Advanced Data Analytics
Lecture week 6: Evidence procedure

Ian T. Nabney

University of BRISTOL

Overview

- Understand approximations involved in evidence procedure
- Application of evidence procedure to Bayesian linear regression
- Application of evidence procedure to PCA

Further reading: Bishop sections 3.5 and 12.2.3. Other models can be found in sections 4.4, 4.5 and 5.7.

Evidence approximation

- Recall the linear basis function model

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

  where the probabilistic form has a Gaussian noise model with zero mean and inverse variance β and a weight prior with zero mean and spherical inverse variance α.
- In a fully Bayesian treatment of the linear basis function model, we would introduce prior distributions over the hyperparameters α and β and make predictions by marginalizing with respect to these hyperparameters as well as with respect to the parameters $\mathbf{w}$.
- However, although we can integrate analytically over either $\mathbf{w}$ or over the hyperparameters, the complete marginalization over all of these variables is analytically intractable.
- In the evidence approximation we set the hyperparameters to specific values determined by maximizing the marginal likelihood function obtained by first integrating over the parameters $\mathbf{w}$.

Evidence procedure framework

- If we introduce hyperpriors over α and β, the predictive distribution is obtained by marginalizing over $\mathbf{w}$, α and β so that

$$p(t|\mathbf{t}) = \iiint p(t|\mathbf{w}, \beta)p(\mathbf{w}|\mathbf{t}, \alpha, \beta)p(\alpha, \beta|\mathbf{t})\, \mathrm{d}\mathbf{w}\mathrm{d}\alpha\mathrm{d}\beta \tag{1}$$

- If the posterior distribution $p(\alpha, \beta|\mathbf{t})$ is sharply peaked around values $\widehat{\alpha}$ and $\widehat{\beta}$, then the predictive distribution is obtained simply by marginalizing over $\mathbf{w}$ in which α and β are fixed to the values $\widehat{\alpha}$ and $\widehat{\beta}$, so that

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \widehat{\alpha}, \widehat{\beta}) = \int p(t|\mathbf{w}, \widehat{\beta})p(\mathbf{w}|\mathbf{t}, \widehat{\alpha}, \widehat{\beta})\, \mathrm{d}\mathbf{w}. \tag{2}$$

The posterior distribution for α and β is given by

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta)p(\alpha, \beta). \tag{3}$$

If the prior is relatively flat, then in the evidence framework the values of $\widehat{\alpha}$ and $\widehat{\beta}$ are obtained by maximizing the marginal likelihood function $p(\mathbf{t}|\alpha, \beta)$.

Evaluating the evidence function

- We can write the evidence function in the form

$$p(\mathbf{t}|\alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \left(\frac{\alpha}{2\pi}\right)^{M/2} \int \exp\{-E(\mathbf{w})\}\,\mathrm{d}\mathbf{w} \tag{4}$$

where M is the dimensionality of $\mathbf{w}$, and we have defined

$$\begin{aligned} E(\mathbf{w}) &= \beta E_D(\mathbf{w}) + \alpha E_W(\mathbf{w}) \\ &= \frac{\beta}{2}\|\mathbf{t} - \boldsymbol{\Phi}\mathbf{w}\|^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}. \end{aligned} \tag{5}$$

- Introduce

$$\mathbf{A} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} = \nabla\nabla E(\mathbf{w}) \tag{6}$$

A is the matrix of second derivatives of the error function, the Hessian.
- The log of the marginal likelihood in the form

$$\ln p(\mathbf{t}|\alpha, \beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi) \tag{7}$$

which is the required expression for the evidence function.

Maximising the evidence function: α

- Defining the following eigenvector equation

$$\left(\beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}\right) \mathbf{u}_i = \lambda_i \mathbf{u}_i. \tag{8}$$

Then $\mathbf{A}$ has eigenvalues $\alpha + \lambda_i$.

