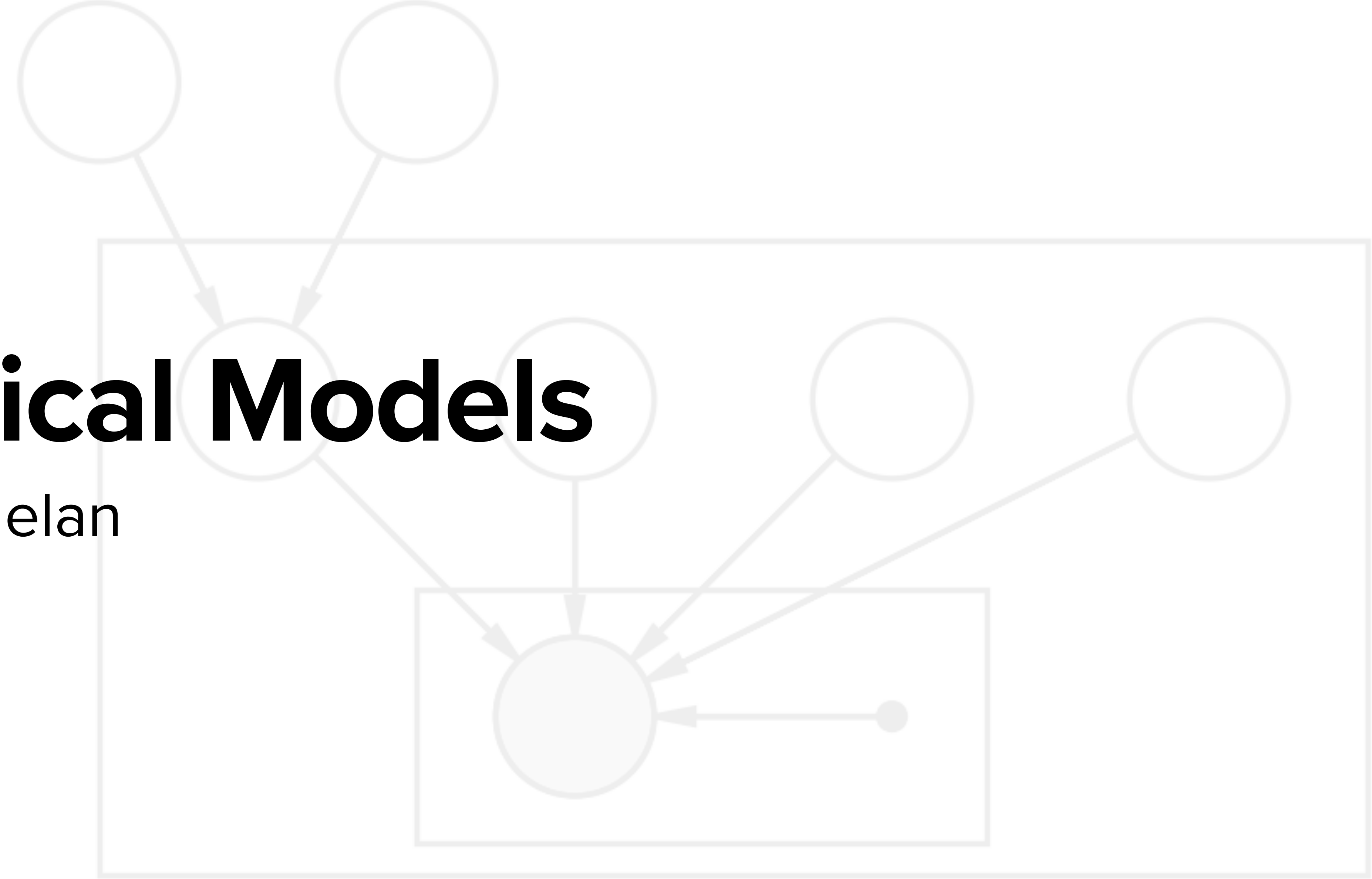
Hierarchical Models

Adrian Price-Whelan

@adrianprw / @adrn



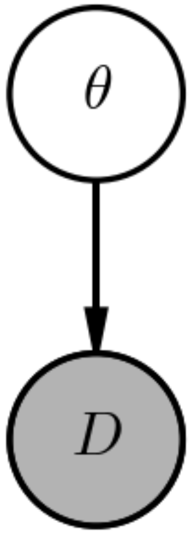
Hierarchical Models

So far, we have worked with *one-level* models: In these cases, the data is generated by some functional model whose parameters are random variables drawn from fixed prior pdfs

In a graphical model, these models have one level of hierarchy

For example:

$$p(D, \theta) = p(D \mid \theta)\, p(\theta)$$

