Field-level inference lecture 3: Bayesian hierarchical models

Cosmology Beyond the Analytic Lamppost course (2025)

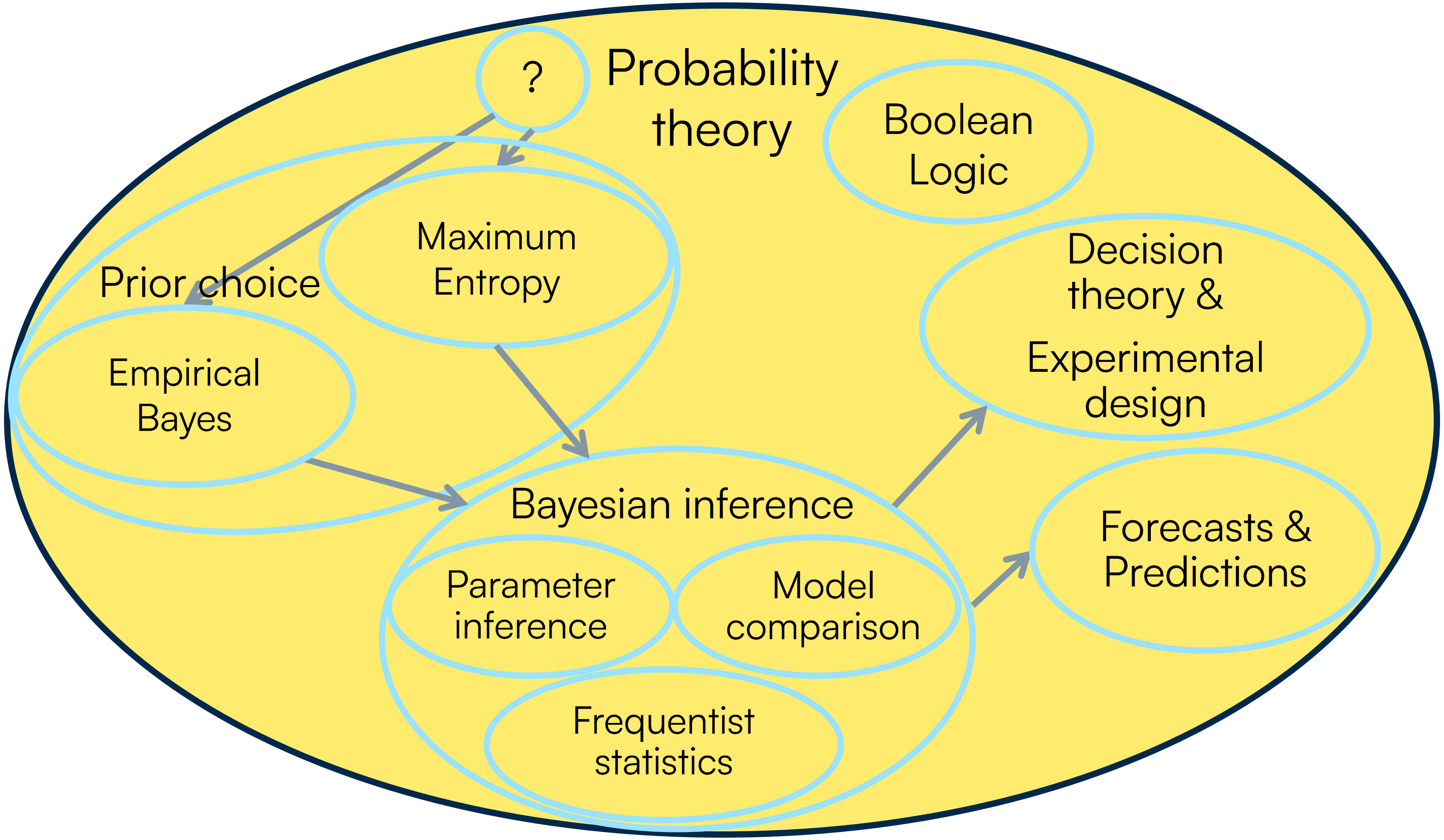
Florent Leclercq

www.florent-leclercq.eu

Institut d'Astrophysique de Paris
CNRS & Sorbonne Université

17 June 2025

Jaynes's "probability theory": an extension of ordinary Boolean logic



03

BAYESIAN DECISION THEORY AND EXPERIMENTAL DESIGN

Bayesian decision theory

- Bayesian decision theory is a framework for optimal decision-making, given a set of possible actions and a state of uncertain knowledge, represented by a pdf $p(\theta|I)$ (usually the posterior from a Bayesian inference prior to decision-making).

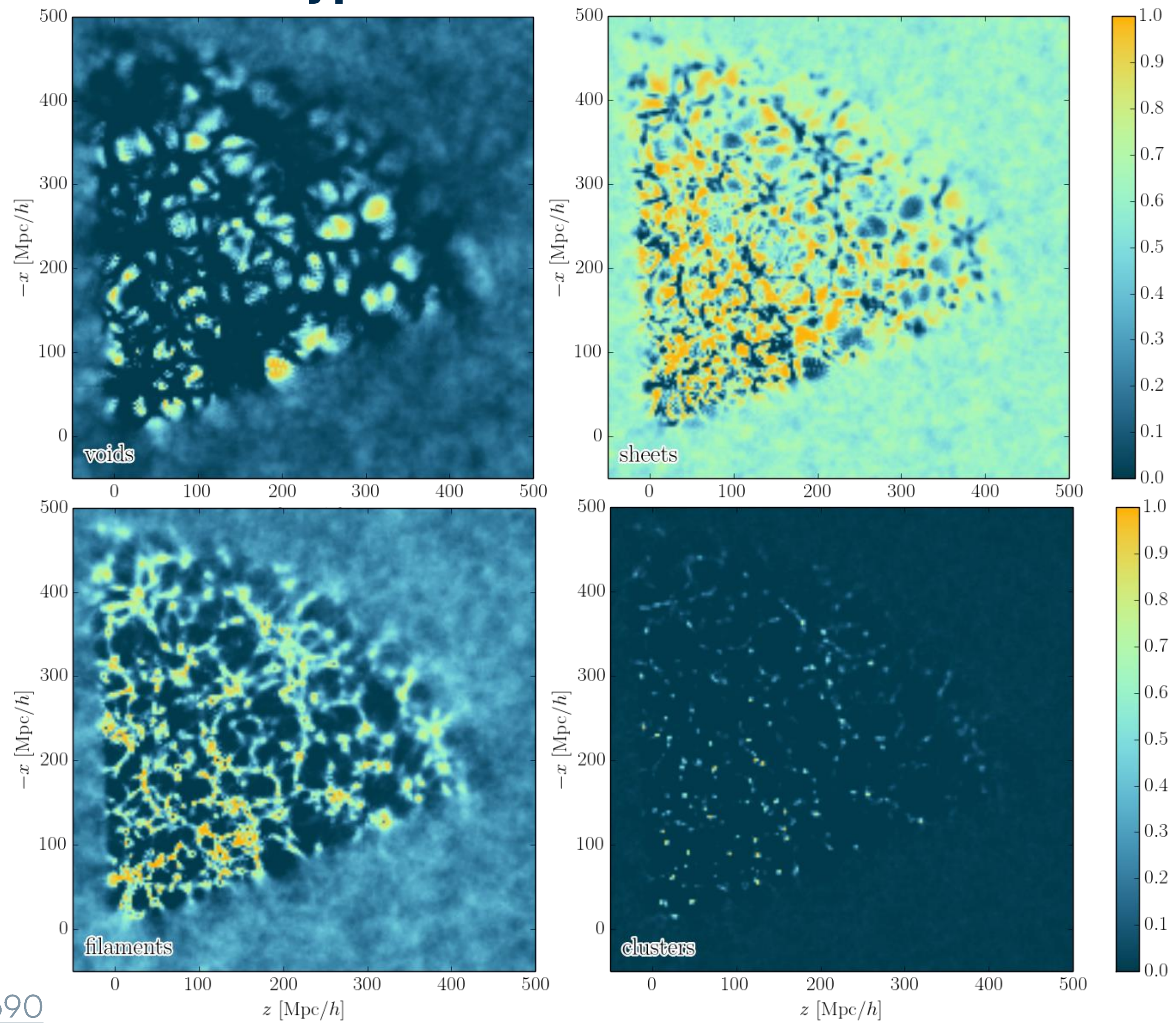
- Notations:
  - $\{\theta\}$ = set of parameters (observed variables)
  - $\{a\}$ = set of possible actions

- Expected utility hypothesis: Given a set of gain functions $G(a|\theta)$, the optimal decision rule consists of performing the action that maximises the expected utility $U(a|I)$, defined by

$$U(a|I) \equiv < G(a|\theta) >_{p(\theta|I)} = \int G(a|\theta) p(\theta|I) \, \mathrm{d}\theta$$

- Thus, one should perform the action $a^* = \mathrm{argmax}_a \, U(a|I)$.

Classification of cosmic web-types



FL, Jasche & Wandelt, 1502.02690

A decision rule for structure classification

- Space of “input features”:

$$\{\mathrm{T}_0 = \text{void}, \mathrm{T}_1 = \text{sheet}, \mathrm{T}_2 = \text{filament}, \mathrm{T}_3 = \text{cluster}\}$$

- Space of “actions”:

$$\{a_0 = \text{“decide void”}, a_1 = \text{“decide sheet”}, a_2 = \text{“decide filament”},\\ a_3 = \text{“decide cluster”}, a_{-1} = \text{“do not decide”}\}$$

- It is thus a problem of Bayesian decision theory: one should take the action that maximises the utility

$$U(a_j(\vec{x}_k)|d) = \sum_{i=0}^{3} G(a_j|\mathrm{T}_i)\,\mathcal{P}(\mathrm{T}_i(\vec{x}_k)|d)$$

- How to write down the gain functions?

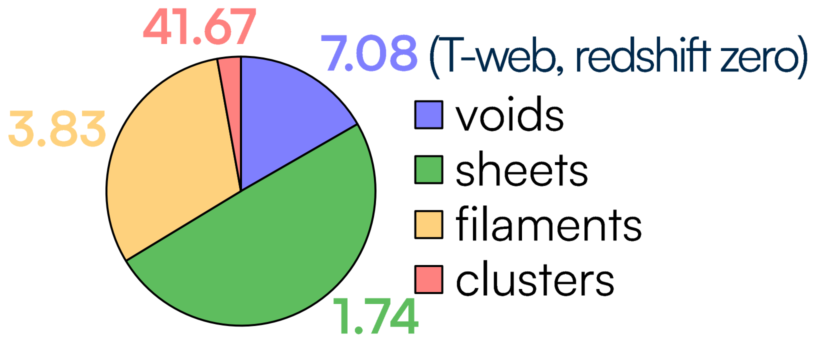
FL, Jasche & Wandelt, 1503.00730

Gambling with the Universe

- One proposal:

$$G(a_j|\mathrm{T}_i) = \begin{cases} \dfrac{1}{\mathcal{P}(\mathrm{T}_i)} - \alpha & \text{if } j \in [\![0,3]\!] \text{ and } i = j \\ -\alpha & \text{if } j \in [\![0,3]\!] \text{ and } i \neq j \\ 0 & \text{if } j = -1. \end{cases}$$

"Winning"
"Losing"
"Not playing"



- Without data, the expected utility is

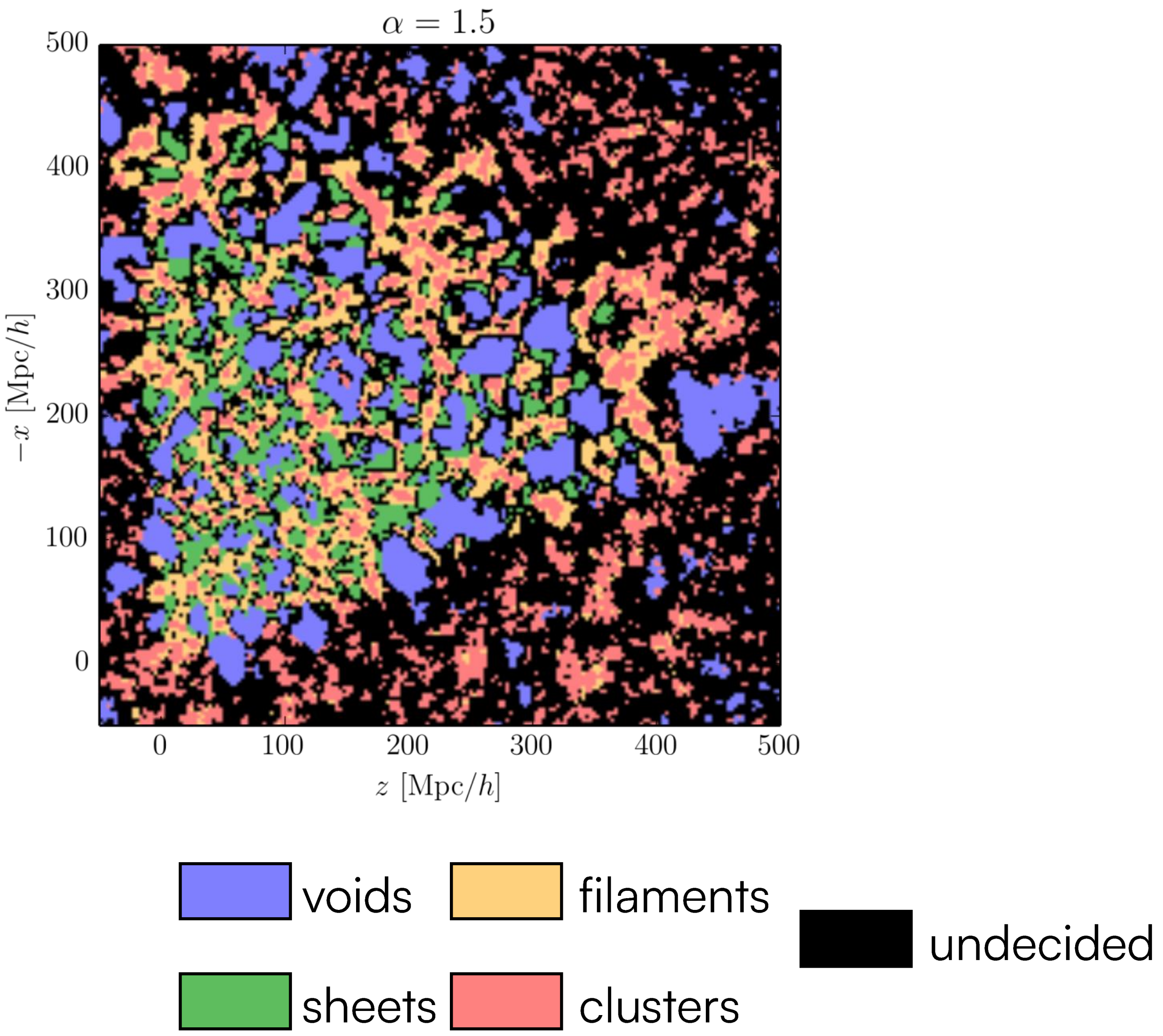
$$U(a_j) = 1 - \alpha \quad \text{if} \quad j \neq -1$$ "Playing the game"

$$U(a_{-1}) = 0$$ "Not playing the game"

- With $\alpha = 1$, it's a fair game ⇒ always play
  ⇒ "speculative map" of the LSS
- Values $\alpha > 1$ represent an aversion for risk
  ⇒ increasingly "conservative maps" of the LSS

FL, Jasche & Wandelt, 1503.00730

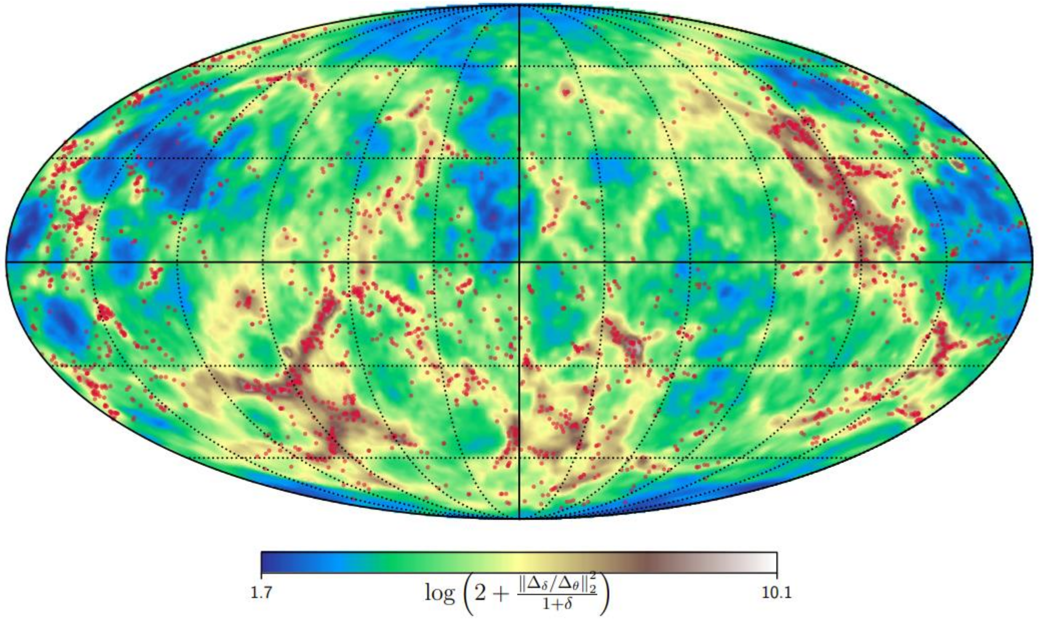
Playing the game...



