Lecture Recording

- **Note: These lectures will be recorded and posted onto the IMPRS website**

- Dear participants,

- We will record all lectures on "*Making sense of data: introduction to statistics for gravitational wave astronomy*", including possible Q&A after the presentation, and we will make the recordings publicly available on the IMPRS lecture website at:

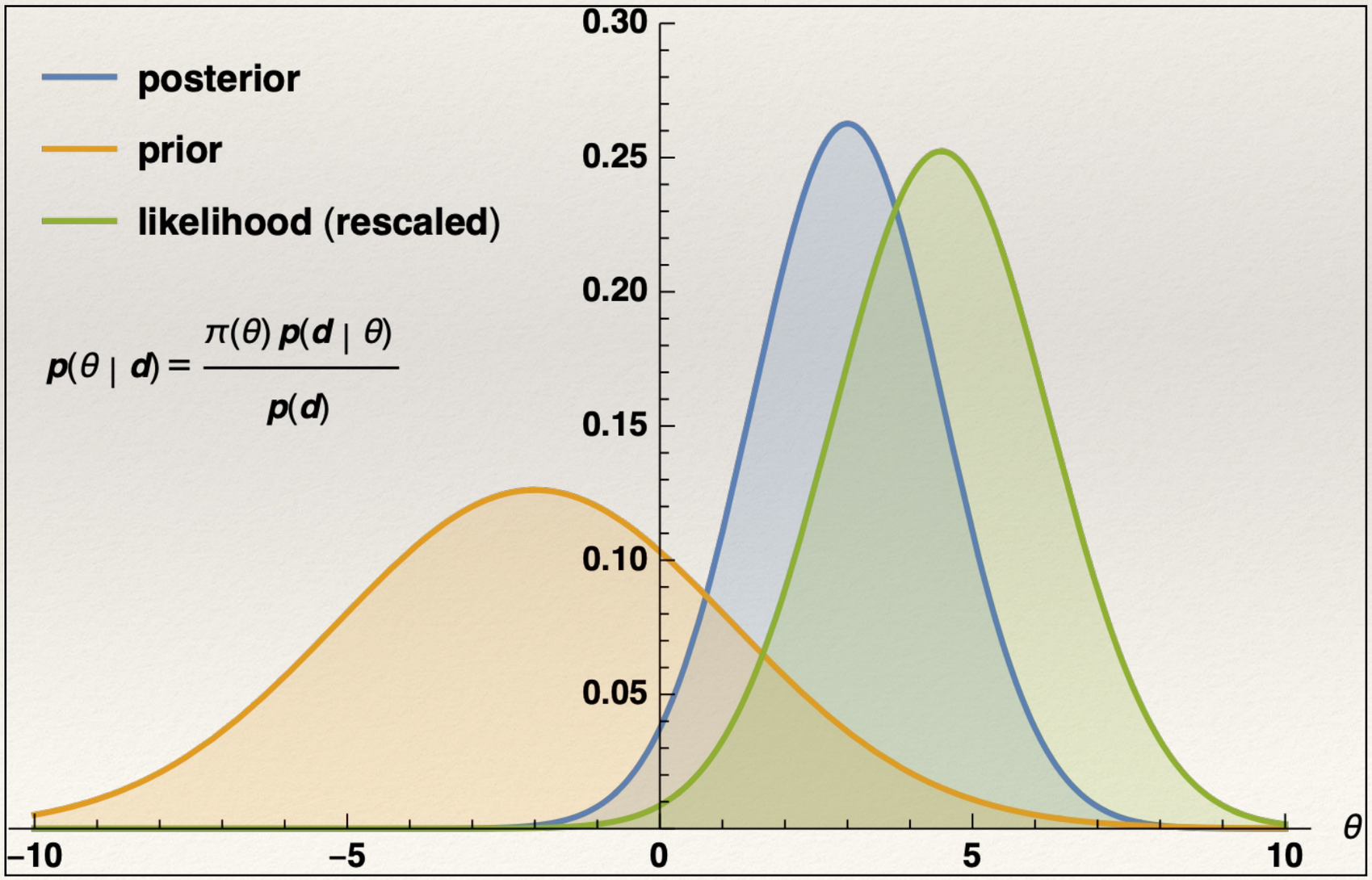
  - ✍ https://imprs-gw-lectures.aei.mpg.de/2023-making-sense-of-data/

- By participating in this Zoom meeting, you are giving your explicit consent to the recording of the lecture and the publication of the recording on the course website.

Making sense of data: introduction to statistics for gravitational wave astronomy
Part II: Bayesian statistics
Lecture 3: Bayesian inference part II

AEI IMPRS Lecture Course

Alexandre Toubiana atoubiana@aei.mpg.de



Bayesian hypothesis testing

- The denominator in Bayes' Theorem

$$p(\theta|x, M) = \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)}$$

- is the **Bayesian evidence**

$$p(x|M) = \int p(x|\theta, M)p(\theta|M)\mathrm{d}\theta$$

- Here we have explicitly introduced the model M to emphasise that the result depends on the model assumed. The evidence is the probability that the observed data would have been produced under the given model and so can be used for **model selection**.

- Models are compared using the **posterior odds ratio**

$$\mathcal{O}_{12} = \frac{p(x|M_1)p(M_1)}{p(x|M_2)p(M_2)}$$

- The first term is the **Bayes Factor**. The second is the **prior odds ratio**.

Bayesian hypothesis testing

- The interpretation of the posterior odds ratio is somewhat arbitrary, but Kass and Rafferty (1995) suggested the following scale:

Bayes Factor	Interpretation
< 3	No evidence of M_1 over M_2
> 3	Positive evidence for M_1
> 20	Strong evidence for M_1
> 150	Very strong evidence for M_1

- Interpreting the Bayes factor as a ratio of probabilities, these thresholds correspond to "p-values" of 0.25, 0.05, 0.007, but the interpretation is different.

