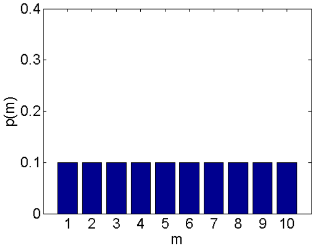
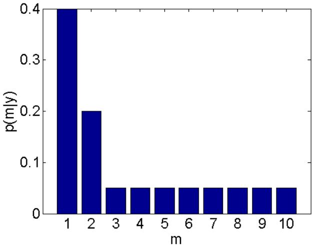
Model Comparison

Course on Bayesian Inference,
WTCN, UCL, February 2013

Bayes rule for models

A prior distribution over model space $p(m)$ (or ‘hypothesis space’) can be updated to a posterior distribution after observing data y.





This is implemented using Bayes rule

$$p(m|y) = \frac{p(y|m)p(m)}{p(y)}$$

where $p(y|m)$ is referred to as the evidence for model m and the denominator is given by

$$p(y) = \sum_{m'} p(y|m')p(m')$$

Model Evidence

The evidence is the denominator from the first (parameter) level of Bayesian inference

$$p(\theta|y,m) = \frac{p(y|\theta,m)p(\theta|m)}{p(y|m)}$$

The model evidence is not, in general, straightforward to compute since computing it involves integrating out the dependence on model parameters

$$p(y|m) = \int p(y|\theta,m)p(\theta|m)d\theta.$$

But for linear, Gaussian models there is an analytic solution.

Posterior Model Probability

Given equal priors, $p(m=i) = p(m=j)$ the posterior model probability is

$$\begin{aligned} p(m=i|y) &= \frac{p(y|m=i)}{p(y|m=i)+p(y|m=j)} \\ &= \frac{1}{1+\frac{p(y|m=j)}{p(y|m=i)}} \end{aligned}$$

Hence

$$p(m=i|y) = \sigma(\log B_{ij})$$

where

$$B_{ij} = \frac{p(y|m=i)}{p(y|m=j)}$$

is the Bayes factor for model 1 versus model 2 and

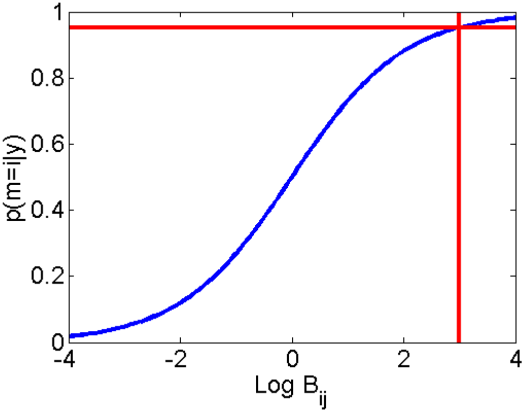
$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

is the sigmoid function.

Bayes factors

The posterior model probability is a sigmoidal function of the log Bayes factor

$$p(m = i|y) = \sigma(\log B_{ij})$$



Bayes factors

The posterior model probability is a sigmoidal function of the log Bayes factor

$$p(m = i|y) = \sigma(\log B_{ij})$$

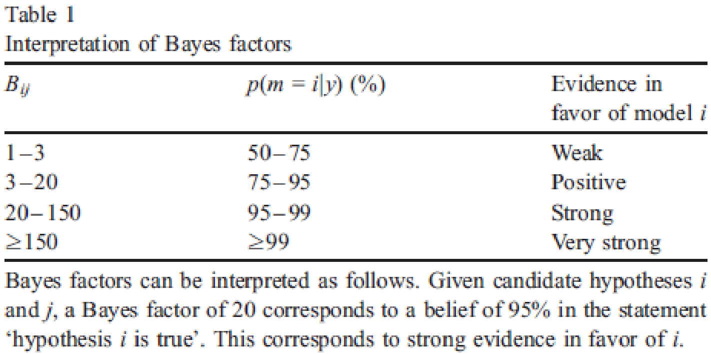


Table 1
Interpretation of Bayes factors

B_{ij}	$p(m = i\|y)$ (%)	Evidence in favor of model i
1–3	50–75	Weak
3–20	75–95	Positive
20–150	95–99	Strong
≥150	≥99	Very strong

Bayes factors can be interpreted as follows. Given candidate hypotheses i and j, a Bayes factor of 20 corresponds to a belief of 95% in the statement 'hypothesis i is true'. This corresponds to strong evidence in favor of i.

