Bayesian Machine Learning

Andrew Gordon Wilson

ORIE 6741
Lecture 3
Stochastic Gradients, Bayesian Inference, and Occam's Razor
https://people.orie.cornell.edu/andrew/orie6741

Cornell University
August 30, 2016

Bayesian Modelling (Theory of Everything)

Everything follows from two simple rules:

Sum rule: $P(x) = \sum_y P(x, y)$

Product rule: $P(x, y) = P(x)P(y|x)$

$$P(\theta|\mathcal{D}, m) = \frac{P(\mathcal{D}|\theta, m)P(\theta|m)}{P(\mathcal{D}|m)}$$

$P(\mathcal{D}\|\theta, m)$	likelihood of parameters θ in model m
$P(\theta\|m)$	prior probability of θ
$P(\theta\|\mathcal{D}, m)$	posterior of θ given data $\mathcal{D}$

Prediction:

$$P(x|\mathcal{D}, m) = \int P(x|\theta, \mathcal{D}, m)P(\theta|\mathcal{D}, m)d\theta$$

Model Comparison:

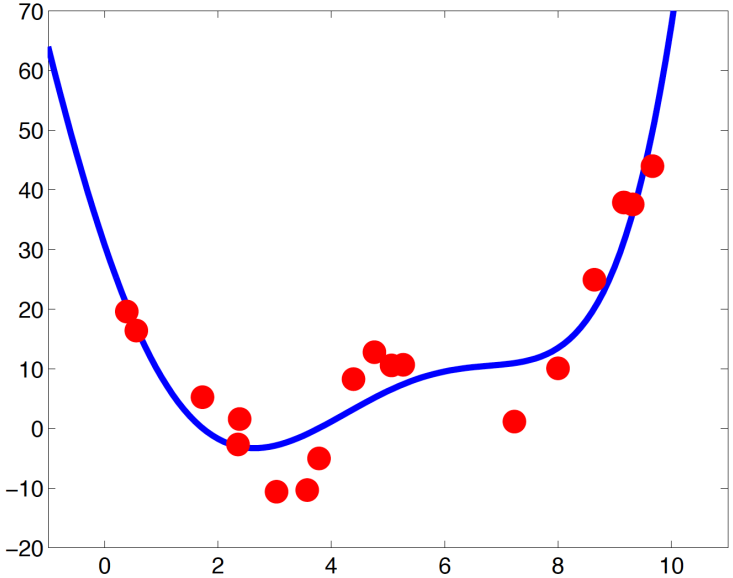
$$P(m|\mathcal{D}) = \frac{P(\mathcal{D}|m)P(m)}{P(\mathcal{D})}$$

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta|m)\, d\theta$$

Slide from Ghahramani (2015).

Worked Example: Basis Regression (Chalkboard)

- We have data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$



- We use the model:

$$y = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) + \epsilon \tag{1}$$

$$\epsilon \sim \mathcal{N}(0, \sigma^2)\,. \tag{2}$$

- We want to make predictions of y_* for any $\mathbf{x}_*$.
- We will consider topics such as *regularization*, *cross-validation*, *Bayesian model averaging* and *conjugate priors*.

Bayesian Linear Basis Regression Results

- We have data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$
- We use the model:

$$y = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}) + \epsilon \tag{3}$$
$$p(\mathbf{w}|\alpha^2) = \mathcal{N}(\mathbf{w}; 0, \alpha^2 I) \tag{4}$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{5}$$

Inference

$$p(\mathbf{w}|\mathbf{y}, X, \alpha^2) \propto p(\mathbf{y}|\mathbf{w}, X) p(\mathbf{w}|\alpha^2) \tag{6}$$
$$p(\mathbf{w}|\mathbf{y}, X) = \mathcal{N}(\mathbf{w}; \mathbf{m}_n, S_n) \tag{7}$$
$$\mathbf{m}_n = \frac{1}{\sigma^2} S_n \Phi^{\mathrm{T}} \mathbf{y} \tag{8}$$
$$S_n^{-1} = \frac{1}{\alpha^2} I + \frac{1}{\sigma^2} \Phi^{\mathrm{T}} \Phi \tag{9}$$
$$\Phi_{ij} = \phi_j(\mathbf{x}_i)\,. \tag{10}$$

