given by $\mathrm{var}\, y(x) = \phi^{\mathrm{T}} \mathbf{A}^{-1} \phi$. Actually we have access to the full covariance information for the entire interpolant, not just the pointwise error bars. It is possible to visualise the *joint* error bars on the interpolant by making typical samples from the posterior distribution, performing a random walk around the posterior 'bubble' in parameter space [60, 64]. Figure 2.8 shows data set Y interpolated by three typical interpolants found by random sampling from the posterior distribution. These error bar properties are found under the assumption that the model is correct; so it is possible for the true interpolant to lie significantly outside the error bars of a poor model.

In this section Bayesian model comparison will be demonstrated first with models differing only in the number of free parameters (for example polynomials of different degrees), then with comparisons between models as disparate as splines, radial basis functions and feedforward neural networks. The characters of some of these models are illustrated in figure 2.9, which shows a typical sample from each. For each individual model, the value of α is optimised, and the evidence is evaluated by integrating over α using the gaussian approximation. All logarithms are to base e.

Legendre polynomials: Occam's razor for the number of basis functions

Figure 2.7a shows the evidence for Legendre polynomials of different degrees for data set X. The basis functions were chosen to be orthonormal on an interval enclosing the data, and a regulariser of the form $E_W = \sum \frac{1}{2} w_h^2$ was used.

Notice that an evidence maximum is obtained: beyond a certain number of terms, the evidence starts to decrease. This is the Bayesian Occam's razor at work. The additional terms make the model more powerful, able to make more predictions. This power is automatically penalised. Notice the characteristic shape of the 'Occam hill.' On the left, the hill is steep as the over–simple models fail to fit the data; the penalty for misfitting the data scales as N, the number of data measurements. The other side of the hill is much less steep; the log Occam factors here only scale as $k \log N$, where k is the number of parameters. We note in table 2.1 the value of the maximum evidence achieved by these models, and move on to alternative models.

The choice of orthonormal Legendre polynomials described above was motivated by a maximum entropy argument [21]. Models using other polynomial basis sets have also been tried. For less well motivated basis sets such as Hermite polynomials, it was found that the Occam factors were far bigger and the evidence was substantially smaller. If the size of the Occam factor increases rapidly with over–parameterisation, it is generally a sign that the space of alternative models is poorly matched to the problem.

Fixed radial basis functions

For a radial basis function or 'kernel' model, the basis functions are $\phi_h(x) = g((x - x_h)/r)/r$; here the x_h are equally spaced over the range of interest. I examine two choices of g: a gaussian and a Cauchy function, $1/1 + x^2$. We can quantitatively compare these alternative models of spatial correlation for any data set by evaluating the evidence. The regulariser is $E_W = \sum \frac{1}{2} w_h^2$. Note that this model includes one new free parameter, r; in these demonstrations this parameter has been set to its most probable value (*i.e.* the value which maximises the evidence). To penalise this free parameter an Occam factor is included, $\sqrt{2\pi} P(\log r) \Delta \log r$, where $\Delta \log r$ = posterior uncertainty in $\log r$, and $P(\log r)$ is the prior on $\log r$, which is subjective to a small degree (I used $P(\log r) = 1/(4 \pm 2)$). This radial basis function model is the same as the 'intrinsic correlation' model of Charter, Gull, Skilling and

so as to allow the data to assign a preference to the alternative models. Notice as pointed out earlier that this modern Bayesian framework includes no emphasis on defining the 'right' prior $\mathcal{R}$ with which we ought to interpolate. Rather, we invent as many priors (regularisers) as we want, and allow the data to tell us which prior is *most probable*. Having said this, experience recommends that the 'maximum entropy principle' and other respected guides should be consulted when inventing these priors (see [21], for example).

