Bayesian Machine Learning

Andrew Gordon Wilson

Occam's Razor and Model Construction

Optimization, Big Data and Applications (OBA)
Veroli Summer School
July 6, 2022

References

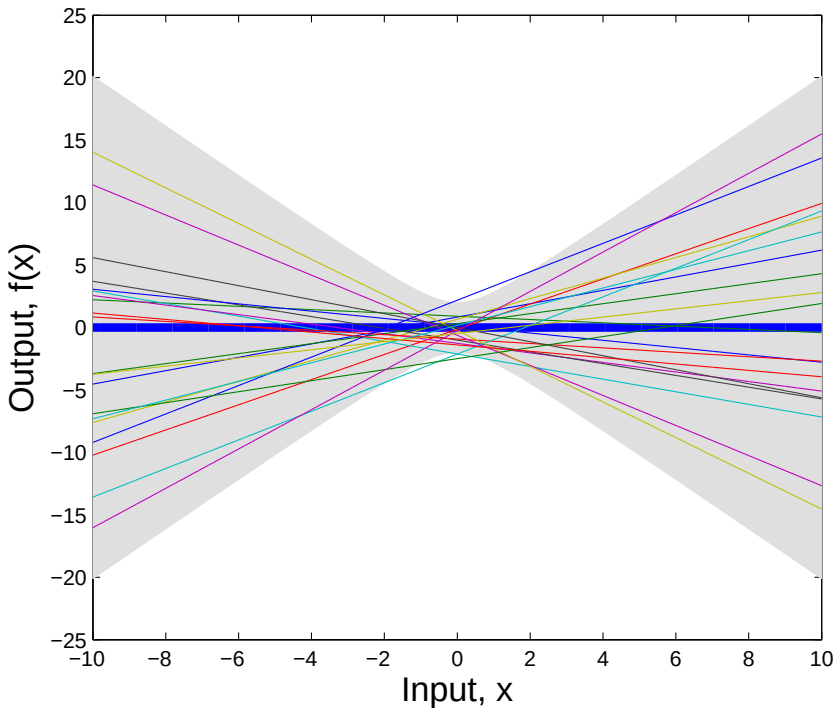
Minka (2000), Rasmussen and Ghahramani (2001), MacKay (2003), Bishop (2006), Ghahramani (2015), Ghahramani (2014), Wilson (2014).