Predictions

$$p(y_*|\mathbf{x}_*, \mathcal{D}, \alpha^2, \sigma^2) = \int p(y_*|\mathbf{w}, \mathbf{x}_*, \sigma^2) p(\mathbf{w}|\mathcal{D}, \alpha^2) \quad = \mathcal{N}(\mathbf{m}_n^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_*), \sigma_n(\mathbf{x}_*))$$
$$\sigma_n(\mathbf{x}_*) = \sigma^2 + \boldsymbol{\phi}(\mathbf{x}_*) S_n \boldsymbol{\phi}(\mathbf{x}_*)\,.$$

Bayesian Linear Basis Regression Results

Learning

$$p(\mathbf{y}|\alpha^2, \sigma^2) = \int p(\mathbf{y}|\mathbf{w}, \sigma^2) p(\mathbf{w}|\alpha) d\mathbf{w} \tag{11}$$

$$\log p(\mathbf{y}|\alpha, \sigma) = -\frac{n}{2}\log(2\pi) - \frac{m}{2}\log\alpha^2 - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\log|A| - E(\mathbf{m}_N)\,, \tag{12}$$

$$E(\mathbf{m}_N) = \frac{1}{2\sigma^2}||\mathbf{y} - \Phi\mathbf{m}_N||^2 + \frac{1}{2\sigma^2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N\,, \tag{13}$$

$$A = \frac{I}{\alpha^2} + \frac{1}{\sigma^2}\Phi^{\mathrm{T}}\Phi = \nabla\nabla E(\mathbf{w})\,, \tag{14}$$

$$\mathbf{m}_N = \frac{1}{\sigma^2}A^{-1}\Phi^{\mathrm{T}}\mathbf{y}\,. \tag{15}$$

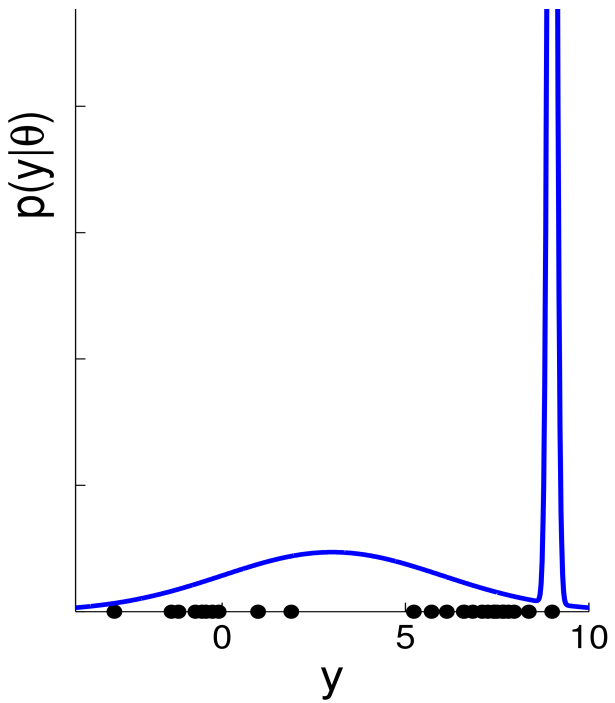
Procedure: Learn α^2 and γ^2 through marginal likelihood optimization. Then condition on these learned parameters to form the predictive distribution: $p(y_*|\mathbf{x}_*, \mathcal{D}, \hat{\alpha}^2, \hat{\sigma}^2)$.

Rant: Regularisation = MAP $\neq$ Bayesian Inference

Example: Density Estimation

- Observations $y_1, \ldots, y_N$ drawn from unknown density $p(y)$.
- Model $p(y|\theta) = w_1 \mathcal{N}(y|\mu_1, \sigma_1^2) + w_2 \mathcal{N}(y|\mu_2, \sigma_2^2)$,
  $\theta = \{w_1, w_2, \mu_1, \mu_2, \sigma_1, \sigma_2\}$.
- Likelihood $p(\mathbf{y}|\theta) = \prod_{i=1}^{N} p(y_i|\theta)$.

Can learn all free parameters θ using maximum likelihood...



