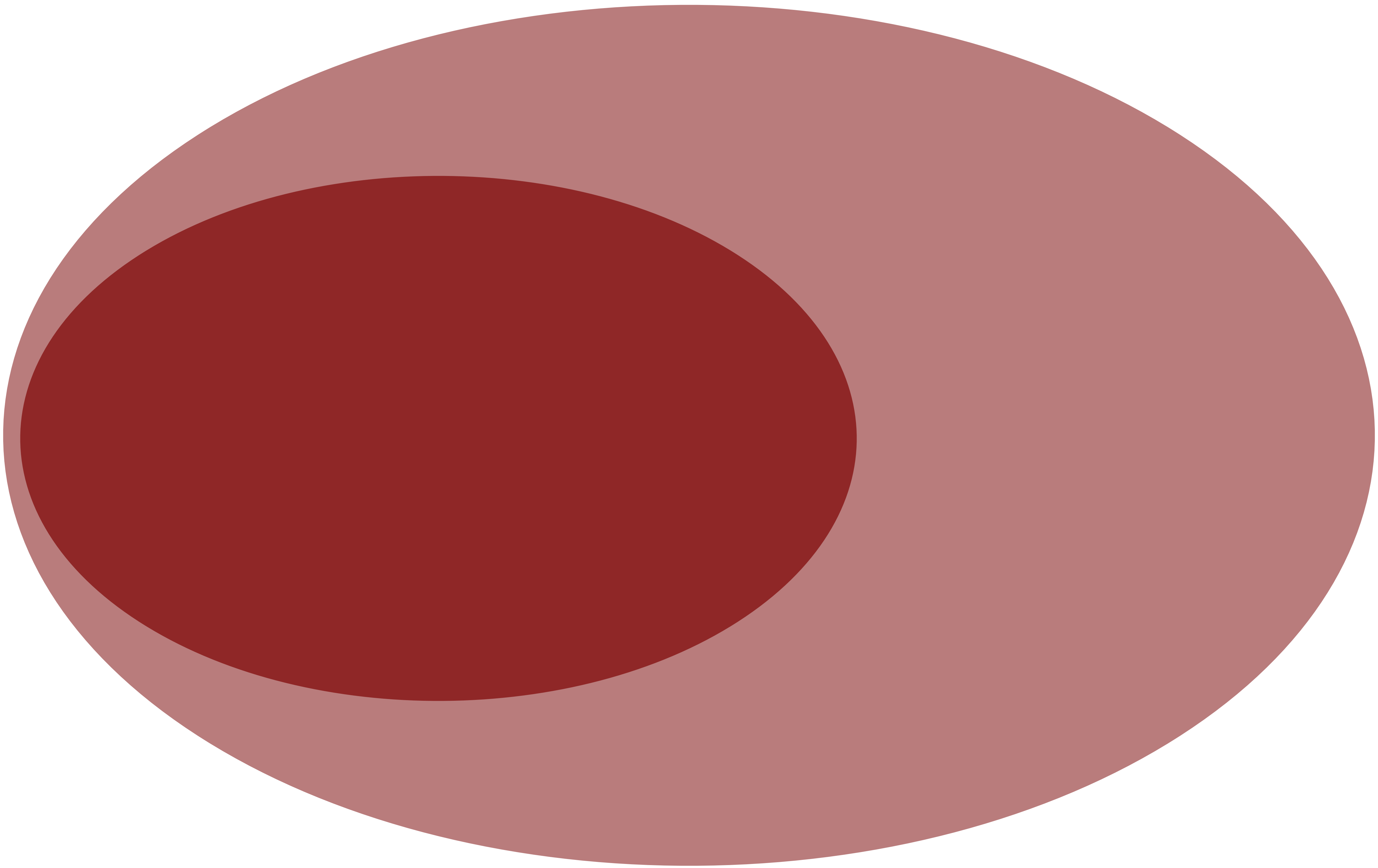
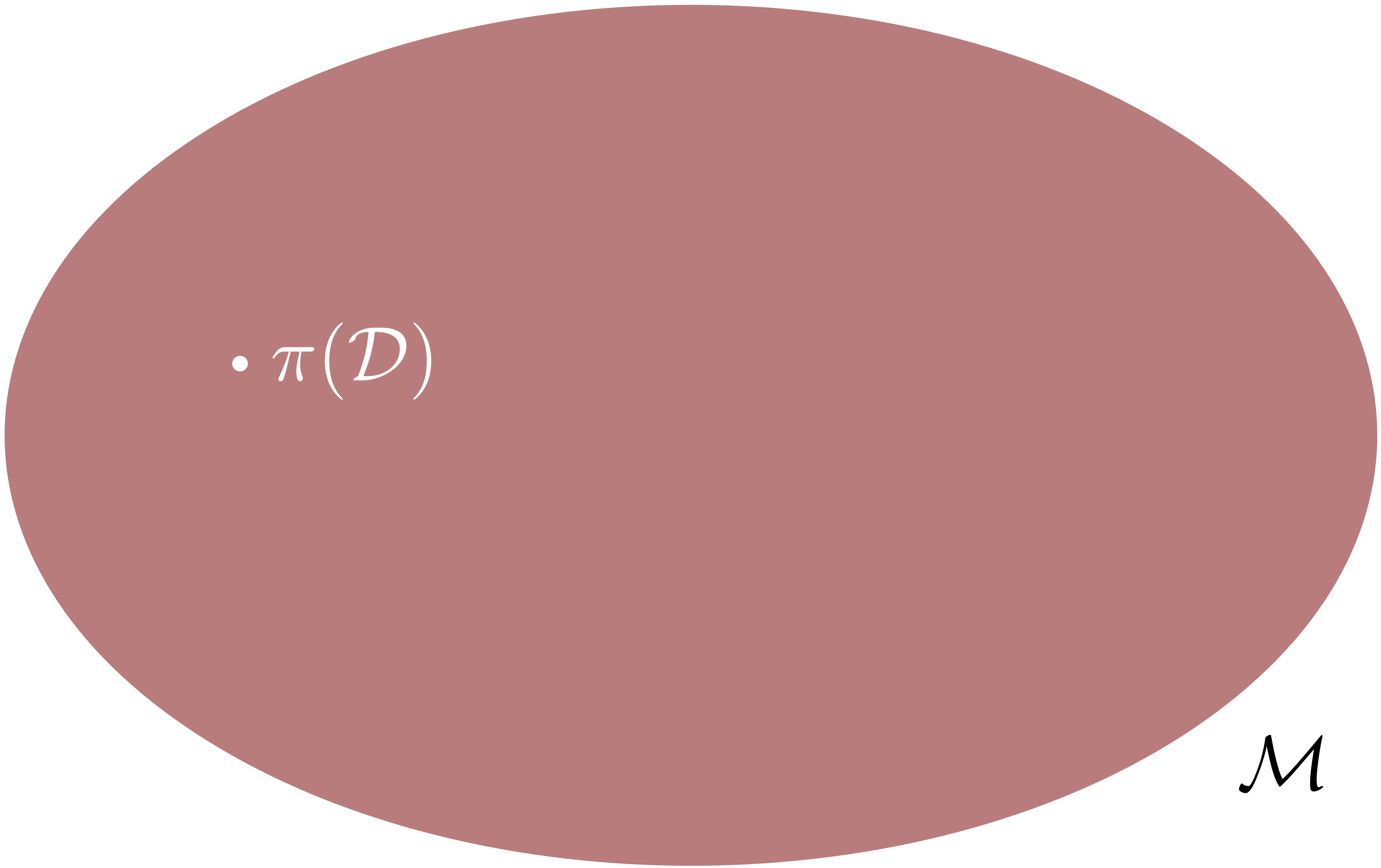
Let's review…



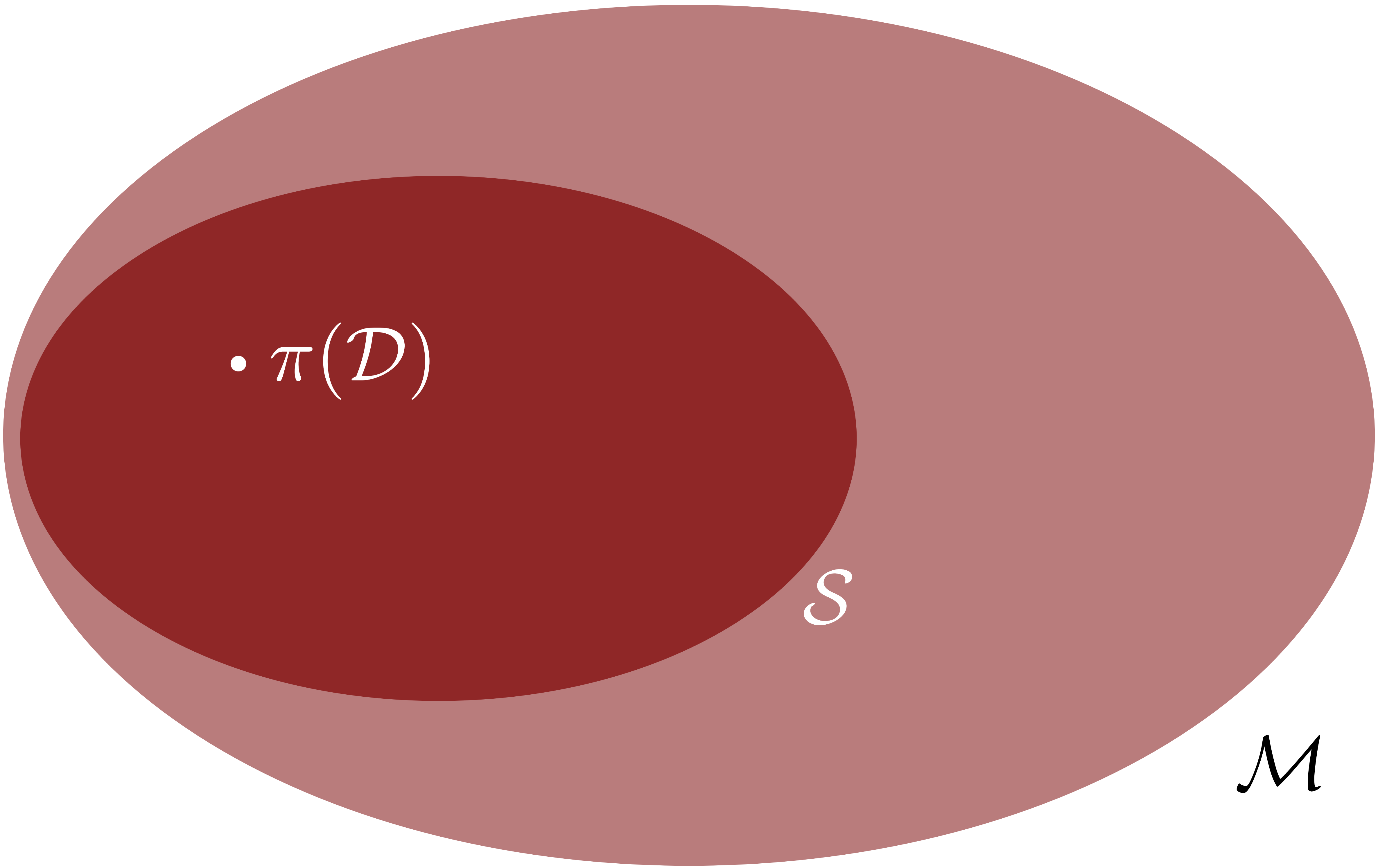
We want to infer some data generating process.

$$\pi(\mathcal{D})$$

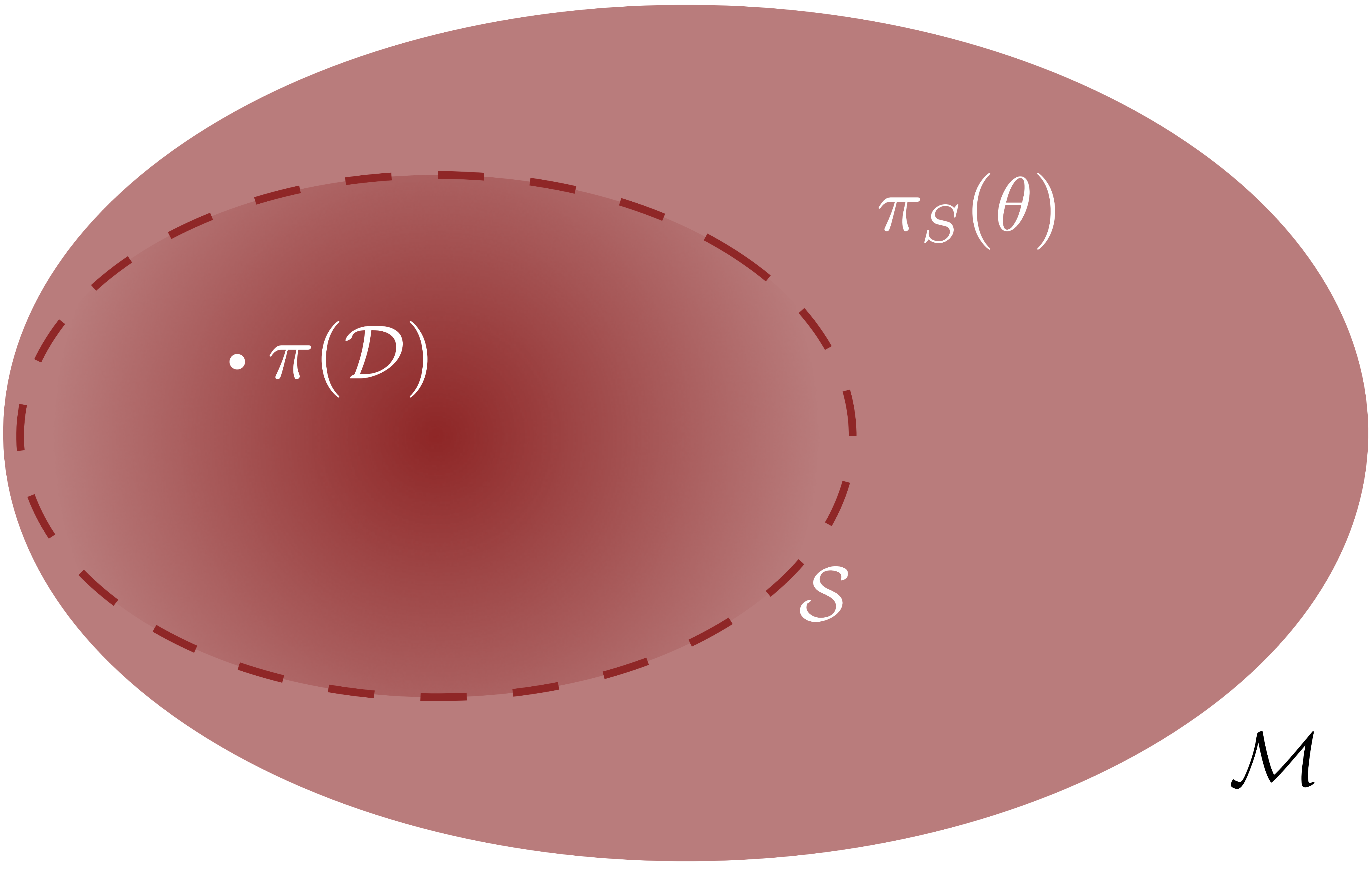
Inference is an identification of this process from the space of all possible data generating processes.