FL, Jasche & Wandelt, 1503.00730

Bayesian experimental design: Information-optimal or entropy-maximal acquisition of future cosmological data

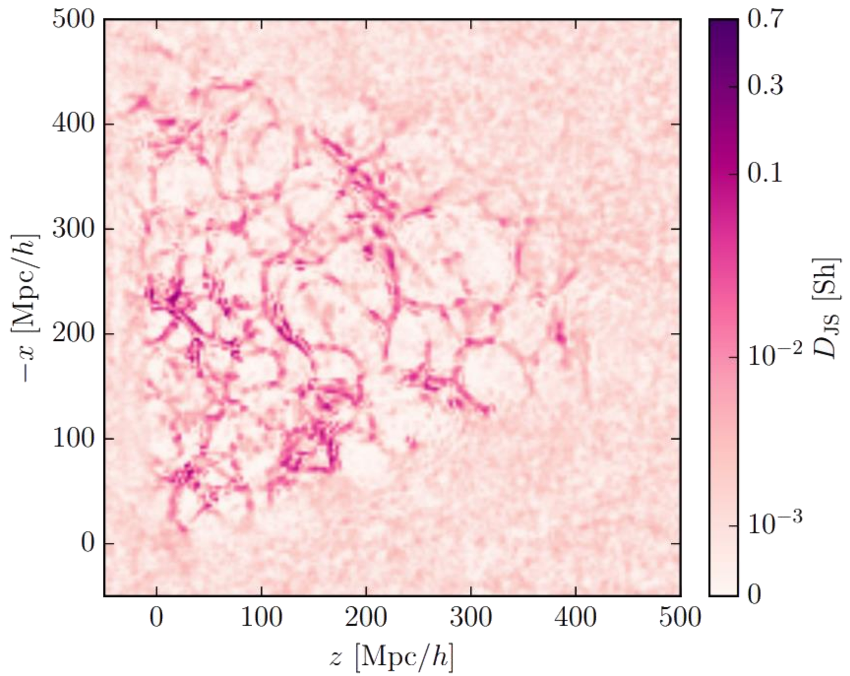
- This is where to look if we want to measure cosmological parameters of ΛCDM…



(Fisher information map for perturbative changes in the cosmological model)

Kostić, Jasche, Kodi Ramanah & Lavaux, 2107.00657

- And this is where to look if we want to learn about dark energy…



(Jensen-Shannon divergence between cosmic web-type posteriors for different values of the dark energy equation of state)

FL, Lavaux, Jasche & Wandelt, 1606.06758

03

BAYESIAN HIERARCHICAL MODELS

BAYESIAN HIERARCHY: AN EXAMPLE

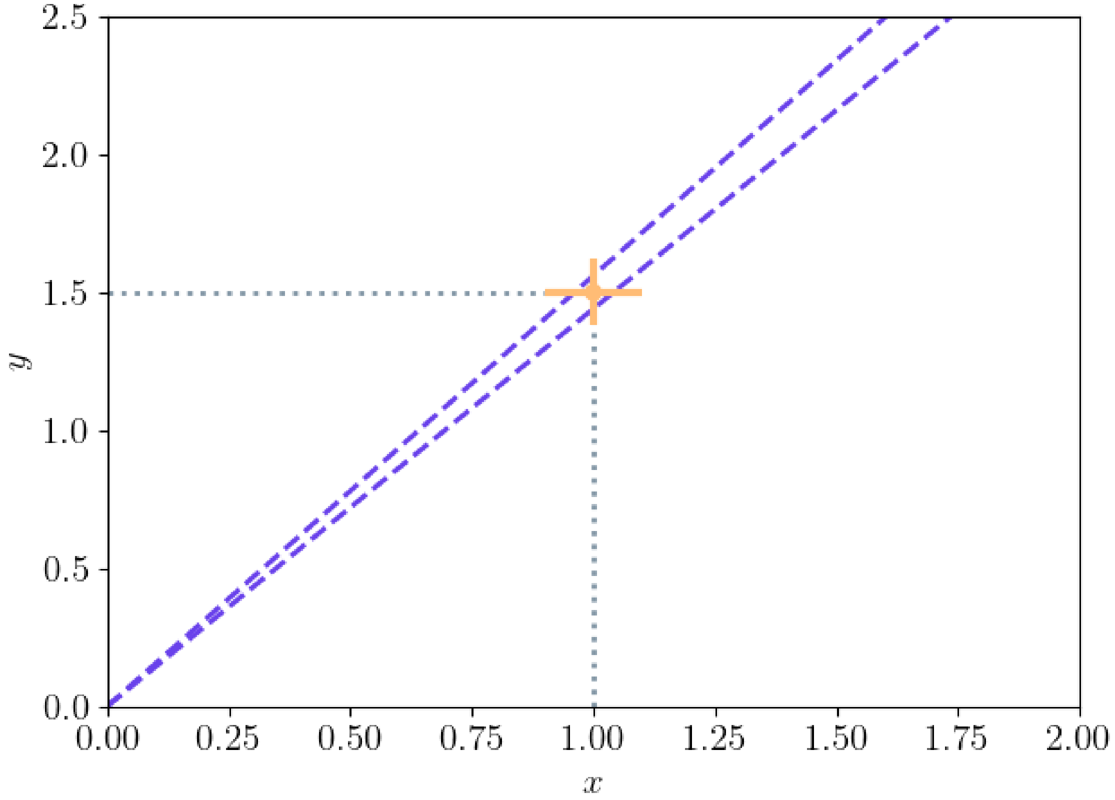
Bayesian hierarchy from latent variables

- Model: $y = mx$
- We measure X, Y, but they both have measurement errors. What is the posterior for the slope m?
- Applying the first rule ("write down what you want to know"): we want to know $p(m|X,Y)$
- There are two unknown ("latent") variables in the problem: the true values x, y
- Full joint pdf of the problem:

$$p(m, x, y, X, Y)$$

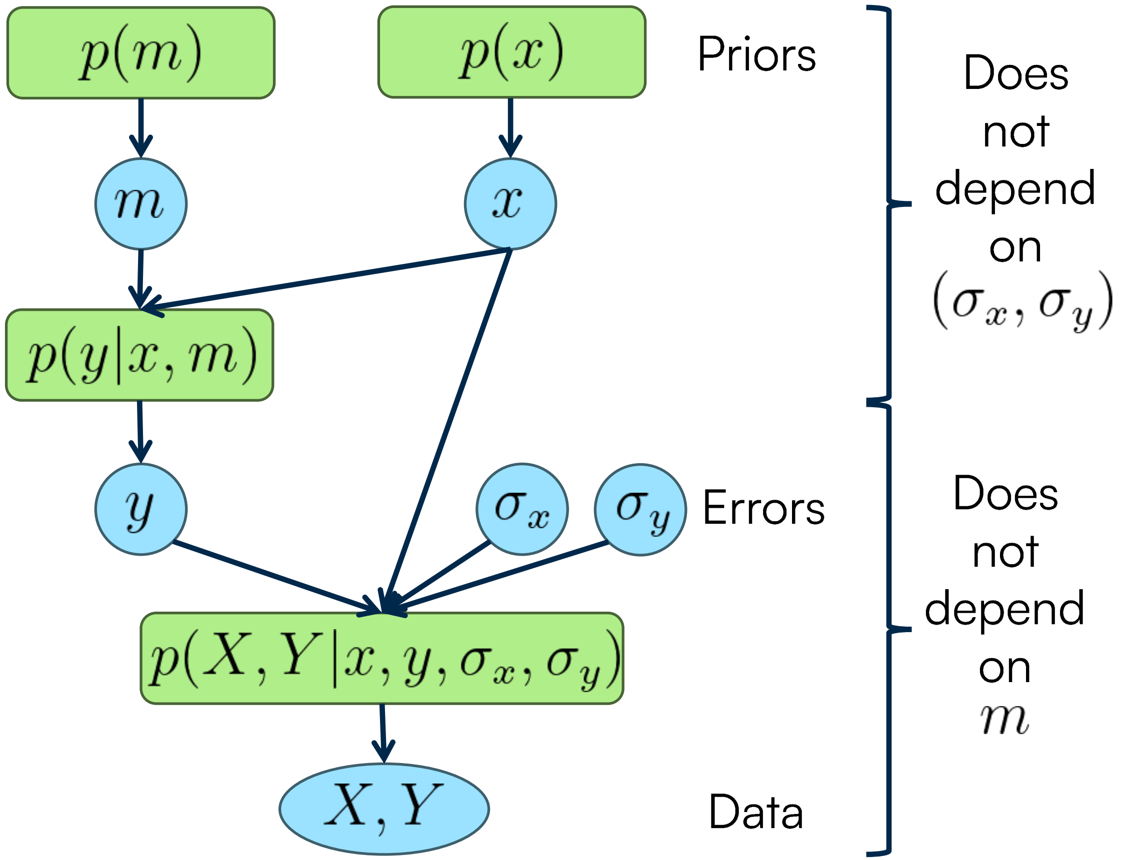
- Joint pdf of the target and observed variables:

$$p(m, X, Y) = \int p(m, x, y, X, Y)\, \mathrm{d}x\, \mathrm{d}y$$



Building the statistical model

- We construct a forward (generative) model of the data graphically:



- Apply Bayes' theorem:

$$p(m|X,Y) \propto p(X,Y|m)p(m)$$

- Introduce the latent variables and marginalise:

$$p(m|X,Y) \propto \iint p(X,Y,x,y|m)p(m)\,\mathrm{d}x\,\mathrm{d}y$$

- Expand first probability with the product rule:

$$p(m|X,Y) \propto \iint p(X,Y|x,y,m)p(x,y|m)p(m)\,\mathrm{d}x\,\mathrm{d}y$$

- Expand second probability with the product rule:

$$p(m|X,Y) \propto \iint p(X,Y|x,y,m)p(y|x,m)p(x|m)p(m)\,\mathrm{d}x\,\mathrm{d}y$$

- Simplify conditional dependencies:

$$p(X,Y|x,y,m) = p(X,Y|x,y)$$
$$p(x|m) = p(x)$$

- Apply physical relation: $p(y|x,m) = \delta_{\mathrm{D}}(y - mx)$
- Integrate to get the final result:

$$p(m|X,Y) \propto \int p(X,Y|x,mx)p(x)p(m)\,\mathrm{d}x$$

Inferring the slope

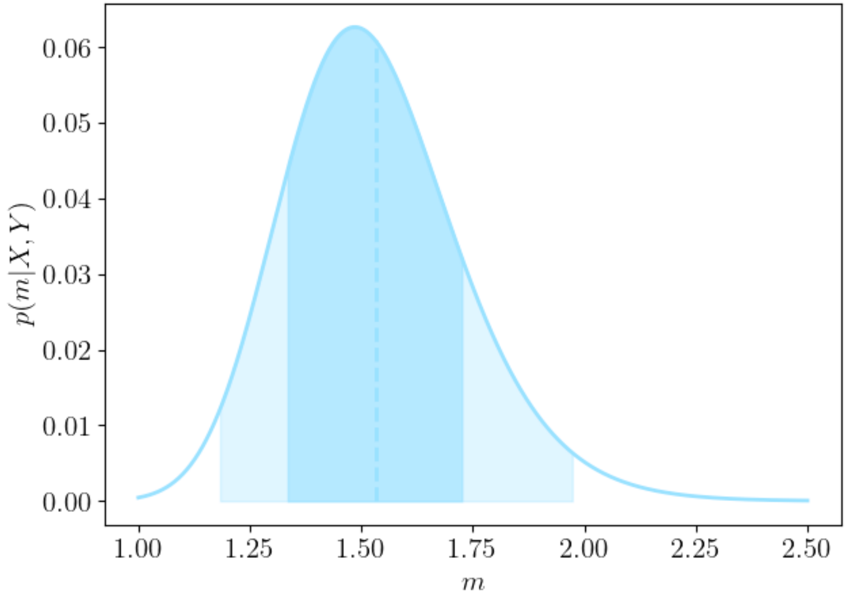
$$p(m|X,Y) \propto \int p(X,Y|x,mx)p(x)p(m)\,\mathrm{d}x$$

- If the error distribution is Gaussian with zero mean, and if we take uniform priors on x and m:

$$p(m|X,Y) \propto \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{1}{2}\frac{(X-x)^2}{\sigma_x^2}}\,\mathrm{e}^{-\frac{1}{2}\frac{(Y-mx)^2}{\sigma_y^2}}\,\mathrm{d}x$$

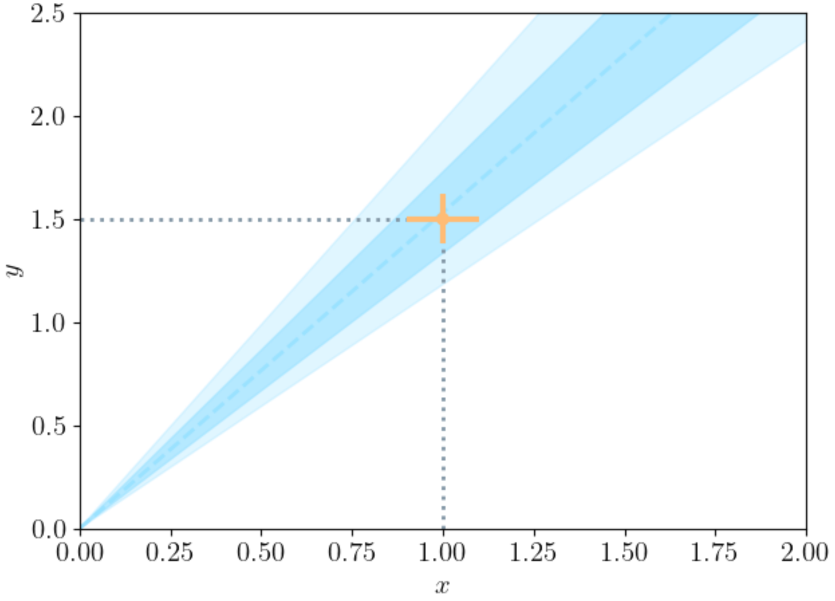
- Completing the square and integrating gives the marginal posterior for m:

$$p(m|X,Y) \propto \frac{\sigma_x\sigma_y}{\sqrt{\sigma_y^2+m^2\sigma_x^2}}\exp\left[-\frac{1}{2}\frac{(Y-mX)^2}{\sigma_y^2+m^2\sigma_x^2}\right]$$



$X = 1.0, Y = 1.5$

$\sigma_x = 0.10, \sigma_y = 0.12$



Inferring the full model and sampling

- The joint posterior for (x, m) is:

$$p(x, m|X, Y) \propto p(X, Y|x, mx)p(x)p(m)$$
$$\propto \mathrm{e}^{-\frac{1}{2}\frac{(X-x)^2}{\sigma_x^2}} \mathrm{e}^{-\frac{1}{2}\frac{(Y-mx)^2}{\sigma_y^2}}$$

- At fixed x:

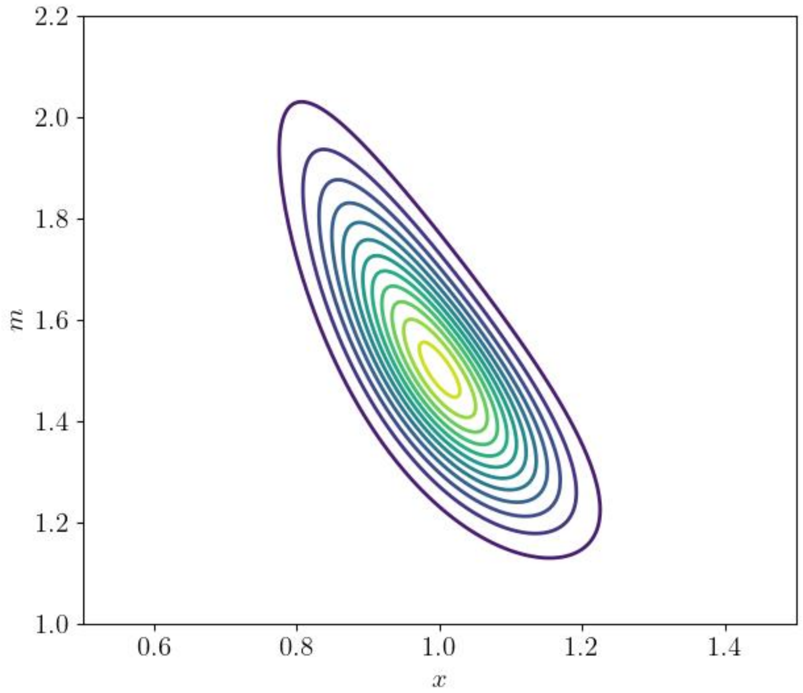
$$p(m|X, Y, x) \propto \exp\left[-\frac{1}{2}\frac{x^2(m-\frac{Y}{x})^2}{\sigma_y^2}\right]$$

i.e. $p(m|X, Y, x) = \mathcal{G}\left(\frac{Y}{x}, \frac{\sigma_y^2}{x^2}\right)$