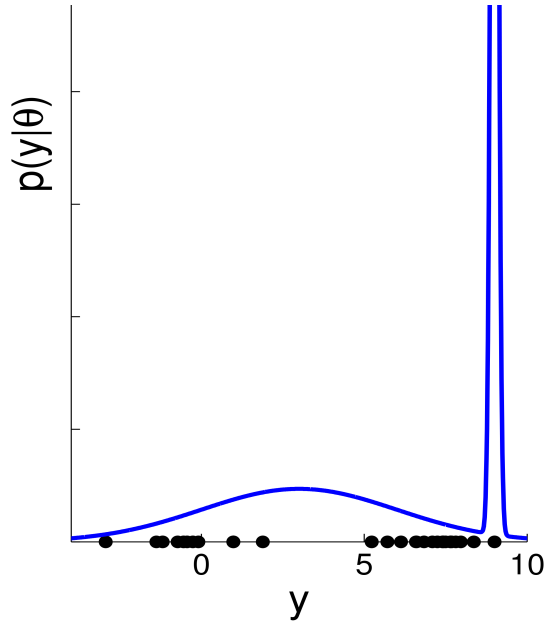
Regularisation = MAP $\neq$ Bayesian Inference

Regularisation or MAP

- Find

$$\text{argmax}_\theta \log p(\theta|\mathbf{y}) \stackrel{c}{=} \overbrace{\log p(\mathbf{y}|\theta)}^{\text{model fit}} + \overbrace{\log p(\theta)}^{\text{complexity penalty}}$$

- Choose $p(\theta)$ such that $p(\theta) \to 0$ faster than $p(\mathbf{y}|\theta) \to \infty$ as σ_1 or $\sigma_2 \to 0$.



Bayesian Inference

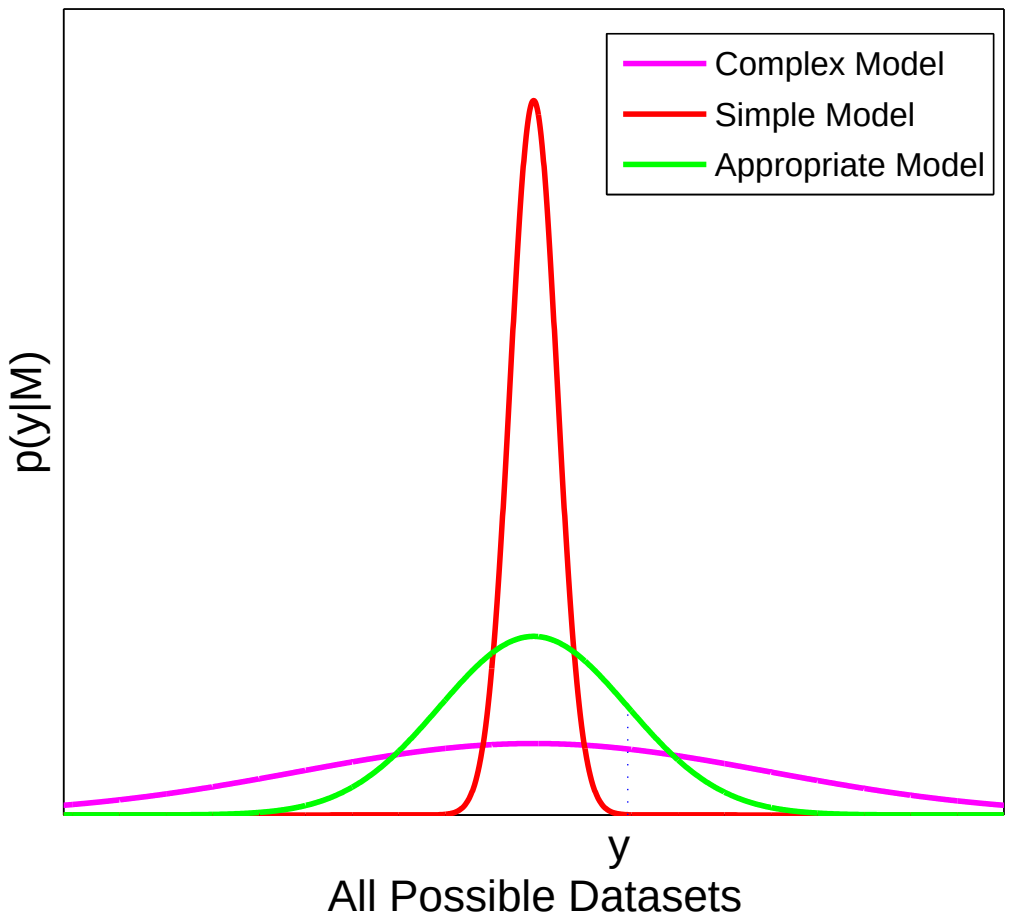
- Predictive Distribution: $p(y_*|\mathbf{y}) = \int p(y_*|\theta)p(\theta|\mathbf{y})d\theta$.
- Parameter Posterior: $p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta)$.
- $p(\theta)$ need not be zero anywhere in order to make reasonable inferences.
- Can use a sampling scheme, with conjugate posterior updates for each separate mixture component, using an inverse Gamma prior on the variances σ_1^2, σ_2^2.

