11

Bayesian inference of polynomials

The following example will be discussed as a computer exercise in the lab course.

Suppose, a detector reports time ordered and independent measurements d_i recorded at times t_i. Every datum has a Gaussian uncertainty of variance σ_i^2. We are interested in finding a time-varying signal

$$s(t_i) = s_i = \sum_{l=0}^{M} \alpha_l \, t_i^l \tag{11.1}$$

which is a polynomial of unknown order M and coefficients α_l. You may assume that you have prior knowledge on the coefficients α_l specified by a mean μ_l^α and uncertainty $(\sigma_l^{\alpha_l})^2$.

The likelihood distribution of the data given the unknown polynomial coefficients for a model of order M can be written as:

$$\begin{aligned} \pi(d_0, ..., d_N | \alpha_0, ..., \alpha_M, M) &= \pi(\{d_i\}|\{\alpha_l\}) \\ &= \prod_i \frac{\mathrm{e}^{-\frac{1}{2}\frac{\left(d_i - \sum_l \alpha_l t_i^l\right)^2}{\sigma_i^2}}}{\sqrt{2\,\pi\,\sigma_i^2}} \\ &= \frac{\mathrm{e}^{-\frac{1}{2}\sum_i \frac{\left(d_i - \sum_l \alpha_l t_i^l\right)^2}{\sigma_i^2}}}{\prod_i \sqrt{2\,\pi\,\sigma_i^2}} \end{aligned} \tag{11.2}$$

We may also have a priori knowledge on the mean μ_l^α and variance $(\sigma_l^\alpha)^2$ of the polynomial coefficients α. The maximum entropy prior is then given

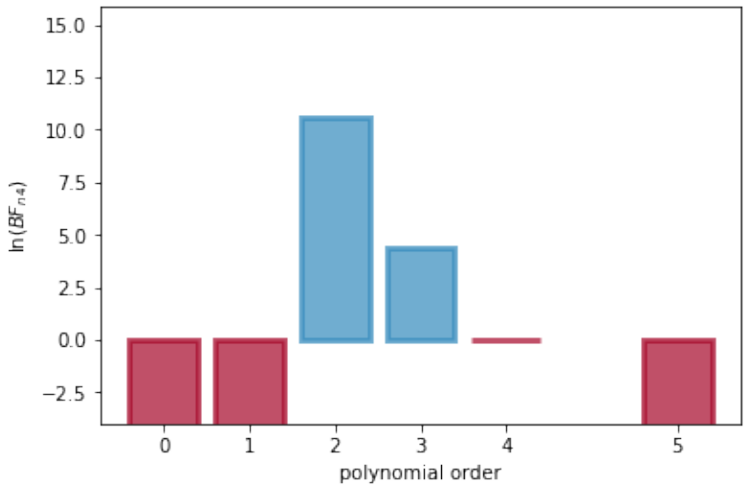

Figure 11.1: The plot shows the comparison of Bayes factors for 6 models of different polynomial order. The model with polynomial order $M = 2$ provides the optimal explanation of observed data.

by a multivariate normal distribution:

$$\pi(\{\alpha_l|M\}) \quad = \quad \frac{\mathrm{e}^{-\frac{1}{2}\sum_l \frac{(\alpha_l-\mu_l^\alpha)^2}{(\sigma_l^\alpha)^2}}}{\prod_l \sqrt{2\,\pi\,(\sigma_l^\alpha)^2}} \tag{11.3}$$

In order to compare models of different polynomial orders we need to estimate the evidence, which, in the following, will be derived from the joint distribution of data and polynomial coefficients:

$$\pi(\{\alpha_l|M\})\,\pi(\{d_i\}|\{\alpha_l\},M) = \frac{\mathrm{e}^{-\frac{1}{2}\sum_l \frac{(\alpha_l-\mu_l^\alpha)^2}{(\sigma_l^\alpha)^2}}}{\prod_l \sqrt{2\,\pi\,(\sigma_l^\alpha)^2}} \frac{\mathrm{e}^{-\frac{1}{2}\sum_i \frac{\left(d_i-\sum_l \alpha_l t_i^l\right)^2}{\sigma_i^2}}}{\prod_i \sqrt{2\,\pi\,\sigma_i^2}} \tag{11.4}$$