From Raftery (1995).

Odds Ratios

If we don't have uniform priors one can work with odds ratios.

The prior and posterior odds ratios are defined as

$$\begin{aligned} \pi_{ij}^{0} &= \frac{p(m=i)}{p(m=j)} \\ \pi_{ij} &= \frac{p(m=i|y)}{p(m=j|y)} \end{aligned}$$

resepectively, and are related by the Bayes Factor

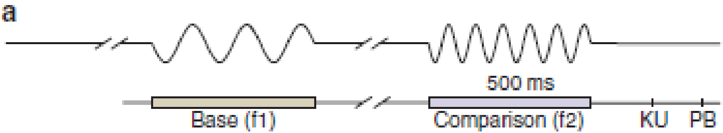
$$\pi_{ij} = B_{ij} \times \pi_{ij}^{0}$$

eg. priors odds of 2 and Bayes factor of 10 leads posterior odds of 20.

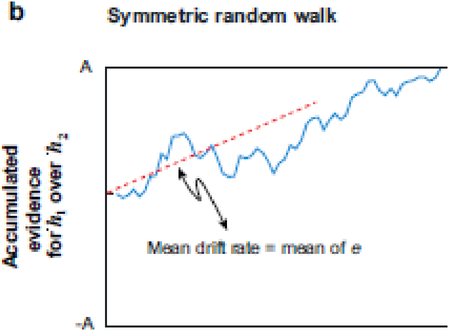
An odds ratio of 20 is 20-1 ON in bookmakers parlance.

Spike rates may reflect log odds ratios

In an vibro-tactile discrimination task (Gold, Ann. Rev. Neuro, 2007) monkey releases lever (KU) and presses one of two buttons (PB)

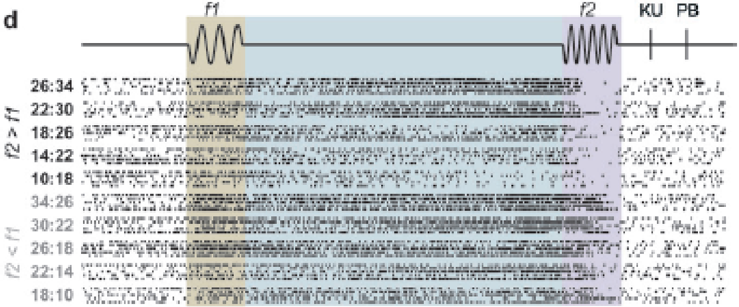


Sequential Likelihood Ratio Test (SLRT)



Spike rates may reflect log odds ratios

Cells in medial and ventral premotor cortices (but not S1) reflect odds ratios



This cell is more active during presentation of 2nd stimulus when $f2 < f1$.

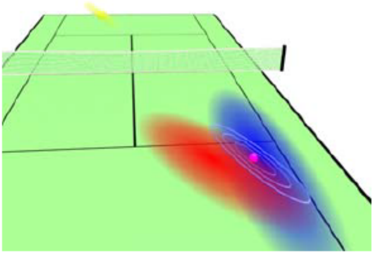
Linear Models

For Linear Models

$$y = Xw + e$$

where X is a design matrix and w are now regression coefficients. The posterior distribution is analytic and given by

$$\begin{aligned} S_w^{-1} &= X^T C_y^{-1} X + C_w^{-1} \\ m_w &= S_w \left(X^T C_y^{-1} y + C_w^{-1} \mu_w \right) \end{aligned}$$