- At fixed m (combining the exponents and completing the square):

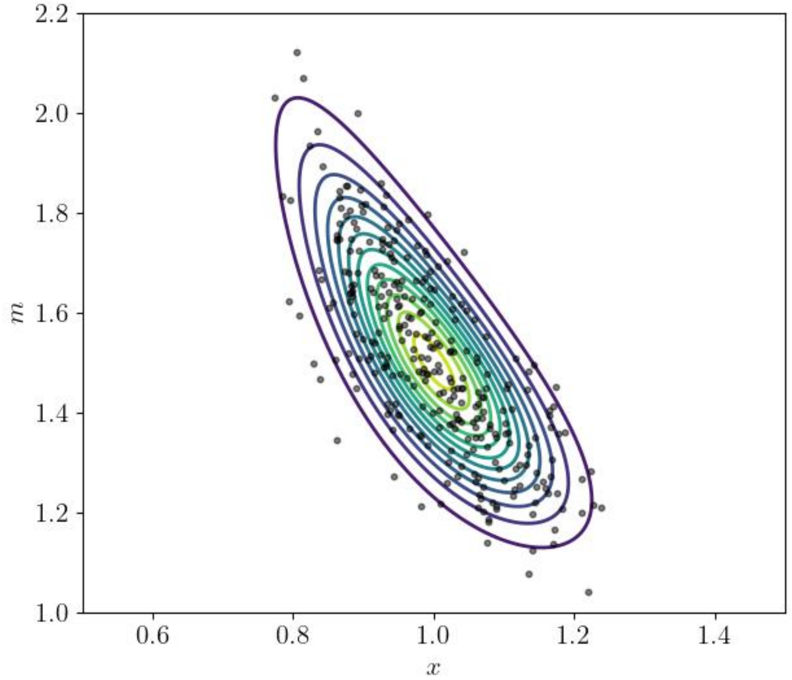
$$p(x|X, Y, m) = \mathcal{G}\left(\frac{\sigma_y^2 X + m\sigma_x^2 Y}{\sigma_y^2 + m^2\sigma_x^2}, \frac{\sigma_y^2\sigma_x^2}{\sigma_y^2 + m^2\sigma_x^2}\right)$$

- We can therefore use Gibbs sampling to draw samples from the joint posterior:
  - $m \curvearrowleft p(m|X, Y, x)$
  - $x \curvearrowleft p(x|X, Y, m)$



$X = 1.0, Y = 1.5$

$\sigma_x = 0.10, \sigma_y = 0.12$



Bayesian hierarchical models and generalised linear regression

- At the heart of the method lies the fundamental problem of (generalised) linear regression, in the presence of measurement errors on both the dependent and the independent variable and intrinsic scatter in the relationship.
- This is a general problem in any field dealing with objects with an intrinsic variability.

- Model to be fitted:

$$y = mx + b$$

- Statistical model:

$x_i \sim p(x\|R_x) = \mathcal{G}(\mu_x, R_x)$	Population distribution
$y_i\|x_i \sim \mathcal{G}(mx_i + b, R_y)$	Intrinsic variability
$X_i, Y_i\|x_i, y_i \sim \mathcal{G}([x_i, y_i], C)$	Measurement error

usually $C = \begin{pmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{pmatrix}$

Gull (1989); Jaynes (1989); Gellman & Hill (2006); Kelly, 0705.2774

BAYESIAN HIERARCHICAL MODELS

Bayesian hierarchical models for adapting the prior

- Simple Bayesian inference:

$$p(\theta|d) \propto p(d|\theta)\overset{\text{prior}}{p(\theta)}$$

- Inference with an adaptive prior depending on a latent variable:

$$p(\theta|d) \propto p(d|\theta)\overset{\text{prior}}{p(\theta|\eta)}\overset{\text{hyperprior}}{p(\eta)}$$

- ... or a full hierarchy of hyperpriors.
  Examples:
  - Cosmic microwave background:

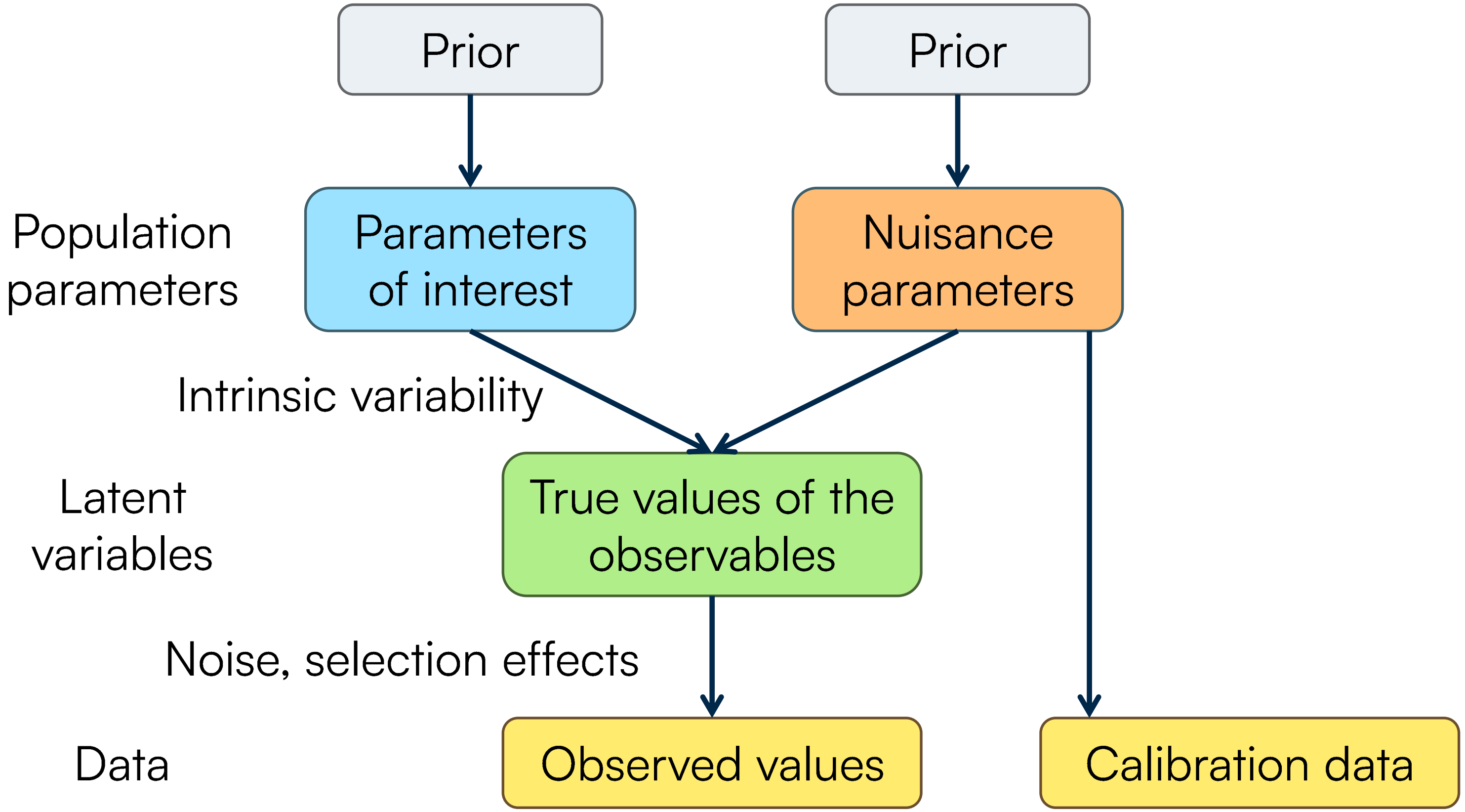
$$p(\{\Omega\}, \{C_\ell\}, s|d) \propto p(d|s)\, p(s|\{C_\ell\})\, p(\{C_\ell\}|\{\Omega\})\, p(\{\Omega\})$$

  - Large-scale structure:

$$p(\{\Omega\}, \phi, g|d) \propto p(d|g)\, p(g|\phi)\, p(\phi|\{\Omega\})\, p(\{\Omega\})$$

Many sources of variability

- You pick a lightbulb, and measure its brightness. What is it?
- There are many reasons why the value might vary:
  - It's picked from a box of bulbs of different brightnesses
  - The manufacturing process is imprecise
  - Measurement error
- *Any or all* of these may apply (and you may not know which).

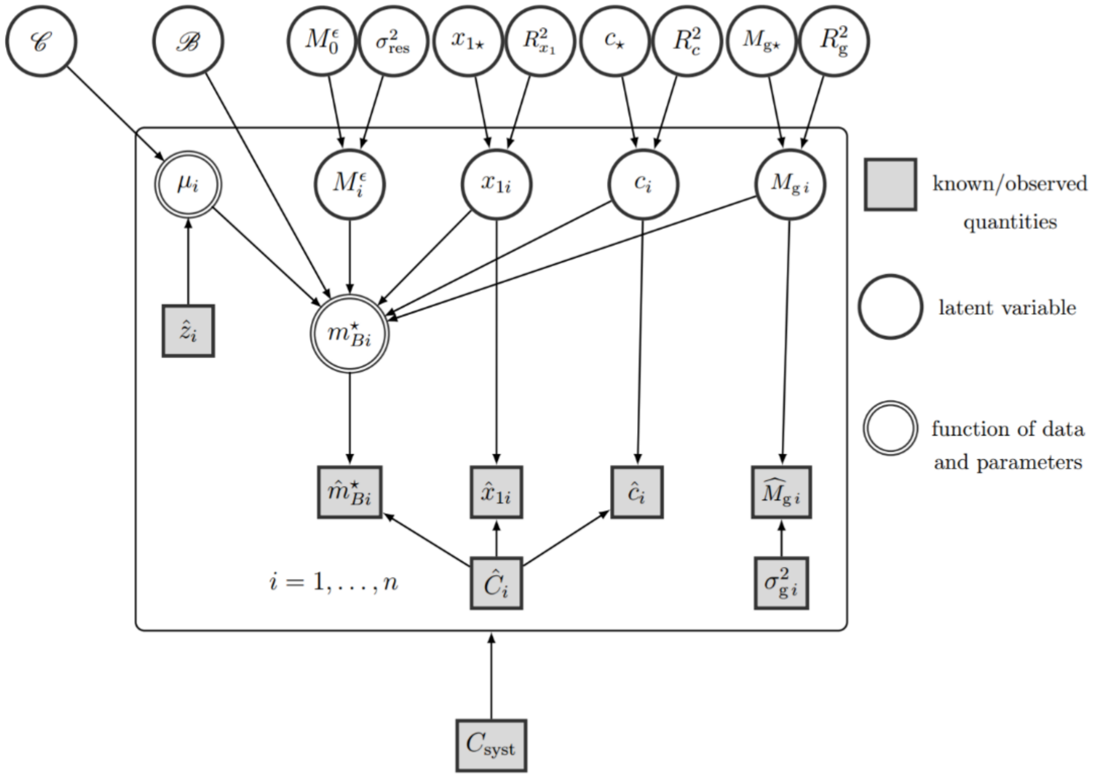


Example courtesy of Alan Heavens

Bayesian hierarchical models for complex problems

- How can we make sure all the errors are propagated correctly to the posterior?
- We split the inference problem into steps, where the full model is made up of a series of sub-models. The aim is to build a complete model of the data. It is a principled way to include systematic errors, selection effects, etc. (everything, really).
- The Bayesian Hierarchical Model (BHM) links the sub-models together, correctly propagating uncertainties in each sub-model from one level to the next.
- It also exposes what you need to know or assume. At each step you will (ideally) know the conditional distributions.
- All of the steps give rise to "latent variables": parameters in sub-models, usually not of interest.
  - A particular sort of "nuisance parameter"
  - They still need to be accounted for (and marginalised over)
    - e.g., contribution of systematic error to a measurement
    - e.g., galactic dust flux in a noisy CMB pixel
  - These might very well be "signal" for a different purpose.
- Therefore, BHMs may have very many parameters.
- When you are using sampling for inference, marginalisation is "trivial": just ignore those variables in the output.
  - Realistically, of course, there is usually some information there!

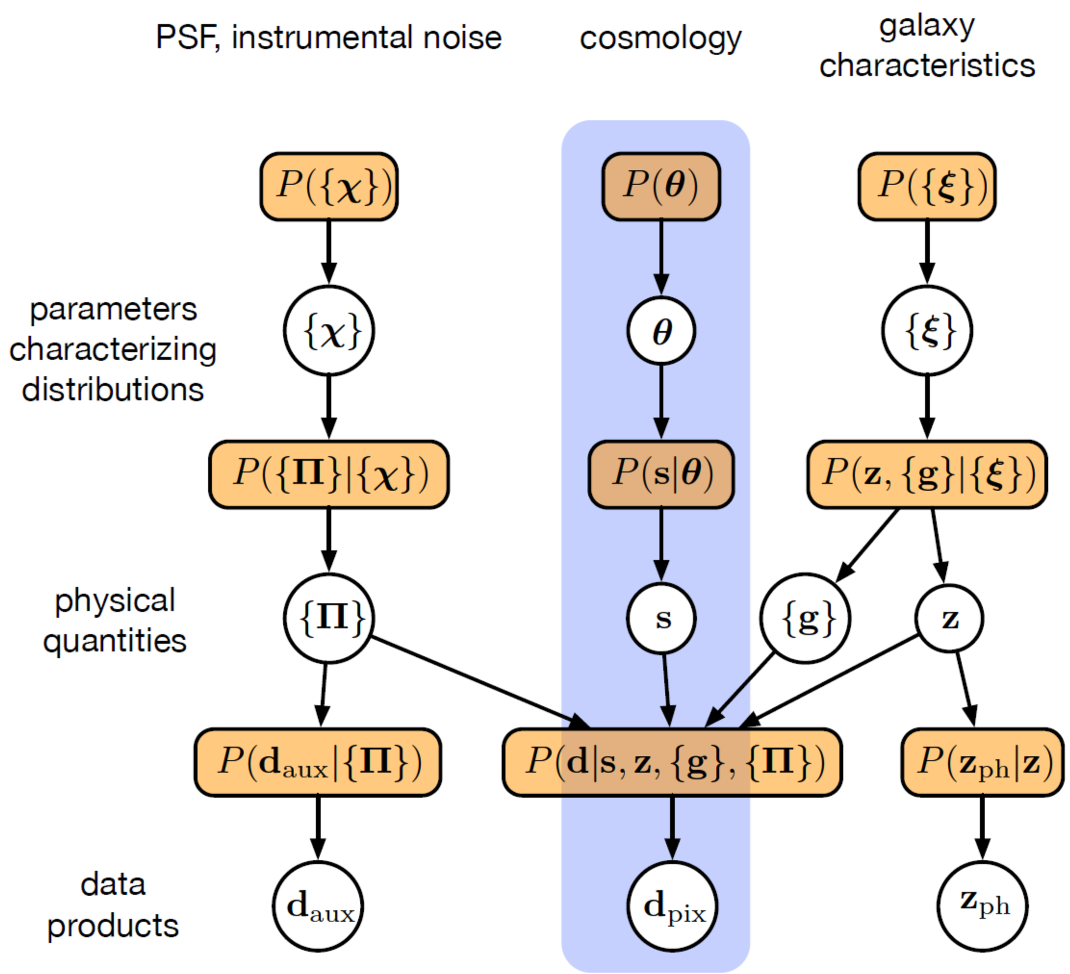
BHM example: supernova cosmology (BAHAMAS)