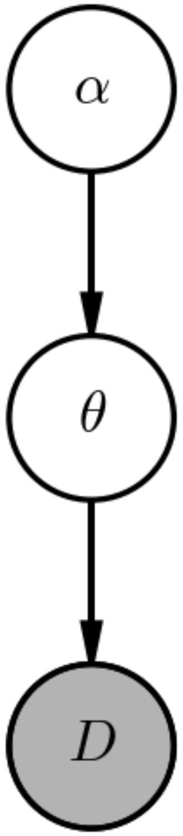


Hierarchical Models

Hierarchical models are also called *multi-level* models because they can have multiple levels of hierarchy, i.e. random variables are generated by distributions whose parameters are random variables generated by distributions ... and so on ...

For example:

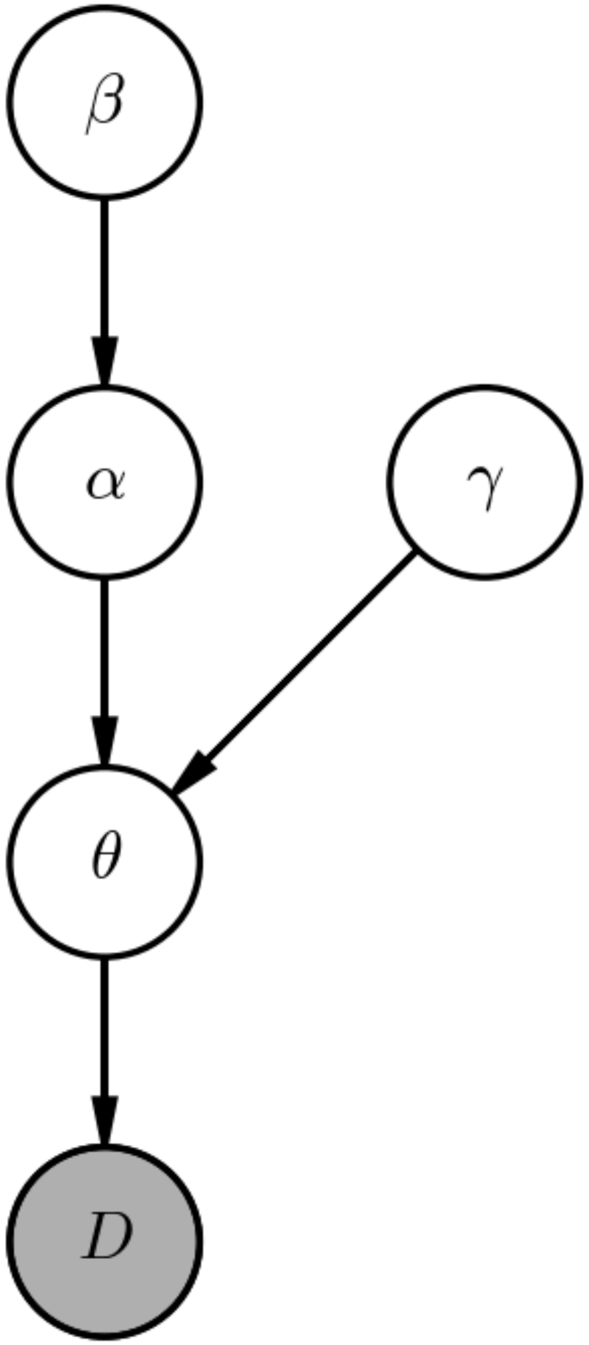
$$p(D, \theta, \alpha) = p(D \mid \theta)\ p(\theta \mid \alpha)\ p(\alpha)$$