Model Comparison

Bayes rule for models

Bayes factors

Spike rates

Linear Models

Model Evidence

Complexity

AIC and BIC

Example

fMRI example

Bayes versus classical inference

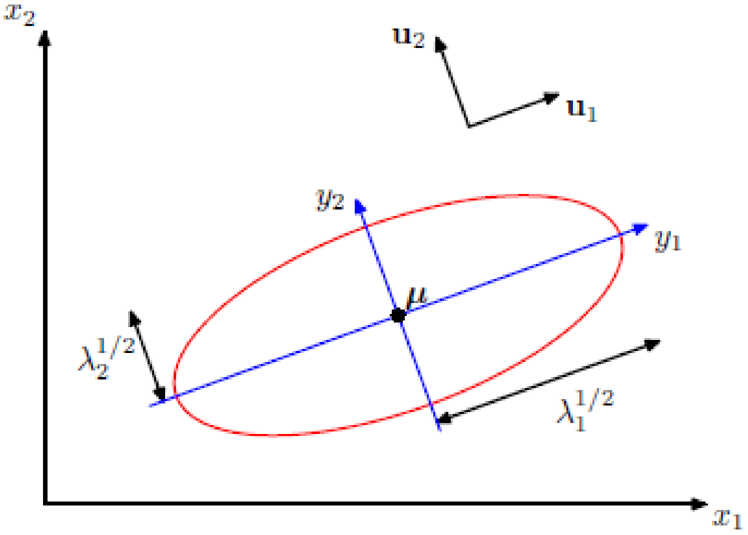
Model evidence in data space

Nonlinear classifiers

References

Covariance matrices

The determinant of a covariance matrix, $|C|$, measures the volume.



Model Evidence

The log model evidence comprises sum squared precision weighted prediction errors and Occam factors

$$\begin{aligned} L &= -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi \\ &- \frac{1}{2} e_w^T C_w^{-1} e_w - \frac{1}{2} \log \frac{|C_w|}{|S_w|} \end{aligned}$$

where prediction errors are the difference between what is expected and what is observed

$$\begin{aligned} e_y &= y - X m_w \\ e_w &= m_w - \mu_w \end{aligned}$$

See Bishop (2006) for derivation.

Accuracy and Complexity

The log evidence for model *m* can be split into an accuracy and a complexity term

$$L(m) = Accuracy(m) - Complexity(m)$$

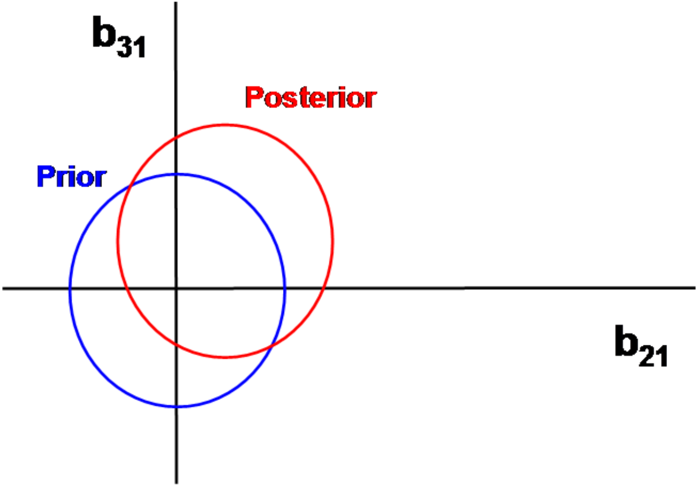
where

$$Accuracy(m) = -\frac{1}{2} e_y^T C_y^{-1} e_y - \frac{1}{2} \log |C_y| - \frac{N_y}{2} \log 2\pi$$

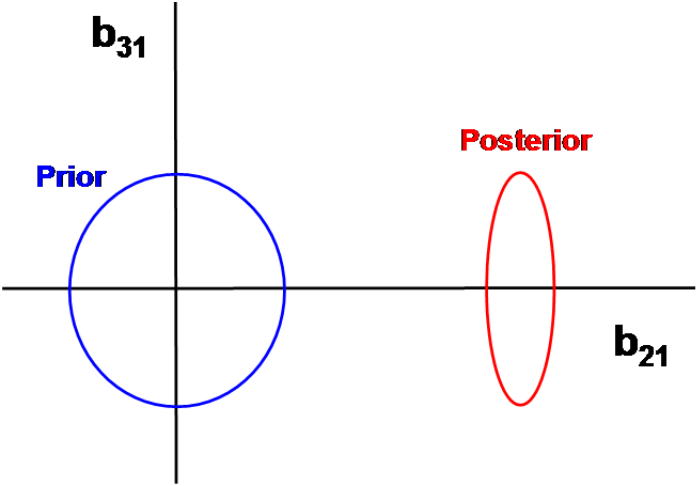
and

$$\begin{aligned} Complexity(m) &= \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|} \\ &\approx KL(prior||posterior) \end{aligned}$$

