Markov Walk Exploration of Model Spaces: Bayesian Selection of Dark Energy Models with Supernovæ

Benedikt Schosser, Tobias Röspel, Björn Malte Schäfer

Why model selection?

- Better fit vs. simpler model
- Model evidence ratio (Bayes factor)

  $B = \ln[p(y|M_1)/p(y|M_2)]$

- Evidence and normalization for every model needed
- Fine for 2, 3, …
- Brutal for 100

How to get a full posterior p(M|y) over all models without exhaustive search?

Bayes theorem for parameters

Parameters

Likelihood

Prior

$$p(\theta|y,M) = \frac{\mathcal{L}(y|\theta,M)\pi(\theta|M)}{p(y|M)}$$

Data

$$p(y|M) = \int \mathrm{d}^n\theta\, \mathcal{L}(y|\theta,M)\pi(\theta|M)$$

Bayesian evidence

Bayesian model selection

$$p(M|y) = \frac{p(y|M)\pi(M)}{p(y)}$$

Model evidence

$$p(y) = \sum_i p(y|M_i)\pi(M_i)$$

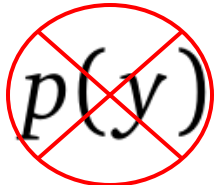
MCMC over model space

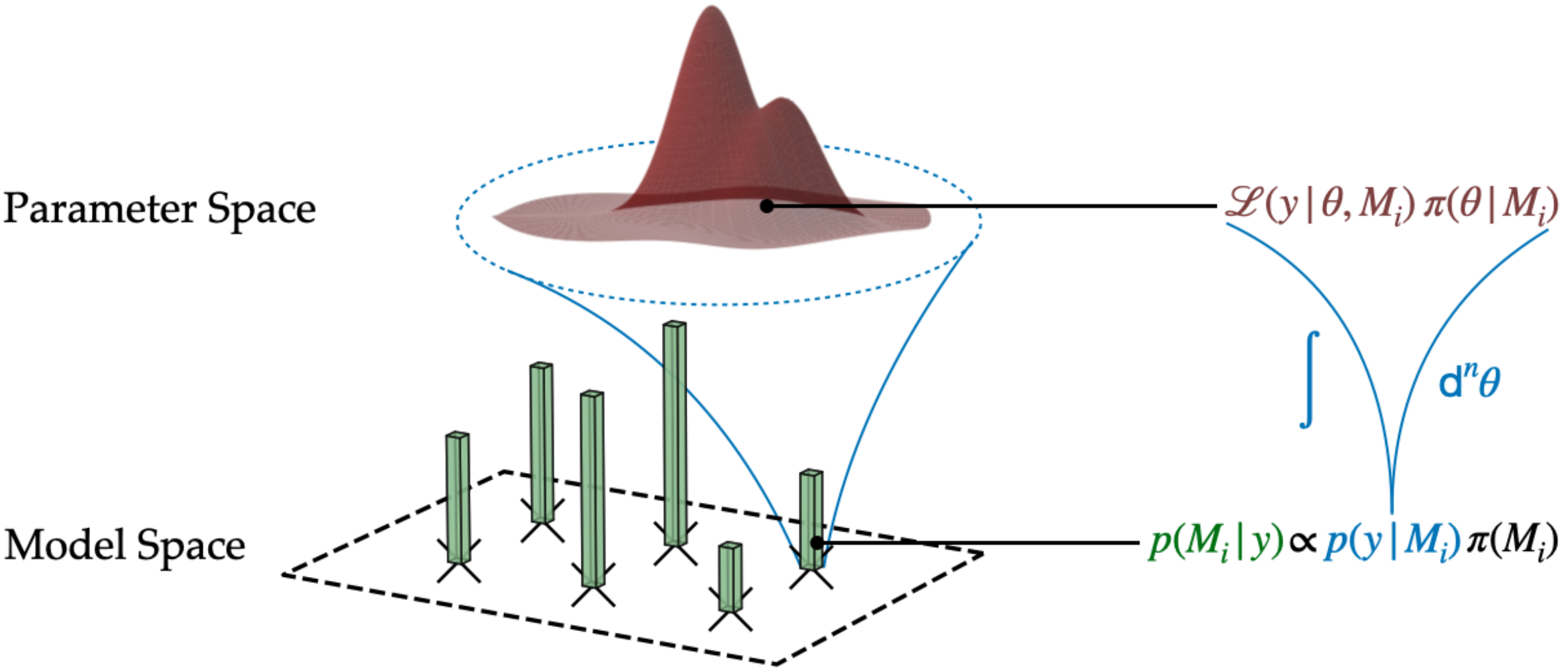
Like for parameters, Metropolis-Hastings walk through **model space**

Transition probability $M_i \rightarrow M_j$:

$$p_{i \rightarrow j} = \min \left[1, \frac{p(y|M_j)\pi(M_j)}{p(y|M_i)\pi(M_i)} \right]$$

Number of visits to a model ∝ posterior probability p(M|y)




Parameter Space
$\mathscr{L}(y \mid \theta, M_i)\, \pi(\theta \mid M_i)$
$\int$
$\mathrm{d}^n\theta$
Model Space
$p(M_i \mid y) \propto p(y \mid M_i)\, \pi(M_i)$