Evaluating the evidence for $\mathcal{H}$

As α and β vary, a single evidence maximum is obtained, at $\hat{\alpha}, \hat{\beta}$ (at least for quadratic E_D and E_W). The evidence maximum is often well approximated[12] by a separable gaussian, and differentiating (2.21) twice we obtain gaussian error bars for $\log \alpha$ and $\log \beta$:

$$(\Delta \log \alpha)^2 \simeq 2/\gamma$$
$$(\Delta \log \beta)^2 \simeq 2/(N-\gamma)$$

Putting these error bars into (2.26), we obtain the evidence.[13]

$$P(D|\mathcal{H}) \simeq P(D|\hat{\alpha},\hat{\beta},\mathcal{H}) P(\hat{\alpha},\hat{\beta}|\mathcal{H})\, 2\pi\, \Delta\log\alpha\, \Delta\log\beta \tag{2.27}$$

How is the prior $P(\hat{\alpha},\hat{\beta}|\mathcal{H})$ assigned? This is the first time in this chapter that we have met one of the infamous 'subjective priors' which are supposed to plague Bayesian methods. Here are some answers to this question. (a) Any other method of assigning a preference to alternatives must implicitly assign such priors. Bayesians adopt the healthy attitude of not sweeping them under the carpet. (b) With some thought, reasonable values can usually be assigned to subjective priors, and the degree of reasonable subjectivity in these assignments can be quantified. For example, a reasonable prior on an unknown standard deviation states that σ is unknown over a range of (3 ± 2) orders of magnitude. This prior contributes a subjectivity of about ± 1 to the value of the log evidence. This degree of subjectivity is often negligible compared to the log evidence differences. (c) In the noisy interpolation example, all models considered include the free parameters α and β. So in this chapter I do not need to assign a value to $P(\hat{\alpha},\hat{\beta}|\mathcal{H})$; I assume that it is a flat prior (flat over $\log\alpha$ and $\log\beta$, since α and β are scale parameters) which cancels out when we compare alternative interpolation models.

2.6 Demonstration

These demonstrations will use two one–dimensional data sets, in imitation of [60]. The first data set, 'X,' has discontinuities in derivative (figure 2.4), and the second is a smoother data set, 'Y' (figure 2.8). In all the demonstrations, β was not left as a free parameter, but was fixed to its known true value.

The Bayesian method of setting α, assuming a single model is correct, has already been demonstrated, and quantified error bars have been placed on the most probable interpolant (figure 2.4). The method of evaluating the error bars is to use the posterior covariance matrix of the parameters w_h, $\mathbf{A}^{-1}$, to get the variance on $y(x)$, which for any x is a linear function of the parameters, $y(x) = \sum_h \phi_h(x) w_h$. The error bars at a single point x are

[12] This approximation is valid when, in the spectrum of eigenvalues of $\beta\mathbf{B}$, the number of eigenvalues within e–fold of $\hat{\alpha}$ is $O(1)$.

[13] There are analytic methods for performing such integrals over β [9].

to is the χ^2 between the *inferred* interpolant and the data, $\chi^2_D = 2\beta E_D$. The 'discrepancy principle' of orthodox statistics states that the model parameters should be adjusted so as to make $\chi^2_D = N$. Work on un–regularised least squares regression suggests that we should estimate the noise level so as to set $\chi^2_D = N - k$, where k is the number of free parameters. Let us find out the opinion of Bayes' rule on this matter.

We differentiate the log evidence (2.21) with respect to β and obtain, setting the derivative to zero:

$$2\beta E_D = N - \gamma \tag{2.24}$$

Thus the most probable noise estimate, $\hat{\beta}$, does not satisfy $\chi^2_D = N$ or $\chi^2_D = N - k$; rather, $\chi^2_D = N - \gamma$. This Bayesian estimate of noise level naturally takes into account the fact that the parameters which have been determined by the data inevitably suppress some of the noise in the data, while the poorly measured parameters do not. Note that the value of χ^2_D only enters into the determination of β: misfit criteria have no role in the Bayesian choice of α [22].

In summary, at the optimum value of α and β, $\chi^2_W = \gamma$, $\chi^2_D = N - \gamma$. Notice that this implies that the total misfit $M = \alpha E_W + \beta E_D$ satisfies the simple equation $2M = N$.