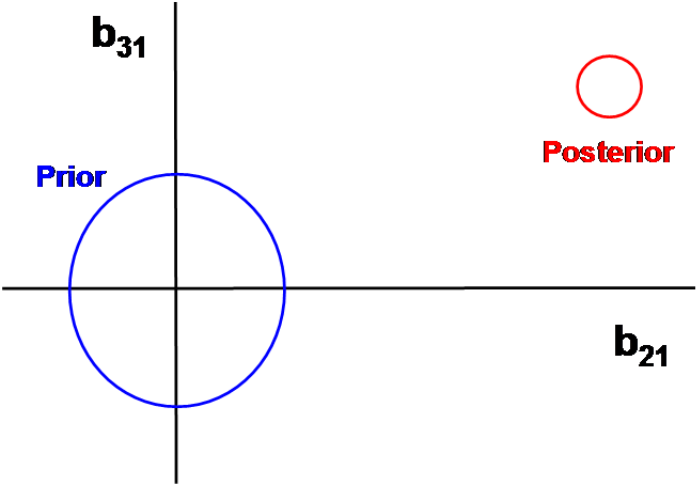
Small KL



Medium KL



Big KL



Complexity

Model complexity will tend to increase with the number of parameters N_w.

For the parameters we have

$$Complexity(m) = \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|}$$

where

$$e_w = m_w - \mu_w$$

But this will only be the case if these extra parameters diverge from their prior values and have smaller posterior (co)variance than prior (co)variance.

Complexity

In the limit that the posterior equals the prior ($e_w = 0, C_w = S_w$), the complexity term equals zero.

$$Complexity(m) = \frac{1}{2} e_w^T C_w^{-1} e_w + \frac{1}{2} \log \frac{|C_w|}{|S_w|}$$

Because the determinant of a matrix corresponds to the volume spanned by its eigenvectors, the last term gets larger and the model evidence smaller as the posterior volume, $|S_w|$, reduces in proportion to the prior volume, $|C_w|$.

Models for which parameters have to specified precisely (small posterior volume) are brittle. They are not good models (complexity is high).

Correlated Parameters

Other factors being equal, models with strong correlation in the posterior are not good models.

For example, given a model with just two parameters the determinant of the posterior covariance is given by

$$|S_w| = (1 - r^2)\sigma_{w_1}^2 \sigma_{w_2}^2$$

where r is the posterior correlation, σ_{w_1} and σ_{w_2} are the posterior standard deviations of the two parameters.

For the case of two parameters having a similar effect on model predictions the posterior correlation will be high, therefore implying a large complexity penalty.

Bayesian Information Criterion

A simple approximation to the log model evidence is given by the Bayesian Information Criterion (Schwarz, 1978)

$$BIC = \log p(y|\hat{w}, m) - \frac{N_w}{2} \log N_y$$

where $\hat{w}$ are the estimated parameters, N_w is the number of parameters, and N_y is the number of data points.

There is a complexity penalty of $\frac{1}{2} \log N_y$ for each parameter.

An Information Criterion

An alternative approximation is Akaike's Information Criterion or 'An Information Criterion' (AIC) - Akaike (1973)

$$AIC = \log p(y|\hat{w}, m) - N_w$$

There is a complexity penalty of 1 for each parameter.

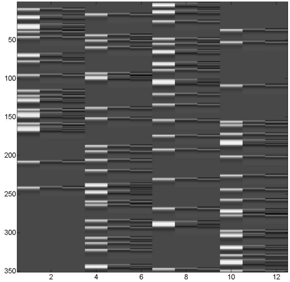
AIC and BIC are attractive because they are so easy to implement. They are also easily applied to nonlinear models.

Linear model of fMRI data

The following example compares AIC, BIC and Bayesian evidence criteria. Full details are given in Penny (2011). We use a linear model

$$y = Xw + e$$

with design matrix from Henson et al (2002).



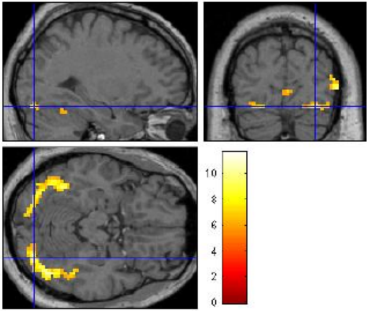
See also SPM Manual. We considered 'full' models (with 12 regressors) and 'nested' models (with last 9 only).

fMRI example

Likelihood

$$p(y|w) = \mathrm{N}(y; Xw, C_y)$$

where $C_y = \sigma_e^2 I_{N_y}$.