Parameter	Notation and Prior Distribution
Cosmological parameters	
Matter density parameter	$\Omega_{\mathrm{m}} \sim \textsc{Uniform}(0,2)$
Cosmological constant density parameter	$\Omega_\Lambda \sim \textsc{Uniform}(0,2)$
Dark energy EOS	$w \sim \textsc{Uniform}(-2,0)$
Hubble parameter	$H_0/\mathrm{km/s/Mpc} = 67.3$
Covariates	
Coefficient of stretch covariate	$\alpha \sim \textsc{Uniform}(0,1)$
Coefficient of color covariate	β (or β_0) $\sim \textsc{Uniform}(0,4)$
Coefficient of interaction of color correction and z	$\beta_1 \sim \textsc{Uniform}(-4,4)$
Jump in coefficient of color covariate	$\Delta\beta \sim \textsc{Uniform}(-1.5,1.5)$
Redshift of jump in color covariate	$z_t \sim \textsc{Uniform}(0.2,1)$
Coefficient of host galaxy mass covariate	$\gamma \sim \textsc{Uniform}(-4,4)$
Population-level distributions	
Mean of absolute magnitude	$M_0^\epsilon \sim \mathcal{N}(-19.3, 2^2)$
Residual scatter after corrections	$\sigma^2_{\mathrm{res}} \sim \textsc{InvGamma}(0.003, 0.003)$
Mean of absolute magnitude, low galaxy mass	$M_0^{\mathrm{lo}} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, low galaxy mass	${\sigma^{\mathrm{lo}}_{\mathrm{res}}}^2 \sim \textsc{InvGamma}(0.003, 0.003)$
Mean of absolute magnitude, high galaxy mass	$M_0^{\mathrm{hi}} \sim \mathcal{N}(-19.3, 2^2)$
SD of absolute magnitude, high galaxy mass	${\sigma^{\mathrm{hi}}_{\mathrm{res}}}^2 \sim \textsc{InvGamma}(0.003, 0.003)$
Mean of stretch	$x_{1\star} \sim \mathcal{N}(0, 10^2)$
SD of stretch	$R_{x_1} \sim \textsc{LogUniform}(-5,2)$
Mean of color	$c_\star \sim \mathcal{N}(0, 1^2)$
SD of color	$R_c \sim \textsc{LogUniform}(-5,2)$
Mean of host galaxy mass	$M_{\mathrm{g}\star} \sim \mathcal{N}(10, 100^2)$
SD of host galaxy mass	$R_{\mathrm{g}} \sim \textsc{LogUniform}(-5,2)$

Shariff *et al.*, 1510.05954

BHM example: weak lensing



Can include:

Mask
Intrinsic alignments
Baryon feedback
Shape measurement
Photometric redshifts

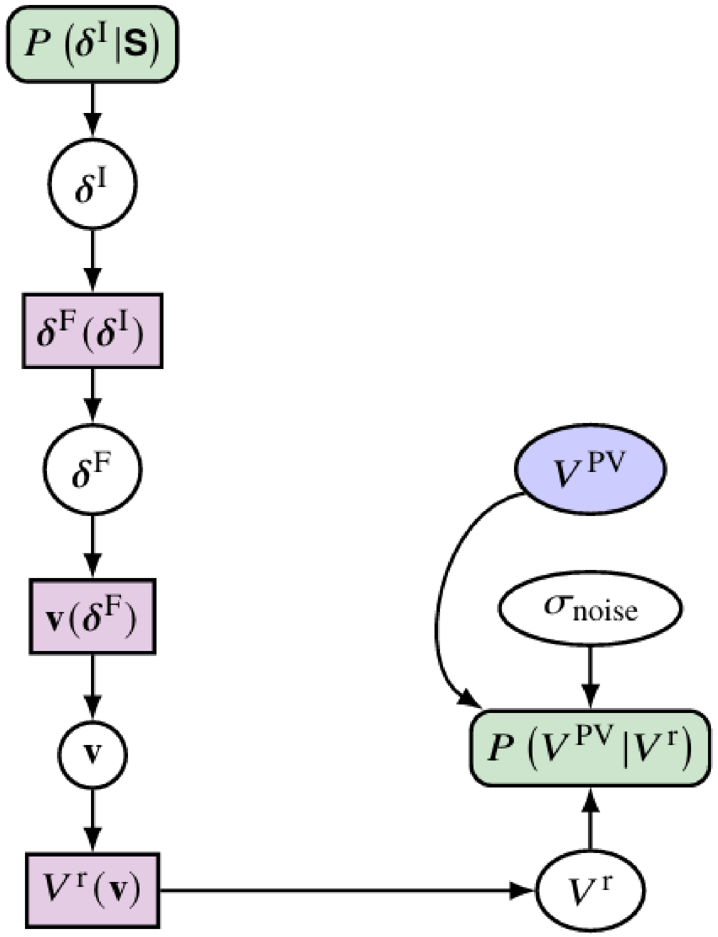
Alsing *et al.*, 1505.07840

BHM example: large-scale structure inference from peculiar velocity tracers

Initial density field

Final density field

Velocity field



Peculiar velocity tracers

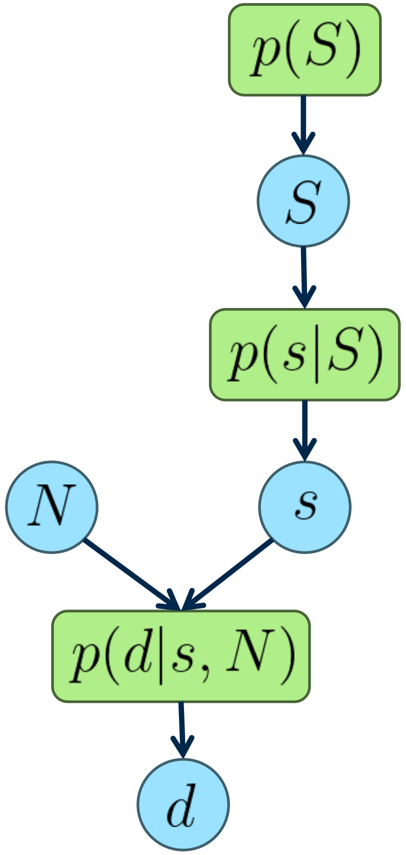
[Prideaux-Ghee, FL *et al.*, 2204.00023]

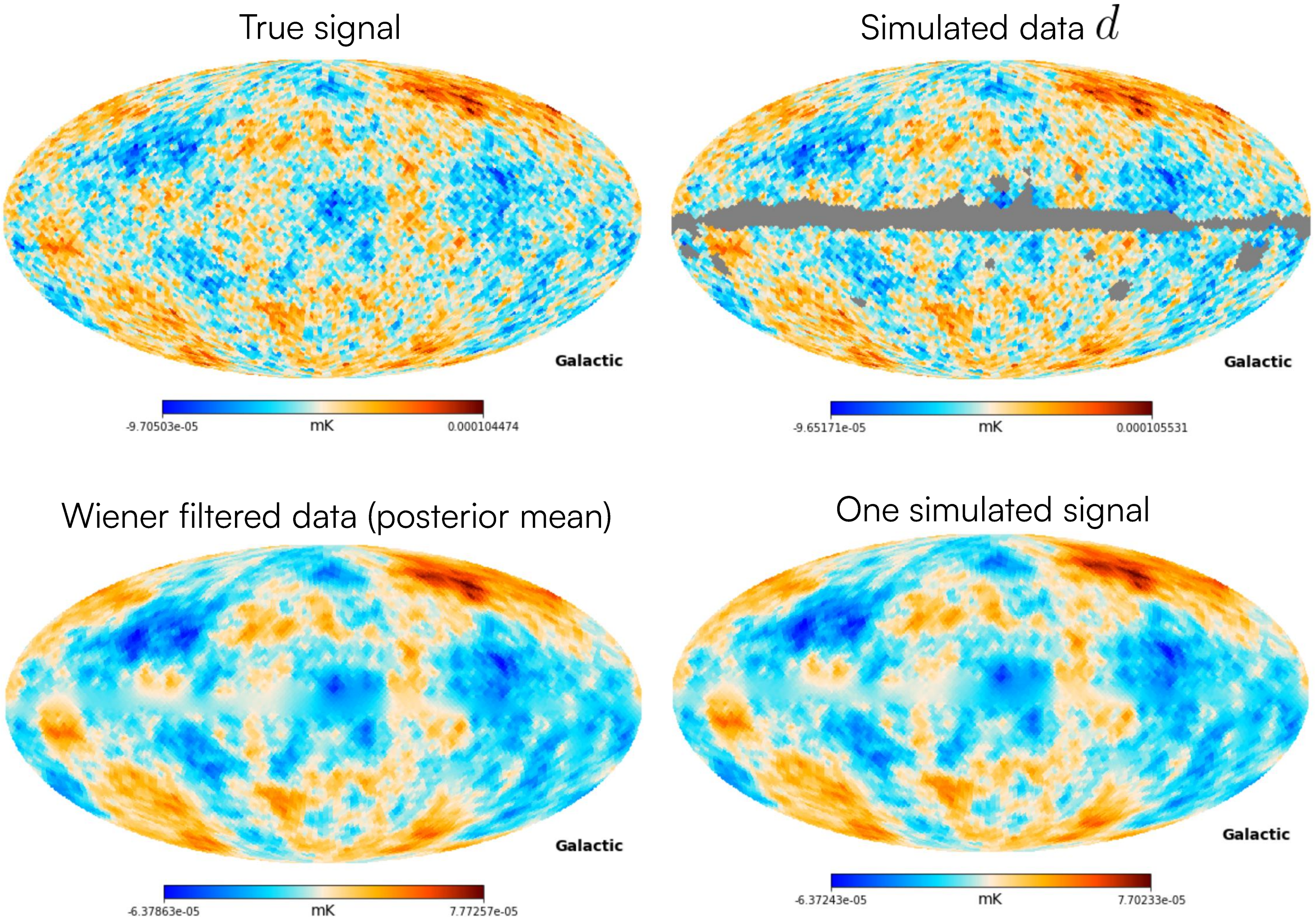
Back to Wiener filtering

$$\mu_{s|d} = \mu_s + S(S+N)^{-1}(d-\mu_d)$$

$$C_{s|d} = S - S(S+N)^{-1}S$$

- As a BHM:



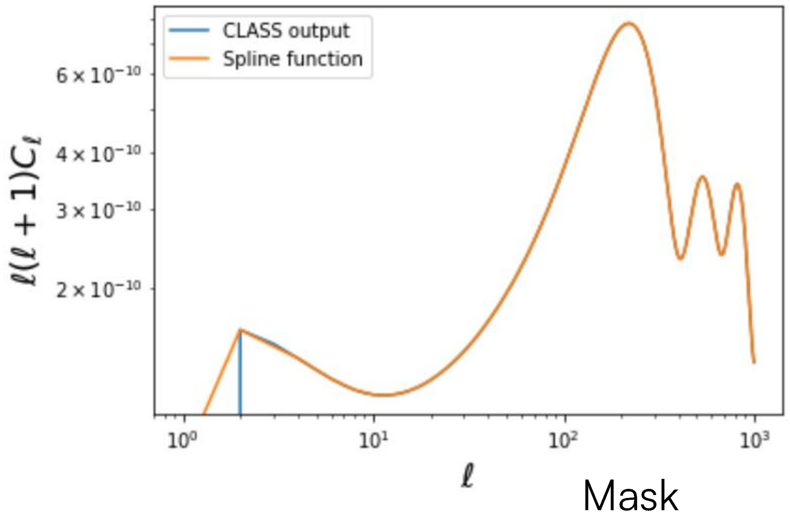


Back to Wiener filtering

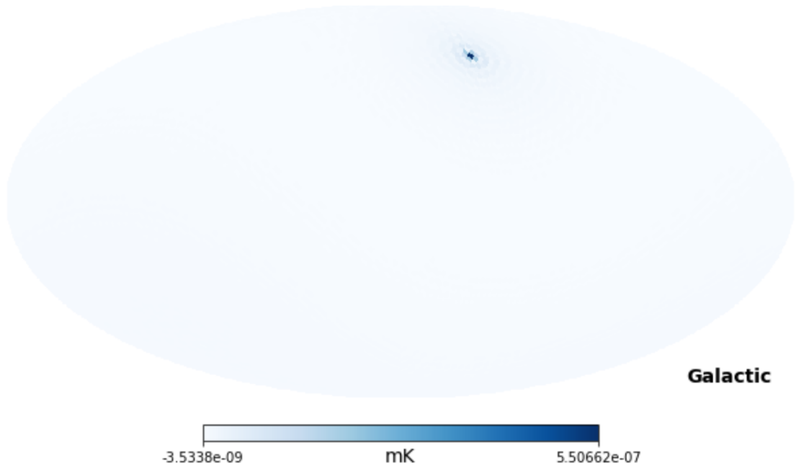
$$\mu_{s|d} = \mu_s + S(S+N)^{-1}(d-\mu_d)$$
$$C_{s|d} = S - S(S+N)^{-1}S$$

- Problem: computing/representing $(S+N)^{-1}$ is difficult because S is sparse in harmonic/ Fourier space and N is sparse in configuration/real space.

S

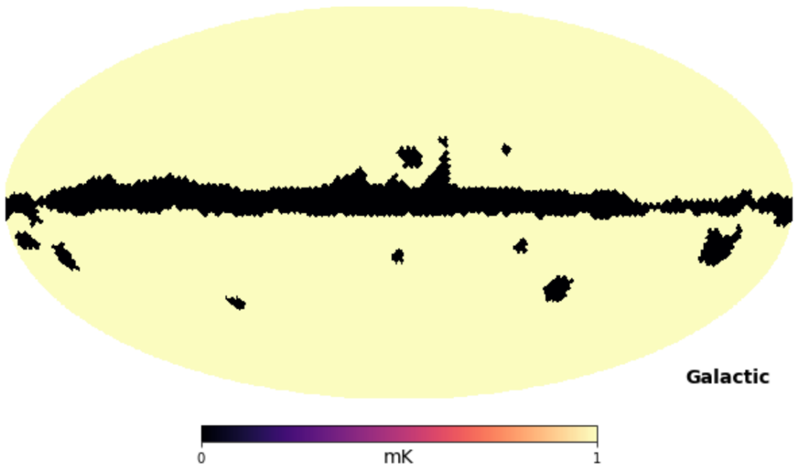


One row of $\sqrt{S}$



Mask

N



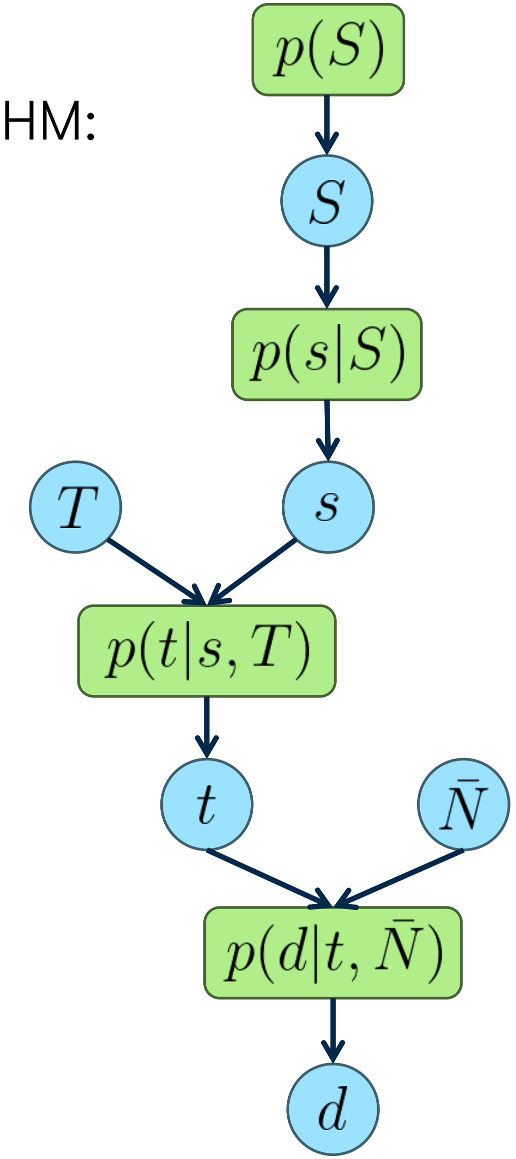
Messenger field and multivariate Wiener filtering

$$\mu_{s|d} = S(S+N)^{-1}d \quad \text{(assuming } \mu_s = \mu_d = 0\text{)}$$

$$C_{s|d} = S - S(S+N)^{-1}S$$

- Messenger field algorithm:
  - Introduce an auxiliary Gaussian random field t with covariance matrix $T \equiv \tau I$.
  - T (isotropic noise covariance matrix) is diagonal in any basis (harmonic/Fourier and configuration/real).
  - Introduce $\bar{N} \equiv N - T$ (residual noise covariance matrix).
- Sampling:
  - Goal: obtain samples of $p(s,t|d)$ via Gibbs sampling. We need the conditionals $p(s|d,t)$ and $p(t|s,d)$.
  - $p(s|d,t) = p(s|t)$ is Gaussian with
    mean: $\mu_{s|t} = (S^{-1} + T^{-1})^{-1} T^{-1} t$ (assuming $\mu_s = \mu_t = 0$)
    covariance: $C_{s|t} = (S^{-1} + T^{-1})^{-1}$
  - $p(t|s,d) \propto p(t|s)p(d|t)$ is Gaussian with
    mean: $\mu_{t|s,d} = (T^{-1} + \bar{N}^{-1})^{-1} T^{-1} s + (T^{-1} + \bar{N}^{-1})^{-1} \bar{N}^{-1} d$
    covariance: $C_{t|s,d} = (T^{-1} + \bar{N}^{-1})^{-1}$

- As a BHM:



Elsner & Wandelt, 1210.4931; Jasche & Lavaux, 1402.1763; Kodi Ramanah *et al.*, 1801.05358

Bayesian hierarchical models: summary

- BHMs are a way to build a statistical model of data by splitting the problem into steps.
- Decomposing into steps exposes what is needed — typically many conditional distributions.
- For complex experiments, this may be the only viable way to build the statistical model of the data.
- The decomposition is usually very natural and logical.
- The model allows the proper propagation of errors from one layer to the next, including a proper treatment of systematics.
- One can often use efficient sampling algorithms to sample from the posterior — precisely what one wants for a Bayesian statistical analysis.

