

Bayesian Model Selection:

Selecting the “best” model in a systematic way

SA AIMS

18th Feb, 2026

Lloyd Fung

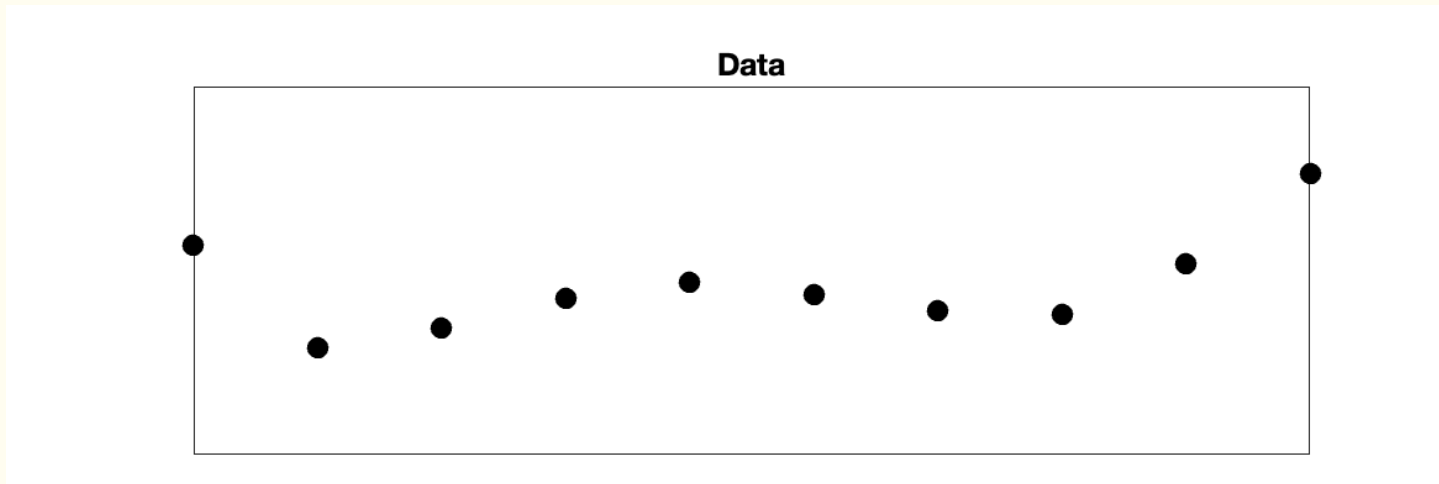
Research Fellow

Imperial College London

Learning Outcome

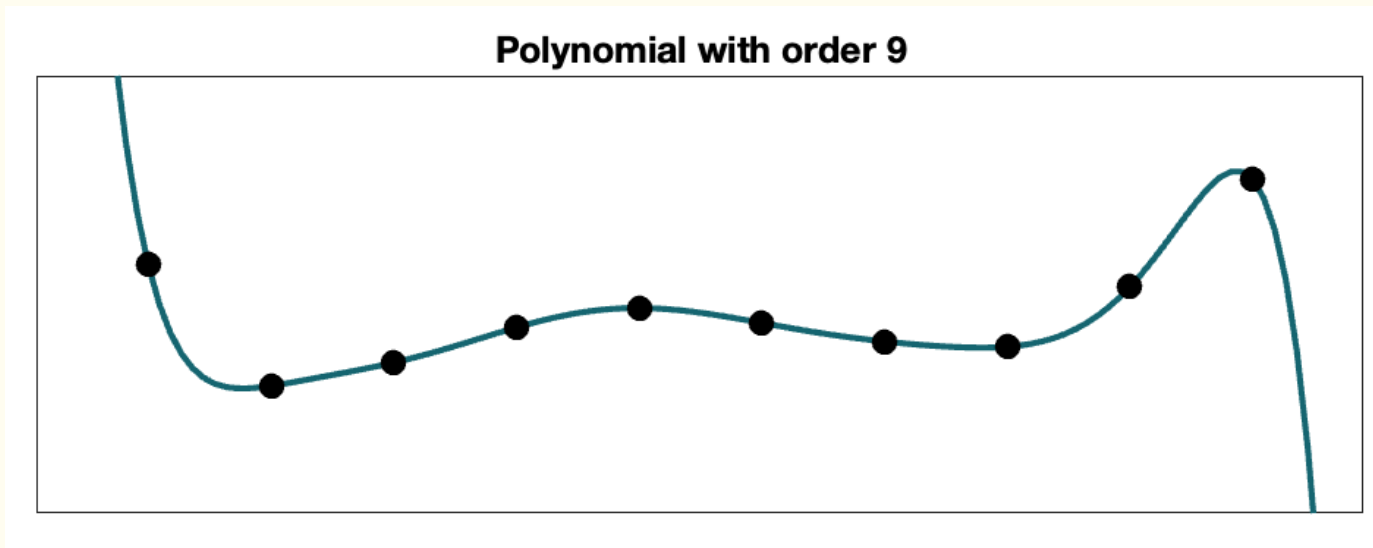
- Understand the principle of Occam razor and its Bayesian interpretation
- Gain an intuition to the scaling of prior and likelihood with data and noise
- Able to interpret Information Criteria from the Bayesian evidence perspective
- (Understand the mechanism of sparsification from Bayes' point of view)
- (Applying sparsifying prior to promote sparsity in model creation)
- (Understand and able to apply the `SparseBayes` algorithm)

What line would you put through these datapoints?



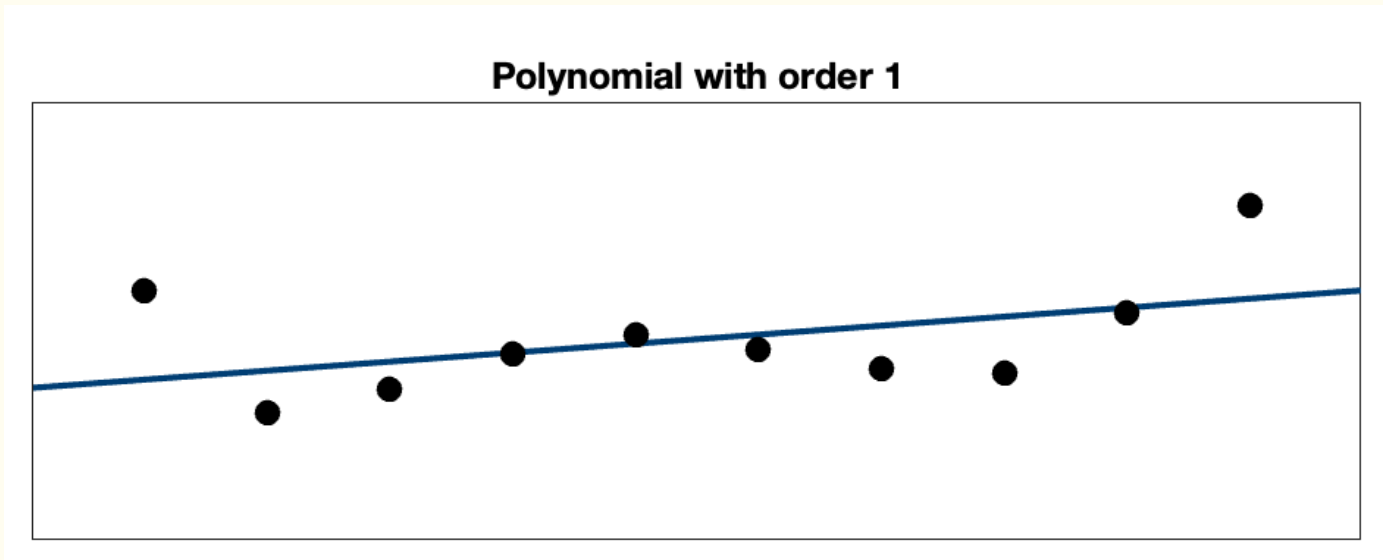
Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

What line would you put through these datapoints?