- In practice, posterior odds ratios can also be used as a test statistic, with significance and power computed via simulation in the usual (frequentist) way.

Bayesian hypothesis testing

- Computing Bayesian evidences is challenging. These can be estimated using the **harmonic mean of the likelihood** of samples from the posterior

$$\frac{1}{p(x|M)} = \int \frac{1}{p(x|\theta, M)} \frac{p(x|\theta, M)p(\theta|M)}{p(x|M)} \mathrm{d}\theta \simeq \frac{1}{N} \sum_{\theta_i \sim p(\theta|x,M)} \frac{1}{p(x|\theta_i, M)}$$

- Necessarily, there are more posterior samples where the likelihood and hence posterior are higher.
- Regions where the likelihood is small are less well sampled and subject to more Monte Carlo error. This makes the above expression very unstable and potentially inaccurate.
- Other techniques, such as **thermodynamic integration** and **nested sampling**, have been developed to overcome these problems and produce robust evidence estimates.

Bayesian hypothesis testing

- **Example**: Normal models. Suppose we have a 2-dimensional likelihood

$$p(x|\theta) = \frac{\sqrt{1-\rho^2}}{2\pi\sigma_1\sigma_2}\exp\left[-\frac{1}{2}\left(\frac{(x_1-\theta_1)^2}{\sigma_1^2} + 2\frac{\rho(x_1-\theta_1)(x_2-\theta_2)}{\sigma_1\sigma_2} + \frac{(x_2-\theta_2)^2}{\sigma_2^2}\right)\right]$$

- and set priors of the form

$$p(\theta_1) = \frac{1}{2\Delta_1}, \qquad p(\theta_2) = \frac{1}{2\Delta_2}$$

- and we want to test the models

$$M_1 : \theta_2 = 0, \qquad M_2 : \theta_2 \in [-\Delta_2, \Delta_2]$$

- The evidence ratio assuming $\Delta_1 \gg \sigma_1$ can be found to be (see lecture notes)

$$\mathcal{O}_{12} = \frac{2\Delta_2}{\sigma_2}\sqrt{1-\rho^2}\exp\left[-\frac{1}{2}\frac{(1-\rho^2)x_2^2}{\sigma_2^2}\right]$$

- size of extra dimension
- improvement to fit

- This can be interpreted as automatically implementing *Occam's Razor*.

Hierarchical Models

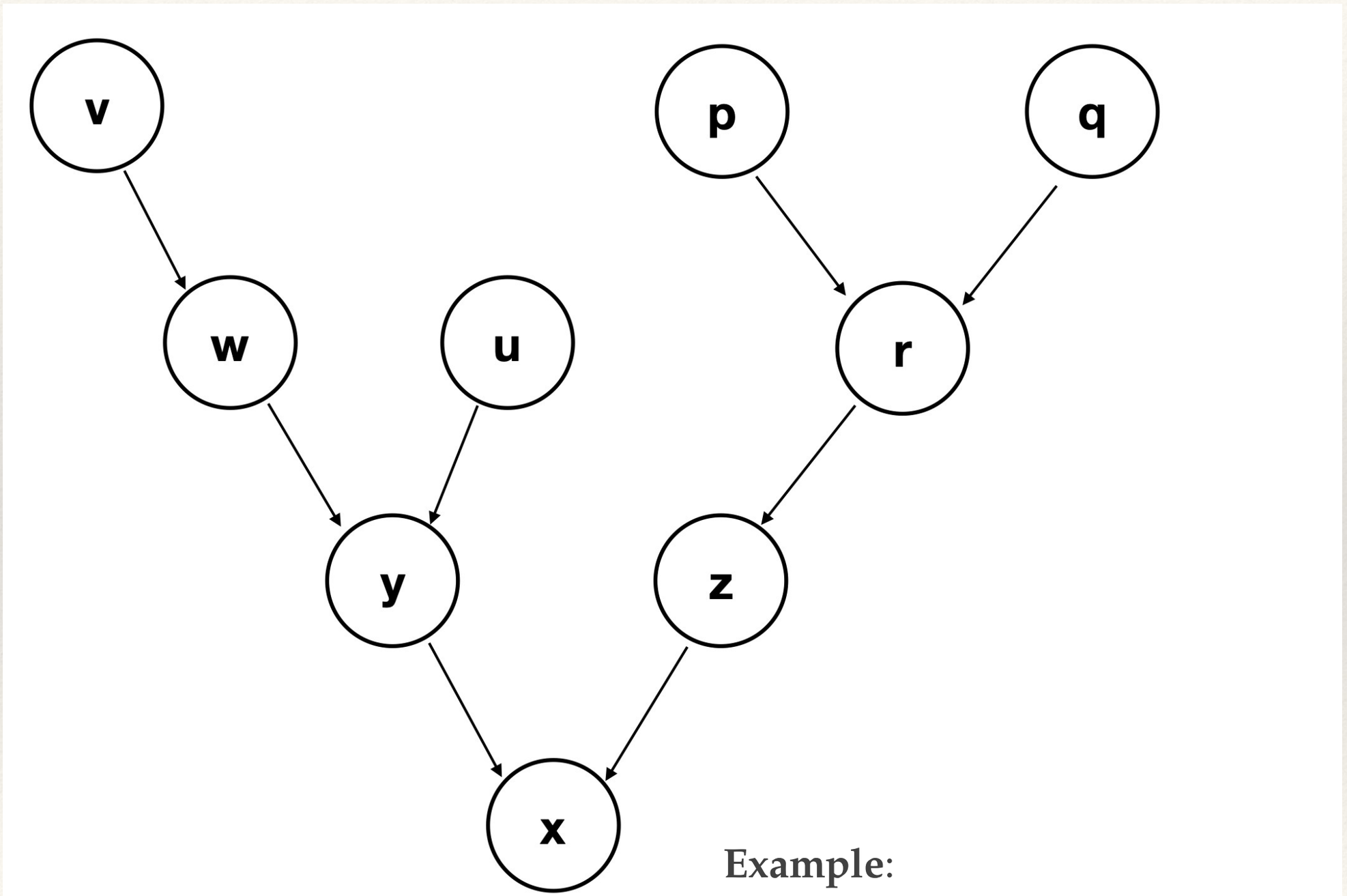
- Often the prior for a single data set represents a model for a population of events, e.g., compact binary coalescences.
- The parameters of that prior encode the details of the population and are also of interest. This leads to the notion of a **hierarchical model**.
- In a hierarchical model, the parameters of the prior (termed **hyperparameters**) are regarded as random variables, on which a **hyperprior** is defined. This can be continued ad infinitum - using another hyperprior on the hyperparameters of the first hyperprior etc.