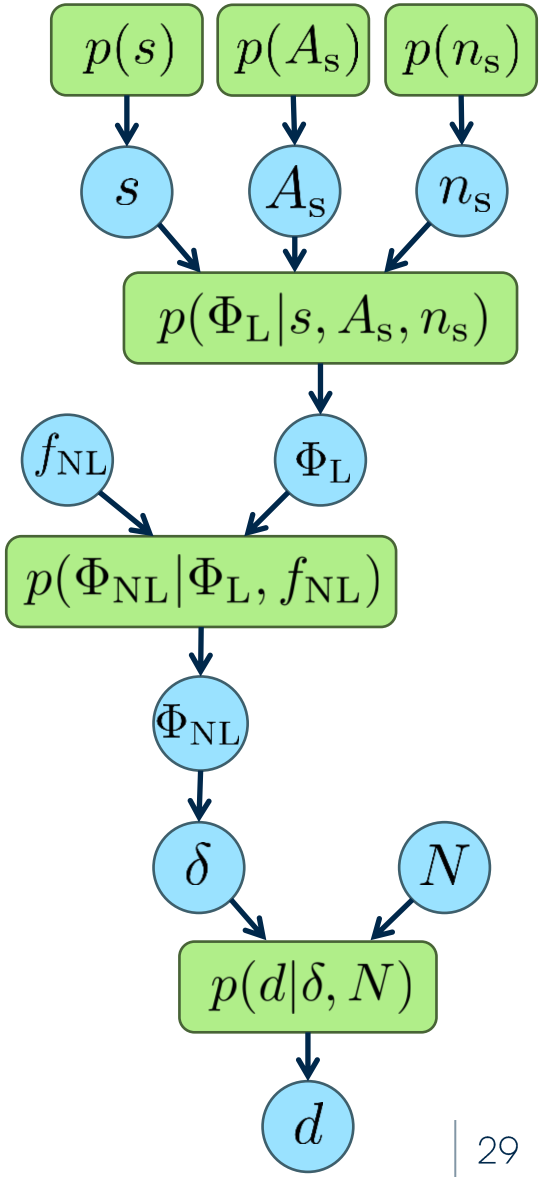
BAYESIAN HIERARCHICAL MODELS FOR JOINT FIELD-PARAMETER SAMPLING

A Bayesian hierarchical field-level model

Exercise: Joint field-parameter sampling

- Model:
  - The signal s is a white noise field.
  - The primordial gravitational potential Φ_L is a Gaussian random field with phases given by s, zero mean and power spectrum
    $$P(k) = A_\mathrm{s} k^{n_\mathrm{s}-1}$$
    (i.e. a diagonal covariance matrix in Fourier space), where A_s and n_s are cosmological parameters.
  - The non-linear gravitational potential Φ_NL follows
    $$\Phi_\mathrm{NL} = \Phi_\mathrm{L} + f_\mathrm{NL}\Phi_\mathrm{L}^2$$
    where f_NL is a parameter.
  - The density contrast δ follows (in Fourier space)
    $$\delta(k) = D_1\, T(k)\, \Phi_\mathrm{NL}(k)$$
    where D_1 is a fixed (arbitrary) coefficient and $T(k)$ is a “BBKS” transfer function.
  - The noise is Gaussian and additive, to give the observed data:
    $$d = \delta(s) + n$$

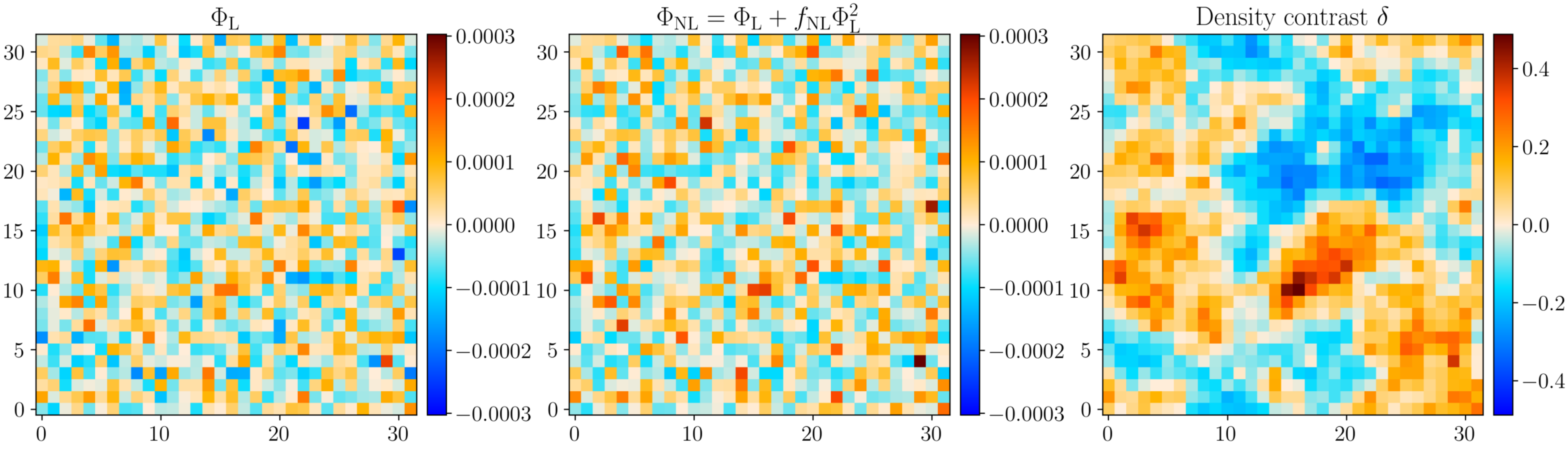
- It is the same model as in the previous lecture, except that $\{A_\mathrm{s}, n_\mathrm{s}, f_\mathrm{NL}\}$ are not fixed. We want to sample them jointly with the signal s.
- As natural byproducts, we will get samples of all of the latent fields of the problem: $\{\Phi_\mathrm{L}, \Phi_\mathrm{NL}, \delta\}$
- The full model is conveniently represented as a Bayesian hierarchical model.



A Bayesian hierarchical field-level model

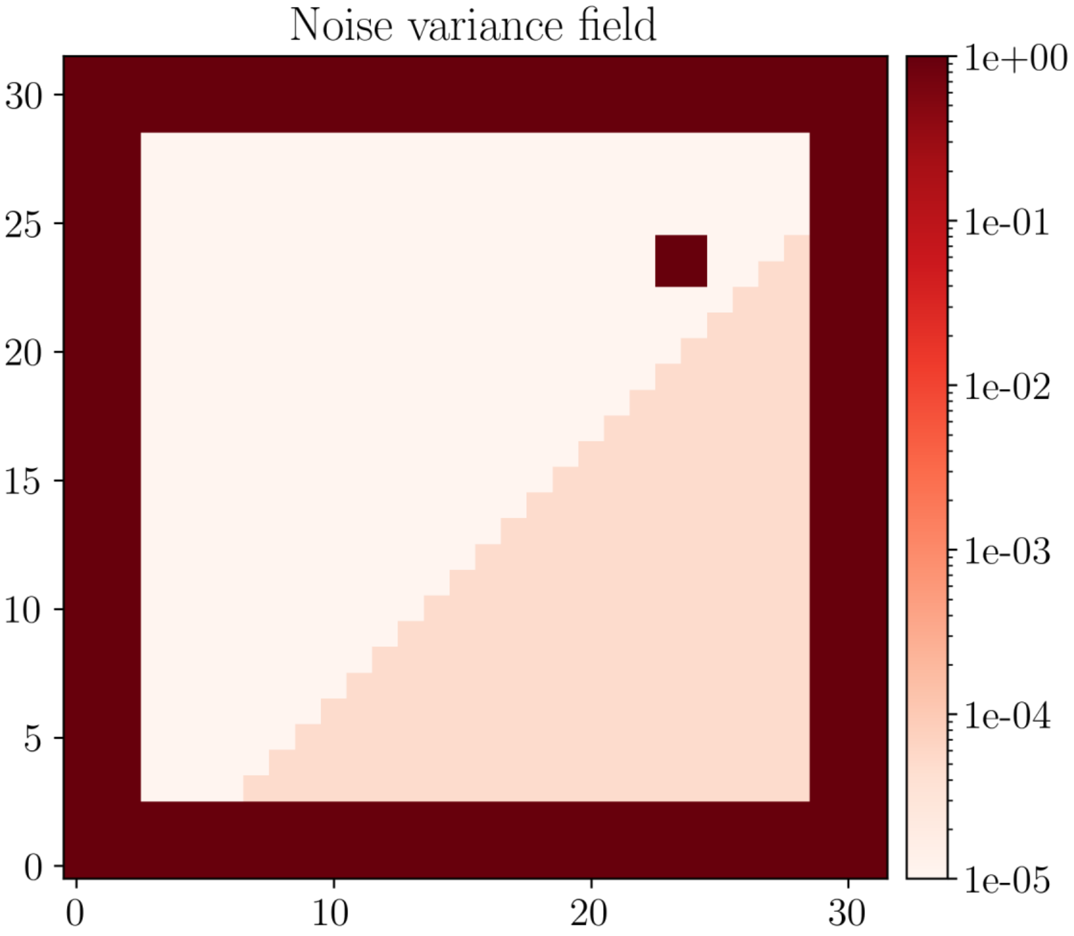
- Generate ground truth cosmological parameters, signal and latent fields

$A_s = 6 \times 10^{-9}$ (arbitrary units) $\quad n_s = 0.96 \quad f_{NL} = 2000$

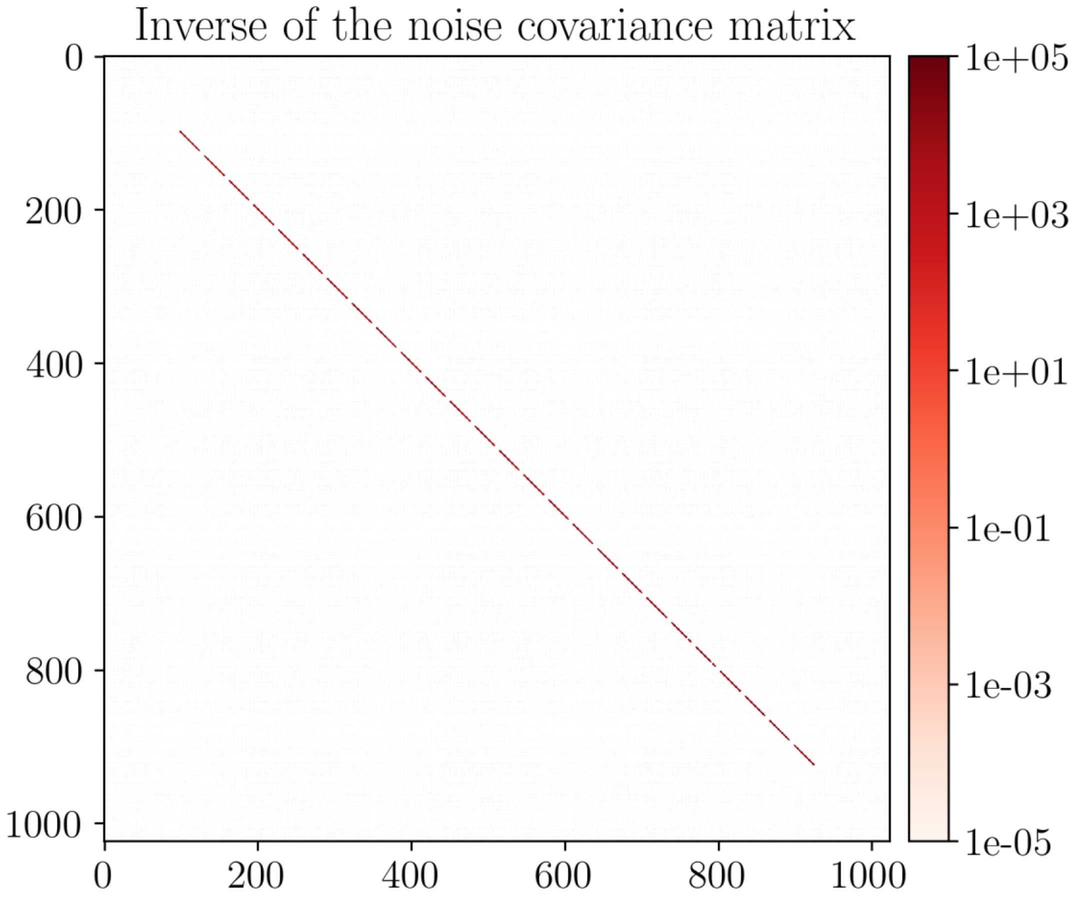


A Bayesian hierarchical field-level model

- Setup Gaussian noise with a covariance matrix as before (diagonal in pixel space)



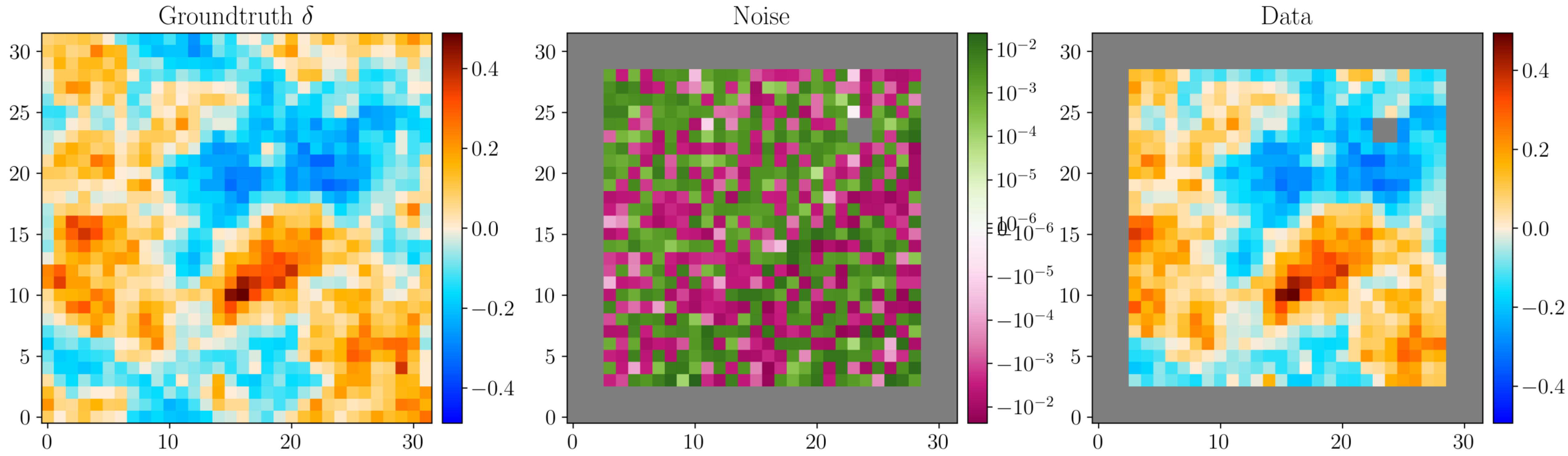
$$N^{-1}$$



A Bayesian hierarchical field-level model

- Generate mock data

$$d = \delta(s) + n$$



Gradients of the Bayesian hierarchical field-level model and its posterior

- Define log-prior, log-likelihood and log-posterior: $\theta = \{A_\mathrm{s}, n_\mathrm{s}, f_\mathrm{NL}, s\}$

