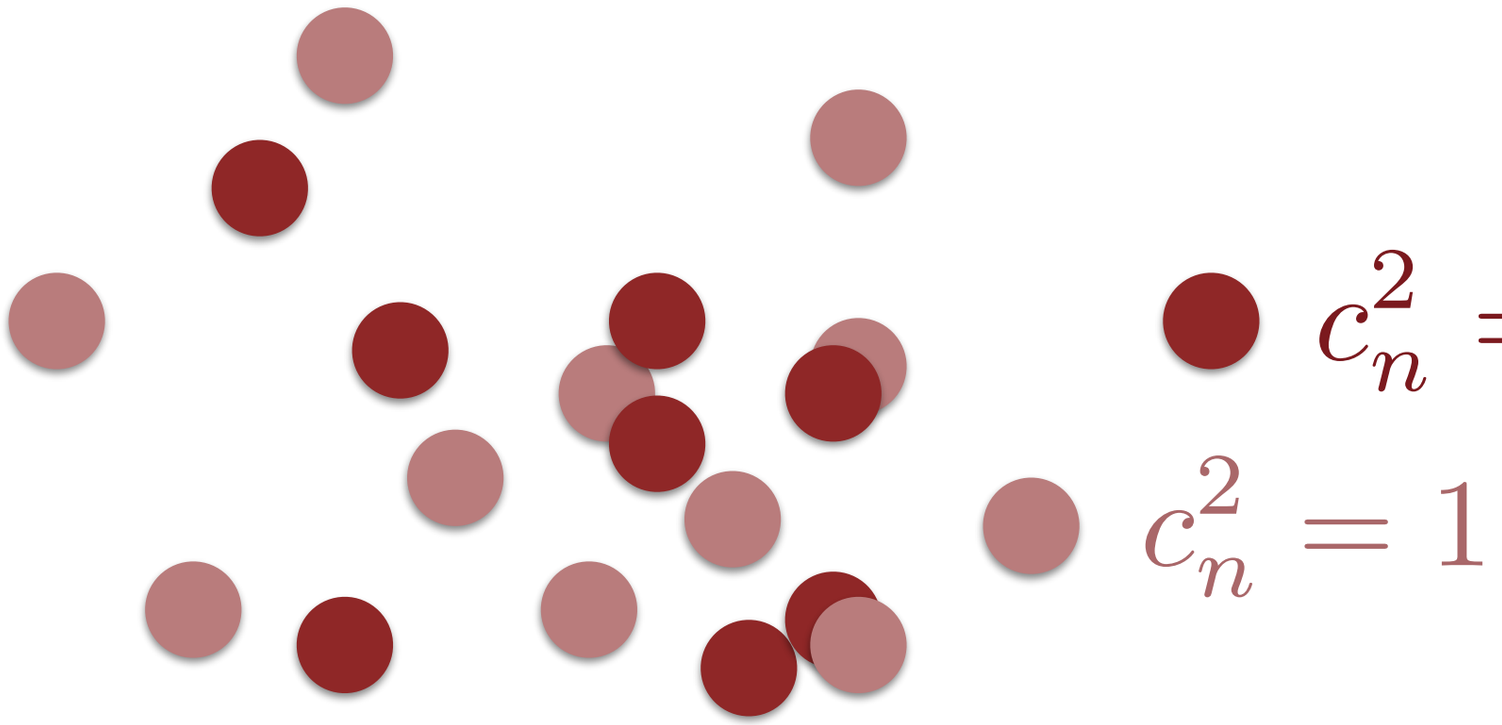
Each level contributes its own slope and intercept.



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

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

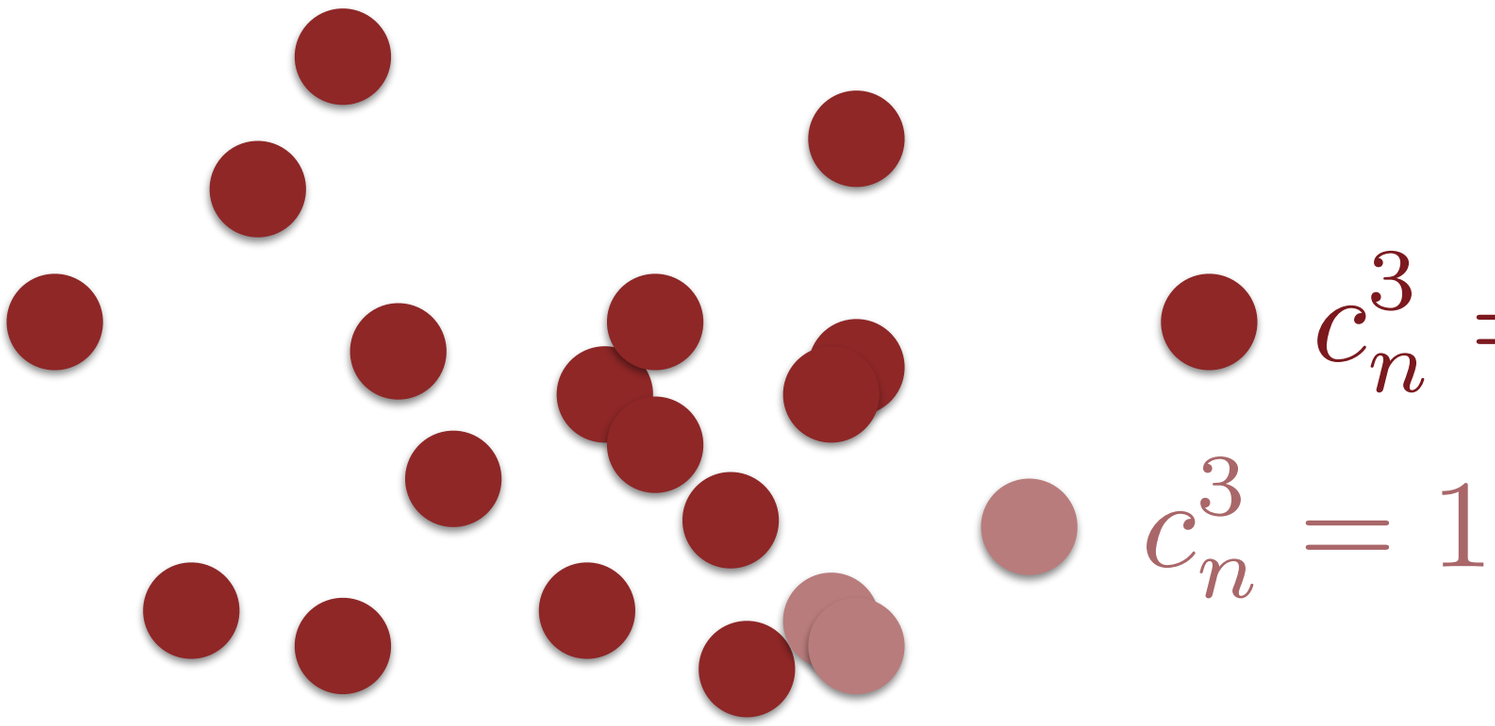
Each level contributes its own slope and intercept.



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

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

Each level contributes its own slope and 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 intercept.

$c_n^1 = 0$

$c_n^2 = 1$

$c_n^3 = 0$

$\alpha_n =$

Each level contributes its own slope and 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 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 intercept.