The interpolant resulting from the Bayesian choice of α is illustrated by figure 2.4b. Figure 2.5b illustrates the functions involved with the Bayesian choice of α, and compares them with the 'test error' approach. Demonstration of the Bayesian choice of β is omitted, since it is straightforward; β is fixed to its true value for the demonstrations in this chapter. Inference of an input–dependent noise level $\beta(x)$ will be demonstrated in a future publication.

These results generalise to the case where there are two or more separate regularisers with independent regularising constants $\{\alpha_c\}$ [22]. In this case, each regulariser has a number of good parameter measurements γ_c associated with it. Multiple regularisers will be used for neural networks in chapter 3.

Finding the evidence maximum with a head on approach would involve evaluating $\det \mathbf{A}$ while searching over α, β; the above results (2.22,2.24) enable us to speed up this search (for example by the use of re–estimation formulae like $\alpha := \gamma / 2E_W$) and replace the evaluation of $\det \mathbf{A}$ by the evaluation of $\mathrm{Trace}\mathbf{A}^{-1}$. For large dimensional problems where this task is demanding, Skilling has developed methods for estimating $\mathrm{Trace}\mathbf{A}^{-1}$ statistically in k^2 time [62].

2.5 Model comparison

To rank alternative basis sets $\mathcal{A}$, noise models $\mathcal{N}$ and regularisers (priors) $\mathcal{R}$ in the light of the data, we examine the posterior probabilities for alternative models $\mathcal{H} = \{\mathcal{A}, \mathcal{N}, \mathcal{R}\}$:

$$P(\mathcal{H}|D) \propto P(D|\mathcal{H})P(\mathcal{H}) \tag{2.25}$$

The data–dependent term, the evidence for $\mathcal{H}$, appeared earlier as the normalising constant in (2.15), and it is evaluated by integrating the evidence for (α, β):

$$P(D|\mathcal{H}) = \int P(D|\alpha, \beta, \mathcal{H}) P(\alpha, \beta|\mathcal{H})\, d\alpha\, d\beta \tag{2.26}$$

Assuming that we have no reason to assign strongly differing priors $P(\mathcal{H})$, alternative models $\mathcal{H}$ are ranked just by examining the evidence. The evidence can also be compared with the evidence found by an equivalent Bayesian analysis of other learning and interpolation models

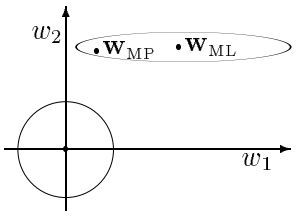

Figure 2.6: **Good and bad parameter measurements**
Let w_1 and w_2 be the components in parameter space in two directions parallel to eigenvectors of the data matrix $\mathbf{B}$. The circle represents the characteristic prior distribution for $\mathbf{w}$. The ellipse represents a characteristic contour of the likelihood, centred on the maximum likelihood solution $\mathbf{w}_{\rm ML}$. $\mathbf{w}_{\rm MP}$ represents the most probable parameter vector. w_1 is a direction in which λ_1 is small compared to α, *i.e.* the data have no strong preference about the value of w_1; w_1 is a poorly measured parameter, and the term $\frac{\lambda_1}{\lambda_1+\alpha}$ is close to zero. w_2 is a direction in which λ_1 is large; w_2 is well determined by the data, and the term $\frac{\lambda_2}{\lambda_2+\alpha}$ is close to one.

The quantity on the left is the dimensionless measure of the amount of structure introduced into the parameters by the data, *i.e.* how much the fitted parameters differ from their null value. It can be interpreted as the χ^2 of the parameters, since it is equal to $\chi^2_W = \sum w_i^2/\sigma_W^2$, with $\alpha = 1/\sigma_W^2$.