$$p(x|\lambda) = \int p(x|\theta)p(\theta|\lambda)\mathrm{d}\theta$$

$$p(\{x\}|\lambda) = \prod_i p(x_i|\lambda)$$

$$\mathrm{Bayes'\ theorem} : p(\lambda|\{x\}) = \frac{p(\{x\}|\lambda)p(\lambda)}{p(\{x\})}$$

Graphical Model



Example:

p(p, q, r, s, t, u, v, w, x, y, z) =

p(x | y, z) p(y | u, w) p(z | r) p(w | v) p(r | p, q) p(v) p(u) p(p) p(q)

Selection Effects

- No instrument is arbitrarily sensitive and therefore some types of source are easier to see than others. This is important to remember in hierarchical models for populations when we are combining only **detected** events.
- There are two ways to think about selection effects.
- One way is to acknowledge that we only include "detected" events in the analysis and then write down a likelihood for detected events. This must integrate to 1 over all "detected" or "above threshold" data sets.

$$p(x|\lambda, \mathrm{obs}) = \frac{p(x|\theta)}{p_s(\lambda)}, \quad \text{where } p_s(\lambda) = \int_{x>\mathrm{threshold}} \int p(x|\theta) p(\theta|\lambda) \mathrm{d}x \mathrm{d}\theta$$

- This framework assumes a priori that the number of detected events contains no information about the parameters of interest.
- Obs: selection effects do not impact the parameter estimation of single events.

Selection Effects

- Alternatively we write down the likelihood for all events, both **detected** events (indexed by i) and **undetected** events (indexed by j)

$$p(\{\theta_i\},\{\theta_j\},\{x_i\},\{x_j\},|\lambda) \propto \left[\prod_i^{N_{\rm obs}} p(x_i|\theta_i)\frac{{\rm d}N}{{\rm d}\theta_i}(\lambda)\right]\left[\prod_j^{N_{\rm nobs}} p(x_j|\theta_j)\frac{{\rm d}N}{{\rm d}\theta_j}(\lambda)\right]\exp[-N(\lambda)]$$

- Marginalising over the unobserved data we obtain

$$p(\{\theta_i\},\{x_i\}|\lambda) \propto \left[\prod_i^{N_{\rm obs}} p(x_i|\theta_i)\frac{{\rm d}N}{{\rm d}\theta_i}(\lambda)\right]\frac{N_{\rm ndet}^{N_{\rm nobs}}}{N_{\rm nobs}!}\exp[-N(\lambda)]$$

$$N_{\rm ndet}(\lambda) = \int_{x<{\rm threshold}}\int p(x|\theta)\frac{{\rm d}N}{{\rm d}\theta}(\lambda){\rm d}x{\rm d}\theta$$

- Marginalising over the unknown number of unobserved events then gives

$$p(\{\theta_i\},\{x_i\}|\lambda) \propto \left[\prod_i^{N_{\rm obs}} p(x_i|\theta_i)\frac{{\rm d}N}{{\rm d}\theta_i}(\lambda)\right]\exp[-N_{\rm det}(\lambda)]$$

Selection Effects

- Writing

$$\frac{\mathrm{d}N}{\mathrm{d}\theta} = Np(\theta|\lambda)$$

- and introducing a scale-invariant prior on the overall rate

$$p(N) \propto \frac{1}{N}$$

- and noting

$$N_{\mathrm{det}}(\lambda) = \int_{x>\mathrm{threshold}} \int p(x|\theta) \frac{\mathrm{d}N}{\mathrm{d}\theta}(\lambda) \mathrm{d}x \mathrm{d}\theta = N p_s(\lambda)$$

- After marginalising over $\{\theta_i\}$, we recover the previous result.