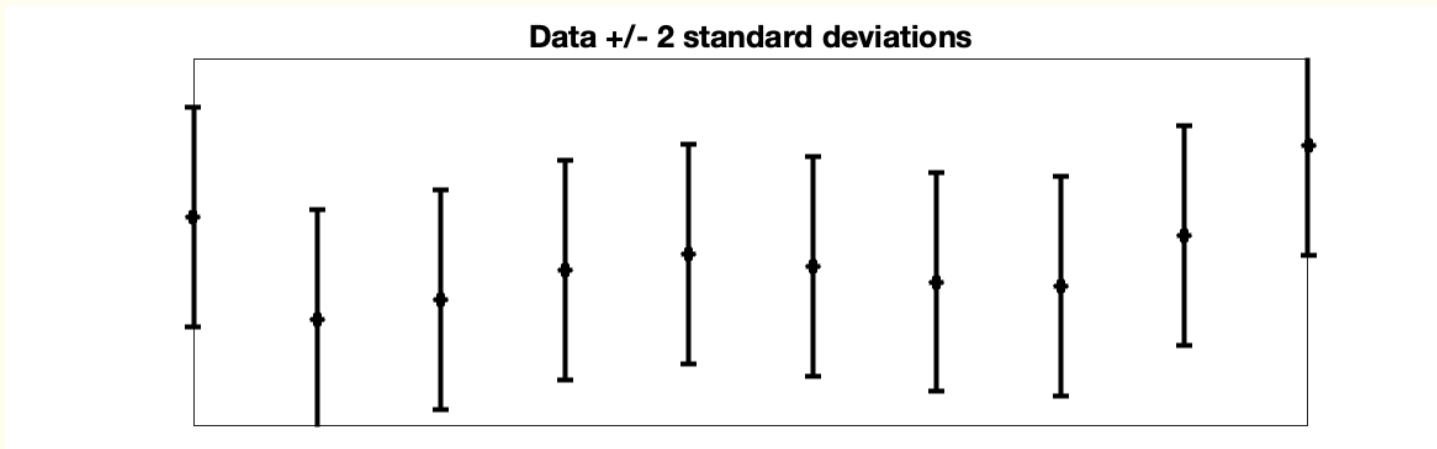


Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

What line would you put through these datapoints?



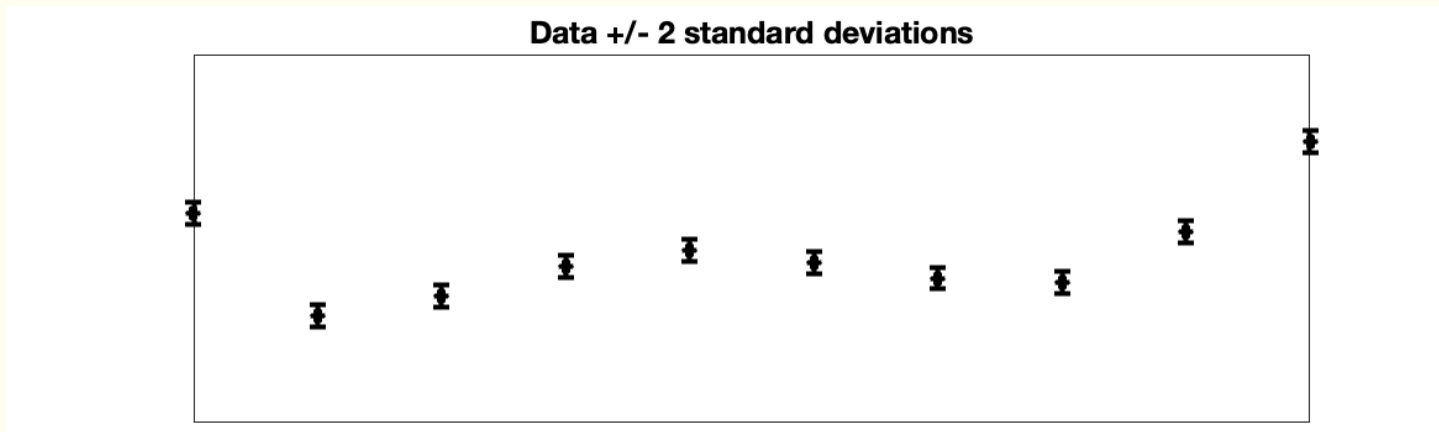
What line would you put through these datapoints?



Juniper (2022)

Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

What line would you put through these datapoints?



Juniper (2022)

Taken from Prof. Matthew Juniper's lecture on Bayesian Data Assimilation

Occam Razor



“When faced with two (or more) possible explanations, the simpler one is the one most likely to be true.”

(Paraphrasing) **William of Ockham**

We want as simple a model as possible (that still fits the data).

Principle of parsimony in modelling



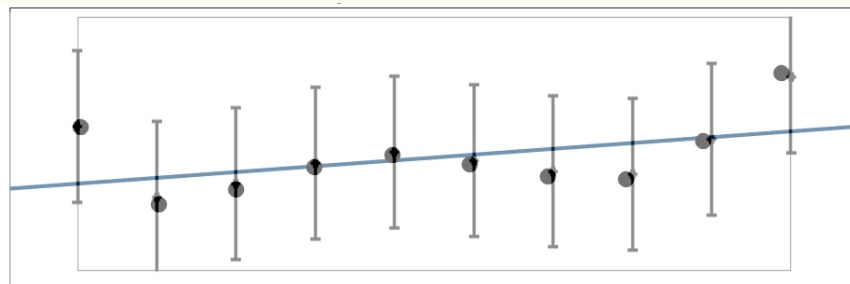
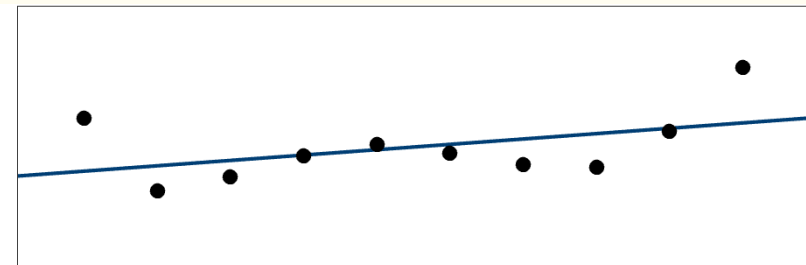
“Truth is much too complicated to allow anything but approximations.”

John von Neumann

We want to balance
model complexity against
accuracy.

What we want is... parsimonious model !

- **Ockham:** We want as simple a model as possible
- **von Neumann:** We want to balance **model complexity** against **accuracy**



How does Bayes help us?

Bayesian perspective
in the pursue of parsimonious models

Bayes' Theorem

$$P(\xi|D) = \frac{\overset{\text{likelihood}}{P(D|\xi)} \overset{\text{prior}}{P(\xi)}}{\underset{\text{evidence}}{P(D)}}$$

posterior

Just a normalising constant?

$D = \{x, y\}$ Data

ξ Fitting parameters

Bayes' Theorem

$$P(\underbrace{\xi_{\mathcal{M}_i}}_{\text{posterior}} | D, \mathcal{M}_i) = \frac{\overbrace{P(D | \xi_{\mathcal{M}_i}, \mathcal{M}_i)}^{\text{likelihood}} \overbrace{P(\xi_{\mathcal{M}_i} | \mathcal{M}_i)}^{\text{prior}}}{\underbrace{P(D | \mathcal{M}_i)}_{\text{evidence}}}$$

Key 'score' to compare btw models !

Model selection under Bayes

$$\operatorname{argmax}_{\mathcal{M}_i} P(\mathcal{M}_i | D)$$

where

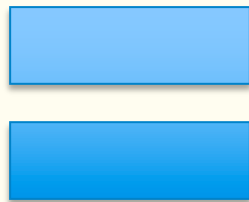
$$P(\mathcal{M}_i | D) \propto P(D | \mathcal{M}_i) P(\mathcal{M}_i)$$

$D = \{\mathbf{x}, \mathbf{y}\}$ Data

\mathcal{M}_i Hypothesis (model basis)

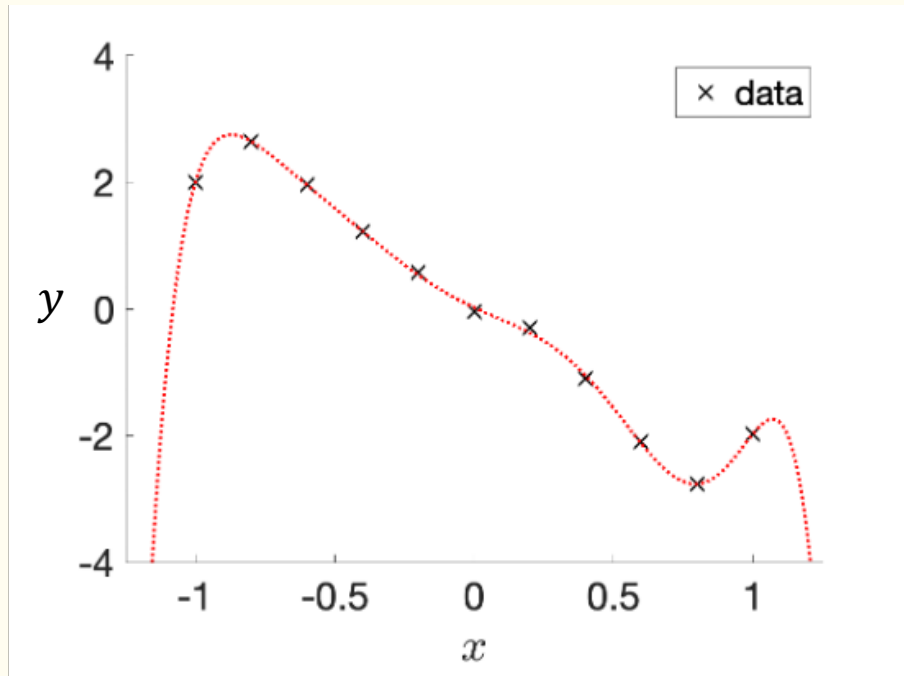
$\xi_{\mathcal{M}_i}$ Fitting parameters