The quantity on the right of (2.22) is called the number of good parameter measurements, γ, and has value between 0 and k. It can be written in terms of the eigenvalues of $\beta\mathbf{B}$, λ_a, where the subscript a runs over the k eigenvectors. The eigenvalues of $\mathbf{A}$ are $\lambda_a + \alpha$, so we have:

$$\gamma = k - \alpha\,\mathrm{Trace}\mathbf{A}^{-1} = k - \sum_{a=1}^{k} \frac{\alpha}{\lambda_a + \alpha} = \sum_{a=1}^{k} \frac{\lambda_a}{\lambda_a + \alpha} \tag{2.23}$$

Each eigenvalue λ_a measures how strongly one parameter is determined by the data. The constant α measures how strongly the parameters are determined by the prior. The ath term $\gamma_a = \lambda_a/(\lambda_a + \alpha)$ is a number between 0 and 1 which measures the strength of the data relative to the prior (figure 2.6): the components of $\mathbf{w}_{\rm MP}$ are given by $\mathbf{w}_{{\rm MP}\,a} = \gamma_a \mathbf{w}_{{\rm ML}\,a}$.

A direction in parameter space for which λ_a is small compared to α does not contribute to the number of good parameter measurements. γ is thus a measure of the effective number of parameters which are well determined by the data. As $\alpha/\beta \to 0$, γ increases from 0 to k. The condition (2.22) for the most probable value of α can therefore be interpreted as an estimation of the variance σ_W^2 of the gaussian distribution from which the weights are drawn, based on γ effective samples from that distribution: $\sigma_W^2 = \sum w_i^2/\gamma$.

This concept is not only important for locating the optimum value of α: it is only the γ good parameter measurements which are expected to contribute to the reduction of the data misfit that occurs when a model is fitted to noisy data. In the process of fitting $\mathbf{w}$ to the data, it is unavoidable that some fitting of the model to noise will occur, because some components of the noise are indistinguishable from real data. Typically, one unit (χ^2) of noise will be fitted for every well–determined parameter. Poorly determined parameters are determined by the regulariser only, so they do not reduce χ^2_D in this way. We will now examine how this concept enters into the Bayesian choice of β.

Recall that the expectation of the χ^2 misfit between the true interpolant and the data is N. However we do not know the true interpolant, and the only misfit measure we have access

If the regulariser $\mathcal{R}$ is a quadratic functional (and the favourites are), then E_D and E_W are quadratic functions of $\mathbf{w}$, and we can evaluate Z_M exactly. Letting $\nabla\nabla E_W = \mathbf{C}$ and $\nabla\nabla E_D = \mathbf{B}$ then using $\mathbf{A} = \alpha\mathbf{C} + \beta\mathbf{B}$, we have:

$$M = M(\mathbf{w}_{\rm MP}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\rm MP})^{\rm T}\mathbf{A}(\mathbf{w} - \mathbf{w}_{\rm MP})$$

where $\mathbf{w}_{\rm MP} = \mathbf{A}^{-1}\mathbf{B}\mathbf{w}_{\rm ML}$. This means that Z_M is the gaussian integral:

$$Z_M = e^{-M_{\rm MP}}(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A} \tag{2.19}$$

In many cases where the regulariser is not quadratic (for example, entropy–based), this gaussian approximation is still servicable [22]. Thus we can write the log evidence for α and β as:

$$\log P(D|\alpha,\beta,\mathcal{H}) = -\alpha E_W^{\rm MP} - \beta E_D^{\rm MP} - \frac{1}{2}\log\det\mathbf{A} - \log Z_W(\alpha) - \log Z_D(\beta) + \frac{k}{2}\log 2\pi \tag{2.20}$$

The term $\beta E_D^{\rm MP}$ represents the misfit of the interpolant to the data. The three terms $-\alpha E_W^{\rm MP} - \frac{1}{2}\log\det\mathbf{A} - \log Z_W(\alpha)$ constitute the log of the 'Occam factor' penalising small values of α: the ratio $(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}/Z_W(\alpha)$ is the ratio of the posterior accessible volume in parameter space to the prior accessible volume, and the term $\alpha E_W^{\rm MP}$ measures how far $\mathbf{w}_{\rm MP}$ is from its null value. Figure 2.5a illustrates the behaviour of these various terms as a function of α for the same radial basis function model as illustrated in figure 2.4.

Now we could just proceed to evaluate the evidence numerically as a function of α and β, but a more deep and fruitful understanding of this problem is possible.

