Bayesian Model Selection

Václav Šmídl

March 5, 2018

Model selection

- Model = Likelihood & Prior.
- Posterior is

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}$$

where $p(d|\theta)$ and $p(\theta)$ is given by ???

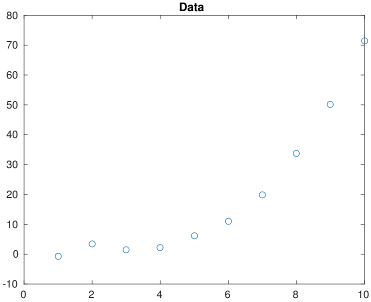
Model selection

- Model = Likelihood & Prior.
- Posterior is

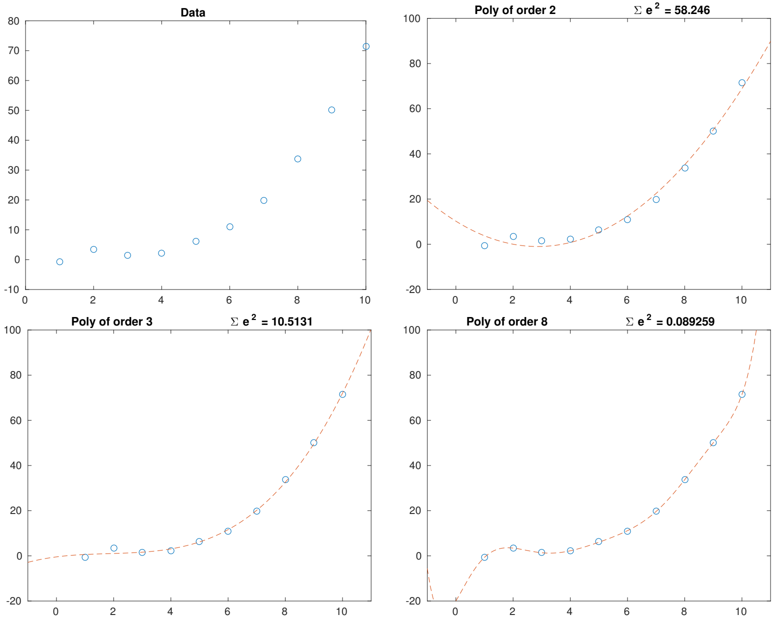
$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}$$

  where $p(d|\theta)$ and $p(\theta)$ is given by ???
- In general, we can not prove that the model is the best.
- We can formulate several candidates and compare them.

Example



Example



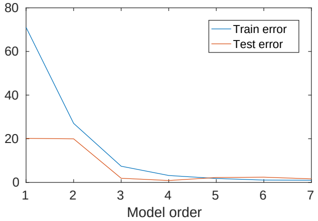
Data
Poly of order 2
Σ e ^{2} = 58.246
Poly of order 3
Σ e ^{2} = 10.5131
Poly of order 8
Σ e ^{2} = 0.089259

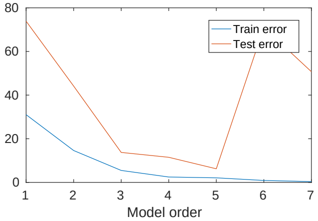
Cross-validation

- Split the data into **training** and **testing** set.
- Fit model parameters on the training set.
- Evaluate model error on the test set.

Cross-validation

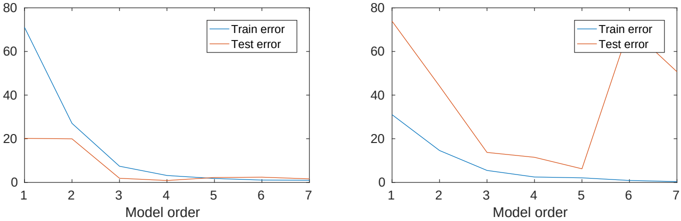
- Split the data into **training** and **testing** set.
- Fit model parameters on the training set.
- Evaluate model error on the test set.





Cross-validation

- Split the data into **training** and **testing** set.
- Fit model parameters on the training set.
- Evaluate model error on the test set.



- Sensitive to sampling,
- May be problematic for hierarchical models.