Hierarchical Bayesian analysis

- Including rates:

$$p(\lambda|\{x\}) \propto p(\lambda) N_{\rm det}(\lambda)^{N_{\rm obs}} \exp[-N_{\rm det}(\lambda)] \left[\prod_i^{N_{\rm obs}} \int p(x_i|\theta_i) p(\theta_i|\lambda) {\rm d}\theta\right]$$

- Marginalising over the rate:

$$p(\lambda|\{x\}) \propto \frac{p(\lambda)}{p_s(\lambda)^{N_{\rm obs}}} \left[\prod_i^{N_{\rm obs}} \int p(x_i|\theta_i) p(\theta_i|\lambda) {\rm d}\theta\right]$$

- We often note $p_{\rm pop}(\theta|\lambda)$ the population prior

Example: measuring deviations

- **Example**: **Constraining GR with GWs** *GR deviations might be too weak to be measured in single events with current detectors, combining information from different observations would increase our measurement power. Different approaches have been proposed, see e.g. arXiv:2204.10742.*

- We consider one deviation parameter α, and assume that the marginalised likelihood on α is Gaussian:

$$p(x|\alpha) = \frac{1}{\sqrt{2\pi}\Sigma} \exp\left[-\frac{1}{2}\frac{(x-\alpha)^2}{\Sigma^2}\right]$$

- When doing single event parameter estimation we assume a flat prior on α:

$$p(\alpha) = \frac{1}{2\Delta_\alpha}$$

- Goal: test if GR is right, i.e. $\alpha = 0$

- Here we do not account for selection effects (i.e we do inference on the observed population)

Example: measuring deviations

- 1st possibility: combine Bayes' factors: $\mathcal{O}^{\rm GR}_{i,{\rm nGR}} = \frac{2\Delta_\alpha}{\sqrt{2\pi}\Sigma}\exp\left[-\frac{1}{2}\frac{x_i^2}{\Sigma^2}\right]$

$$\mathcal{O}^{\rm GR}_{\rm tot,nGR} = \prod_i^{N_{\rm obs}} O^{\rm GR}_{i,{\rm nGR}} = \exp\left[\frac{N_{\rm obs}}{2}\left(\log\left(\frac{2\Delta_\alpha}{\sqrt{2\pi}\Sigma}\right) - \frac{<\{x\}>^2}{\Sigma^2} - \frac{{\rm var}(\{x\})}{\Sigma^2}\right)\right]$$

- Taking $\Delta_\alpha = 5\Sigma : \mathcal{O}^{\rm GR}_{\rm tot,nGR} \simeq \exp\left[\frac{N_{\rm obs}}{2}\left(1.3 - \frac{<\{x\}>^2}{\Sigma^2} - \frac{{\rm var}(\{x\})}{\Sigma^2}\right)\right]$

- We might erroneously build confidence that GR is right !

Example: measuring deviations

- We can re-interpret this result within a hierarchical framework. We treat $(\alpha_1, \alpha_2, ..., \alpha_{N_{\rm obs}})$ as $\mathrm{N_{obs}}$ independent hyperparamers. Then:

$$p(\alpha_1, \alpha_2, ..., \alpha_n | \{x\}) = \frac{1}{(2\pi)^{\frac{N_{\rm obs}}{2}} \sum^{N_{\rm obs}}} \exp\left[-\frac{1}{2}\sum_i^{N_{\rm obs}} \frac{(x_i - \alpha_i)^2}{\Sigma^2}\right]$$

- GR coresponds to $\alpha_1 = \alpha_2 = ... = \alpha_{N_{\rm obs}} = 0$, the Bayes' factor for it is:

$$\mathcal{O}^{\rm GR}_{\rm nGR} = \exp\left[\frac{N_{\rm obs}}{2}\left(\log\left(\frac{2\Delta_\alpha}{\sqrt{2\pi}\Sigma}\right) - \frac{<\{x\}>^2}{\Sigma^2} - \frac{\mathrm{var}(\{x\})}{\Sigma^2}\right)\right]$$

Example: measuring deviations

- Now, we assume $\alpha \sim \mathcal{N}(\mu, \sigma)$. GR coresponds to $\mu = \sigma^2 = 0$.
- We estimate (μ, σ^2)using a hierrachical bayesian analysis (taking a flat prior):

$$p(\mu, \sigma^2|\{x\}) \propto \frac{1}{(2\pi(\sigma^2+\Sigma^2))^{\frac{N_{\rm obs}}{2}}} \exp\left[-\frac{N_{\rm obs}}{2}\frac{(\mu - <\{x\}>)^2 + {\rm var}(\{x\})}{\sigma^2+\Sigma^2}\right]$$

- It is maximum for $\mu = <\{x\}>,\ \sigma^2 = {\rm var}(\{x\}) - \Sigma^2$.
- In the limit of large $N_{\rm obs}$, Gaussian approximation:

$$p(\mu, \sigma^2|\{x\}) = \frac{N_{\rm obs}}{2\sqrt{2}\pi{\rm var}(\{x\})^{\frac{3}{2}}} \exp\left[-\frac{N_{\rm obs}}{2}\left(\frac{(\mu - <\{x\}>)^2}{{\rm var}(\{x\})} + \frac{(\sigma^2+\Sigma^2-{\rm var}(\{x\}))^2}{2{\rm var}(\{x\})^2}\right)\right]$$

- The Bayes' factor is then:
  - smaller dimensionality penalty

$$\mathcal{O}^{\rm GR}_{\rm nGR} = \frac{4N_{\rm obs}\Delta_\mu\Delta_{\sigma^2}}{2\sqrt{2}\pi{\rm var}(\{x\})^{\frac{3}{2}}} \exp\left[-\frac{N_{\rm obs}}{2}\left(\frac{<\{x\}>^2}{{\rm var}(\{x\})} + \frac{(\Sigma^2-{\rm var}(\{x\})^2)^2}{2{\rm var}(\{x\})^2}\right)\right]$$

Example: measuring deviations

- Alternatively, we can compute a generalised quantile to quantify agreement with GR:

$$\mathcal{Q}_{\rm GR} = \int_{p(\lambda|\{x\})<p(\lambda_{\rm GR}|\{x\})} p(\lambda|\{x\}){\rm d}\lambda$$

- The closer it is to 1, the better the agreement with GR.