Model with max. $P(D|\mathcal{M}_i)$



Most parsimonious model

An example with linear regression – Model 1 (Complex)

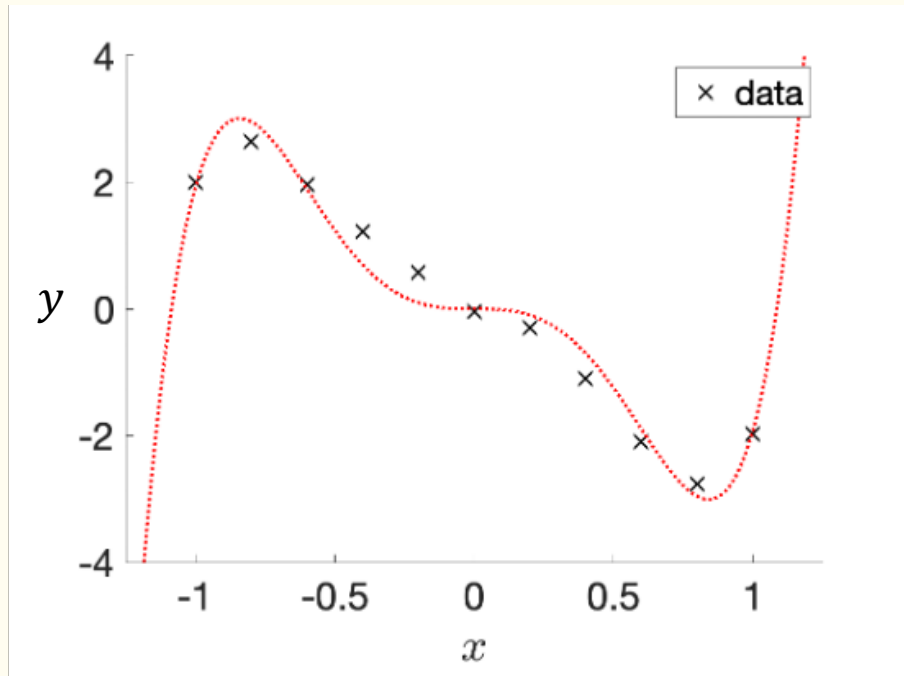


$$\begin{array}{lcl} \text{target} & = & \text{Polynomial Functions} \cdot \text{Learnt Coefficients} \\ \mathbf{y} & = & \Theta(\mathbf{x}) \cdot \boldsymbol{\xi} \end{array}$$

$$\begin{bmatrix} \text{red bar} \\ \text{red bar} \end{bmatrix} = \begin{bmatrix} \text{red bar} & \text{red bar} & \text{red bar} & \text{red bar} & \text{red bar} & \text{red bar} & \text{red bar} & \text{red bar} & \text{red bar} & \dots \end{bmatrix} \begin{bmatrix} \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \\ \text{red dot} \end{bmatrix}$$

$\mathbf{y} = \begin{bmatrix} 1 & x & x^2 & x^3 & x^4 & x^5 & x^6 & x^7 & x^8 & \dots \end{bmatrix} \boldsymbol{\xi}$

An example with linear regression – Model 2 (parsimonious)



$$\begin{array}{rcl}
 \text{target} & = & \text{Polynomial Functions} \cdot \text{Learnt Coefficients} \\
 \mathbf{y} & = & \Theta(x) \cdot \xi \\
 \begin{bmatrix} \text{red bar} \end{bmatrix} & = & \begin{bmatrix} \text{gray bar} & \text{gray bar} & \text{gray bar} & \text{red bar} & \text{gray bar} & \text{red bar} & \text{gray bar} & \text{gray bar} & \text{gray bar} & \dots \end{bmatrix} \begin{bmatrix} \text{gray bar} \\ \text{red dot} \\ \text{red dot} \\ \text{gray bar} \end{bmatrix} \\
 \mathbf{y} & & \begin{matrix} 1 \quad x \quad x^2 \quad x^3 \quad x^4 \quad x^5 \quad x^6 \quad x^7 \quad x^8 \end{matrix} \quad \xi
 \end{array}$$

Why is max. evidence = Occam Razor?

An example with linear regression

Bayes' rule

$$P(\xi|D, \mathcal{M}_i) = \frac{\overset{\text{likelihood}}{P(D|\xi, \mathcal{M}_i)} \overset{\text{prior}}{P(\xi|\mathcal{M}_i)}}{\underset{\text{evidence}}{P(D|\mathcal{M}_i)}}$$

posterior

$D = \{\mathbf{x}, \mathbf{y}\}$	Data
\mathcal{M}_i	Model
$\Theta_{\mathcal{M}}(\mathbf{x})$	Model Basis
ξ	Fitting parameters

$$-\ln P(\xi|D, \mathcal{M}_i) = \underbrace{\frac{1}{2\sigma^2} \|\mathbf{y} - \Theta_{\mathcal{M}}(\mathbf{x}) \cdot \xi\|_F^2}_{\text{likelihood}} + \underbrace{\frac{1}{2\sigma_p^2} \|\xi\|_2^2}_{\text{Prior (L}_2 \text{ regularisation)}} + \text{normalise}$$

posterior

error

Gaussian Linear model:
 $\mathbf{y} \sim \mathcal{N}(\Theta_{\mathcal{M}}(\mathbf{x}) \cdot \xi, \sigma^2 \mathbf{I})$

Maximum a posteriori (MAP):

$$\xi_{MAP} = \underset{\xi}{\operatorname{argmax}} P(\xi|D, \mathcal{M}_i)$$



Regularised Regression
 (Regularisation \leftrightarrow prior)

Why is max. evidence = Occam Razor?

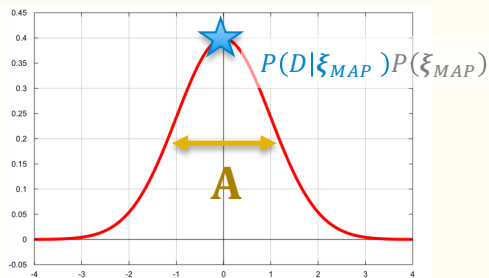
$$\mathbf{A} = \sigma_p^{-2} \mathbf{I} + \sigma^{-2} \mathbf{\Theta}_{\mathcal{M}}^T \mathbf{\Theta}_{\mathcal{M}}$$

An example with linear regression

Assuming the prior to all models $P(\mathcal{M}_i)$ is uniform:

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} P(D|\xi, \mathcal{M}_i) P(\xi|\mathcal{M}_i) d\xi \quad \leftarrow \text{Difficult to eval.}$$

Posterior:



$D = \{x, y\}$ Data

\mathcal{M}_i Model

ξ Fitting parameters

\mathbf{A} Parameter Covariance

Why is max. evidence = Occam Razor?

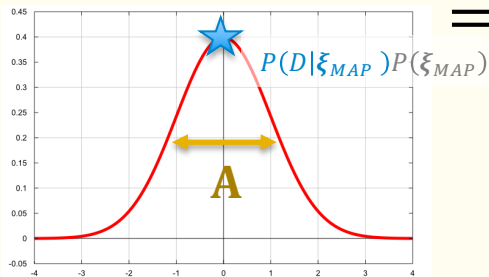
$$\mathbf{A} = \sigma_p^{-2} \mathbf{I} + \sigma^{-2} \mathbf{\Theta}_{\mathcal{M}}^T \mathbf{\Theta}_{\mathcal{M}}$$

An example with linear regression

Assuming the prior to all models $P(\mathcal{M}_i)$ is uniform:

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} P(D|\xi, \mathcal{M}_i) P(\xi|\mathcal{M}_i) d\xi \quad \text{Occam Razor!}$$

Posterior:



$$= P(D|\xi_{MAP}, \mathcal{M}_i)$$

Likelihood @ MAP

$$P(\xi_{MAP}|\mathcal{M}_i) \left[\det \left(\frac{\mathbf{A}}{2\pi} \right) \right]^{-1/2}$$

Prior on ξ @ MAP

Uncertainty in ξ around MAP

$\uparrow \# \xi \Rightarrow$
 $\uparrow \text{likelihood}$

$\uparrow \# \xi \Rightarrow$
 $\downarrow \text{prior}$

$\uparrow \# \xi \Rightarrow$
 $\uparrow \text{rank}(\mathbf{A}) \Rightarrow$
 $\downarrow |\mathbf{A}|^{-1/2}$

$D = \{\mathbf{x}, \mathbf{y}\}$ Data

\mathcal{M}_i Model

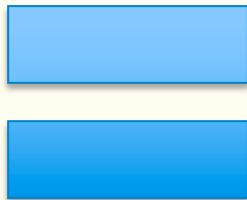
ξ Fitting parameters

\mathbf{A} Parameter Covariance

How well \mathcal{M}_i fit

Penalise models with too many param.

Model with max. $P(D|\mathcal{M}_i)$



Indeed!