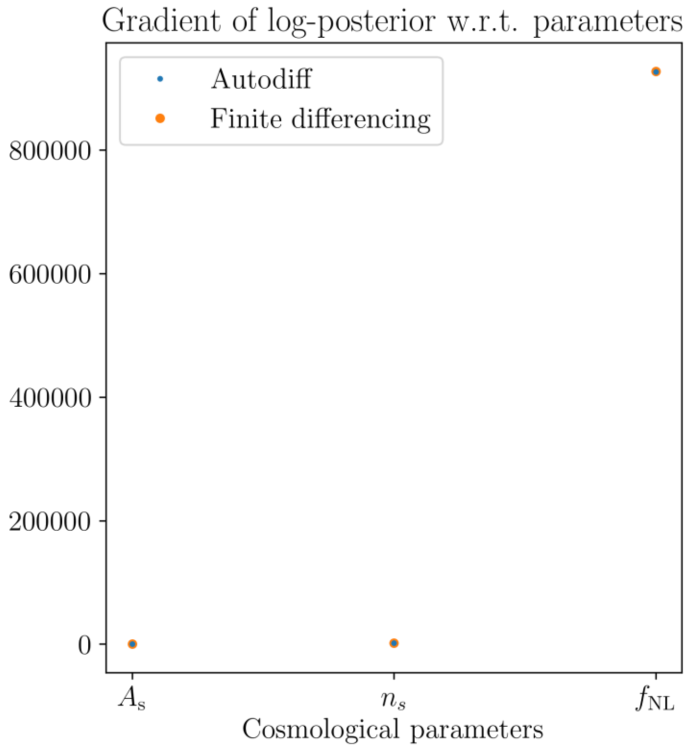
$$\begin{cases} \ln p(A_\mathrm{s}) = -\frac{1}{2}\frac{(A_\mathrm{s}-\mu_{A_\mathrm{s}})^2}{\sigma_{A_\mathrm{s}}^2} + \text{const.} \\ \ln p(n_\mathrm{s}) = -\frac{1}{2}\frac{(n_\mathrm{s}-\mu_{n_\mathrm{s}})^2}{\sigma_{n_\mathrm{s}}^2} + \text{const.} \\ \ln p(f_\mathrm{NL}) = -\frac{1}{2}\frac{(f_\mathrm{NL}-\mu_{f_\mathrm{NL}})^2}{\sigma_{f_\mathrm{NL}}^2} + \text{const.} \\ \ln p(s) = -\frac{1}{2} s^\intercal s + \text{const.} \end{cases}$$

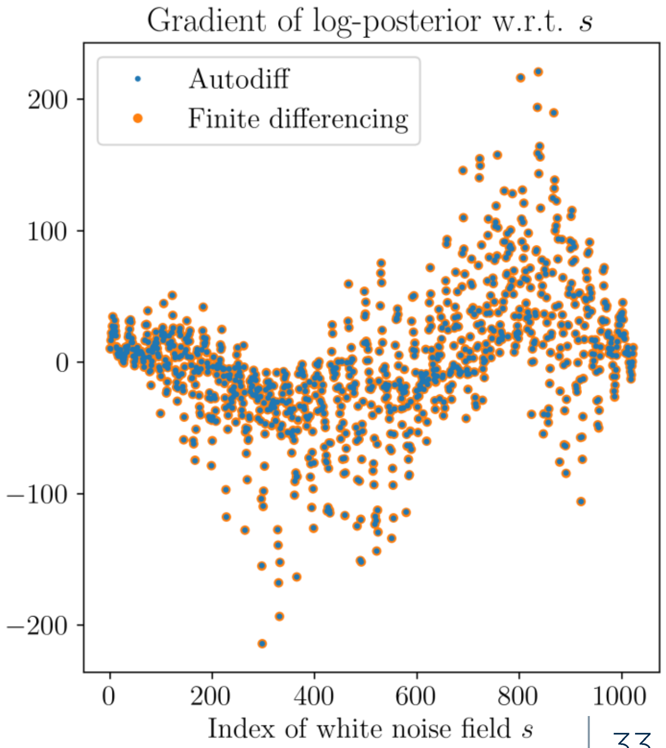
$$\ln p(\theta) = \ln p(A_\mathrm{s}) + \ln p(n_\mathrm{s}) + \ln p(f_\mathrm{NL}) + \ln p(s)$$

$$\ln p(d|\delta) = -\frac{1}{2}\left[\delta - d\right]^\intercal N^{-1} \left[\delta - d\right] + \text{const.}$$

$$\ln p(\theta|d) = \ln p(\theta) + \ln p(d|\theta) + \text{const.}$$

- Compute their gradients (analytically, “manual” differentiation, automatic differentiation).
- Checking the gradient code versus finite differencing is always a useful sanity test.
- Depending on the sampling strategy, one may not need the gradient w.r.t. cosmological parameters (but the gradients w.r.t. to the field are always needed).



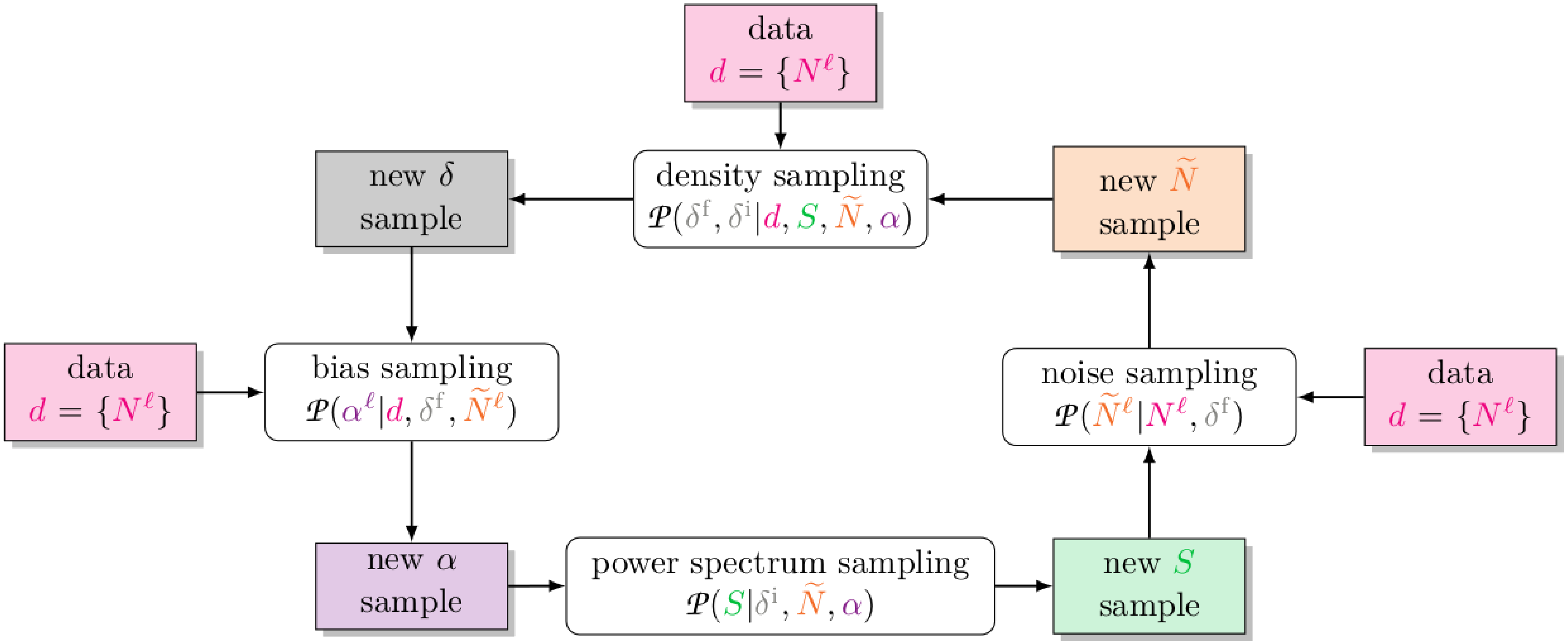


Bayesian hierarchical field-level models: sampling strategy

- Either sample everything together in the same HMC/NUTS (strategy adopted in the Almanac code)
  - Pros: no need to know conditionals, no need to think about different samplers for different parameters.
  - Cons: difficult to tune the mass matrix when field and other parameters are of different nature (and therefore have different effects on the data) and may have different scales.
- Either use Gibbs sampling: Hamiltonian-within-Gibbs for the field and some-sampler-within-Gibbs for the cosmological/other parameters (strategy adopted in the BORG code)
  - Pros: easy to implement, since we usually know the conditional distributions. We can use e.g. slice sampling for cosmological/other parameters, which is rejection-free.
  - Cons: we cannot take diagonal steps in the joint field-parameter space, which can make sampling inefficient.
- Pro tip: work with white noise as the target field parameters, and "whiten" all other target parameters (rescale and shift them so that they take their values in $[0, 1]$) — particularly if you adopt the first strategy.

Modular probabilistic programming: example

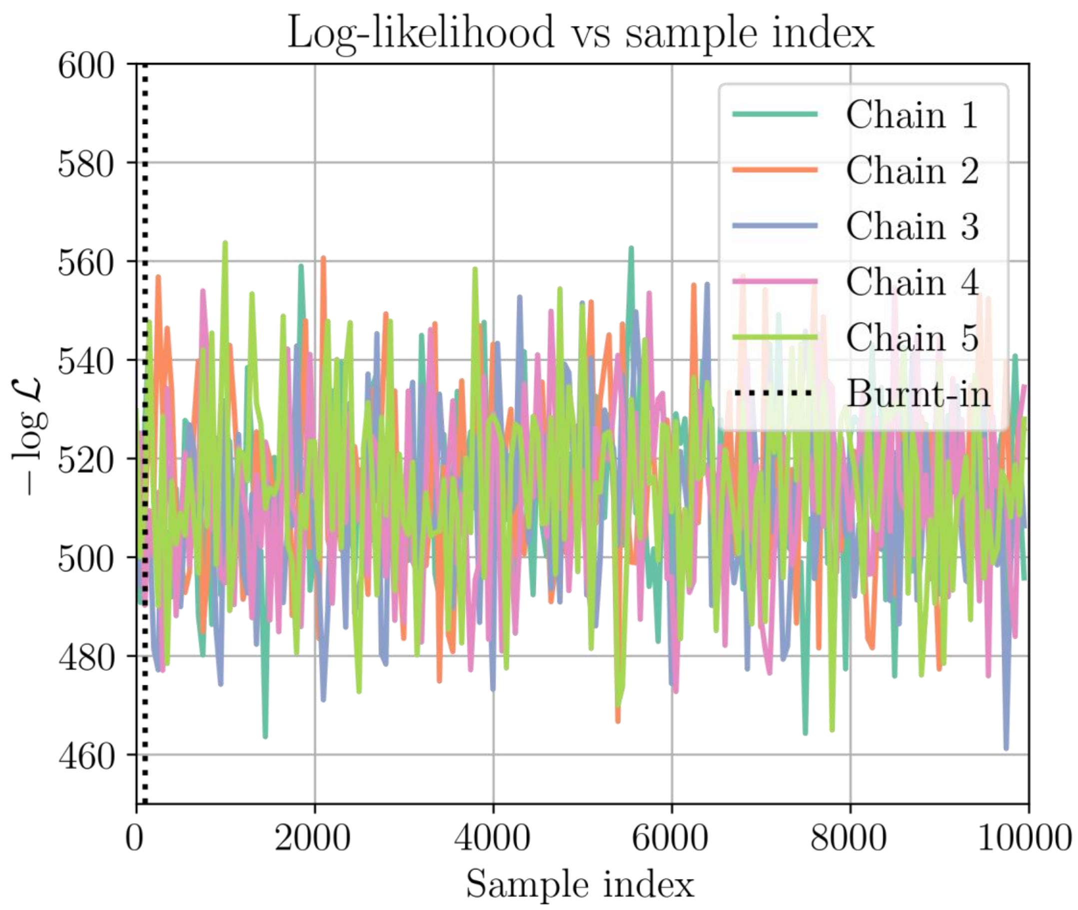
- ARES: Algorithm for Reconstruction and Sampling



Jasche, Kitaura, Wandelt & Enßlin, 0911.2493; Jasche & Wandelt, 1306.1821

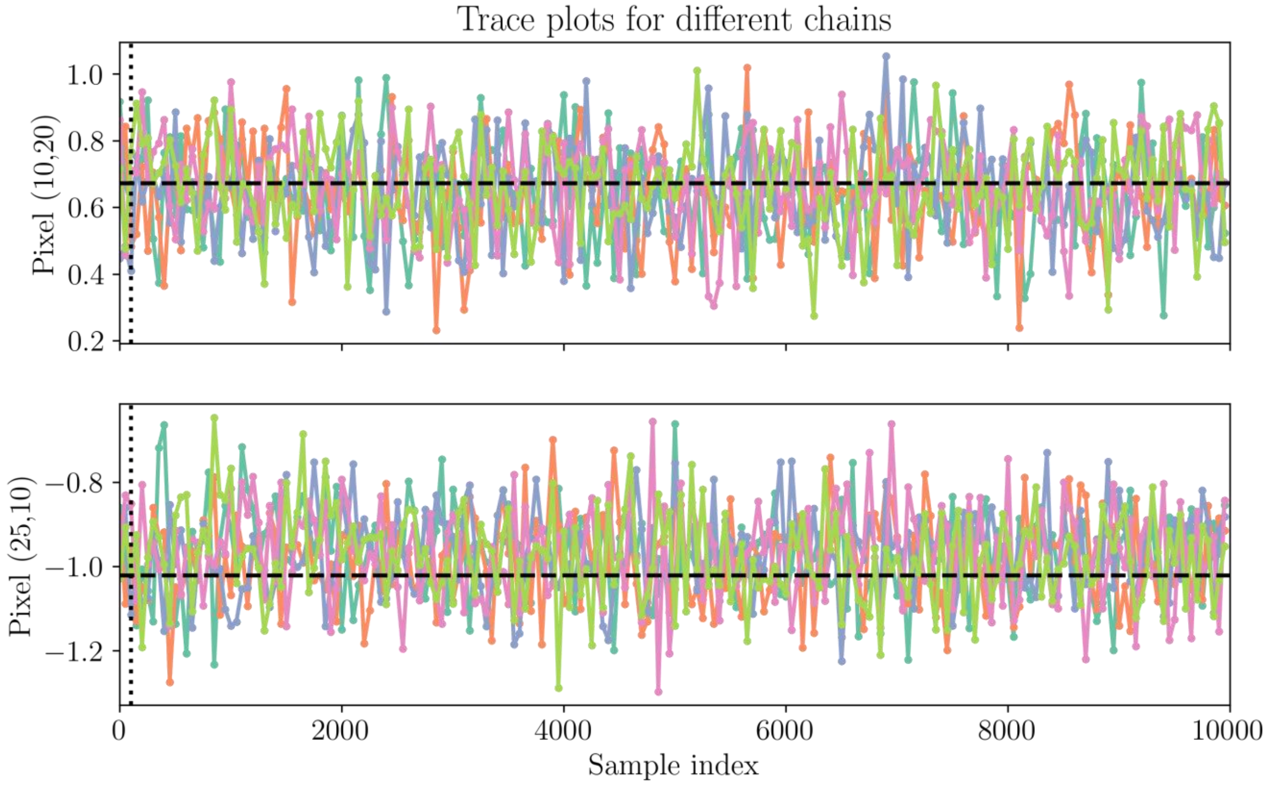
Field-level inference with hierarchical models: sampler tuning and burn-in

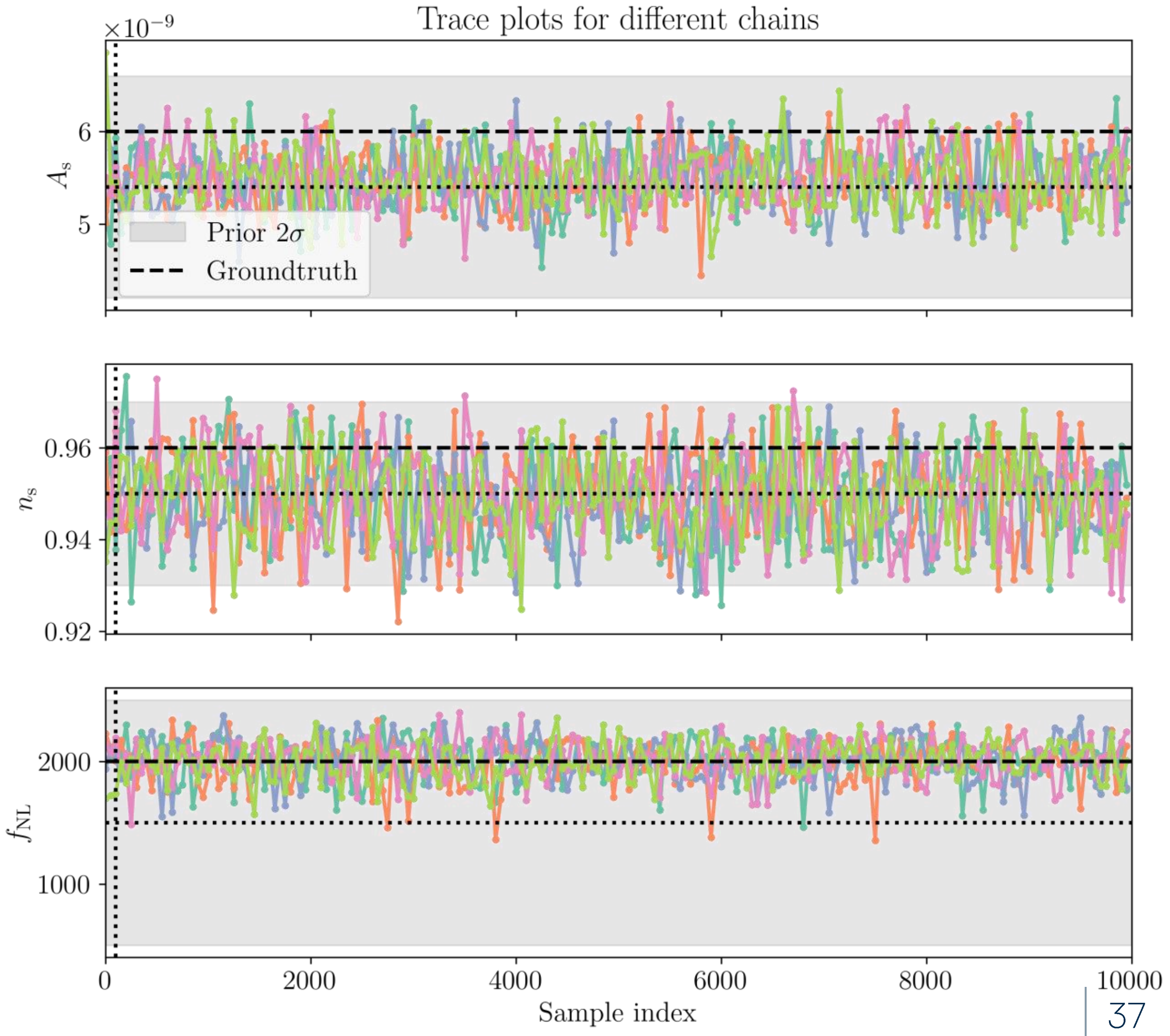
- For a well-tuned sampler, the log-likelihood vs sample index should oscillate around some value.