A Function-Space View

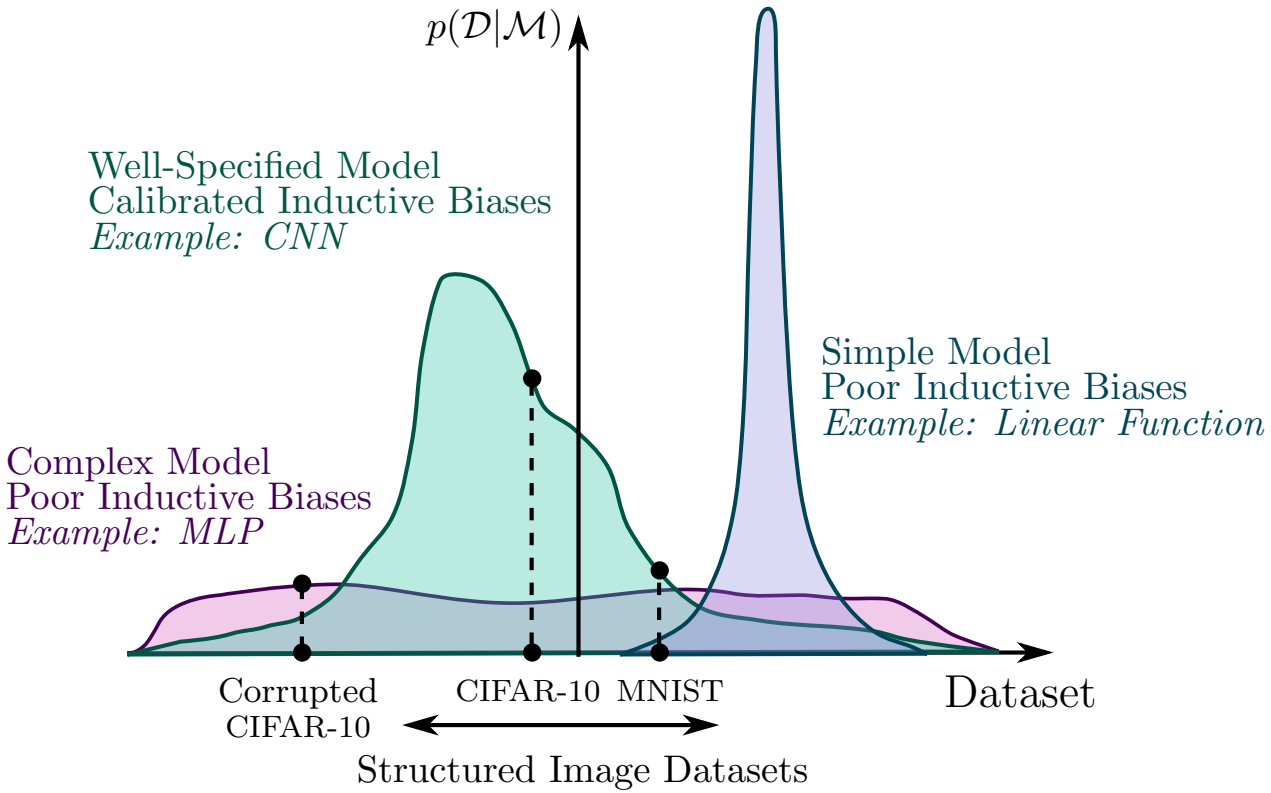
Consider the simple linear model,

$$f(x) = w_0 + w_1 x\,, \tag{1}$$

$$w_0, w_1 \sim \mathcal{N}(0, 1)\,. \tag{2}$$



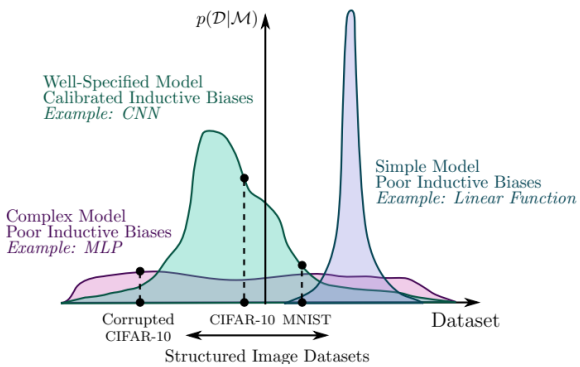
Model Construction and Generalization

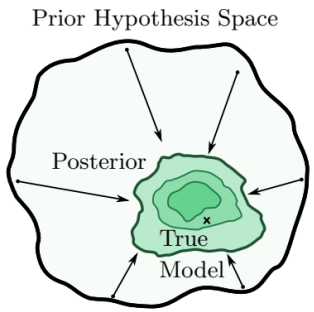


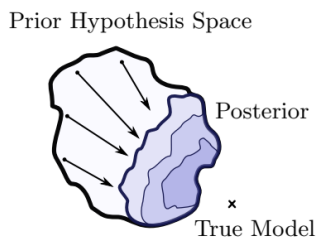
$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\mathcal{M}, w)p(w)dw \tag{3}$$

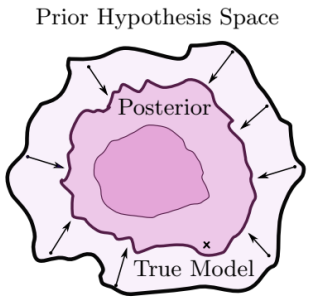
How do we learn?

- The ability for a system to learn is determined by its *support* (which solutions are a priori possible) and *inductive biases* (which solutions are a priori likely).
- We should not conflate *flexibility* and *complexity*.
- An influx of new *massive* datasets provide great opportunities to automatically learn rich statistical structure, leading to new scientific discoveries.