Bayesian Model selection

- Assume a fixed set of available models $M \in \{M_1, M_2, \ldots M_m\}$

$$M_i : \quad p_i(d|\theta_i)p_i(\theta_i)$$

- Compute the probability that the data were generated from each model:

$$p(M = M_i|d) \quad \propto \quad p(d|M_i)p(M_i)$$

Bayesian Model selection

- Assume a fixed set of available models $M \in \{M_1, M_2, \dots M_m\}$

$$M_i : \quad p_i(d|\theta_i)p_i(\theta_i)$$

- Compute the probability that the data were generated from each model:

$$\begin{aligned} p(M = M_i|d) &\propto p(d|M_i)p(M_i) \\ &= p(M_i) \int p(d|\theta_i, M_i)p(\theta_i)d\theta, \end{aligned}$$

Bayesian Model selection

- Assume a fixed set of available models $M \in \{M_1, M_2, \dots M_m\}$

$$M_i : \quad p_i(d|\theta_i)p_i(\theta_i)$$

- Compute the probability that the data were generated from each model:

$$\begin{aligned} p(M = M_i|d) &\propto p(d|M_i)p(M_i) \\ &= p(M_i) \int p(d|\theta_i, M_i)p(\theta_i)d\theta, \end{aligned}$$

- Marginal likelihood (evidence) $p(d|M)$ is the normalizing constant of the Bayes rule.
- Can be either readily available (Laplace) or hard to find (Gibbs).

Bayes factor

- Comparison of two models (hypotheses):

$$K = \frac{p(d|M_1)}{p(d|M_2)} = \frac{\int p(\theta_1|M_1)p(d|\theta_1, M_1)\, d\theta_1}{\int p(\theta_2|M_2)p(d|\theta_2, M_2)\, d\theta_2} = \frac{p(M_1|d)}{p(M_2|d)}\frac{p(M_2)}{p(M_1)}.$$

  Typically assumed that $p(M_1) = p(M_2)$.

Bayes factor

- Comparison of two models (hypotheses):

$$K = \frac{p(d|M_1)}{p(d|M_2)} = \frac{\int p(\theta_1|M_1)p(d|\theta_1, M_1)\, d\theta_1}{\int p(\theta_2|M_2)p(d|\theta_2, M_2)\, d\theta_2} = \frac{p(M_1|d)}{p(M_2|d)}\frac{p(M_2)}{p(M_1)}.$$

  Typically assumed that $p(M_1) = p(M_2)$.

- Interpretation

K	Strength of evidence
1 to 3	not worth more than a bare mention
3 to 20	positive
20 to 150	strong
$>$150	very strong

- Often reported only as: $\log p(M_i|d)$

Challenge: toy example

- Noisy observation:

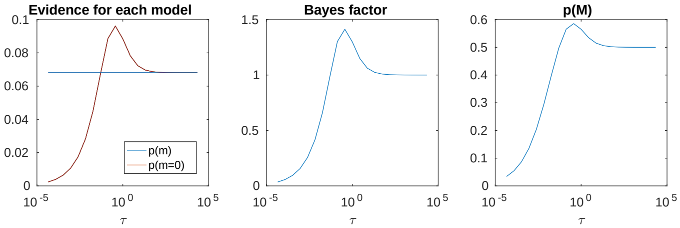
$$\begin{aligned} M_1 : &\quad d = m + e, \\ M_2 : &\quad d = e, \end{aligned}$$

where:

$$\begin{aligned} p(d_i) &= \mathcal{N}(m, \omega^{-1}), \\ p(m|\omega) &= \mathcal{N}(0, \tau\omega), \\ p(\omega) &= G(\alpha, \beta). \end{aligned}$$

- $d = 2$, $\alpha = 1$, $\beta = 1$
- K=?
- $p(M_1)$=?

Challenge: toy example



Asymptotic Approximations

Bayesian information criterion (BIC) or Schwarz criterion (also SBC) [Schwarz, 1978]:

$$-2 \cdot \ln p(d|M) \approx \mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot (\ln(n) - \ln(2\pi)).$$
$$\hat{L} = p(d|\hat{\theta}),$$

where $\hat{\theta}$ is maximum likelihood estimate of parameter θ, n is the number of data, k is the number of parameters.

- Penalization of model complexity

Asymptotic Approximations

Bayesian information criterion (BIC) or Schwarz criterion (also SBC) [Schwarz, 1978]:

$$-2 \cdot \ln p(d|M) \approx \mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot (\ln(n) - \ln(2\pi)).$$
$$\hat{L} = p(d|\hat{\theta}),$$

where $\hat{\theta}$ is maximum likelihood estimate of parameter θ, n is the number of data, k is the number of parameters.