MCMC conditions

1. Ergodicity – each state is reachable in finite number of steps
2. Detailed balance -"time reversibility"

symmetric proposal distribution

$g(M_i | M_j) = g(M_j | M_i)$

Asymmetric distribution

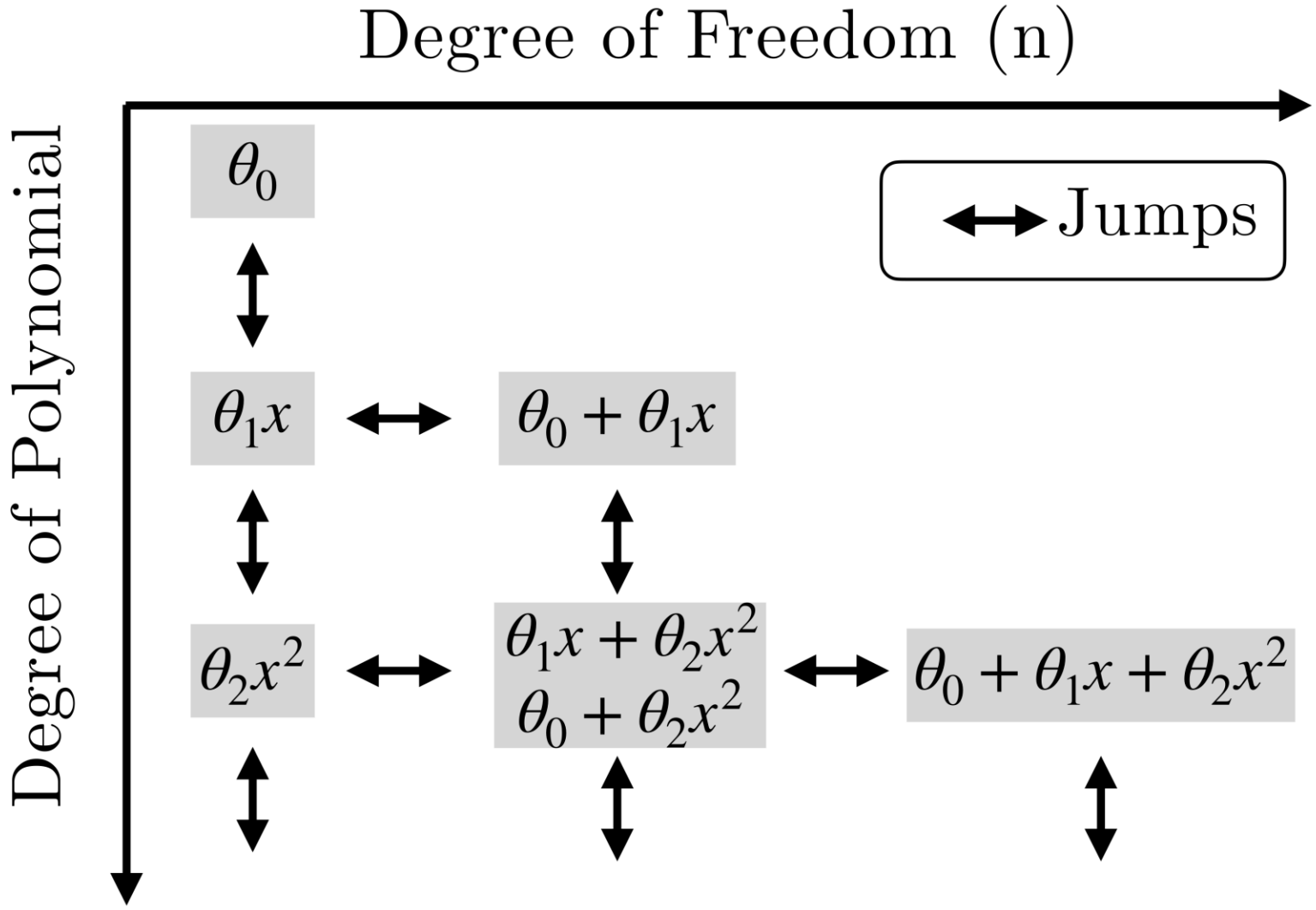
$g(M_i | M_j) \neq g(M_j | M_i)$

Acceptance probability

$$A(M_i, M_j) = \min\left[1, q_{i\to j}\right]$$

$$q_{i\to j} = \frac{p(y|M_j)\pi(M_j)}{p(y|M_i)\pi(M_i)} \frac{g(M_i|M_j)}{g(M_j|M_i)}$$

Model space


Degree of Freedom (n)
Degree of Polynomial
Jumps
θ_0
$\theta_1 x$
$\theta_0 + \theta_1 x$
$\theta_2 x^2$
$\theta_1 x + \theta_2 x^2$
$\theta_0 + \theta_2 x^2$
$\theta_0 + \theta_1 x + \theta_2 x^2$