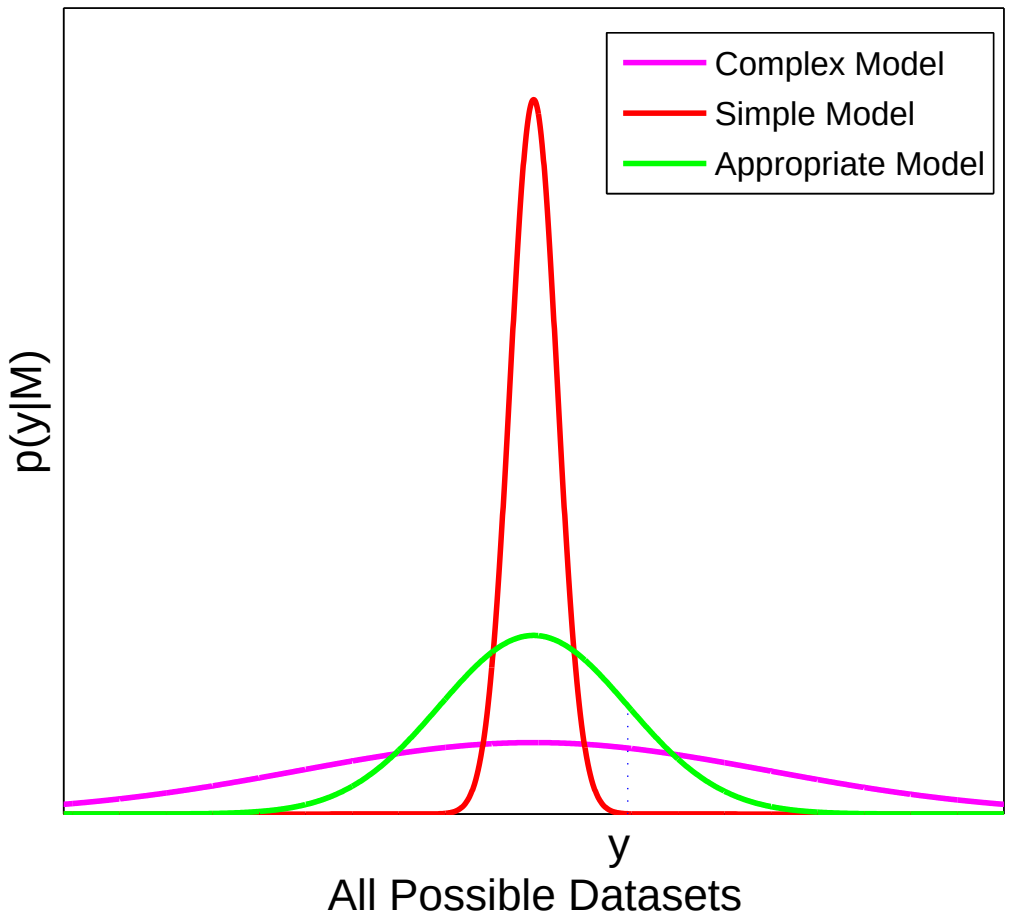




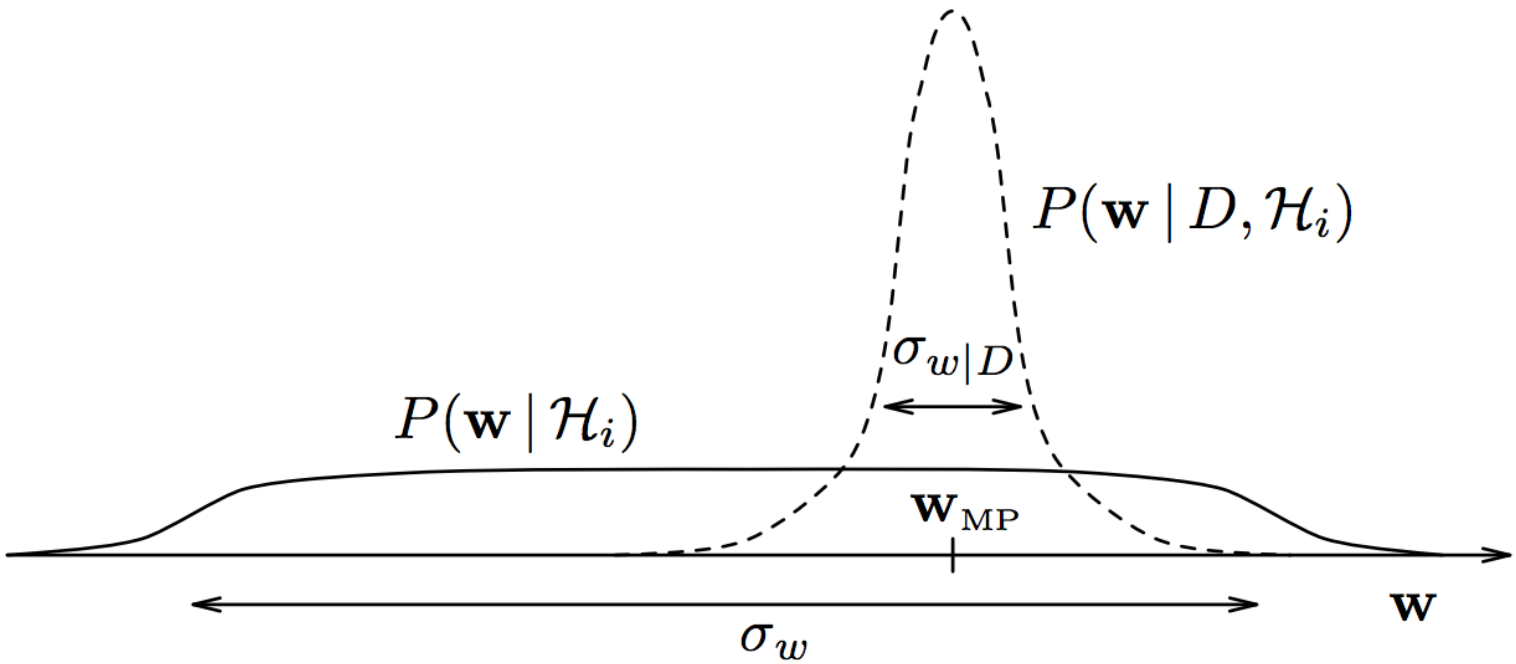
Bayesian Deep Learning and a Probabilistic Perspective of Generalization
Wilson and Izmailov, 2020
arXiv 2002.08791

Model Selection and Marginal Likelihood

$$p(\mathbf{y}|\mathcal{M}_1, X) = \int p(\mathbf{y}|f_1(x, \mathbf{w}))p(\mathbf{w})d\mathbf{w} \tag{4}$$