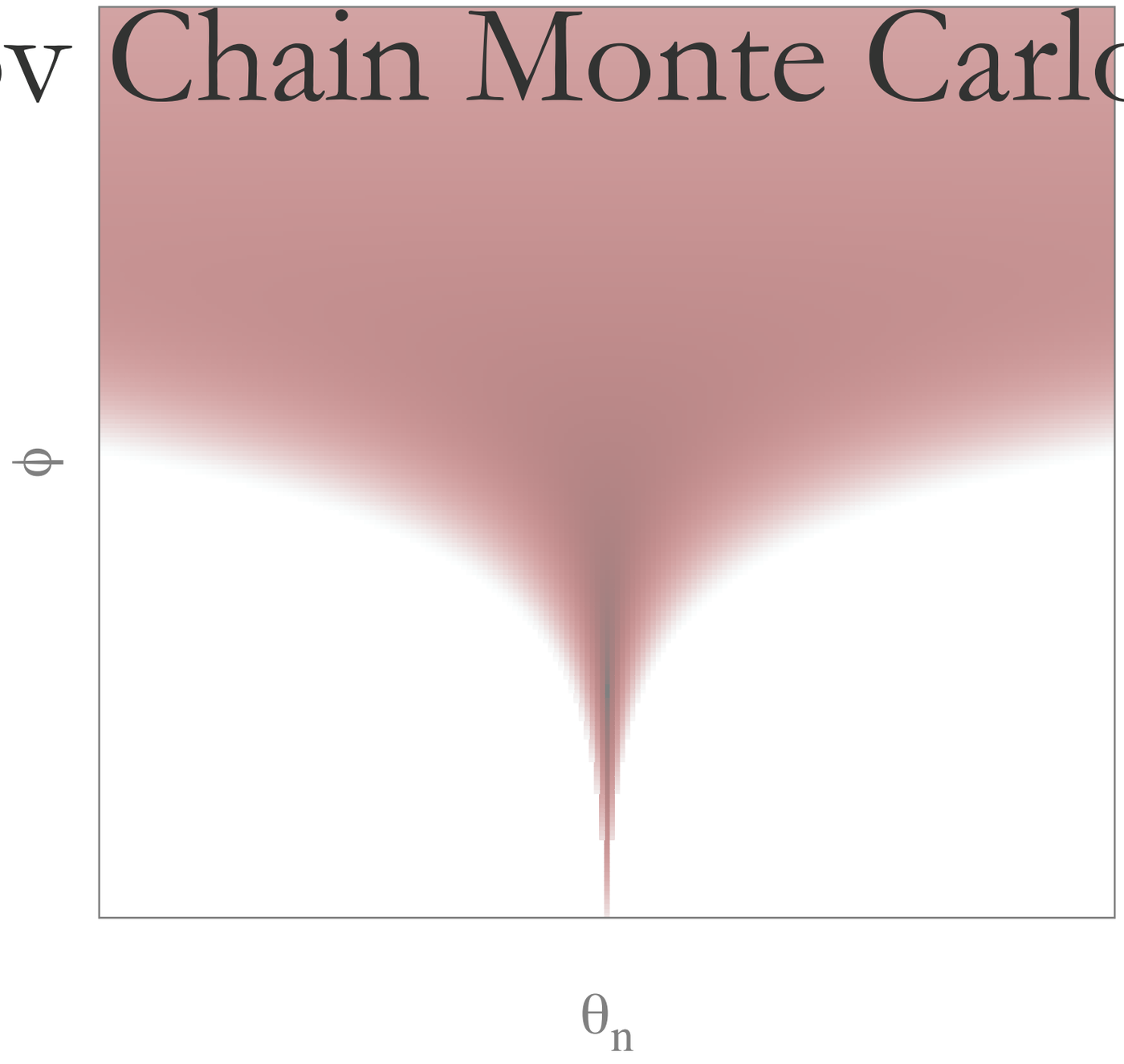
$c_n^1 = 0$

$c_n^2 = 1$

$c_n^3 = 0$

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

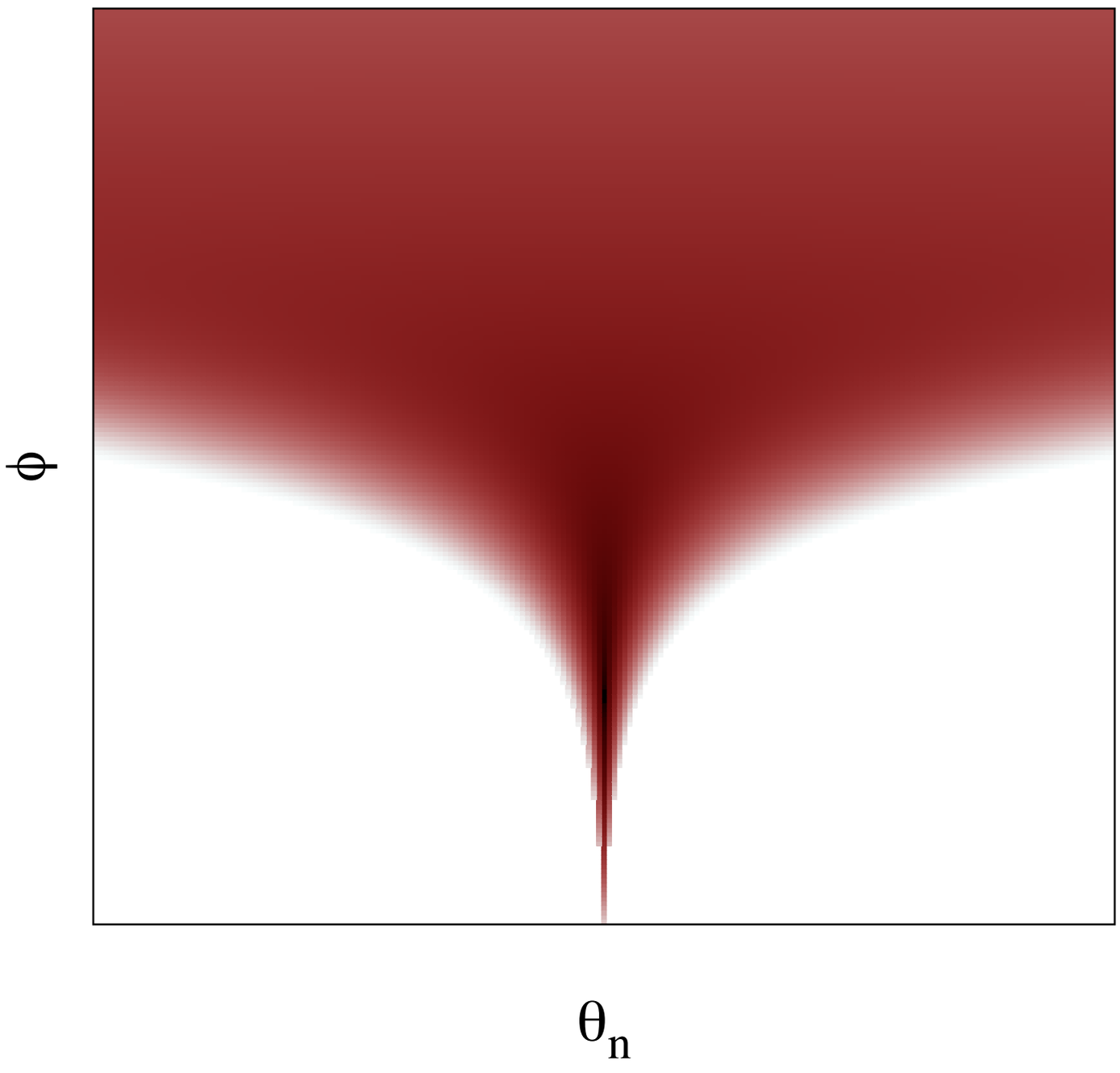
General Linear Hierarchical Models and Markov Chain Monte Carlo



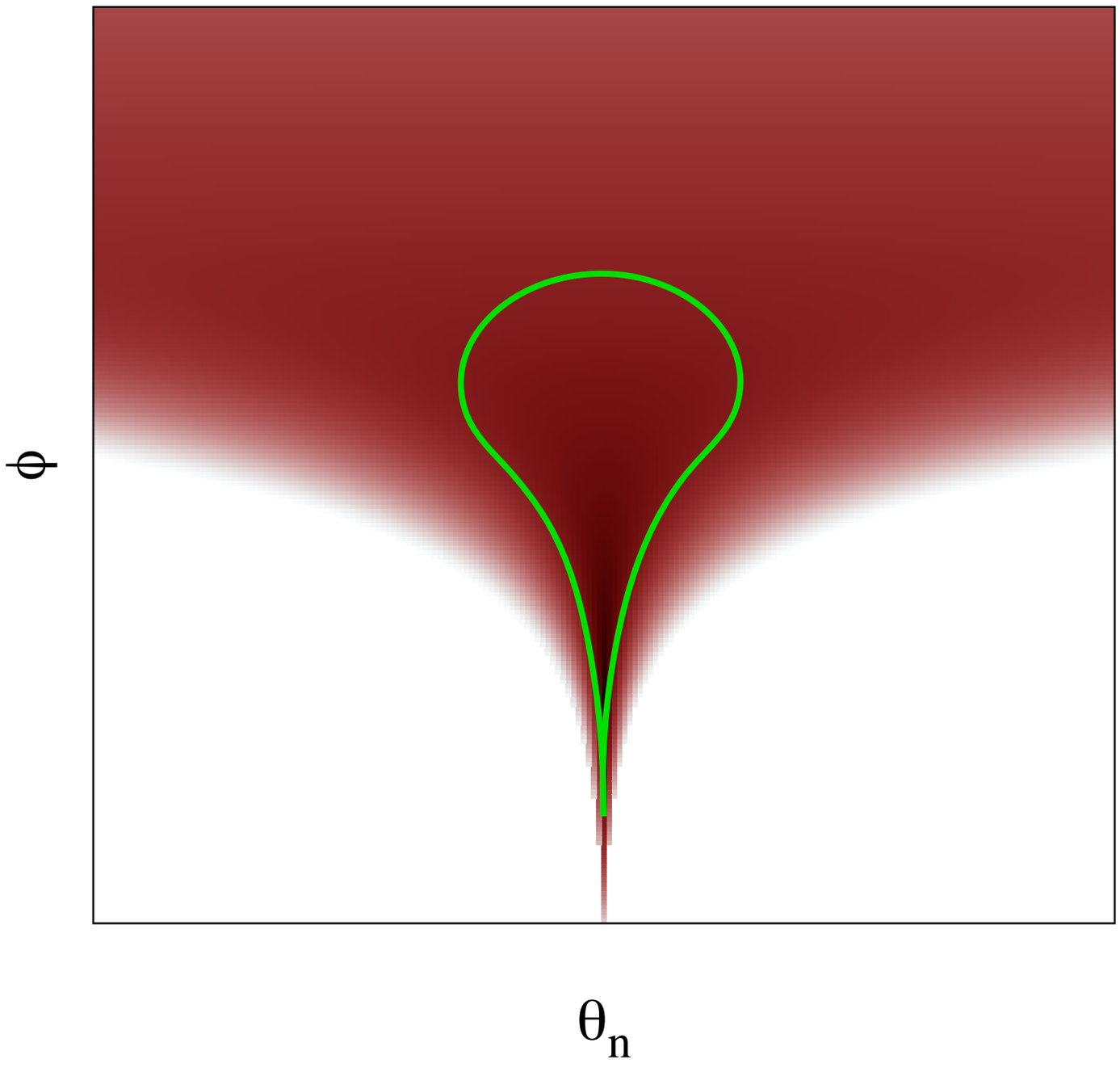
Neal's funnel distribution illustrates some of the problems with sampling from hierarchical distributions.

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

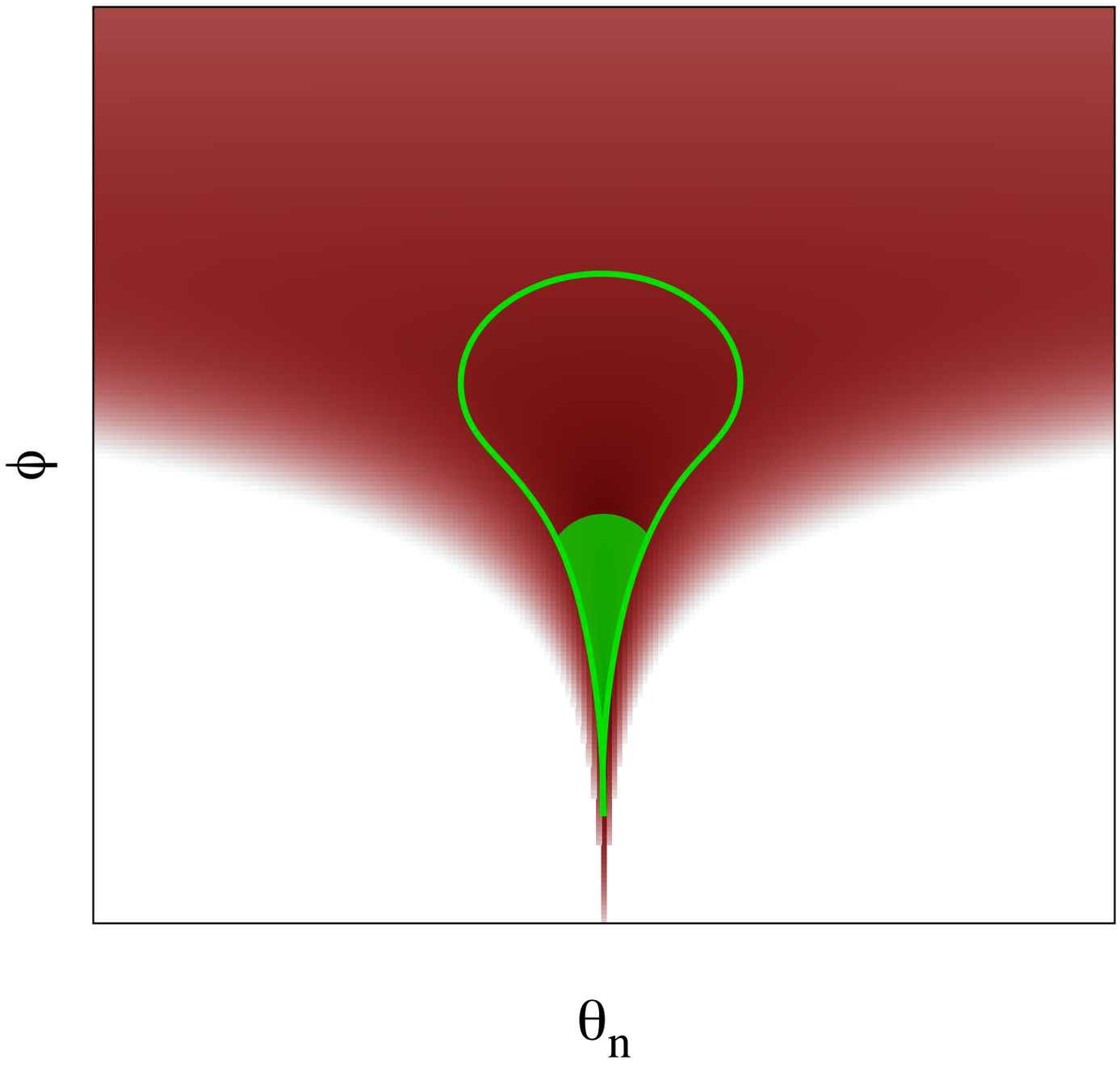
Neal's funnel distribution illustrates some of the problems with sampling from hierarchical priors.