Properties of the evidence maximum

The maximum over α,β of $P(D|\alpha,\beta,\mathcal{H}) = Z_M(\alpha,\beta)/(Z_W(\alpha)Z_D(\beta))$ has some remarkable properties which give deeper insight into this Bayesian approach. The results of this section are useful both numerically and intuitively.

Following Gull [22], we transform to the basis in which the Hessian of E_W is the identity, $\nabla\nabla E_W = \mathbf{I}$. This transformation is simple in the case of quadratic E_W: rotate into the eigenvector basis of $\mathbf{C}$ and stretch the axes so that the quadratic form E_W becomes homogeneous. This is the natural basis for the prior. I will continue to refer to the parameter vector in this basis as $\mathbf{w}$, so from here on $E_W = \frac{1}{2}\sum w_i^2$. Using $\nabla\nabla M = \mathbf{A}$ and $\nabla\nabla E_D = \mathbf{B}$ as above, we differentiate the log evidence with respect to α and β so as to find the condition that is satisfied at the maximum. The log evidence, from (2.20), is:

$$\log P(D|\alpha,\beta,\mathcal{H}) = -\alpha E_W^{\rm MP} - \beta E_D^{\rm MP} - \frac{1}{2}\log\det\mathbf{A} + \frac{k}{2}\log\alpha + \frac{N}{2}\log\beta - \frac{N}{2}\log 2\pi \tag{2.21}$$

First, differentiating with respect to α, we need to evaluate $\frac{d}{d\alpha}\log\det\mathbf{A}$. Using $\mathbf{A} = \alpha\mathbf{I} + \beta\mathbf{B}$,

$$\begin{aligned}\frac{d}{d\alpha}\log\det\mathbf{A} &= \mathrm{Trace}\left(\mathbf{A}^{-1}\frac{d\mathbf{A}}{d\alpha}\right)\\ &= \mathrm{Trace}\left(\mathbf{A}^{-1}\mathbf{I}\right) = \mathrm{Trace}\mathbf{A}^{-1}\end{aligned}$$

This result is exact if E_W and E_D are quadratic. Otherwise this result is an approximation, omitting terms in $\partial\mathbf{B}/\partial\alpha$. Now, differentiating (2.21) and setting the derivative to zero, we obtain the following condition for the most probable value of α:

$$2\alpha E_W^{\rm MP} = k - \alpha\mathrm{Trace}\mathbf{A}^{-1} \tag{2.22}$$

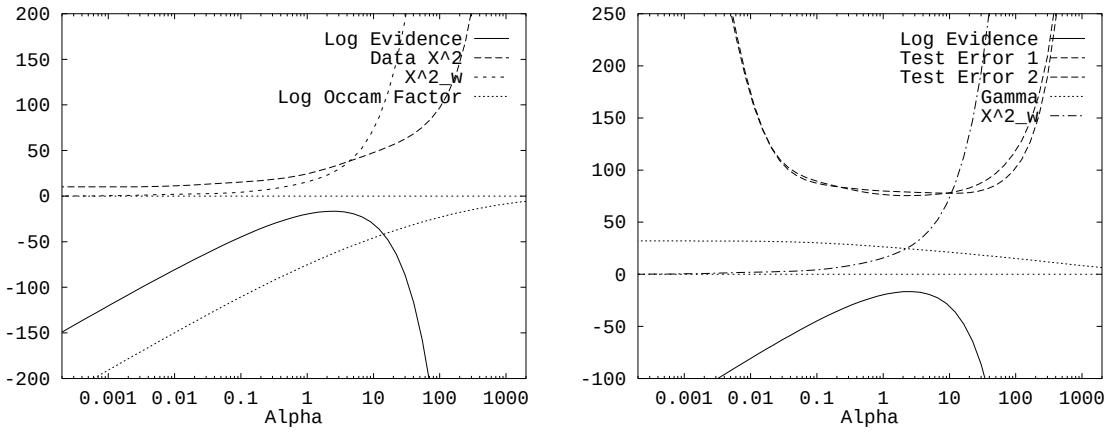

Figure 2.5: **Choosing** α

a) **The evidence as a function of** α**:** Using the same radial basis function model as in figure 2.4, this graph shows the log evidence as a function of α, and shows the functions which make up the log evidence, namely the data misfit $\chi^2_D = 2\beta E_D$, the weight penalty term $\chi^2_W = 2\alpha E_W$, and the log of the Occam factor $(2\pi)^{k/2}\det{}^{-\frac{1}{2}}\mathbf{A}/Z_W(\alpha)$.