- Now consider the derivative of the term involving $\ln |\mathbf{A}|$ in (7) with respect to α. We have

$$\frac{d}{d\alpha} \ln |\mathbf{A}| = \frac{d}{d\alpha} \ln \prod_i (\lambda_i + \alpha) = \frac{d}{d\alpha} \sum_i \ln(\lambda_i + \alpha) = \sum_i \frac{1}{\lambda_i + \alpha}. \tag{9}$$

Thus the stationary points of (7) with respect to α satisfy

$$0 = \frac{M}{2\alpha} - \frac{1}{2}\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N - \frac{1}{2}\sum_i \frac{1}{\lambda_i + \alpha}. \tag{10}$$

Write

$$\gamma = M - \frac{1}{\lambda_i + \alpha} = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}. \tag{11}$$

- So the value of α that maximizes the marginal likelihood satisfies (11)

$$\alpha = \frac{\gamma}{\mathbf{m}_N^{\mathrm{T}}\mathbf{m}_N}. \tag{12}$$

This is an implicit solution for α not only because γ depends on α, but also because the mode $\mathbf{m}_N$ of the posterior distribution depends on the choice of α.

Maximising the evidence function: β

- The eigenvalues λ_i are proportional to β, and hence $d\lambda_i/d\beta = \lambda_i/\beta$ giving

$$\frac{d}{d\beta}\ln|\mathbf{A}| = \frac{d}{d\beta}\sum_i \ln(\lambda_i + \alpha) = \frac{1}{\beta}\sum_i \frac{\lambda_i}{\lambda_i + \alpha} = \frac{\gamma}{\beta}. \tag{13}$$

- The stationary point of the marginal likelihood therefore satisfies

$$0 = \frac{N}{2\beta} - \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2 - \frac{\gamma}{2\beta} \tag{14}$$

and rearranging we obtain (15)

$$\frac{1}{\beta} = \frac{1}{N-\gamma}\sum_{n=1}^{N}\left\{t_n - \mathbf{m}_N^{\mathrm{T}}\phi(\mathbf{x}_n)\right\}^2. \tag{15}$$

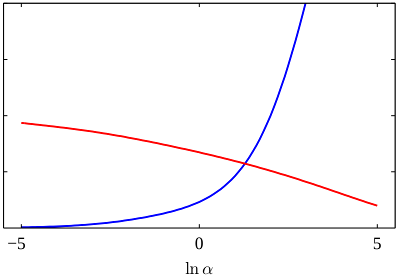
Again, this is an implicit solution for β.

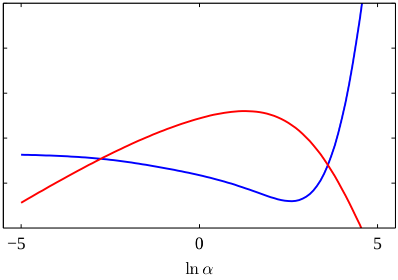
Evidence procedure algorithm

1. Compute the eigenvalues of $\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}$ (note that it is fixed).
2. Initialise α and β.
3. Compute γ γ, with (11)
4. Compute $\mathbf{m}_N$, which is given by (11). These values are then used to re-estimate α using (12).
5. Use the current value of β to calculate $\mathbf{m}_N$ and γ and then re-estimate β using (15)
6. Iterate from step 3 until convergence.