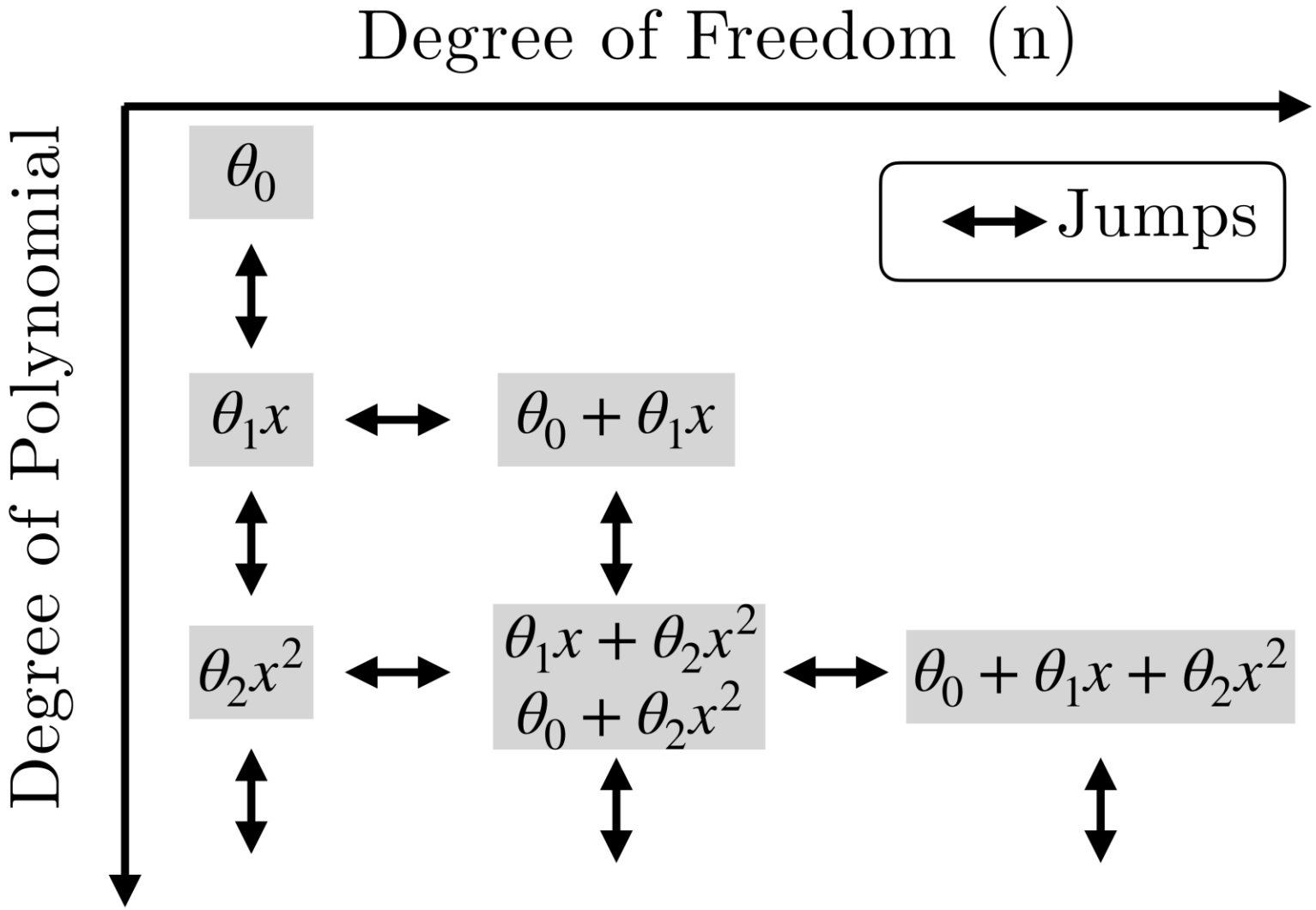
Model space

Number of models at one state

row $\binom{d}{\tilde{n}}$ column

$n = \tilde{n} + 1$

d.o.f.



$$g(M_i|M_j) \neq g(M_j|M_i)$$

Rescaled proposal function

$$g(M_i|M_j) \neq g(M_j|M_i)$$

$$\tilde{g}(j|i) = c_{i\to j}\, g(j|i) = \text{const.} = c_{j\to i}\, g(i|j) = \tilde{g}(i|j) \quad \forall i, j\,,$$

$g(j|i) = g(M_j|M_i)$

$c_{i\to j}$ correction prefactor

$$c_{i\to j}\big|_{\mathcal{P}} = c_{i\to M_1}\cdots c_{M_N\to j} = \left(\frac{1}{2}\right)^{b_i}\binom{d_{M_1}}{\tilde{n}_{M_1}}\cdots\binom{d_{M_N}}{\tilde{n}_{M_N}}\binom{d_j}{\tilde{n}_j}$$

$$c_{j\to i}\big|_{\mathcal{P}} = c_{j\to M_N}\cdots c_{M_1\to i} = \left(\frac{1}{2}\right)^{b_j}\binom{d_{M_N}}{\tilde{n}_{M_N}}\cdots\binom{d_{M_1}}{\tilde{n}_{M_1}}\binom{d_i}{\tilde{n}_i},$$

b_i number of borders

$$\frac{g(j|i)}{g(i|j)} = \left(\frac{1}{2}\right)^{b_j-b_i}\frac{\binom{d_i}{\tilde{n}_i}}{\binom{d_j}{\tilde{n}_j}}$$