- Penalization of model complexity

Akaike information criterion (AIC) [Akaike, 1974] asymptotic of cross-validation, goodness of fit versus model simplicity:

$$\mathrm{AIC} = -2 \cdot \ln \hat{L} + k \cdot 2.$$

Asymptotic Approximations

Bayesian information criterion (BIC) or Schwarz criterion (also SBC) [Schwarz, 1978]:

$$-2 \cdot \ln p(d|M) \approx \mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot (\ln(n) - \ln(2\pi)).$$
$$\hat{L} = p(d|\hat{\theta}),$$

where $\hat{\theta}$ is maximum likelihood estimate of parameter θ, n is the number of data, k is the number of parameters.

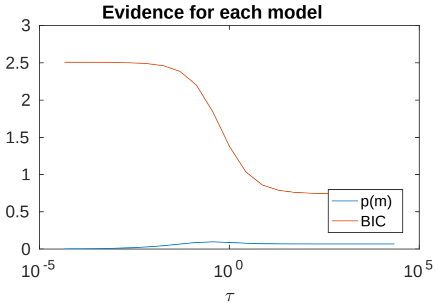
- Penalization of model complexity

Akaike information criterion (AIC) [Akaike, 1974] asymptotic of cross-validation, goodness of fit versus model simplicity:

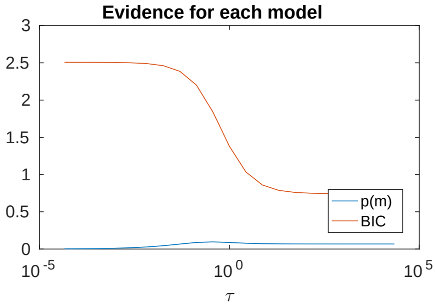
$$\mathrm{AIC} = -2 \cdot \ln \hat{L} + k \cdot 2.$$

WAIC, DIC, ...

Asymptotic Approximations



Asymptotic Approximations



- not asymptotics

Laplace approximation

Laplace approximation (derived without normalization):

$$
\begin{aligned}
p(\theta|d) &\approx \mathcal{N}(\hat{\theta}, \Sigma), \\
\hat{\theta} &= \arg\max p(\theta, d), \\
\Sigma &= (-\nabla\nabla \log p(\hat{\theta}))^{-1}
\end{aligned}
$$

Evidence (normalization constant) [Kass, Raftery, 1995]:

$$p(d) = (2\pi)^{d/2}|\Sigma|^{1/2} p(d|\hat{\theta}) p(\hat{\theta})$$

Often used with large datasets.

Variational Bayes

Original Variational Bayes derived without normalization.

$$p(\theta_1, \theta_2) \approx q(\theta_1)q(\theta_2)$$

Considering joint model $p(d|\theta_1, M_1)$ and $p(d|\theta_2, M_2)$ we can not split $q(M)q(\theta_?)$.

Considering $q(Z|M)q(M)$, the solution is [Bishop, 2006]:

$$q(M|d) \propto p(M)\exp(\mathcal{L}_M)$$
$$\mathcal{L}_M = KL\left(q(\theta|M)||p(d,\theta|M)\right)$$

where $q(\theta|M)$ are results of the standard Variational Bayes for each model.

Toy: Variational Bayes

General rule: $q(x_1|\theta_1) \propto \exp\left(E_{qx(1)}\left[\log p(x_1, x_2)\right]\right)$
Toy (with constant c)

$$\log p(m, \omega, d) = \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2$$
$$\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha-1)\log\omega - \beta\omega + c$$

Toy: Variational Bayes

General rule: $q(x_1|\theta_1) \propto \exp\left(E_{qx(1)}\left[\log p(x_1, x_2)\right]\right)$
Toy (with constant c)

$$\log p(m, \omega, d) = \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2$$
$$\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log\omega - \beta\omega + c$$

Factor $q(\omega|d)$:

$$\log q(\omega|d) \propto E_{q(m)}\left[\alpha\log\omega - \frac{1}{2}\omega\left((d-m)^2 + \tau m^2 + \beta\right)\right]$$

Toy: Variational Bayes

General rule: $q(x_1|\theta_1) \propto \exp\left(E_{qx(1)}\left[\log p(x_1, x_2)\right]\right)$
Toy (with constant c)

$$\log p(m, \omega, d) = \frac{1}{2}\log\omega - \frac{1}{2}\omega(d-m)^2$$
$$\frac{1}{2}\log\omega - \frac{1}{2}\tau\omega m^2 + (\alpha - 1)\log\omega - \beta\omega + c$$