The logarithm of the joint distribution is given as:

$$\begin{aligned}
\ln\left(\pi(\{\alpha_l|M\})\,\pi(\{d_i\}|\{\alpha_l\},M)\right) &= -\frac{1}{2}\sum_l \frac{\left(\alpha_l-\mu_l^\alpha\right)^2}{(\sigma_l^\alpha)^2} - \frac{1}{2}\sum_i \frac{\left(d_i-\sum_l \alpha_l t_i^l\right)^2}{\sigma_i^2} \\
&\quad -\frac{1}{2}\sum_l \ln\left((\sigma_l^\alpha)^2\right) - \frac{1}{2}\sum_i \ln\left(\sigma_i^2\right) \\
&= -\frac{1}{2}\left[\sum_l\sum_m \alpha_l\left(\sum_i \frac{t_i^l\,t_i^m}{\sigma_i^2} + \frac{\delta_{lm}^K}{(\sigma_l^\alpha)^2}\right)\alpha_m\right. \\
&\quad -2\sum_l \alpha_l\left(\frac{\mu_l^\alpha}{(\sigma_l^\alpha)^2} + \sum_i \frac{t_i^l\,d_i}{\sigma_i^2}\right) \\
&\quad +\sum_l \frac{\left(\mu_l^\alpha\right)^2}{(\sigma_l^\alpha)^2} + \sum_l \ln\left((\sigma_l^\alpha)^2\right) \\
&\quad \left.+\sum_i \frac{d_i^2}{\sigma_i^2} + \sum_i \ln\left(\sigma_i^2\right)\right]\,,
\end{aligned} \tag{11.5}$$

which can be simplified by introducing the inverse posterior covariance matrix:

$$\left(D_{lm}^P\right)^{-1} = \frac{\delta_{lm}^K}{(\sigma_l^\alpha)^2} + \sum_i \frac{t_i^l\,t_i^m}{\sigma_i^2} \tag{11.6}$$

where δ_{lm}^K is the Kronecker Delta. Furthermore let:

$$A_l = \frac{\mu_l^\alpha}{(\sigma_l^\alpha)^2} + \sum_i \frac{t_i^l\,d_i}{\sigma_i^2}\,, \tag{11.7}$$

Then the posterior mean can be expressed as:

$$\mu_l^P = \sum_m D_{lm}^P A_m\,, \tag{11.8}$$

which amounts to solving a linear system of equations. Then:

$$\begin{aligned}\ln\left(\pi(\{\alpha_l|M\})\,\pi(\{d_i\}|\{\alpha_l\},M)\right) &= -\frac{1}{2}\left[\sum_{lm}\left(\alpha_l-\mu_l^P\right)\left(D_{lm}^P\right)^{-1}\left(\alpha_m-\mu_m^P\right)\right.\\ &\quad -\sum_{lm}\left(\mu_l^P\right)\left(D_{lm}^P\right)^{-1}\left(\mu_m^P\right)\\ &\quad +\sum_l\frac{\left(\mu_l^\alpha\right)^2}{(\sigma_l^\alpha)^2}+\sum_l\ln\left((\sigma_l^\alpha)^2\right)\\ &\quad \left.+\sum_i\frac{d_i^2}{\sigma_i^2}+\sum_i\ln\left(\sigma_i^2\right)\right],\end{aligned} \tag{11.9}$$