- Exploiting the fact that for a $n-$dimensional Gaussian distribution $p_n(\lambda)$, $2(\max(\ln(p_n)) - \ln(p_n)) \sim \chi^2(n)$:

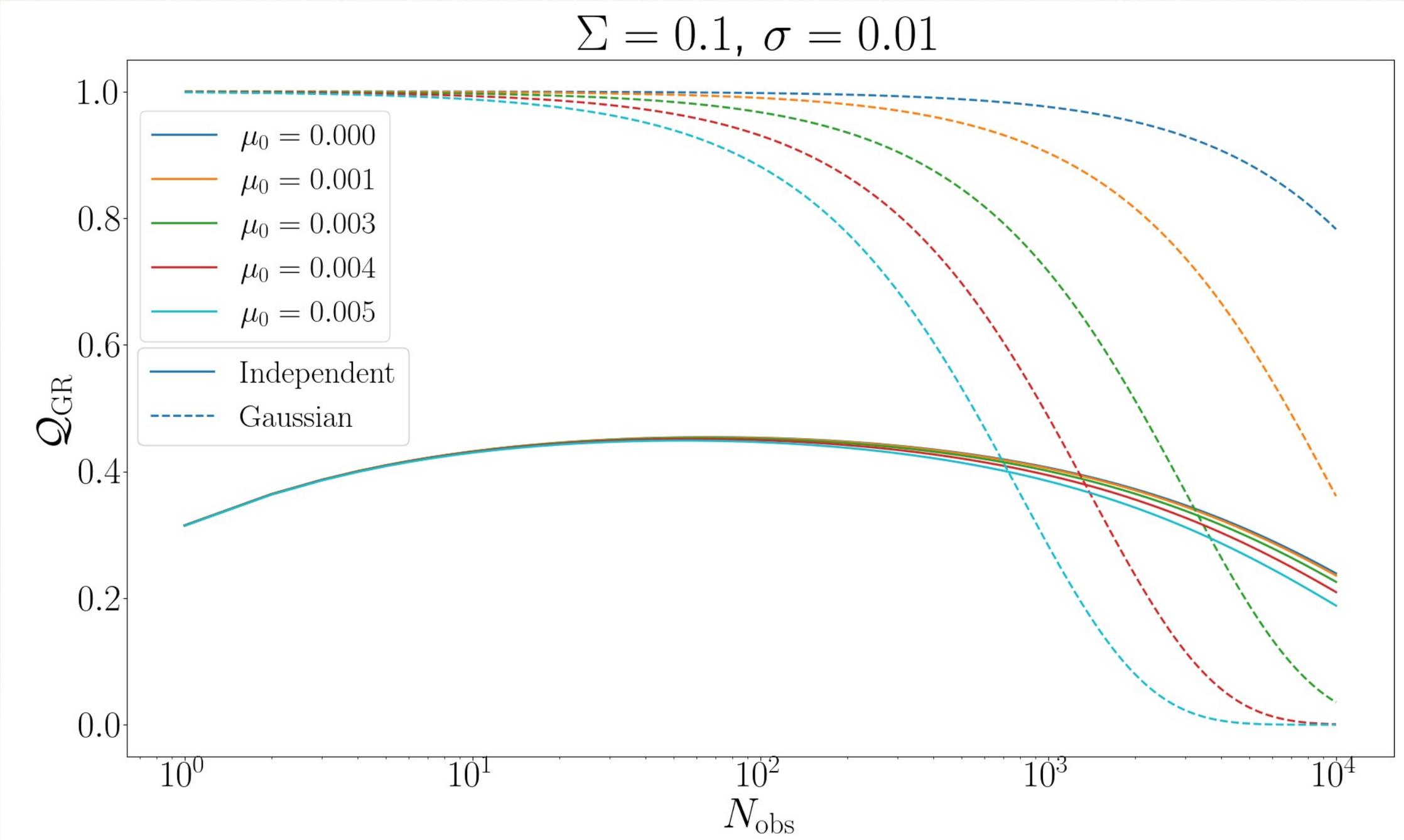
  - Taking $(\alpha_1, \alpha_2, ..., \alpha_{N_{\rm obs}})$ as $\rm N_{obs}$ independent hyperparamers:

$$\mathcal{Q}_{\rm GR} \simeq \frac{1}{2\sqrt{\pi}} \exp\left[-\frac{N_{\rm obs}}{4}\left(\frac{<\{x\}>^2 + {\rm var}(\{x\})}{\Sigma^2} - 1\right)^2\right] \left(\frac{\sqrt{N_{\rm obs}}}{2}\left(\frac{<\{x\}>^2 + {\rm var}(\{x\})}{\Sigma^2} - 1\right)\right)^{-1}$$

  - Taking $\alpha \sim \mathcal{N}(\mu, \sigma)$: For small deviation, and large $N_{\rm obs}$

$$\mathcal{Q}_{\rm GR} \simeq \exp\left[-\frac{N_{\rm obs}}{2}\left(\frac{<\{x\}>^2}{{\rm var}(\{x\})} + \frac{(\Sigma^2 - {\rm var}(\{x\}))^2}{2{\rm var}(\{x\})^2}\right)\right]$$

Example: measuring deviations



Example: measuring deviations

- Priors (in the broad sense) matter a lot!
  - The difference is not only $\mathcal{Q}_{\rm GR}$ vs $\mathcal{O}^{\rm GR}_{\rm nGR}$ (see *arXiv:2204.10742*)
  - Gaussian model reduces dimensionality at the cost of a strong assumption.
- The generalised quantile approach is valid only for “nested models”.
- Dimensionality penalty is incorporated in reversible-jump MCMC via the acceptance ratio:

$$\frac{p(\mathbf{d}|\Lambda_{k+1})q(\lambda_{k+1})}{p(\mathbf{d}|\Lambda_k)p(\lambda_{k+1})}$$