Most parsimonious model

Computing Evidence – Gaussian case

Aim: $\operatorname{argmax}_{\mathcal{M}_i} P(\mathcal{M}_i|D)$

$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} P(D|\xi, \mathcal{M}_i) P(\xi|\mathcal{M}_i) d\xi$$

$$= P(D|\xi_{MAP}, \mathcal{M}_i) \quad P(\xi_{MAP}|\mathcal{M}_i) \quad \left[\det\left(\frac{\mathbf{A}}{2\pi}\right) \right]^{-1/2}$$

Likelihood @ MAP

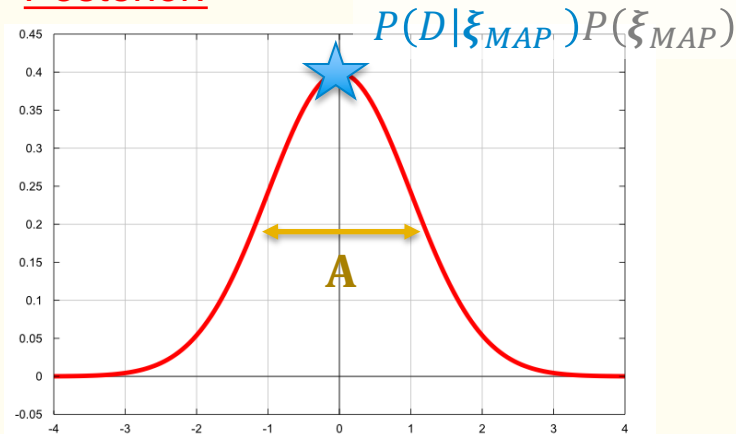
Prior on ξ @ MAP

Uncertainty in ξ around MAP

Obtained from
linear regression

Occam Factor

Posterior:



Penalizes models
with too many
param.

$D = \{x, y\}$ Data

\mathcal{M}_i Model

ξ Fitting parameters

\mathbf{A} Parameter Covariance

$$\mathbf{A} = \sigma_p^{-2} \mathbf{I} + \sigma^{-2} \mathbf{\Theta}_{\mathcal{M}}^T \mathbf{\Theta}_{\mathcal{M}}$$

Approximating Evidence – Laplace Method

Aim: $\operatorname{argmax}_{\mathcal{M}_i} P(\mathcal{M}_i|D)$

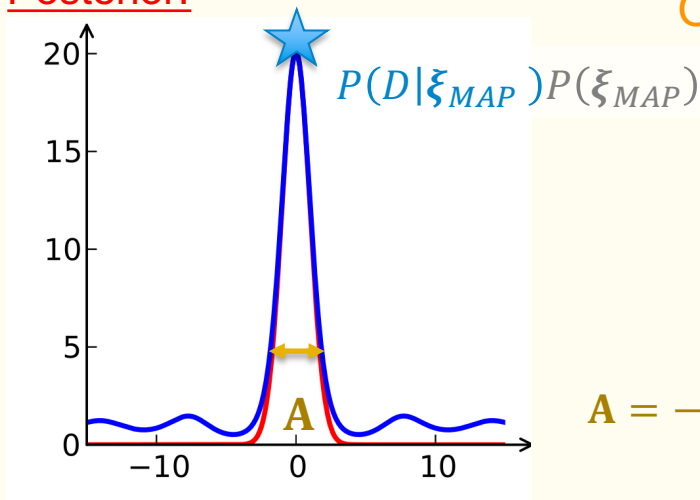
$$P(\mathcal{M}_i|D) \propto P(D|\mathcal{M}_i) = \int_{\xi} \underbrace{P(D|\xi, \mathcal{M}_i)P(\xi|\mathcal{M}_i)}_{\propto P(\xi|D)} d\xi$$

Difficult to eval.

$$\approx \underbrace{P(D|\xi_{MAP}, \mathcal{M}_i)}_{\text{Likelihood @ MAP}} \underbrace{P(\xi_{MAP}|\mathcal{M}_i)}_{\text{Prior on } \xi \text{ @ MAP}} \underbrace{\left[\det \left(\frac{\mathbf{A}}{2\pi} \right) \right]^{-1/2}}_{\text{Uncertainty in } \xi \text{ around MAP}}$$

Obtained from
variational inference

Posterior:



Occam Factor

Penalizes models
with too many
param.

$D = \{x, y\}$ Data

\mathcal{M}_i Model

ξ Fitting parameters

\mathbf{A} Parameter Uncertainty

$$\mathbf{A} = - \left. \frac{d^2 \ln P(D|\xi, \mathcal{M}_i)P(\xi|\mathcal{M}_i)}{d\xi^2} \right|_{\xi_{MAP}}$$

Scaling of Likelihood and Occam Factor and BIC

\mathbf{A} is a $k \times k$ matrix
scales with n

$$\ln P(D|\mathcal{M}_i) \approx \underbrace{\ln P(D|\xi_{MAP}, \mathcal{M}_i)}_{\sim O(n)} + \underbrace{\ln P(\xi_{MAP}|\mathcal{M}_i)}_{\sim O(k)} + \underbrace{\ln \left[\det \left(\frac{\mathbf{A}}{2\pi} \right) \right]^{-1/2}}_{\sim O(-\frac{k}{2} \ln \frac{n}{2\pi})}$$

Likelihood @ MAP Occam Factor

When $n \rightarrow \infty \gg k$,

$$-2 \ln P(D|\mathcal{M}_i) \approx -2 \ln \hat{L} + k \ln n + O(1)$$

$$BIC = k \ln n - 2 \ln \hat{L}$$

\therefore Minimise BIC \approx Maximise Evidence

n	Number of data
k	Number of parameters
D	Data
\mathcal{M}_i	Model
ξ	Fitting parameters
\mathbf{A}	Parameter Covariance

Model Sparsification through priors

Selecting a model in a set of models defined by combinations of terms

The Role of Prior in Regression

Bayes' rule

$$P(\xi|D, \mathcal{M}_i) = \frac{\overset{\text{likelihood}}{P(D|\xi, \mathcal{M}_i)} \overset{\text{prior}}{P(\xi|\mathcal{M}_i)}}{\underset{\text{evidence}}{P(D|\mathcal{M}_i)}}$$

posterior

$D = \{\mathbf{x}, \mathbf{y}\}$ Data

\mathcal{M}_i Model

$\Theta_{\mathcal{M}}(\mathbf{x})$ Model Basis

ξ Fitting parameters

Gaussian Linear model:
 $\mathbf{y} \sim \mathcal{N}(\Theta_{\mathcal{M}}(\mathbf{x}) \cdot \xi, \sigma^2 \mathbf{I})$

$$-\ln P(\xi|D, \mathcal{M}_i) = \underbrace{\frac{1}{2\sigma^2} \|\mathbf{y} - \Theta_{\mathcal{M}}(\mathbf{x}) \cdot \xi\|_F^2}_{\text{likelihood}} + \underbrace{\gamma \|\xi\|}_{\text{Prior (Regularisation)}} + \text{normalise}$$

posterior

Maximum a posteriori (MAP):

$$\xi_{MAP} = \underset{\xi}{\operatorname{argmax}} P(\xi|D, \mathcal{M}_i)$$

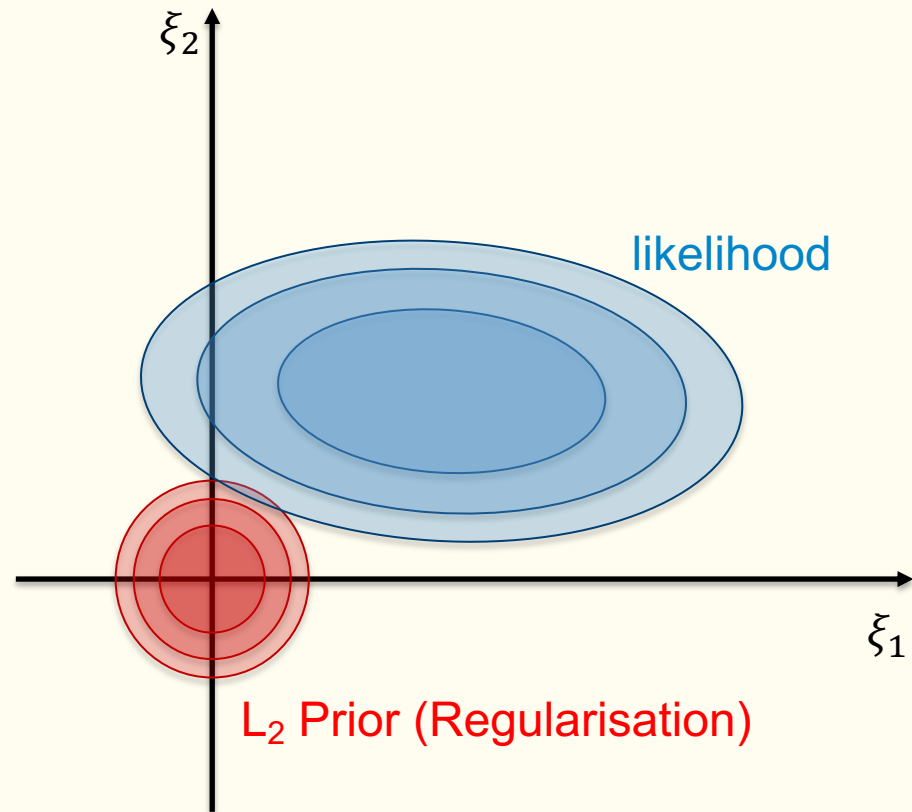


Regularised Regression
 (Regularisation \leftrightarrow prior)