By exponentiating the logarithm we obtain:

$$\begin{aligned}\pi(\{\alpha_l|M\})\,\pi(\{d_i\}|\{\alpha_l\},M) &= \mathrm{e}^{-\frac{1}{2}\sum_{lm}\left(\alpha_l-\mu_l^P\right)\left(D_{lm}^P\right)^{-1}\left(\alpha_m-\mu_m^P\right)}\\ &\quad\times\mathrm{e}^{\frac{1}{2}\sum_{lm}\left(\mu_l^P\right)\left(D_{lm}^P\right)^{-1}\left(\mu_m^P\right)}\\ &\quad\times\prod_l\frac{\mathrm{e}^{-\frac{1}{2}\frac{\left(\mu_l^\alpha\right)^2}{(\sigma_l^\alpha)^2}}}{\sqrt{2\pi(\sigma_l^\alpha)^2}}\\ &\quad\times\prod_i\frac{\mathrm{e}^{-\frac{1}{2}\frac{d_i^2}{\sigma_i^2}}}{\sqrt{2\pi\sigma_i^2}}\end{aligned} \tag{11.10}$$

The evidence can then be obtained by marginalizing over the polynomial

coefficients $\{\alpha_l\}$:

$$\begin{aligned}\pi(\{d_i\}|M) &= \int \mathrm{d}\{\alpha_l\}\, \pi(\{\alpha_l|M\})\, \pi(\{d_i\}|\{\alpha_l\}, M) \\ &= \sqrt{\det(2\,\pi D^P)} \\ &\quad \times \mathrm{e}^{\frac{1}{2}\sum_{lm}(\mu_l^P)\left(D_{lm}^P\right)^{-1}(\mu_m^P)} \\ &\quad \times \prod_l \frac{\mathrm{e}^{-\frac{1}{2}\frac{(\mu_l^\alpha)^2}{(\sigma_l^\alpha)^2}}}{\sqrt{2\pi(\sigma_l^\alpha)^2}} \\ &\quad \times \prod_i \frac{\mathrm{e}^{-\frac{1}{2}\frac{d_i^2}{\sigma_i^2}}}{\sqrt{2\pi\sigma_i^2}}\end{aligned} \tag{11.11}$$

The posterior distribution then yields a simple multivariate normal distribution:

$$\begin{aligned}\pi(\{\alpha_l\}|\{d_i\}, M) &= \pi(\{\alpha_l|M\})\, \frac{\pi(\{d_i\}|\{\alpha_l\}, M)}{\pi(\{d_i\}|M)} \\ &\quad \frac{\mathrm{e}^{-\frac{1}{2}\sum_{lm}\left(\alpha_l-\mu_l^P\right)\left(D_{lm}^P\right)^{-1}\left(\alpha_m-\mu_m^P\right)}}{\sqrt{\det(2\,\pi D^P)}}\,.\end{aligned} \tag{11.12}$$

This Gaussian posterior is completely determined by the posterior mean μ_m^{P} of eq 11.8 and covariance matrix D_{lm}^P 11.6. In particular the mean μ_m^{P} corresponds to the Maximum A Posteriori (MAP) estimate for the coefficients α_l. A simple posterior prediction for the time dependent signal can then be obtained by using the polynomial equation:

$$s(t) = \sum_m \mu_m^{\mathrm{P}}\, t^m\,. \tag{11.13}$$

For an illustration of posterior predicted results see Fig. 5.4 for analysis of six models of different polynomial order M.

Models of different polynomial order, e.g. $M = 2$ and $M = 4$ can then be selected by evaluating the Bayes factor:

$$BF_{24} = \frac{\pi(\{d_i\})|M = 2)}{\pi(\{d_i\})|M = 4)} \tag{11.14}$$

where $\pi(\{d_i\})|M = m)$ is given by eq 11.11, the evidence for a model of polynomial order m. For an illustration of Bayes factors for models of different polynomial order see Fig. 11.1.

12

A running example

For the sake of the lecture, and to contrast Bayesian from Frequentists results we will design a small standard example. In an experiment N independent measurements d_i have been taken. We assume the following data model:

$$d_i = x + \epsilon_i \,, \tag{12.1}$$

where x is the target or signal and ϵ_i is a normally distributed zero-mean normal noise contribution with variance σ_n^2.