Evaluating the evidence

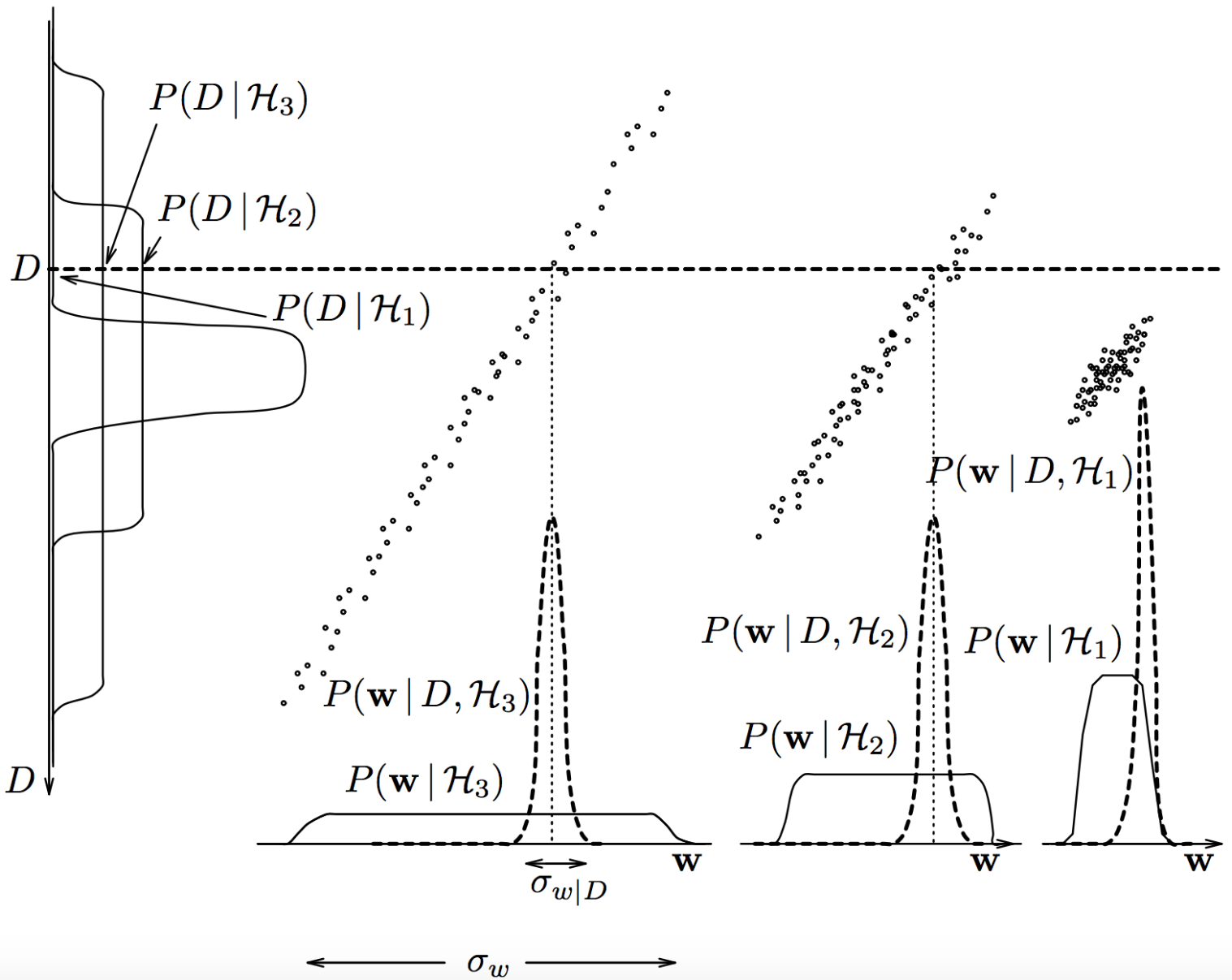


$$p(\mathcal{D}|\mathcal{H}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{H}_i) p(\mathbf{w}|\mathcal{H}_i) d\mathbf{w} \tag{5}$$

$$p(\mathcal{D}|\mathcal{H}_i) = \underbrace{p(\mathcal{D}|\mathbf{w}_{\mathrm{MP}})}_{\text{best fit likelihood}} \times \underbrace{p(\mathbf{w}_{\mathrm{MP}}|\mathcal{H}_i)\sigma_{w|\mathcal{D}}}_{\text{Occam factor}} \tag{6}$$

Occam factor = $\frac{\sigma_{w|\mathcal{D}}}{\sigma_w}$

Model Comparison



Occam's factor for Gaussian posteriors

If the posterior over $\mathbf{w}$ is well approximated by a Gaussian, then the Occam factor is described by the log determinant of the prior covariance matrix:

$$p(\mathcal{D}|\mathcal{H}_i) = \underbrace{p(\mathcal{D}|\mathbf{w}_{\mathrm{MP}})}_{\text{best fit likelihood}} \times \underbrace{p(\mathbf{w}_{\mathrm{MP}}|\mathcal{H}_i)\det(\frac{A}{2\pi})^{-\frac{1}{2}}}_{\text{Occam factor}} \tag{7}$$

where $A = -\nabla\nabla \log p(\mathbf{w}|\mathcal{D}, \mathcal{H}_i)$.