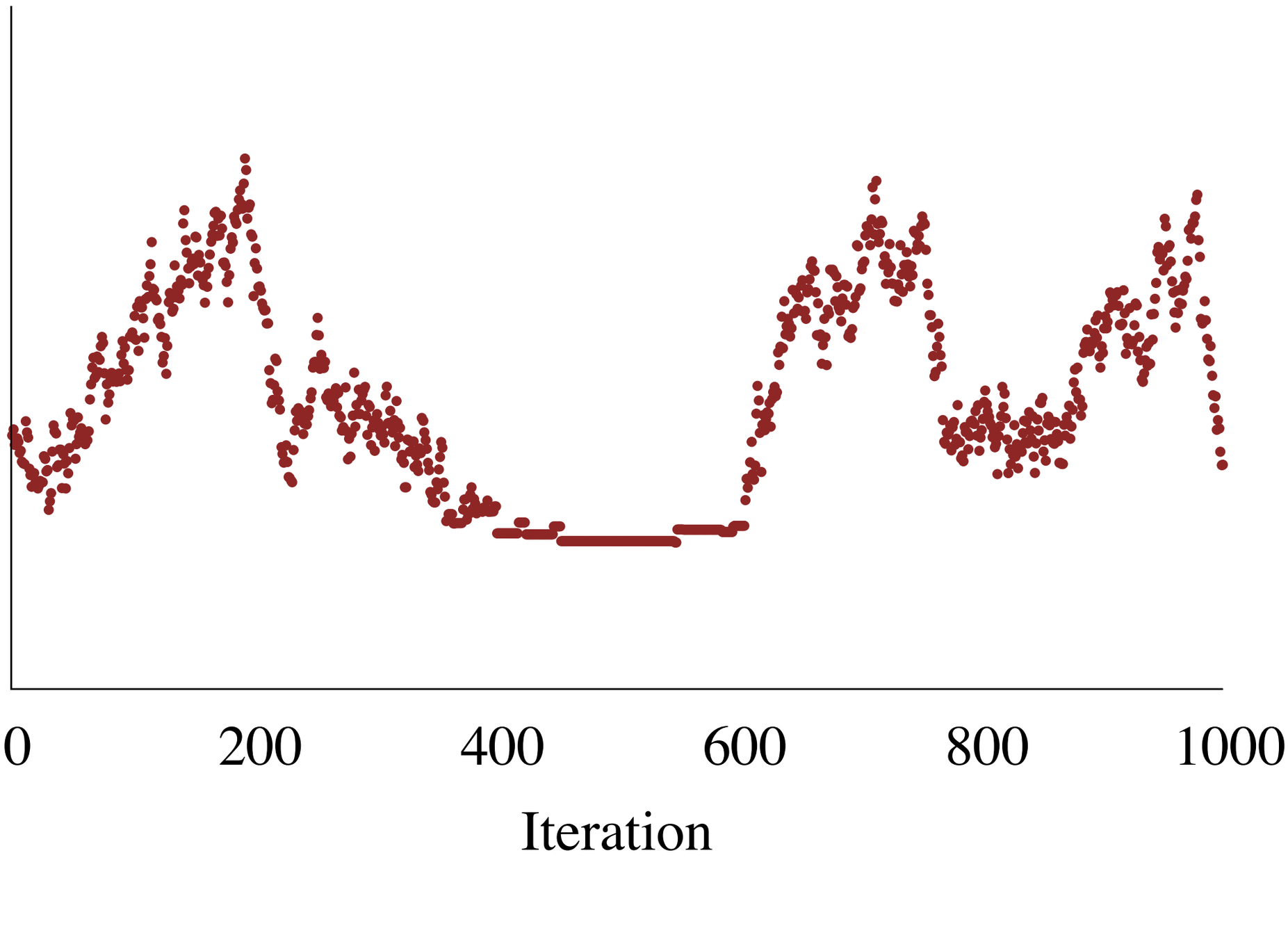
Neal's funnel distribution illustrates some of the problems with sampling from hierarchical priors.



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.

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

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

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

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

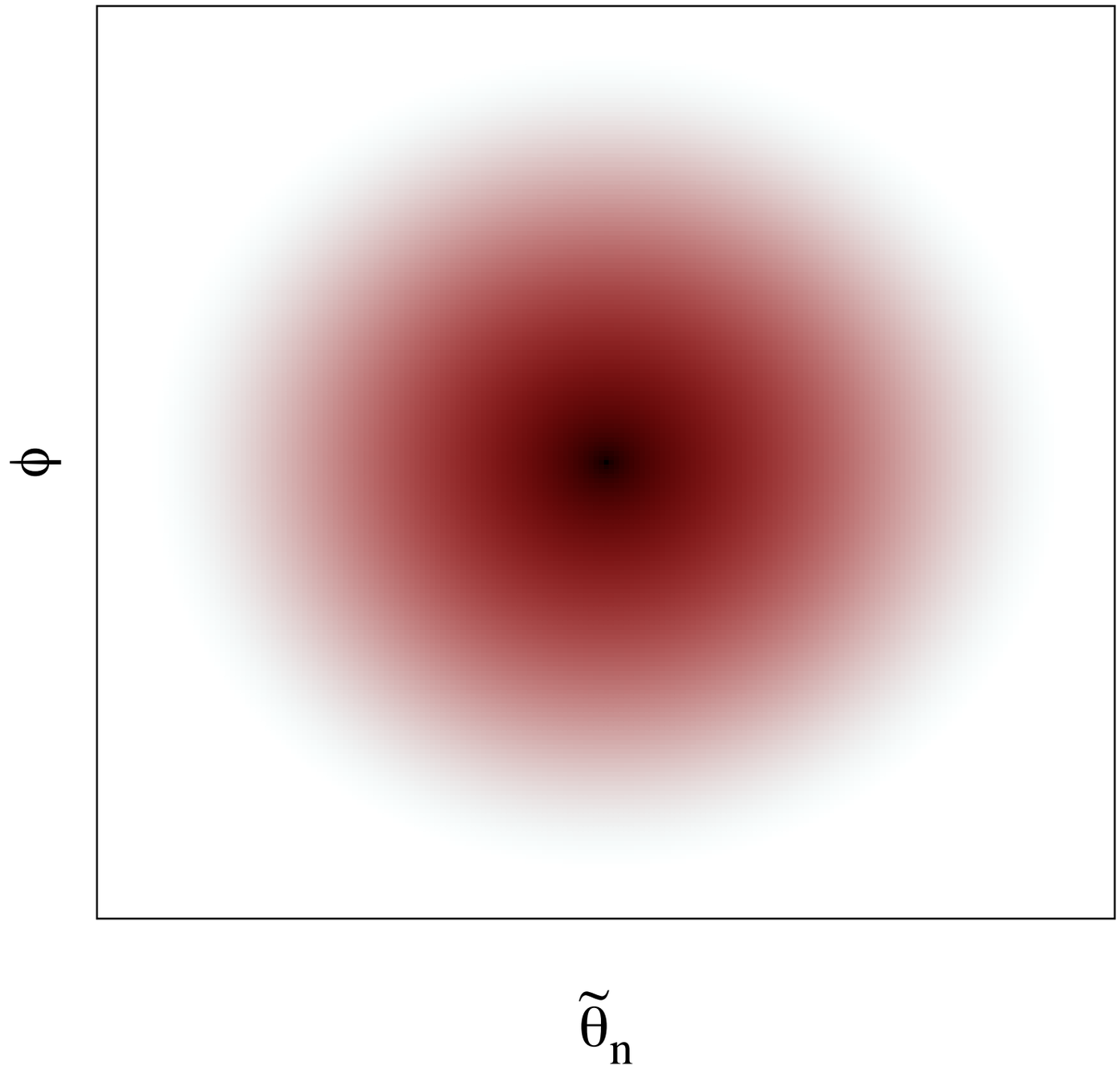
The geometry of the funnel, however, dramatically improves when we sample from auxiliary parameters.

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

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

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

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



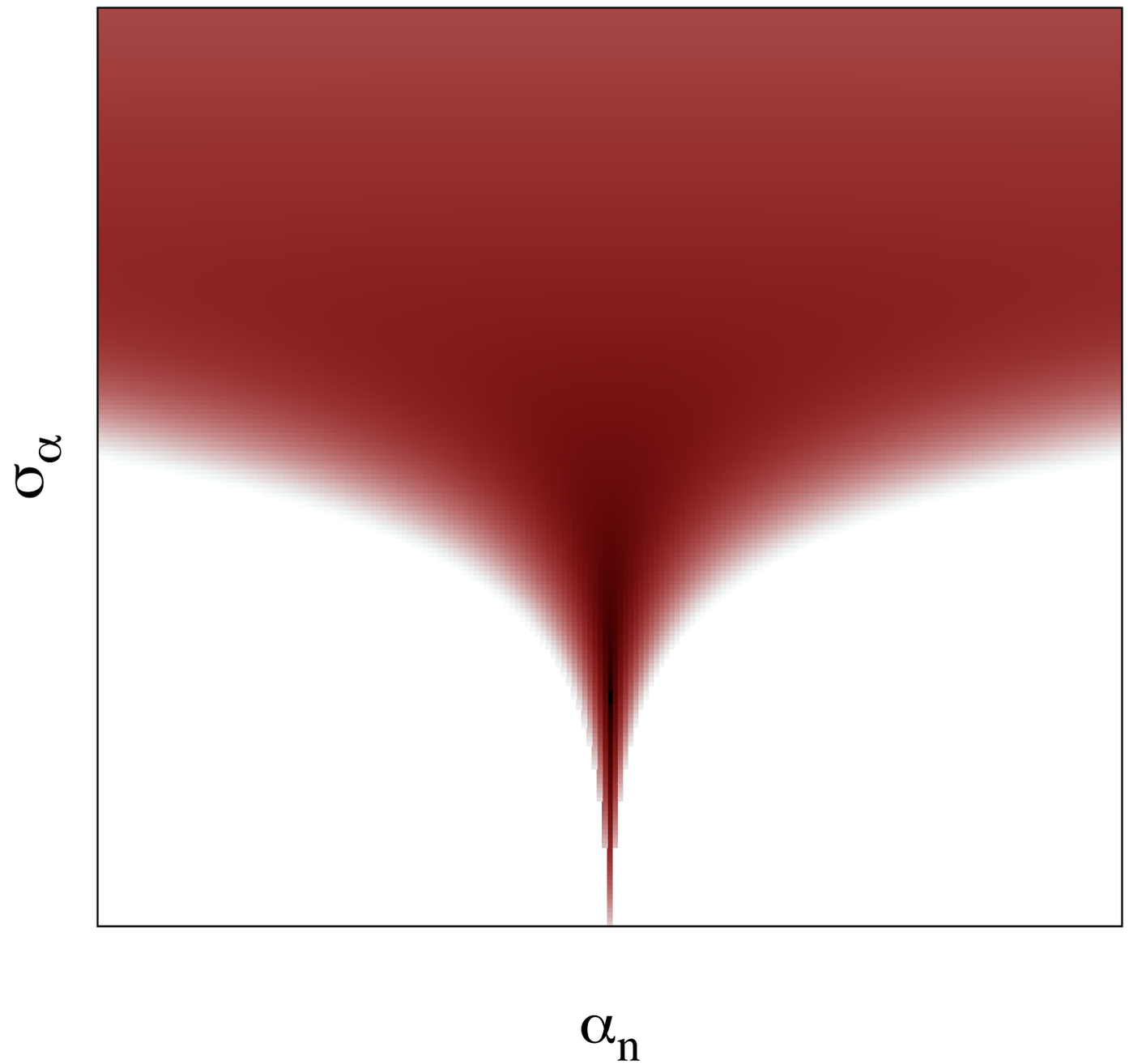
The most natural representation of a general linear hierarchical model is a *centered parameterization*.