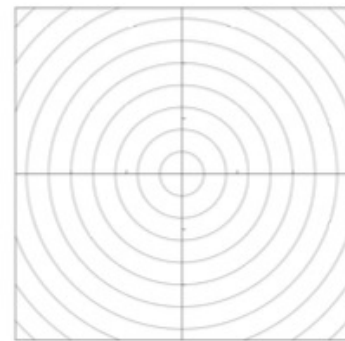
Types of prior

Regression	Loss	Regulariser	Prior
Ridge	$\frac{1}{2\sigma^2} \ \mathbf{y} - \mathbf{\Theta}_{\mathcal{M}}(\mathbf{x}) \cdot \boldsymbol{\xi}\ _F^2 + \frac{\gamma}{2} \ \boldsymbol{\xi}\ _2^2$	$\frac{1}{2} \ \boldsymbol{\xi}\ _2^2 = \frac{1}{2} \sum_i \xi_i^2$	Gaussian (L2) $\boldsymbol{\xi} \sim \mathcal{N}(0,1)$
LASSO	$\frac{1}{2\sigma^2} \ \mathbf{y} - \mathbf{\Theta}_{\mathcal{M}}(\mathbf{x}) \cdot \boldsymbol{\xi}\ _F^2 + \gamma \ \boldsymbol{\xi}\ _1$	$\ \boldsymbol{\xi}\ _1 = \sum_i \xi_i $	Laplace (L1) $\boldsymbol{\xi} \sim \text{Laplace}(0,1)$
Sequential Threshold	$\frac{1}{2\sigma^2} \ \mathbf{y} - \mathbf{\Theta}_{\mathcal{M}}(\mathbf{x}) \cdot \boldsymbol{\xi}\ _F^2 + \gamma \ \boldsymbol{\xi}\ _0$	$\ \boldsymbol{\xi}\ _0$ $= \# \text{ of nonzero}$	(No formal prior, L0)
Spike and Slab	(Ridge with $\ \boldsymbol{\xi}\ _2^2$ + loss for λ)	$\boldsymbol{\xi} \lambda \sim \mathcal{N}(0,1)\lambda + \mathcal{N}(0, \epsilon^2)(1 - \lambda)$ $\lambda \sim \text{Ber}(\pi)$	
Normal-Gamma	(Ridge with $\ \boldsymbol{\xi}\ _2^2$ + loss for α)	$\boldsymbol{\xi} \alpha \sim \mathcal{N}(0, \text{diag}(\alpha^{-1}))$ $\alpha \sim \text{Gamma}(k, \theta)$	

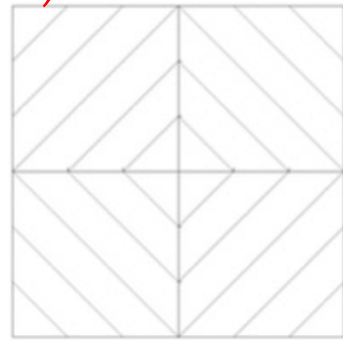
Prior as sparsifier



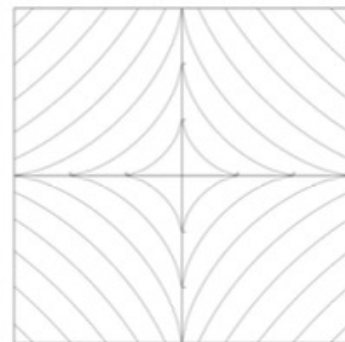
$$\|\xi\|_p = \left(\sum_i |\xi_i|^p \right)^{1/p}$$



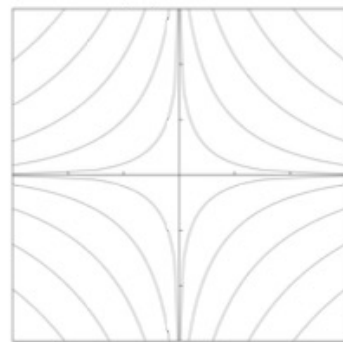
(a) $p = 2$



(b) $p = 1$

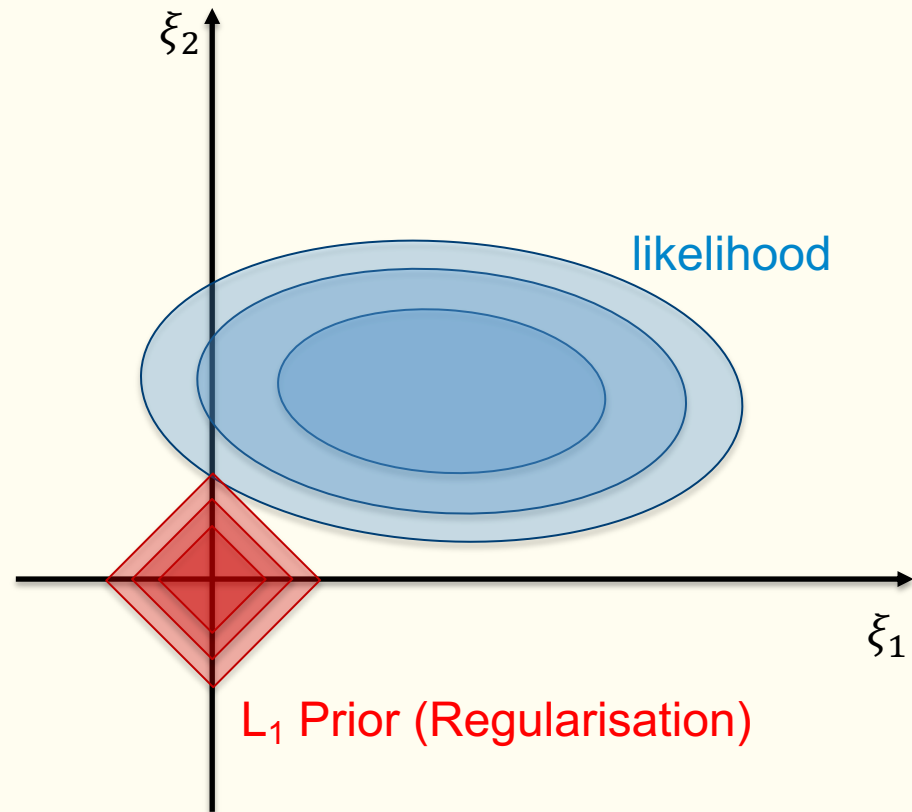


(c) $p = 2/3$

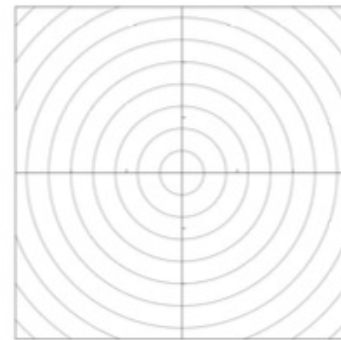


(d) $p = 1/3$

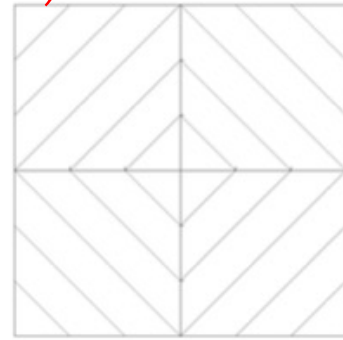
Prior as sparsifier



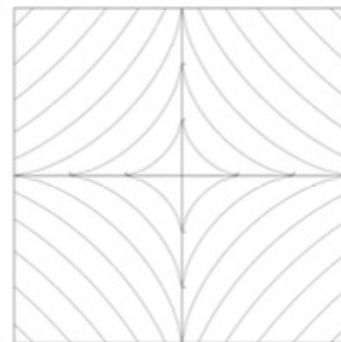
$$\|\xi\|_p = \left(\sum_i |\xi_i|^p \right)^{1/p}$$