Parameters were drawn from the prior

$$p(w) = \mathrm{N}(w; \mu_w, C_w)$$

with $\mu_w = 0$ and $C_w = \sigma_p^2 I_p$ with set σ_p to correspond to the magnitude of coefficients in a face responsive area.

fMRI example

The parameter σ_e (observation noise SD) was set to give a range of SNRs where

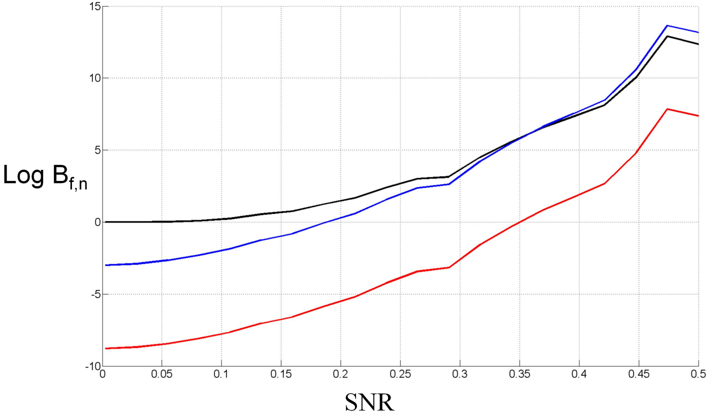
$$SNR = \frac{std(g)}{\sigma_e}$$

and $g = Xw$ is the signal. For each SNR we generated 100 data sets.

We first look at model comparison behaviours when the true model is full. For each generated data set we fitted both full and nested models and computed the log Bayes factor. We then averaged this over runs.

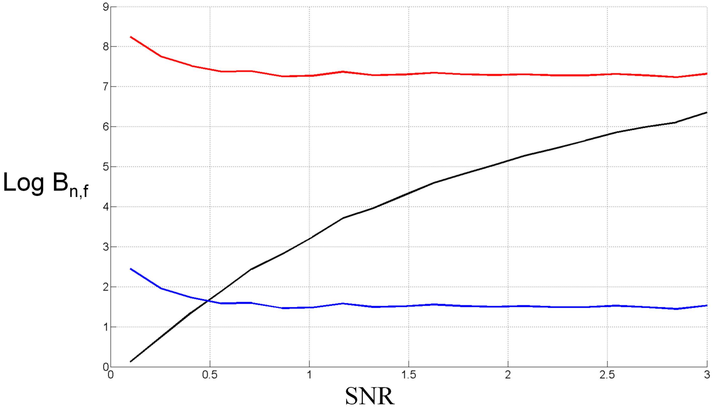
True Model: Full GLM

Log Bayes factor of full versus nested model versus the signal to noise ratio, SNR, when true model is the full GLM for Bayesian Log Evidence (black), AIC (blue) and BIC (red).



True Model: Nested GLM

Log Bayes factor of nested versus full model versus the signal to noise ratio, SNR, when true model is the nested GLM for Bayesian Log Evidence (black), AIC (blue) and BIC (red).



Evidence versus AIC, BIC

We have seen that Bayesian model evidence is more accurate than AIC or BIC.

Similar results have been found for other types of models

- Nonlinear autoregressive models (Roberts and Penny, 2002)
- Hidden Markov Models (Valente and Wellekens)
- Dynamic Causal Models (Penny, 2011)
- Graphical Models (Beal, 2003)

Bayes versus classical inference

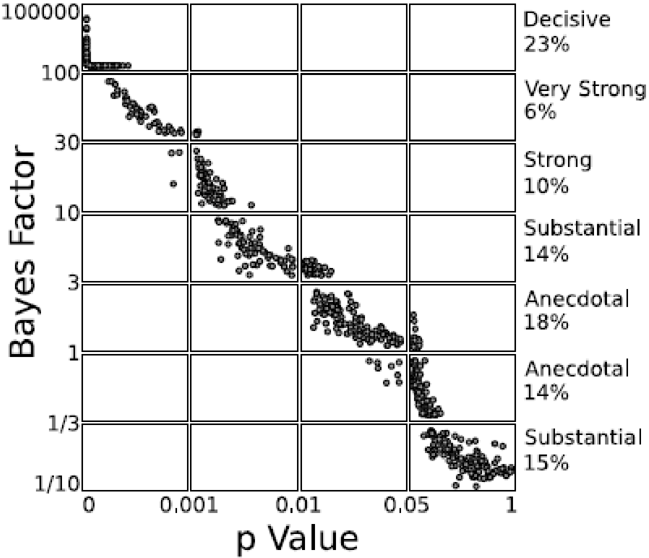
For nested model comparisons in linear models one can use classical inference based on F-tests.