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

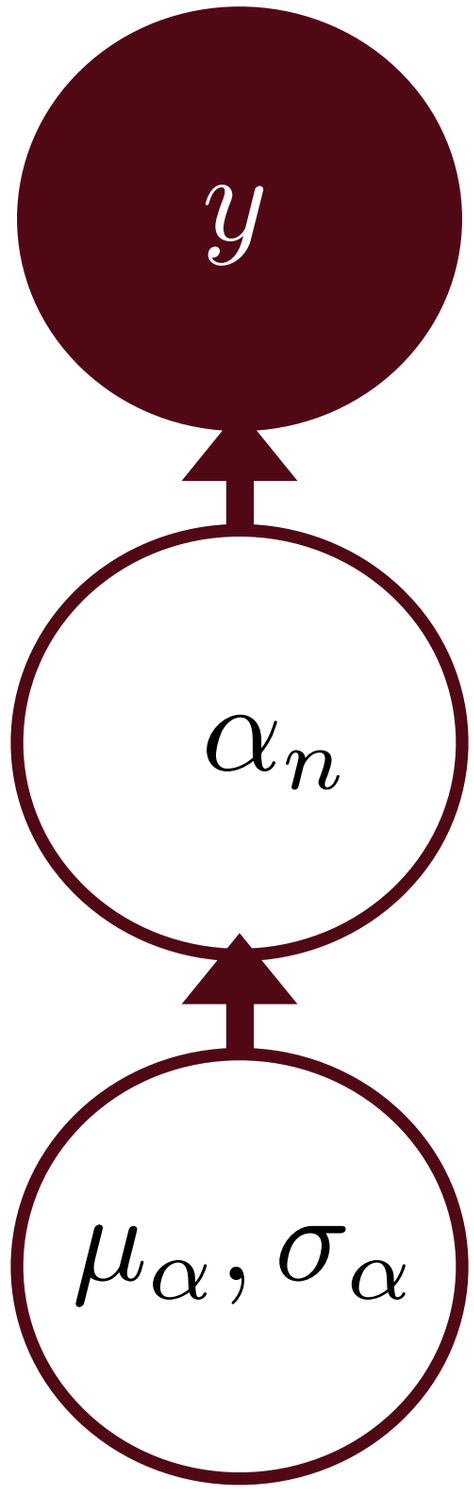
$$\boldsymbol{\beta}_n \sim \mathcal{N}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)$$

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

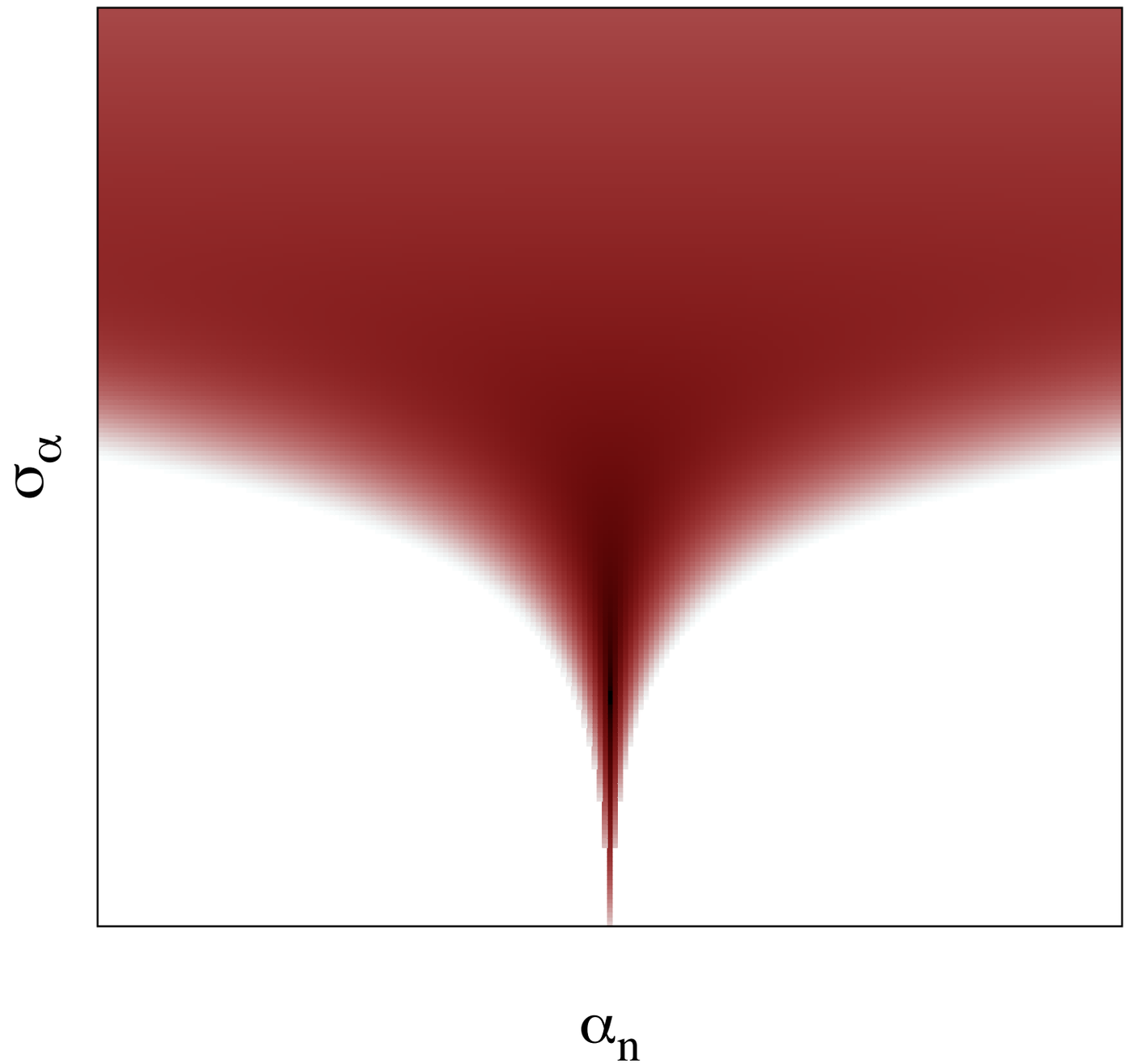
When data are sparse and uninformative, the centered parameterization of the posterior resembles the funnel.



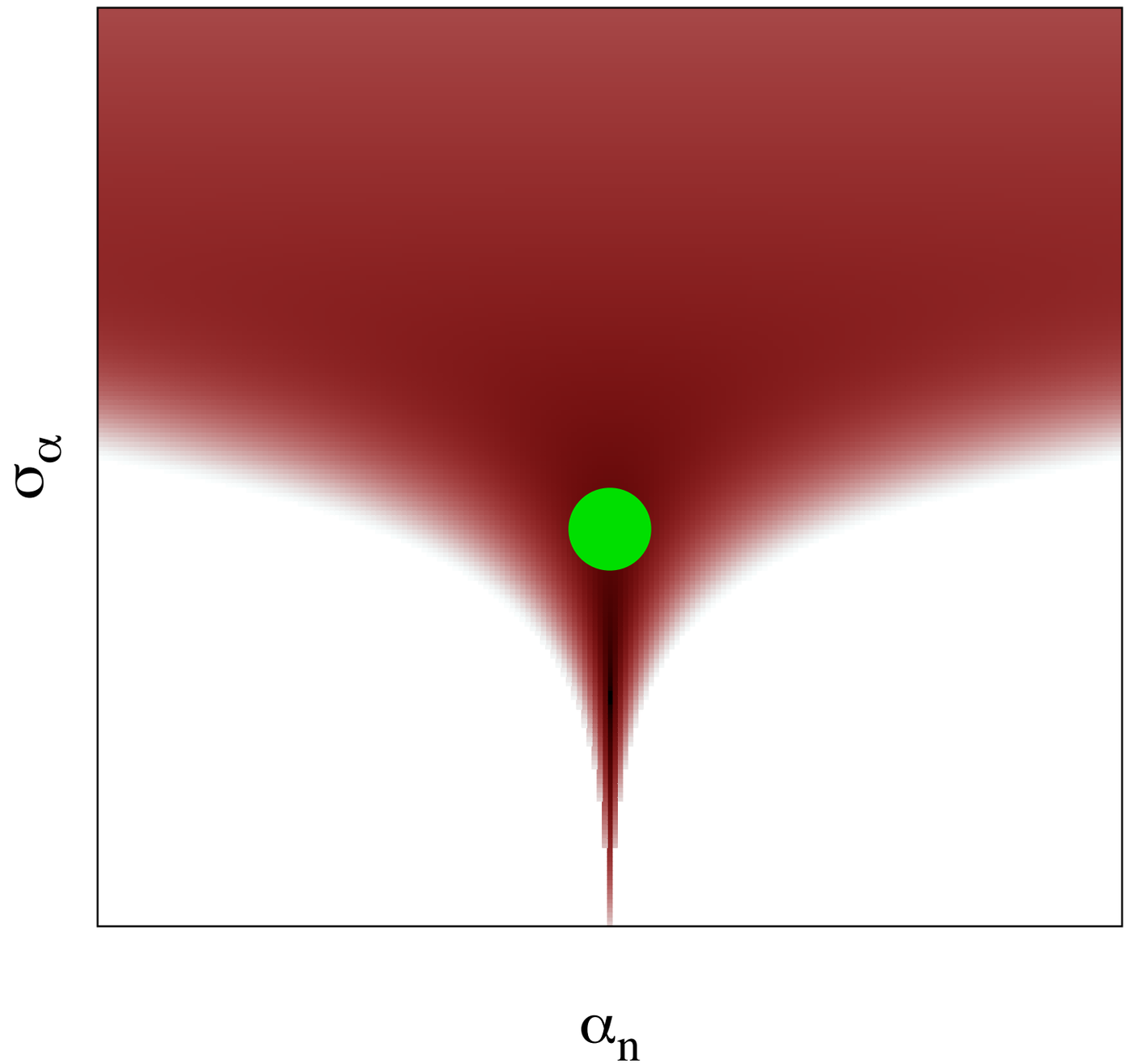
When data are sparse and uninformative, the centered parameterization of the posterior resembles the funnel.