Binary keys

$$\kappa = (0, 1, 1)$$

$$d = len(\kappa)$$

$$n = sum(\kappa)$$

Priors

$$\pi(\theta \mid M_i)$$

Parameters

$$\pi(M_i)$$

Model

Model priors

$$\pi_{\mathrm{OVN}}(M_i) = \frac{1}{n}$$

one over n

$$\pi_{\mathrm{NP}}(d, n) = \frac{1}{(d+1)^{n+1}}$$

normalisable prior

	AIC	BIC	OVN	NP	U
IC	$\hat{\chi}^2 + 2n$	$\hat{\chi}^2 + 2n \ln N$	$\hat{\chi}^2 + 2 \ln n$	$\hat{\chi}^2 + (n+1) \ln(d+1)$	$\hat{\chi}^2$
π	$\exp(-n)$	$N^{-n/2}$	$1/n$	$1/(d+1)^{n+1}$	1

Algorithm 1 Markov Walk Exploration of Model Spaces

Require: Initial model $M^{(0)}$, data y, model prior $\pi(M)$, parameter priors $\pi(\theta \mid M)$, Poisson rate λ, number of iterations N

Ensure: Empirical estimate of $p(M \mid y)$

Initialize storage for evidence estimates $\{\hat{p}(y \mid M)\}$
for $t = 1$ to N **do**
    $M_{\text{curr}} \leftarrow M^{(t-1)}$
    Draw $K \sim \text{Poisson}(\lambda)$
    $M_{\text{prop}} \leftarrow M_{\text{curr}}$
    for $k = 1$ to K **do**
        Propose a single-edge move $M_{\text{prop}} \to M'$ among its neighbors
        $M_{\text{prop}} \leftarrow M'$
    if $\hat{p}(y \mid M_{\text{prop}})$ not computed **then**
        Compute $\hat{p}(y \mid M_{\text{prop}})$ (e.g. via nested sampling)
    Compute Hastings ratio:

$$r = \frac{\hat{p}(y \mid M_{\text{prop}})\,\pi(M_{\text{prop}})}{\hat{p}(y \mid M_{\text{curr}})\,\pi(M_{\text{curr}})} \times \frac{g(M_{\text{curr}} \mid M_{\text{prop}})}{g(M_{\text{prop}} \mid M_{\text{curr}})}$$

    $A \leftarrow \min\{1, r\}$
    Draw $u \sim \text{Uniform}(0, 1)$
    if $u < A$ **then**
        $M^{(t)} \leftarrow M_{\text{prop}}$
    else
        $M^{(t)} \leftarrow M_{\text{curr}}$
return Empirical frequencies of $\{M^{(t)}\}$ as estimate of $p(M \mid y)$

Toy model

Polynomial

$$y^i_{\text{Model}} = A^i{}_\alpha \theta^\alpha .$$

$$p(\theta|y,M) = \frac{\mathcal{L}(y|\theta,M)\pi(\theta|M)}{p(y|M)}$$

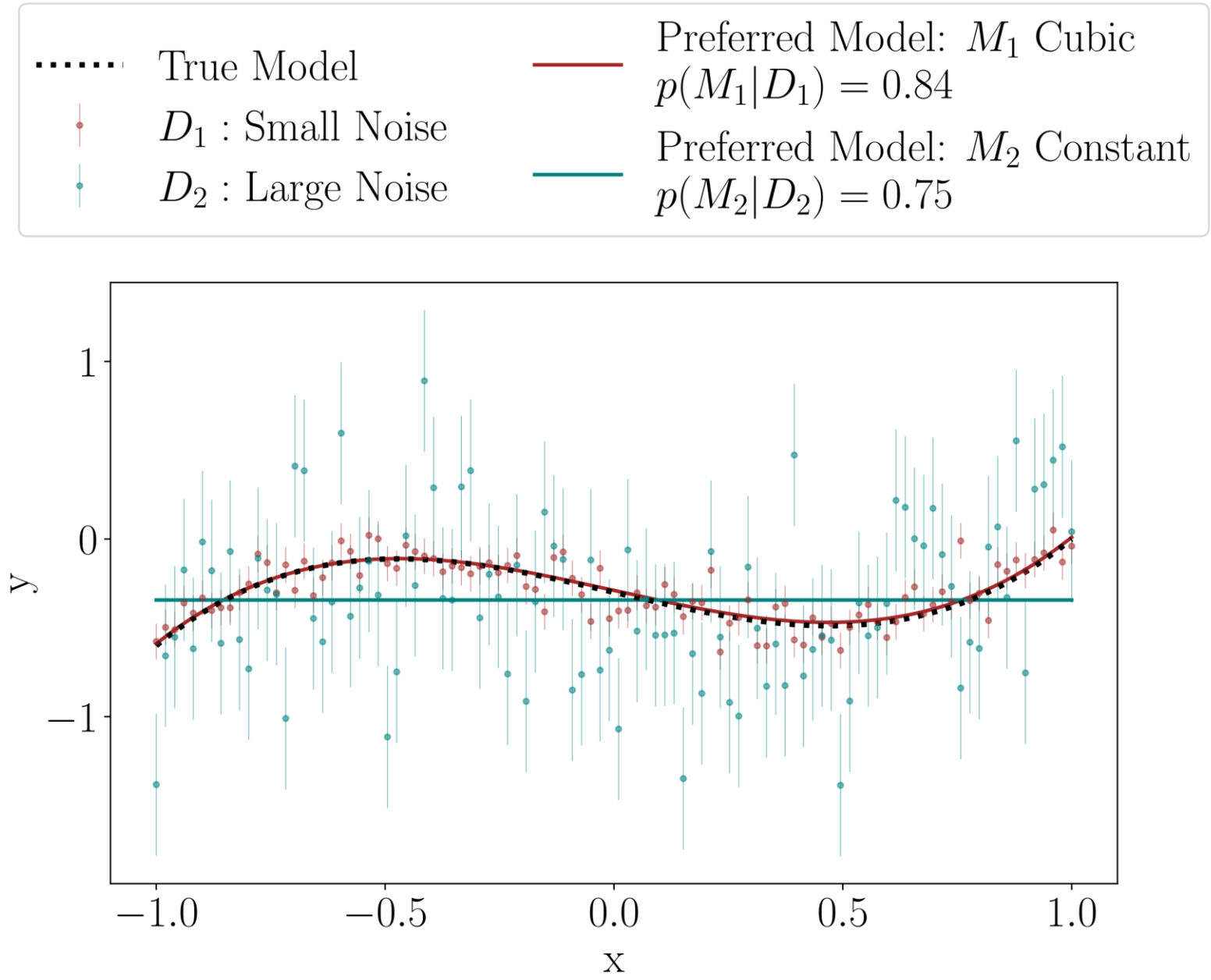
$$\begin{aligned}\chi^2 &= (y^i - A^i{}_\alpha \theta^\alpha) C_{ij} (y^j - A^j{}_\beta \theta^\beta) \\ &= y^i C_{ij} y^j - 2 \underbrace{y^i C_{ij} A^j{}_\alpha}_{:=Q_\alpha} \theta^\alpha + \underbrace{A^i{}_\alpha C_{ij} A^j{}_\beta}_{:=F_{\alpha\beta}} \theta^\alpha \theta^\beta ,\end{aligned}$$