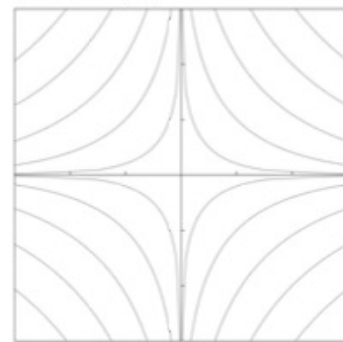
(a) $p = 2$



(b) $p = 1$

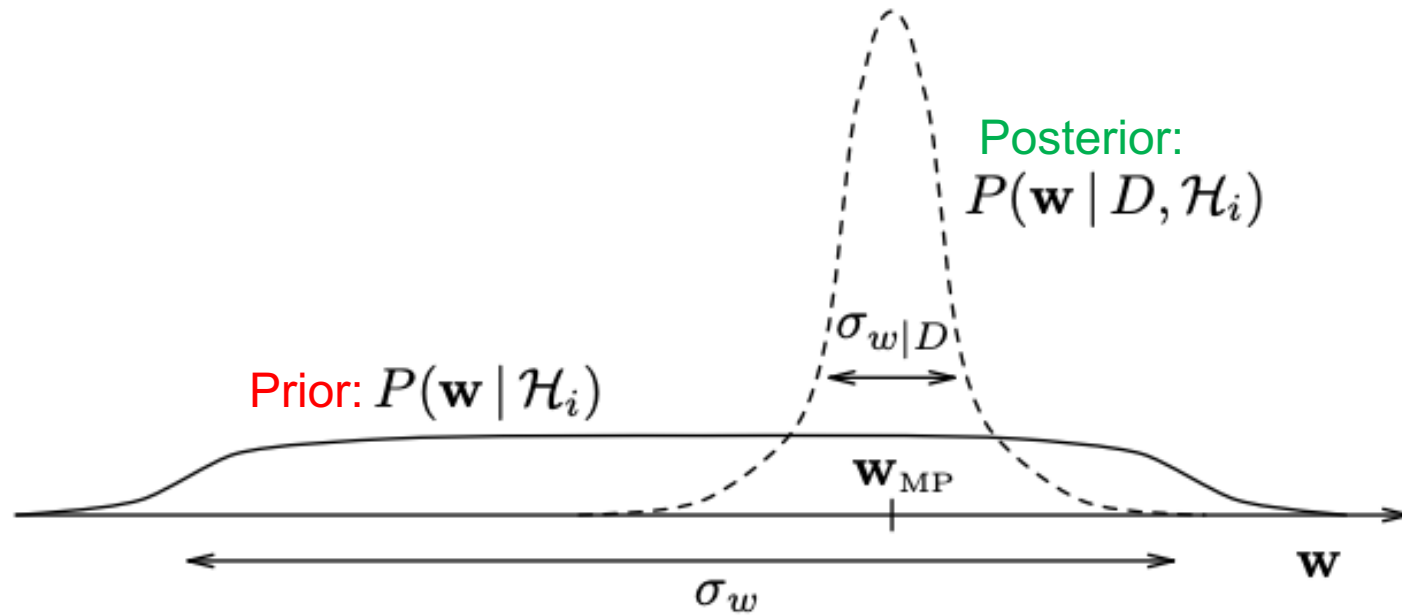


(c) $p = 2/3$



(d) $p = 1/3$

Prior as sparsifier – Hierarchical Prior

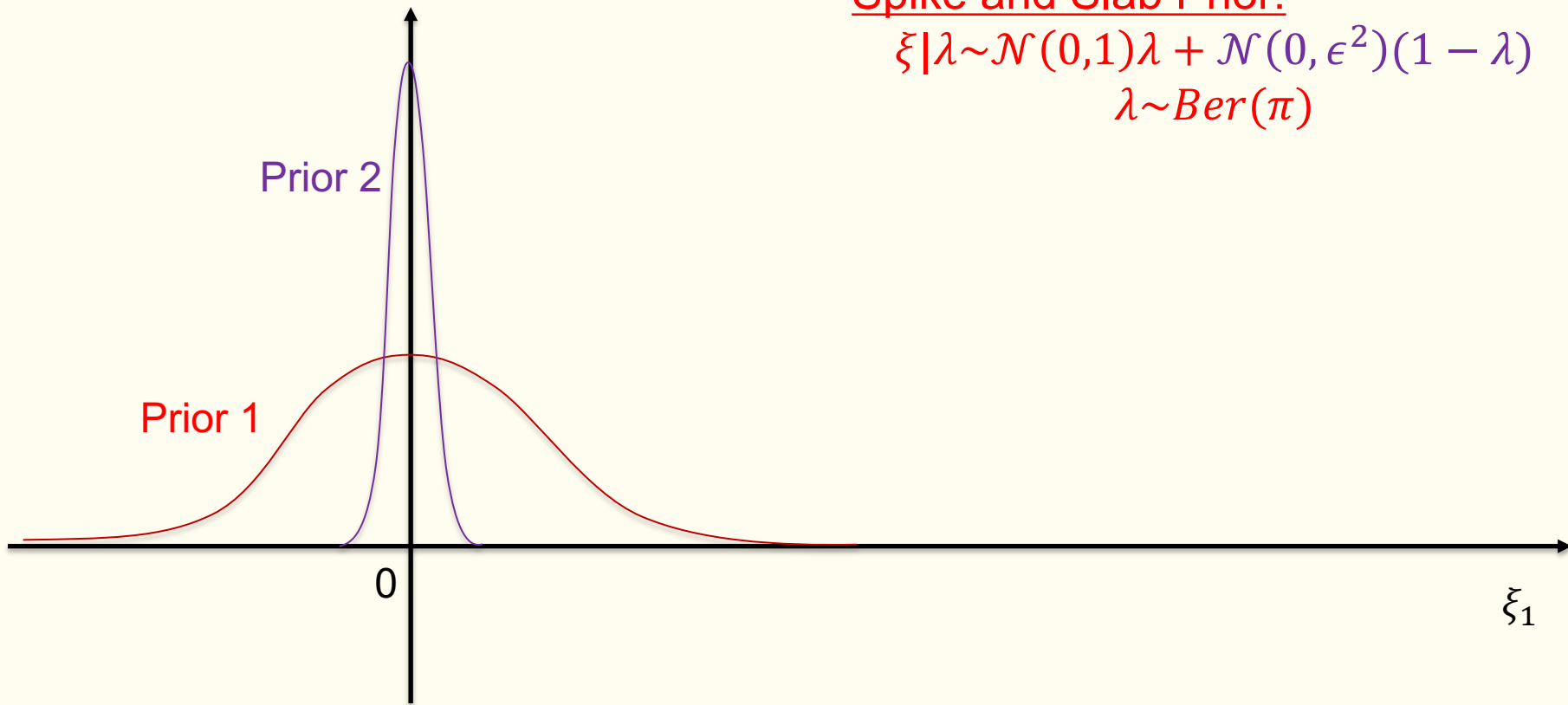


Prior as sparsifier – Hierarchical Prior

Spike and Slab Prior:

$$\xi | \lambda \sim \mathcal{N}(0, 1) \lambda + \mathcal{N}(0, \epsilon^2) (1 - \lambda)$$