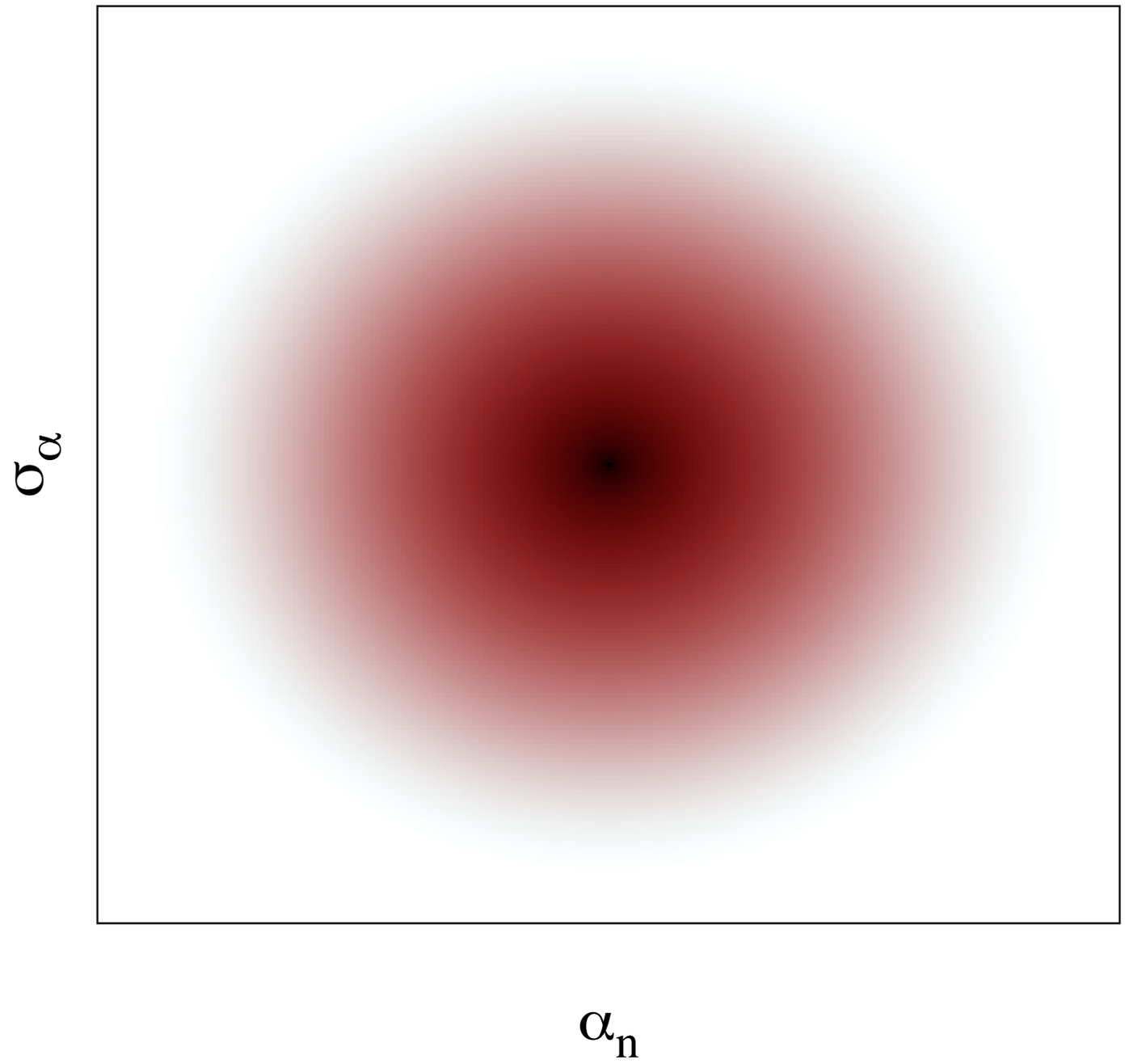
But when the data are more informative the posterior geometry is much nicer.



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.



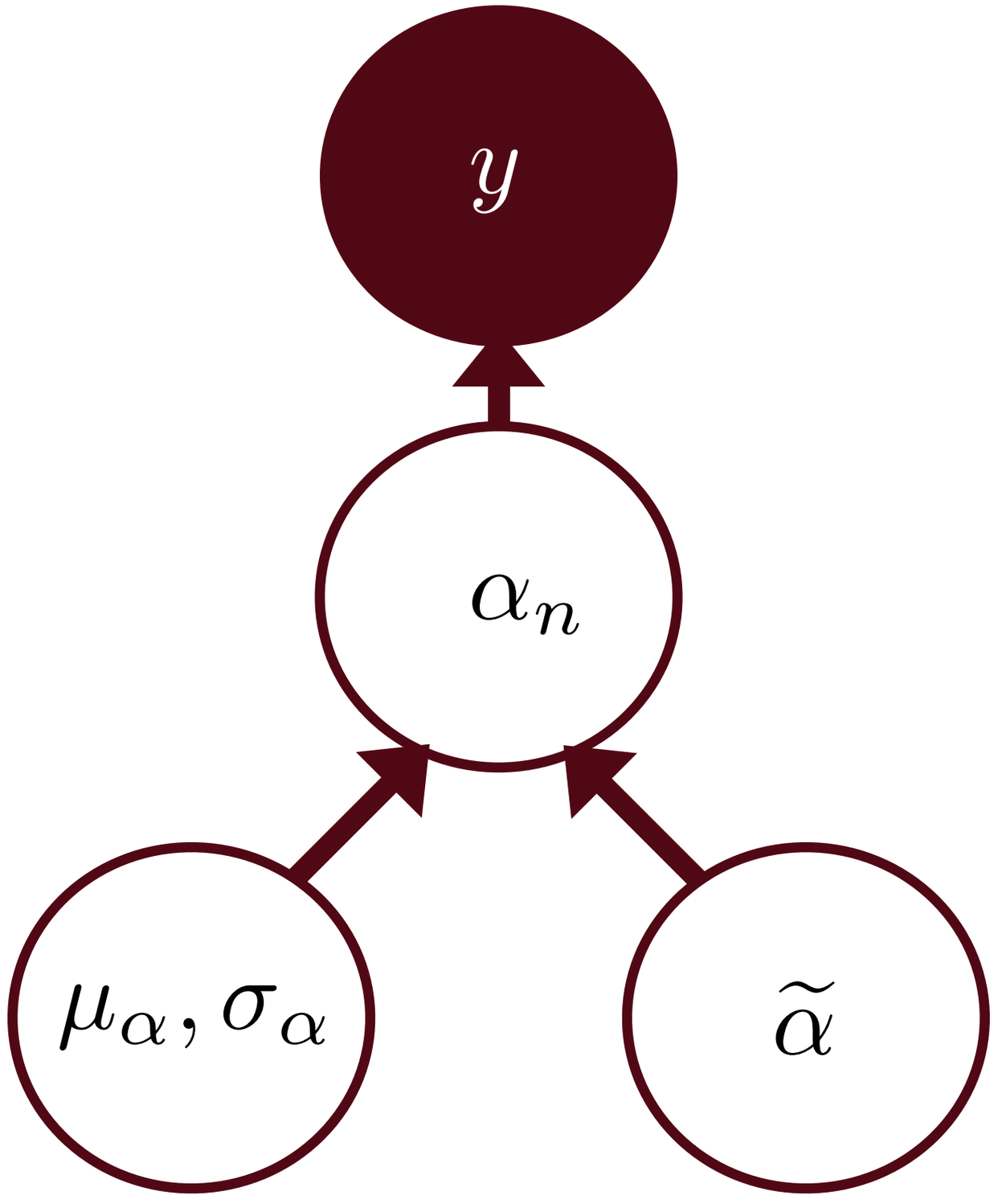
The same general linear hierarchical model can also be implemented with a *non-centered parameterization*.

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

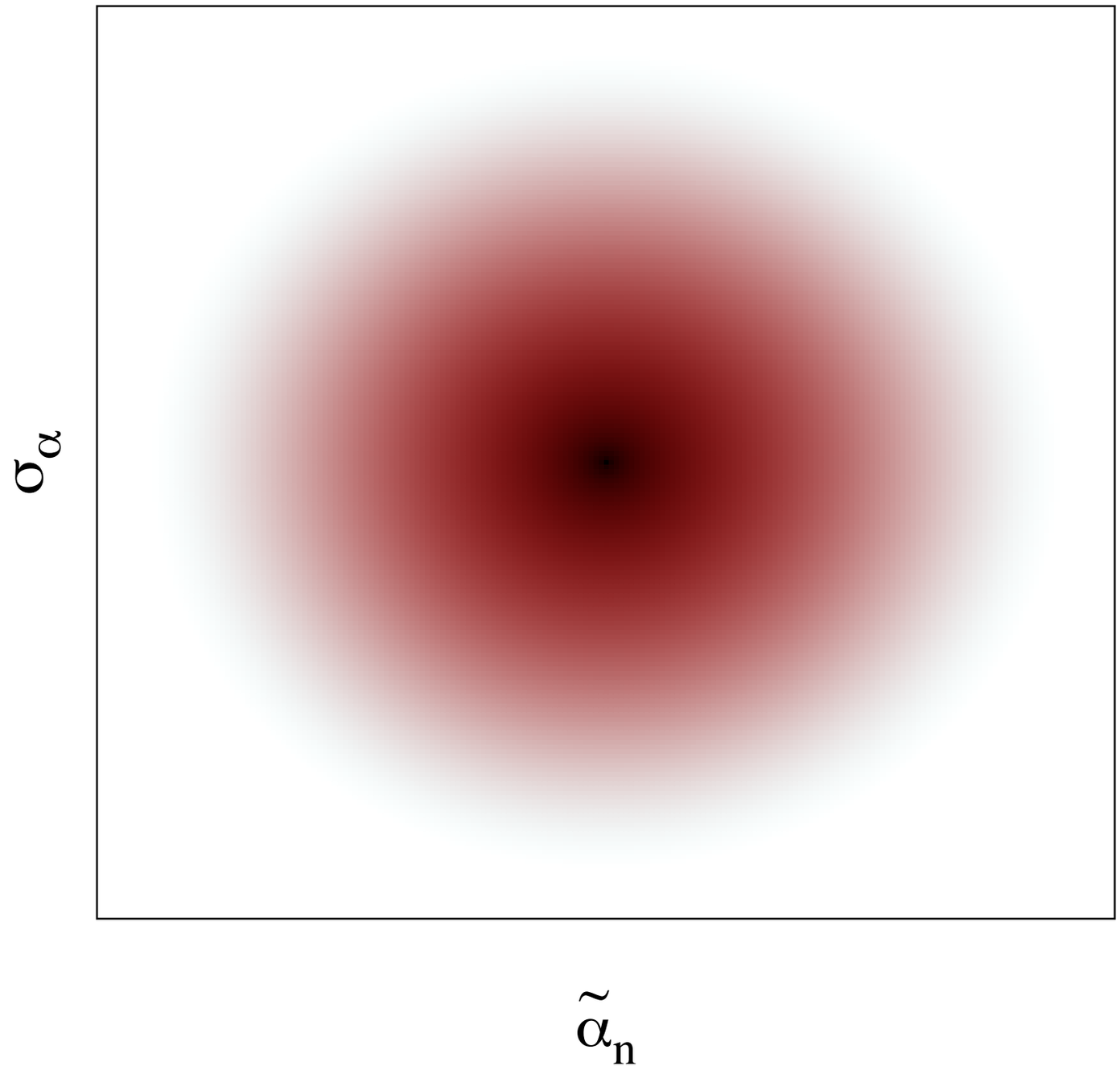
$$\tilde{\boldsymbol{\beta}}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\boldsymbol{\beta}_n = \boldsymbol{\mu}_\beta + \mathbf{L}\tilde{\boldsymbol{\beta}}_n,\ \mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}_\beta$$