Bayesian inference has three advantages over classical inference here

- The models do not have to be nested. The regressors in each of the models can be completely different.
- The 'null model' can be accepted if the Bayes factor is sufficiently high. In classical inference you can never accept the null.
- You can compare as many models as you like.

See Dienes (2011) for benefits over classical inference.

Bayes versus classical two-sample t-tests



Wetzels et al. Persp. Psych. Science, 2011.

Model evidence in data space

There is an alternative form of the model evidence which is useful if the dimension of the data points d is less than that of the regression coefficients p. This is not the case for most regression models but is the case, for example, for MEG source reconstruction.

Given the linear model $y = Xw + e$ with prior mean and covariance μ_w and C_w on the regression coefficients and observation noise covariance C_y the mean and covariance of the data are

$$\begin{aligned} m_d &= X\mu_w \\ C_d &= XC_wX^T + C_y \end{aligned}$$

Model evidence in data space

The log model evidence is then just the log of the probability of the data under the Gaussian density with the above mean and covariance

$$L = -\frac{d}{2}\log 2\pi - \frac{1}{2}\log|C_d| - \frac{1}{2}e_d^T C_d^{-1} e_d$$

and the prediction errors e_d are

$$e_d = y - m_d$$

The above expression requires inversion and determinants of $d \times d$ matrices rather than $p \times p$.

Nonlinear classifiers

Nonlinear classification with Multi-Layer Perceptrons (Penny and Roberts, 1998)

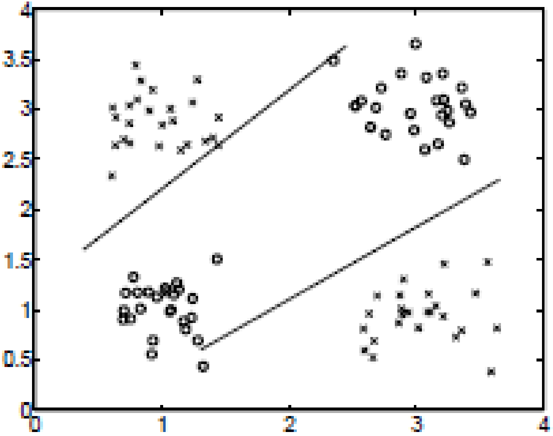
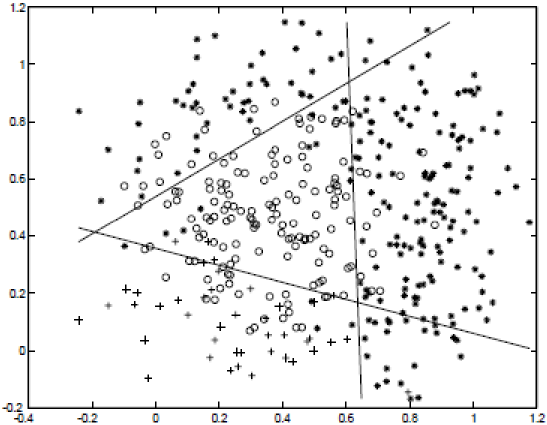


Fig. 1. XOR data: the solid lines show the orientation of each hidden unit in a two-hidden-unit MLP.