Model Comparison

$$\frac{p(\mathcal{H}_1|\mathcal{D})}{p(\mathcal{H}_2|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_2)} \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)} \,. \tag{8}$$

Blackboard: Examples of Occam's Razor in Everyday Inferences

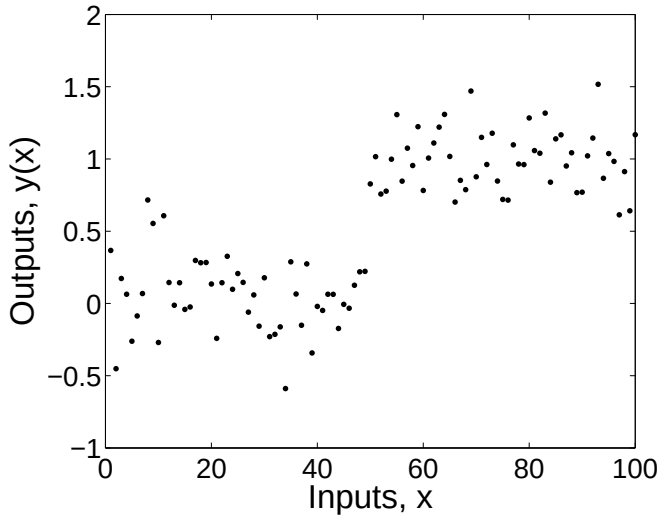
For further reading, see MacKay (2003) textbook, *Information Theory, Inference, and Learning Algorithms*.

Occam's Razor Example

-1, 3, 7, 11, ??, ??

- H_1: the sequence is an arithmetic progression, add n, where n is an integer.
- H_2: the sequence is generated by a cubic function of the form $cx^3 + dx^2 + e$, where c, d, and e are fractions. $(-\frac{1}{11}x^3 + \frac{9}{11}x^2 + \frac{23}{11})$

Model Selection



Observations $y(x)$. Assume $p(y(x)|f(x)) \sim \mathcal{N}(y(x); f(x), \sigma^2)$. Consider polynomials of different orders. As always, observations are out of the chosen model class! Which model should we choose?

$$f_0(x) = a_0\,, \tag{9}$$

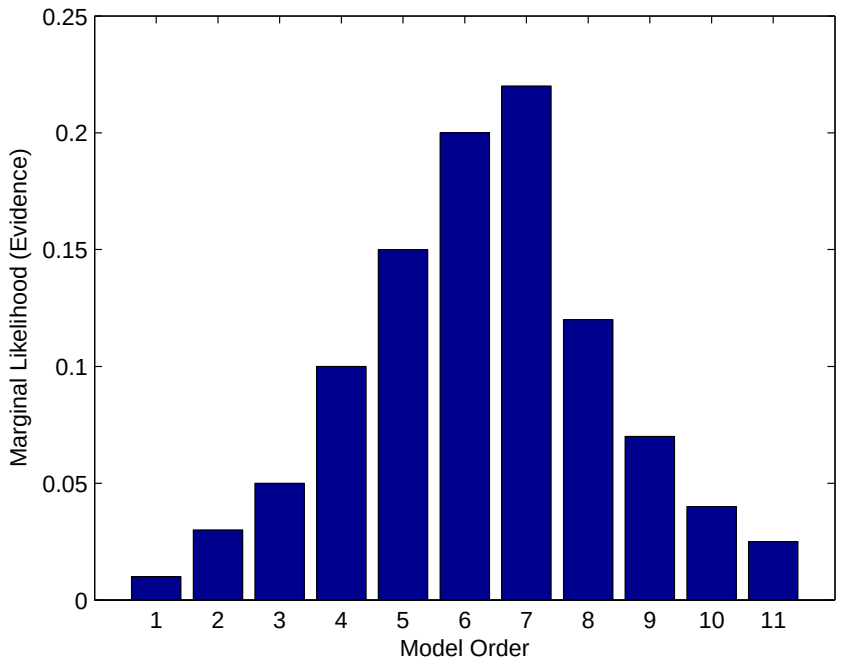
$$f_1(x) = a_0 + a_1 x\,, \tag{10}$$

$$f_2(x) = a_0 + a_1 x + a_2 x^2\,, \tag{11}$$

$$\vdots \tag{12}$$

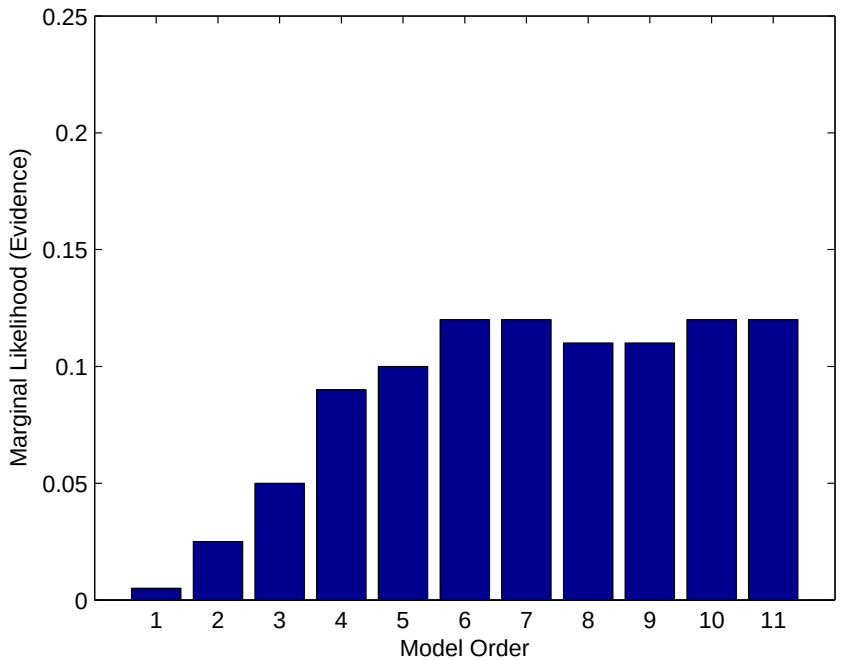
$$f_J(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_J x^J\,. \tag{13}$$