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

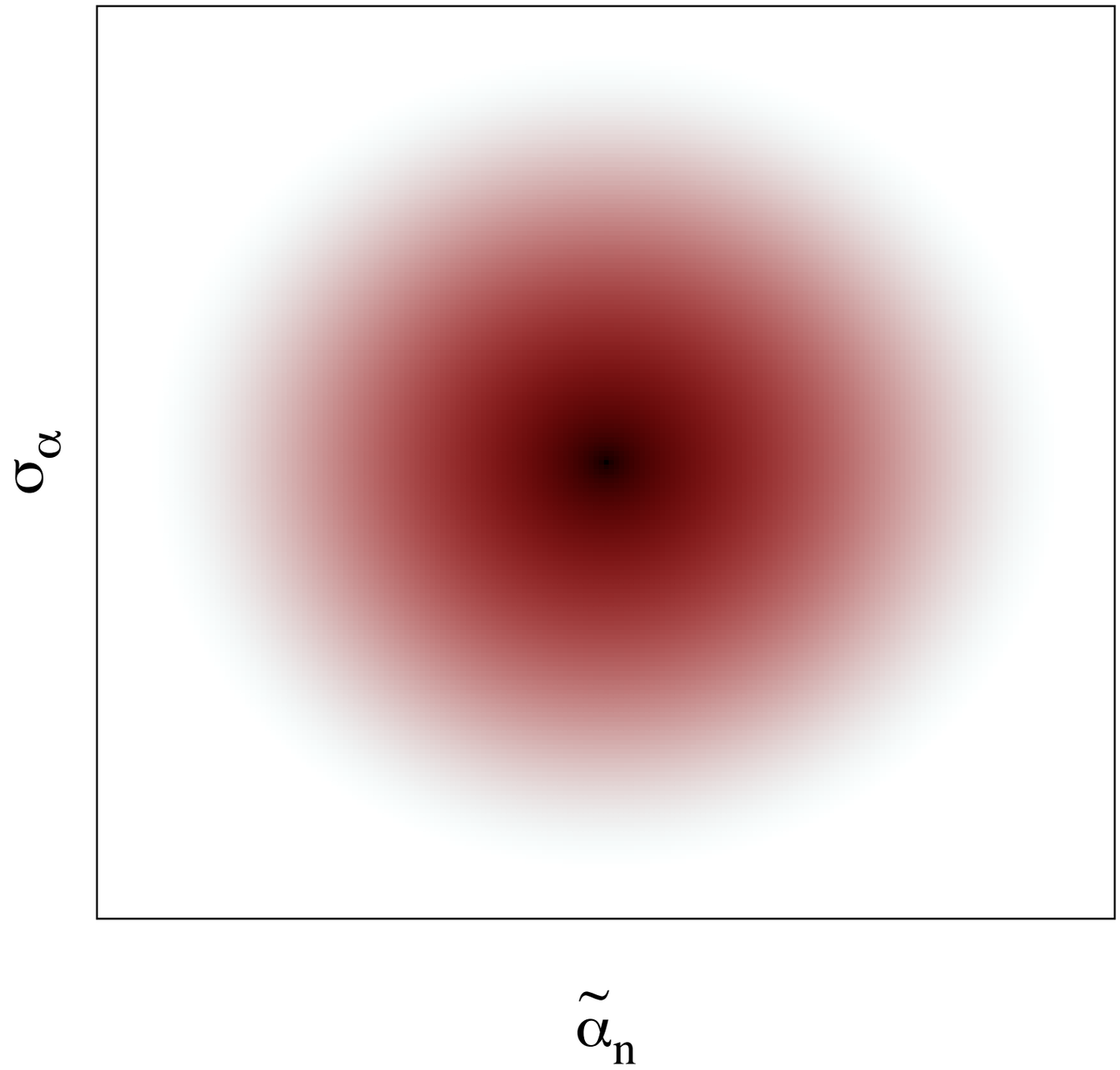
When data are sparse, the non-centered parameterization of the posterior yields a favorable geometry.



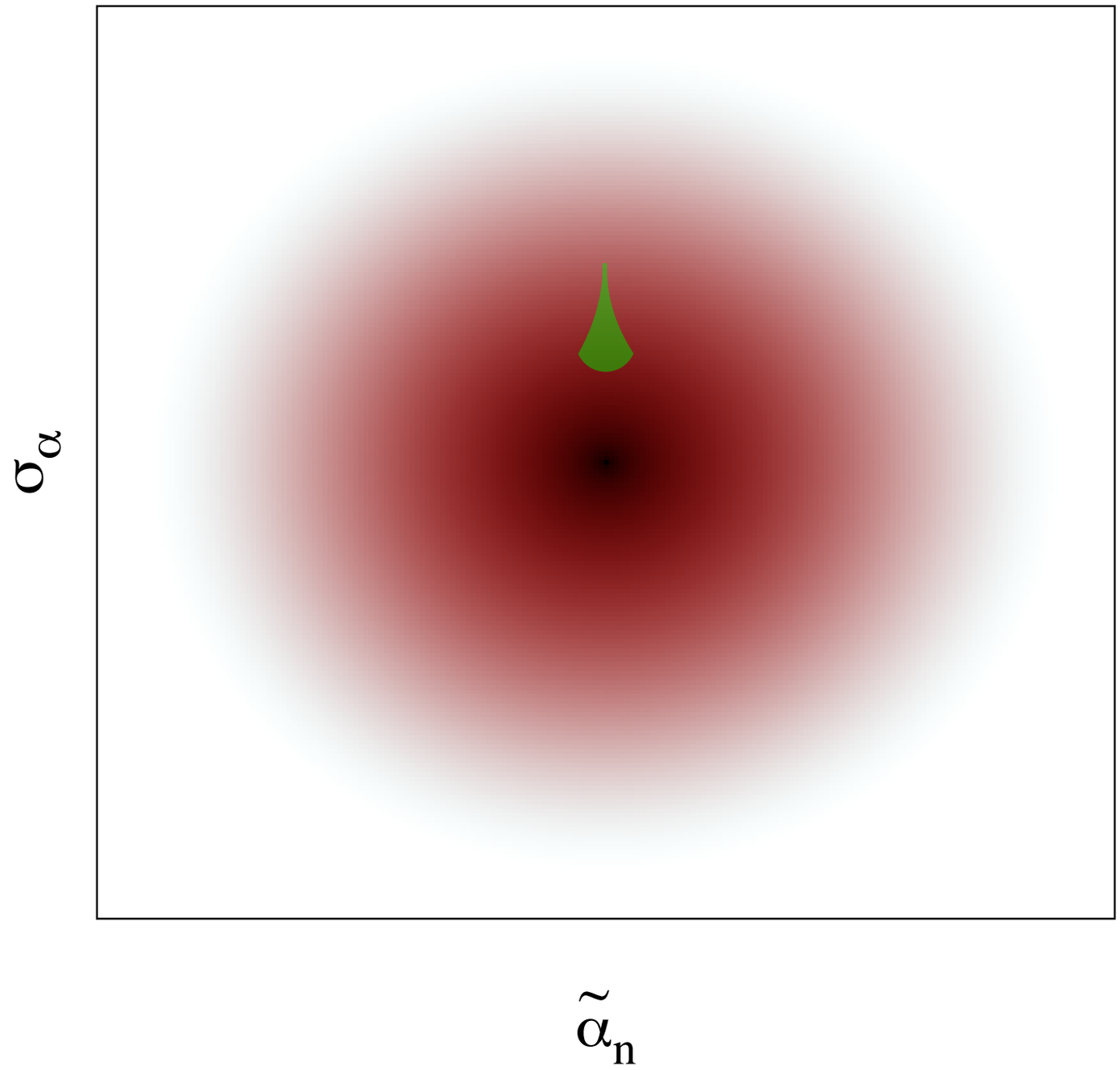
When data are sparse, the non-centered parameterization of the posterior yields a favorable geometry.



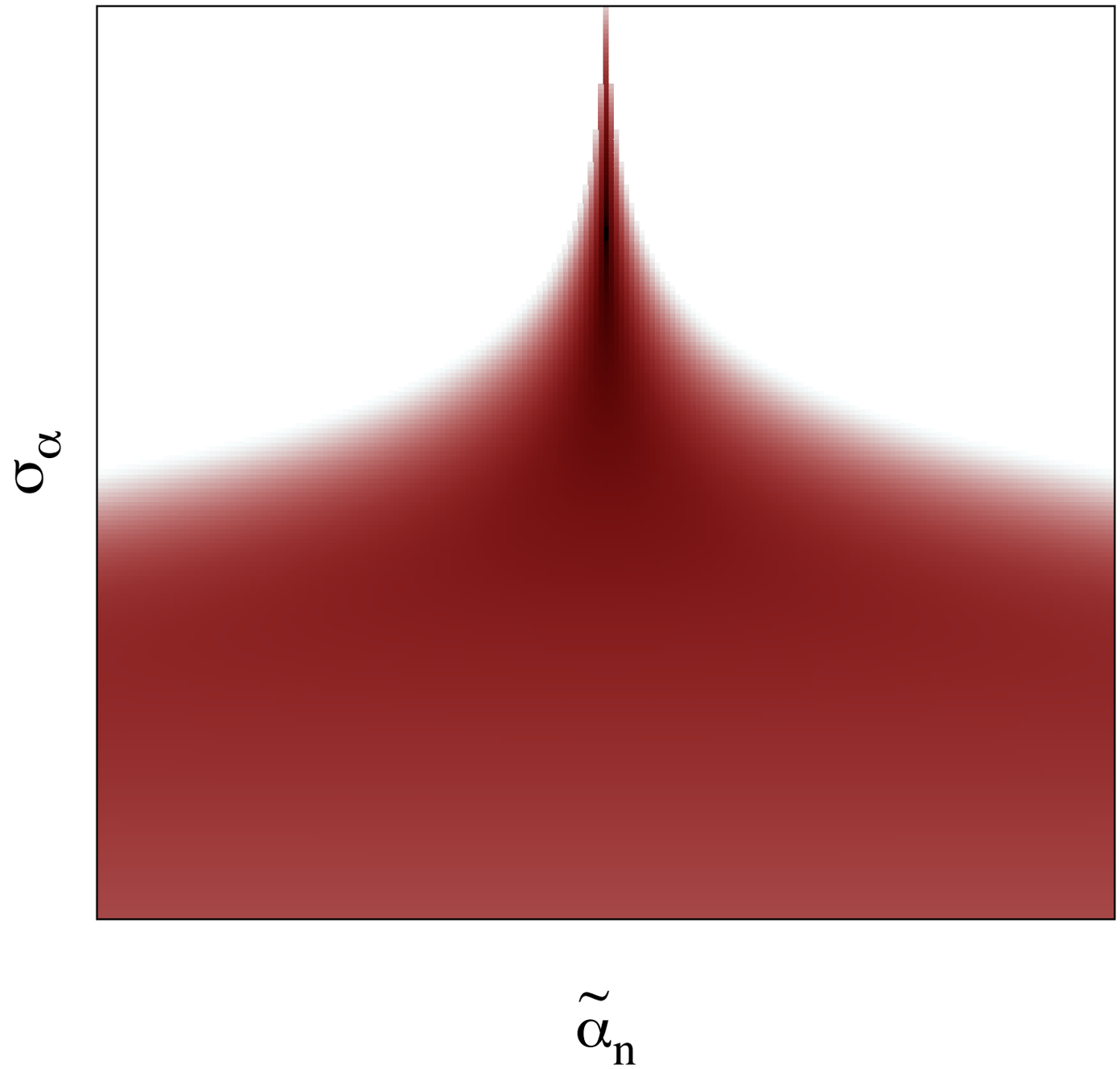
The constraints imposed by the data, however, are more complex, yielding pathological posteriors with more data.