Learning with Stochastic Gradient Descent

Chalkboard.

Model Selection and Marginal Likelihood

$$p(\mathbf{y}|\mathcal{M}_1, X) = \int p(\mathbf{y}|f_1(x, \mathbf{w}))p(\mathbf{w})d\mathbf{w} \tag{16}$$



Model Comparison

$$\frac{p(\mathcal{H}_1|\mathcal{D})}{p(\mathcal{H}_2|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H}_1)}{p(\mathcal{D}|\mathcal{H}_2)} \frac{p(\mathcal{H}_1)}{p(\mathcal{H}_2)} \,. \tag{17}$$

Blackboard: Examples of Occam's Razor in Everyday Inferences

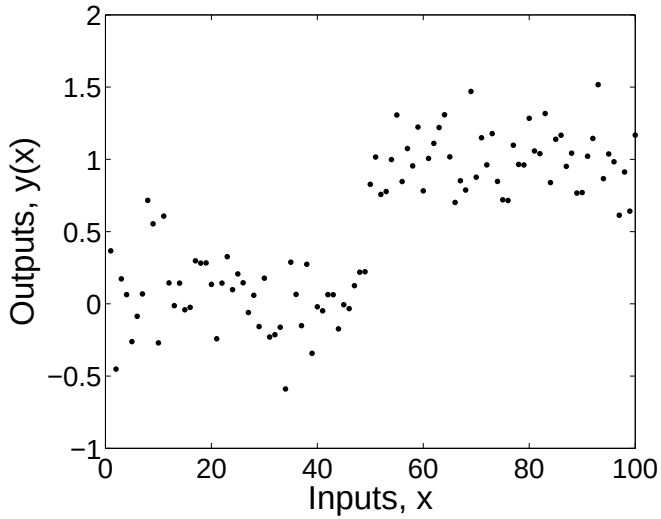
For further reading, see MacKay (2003) textbook, *Information Theory, Inference, and Learning Algorithms*.

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?