Model Selection: Occam's Hill



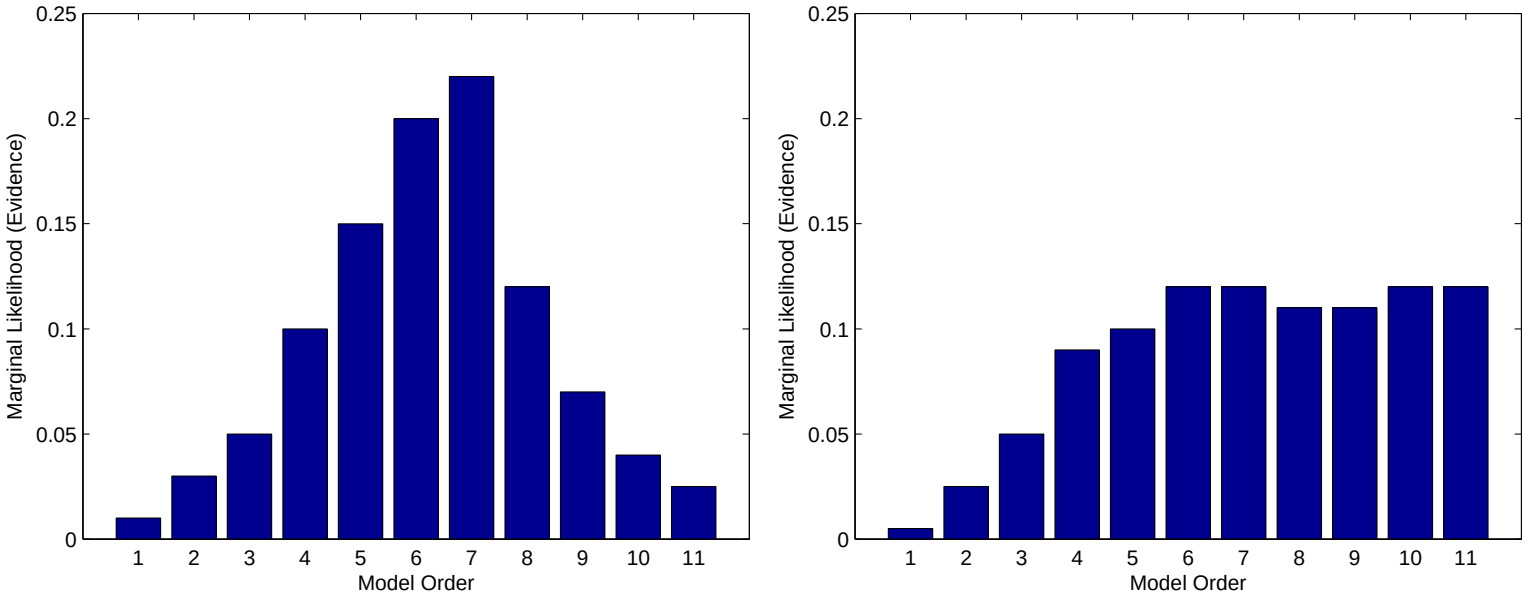
Marginal likelihood (evidence) as a function of model order, using an isotropic prior $p(a) = \mathcal{N}(0, \sigma^2 I)$.

Model Selection: Occam's Asymptote



Marginal likelihood (evidence) as a function of model order, using an anisotropic prior $p(a_i) = \mathcal{N}(0, \gamma^{-i})$, with γ learned from the data.

Occam's Razor



(a) Isotropic Gaussian Prior

(b) Anisotropic Gaussian Prior

For further reading, see Rasmussen and Ghahramani (2001) (*Occam's Razor*), Kass and Raftery (1995) (*Bayes Factors*), and MacKay (2003), Chapter 28.

Automatic Choice of Dimensionality for PCA

- PCA projects a d dimensional vector $\mathbf{x}$ into a $k \leq d$ dimensional space in a way that maximizes the variance of the projection.
- How do we choose k?

Probabilistic PCA

- Formulate dimensionality reduction as a probabilistic model:

$$\mathbf{x} = \sum_{j=1}^{k} \mathbf{h}_j w_j + \mathbf{m} + \boldsymbol{\epsilon}\,, \tag{14}$$

$$= H\mathbf{w} + \mathbf{m} + \boldsymbol{\epsilon}\,, \tag{15}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, V)\,. \tag{16}$$

- Let $V = vI_d$ and $p(\mathbf{w}) \sim \mathcal{N}(0, I_k)$.
- The maximum likelihood solution for H, given data $\mathcal{D} = \{\mathbf{x}_1, \ldots \mathbf{x}_N\}$ is exactly equal to the PCA solution!
- Let's place probability distributions over $H, \mathbf{m}$, integrate away from the likelihood, then use the *evidence* $p(\mathcal{D}|k)$ to determine the value of k. As $N \to \infty$, the evidence will collapse onto the true value of k.