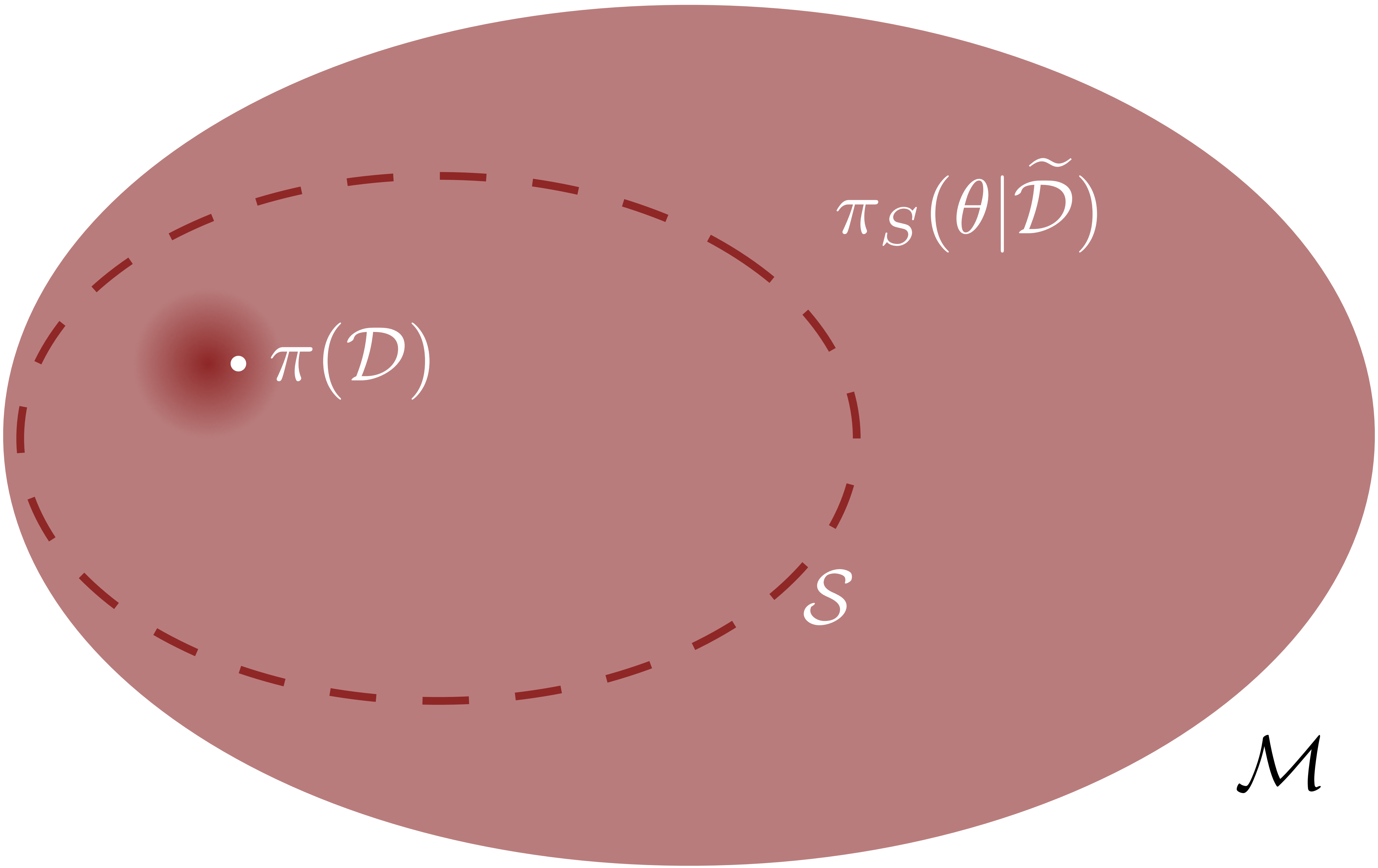
The space of all data generating processes is too big so we restrict our analysis to a *small world*.



In Bayesian inference we use probability distributions to quantify our information about the *small world*.



In Bayesian inference we use probability distributions to quantify our information about the *small world*.



All Bayesian computations reduce to expectations with respect to the the posterior distribution.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta | \widetilde{\mathcal{D}}) \, f(\theta)$$

Best Practices

1a. Maintain reproducibility by saving the model, data, and inits in files and the R commands in scripts.

1a. Maintain reproducibility by saving the model, data, and inits in files and the R commands in scripts.

1b. Use version control on your files and scripts.

1a. Maintain reproducibility by saving the model, data, and inits in files and the R commands in scripts.

1b. Use version control on your files and scripts.

2. Start simple! Build your model in stages, ensuring good fits at each stage.

1a. Maintain reproducibility by saving the model, data, and inits in files and the R commands in scripts.

1b. Use version control on your files and scripts.

2. Start simple! Build your model in stages, ensuring good fits at each stage.

3. Fit your model to simulated data to ensure that Stan can recover the true values.

1a. Maintain reproducibility by saving the model, data, and inits in files and the R commands in scripts.

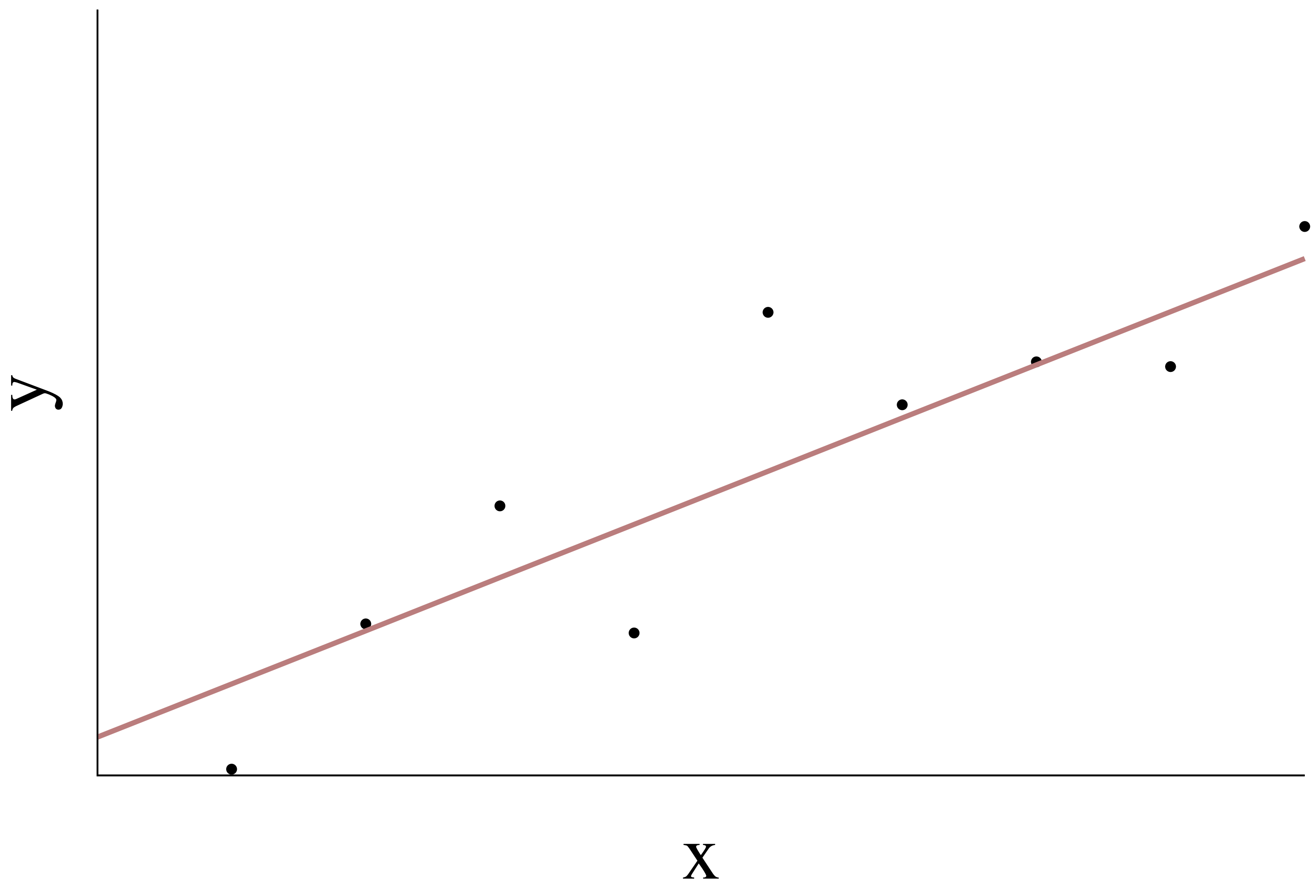
1b. Use version control on your files and scripts.

2. Start simple! Build your model in stages, ensuring good fits at each stage.

3. Fit your model to simulated data to ensure that Stan can recover the true values.

4. Keep an eye on the diagnostics!

Regression Modeling



Recall that in Bayesian inference we build up an inferential model by specifying a prior and a likelihood.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) \propto \pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)$$

Recall that in Bayesian inference we build up an inferential model by specifying a prior and a likelihood.

$$\pi \;\; (\theta|\mathcal{D}) \propto \pi \;\; (\mathcal{D}|\theta)\pi \;\; (\theta)$$

Likelihoods model the measurement process
and are most naturally specified *generatively*.

$$\pi(\mathcal{D}_1, \ldots, \mathcal{D}_4 \mid \theta)$$

Likelihoods model the measurement process and are most naturally specified *generatively*.

$$
\begin{aligned}
\pi(\mathcal{D}_1, \ldots, \mathcal{D}_4 \mid \theta) \propto \quad & \pi(\mathcal{D}_1 \mid \mathcal{D}_2, \ldots, \mathcal{D}_4 \theta) \\
& \times \pi(\mathcal{D}_2 \mid \mathcal{D}_3, \mathcal{D}_4, \theta) \\
& \times \pi(\mathcal{D}_3 \mid \mathcal{D}_4, \theta) \\
& \times \pi(\mathcal{D}_4 \mid \theta)
\end{aligned}
$$

Likelihoods model the measurement process and are most naturally specified *generatively*.

$$
\begin{aligned}
\pi(\mathcal{D}_1, \ldots, \mathcal{D}_4 \mid \theta) \propto \quad & \pi(\mathcal{D}_1 \mid \mathcal{D}_2, \theta) \\
& \times \pi(\mathcal{D}_2 \mid \mathcal{D}_3, \theta) \\
& \times \pi(\mathcal{D}_3 \mid \mathcal{D}_4, \theta) \\
& \times \pi(\mathcal{D}_4 \mid \theta)
\end{aligned}
$$

Similarly, we can also model unobserved variables, such as the parameters, *generatively*.

$$
\begin{aligned}
\pi(\theta_1, \ldots, \theta_4) \propto \quad & \pi(\theta_1 \mid \theta_2, \ldots, \theta_4) \\
& \times \pi(\theta_2 \mid \theta_3, \theta_4) \\
& \times \pi(\theta_3 \mid \theta_4) \\
& \times \pi(\theta_4)
\end{aligned}
$$

Similarly, we can also model unobserved variables, such as the parameters, *generatively*.

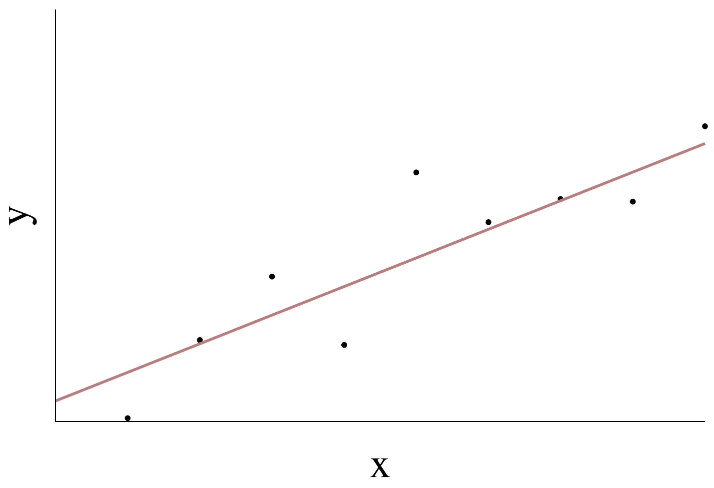
$$
\begin{aligned}
\pi(\theta_1, \ldots, \theta_4) \propto \quad & \pi(\theta_1 \mid \theta_2) \\
& \times \pi(\theta_2 \mid \theta_3) \\
& \times \pi(\theta_3 \mid \theta_4) \\
& \times \pi(\theta_4)
\end{aligned}
$$

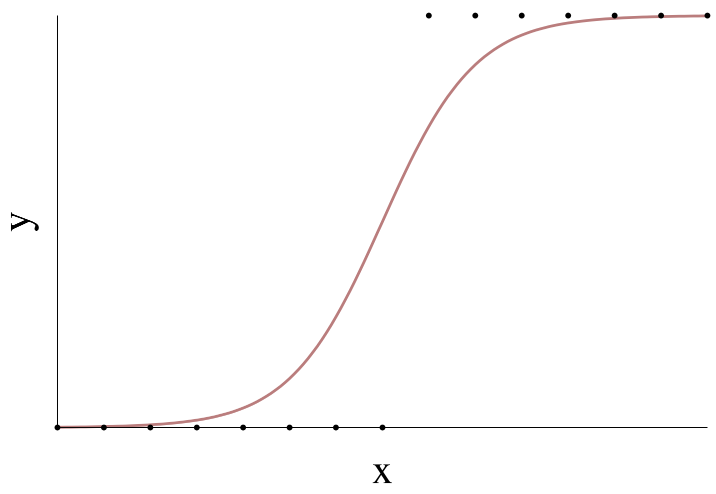
This generative decomposition allows us to focus on *modular* modeling components.