The corresponding likelihood distribution for N independent measurements is:

$$\pi\left(d|x,\sigma_n^2\right) = \prod_{i=0}^{N} \frac{\mathrm{e}^{-\frac{1}{2}\frac{(x-d_i)^2}{\sigma_n^2}}}{\sqrt{2\pi\sigma_n^2}} = \frac{\mathrm{e}^{-\frac{1}{2}\sum_{i=0}^{N}\frac{(x-d_i)^2}{\sigma_n^2}}}{\left(2\pi\sigma_n^2\right)^{N/2}} \tag{12.2}$$

For various a priori arguments (e.g. x could indicate a physical mass) we will assume in our case that $x \in [0,\infty]$. In Bayesian statistics, these prior assumptions are made explicit by formulating a corresponding prior distribution $\pi(x)$. In our case the corresponding prior would be the Heaviside function:

$$\Theta(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

In a Bayesian context, the combination of the likelihood and the prior

gives the posterior distribution:

$$\begin{aligned}\pi\left(x|d,\sigma_n^2\right) &= \pi\left(x\right)\frac{\pi\left(d|x,\sigma_n^2\right)}{\pi\left(d\right)} \\ &\propto \Theta(x)\frac{\mathrm{e}^{-\frac{1}{2}\sum_{i=0}^{N}\frac{\left(x-d_i\right)^2}{\sigma_n^2}}}{\left(2\pi\sigma_n^2\right)^{N/2}} \\ &\propto \Theta(x)\frac{\mathrm{e}^{-\frac{N}{2\sigma_n^2}\left(x-\frac{\sum_{i=0}^{N}d_i}{N}\right)^2}}{\left(2\pi\sigma_n^2\right)^{N/2}} \end{aligned} \tag{12.3}$$

where we used simple algebraic transformations and we ignored several constant factors. The product of Heaviside function and Gaussian likelihood yields a truncated normal distribution. Let:

$$\Phi(x|\mu,\sigma^2) = \frac{1}{2}\left[1-\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \tag{12.4}$$

be the cumulative probability function of a normal distribution with mean μ and variance σ^2.

The properly normalized posterior distribution is then given by:

$$\pi\left(x|d,\sigma_n^2\right) = \Theta(x)\frac{\mathrm{e}^{-\frac{N}{2\sigma_n^2}\left(x-\frac{\sum_{i=0}^{N}d_i}{N}\right)^2}}{\frac{1}{2}\left[1-\mathrm{erf}\left(\frac{-\sum_{i=0}^{N}d_i}{\sqrt{2N\sigma_n^2}}\right)\right]} \tag{12.5}$$

12.0.1 Statistical Summaries for the running model

In Bayesian statistics, we do not report estimators, but we attempt to summarize the posterior distributions via a number of statistical summaries. This is because in Bayesian statistics we do not assume the parameter x to be fixed but to be a random variate. A random variate can only be fully described when reporting its probability distribution. Whether or not a few statistical summaries are sufficient to capture all details of the distribution has to be decided on a case by case basis. Correspondingly care has to be taken, when interpreting the random variate x based only on a finite number of possibly insufficient statistical summaries.

For the truncated Gaussian distribution we will report a few statistical summaries in the following:

1. **Mode a.k.a the Maximum A Posteriori (MAP) value**:

$$x_{\rm MAP} = \begin{cases} \frac{\sum_{i=0}^{N} d_i}{N}, & \text{if } \frac{\sum_{i=0}^{N} d_i}{N} \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

   Note, that $x_{\rm MAP}$ is an "unbiased estimator" in the Frequentist sense, i.e. it converges to the correct results in the large sample limit if $x \in [0, \infty]$.

2. **Mean**:

$$\langle x \rangle = \frac{\sum_{i=0}^{N} d_i}{N} + \frac{e^{-\frac{N}{2\sigma_n^2}\left(\frac{\sum_{i=0}^{N} d_i}{N}\right)^2}}{\frac{1}{2}\left[1 - \text{erf}\left(\frac{-\sum_{i=0}^{N} d_i}{\sqrt{2N\sigma_n^2}}\right)\right]} \frac{\sigma_n}{\sqrt{N}} \tag{12.6}$$

   Now, the mean is not an "unbiased estimator" in the Frequentist sense, since the second term will always introduce a slight bias. In particular, the second term will decay slower as the first term due to the $1/\sqrt{N}$ dependence that is why the bias term will always contribute.