$$\pi(\theta|M) = \frac{1}{\sqrt{(2\pi)^n \det G^{\alpha\beta}}} \exp\left(-\frac{1}{2}(\theta-\bar{\theta})^\alpha G_{\alpha\beta} (\theta-\bar{\theta})^\beta\right)$$

$$\begin{aligned}\ln p(y|M) = &-\ln \mathcal{N}_{\mathcal{L}} - \frac{n}{2}\ln 2\pi + \frac{1}{2}\ln\det G_{\alpha\beta} \\ &- \frac{1}{2} y^i C_{ij} y^j - \frac{1}{2} G_{\alpha\beta}\bar{\theta}^\alpha \bar{\theta}^\beta \\ &+ \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\det\left(F_{\alpha\beta} + G_{\alpha\beta}\right) \\ &+ \frac{1}{2}(F+G)^{-1\,\alpha\beta}(Q_\alpha + G_{\alpha\gamma}\bar{\theta}^\gamma)(Q_\beta + G_{\beta\nu}\bar{\theta}^\nu) .\end{aligned}$$

Toy model

Polynomial



Toy model

Polynomial

Three datasets with $\sigma = [1, 0.1, 0.01]$

- For each model random choice whether a monomial is present or not
- Every 40 points allowed maximal polynomial degree increases (up to 5 for 200 points)

Toy model

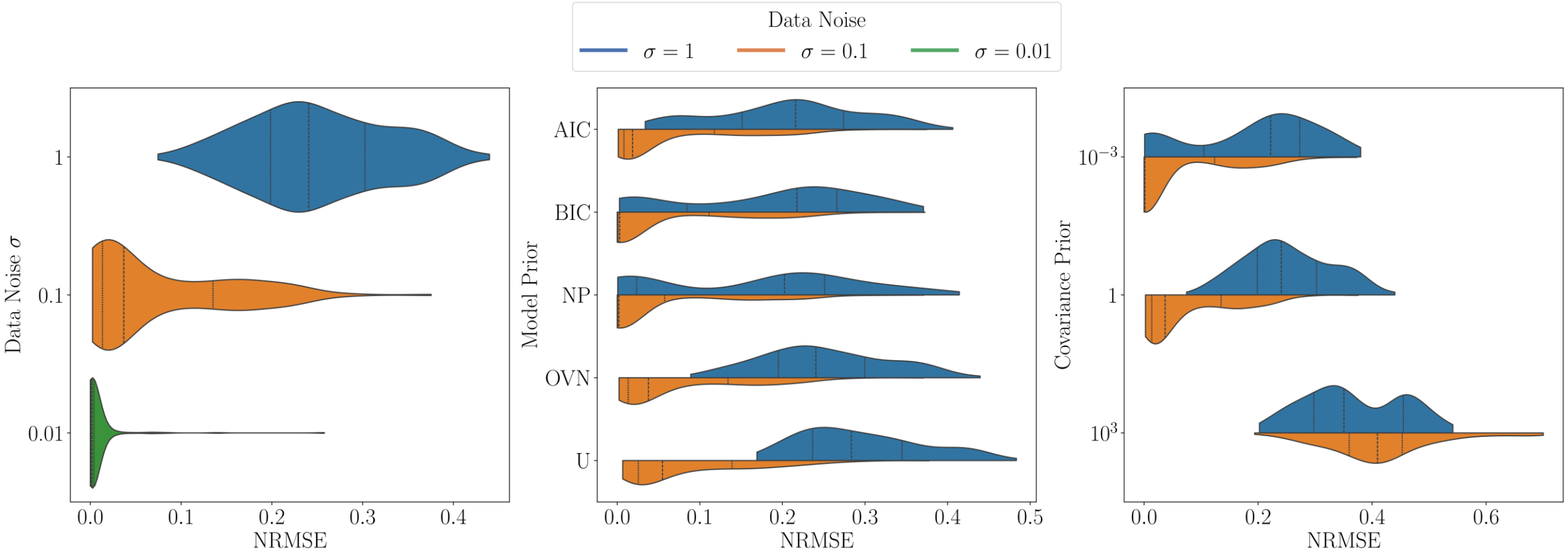
Polynomial

$$y^i_{\text{Model}} = A^i{}_\alpha \theta^\alpha .$$

Normalised Root Mean Square Error (NRMSE)