Automatically Learning the Dimensionality of PCA (Minka, 2001).

Automatically Learning the Dimensionality of PCA

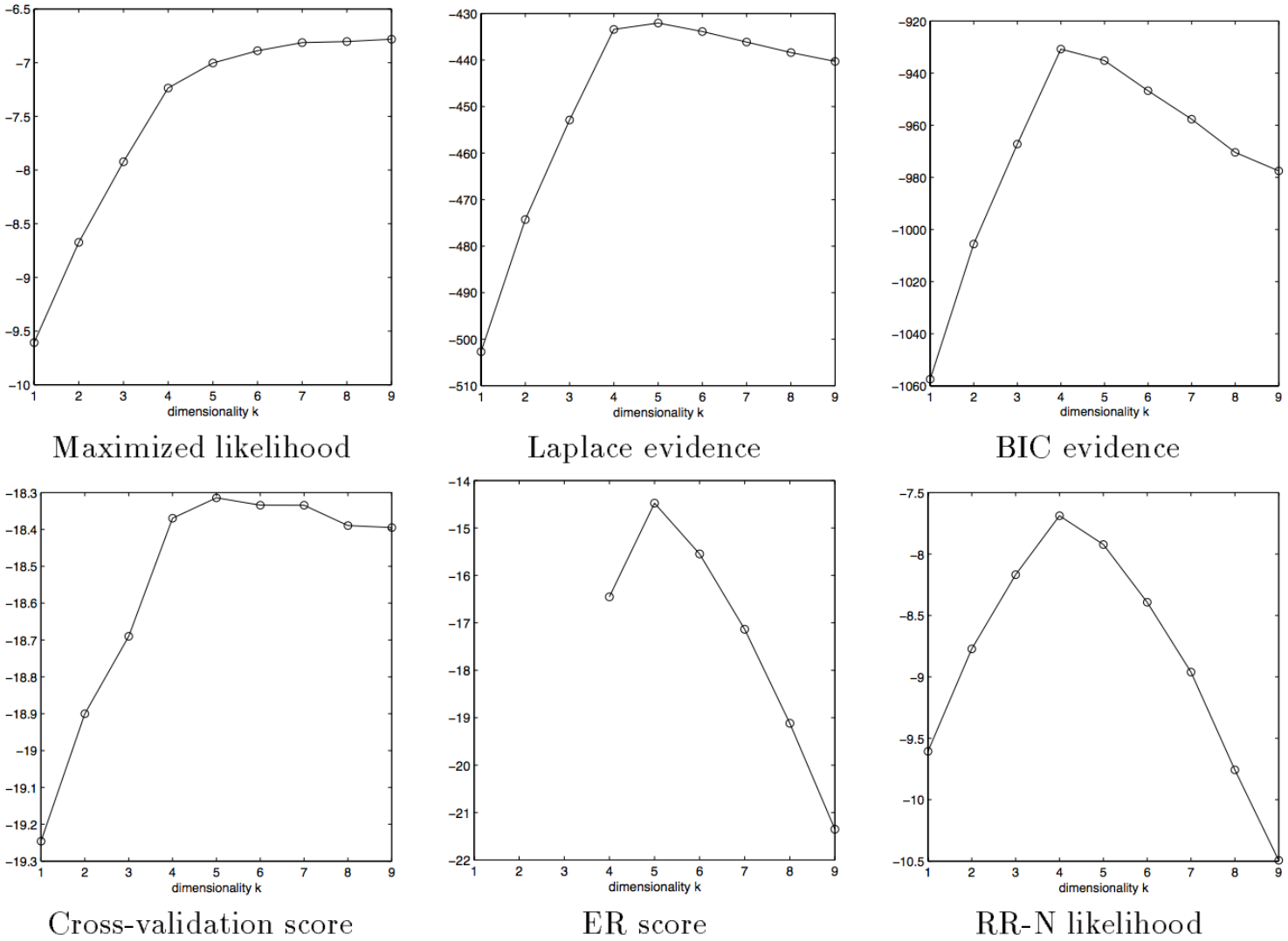


Figure 4: The score for each dimensionality, evaluated in six different ways. The true value is $k = 5$.