Field-level inference with hierarchical models: sampler tuning and burn-in

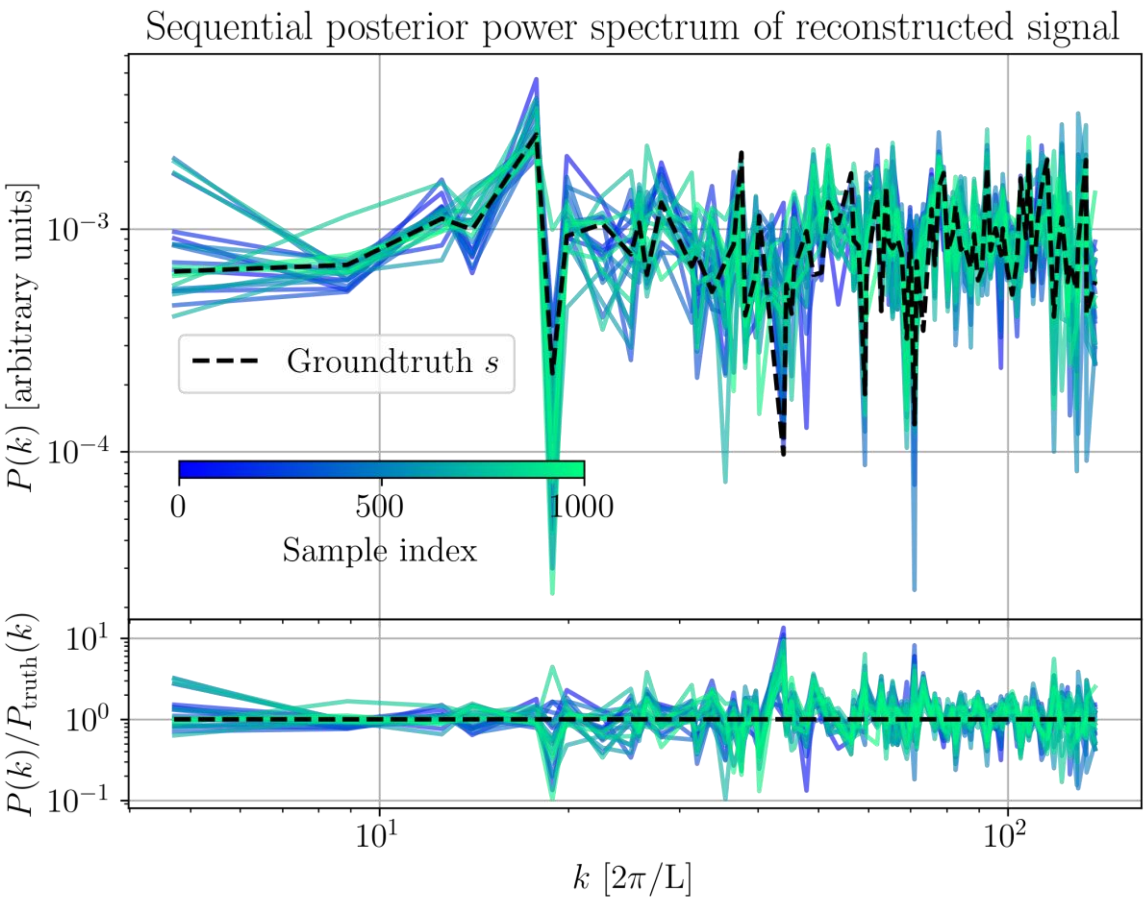
- In addition to the trace plot for pixels values, one should check trace plots of cosmological parameters.

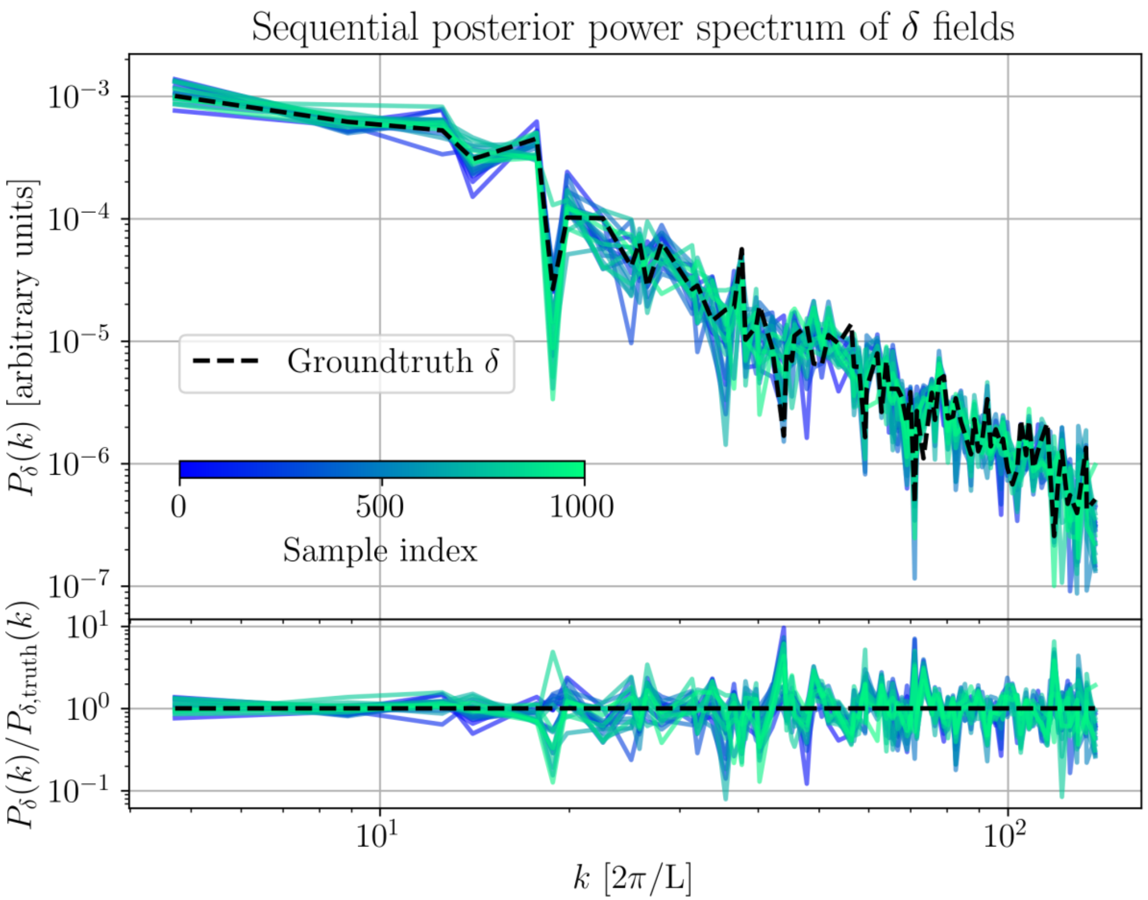




Field-level inference with hierarchical models: sampler tuning and burn-in

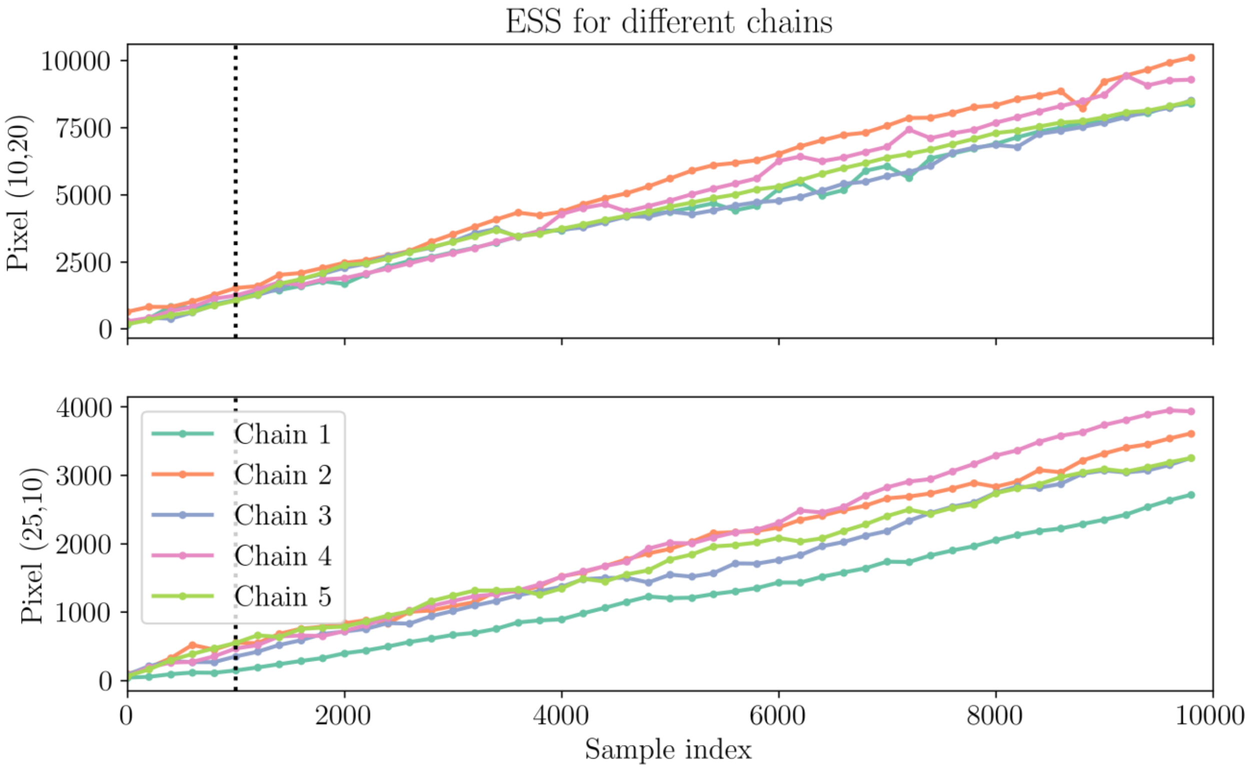
- It remains a good idea to start the chain from an over-dispersed state (for the field) and check the sequential posterior power spectrum of samples.

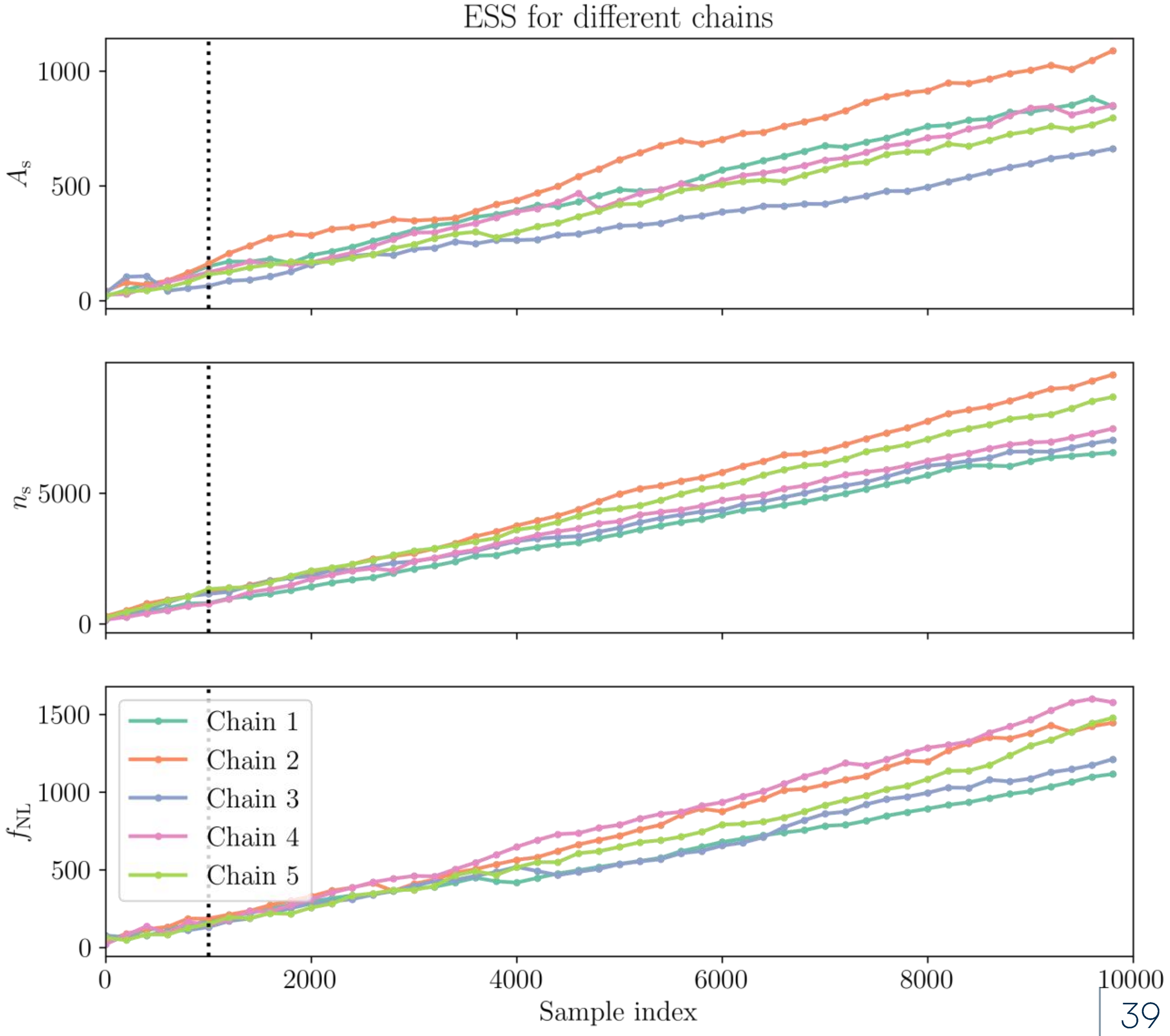




Field-level inference with hierarchical models: autocorrelation and convergence

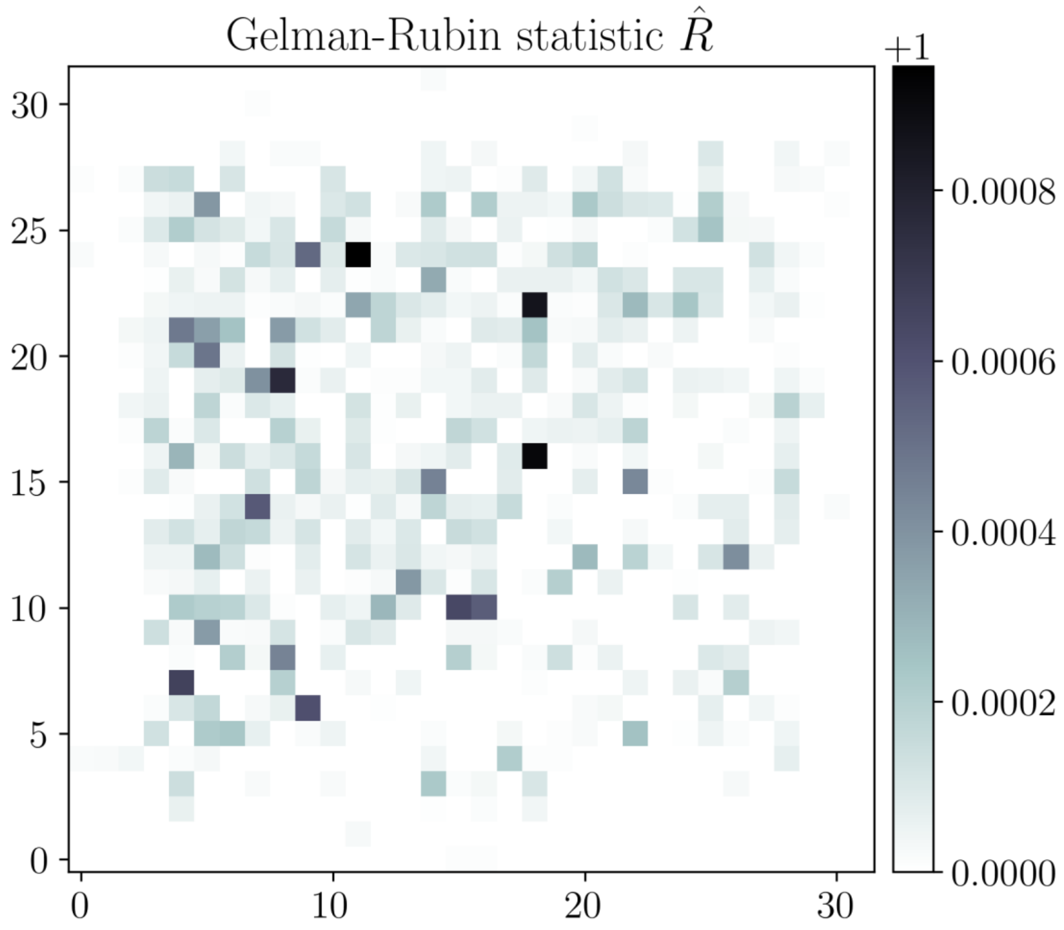
- The effective sample size (ESS) gives number of independent samples. It is especially important for posteriors of cosmological parameters!