$$\lambda \sim \text{Ber}(\pi)$$

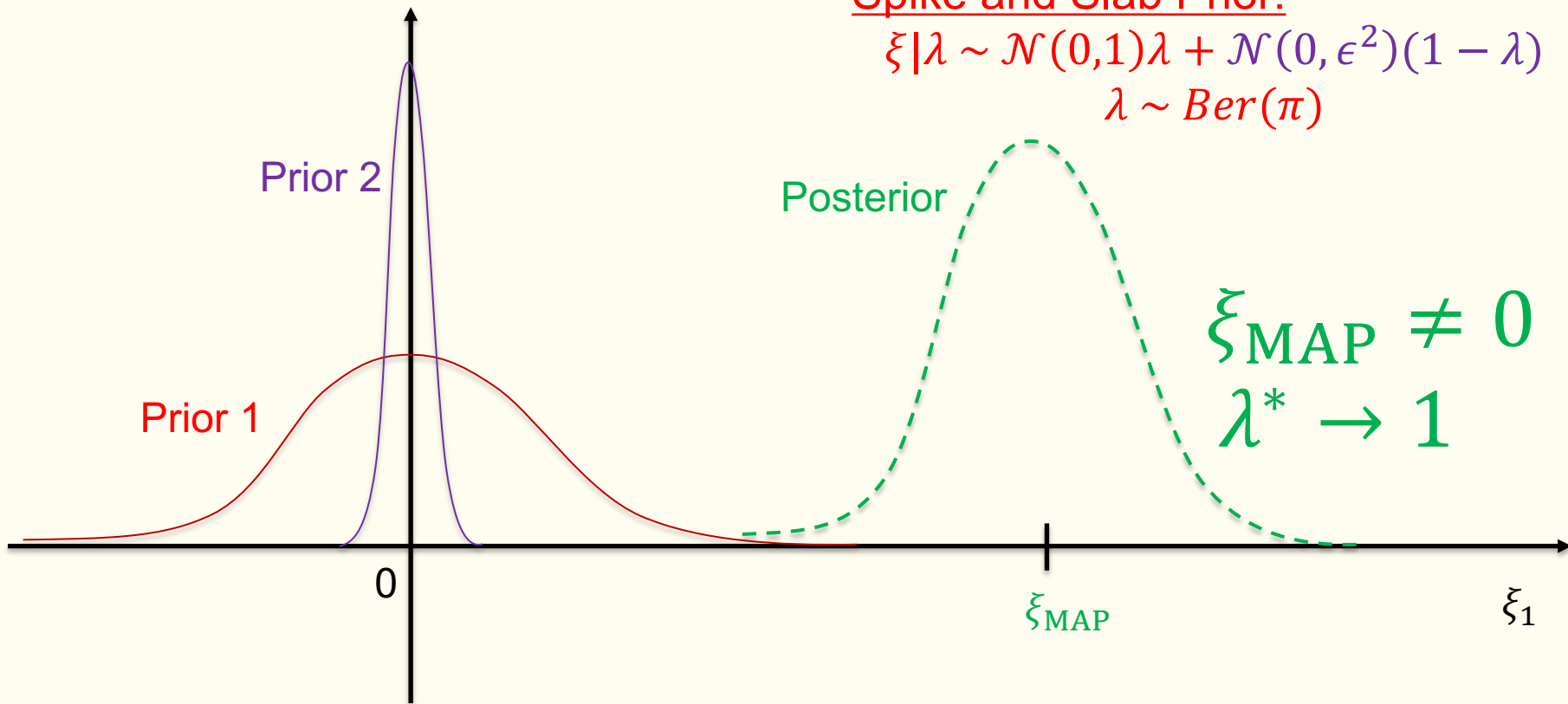


Prior as sparsifier – Hierarchical Prior

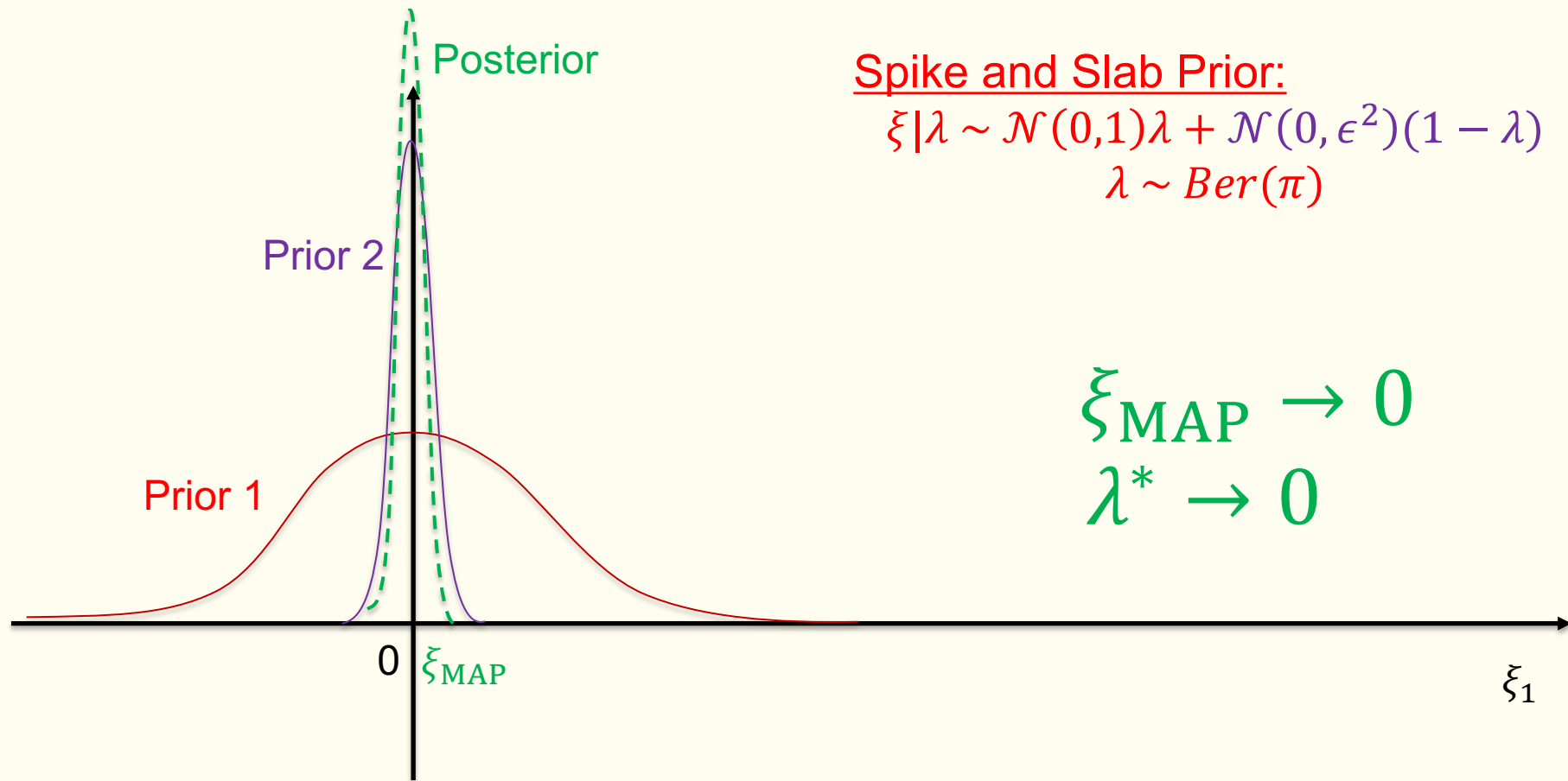
Spike and Slab Prior:

$$\xi | \lambda \sim \mathcal{N}(0, 1) \lambda + \mathcal{N}(0, \epsilon^2) (1 - \lambda)$$