Automatically Learning the Dimensionality of PCA

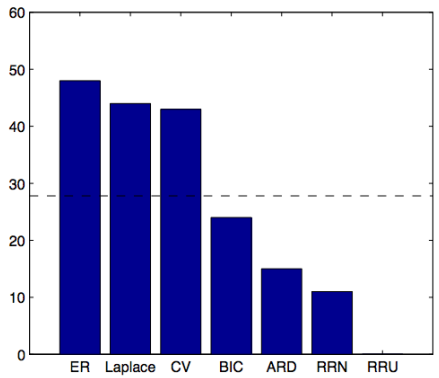


Figure 5: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 10, k = 5, N = 100$)

Automatically Learning the Dimensionality of PCA

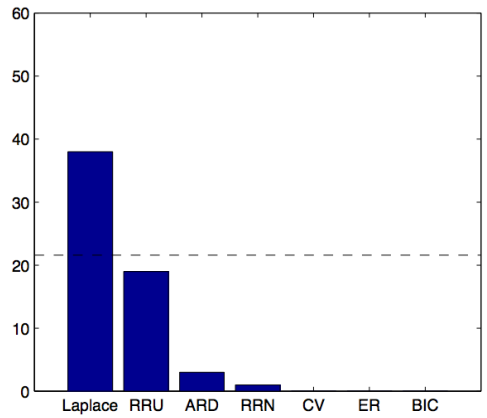


Figure 6: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 15, k = 5, N = 10$)

Automatically Learning the Dimensionality of PCA

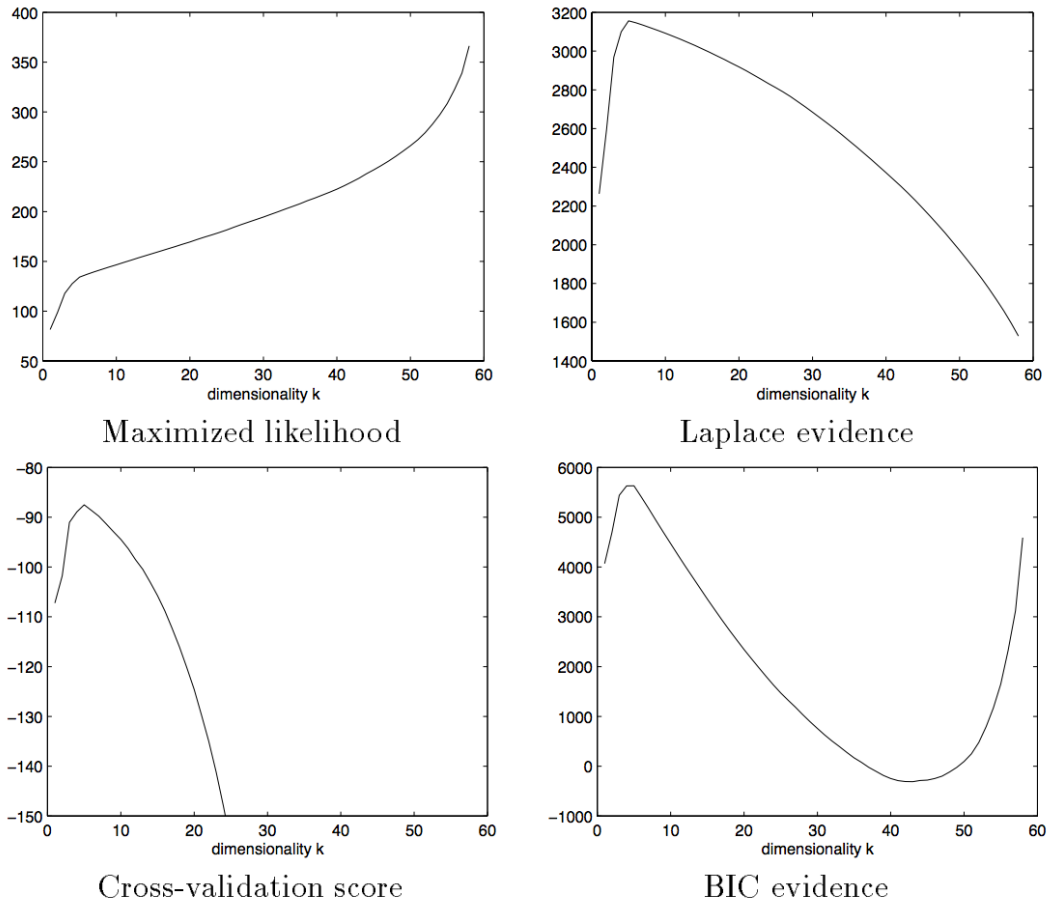


Figure 8: The score for each dimensionality, evaluated in four different ways. The cross-validation curve drops off quickly after $k = 15$. All except the likelihood peak at the true value in this case.

Automatically Learning the Dimensionality of PCA

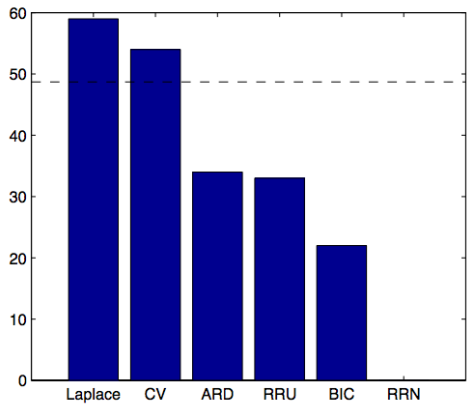


Figure 9: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 100, k = 5, N = 60$)

Model Construction: Support and Inductive Biases

- Support: which datasets (hypotheses) are a priori possible.
- Inductive Biases: which datasets are a priori likely.

Want to make the *support* of our model as big as possible, with inductive biases which are calibrated to particular applications, so as to not rule out potential explanations of the data, while at the same time quickly learn from a finite amount of information on a particular application.