$$f_0(x) = a_0\,, \tag{18}$$

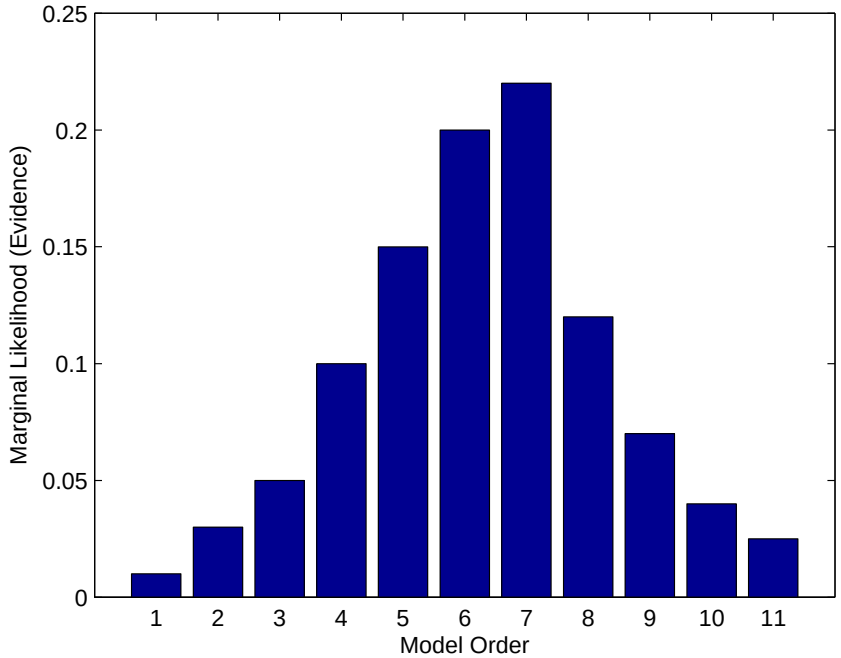
$$f_1(x) = a_0 + a_1 x\,, \tag{19}$$

$$f_2(x) = a_0 + a_1 x + a_2 x^2\,, \tag{20}$$

$$\vdots \tag{21}$$

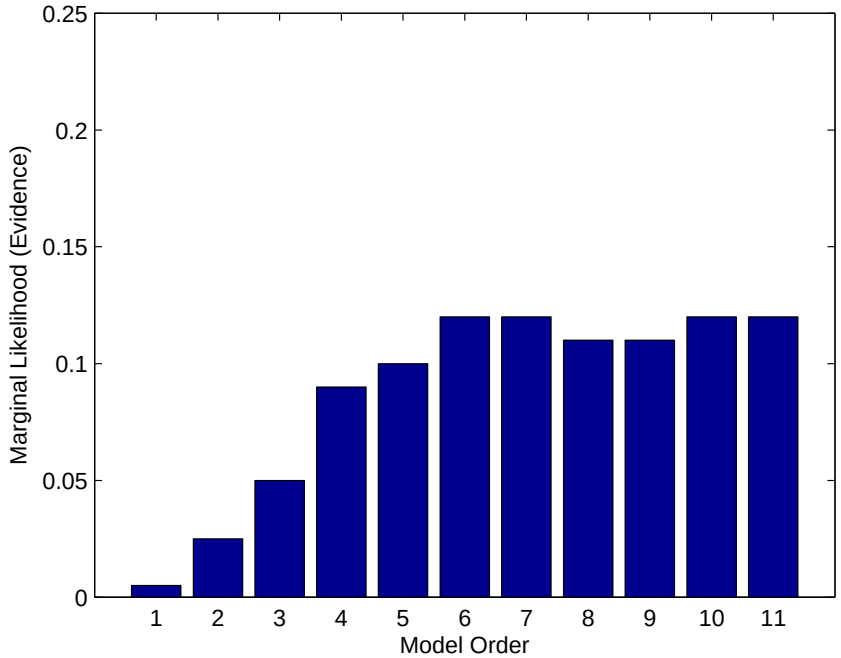
$$f_J(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_J x^J\,. \tag{22}$$