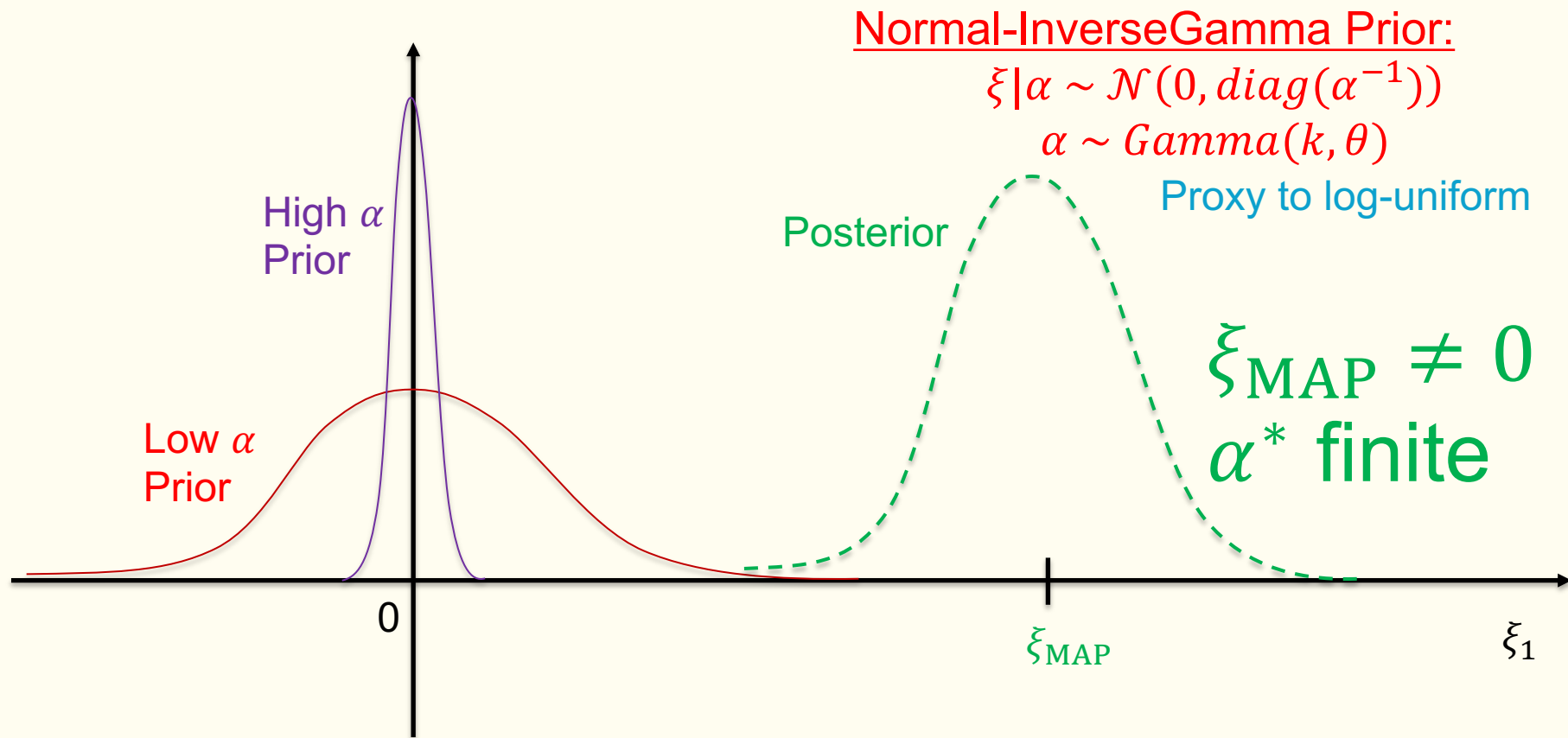
$$\lambda \sim \text{Ber}(\pi)$$



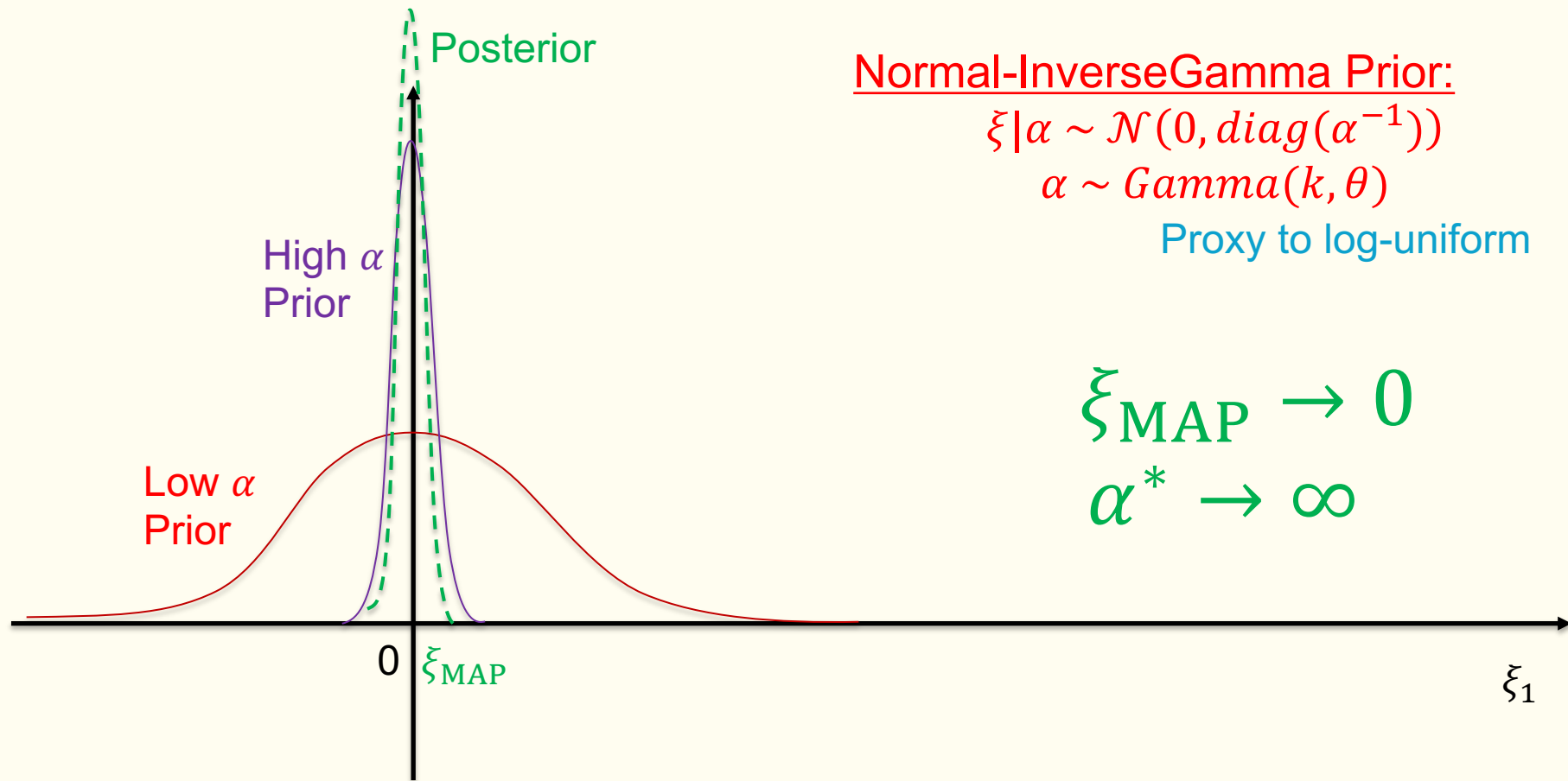
Prior as sparsifier – Hierarchical Prior



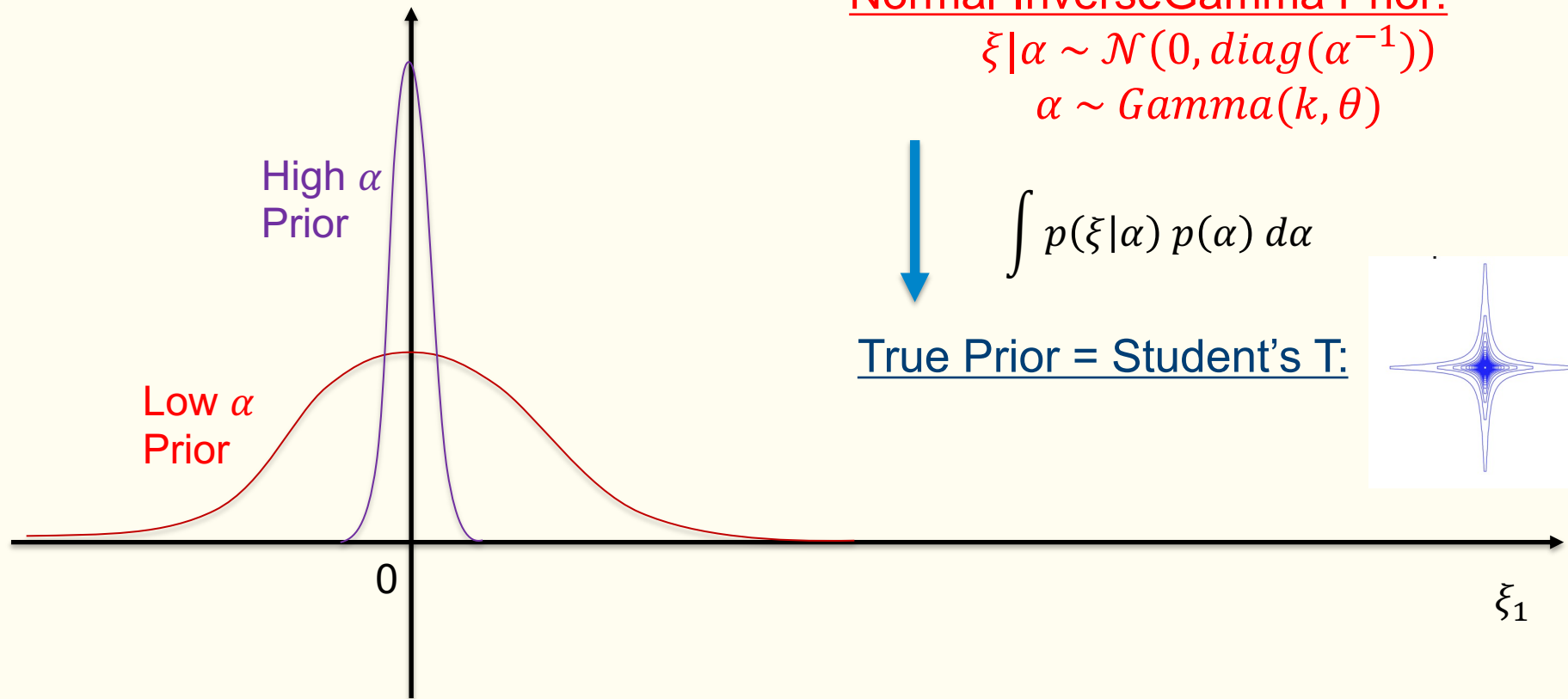
Prior as sparsifier – Hierarchical Prior



Prior as sparsifier – Hierarchical Prior



Prior as sparsifier – Hierarchical Prior



Final remark on sparsifying priors VS Bayesian evidence

- Selecting model by Bayesian evidence is to formally think about

$$P(\mathcal{M}_i|D) = P(D|\mathcal{M}_i)P(\mathcal{M}_i)$$

- FORMAL!
 - Fundamentally Bayesian, Fundamentally probabilistic
 - Can deal with heterogeneous datasets (parameters are marginalised out)
- Sparsifying (selecting) model by sparsifying prior
 - Rooted in “regression” typed problem
 - (not as) formal
 - More popular (because they’re “faster”?)

Tutorial: 1D Polynomial Regression

Data Generation

- Generate $\{x,y\}$ data pair from some randomly selected sum of polynomial terms
- Add Gaussian noise to y with known variance σ^2

Regression

- Perform linear regression (Frequentist: OLS; Bayesian: Ridge) on a polynomial library
- Find Evidence
- Remove terms in the library and redo regression; Find the combination of terms (model) that give max evidence

Experimentation

- How does changing σ^2 in the input of the regression changes the model selection?
 - How does that correspond to the intuition of Occam Razor?
- How can one further optimise the prior variance (a hyperparameter)?