$$\text{NRMSE}(\kappa_{\text{gt}}, \kappa_{\text{fit}}) = \frac{\sqrt{\sum_\alpha (\kappa^\alpha_{\text{gt}} - \kappa^\alpha_{\text{fit}})^2}}{\sqrt{\sum_\alpha (\kappa^\alpha_{\text{gt}})^2}}$$

Toy model


Data Noise
$\sigma = 1$
$\sigma = 0.1$
$\sigma = 0.01$
Data Noise σ
1
0.1
0.01
0.0
0.1
0.2
0.3
0.4
NRMSE
Model Prior
AIC
BIC
NP
OVN
U
0.0
0.1
0.2
0.3
0.4
0.5
NRMSE
Covariance Prior
10^{-3}
1
10^{3}
0.0
0.2
0.4
0.6
NRMSE

Application to cosmology: Ia supernovae

Friedmann-Lemaître-Robertson-Walker spacetimes

$$w(a) = w_0 + w_1(1-a) + \frac{w_2}{2}(1-a)^2 + \cdots = \sum_{j=0}^{d} \frac{w_j}{j!}(1-a)^j$$

For a given Hubble function:

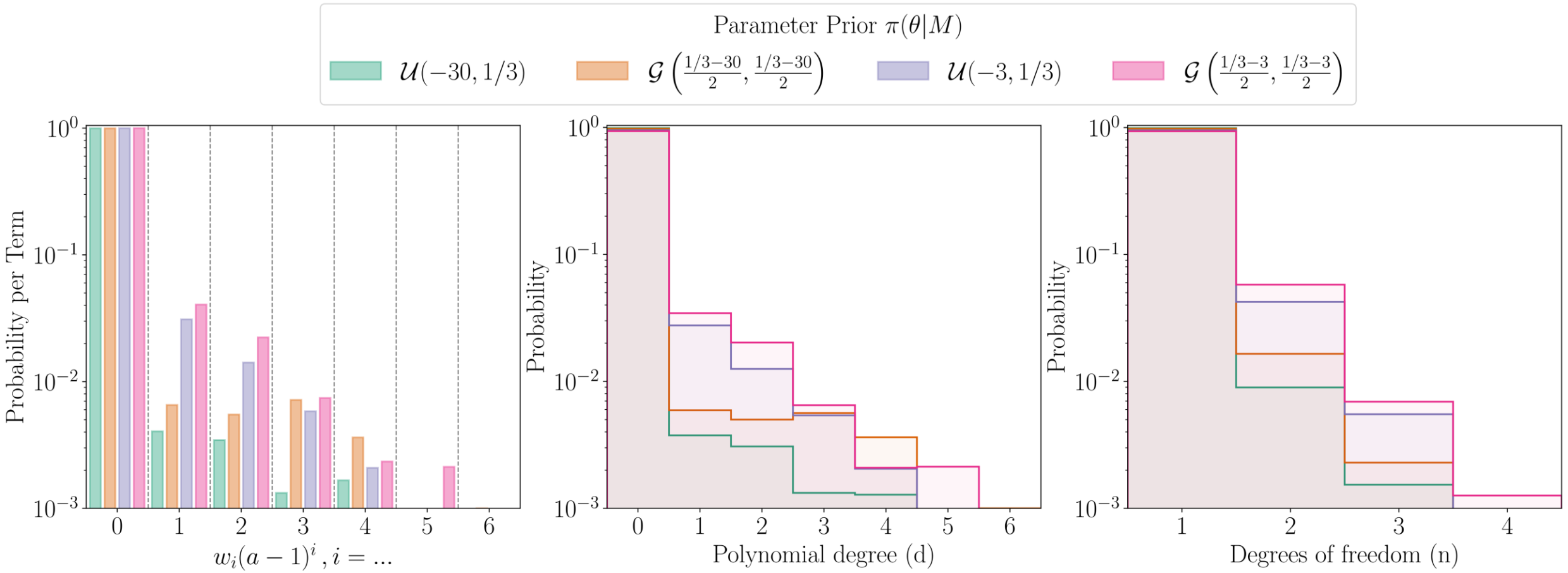
$$\mu = m - M = 5\log_{10}(d_L(a,\theta)) + 10\,, \qquad d_L(a,\theta) = \frac{c}{a}\int_a^1 \mathrm{d}a' \frac{1}{H(a',\theta)}\,,$$