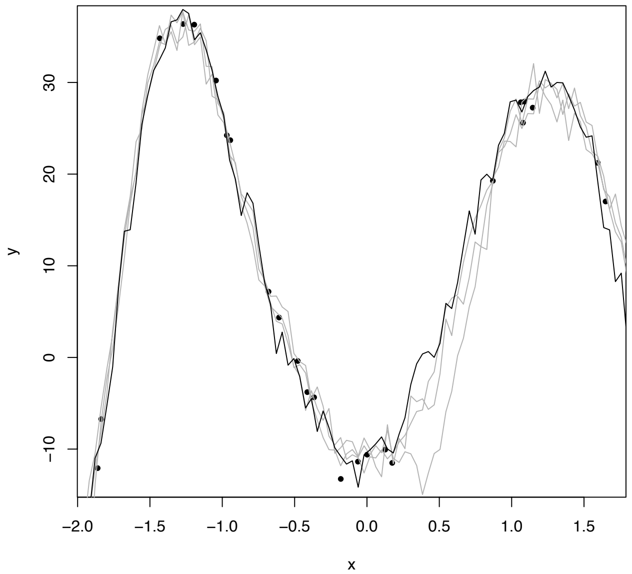
$$
\begin{aligned}
\pi(\theta_1, \ldots, \theta_4) \propto \quad & \pi(\theta_1 \mid \theta_2) \\
& \times \pi(\theta_2 \mid \theta_3) \\
& \times \pi(\theta_3 \mid \theta_4) \\
& \times \pi(\theta_4)
\end{aligned}
$$

$$\pi(\theta_n | \theta_{n+1})$$

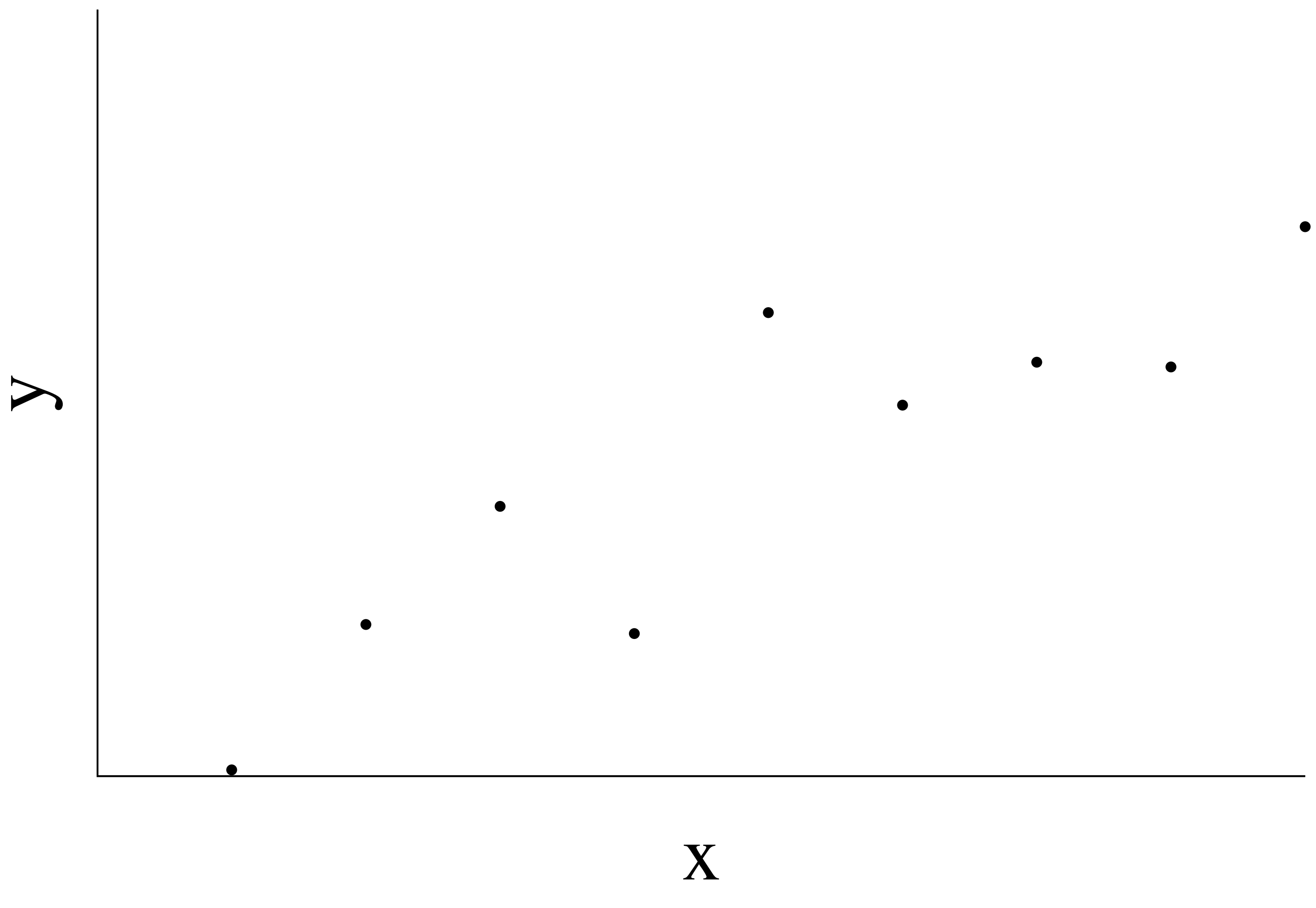
Many of the most common and useful modeling techniques are forms of *regression*.







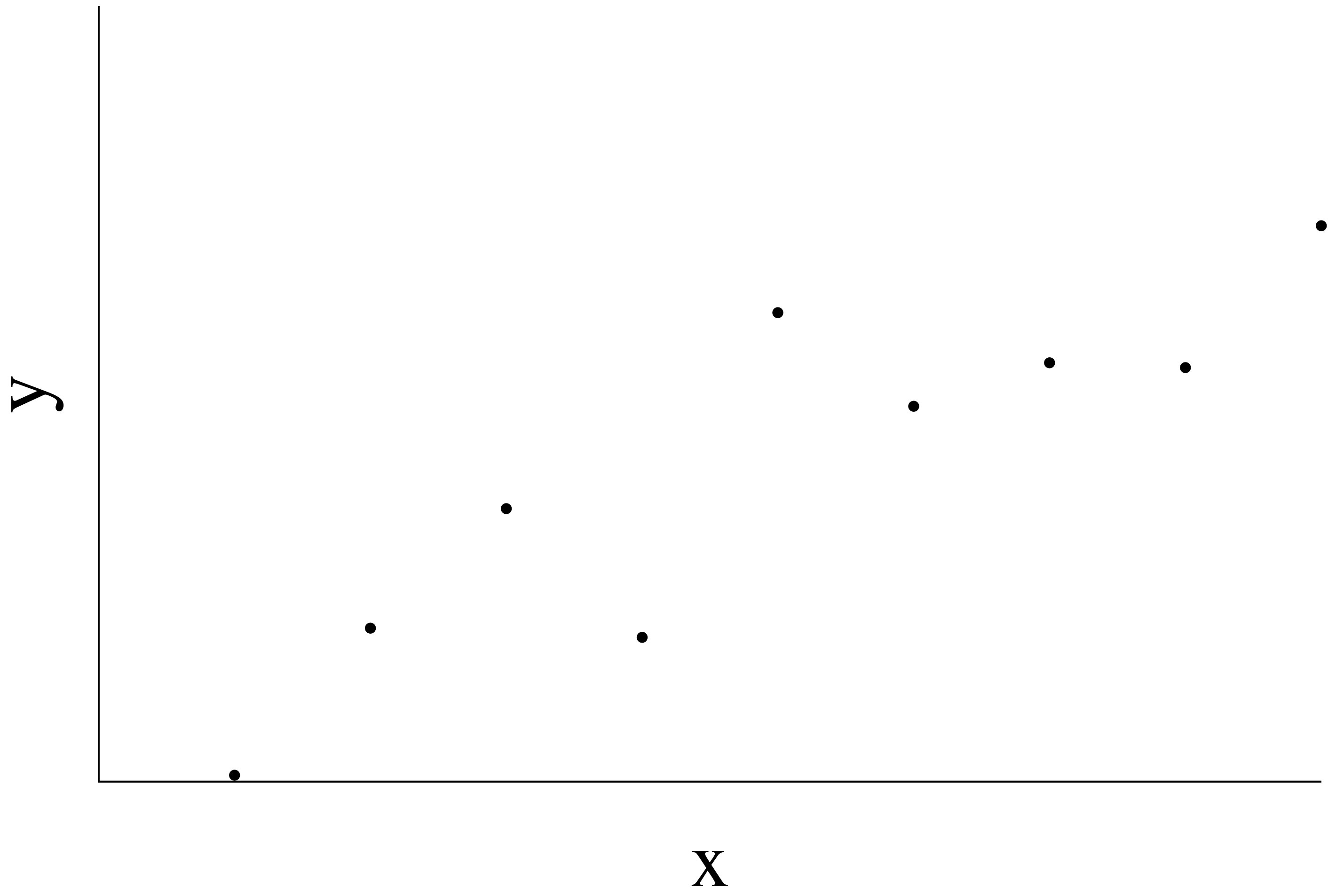
Foundations of Regression



Often the data naturally separate into *variates*, y, and *covariates*, x.

$$\mathcal{D} \to \{y, x\}$$

Regression models the statistical relationship between the variates and the covariates.



Regression models the statistical relationship between the variates and the covariates.

$$\pi(y, x | \theta) = \pi(y | x, \theta) \, \pi(x | \theta)$$

Regression models the statistical relationship between the variates and the covariates.

$$\pi(y, x|\theta) = \pi(y|x, \theta)\, \pi(x|\theta)$$

We typically assume that the covariates are independent of the model parameters.

$$\pi(x|\theta) = \pi(x)$$

In which case the likelihood becomes a model of the variates conditional on the covariates.

$$\pi(y, x|\theta) = \pi(y|x, \theta)\, \pi(x|\theta)$$

In which case the likelihood becomes a model of the variates conditional on the covariates.

$$\pi(y, x|\theta) = \pi(y|x, \theta)\,\pi(x|\theta)$$

$$\pi(y, x|\theta) = \pi(y|x, \theta)\,\pi(x)$$

In which case the likelihood becomes a model of the variates conditional on the covariates.

$$\pi(y, x|\theta) = \pi(y|x, \theta)\,\pi(x|\theta)$$

$$\pi(y, x|\theta) = \pi(y|x, \theta)\,\pi(x)$$

$$\pi(y, x|\theta) \propto \pi(y|x, \theta)$$

Covariates are often restricted to a single effective parameter through a deterministic mapping.

$$\pi(y|x,\theta) = \pi(y|f(x,\theta_1),\theta_2)$$

Covariates are often restricted to a single effective parameter through a deterministic mapping.

$$\pi(y|x,\theta) = \pi(y|f(x,\theta_1),\theta_2)$$

$$\pi(y|x,\theta) = \mathcal{N}(y|f(x,\theta),\sigma)$$

Covariates are often restricted to a single effective parameter through a deterministic mapping.

$$\pi(y|x,\theta) = \pi(y|f(x,\theta_1),\theta_2)$$

$$\pi(y|x,\theta) = \mathcal{N}(y|f(x,\theta),\sigma)$$

$$\pi(y|x,\theta) = \mathrm{Bin}(y|f(x,\theta),N)$$

This immediately generalizes to multiple effective parameters.

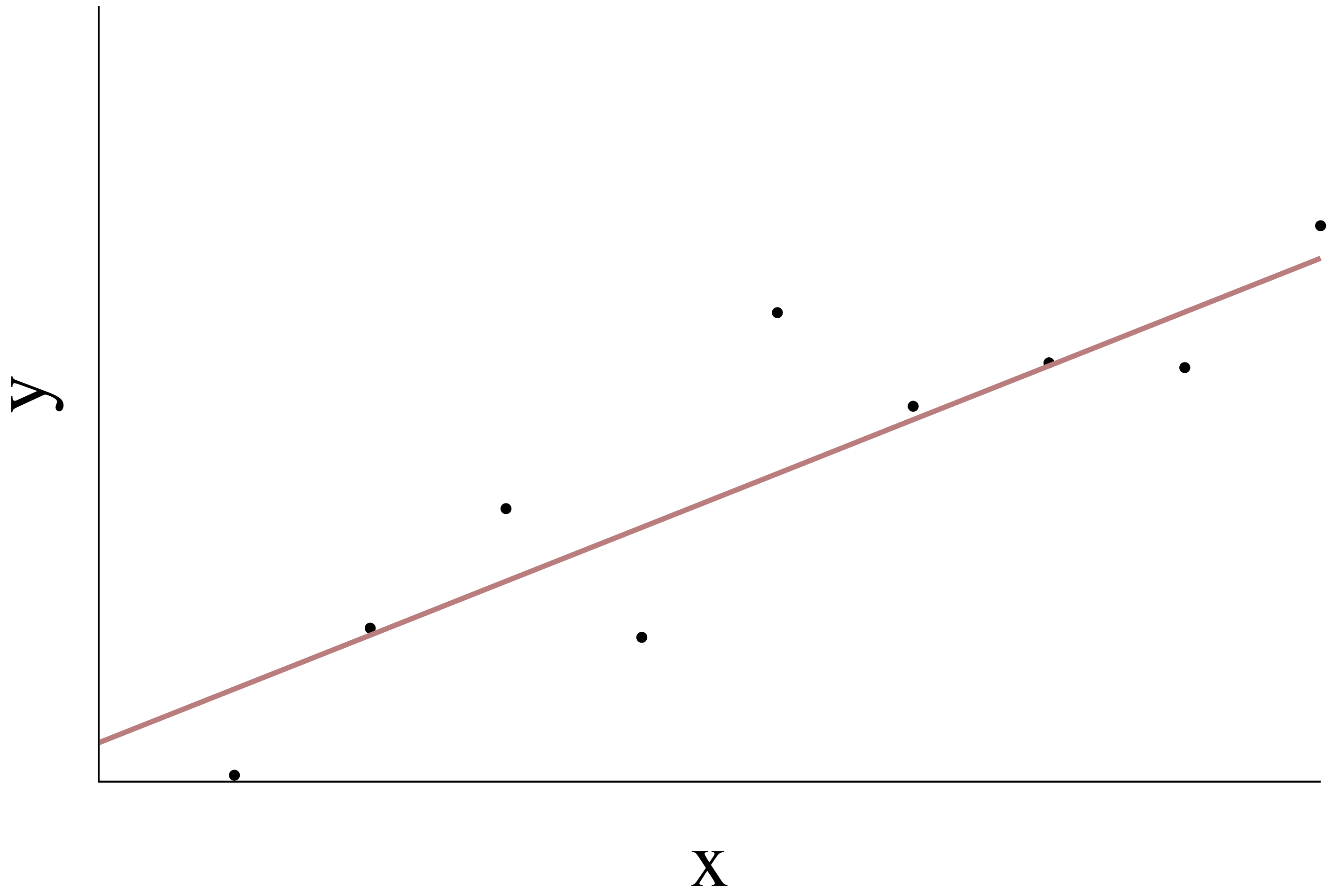
$$\pi(y|x,\theta) = \pi(y|f_1(x,\theta_1)\,,f_2(x,\theta_2)\,,\theta_3)$$

This immediately generalizes to multiple effective parameters.

$$\pi(y|x,\theta) = \pi(y|f_1(x,\theta_1)\,, f_2(x,\theta_2)\,, \theta_3)$$

$$\pi(y|x,\theta) = \mathcal{G}(y|\alpha(x,\theta_1)\,, \beta(x,\theta_2))$$

Linear Models



When an effective parameter is unconstrained we can model it with a linear mapping.

$$f(x, \alpha, \beta) = \beta \cdot x + \alpha$$

Multiple covariates are commonly encapsulated in a *design matrix*.