Predictive checking

- It is natural to want to test if the assumed model is a good fit to the data. In a Bayesian context this is achieved through **predictive checking**.
- The **prior predictive distribution** is defined by

$$p(x) = \int p(x|\lambda)p(\lambda)\mathrm{d}\lambda$$

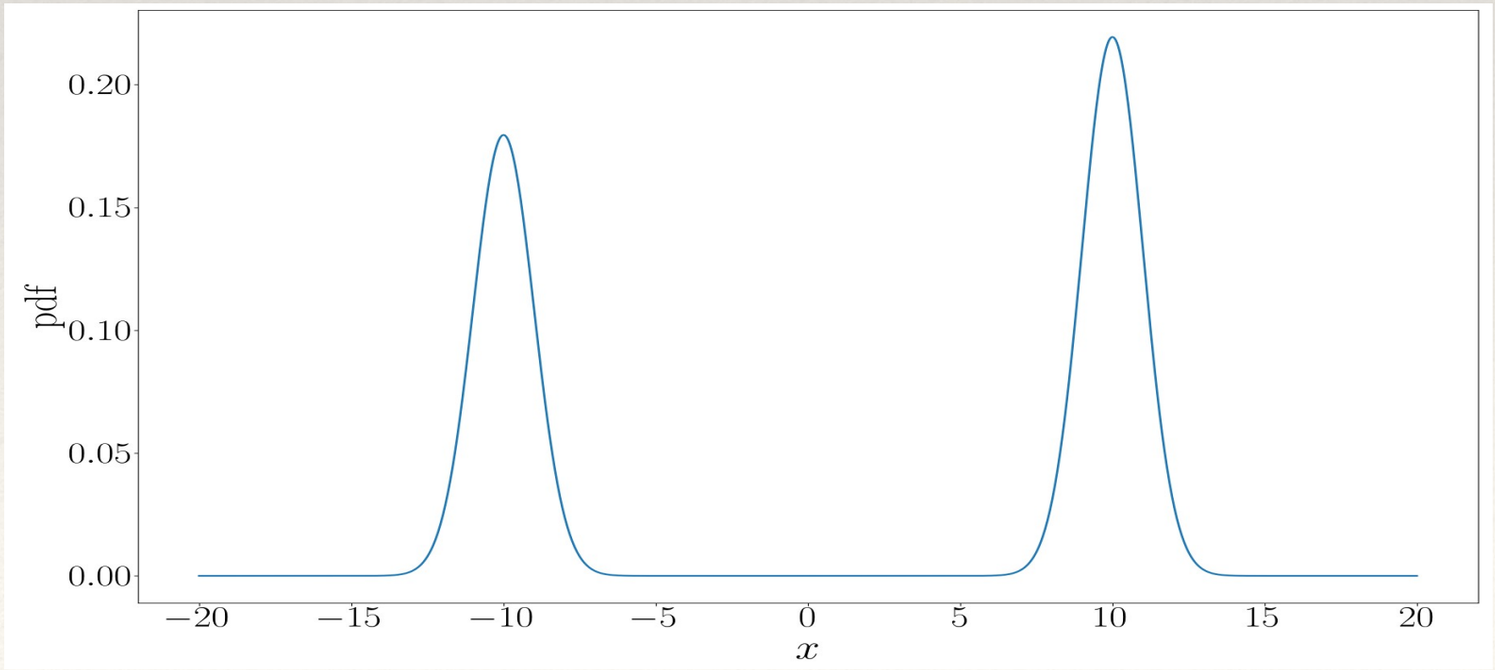
- This is the distribution of observed data sets within the model assumed in the prior. If the observed data is not very consistent with this distribution, the prior parameters might need to be adjusted.
- The **posterior predictive distribution** is defined similarly

$$p(x_{\mathrm{new}}|\{x\}_{\mathrm{old}}) = \int p(x_{\mathrm{new}}|\lambda)p(\lambda|\{x\}_{\mathrm{old}})\mathrm{d}\lambda$$