b) **Criteria for optimising** α**:** This graph shows the log evidence as a function of α, and the functions whose intersection locates the evidence maximum: the number of good parameter measurements γ, and χ^2_W. Also shown is the test error (rescaled) on two test sets; finding the test error minimum is an alternative criterion for setting α. Both test sets were more than twice as large in size as the interpolated data set. Note how the point at which $\chi^2_W = \gamma$ is clear and unambiguous, which cannot be said for the minima of the test energies. The evidence gives α a 1-σ confidence interval of $[1.3, 5.0]$. The test error minima are more widely distributed because of finite sample noise.

defined earlier by

$$P(D|\alpha,\beta,\mathcal{H}) = \frac{Z_M(\alpha,\beta)}{Z_W(\alpha)Z_D(\beta)}. \tag{2.16}$$

Occam's razor is implicit in this formula: if α is small, the large freedom in the range of possible values of $\mathbf{w}$ is automatically penalised by the consequent large value of Z_W; models that fit the data well achieve a large value of Z_M. The optimum value of α achieves a compromise between fitting the data well and being a simple model.

Now to assign a preference to (α,β), our computational task is to evaluate the three integrals Z_M, Z_W and Z_D. We will come back to this task in a moment.

But that sounds like determining your prior after the data have arrived!

When I first heard the preceding explanation of Bayesian regularisation I was discontent because it seemed that the prior is being chosen from an ensemble of possible priors *after* the data have arrived. To be precise, as described above, the most probable value of α is selected; then the prior corresponding to that value of α alone is used to infer what the interpolant might be. This is not how Bayes would have us infer what the interpolant is. It is the combined ensemble of priors that define our prior, and we should integrate over this ensemble when we do inference.[11] Let us work out what happens if we follow this proper approach. The preceding method of using only the most probable prior will emerge as a good approximation.

The true posterior $P(\mathbf{w}|D,\mathcal{H})$ is obtained by integrating over α and β:

$$P(\mathbf{w}|D,\mathcal{H}) = \int P(\mathbf{w}|D,\alpha,\beta,\mathcal{H})P(\alpha,\beta|D,\mathcal{H})\,d\alpha\,d\beta \tag{2.17}$$

In words, the posterior probability over $\mathbf{w}$ can be written as a linear combination of the posteriors for all values of α,β. Each posterior density is weighted by the probability of α,β given the data, which appeared in (2.15). This means that if $P(\alpha,\beta|D,\mathcal{H})$ has a dominant peak at $\hat{\alpha},\hat{\beta}$, then the true posterior $P(\mathbf{w}|D,\mathcal{H})$ will be dominated by the density $P(\mathbf{w}|D,\hat{\alpha},\hat{\beta},\mathcal{H})$. As long as the properties of the posterior $P(\mathbf{w}|D,\alpha,\beta,\mathcal{H})$ do not change rapidly with α,β near $\hat{\alpha},\hat{\beta}$ and the peak in $P(\alpha,\beta|D,\mathcal{H})$ is strong, we are justified in using the approximation:

$$P(\mathbf{w}|D,\mathcal{H}) \simeq P(\mathbf{w}|D,\hat{\alpha},\hat{\beta},\mathcal{H}) \tag{2.18}$$

This approximation is valid under the same conditions as in footnote 12. It is a matter of ongoing research to develop computational methods for cases where this approximation is invalid (Sibisi and Skilling, personal communication).