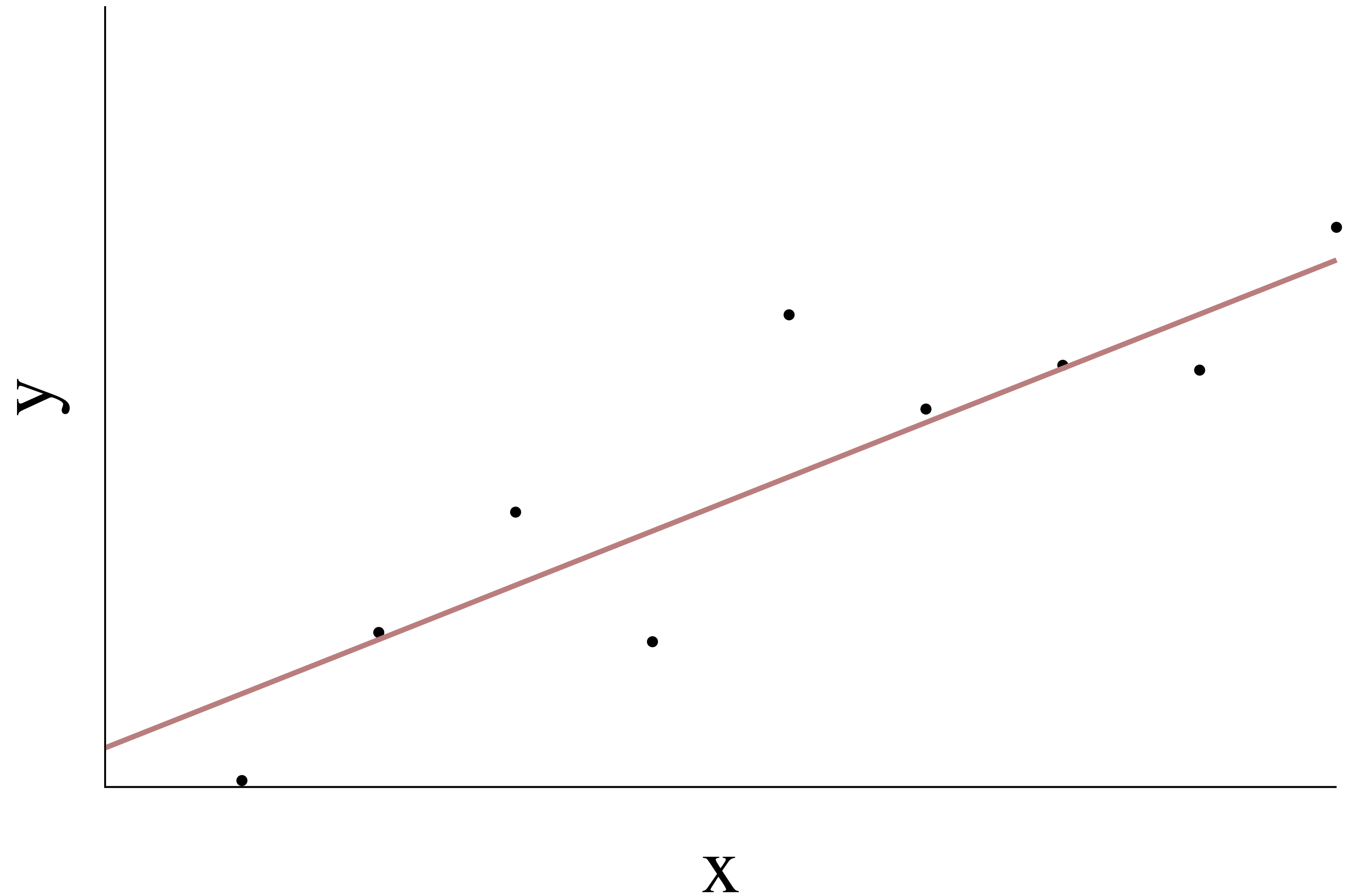
$$f(x, \alpha, \boldsymbol{\beta}) = \sum_{n,i} X_{in} \beta_i + \alpha$$

Multiple covariates are commonly encapsulated in a *design matrix*.

$$f(x, \alpha, \boldsymbol{\beta}) = \sum_{n,i} X_{in}\beta_i + \alpha$$

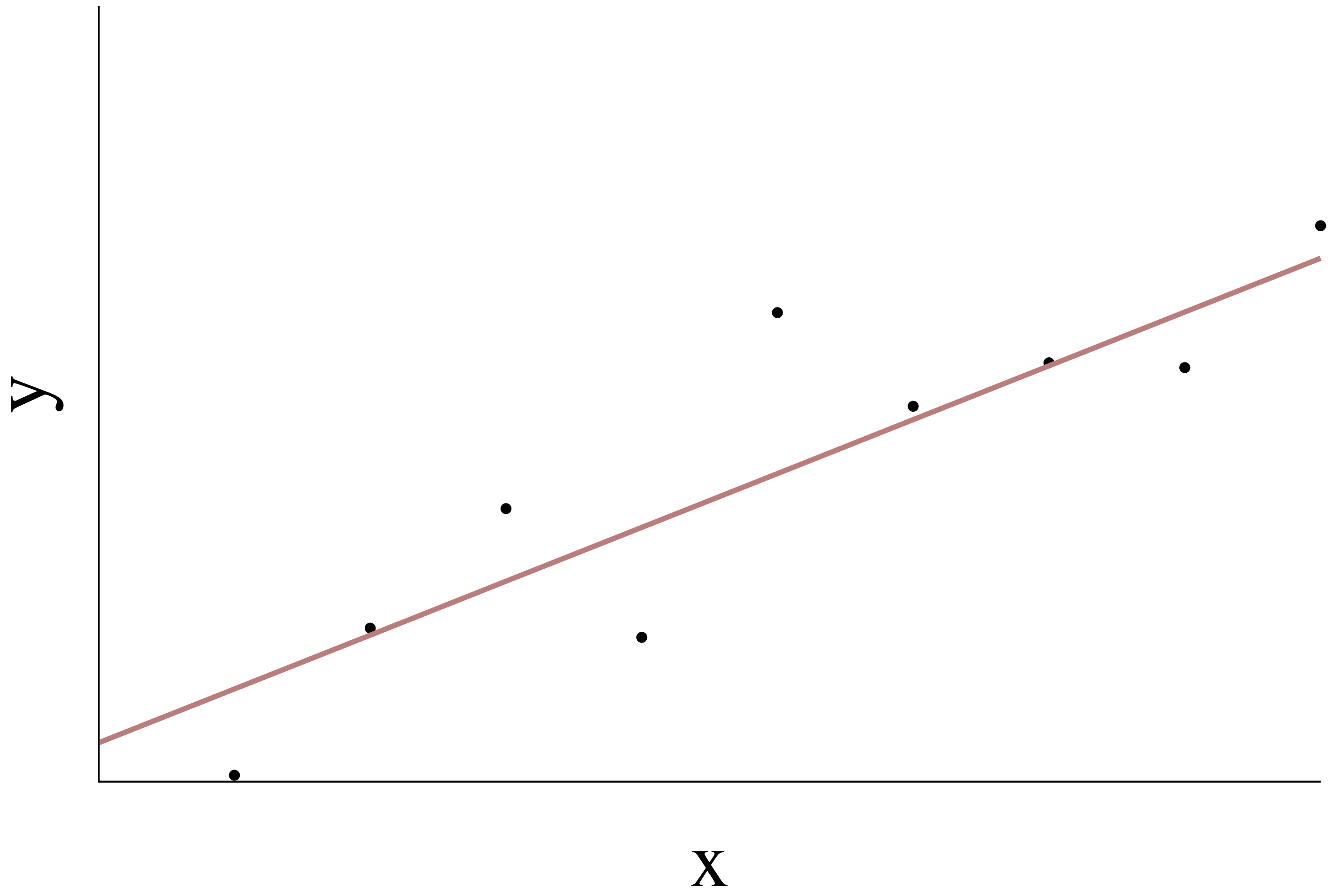
$$f(x, \alpha, \boldsymbol{\beta}) = \mathbf{X}^T\boldsymbol{\beta} + \alpha$$

When the measurement model is Gaussian we recover the ubiquitous Gaussian-Linear model.

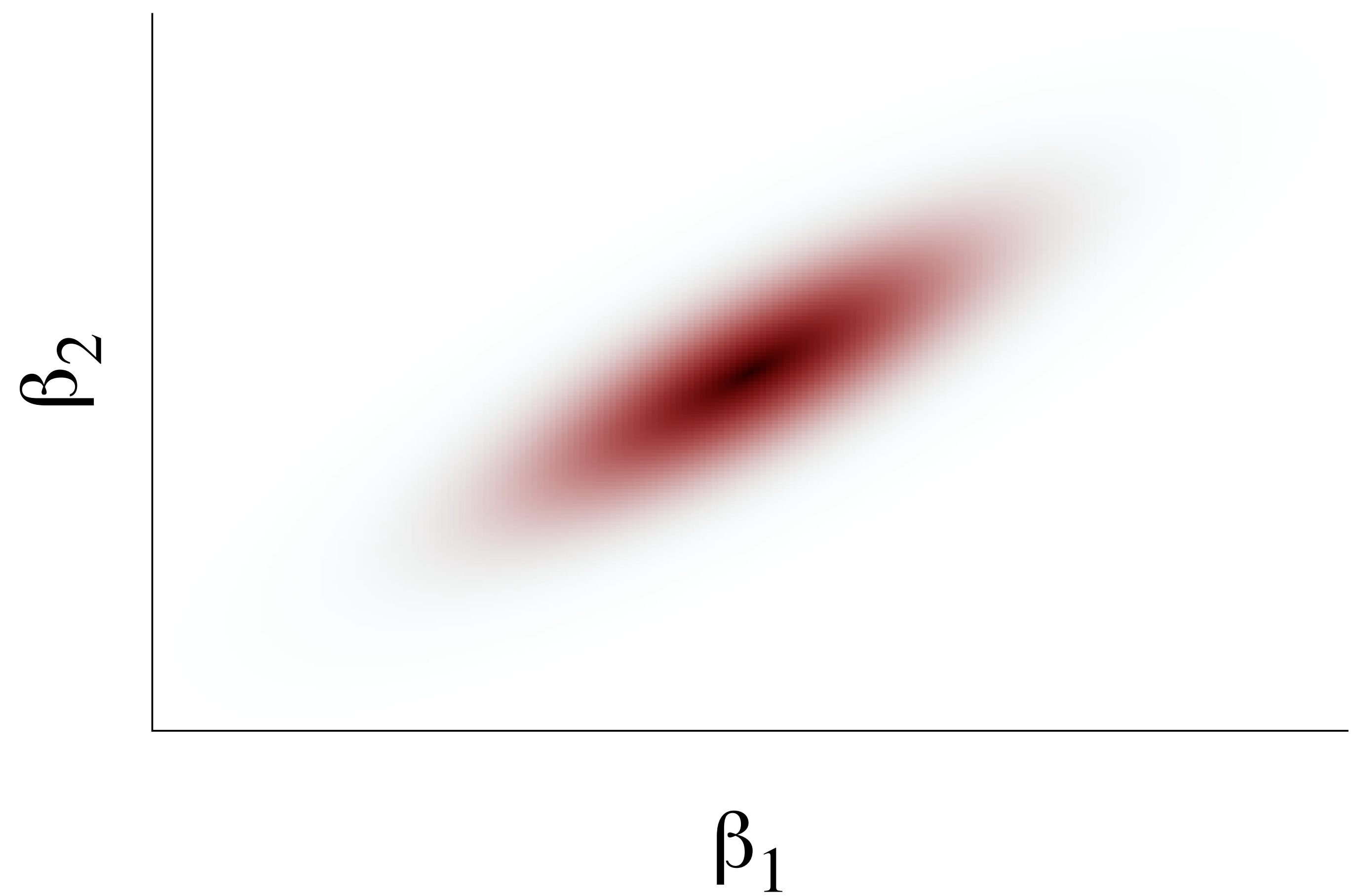


$$\pi(y|\mathbf{X}, \alpha, \boldsymbol{\beta}, \sigma) = \mathcal{N}\left(y|\mathbf{X}^{T}\boldsymbol{\beta} + \alpha, \sigma\right)$$