University of BRISTOL

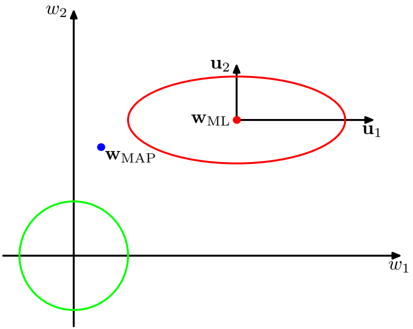
Evidence procedure results





Density contours

- Contours of the likelihood function (red) and the prior (green) in which the axes in parameter space have been rotated to align with the eigenvectors $\mathbf{u}_i$ of the Hessian.
- For $\alpha = 0$, the mode of the posterior is given by the maximum likelihood solution $\mathbf{w}_{\mathrm{ML}}$, whereas for nonzero α the mode is at $\mathbf{w}_{\mathrm{MAP}} = \mathbf{m}_N$. In the direction w_1 the eigenvalue λ_1 is small compared with α and so the quantity $\lambda_1/(\lambda_1 + \alpha)$ is close to zero, and the corresponding MAP value of w_1 is also close to zero.
- By contrast, in the direction w_2 the eigenvalue λ_2 is large compared with α and so the quantity $\lambda_2/(\lambda_2 + \alpha)$ is close to unity, and the MAP value of w_2 is close to its maximum likelihood value.

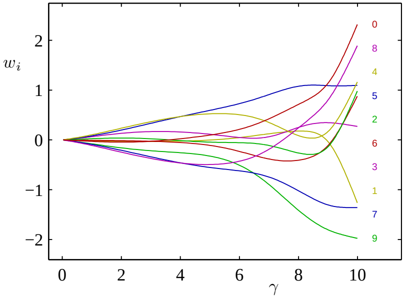


Effective number of parameters

- Because $\beta \mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi}$ is a positive definite matrix, it will have positive eigenvalues, and so the ratio $\lambda_i / (\lambda_i + \alpha)$ will lie between 0 and 1 and so $0 \leqslant \gamma \leqslant M$.
- For directions in which $\lambda_i \gg \alpha$, the corresponding parameter w_i will be close to its maximum likelihood value, and the ratio $\lambda_i / (\lambda_i + \alpha)$ will be close to 1. Such parameters are called well determined because their values are tightly constrained by the data.
- Conversely, for directions in which $\lambda_i \ll \alpha$, the corresponding parameters w_i will be close to zero, as will the ratios $\lambda_i / (\lambda_i + \alpha)$. These are directions in which the likelihood function is relatively insensitive to the parameter value and so the parameter has been set to a small value by the prior.
- The quantity γ therefore measures the effective total number of well determined parameters.

Effective number of parameters and α

- See how the parameter α controls the magnitude of the 10 parameters $\{w_i\}$, by plotting the individual parameters versus the effective number γ of parameters.
- The hyperparameter α is varied in the range $0 \leqslant \alpha \leqslant \infty$ causing γ to vary in the range $0 \leqslant \gamma \leqslant M$.



Laplace approximation

- For linear basis models, the posterior distribution over $\mathbf{w}$ is Gaussian.
- For nonlinear models, such as neural networks, this will no longer be the case. For such models we can use the Laplace approximation which is based on a local Gaussian approximation to the true posterior, and combine this with a local linear approximation to the model function.
- The Gaussian is fitted to the peak of the distribution (its mode) with variance given by the curvature (second derivative or Hessian) at that peak.
- For the linear model discussed, the posterior distribution is already Gaussian and so the Laplace approximation is exact.
- We can apply this method to other models, such as logistic regression (Sections 4.4, 4.5) and neural networks (Section 5.7).

Dimension of PCA

- So far in our discussion of PCA, we have assumed that the value M for the dimensionality of the principal subspace is given. In practice, we must choose a suitable value according to the application.
- One approach is to plot the eigenvalue spectrum for the data set and look to see if the eigenvalues naturally form two groups comprising a set of small values separated by a significant gap from a set of relatively large values, indicating a natural choice for M. In practice, such a gap is often not seen.
- Because the probabilistic PCA model has a well-defined likelihood function, we could employ cross-validation to determine the value of dimensionality by selecting the largest log likelihood on a validation data set. Such an approach, however, can become computationally costly.
- It is also infeasible if we consider a probabilistic mixture of PCA models in which we seek to determine the appropriate dimensionality separately for each component in the mixture.