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.

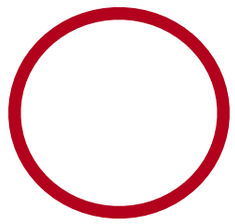
Centered Parameterization

Consequently the two parameterizations are each advantageous in different circumstances.

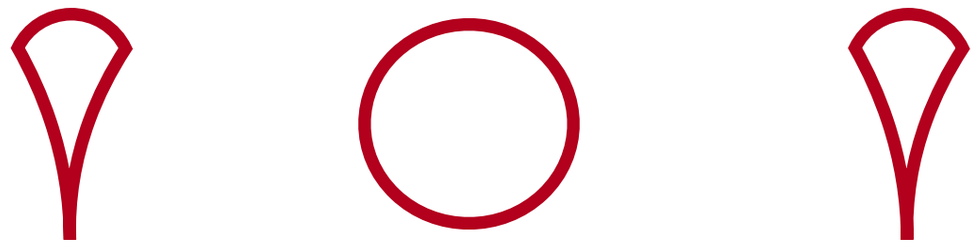
Hierarchical
Geometry

Centered
Parameterization

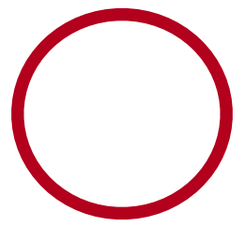
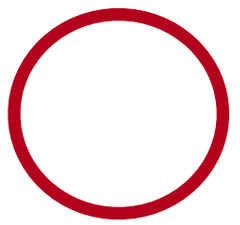
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry
Centered Parameterization		

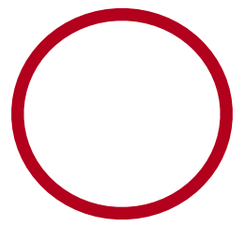
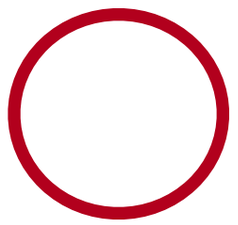
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data
Centered Parameterization			

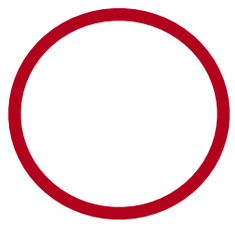
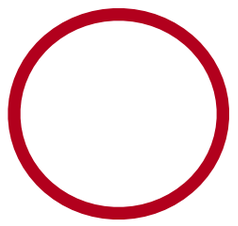
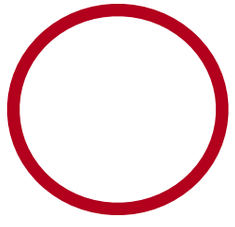
Consequently the two parameterizations are each advantageous in different circumstances.

	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				

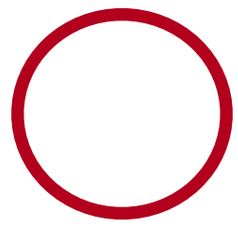
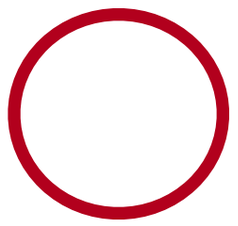
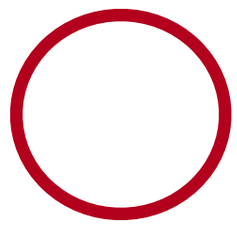
Consequently the two parameterizations are each advantageous in different circumstances.

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

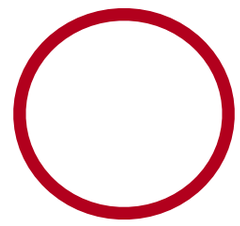
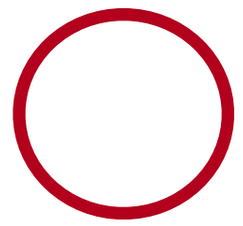
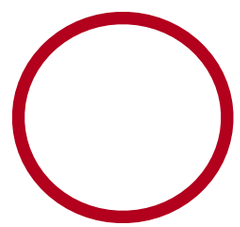
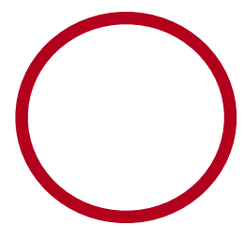
Consequently the two parameterizations are each advantageous in different circumstances.

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

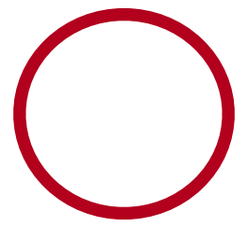
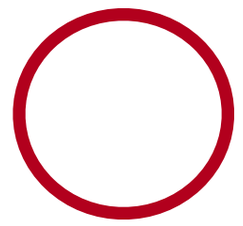
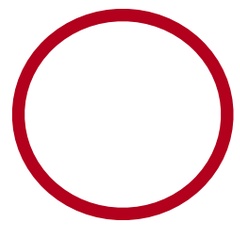
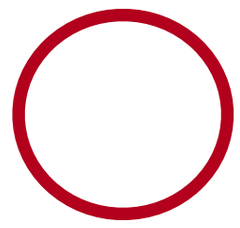
Consequently the two parameterizations are each advantageous in different circumstances.

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

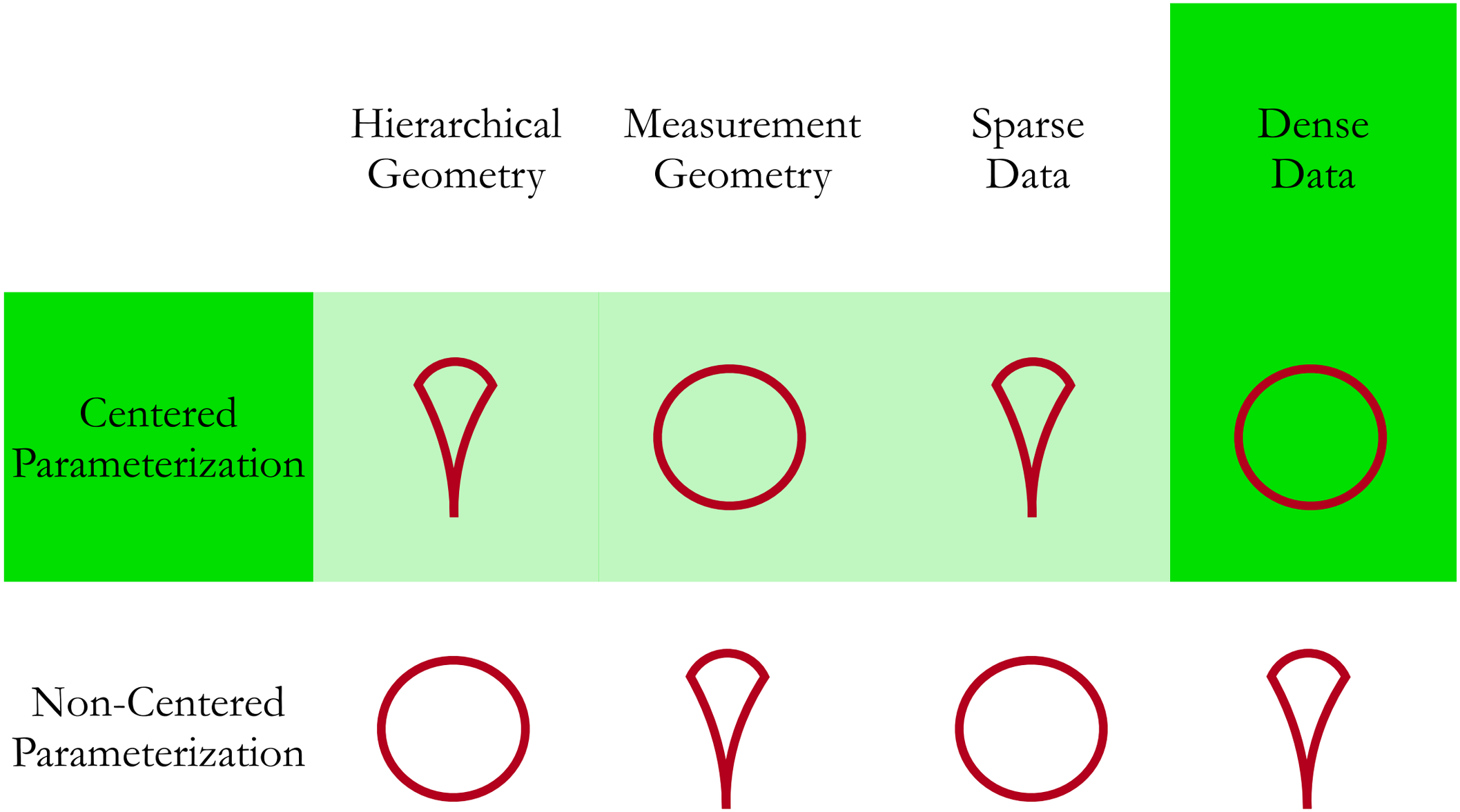
Consequently the two parameterizations are each advantageous in different circumstances.

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

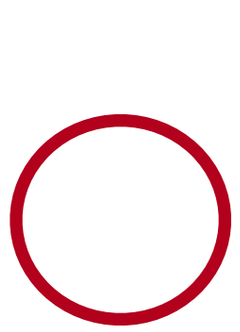
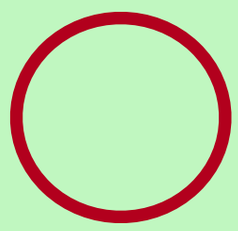
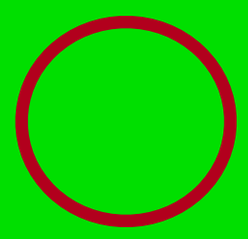
Consequently the two parameterizations are each advantageous in different circumstances.