Model Selection: Occam's Hill



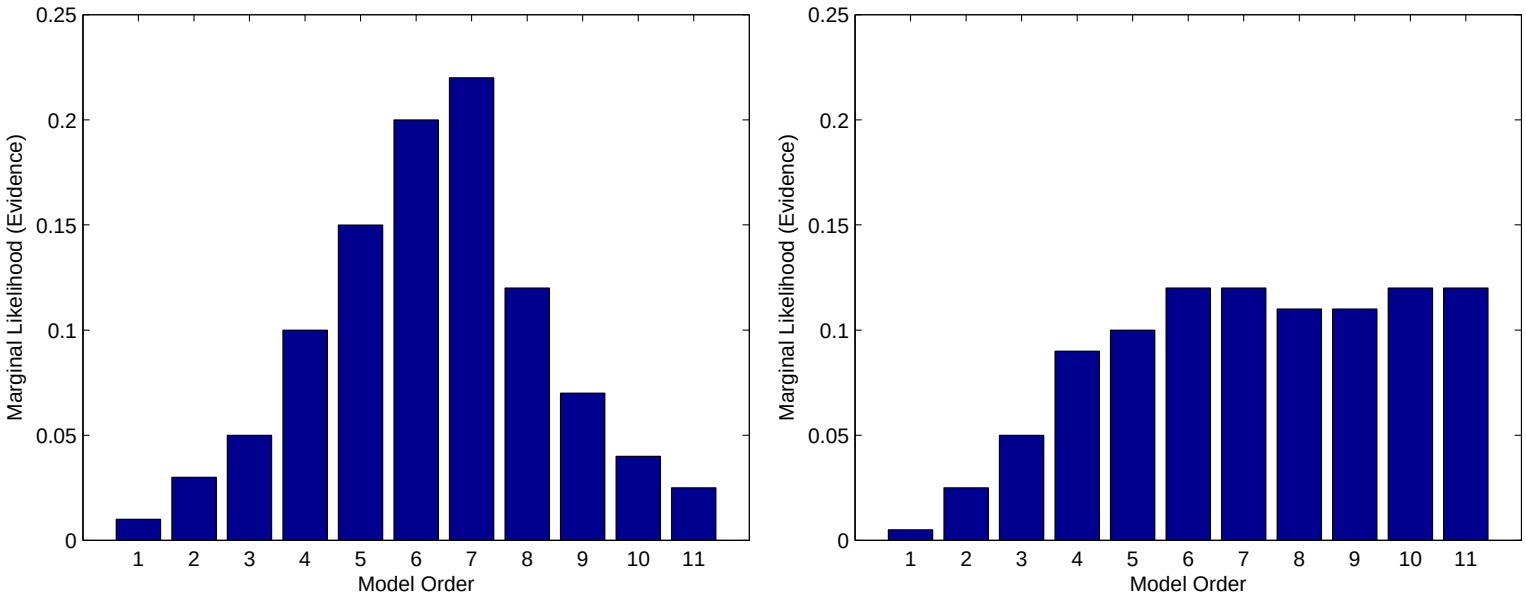
Marginal likelihood (evidence) as a function of model order, using an isotropic prior $p(a) = \mathcal{N}(0, \sigma^2 I)$.

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:

$$\mathbf{x} = \sum_{j=1}^{k} \mathbf{h}_j w_j + \mathbf{m} + \boldsymbol{\epsilon}\,, \tag{23}$$

$$= H\mathbf{w} + \mathbf{m} + \boldsymbol{\epsilon}\,, \tag{24}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, V)\,. \tag{25}$$