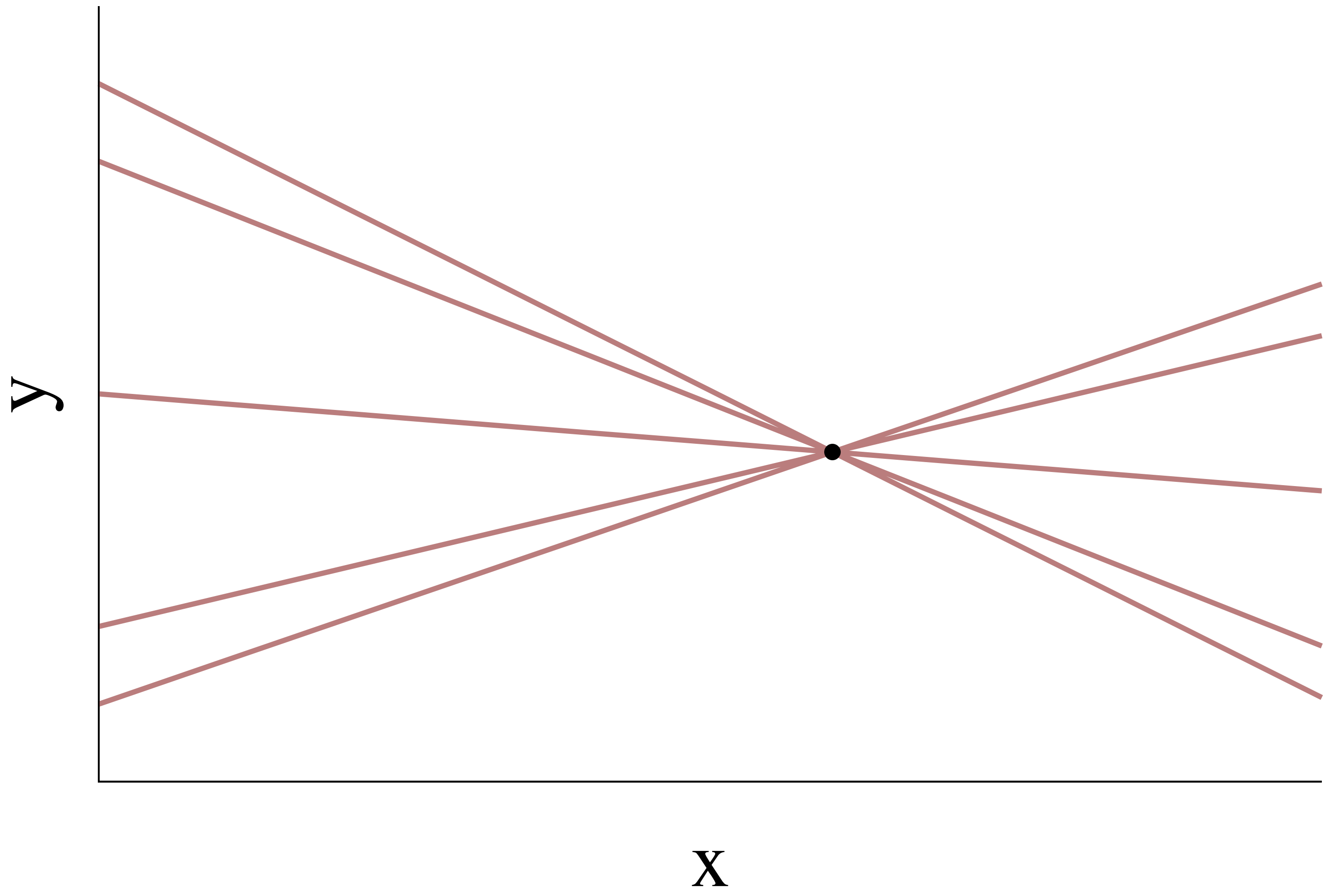
Given enough data, linear models are over-constrained and all the slopes can be fit well.



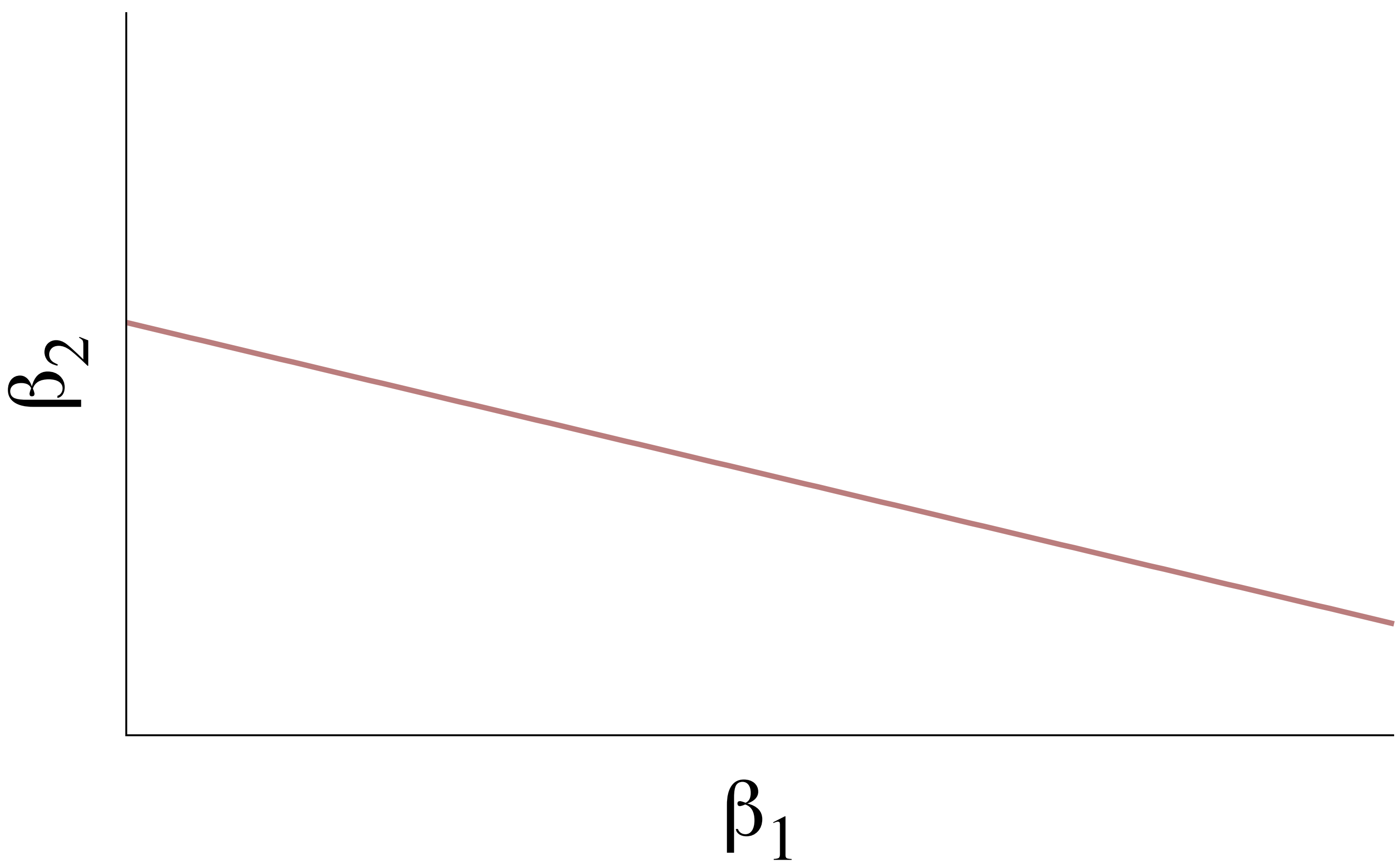
Given enough data, linear models are over-constrained and all the slopes can be fit well.



When there are fewer data than covariates, however, linear models are subject to collinearity.



In collinearity some of the slopes are fully determined while the others are completely undetermined.

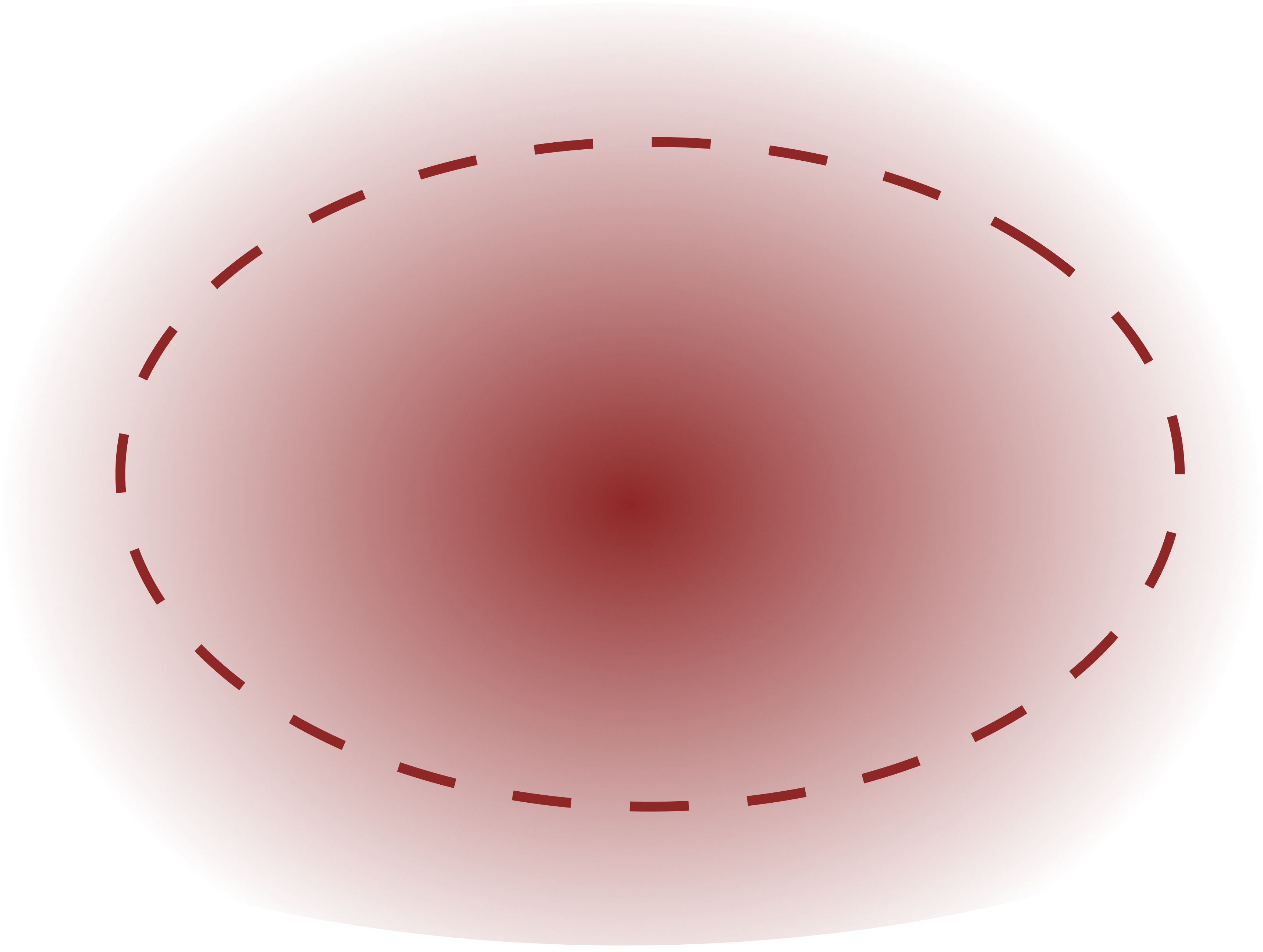


Consequently (weakly) informative priors are critical for building robust linear models.

Consequently (weakly) informative priors are critical for building robust linear models.

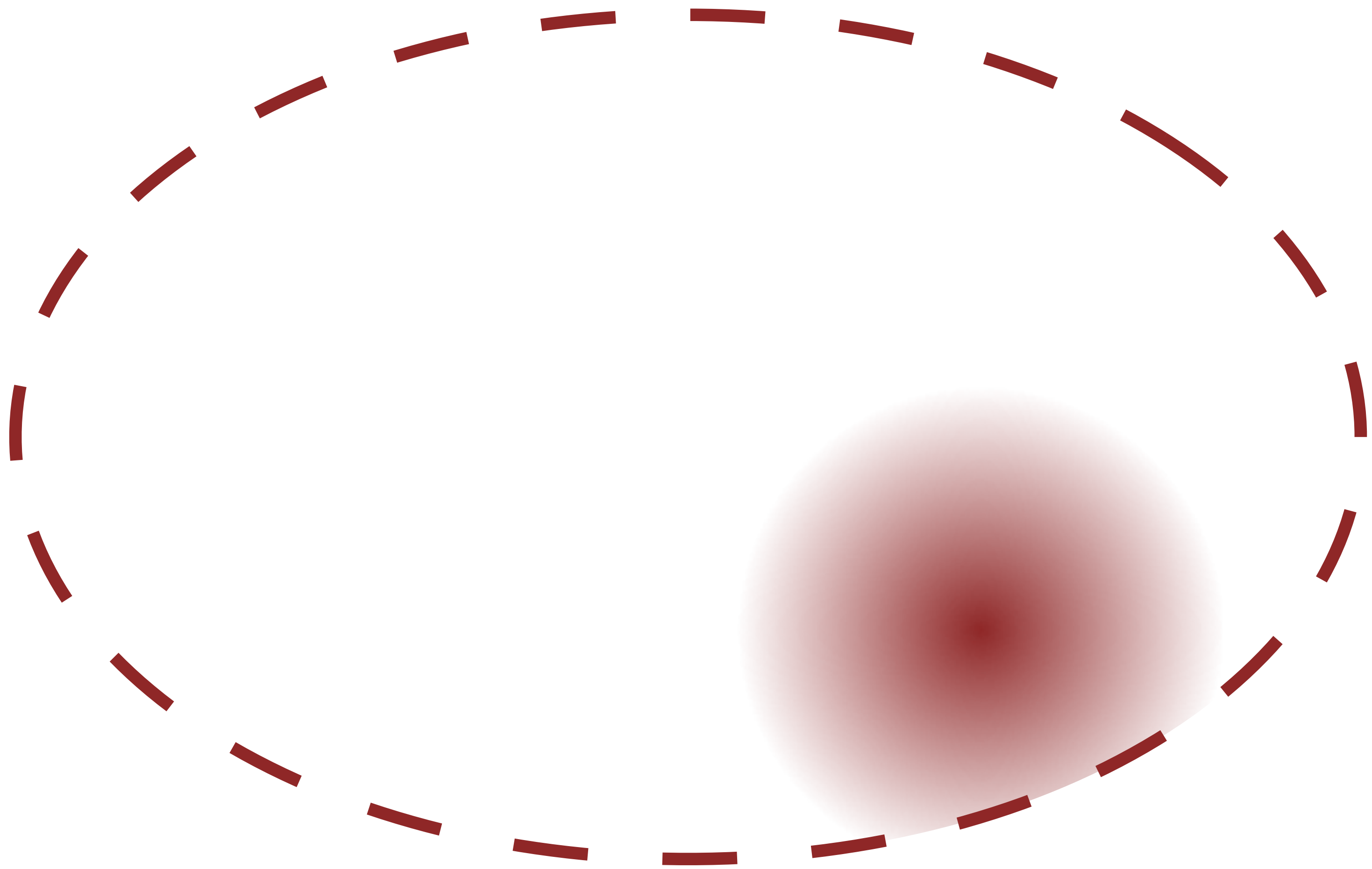
Consequently (weakly) informative priors are critical for building robust linear models.