Evaluating the evidence

Let us return to our train of thought at equation (2.16). To evaluate the evidence for α,β, we want to find the integrals Z_M, Z_W and Z_D. Typically the most difficult integral to evaluate is Z_M.

$$Z_M(\alpha,\beta) = \int d^k\mathbf{w}\ \exp(-M(\mathbf{w},\alpha,\beta))$$

[11] It is remarkable that Laplace almost got this right in 1774 [69]; when inferring the mean of a Laplacian distribution, he both inferred the posterior probability of a nuisance parameter like β in (2.15), and then attempted to integrate out the nuisance parameter as in equation (2.17).

use of test data, and cross–validation. Gull has demonstrated why the popular use of misfit criteria is incorrect and how Bayes sets these parameters [22]. The use of test data may be an unreliable technique unless large quantities of data are available. Cross–validation, the orthodox 'method of choice' [17], will be discussed more in section 2.6 and chapter 3. I will explain the Bayesian method of inferring α and β after first reviewing some statistics of misfit.

Misfit, χ^2, and the effect of parameter measurements

For N independent gaussian variables with mean μ and standard deviation σ, the statistic $\chi^2 = \sum(x-\mu)^2/\sigma^2$ is a measure of misfit. If μ is known *a priori*, χ^2 has expectation $N \pm \sqrt{N}$. However, if μ is fitted from the data by setting $\mu = \bar{x}$, we 'use up a degree of freedom,' and χ^2 has expectation $N-1$. In the second case μ is a 'well–measured parameter.' When a parameter is determined by the data in this way it is unavoidable that the parameter fits some of the noise in the data as well. That is why the expectation of χ^2 is reduced by one. This is the basis of the distinction between the σ_N and σ_{N-1} buttons on your calculator. It is common for this distinction to be ignored, but in cases such as interpolation where the number of free parameters is similar to the number of data points, it is essential to find and make the analogous distinction. It will be demonstrated that the Bayesian choices of both α and β are most simply expressed in terms of the effective number of well–measured parameters, γ, to be derived below.

Misfit criteria are 'principles' which set parameters like α and β by requiring that χ^2 should have a particular value. The discrepancy principle requires $\chi^2 = N$. Another principle requires $\chi^2 = N - k$, where k is the number of free parameters. We will find that an intuitive misfit criterion arises for the most probable value of β; on the other hand, the Bayesian choice of α will be unrelated to the value of the misfit.

Bayesian choice of α and β

To infer from the data what value α and β should have,[9] Bayesians evaluate the posterior probability distribution:

$$P(\alpha, \beta | D, \mathcal{H}) = \frac{P(D|\alpha, \beta, \mathcal{H}) P(\alpha, \beta | \mathcal{H})}{P(D|\mathcal{H})} \tag{2.15}$$

The data dependent term $P(D|\alpha, \beta, \mathcal{H})$ has already appeared earlier as the normalising constant in equation (2.12), and it is called the evidence for α and β. Similarly the normalising constant of (2.15) is called the evidence for $\mathcal{H}$, and it will turn up later when we compare alternative models $\mathcal{H} = \{\mathcal{A}, \mathcal{N}, \mathcal{R}\}$ in the light of the data.

If $P(\alpha, \beta | \mathcal{H})$ is a flat prior[10] (which corresponds to the statement that we don't know what value α and β should have), the evidence is the function that we use to assign a preference to alternative values of α and β. It is given in terms of the normalising constants

[9] Note that it is not satisfactory to simply maximise the likelihood simultaneously over $\mathbf{w}$, α and β; the likelihood has a skew peak such that the maximum likelihood value for the parameters is not in the same place as most of the posterior probability [22]. To get a feeling for this here is a more familiar problem: examine the posterior probability for the parameters of a gaussian (μ, σ) given N samples: the maximum likelihood value for σ is σ_N, but the most probable value for σ (found by integrating over μ) is σ_{N-1}. It should be emphasised that this distinction has nothing to do with the prior over the parameters, which is flat here. It is the process of marginalisation that corrects the bias of maximum likelihood.

[10] Since α and β are scale parameters, this prior should be understood as a flat prior over $\log \alpha$ and $\log \beta$.