- 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

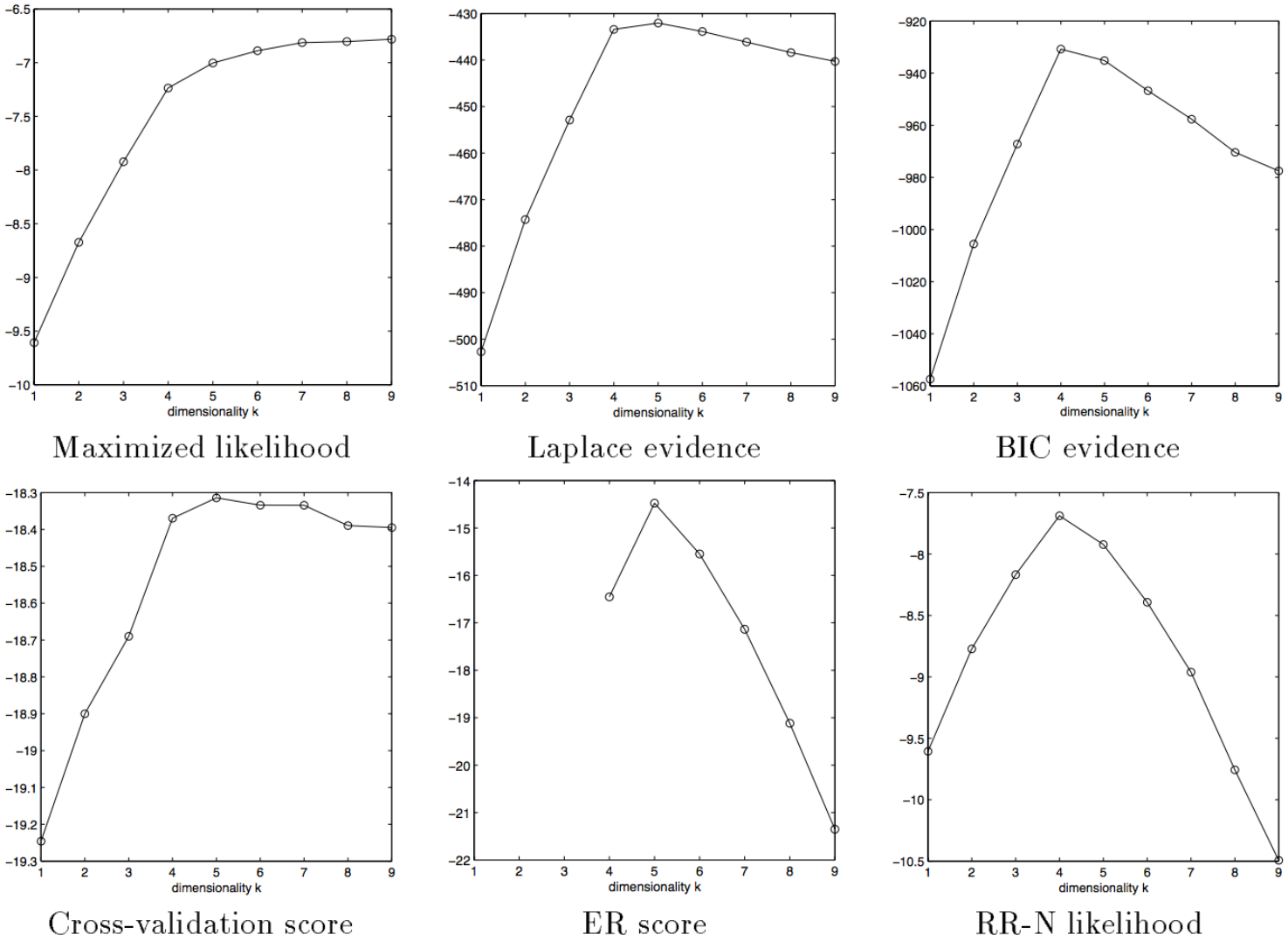


Figure 4: The score for each dimensionality, evaluated in six different ways. The true value is $k = 5$.

Automatically Learning the Dimensionality of PCA

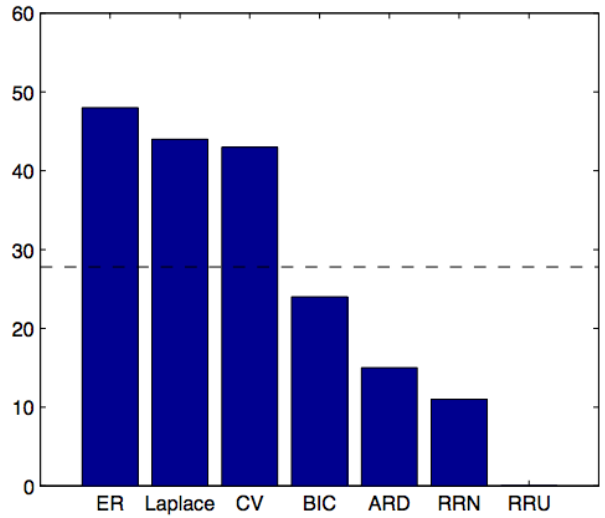


Figure 5: The number of times each estimator picked the correct dimensionality in 60 replications ($5, N = 100$)

Automatically Learning the Dimensionality of PCA

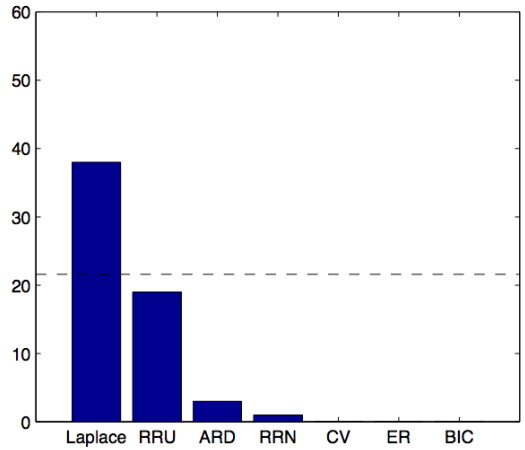


Figure 6: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 15, k = 5, N = 10$)