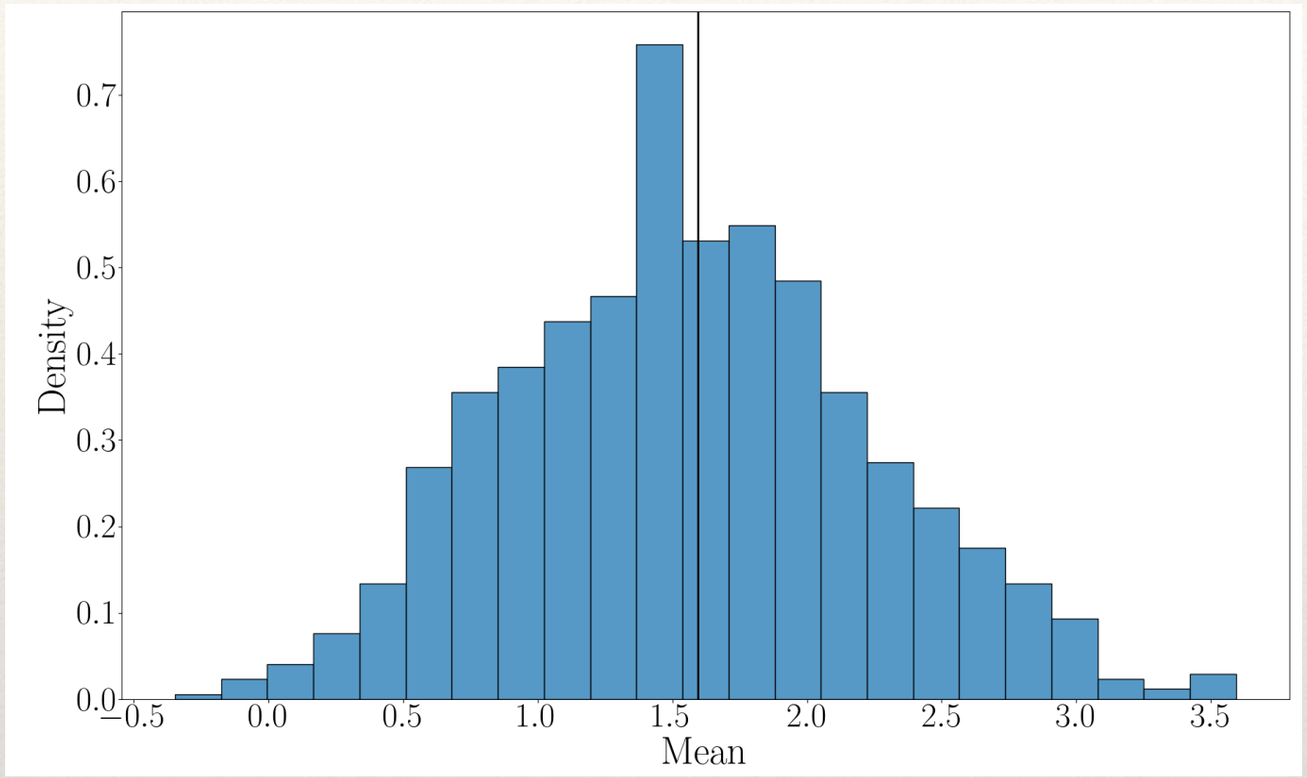
- This is the distribution of new datasets based on the model fitted to the data. The observed data should lie within the body of this distribution if the model is good.

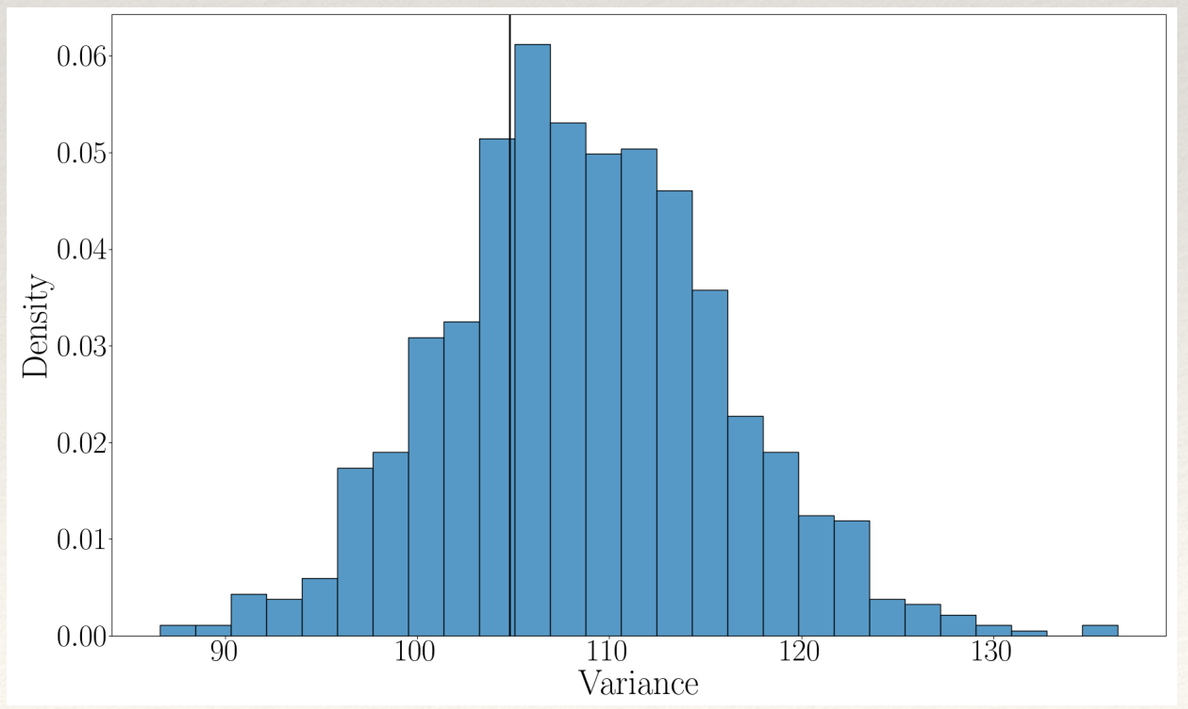
Example: gaussian fit

- The predictive distribution can be used to compute the distribution of summary quantities. The value of those summary quantities in the observed data can then be compared to these distributions.
- It is better to choose quantities that are somewhat “orthogonal” to what is adjusted to fit the data.
- Example: we try to fit a Gaussian to the following distribution:
- We assume a Gaussian measurement error of 2