Factor $q(\omega|d)$:

$$\log q(\omega|d) \propto E_{q(m)}\left[\alpha\log\omega - \frac{1}{2}\omega\left((d-m)^2 + \tau m^2 + \beta\right)\right]$$

Log-likelihood $q(d|M) = \mathcal{L}_M$:

$$\begin{aligned}\mathcal{L}_M &= E_{q(\omega)q(m)}\left[\alpha\log\omega - \frac{1}{2}\omega\left((d-m)^2 + \tau m^2 + \beta\right) + c\right]\\ &= \alpha\left\langle\log\omega\right\rangle - \frac{1}{2}\left\langle\omega\right\rangle\left\langle(d-m)^2 + \tau m^2 + \beta\right\rangle + c\end{aligned}$$

where $\langle\log\omega\rangle$ needs to be computed for the first time!

Monte Carlo methods

Approximation of a distribution by "Dirac train"

$$p(x) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(x - x^{(i)}).$$

Approximation of moments, cumulative density.
We seek integral

$$p(d|M) = \int p(d|\theta, M)p(\theta|M)d\theta,$$

which can be solved by sampling from $p(\theta|M)$.

Monte Carlo methods

Approximation of a distribution by "Dirac train"

$$p(x) \approx \frac{1}{N}\sum_{i=1}^{N}\delta(x - x^{(i)}).$$

Approximation of moments, cumulative density.
We seek integral

$$p(d|M) = \int p(d|\theta, M)p(\theta|M)d\theta,$$

which can be solved by sampling from $p(\theta|M)$. Inefficient, numerically unstable.

- Importance sampling [Perrakis, Ntzoufras, and Tsionas, 2014],
- Gibbs sampler [Chib, 1995] using

$$\ln\left(p\left(d|M\right)\right) = \ln\left(p\left(d|M,\theta\right)\right) + \ln\left(p\left(\theta\right)\right) - \ln\left(p\left(\theta|d\right)\right),$$

evaluated point-wise.

Searching of the model space

Consider a set of models with m binary options, forming a space of 2^m hypothesis.

- Marginal likelihood for each hypotheses can be evaluated
- How to efficiently find the best?

Searching of the model space

Consider a set of models with m binary options, forming a space of 2^m hypothesis.

- Marginal likelihood for each hypotheses can be evaluated
- How to efficiently find the best?
- Combinatorial optimization
  - Genetic algorithms,
  - Simulated annealing,
- Can we use some information about the search space? The evaluations are not completely independent.

MCMC approximation

- Is it possible to treat M as a “normal” random variable?

MCMC approximation

- Is it possible to treat M as a "normal" random variable?
- In general MCMC, we need to define a transition kernel from which we sample x^* and accept/reject it
- With different M, dimension of x is changing

MCMC approximation

- Is it possible to treat M as a "normal" random variable?
- In general MCMC, we need to define a transition kernel from which we sample x^* and accept/reject it
- With different M, dimension of x is changing

Reversible jump MCMC standard sample x is complemented by vector of random numbers u such that couples (x, u) and (x', u') can be reversibly mapped. MH is then extended [Green, Hastie, 2009]

$$\alpha(x, x') = \min \left\{ 1, \frac{\pi(x')j(x')g'(u')}{\pi(x)j(x)g(u)} \left| \frac{\partial(\theta'_{k'}, u')}{\partial(\theta_k, u)} \right| \right\},$$

exploring space of hypothesis (not necessarily finite).

Model selection

- Evidence or marginal likelihood is an important quantity for model selection,
- Provides an alternative to cross-validation
- In machine learning, many benchmark data sets are compared using log-likelihood

Model	$\log p(x) \geq$
NF (k=80) [Rezende et al., 2015]	-85.1
PixelRNN [Oord et al., 2016]	-79.2
AVB [Mescheder et al., 2016]	-79.5
ASVAE [Pu et al., 2017]	-81.14
GAN [Goodfellow et al., 2014]	-114.25 †
WGAN-GP [Ishaan Gulrajani, 2017]	-79.92 †
DCGAN [Radford et al., 2016]	-79.47 †
sVAE (ours)	-80.42 †
sVAE-r (ours)	-79.26 †