"hyper-parameters"

Hierarchical Models

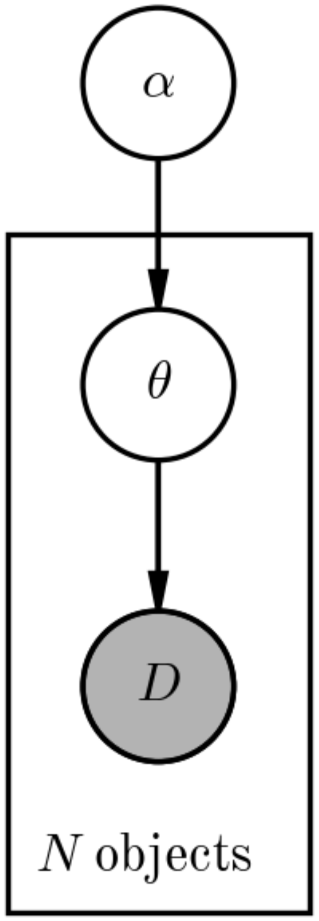
Question: How could we write the factorized probability distribution of this hierarchical model?



Population Models are Hierarchical Models

In astronomy, you will frequently see population models of this form, in which the θ parameters exist for each object in a sample and the α parameters control the population distribution

In these cases, using our (very general) terminology:



Population Models are Hierarchical Models

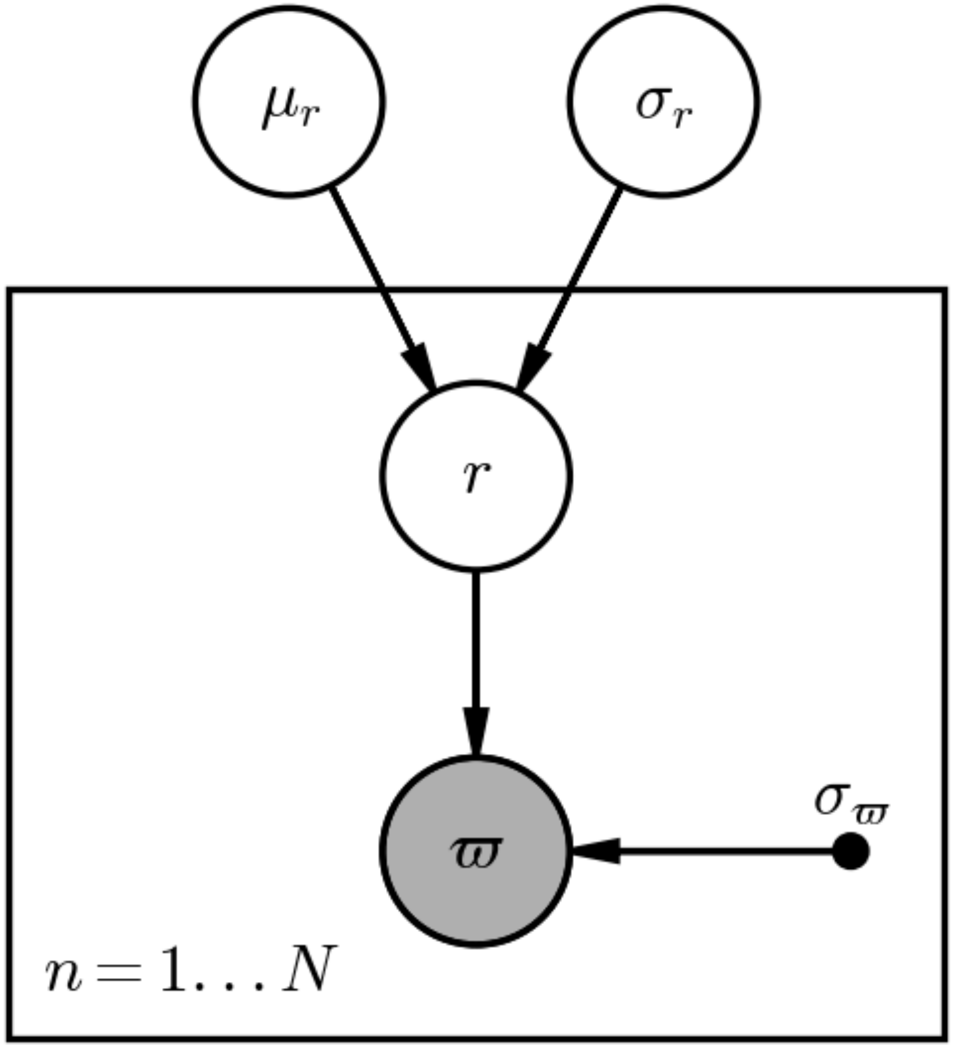
Examples:

- **Exoplanet (or binary-star) period or eccentricity distribution**: The raw data is usually light curves with transits or eclipses, but the parameters one wants to know are the properties of the period distribution or eccentricity distribution of the systems. A model therefore has the true period/eccentricity for each system, but then some parameters that govern the distribution of periods/eccentricities.
- **Velocity dispersion of a galaxy cluster**: The observed data are redshifts of members of a galaxy cluster, but one wants to know the velocity dispersion of the cluster. One way of doing this inference would be to construct a hierarchical model that predicts redshifts given the mean velocity and dispersion of a galaxy cluster by simultaneously inferring the true velocity of each member galaxy given its (noisy) redshift measurement.
- **Color–magnitude fitting / stellar population modeling**: The observed data here are the magnitudes of stars in a number of photometric bands, and maybe also spectroscopic parameters, and the parameters one wants to infer are properties of the stellar population like the distribution of ages, metallicities, and stellar evolutionary stages.

Example: The mean distance of a star cluster

Question: Write an expression for the posterior pdf

$$p(\mu_r, \sigma_r, \boldsymbol{r} | \boldsymbol{\varpi})$$



Example: The mean distance of a star cluster

(switching to Jupyter)

Mixture Models

So far, we have assumed that all objects for which we have data are true members of the population we want to model

(For example, in the star cluster example, we assumed that all stars are members of the cluster)

We very rarely have this luxury!

We typically have either:

- a superposition of multiple structures,
- or a population of interest + a "field" population,
- or a trend of interest + "outliers"

In these contexts (i.e. most contexts!) the models we have used so far are not sufficient for performing population inferences

Mixture Models

One way of dealing with these cases is to use mixture models

Mixture models are one of the most important statistical modeling structures you can learn as a data-oriented researcher!

Let's start with a particular subclass: Gaussian Mixture Models

Gaussian Mixture Models

GMMs are a general and commonly-seen tool for many statistics and machine learning problems

In what we have seen so far, we are used to thinking about single-component normal probability distributions, like:

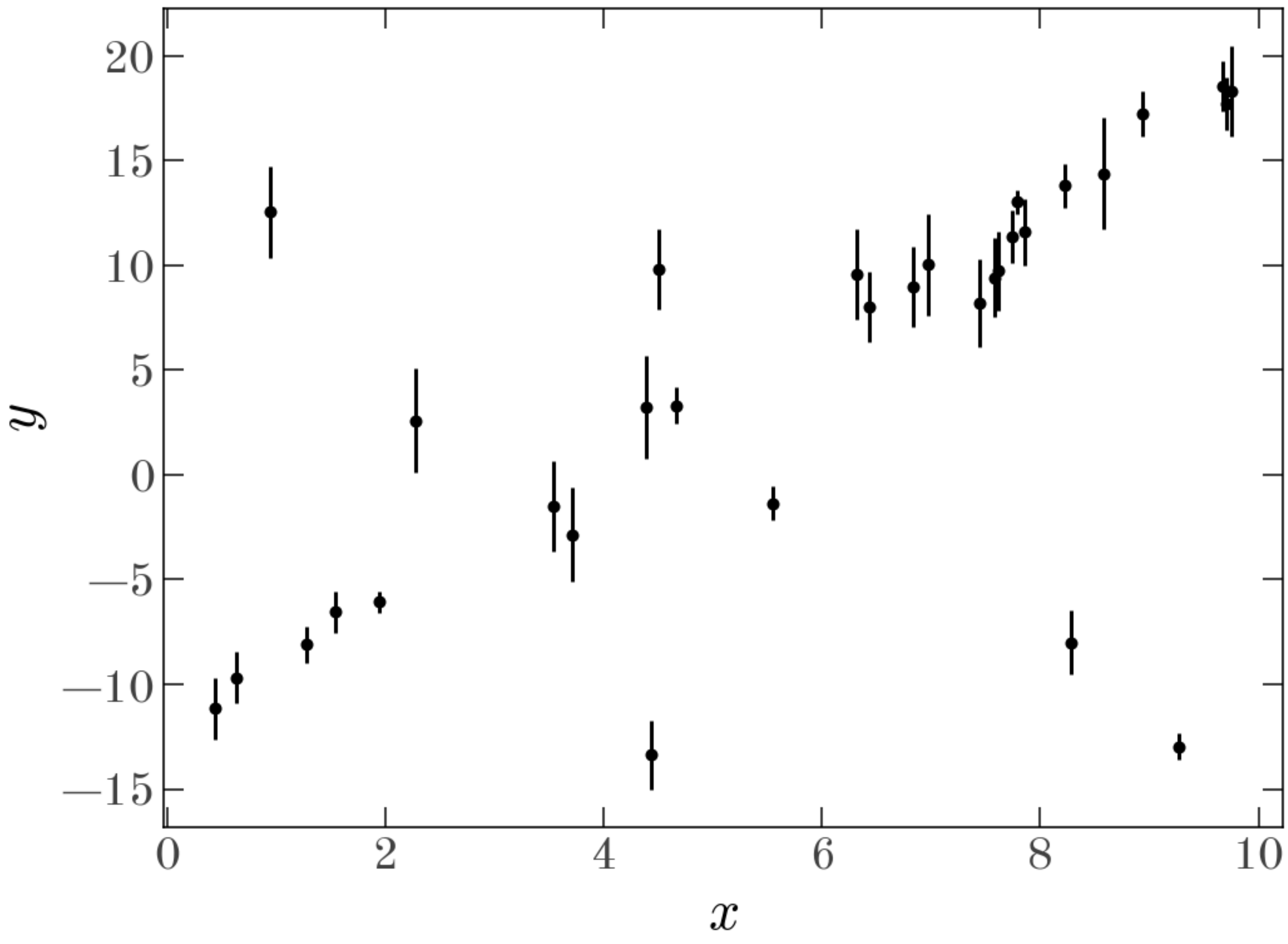
$$p(x \mid \mu, \sigma) = \mathcal{N}(x \mid \mu, \sigma^2)$$

GMMs are instead *sums* of normal distributions, added together with weights such that the integrated density is still 1 (i.e. the mixture pdf is still a valid probability distribution!):

$$p(x|\boldsymbol{\mu}, \boldsymbol{\sigma}) = \sum_{k}^{K} w_k \, \mathcal{N}(x|\mu_k, \sigma_k^2)$$

$$\sum_{k}^{K} w_k = 1$$

Example: Fitting a line to data with outliers



Example: Fitting a line to data with outliers

(switching to Jupyter)