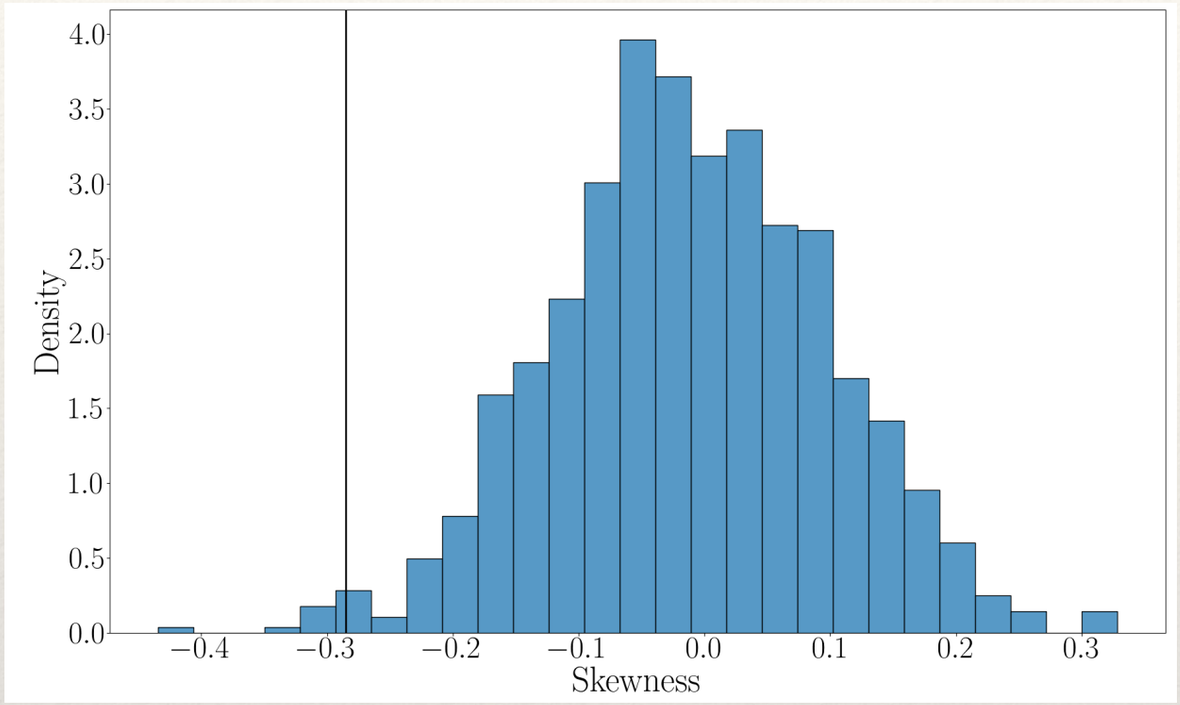


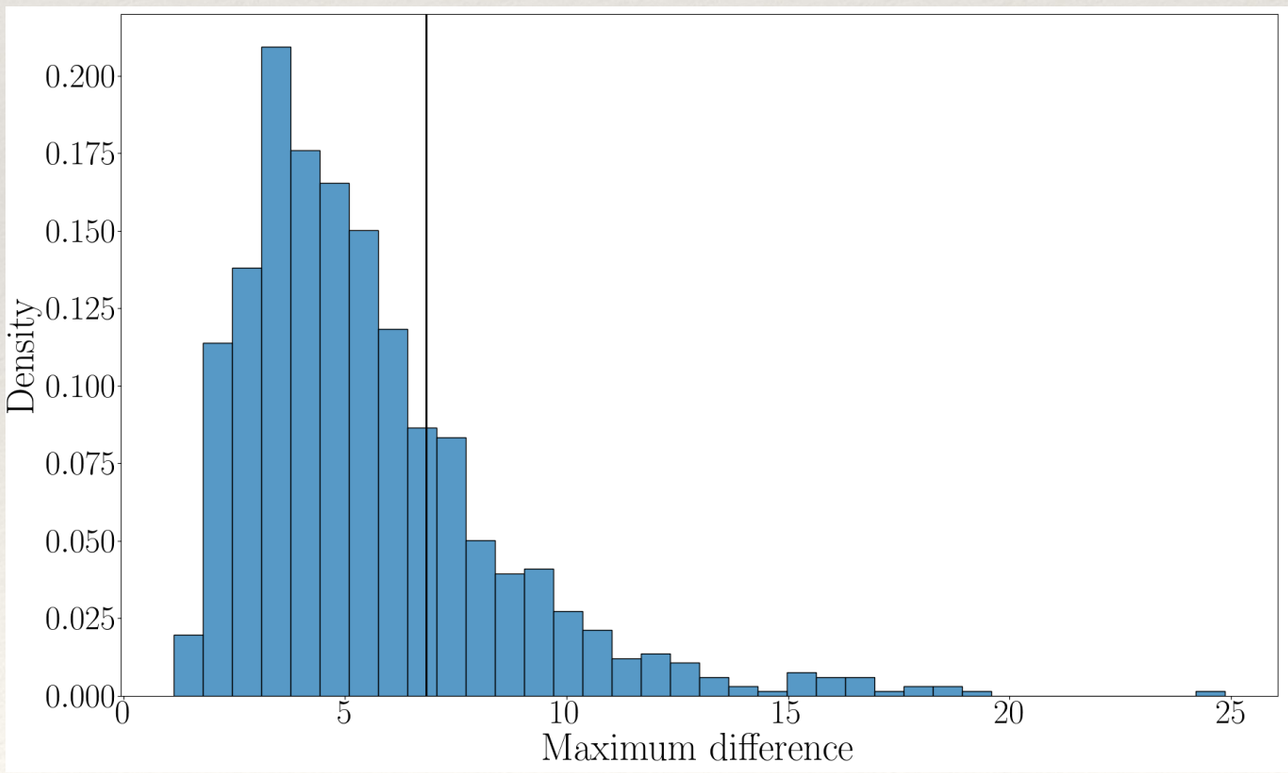
Example: gaussian fit





Example: gaussian fit





Predictive checking

- Posterior predictive checks are “good practice”.
- Can help build intuition how to improve models.
- But are often computationally expensive...