sample

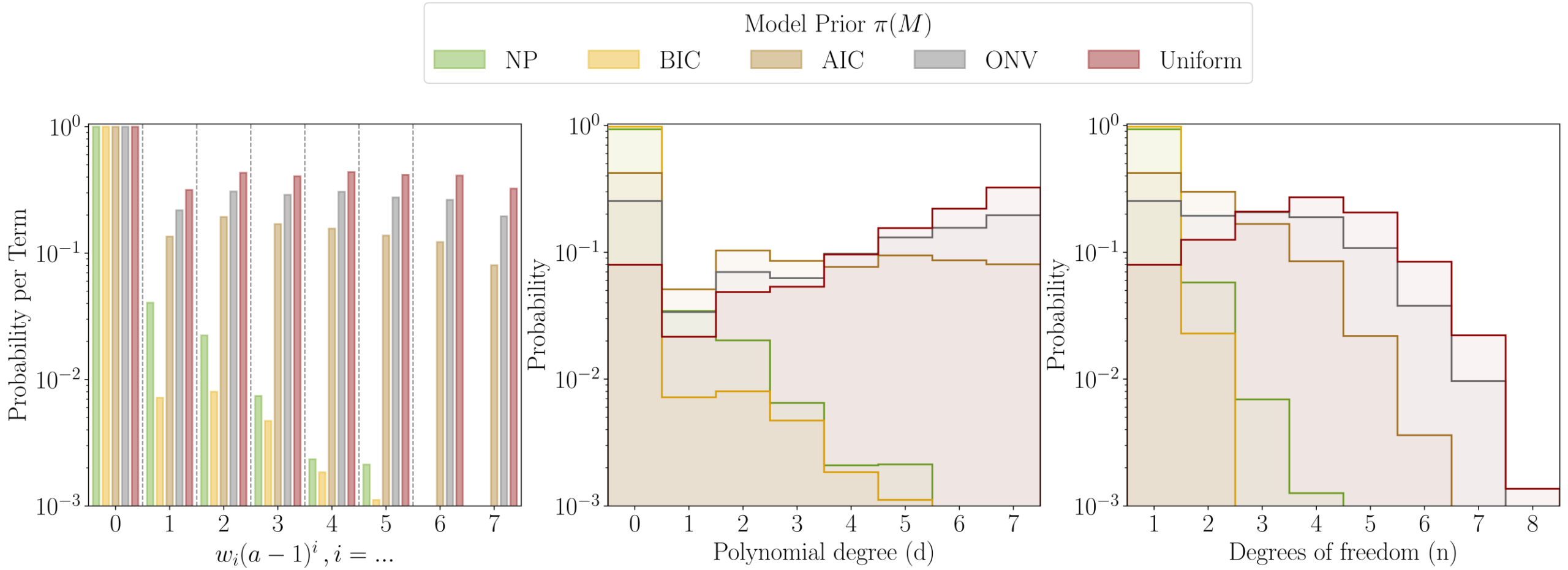
fix

$$\frac{H(a,\theta)^2}{H_0^2} = \frac{\Omega_m}{a^3} + (1-\Omega_m)\exp\left[-3\int_1^a \mathrm{d}a' \frac{1+w(a')}{a'}\right]$$