Automatically Learning the Dimensionality of PCA

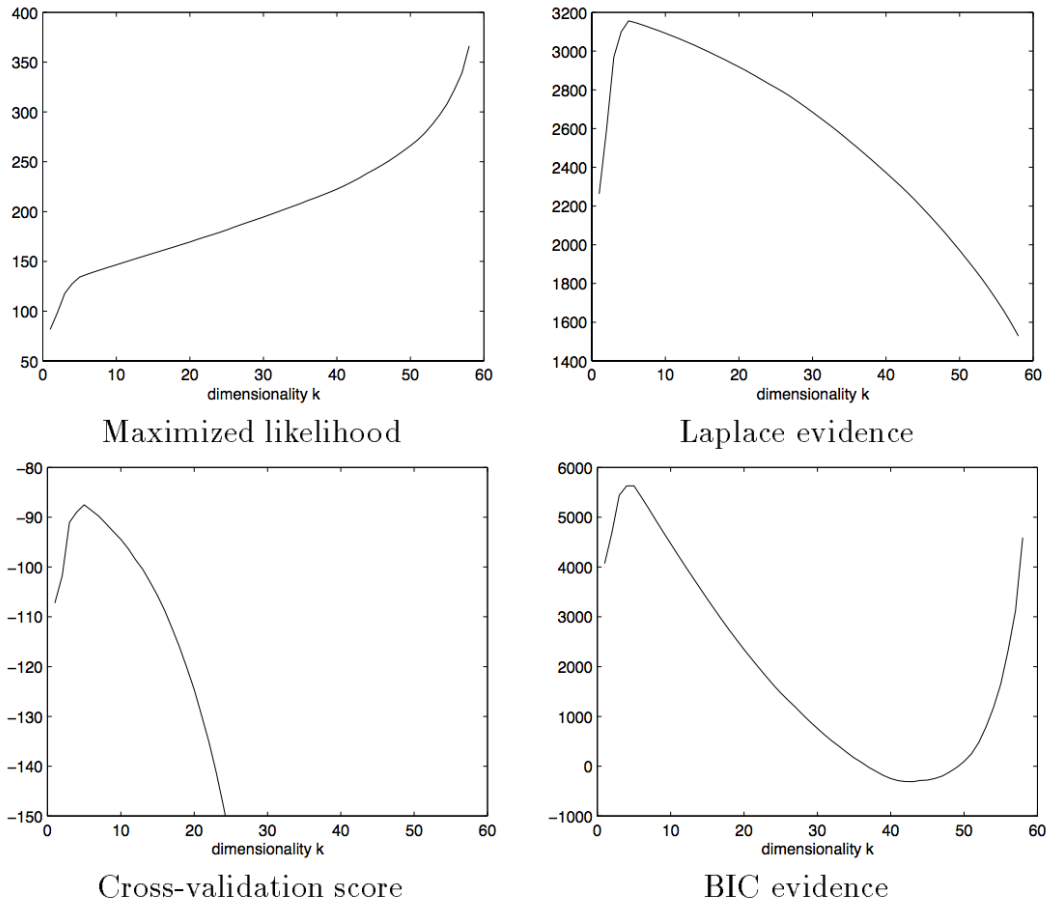


Figure 8: The score for each dimensionality, evaluated in four different ways. The cross-validation curve drops off quickly after $k = 15$. All except the likelihood peak at the true value in this case.

Automatically Learning the Dimensionality of PCA

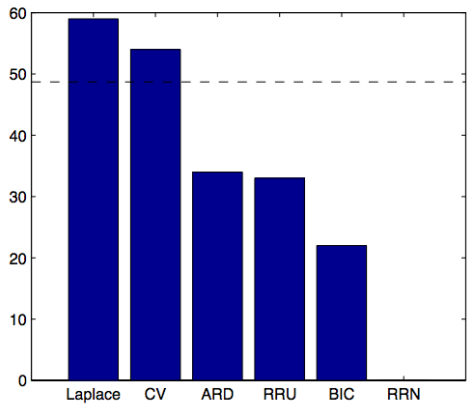


Figure 9: The number of times each estimator picked the correct dimensionality in 60 replications. ($d = 100, k = 5, N = 60$)