6.9 Marginalization

There is some sample data for this problem, a "spectrum" 100 samples long containing a Gaussian line and a variable background.
I chose a model of the form

$$m(x) = \alpha \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + \beta + \gamma x$$

and assumed Gaussian residuals.
The first thing to notice is that no-one has told us what the noise level is, and without this it seems that we cannot set up the problem.
There are various approaches; I will take the more familiar ones first.
Ideally, we would repeat the measurement. In this exercise, we can't do that, and in real life it is often not possible either.
We can try to estimate the noise level from what looks like a flat bit of the background. This is not very easy because, just as also often happens in real life, there doesn't seem to be anywhere that is very flat.
A common approach (the one we will use in this example) is to *assume* that we have the correct model, and then *derive* the noise level as the standard deviation of the residuals (data - model).
This is a useful pragmatic technique, but of course it rules out any model testing – it assumes the model is right. If we have the wrong model, we will just infer a large noise level and the model will still be probable, formally. Also, the standard deviation which we put into our likelihood function is now a statistic – it is not a known parameter. To analyze this properly is involved. We may escape by asserting (hoping?) that we have enough data so that the noise level is well-determined; and by hoping further that in this limit we can use the asymptotic Gaussian form of the likelihood function, which greatly simplifies analysis. There are some substantial compromises here; we will have to abandon priors, because we are equating the posterior probability distribution to the likelihood alone.
Now going through the full panoply of deriving a distribution function is rather involved, and it is a indeed lot easier to use the asymptotic limit of the likelihood. That is, we find the values of the free parameters $(\alpha, \mu, \sigma, \beta, \gamma)$ using a least-squares minimization algorithm.
Note that all of this will assume that the residuals are statistically independent from pixel to pixel. This means we can just multiply terms together to get the likelihood.
Below is the result of this fit; the fitted function is

$$-9.54 + +0.18x + 115.12 \exp\left(-0.01945(-29.94 + x)^2\right)$$

I used *MATHEMATICA*'s routine NonlinearFit to get this answer. Like most such routines, it can be persuaded to tell you the Hessian matrix at the minimum, which is what

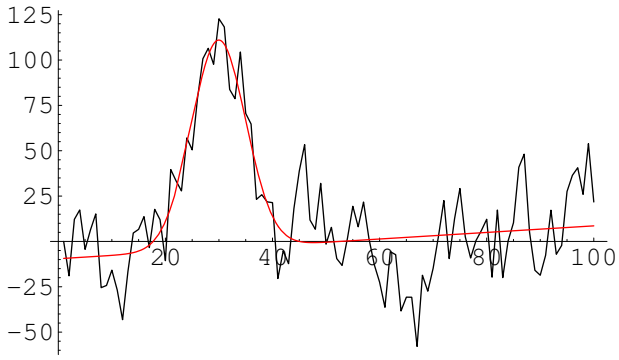

Figure 1: The least-squares fit of a Gaussian line to the data.

we want to get the covariance matrix of the parameters in the asymptotic approximation that they are distributed as a multivariate Gaussian. Note that we also have to put in an estimate of the noise level, defined as the standard deviation of the residuals. This matrix is (with rows and columns ordered $\alpha, \mu, \sigma, \beta, \gamma$ –

$$C = \begin{bmatrix} 77.91 & -0.08157 & -1.637 & -11.7 & 0.1292 \\ -0.08157 & 0.1897 & -0.007184 & 0.238 & -0.004098 \\ -1.637 & -0.007184 & 0.2301 & -1.031 & 0.01138 \\ -11.7 & 0.238 & -1.031 & 27.83 & -0.377 \\ 0.1292 & -0.004098 & 0.01138 & -0.377 & 0.00649 \end{bmatrix}.$$

It is now extremely easy to marginalize out baseline parameters; we have a Gaussian distribution for the vector $\alpha, \mu, \sigma, \beta, \gamma$ that is described fully by C. The covariance matrix for C', which describes the multivariate Gaussian distribution of α, μ, σ, is just the original matrix C with the rows and columns deleted that correspond to β, γ. The integrations can be done analytically; the file `jaynes_appendix.ps` contains most of the details necessary for a proof. (This is part of the on-line version of Jaynes' last book; it did not appear in the printed version, perhaps because Jaynes does not seem to have realized how simple an answer he had almost proved.)
We therefore have

$$C' = \begin{bmatrix} 77.91 & -0.08157 & -1.637 \\ -0.08157 & 0.1897 & -0.007184 \\ -1.637 & -0.007184 & 0.2301 \end{bmatrix}.$$

The correlation coefficient between height and width is about 0.39, and the standard deviation on the height, marginalizing out width and position, is $\sqrt{77.91}$; and similarly for the other variables.
All of this (the machinery, if not the interpretation of the results) is standard least-squares theory, and is described lucidly and fully in *Numerical Recipes*.

A specifically Bayesian ingredient is to use a prior to get around the problem of "not knowing" the noise level. Assuming still that we have Gaussian residuals, with standard deviation ξ (don't confuse this with the standard deviation of the line!), each term in the likelihood product is

$$\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(Y_i - m(X_i))^2}{2\xi^2}\right)$$

where the Y_i are the data.
As usual, we can make the likelihood

$$\mathcal{L}(\alpha, \mu, \sigma, \beta, \gamma, \xi|\text{data})$$

which now includes the noise level.
Suppose we use the Jeffreys prior $\text{prob}(\xi) \propto 1/\xi$, and marginalize over ξ – in other words, average over the range of (prior) possibilities for ξ. The posterior probability is then proportional to

$$\int_0^\infty \frac{d\xi}{\xi}\mathcal{L}(\alpha, \mu, \sigma, \beta, \gamma, \xi|\text{data}).$$

Our likelihood may appear formidable, but from the point of view of integrating over ξ it is simply of the form

$$\frac{1}{\xi^{n+1}}\exp\left(-\frac{\sum_i (Y_i - m(X_i))^2}{2\xi^2}\right)$$

and its integral, as above, is proportional to

$$\left(\frac{1}{\sum_i (Y_i - m(X_i))^2}\right)^{n/2}.$$

For small amounts of data (small n) this is where the analysis would stop, with an exact form of the posterior distribution of the fitting parameters. In our case, for large n, we can go further. If m is linear in the parameters $a, b \ldots$ this is a (multivariate) t-distribution; and for n large, the t-distribution is close to a Gaussian. If we have enough data, we will only be interested in small ranges of the parameters (which will be well-determined) so a linear approximation will suffice. This gets us back to our original approximation, which is that the parameters have a multivariate Gaussian distribution. Intuitively this is right; as suggested before, if we have "enough" data, then we ought to be able to estimate the noise level from the data.
However, we have gained something - we can compare models (various m). Because we have put in a prior for the noise level ξ, not all noise levels are equally acceptable. It follows that we can compare various models in the standard Bayesian way, using the Bayes factor. Some caution is needed with the normalization if you use the Jeffreys prior – see the Errata.