Bayesian PCA

- Given that we have a probabilistic formulation of PCA, it seems natural to seek a Bayesian approach to model selection. To do this, we need to marginalize out the model parameters $\boldsymbol{\mu}$, $\mathbf{W}$, and σ^2 with respect to appropriate prior distributions.
- This can be done by using a variational framework to approximate the analytically intractable marginalizations.
- Here we consider a simpler approach introduced by Minka based on the evidence approximation, which is appropriate when the number of data points is relatively large and the corresponding posterior distribution is tightly peaked.

Priors and optimisation

- We make a specific choice of prior over $\mathbf{W}$ that allows surplus dimensions in the principal subspace to be pruned out of the model. This corresponds to ARD.
- We define an independent Gaussian prior over each column of $\mathbf{W}$, which represent the vectors defining the principal subspace. Each such Gaussian has an independent variance governed by a precision hyperparameter α_i so that

$$p(\mathbf{W}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \left(\frac{\alpha_i}{2\pi}\right)^{D/2} \exp\left\{-\frac{1}{2}\alpha_i \mathbf{w}_i^{\mathrm{T}} \mathbf{w}_i\right\} \tag{16}$$

  where $\mathbf{w}_i$ is the i^{th} column of $\mathbf{W}$.
- The values for α_i are found iteratively by maximizing the marginal likelihood function in which $\mathbf{W}$ has been integrated out. As a result of this optimization, some of the α_i may be driven to infinity, with the corresponding parameters vector $\mathbf{w}_i$ being driven to zero (the posterior distribution becomes a delta function at the origin) giving a sparse solution.

Bayesian PCA algorithm

- The values of α_i are re-estimated during training by maximizing the log marginal likelihood given by

$$p(\mathbf{X}|\boldsymbol{\alpha}, \boldsymbol{\mu}, \sigma^2) = \int p(\mathbf{X}|\mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{W}|\boldsymbol{\alpha}) \, \mathrm{d}\mathbf{W} \tag{17}$$

  For simplicity we also treat $\boldsymbol{\mu}$ and σ^2 as parameters to be estimated, rather than defining priors over these parameters.

- Because this integration is intractable, we make use of the Laplace approximation. The re-estimation equations obtained by maximizing the marginal likelihood with respect to α_i take the simple form

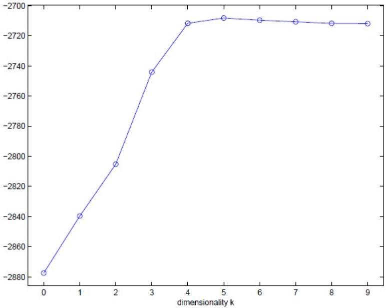
$$\alpha_i^{\mathrm{new}} = \frac{D}{\mathbf{w}_i^{\mathrm{T}} \mathbf{w}_i} \tag{18}$$

  which follows from (12), noting that the dimensionality of $\mathbf{w}_i$ is D.

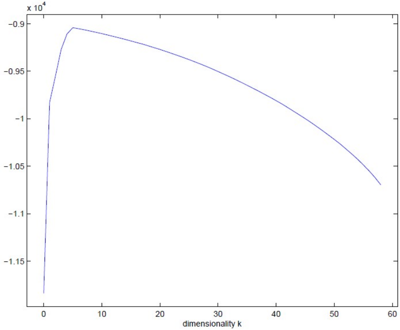
- These re-estimations are interleaved with the EM algorithm updates for determining $\mathbf{W}$ and σ^2.

University of BRISTOL

Experimental results



Data-rich case $N \gg D$ is generated from a 10-dimensional Gaussian with variance in 5 directions given by [10 8 6 4 2] and variance 1 in the remaining 5 directions.



Data-rich case $N \gg D$ is generated from a 100-dimensional Gaussian with variance in 5 directions given by [10 8 6 4 2] and variance 1 in the remaining 95 directions.

University of BRISTOL

- Understand approximations involved in evidence procedure
- Application of evidence procedure to Bayesian linear regression
- Application of evidence procedure to PCA

Further reading: Bishop sections 3.5 and 12.2.3. Other models can be found in sections 4.4, 4.5 and 5.7.