Consequently (weakly) informative priors are critical for building robust linear models.



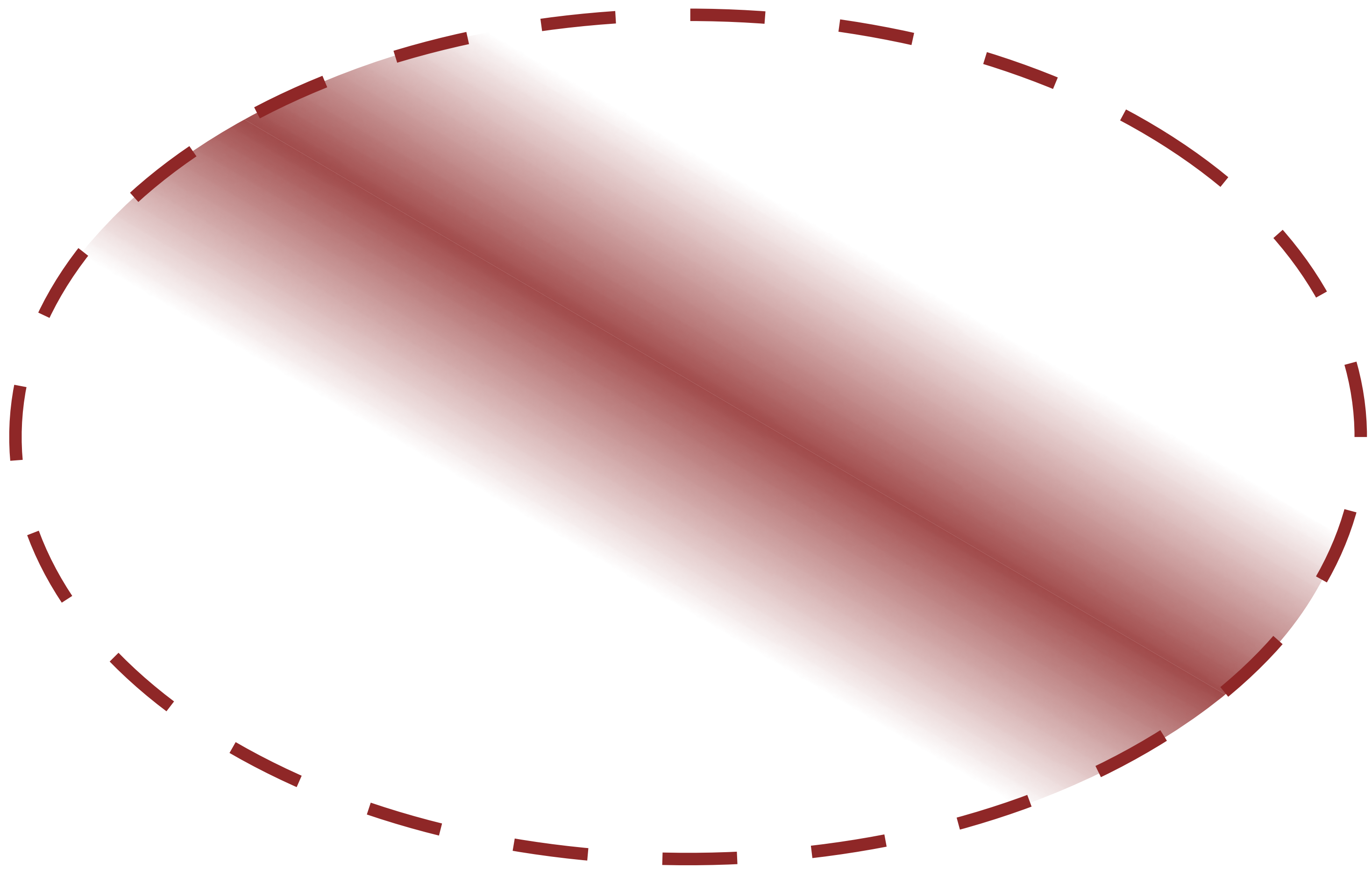
Consequently (weakly) informative priors are critical for building robust linear models.

Consequently (weakly) informative priors are critical for building robust linear models.

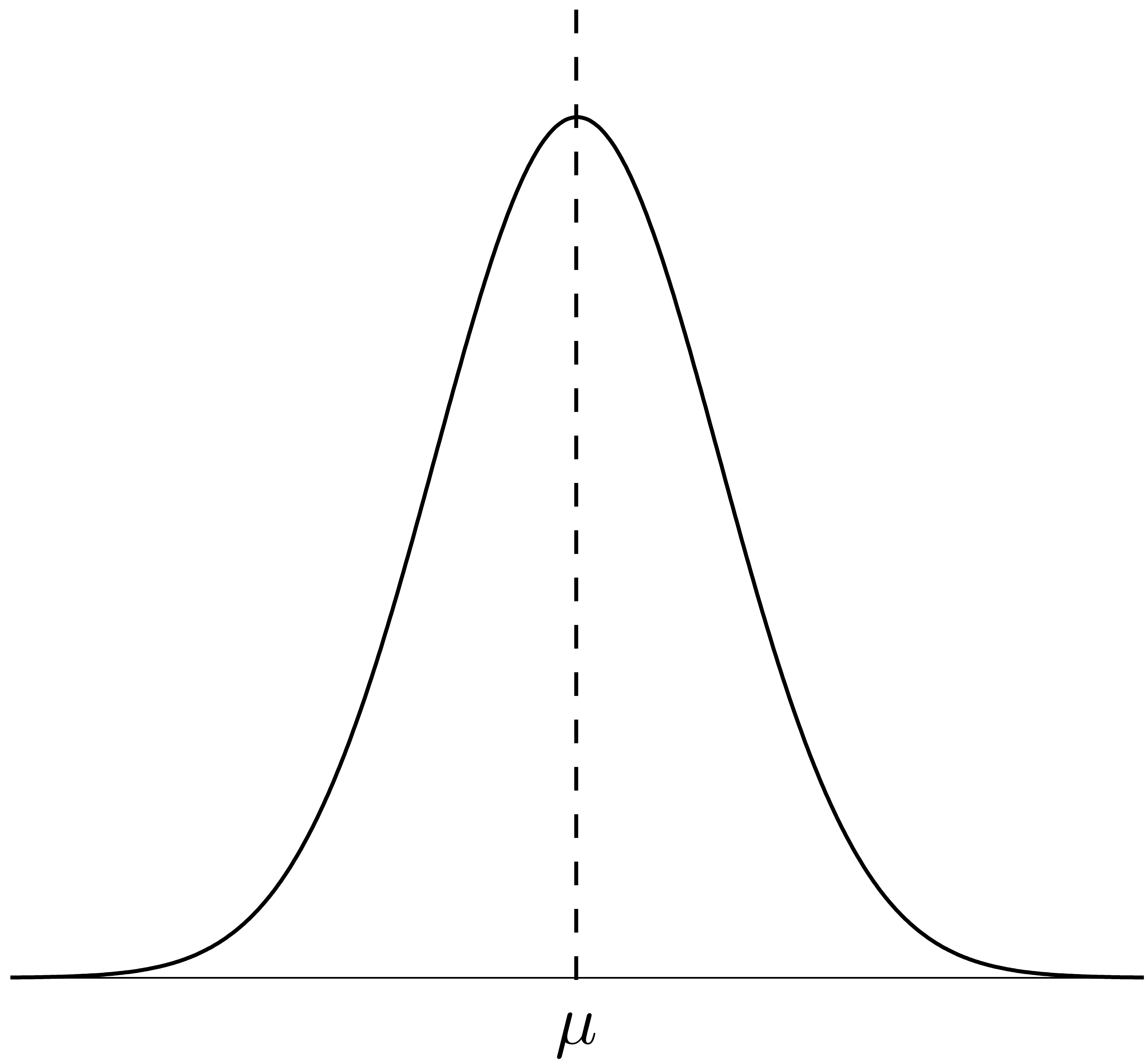


Consequently (weakly) informative priors are critical for building robust linear models.

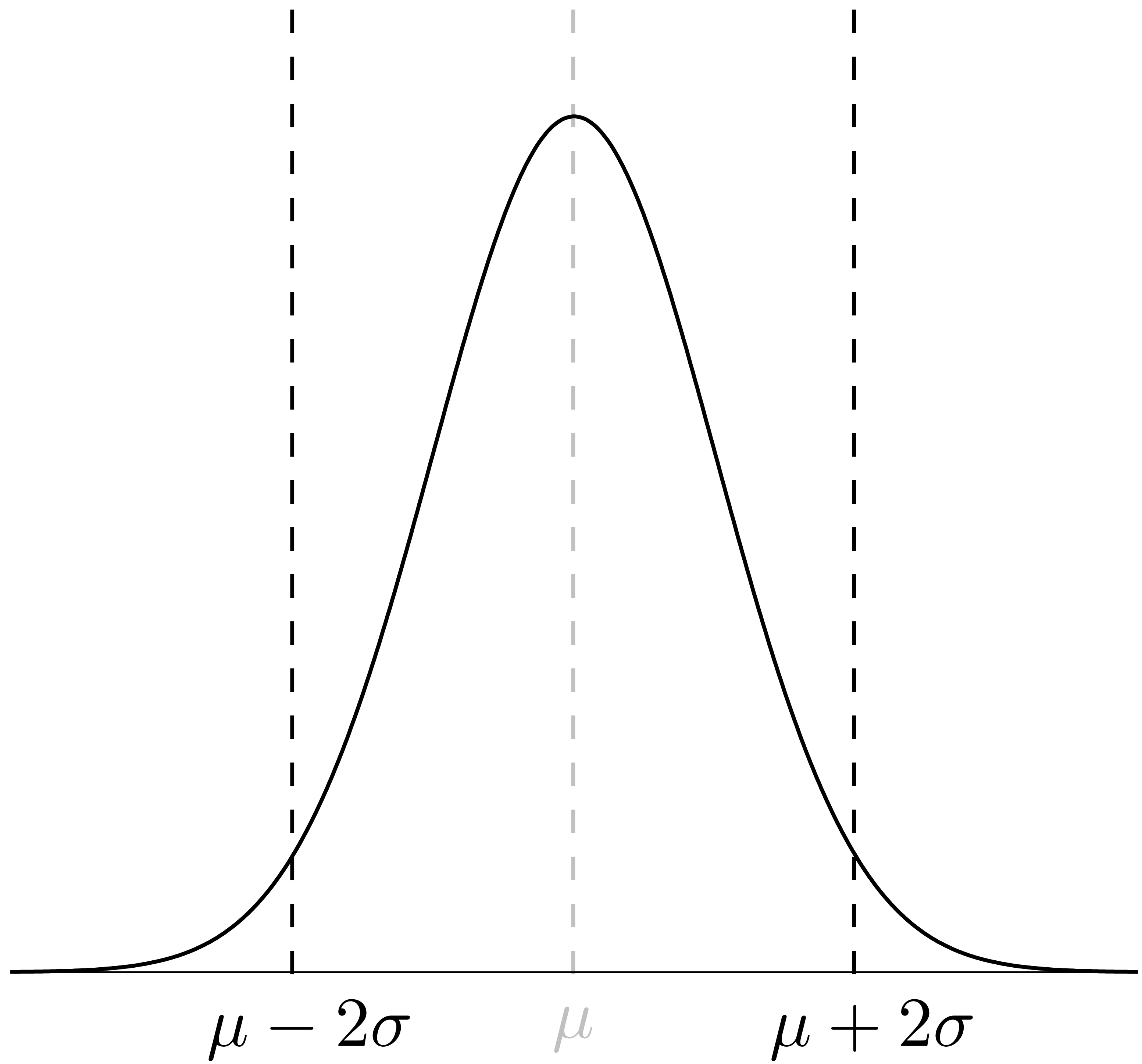
Consequently (weakly) informative priors are critical for building robust linear models.



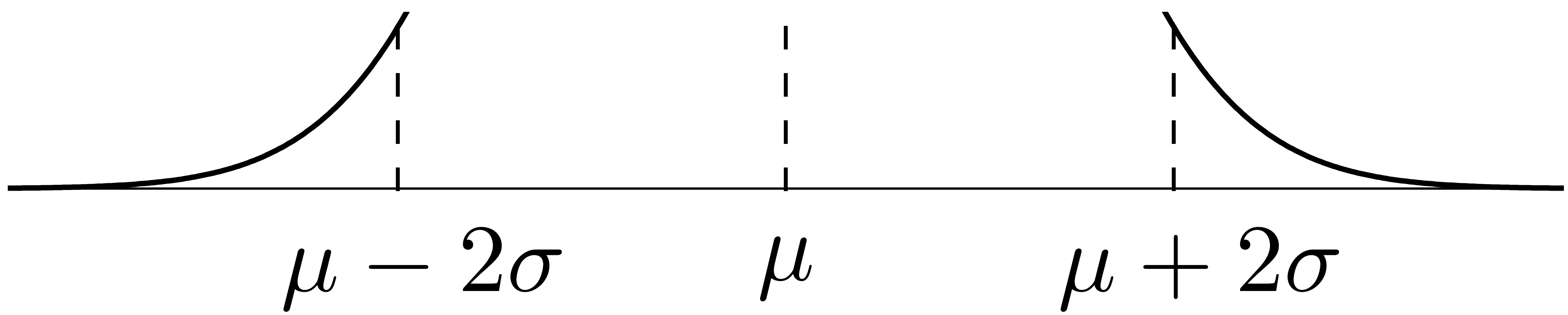
These weakly-informative *containment* priors are typically defined in terms of Gaussian distributions.



These weakly-informative *containment* priors are typically defined in terms of Gaussian distributions.



The breadth of these priors is motivated by reasoning about plausible variations. Think *units*.



The breadth of these priors is motivated by reasoning about plausible variations. Think *units*.