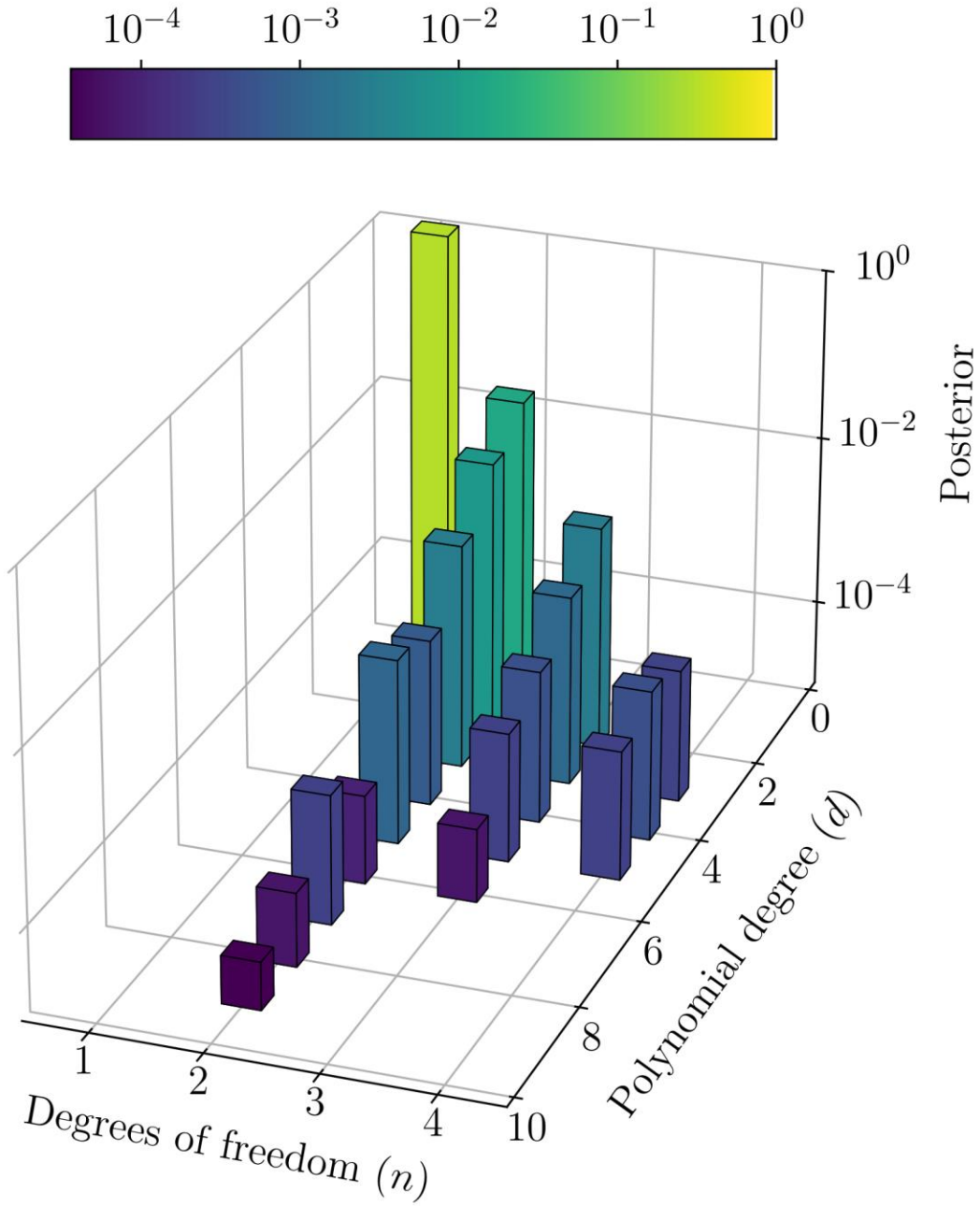
Parameter Prior



Model Prior


Model Prior $\pi(M)$
NP
BIC
AIC
ONV
Uniform
Probability per Term
$w_i(a-1)^i\,, i = ...$
Probability
Polynomial degree (d)
Probability
Degrees of freedom (n)

Posterior


10^{-4}
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{0}
10^{-2}
10^{-4}
Posterior
1
2
3
4
Degrees of freedom (n)
0
2
4
6
8
10
Polynomial degree (d)

Conclusions

- Generalizes pairwise Bayes factor comparison to full model posteriors
- Efficient: skips bad models, reuses computed evidences
- Only polynomial models considered
- Sensitivity to prior choice on parameters non-trivial
- Prior choice of $\pi(M_i)$ - should follow principle of maximised entropy similarly as in the choice of $\pi(\theta|M_i)$

Partition function

temperature

$$Z[\beta, J|M] = \int \mathrm{d}^n\theta \; [\mathcal{L}(y|\theta, M)\pi(\theta|M) \exp(J_\alpha \theta^\alpha)]^\beta \; ,$$

$$\beta = 1 \text{ and } J = 0$$

Bayesian evidence $p(y|M) = \int \mathrm{d}^n\theta \; \mathcal{L}(y|\theta, M)\pi(\theta|M)$

$$F(\beta, J|M) = -\frac{1}{\beta} \ln Z[\beta, J|M] \quad \Rightarrow \quad p(\theta|y, M)$$

Helmholtz free energy

Canonical partitions on model spaces

temperature

$$Y[\zeta] = \sum_i [p(y|M_i)\pi(M_i)]^{\zeta},$$

$$\zeta = 1$$

Model evidence $p(y) = \sum_i p(y|M_i)\pi(M_i)$

Partition of partition

$$Y[\zeta, \beta] = \sum_i \left(\int \mathrm{d}^n\theta \, [\mathcal{L}(y|\theta, M_i)\pi(\theta|M_i)]^{\beta} \, \pi(M_i) \right)^{\zeta}$$

$$= \sum_i (Z[\beta|M_i]\pi(M_i))^{\zeta},$$

Entropy, specific heat

Thermodynamic potential:

$$\Phi(\zeta) = -\frac{1}{\zeta} \ln Y[\zeta]$$

Entropy:

$$S(\zeta) = -\frac{1}{\zeta^2} \frac{\partial \Phi(\zeta)}{\partial \zeta}$$

Shannon's information entropy

$$S = -\sum_i p(M_i|y) \ln p(M_i|y) = \langle \ln p(M_i|y) \rangle \quad \text{at} \quad \zeta = 1\,.$$

Specific heat:

$$C = -\zeta \frac{\partial S}{\partial \zeta} = \left\langle \ln^2 p(M_i|y) \right\rangle - \langle \ln p(M_i|y) \rangle^2 \quad \text{at} \quad \zeta = 1\,,$$

Related partition:

$$Y'[\zeta] = \sum_i p(y|M_i)^\zeta \pi(M_i)^{(\zeta+1/\zeta)/2}$$

$$\frac{d}{d\zeta} \ln\left(\frac{Y'}{\zeta}\right)$$

$$\Delta S = \sum_i p(M_i|y) \ln \frac{p(M_i|y)}{\pi(M_i)} \quad \text{at} \quad \zeta = 1\,,$$

Kullback-Leibler (KL) divergence between the prior and posterior

For binary keys

$$Y[\zeta, K] = \sum_{\text{keyspace}} [p(y|\kappa)\pi(\kappa)]^{\zeta} \exp(\zeta K_{\alpha} \kappa^{\alpha}),$$

expectation value of individual key bits

$$\frac{\partial}{\partial K_{\alpha}} \frac{\ln Y[\zeta, K]}{\zeta} = \left. \frac{\sum_{\text{keyspace}} p(\kappa^{\alpha}|y)\kappa^{\alpha}}{\sum_{\text{keyspace}} p(\kappa|y)} \right|_{K=0,\zeta=1} = \langle \kappa |^{\alpha} | \rangle ,.$$