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

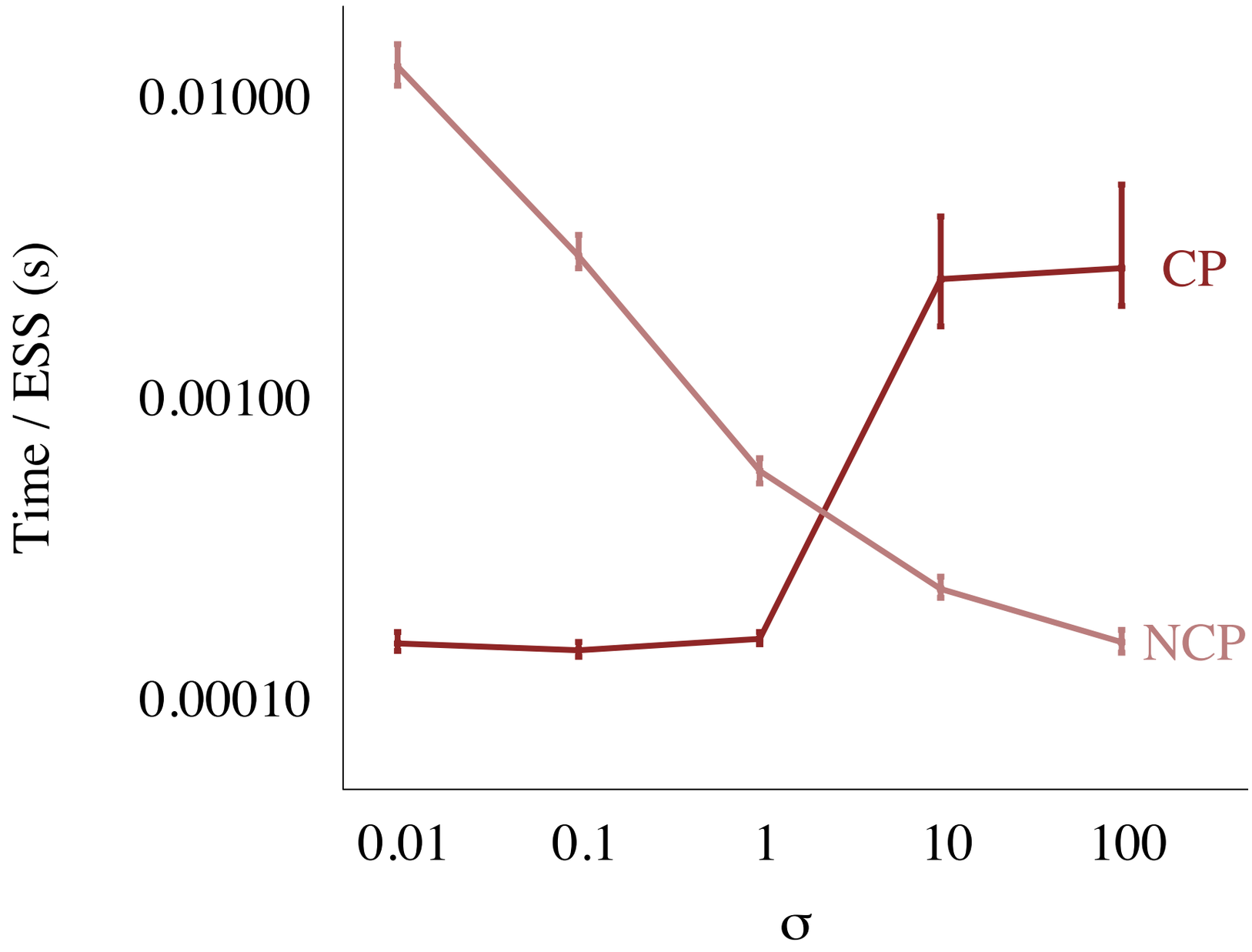
Consequently the two parameterizations are each advantageous in different circumstances.



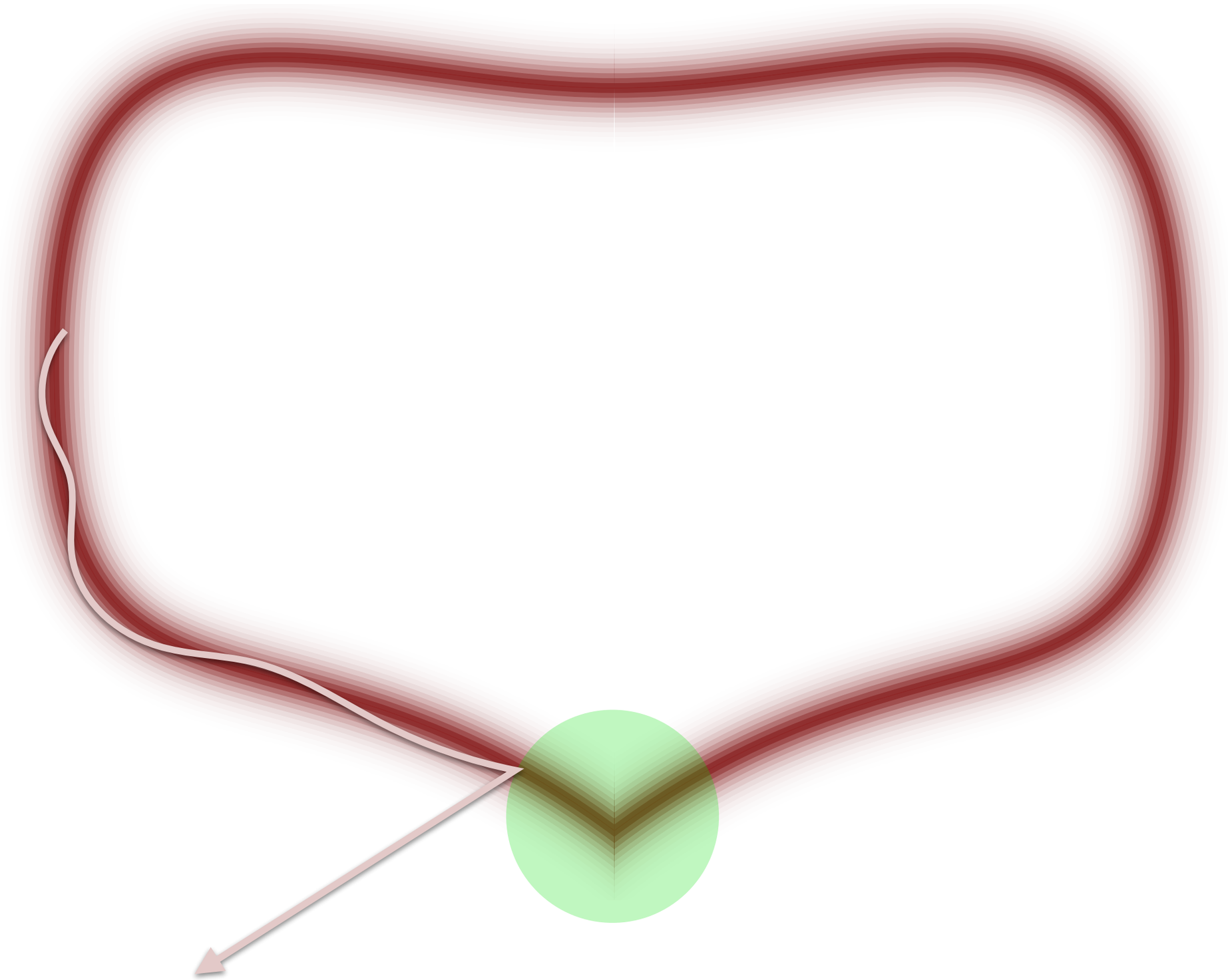
Consequently the two parameterizations are each advantageous in different circumstances.

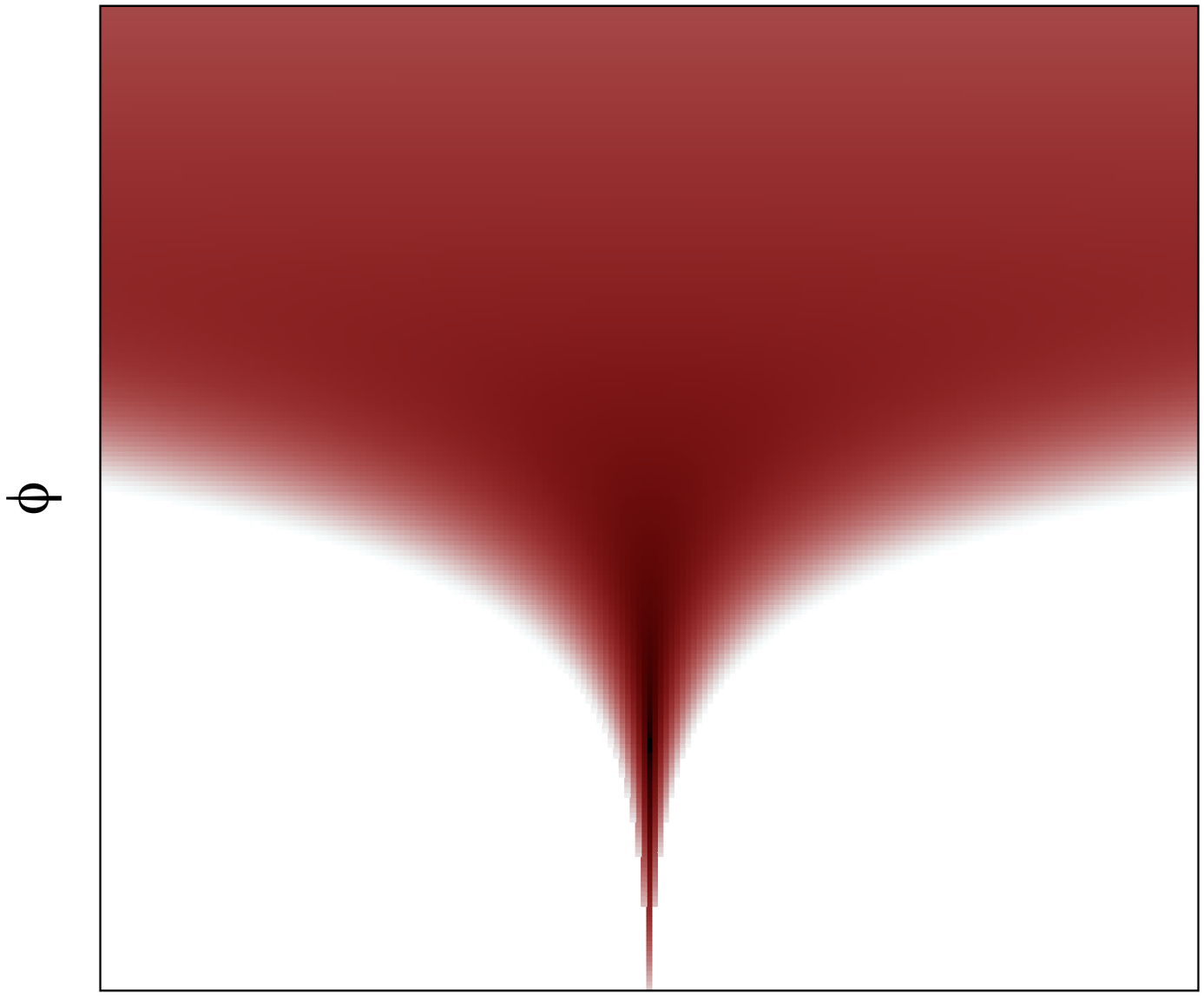
	Hierarchical Geometry	Measurement Geometry	Sparse Data	Dense Data
Centered Parameterization				
Non-Centered Parameterization				

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