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$$

The breadth of these priors is motivated by reasoning about plausible variations. Think *units*.

$$\beta_i \sim \mathcal{N}(\mu_i, \omega_i)$$

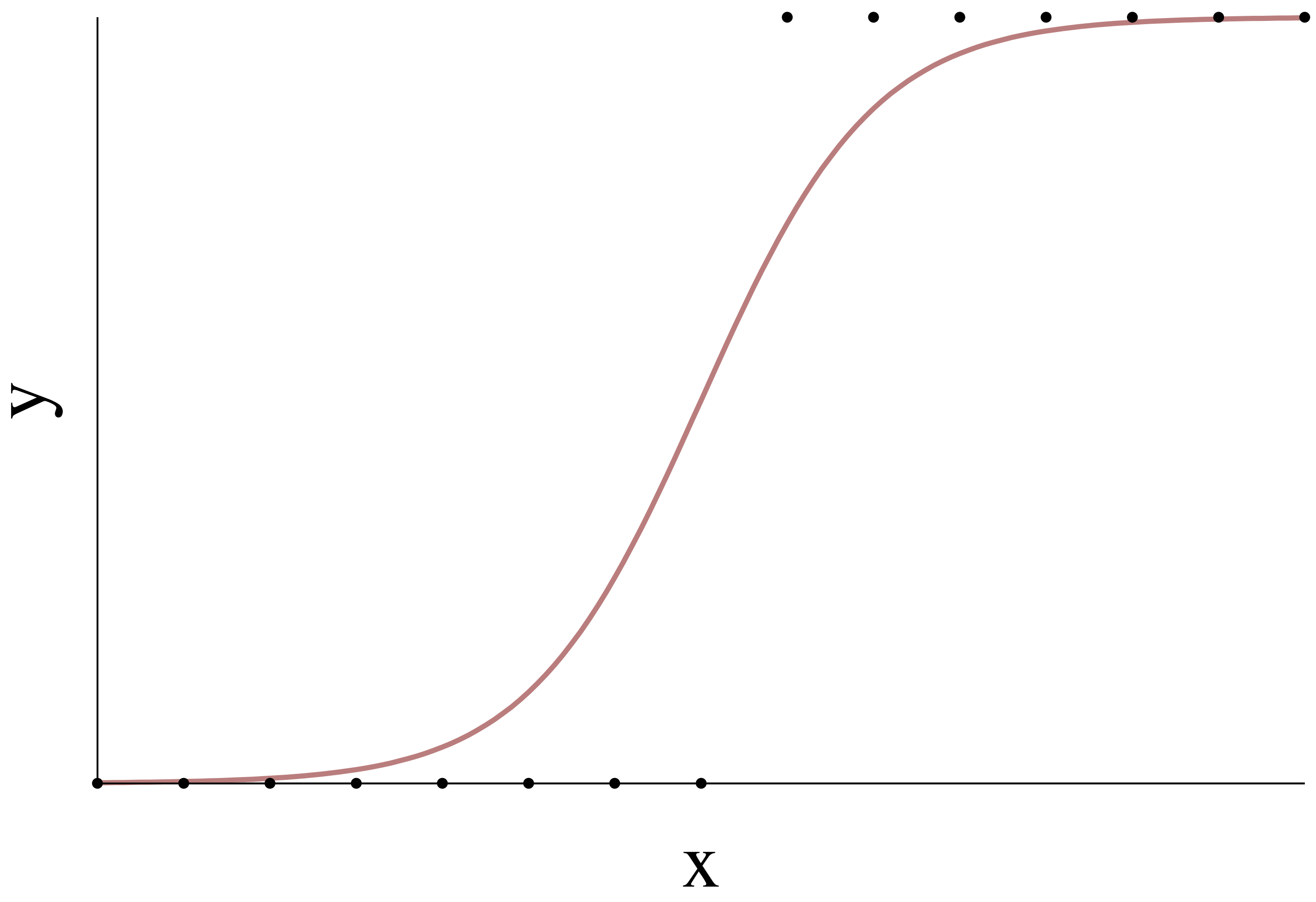
The breadth of these priors is motivated by reasoning about plausible variations. Think *units*.

$$\beta_i \sim \mathcal{N}(0, \omega_i)$$

As with the linear model parameters, weak prior information for the measurement variability is critical.

$$\pi(\sigma) = \text{Half-}\mathcal{N}(0, \tau)$$

General Linear Models



Constrained effective parameters are
not amenable to linear models.

$$\theta \in (a, b)$$

$$\mathbf{X}^T \boldsymbol{\beta} + \alpha \in (-\infty, \infty)$$

But we can generalize linear models with a *link function*.

$$\theta \in (a, b)$$

$$g(\mathbf{X}^T \boldsymbol{\beta} + \alpha) \in (a, b)$$

In the statistics literature link functions are defined by the un-constraining map.

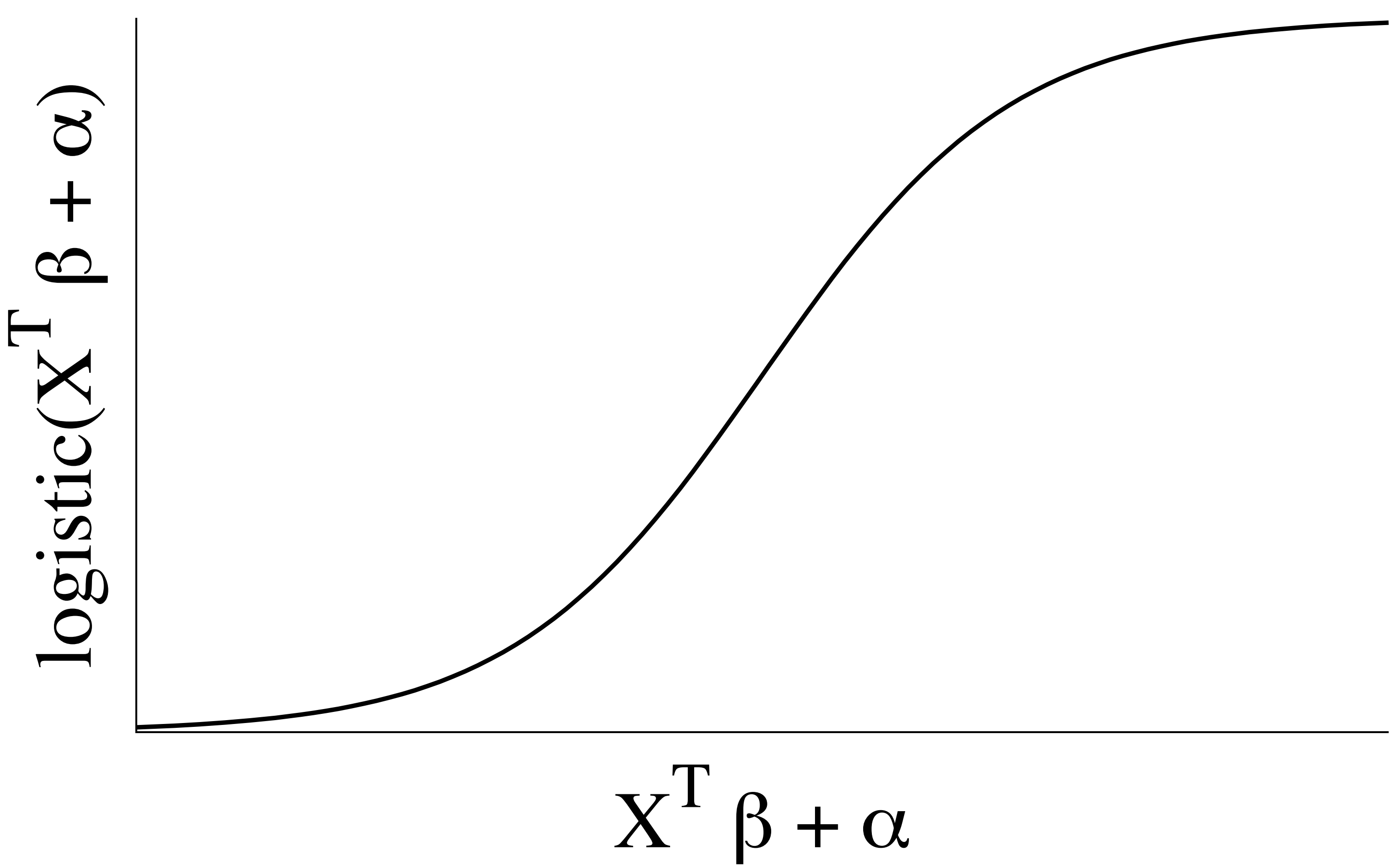
$$g^{-1} : (a, b) \to (-\infty, \infty)$$

While bounded parameters are modeled with the *logit* link function.

$$\mathrm{logit} : (0, 1) \to (-\infty, \infty)$$

$$\mathrm{logistic}\left(\mathbf{X}^T \boldsymbol{\beta} + \alpha\right) \in (0, 1)$$

While bounded parameters are modeled with the *logit* link function.



Success/failure data subject to covariates can be modeled with generalized binomial/Bernoulli models.

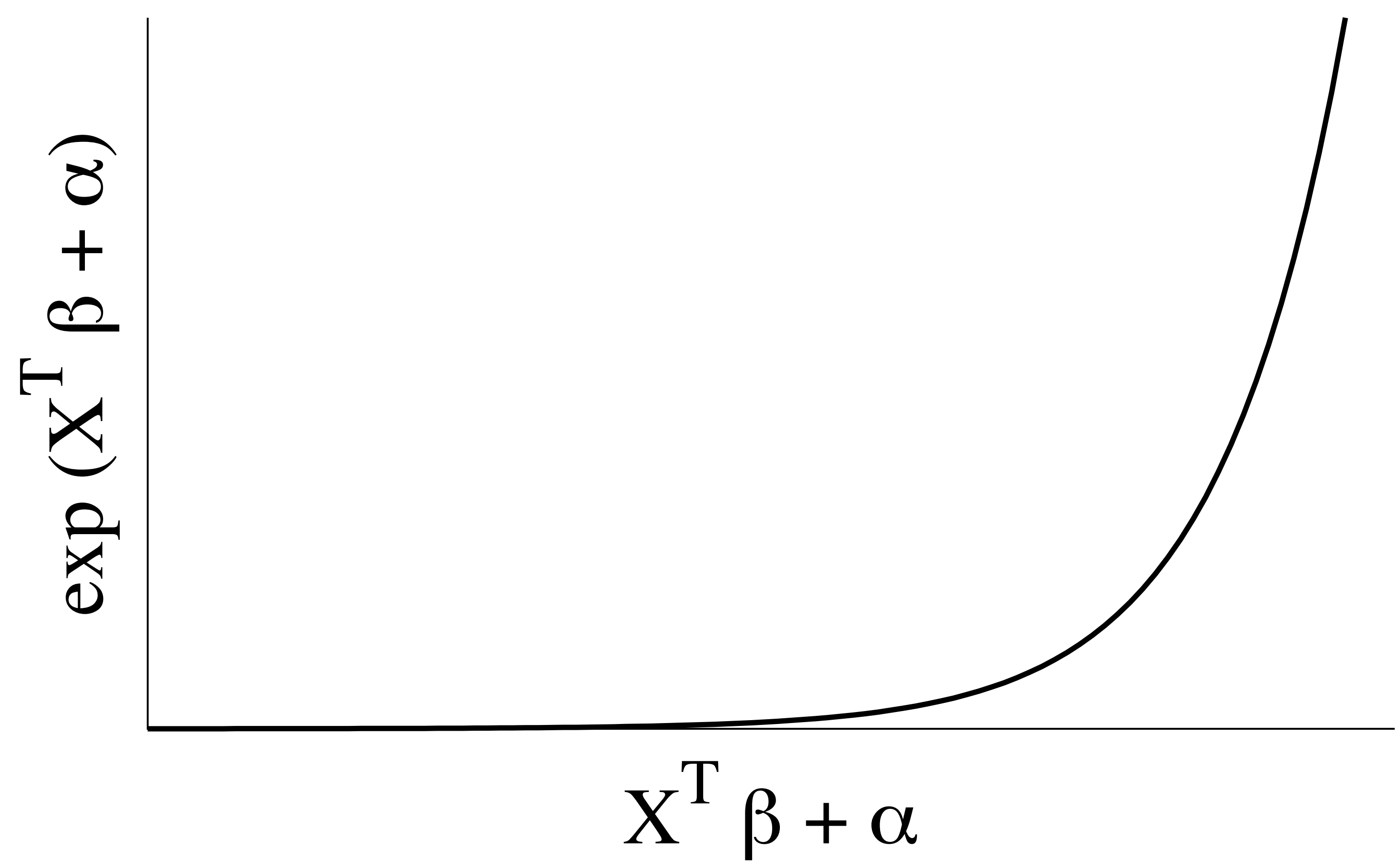
$$\pi(y|\mathbf{X}, \alpha, \boldsymbol{\beta}) = \mathrm{Ber}\big(y|\mathrm{logistic}\big(\mathbf{X}^T \boldsymbol{\beta} + \alpha\big)\big)$$

Positive parameters are modeled with the *log* link function.

$$\log : (0, \infty) \to (-\infty, \infty)$$

$$\exp\left(\mathbf{X}^T \boldsymbol{\beta} + \alpha\right) \in (0, \infty)$$

Positive parameters are modeled with the *log* link function.