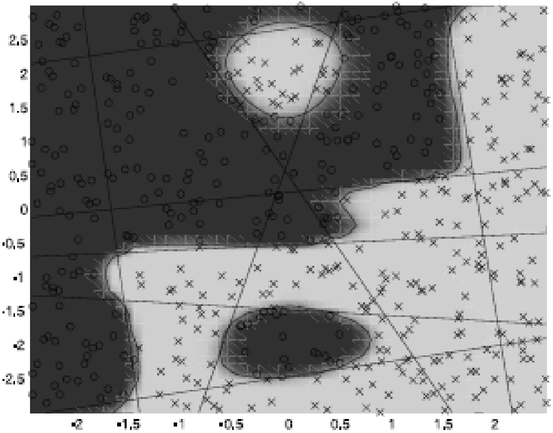
Nonlinear classifiers

Nonlinear classification with Multi-Layer Perceptrons



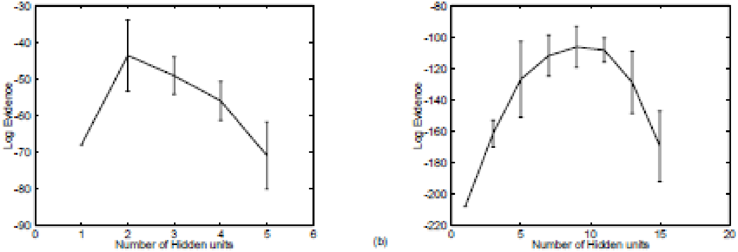
Nonlinear classifiers

Nonlinear classification with Multi-Layer Perceptrons



Nonlinear classifiers

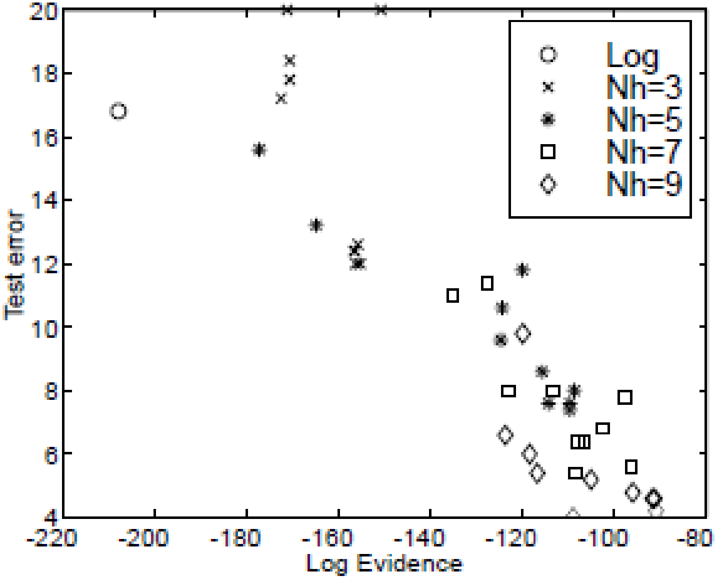
How many hidden units should we use in our MLP ?



For nonlinear models we can only approximate the model evidence using eg. a Laplace approximation (Bishop, 2006).

We can also use Cross-Validation (CV) to do this. Does CV agree with model evidence ?

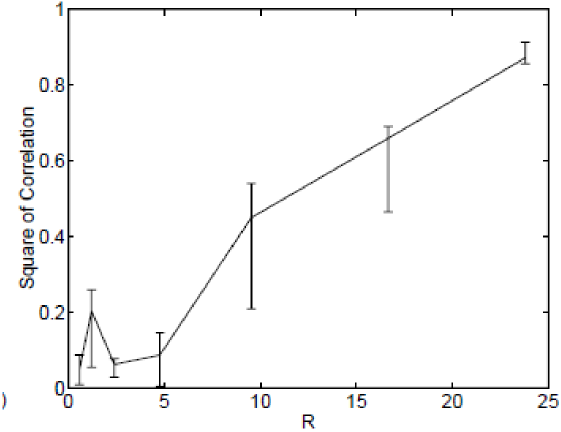
Cross validation versus model evidence



They are highly correlated.

Nonlinear classifiers

Correlation between CV and evidence versus R, the proportion of data points to model parameters.



References

H Akaike (1973) Information measures and model selection. Bull. Inst. Int. Stat 50, 277-290.

C. Bishop (2006) Pattern Recognition and Machine Learning. Springer.

Z. Dienes (2011) Bayesian versus orthodox statistics: which side are you on ? Perspectives on Psychological Science 6(3):274-290.

A. Gelman et al. (1995) Bayesian Data Analysis. Chapman and Hall.

R. Henson et al (2002) Face repetition effects in implicit and explicit memory tests as measured by fMRI. Cerebral Cortex, 12, 178-186.

W. Penny (2011) Comparing Dynamic Causal Models using AIC, BIC and Free Energy. Neuroimage Available online 27 July 2011.

A Raftery (1995) Bayesian model selection in social research. In Marsden, P (Ed) Sociological Methodology, 111-196, Cambridge.

S. Roberts and W. Penny (2002). Variational Bayes for generalised autoregressive models. IEEE transactions on signal processing. 50(9), 2245-2257.

G. Schwarz (1978) Estimating the dimension of a model. Ann. Stat. 6, 461-464.

D. Valente and C. Wellekens (2004) Scoring unknown speaker clustering: VB versus BIC. ICSLP 2004, Korea.