Field-level inference with hierarchical models: autocorrelation and convergence

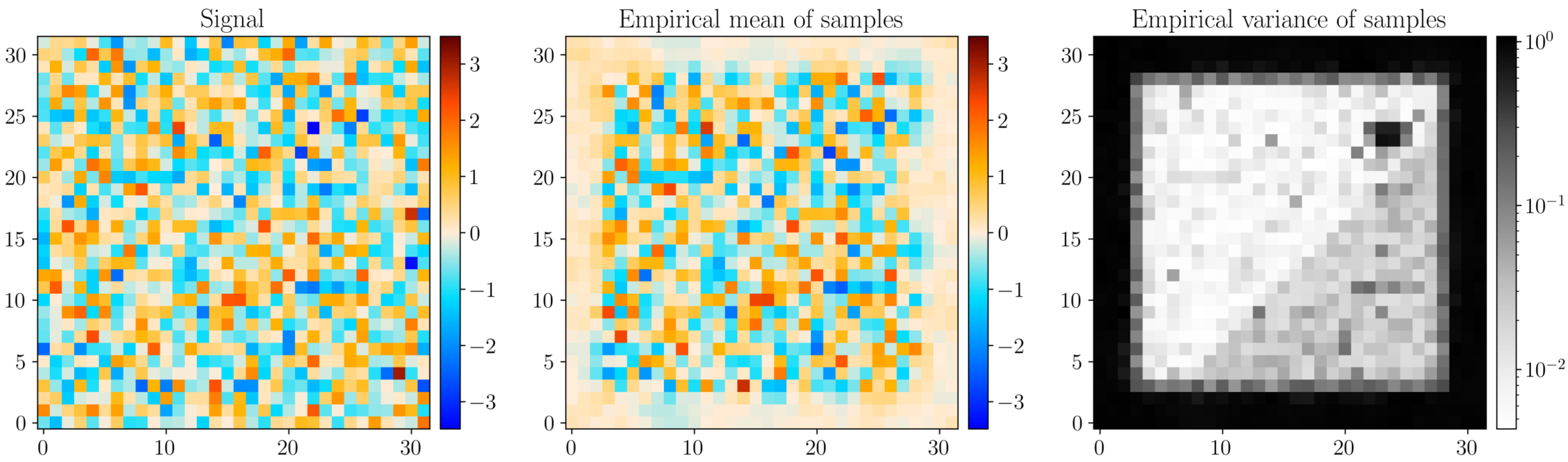
- One can also compute the Gelman-Rubin statistic for the field and for cosmological parameters.
  - Note: the chain can appear converged for some parameters when it is far from converged for the field!



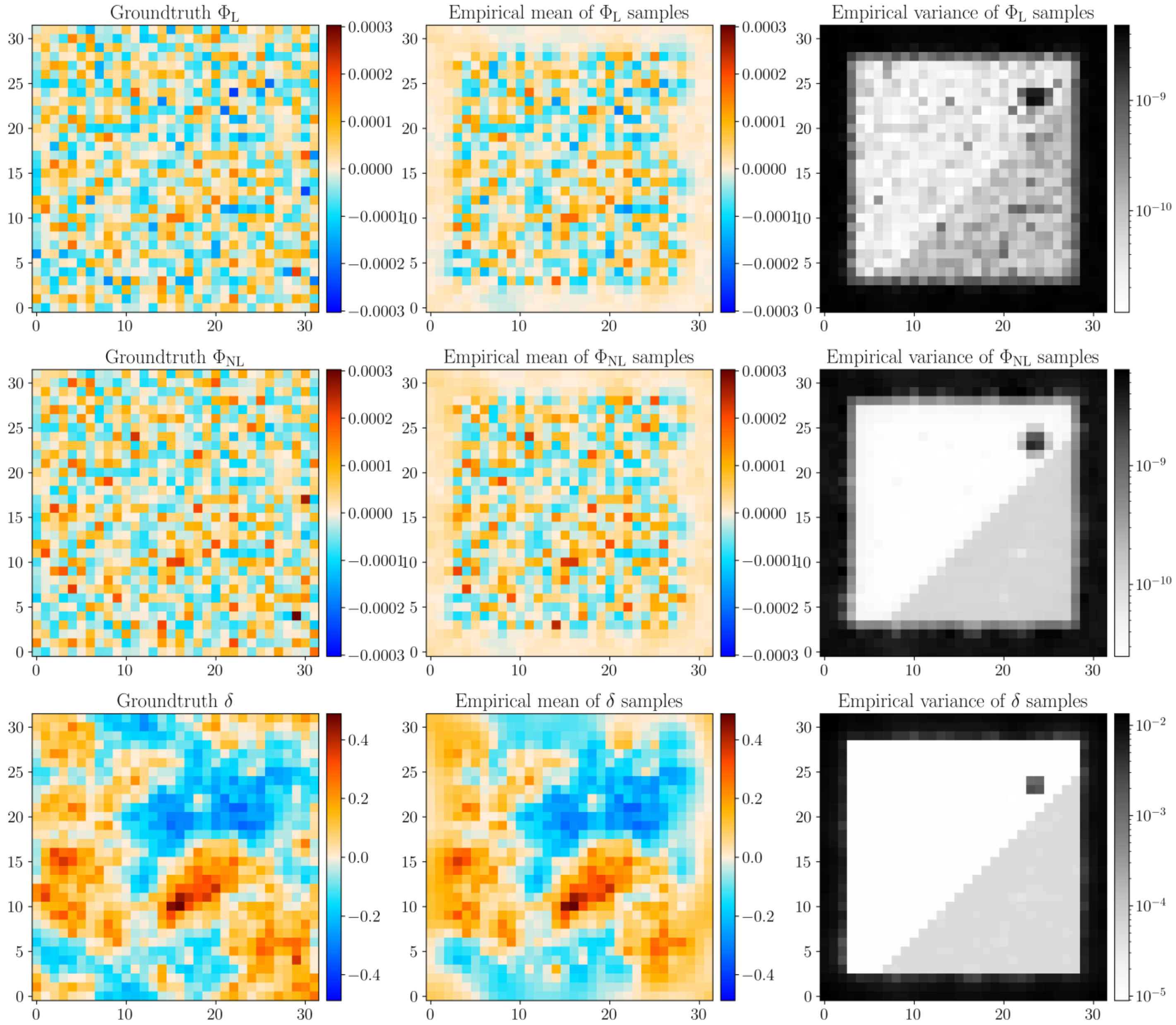
$$\hat{R} - 1 \lesssim 10^{-4} \quad \text{for} \quad \{A_\mathrm{s}, n_\mathrm{s}, f_\mathrm{NL}\}$$

Field-level inference with hierarchical models

- Visualise the empirical mean and empirical variance of samples for the reconstructed signal (these are the target parameters of the problem).



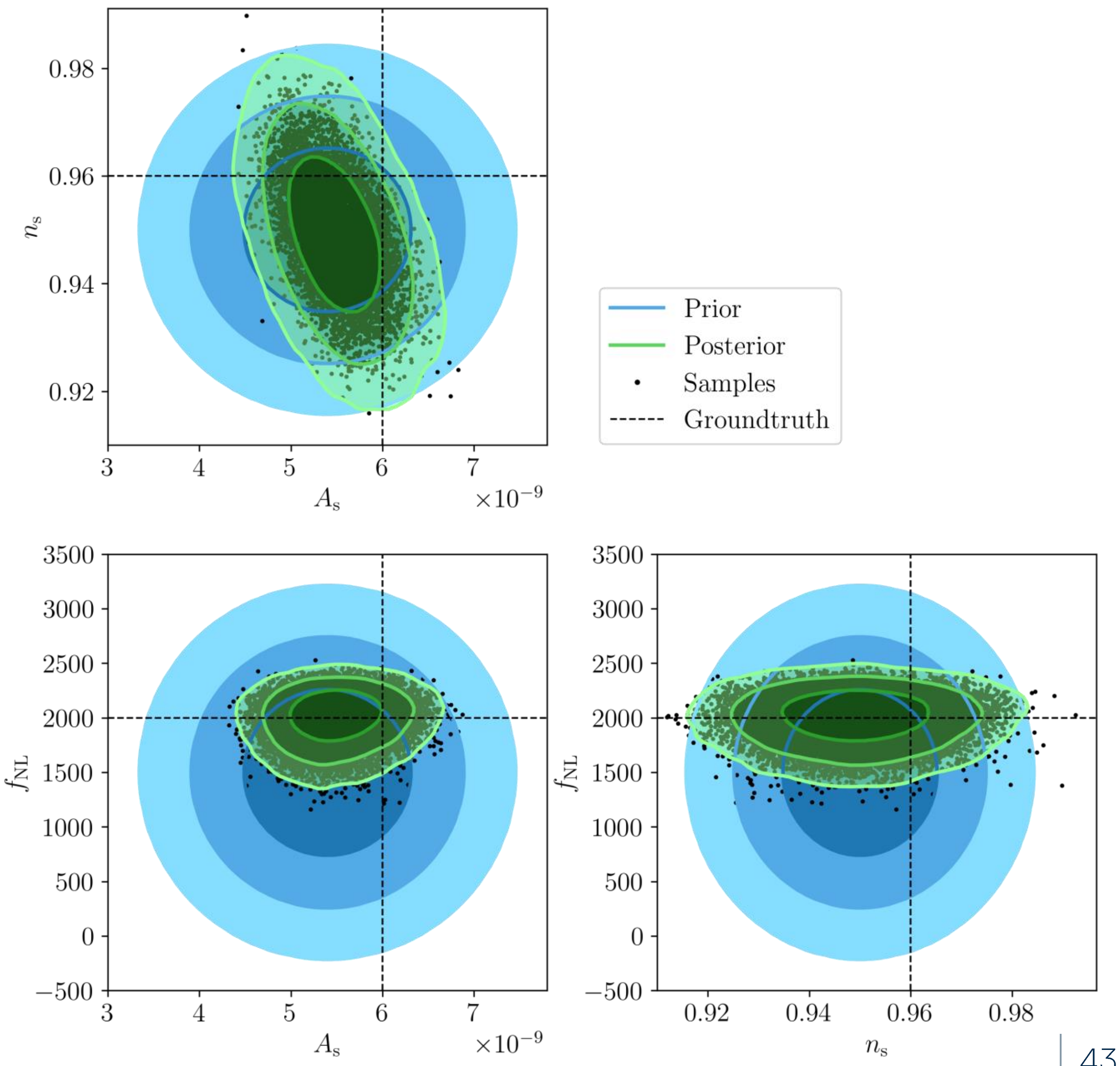
Field-level inference with hierarchical models



- It is also usual to show the empirical mean and empirical variance of samples for latent fields that have an interpretable physical meaning.

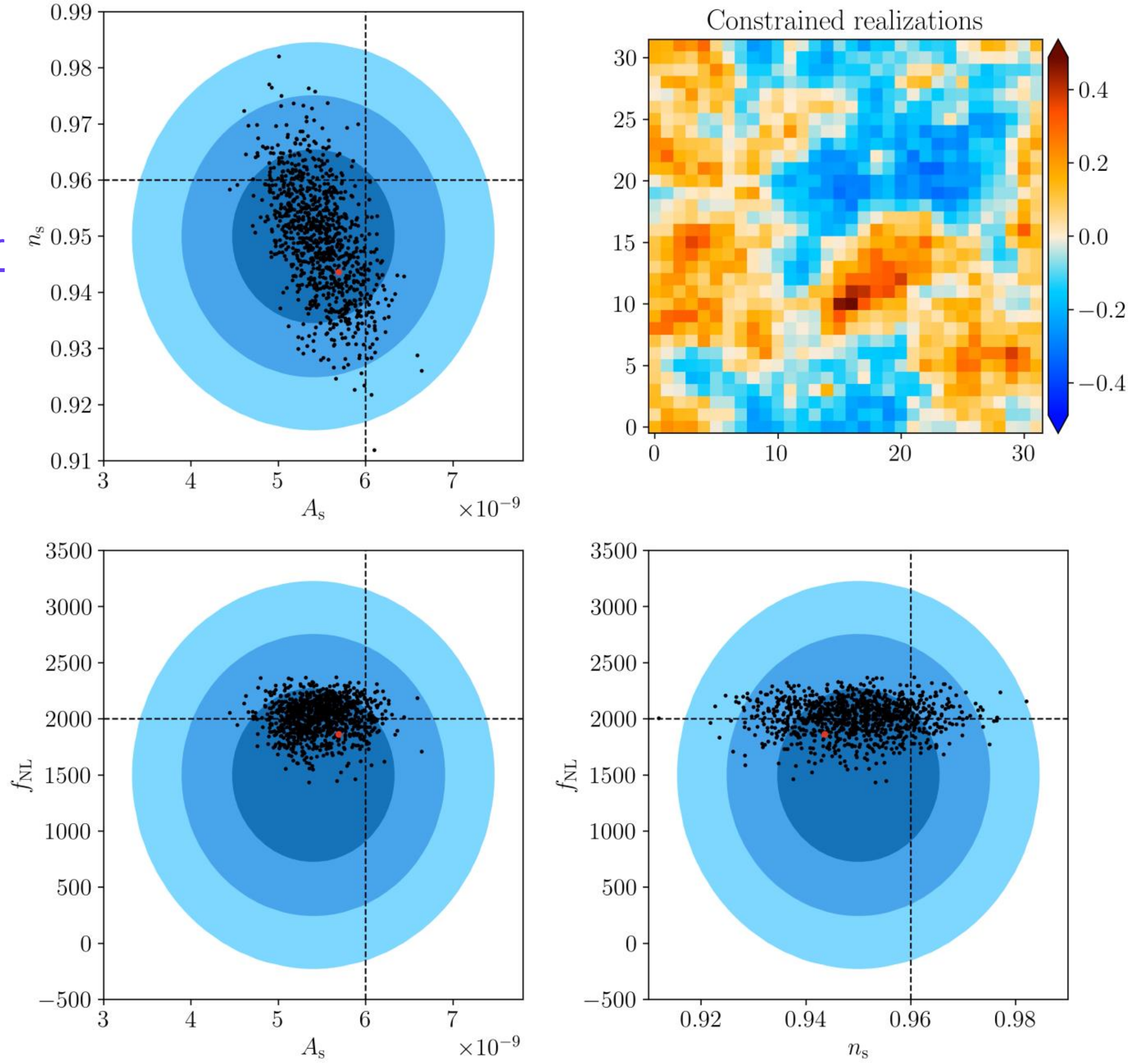
Field-level inference with hierarchical models

- Finally, the posterior for cosmological parameters is usually shown as a corner plot (i.e. contours of the two-dimensional marginals for pairs of parameters).



Field-level inference with hierarchical models

- Each sample of the chain is a constrained realisation of the corresponding field, given the data.
- The chain explores jointly the posterior distribution of cosmological parameters.



References and further reading

References:

- A. Gelman *et al.* (2021), *Bayesian Data Analysis, Third edition*
- C. Geyer (2011), *Introduction to Markov Chain Monte Carlo*
- R. M. Neal (2011), 1206.1901, *MCMC using Hamiltonian Dynamics*

https://florent-leclercq.eu/teaching.php