Count data whose rate depends on covariates can be modeled with a generalized Poisson model.

$$\pi(y|\mathbf{X}, \alpha, \boldsymbol{\beta}) = \\ \text{Poisson}\big(y\big|\exp\big(\mathbf{X}^T \boldsymbol{\beta} + \alpha\big)\big)$$

In some applications the Poisson likelihood is too restrictive.

$$\text{Poisson}(y|\lambda)$$

$$\mathbb{E}[y] = \lambda$$

$$\text{Var}[y] = \lambda$$

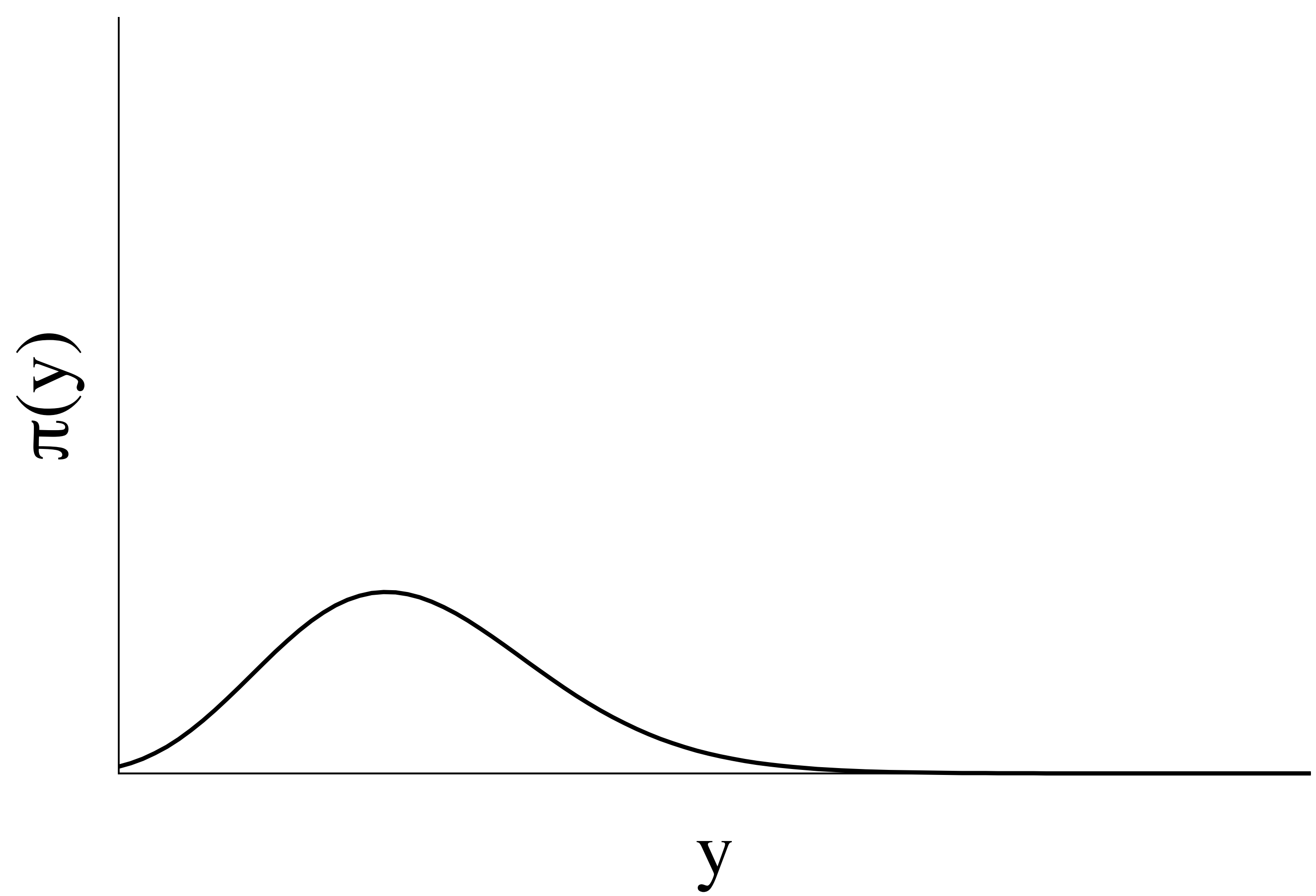
But we can incorporate overdispersion with a generalized negative binomial model.

$$\mathrm{Poisson}(y|\lambda) \qquad \mathrm{NegBin2}(y|\mu,\phi)$$

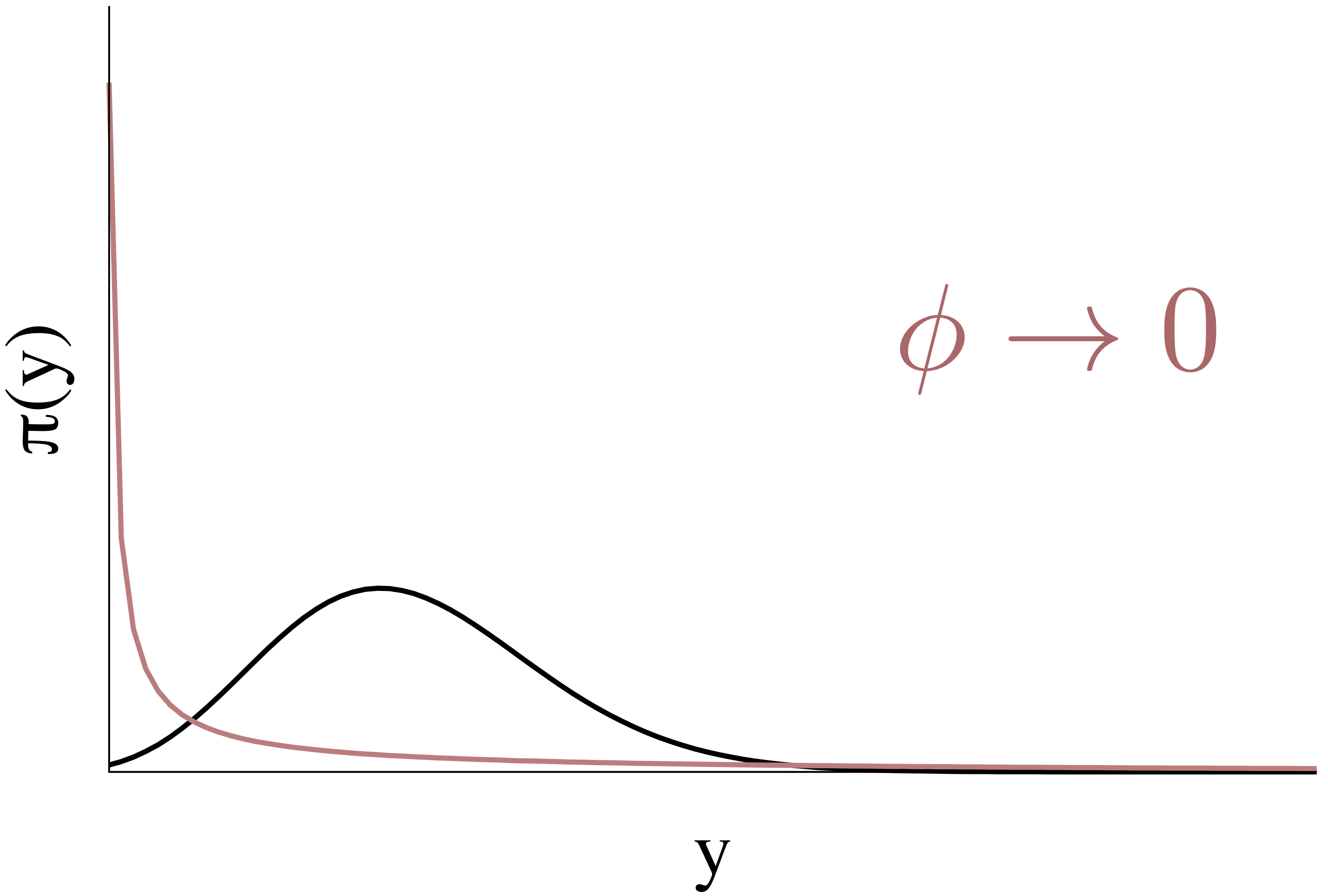
$$\mathbb{E}[y] = \lambda \qquad \mathbb{E}[y] = \mu$$

$$\mathrm{Var}[y] = \lambda \qquad \mathrm{Var}[y] = \mu + \mu^2/\phi$$

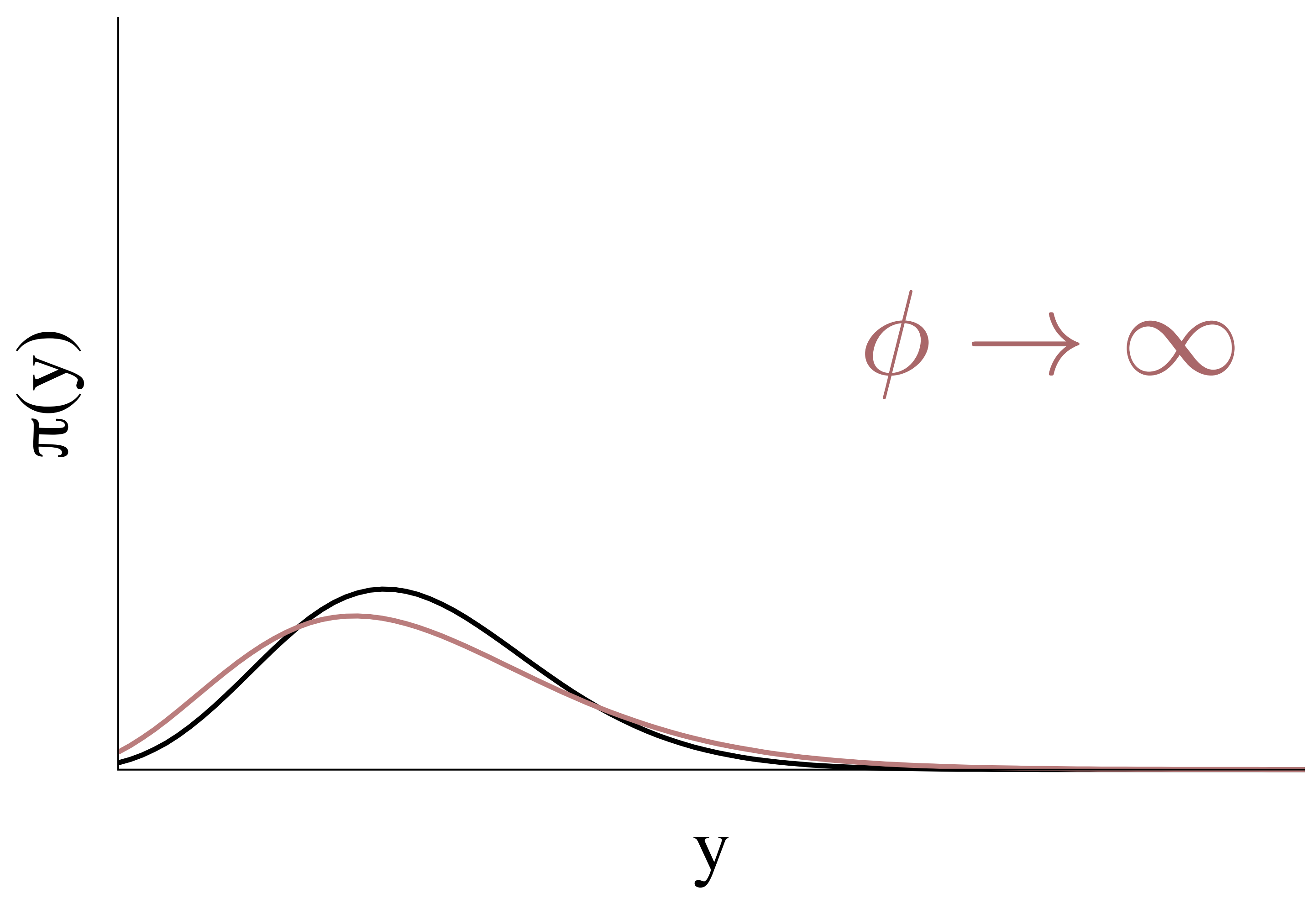
But we can incorporate overdispersion with a generalized negative binomial model.

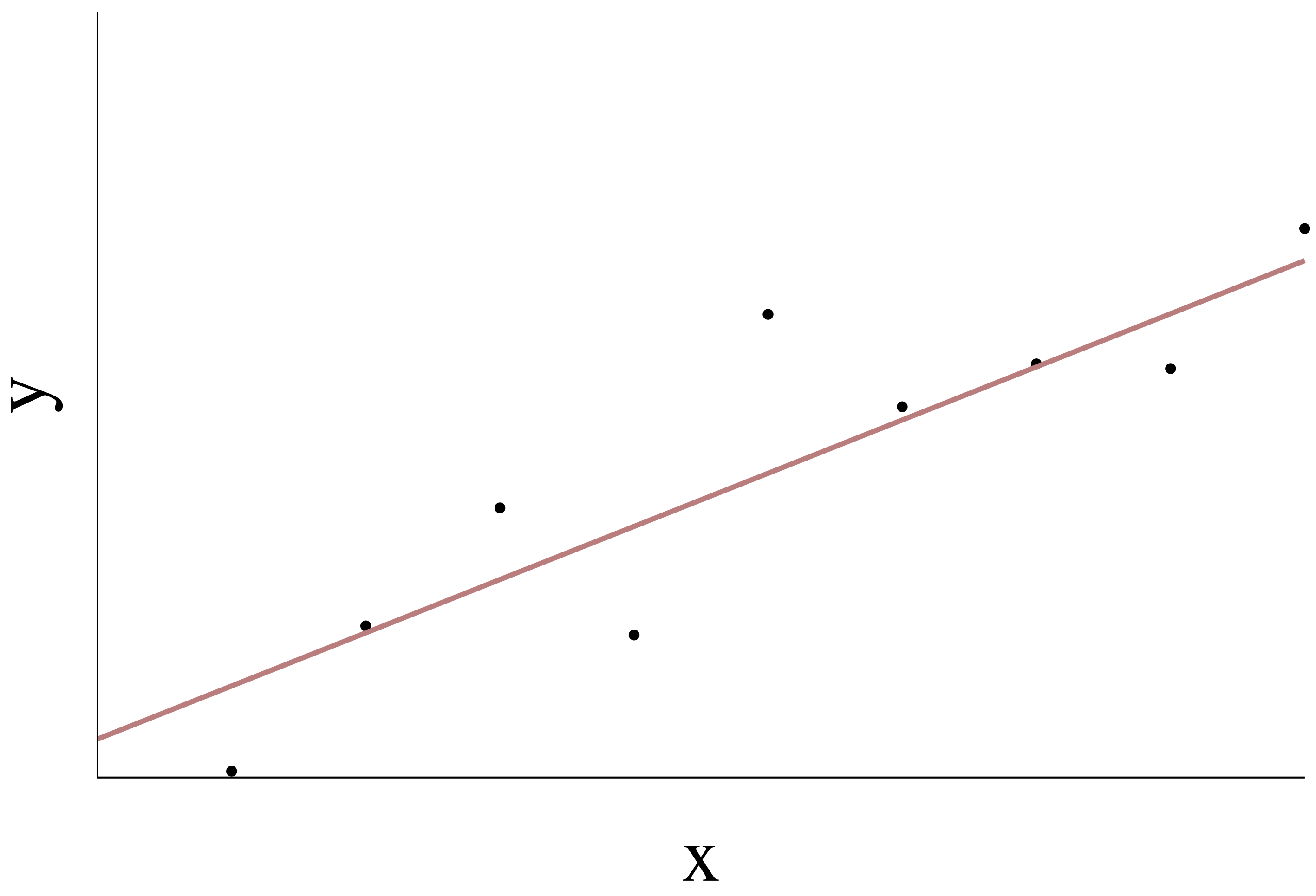


But we can incorporate overdispersion with a generalized negative binomial model.



But we can incorporate overdispersion with a generalized negative binomial model.




y
x