no prior knowledge of $\underline{\theta}_0$, and that we are concerned with location parameters, for which a non-informative uniform distribution ought to be considered. These conditions lead us to the posterior-distribution formula $p(\underline{\theta}|\underline{\eta},\underline{V})d\underline{\theta} \propto \exp[(-1/2)(\underline{\eta} - \underline{D}\,\underline{\theta})^+ \underline{V}^{-1}(\underline{\eta} - \underline{D}\,\underline{\theta})]d\underline{\theta}$ which is indeed normal in $\underline{\theta}$. The method of least squares, as discussed in Chap. 12, can then be used without ambiguity to obtain a best estimate for $\underline{\theta}$.

Example 11.16

Suppose that all of the parameters to be estimated, and all of the observables to which they are related, are inherently positive, i.e., $\theta_k > 0$, for k = 1,m, and $y_i > 0$, for i = 1,n. Furthermore, assume that the parameters and observables are related through products and quotients of factors, i.e., $y_i = \Pi_{k=1,m}\theta_k{}^b$, where (for convenience) $b = b_{ki}$, for k = 1,m and i = 1,n, is any real number which can vary arbitrarily with i and k. We also assume that we are considering location parameters, for which we have no prior knowledge. As in Ex. 11.15, we suppose that $\underline{\eta}$ and $\underline{V}$ are obtained from experimental measurements of the observables $\underline{y}$. The relationship between $\underline{\theta}$ and $\underline{y}$ is not linear, so the posterior distribution derived from the Bayesian method is not a Gaussian. However, if we make the variable transformations $z_i = \mathit{ln}\ y_i$, for i = 1,n, $s_k = \mathit{ln}\ \theta_k$, for k = 1,m, and $\xi_i = \mathit{ln}\ \eta_i$, for i = 1,n, we then have the expression $z_i = \mathit{ln}\ y_i = \Sigma_{k=1,m}\ b_{ki}\ \mathit{ln}\ \theta_k = \Sigma_{k=1,m}\ b_{ki}s_k$, for i = 1,n. The covariance matrix for $\underline{\xi}$, which we shall call $\underline{W}$, is simply the relative covariance matrix corresponding to $\underline{V}$. There is a linear relationship between $\underline{z}$ and $\underline{s}$, which we can denote by $\underline{z} = \underline{D}\,\underline{s}$. We are not completely past our difficulties, however, for we still have to consider the matter of the prior distribution for $\underline{s}$ (the transformed parameter set). At this point, it is necessary to make an approximation. We suppose that the likelihood $L(\underline{\eta}|\underline{z})$ is well localized (i.e., that $\underline{V}$ implies small errors). Then, we can assume that the non–informative prior distribution for $\underline{s}$ is approximately uniform over the region of significant likelihood. What is involved in this assumption? Since $s_k = \mathit{ln}\ \theta_k$, $ds_k = d\theta_k/\theta_k$. We require that $p(\theta_k)d\theta_k = p(s_k)ds_k$ in order to preserve probability in the transformation of variables. Actually, $p(\theta_k)d\theta_k \propto d\theta_k$. Therefore, $p(s_k)ds_k \propto \exp(s_k)ds_k$. Our assumption that $p(s_k)$ is constant is thus equivalent to assuming that $\exp(s_k)$ is approximately constant over the region of significant likelihood. The transformation we have been describing, which leads to a log–normal distribution (see Sec. 7.1.11), is quite handy whenever it is applicable. It conveniently avoids some very serious difficulties, like the one to be discussed below.

We now turn to a discussion of certain hazards associated with attempting to circumvent the rigorous Bayesian approach to parameter estimation. It is convenient to use a particular example for demonstration purposes. This example was suggested by R. Peelle [Pee87], and it has come to be known as "Peelle's Pertinent Puzzle" ("PPP" for short). Several individuals participated recently in a detailed examination of PPP and its implications for nuclear applications of statistical methodology [Chi90, Fro90, Mea90, Per90, Von90, Zha90]. The present discussion of this issue is derived largely from the unpublished memoranda which emerged out of this debate. The viewpoints that have been expressed on this issue diverge significantly. Since no consensus exists at present on precisely how to deal with the specific problem raised by PPP (which, in fact, is a relatively simple one), it is clear that there is much work left to be done before the nuclear community can agree on suitable approximation methods to use for parameter–estimation within the general framework of Bayesian methodology. We shall now examine PPP in more detail:

Example 11.17

Let us suppose that our objective is to determine best values for a parameter x and its variance. However, x is not directly observable but is derived as the quotient of two positive observable quantities

a and c, i.e., $x = a/c$. We suppose that there exist two independent determinations of a, namely, $a_1 \pm \sigma_1$ and $a_2 \pm \sigma_2$. On the other hand, c has been measured only once, with the result $c_0 \pm \sigma_0$. The measurement of c is independent of those for a. We further suppose that both a and c are location parameters and that there exists no prior information on any of these parameters. Peelle's Pertinent Puzzle (PPP) emerges in addressing the matter of finding a best estimate for x, not one for a or c. It deals with a fundamental dilemma which arises when considering the relationships between the various processes of measurement, data reporting, and data evaluation. For the sake of argument, let us assign values to the measured results: $a_1 = 1.5$, $\sigma_1 = 0.15$, $a_2 = 1.0$, $\sigma_2 = 0.1$, $c_0 = 1.0$, and $\sigma_0 = 0.2$. We note at the outset that the experimental determinations of a are quite discrepant (differ significantly relative to the given errors). Furthermore, the errors in all the measured results, particularly for c_0, are quite substantial. The Bayesian procedure which we have been describing in this section leads us immediately to the posterior distribution function

$$p(a,c \mid a_1,\sigma_1,a_2,\sigma_2,c_0,\sigma_0)dadc \propto$$

$$\exp\{(-1/2)[\Sigma_{i=1,2}(a_i - a)^2/\sigma_i^2] + (-1/2)[(c_0 - c)^2/\sigma_0^2]\}dadc$$

From this distribution function we can easily compute the expected values for a and c and their respective variances, namely, $\langle a \rangle = (a_1\sigma_1^{-2} + a_2\sigma_2^{-2})/(\sigma_1^{-2} + \sigma_2^{-2})$, $\mathrm{var}(a) = (\sigma_1^{-2} + \sigma_2^{-2})^{-1}$, $\langle c \rangle = c_0$, and $\mathrm{var}(c) = \sigma_0^2$. The results for a and var(a) conform completely with Ex. 11.14. If we now insert numerical values for the data, we find that $\langle a \rangle = 1.154$ and $\mathrm{var}(a) = 0.006923$. But, we must not get sidetracked here; our objective was to determine x and var(x). In order to accomplish this within the framework of Bayesian methodology, we shall make a change of variables and transform the posterior distribution accordingly. We transform the variable set (a,c) to the equivalent set (x,c). Since $x = a/c$, $a = cx$. For convenience, we let "$\mathscr{D}$" stand for all the experimental data ($a_1,\sigma_1,a_2,\sigma_2,c_0$, and σ_0). Our task is then to transform $p(a,c \mid \mathscr{D})dadc$ to the equivalent $p(x,c \mid \mathscr{D})dxdc$. Referring to Sec. 6.2 and, specifically, to Eq. (6.14), we see that $p(x,c \mid \mathscr{D}) = p(a,c \mid \mathscr{D})/|J|$, where $|J|$ is the absolute value of the Jacobian J, for this transformation. In this simple case, J is quite easy to calculate. Thus, $|J| = (1/c)$ and $p(x,c \mid \mathscr{D}) \propto c \exp\{(-1/2)[\Sigma_{i=1,2}(a_i - cx)^2/\sigma_i^2] + (-1/2)[(c_0 - c)^2/\sigma_0^2]\}$. This is the correct posterior distribution for the evaluation of best estimates for x and c. This distribution is <u>not</u> normal in x and c. There is no easy way to "read off" the answers we desired. What we should do, for completeness, is to evaluate $\langle x \rangle$, $\langle c \rangle$, and their associated 2×2 covariance matrix $\underline{V}_{xc}$ consisting of the elements $V_{xx} = \langle \delta x^2 \rangle$, $V_{xc} = V_{cx} = \langle \delta x \delta c \rangle = \langle \delta c \delta x \rangle$, and $V_{cc} = \langle \delta c^2 \rangle$, where $\delta x = x - \langle x \rangle$ and $\delta c = c - \langle c \rangle$. We are really concerned only with $\langle x \rangle$ and $\mathrm{var}(x) = \langle \delta x^2 \rangle$. Therefore, we turn to "brute force" numerical integration to evaluate them. Our result is $\langle x \rangle = 1.21$ and $\mathrm{var}(x) = 0.09$ (Method 1). This is the rigorous solution to PPP, in the context of the Bayesian formation. The marginal probability distribution for x (i.e., the posterior distribution integrated with respect to c) is skewed considerably toward larger x, relative to what our naive intuition might lead us to think is the best value for x, namely, $\langle a \rangle / c_0 = 1.154$. Since $x = a/c$, we have a simpler (albeit approximate) option for dealing with this problem, i.e., the logarithmic transformation approach described in Ex. 11.16. This transformation eliminates the non–linearity introduced by the quotient $x = a/c$, since $\ln x - \ln a = -\ln c$. We shall not go into the details, but it can be shown that this leads to a best estimate for x of 1.225 ± 0.260 (Method 2), a result which does not differ by too much from a completely rigorous treatment involving numerical integration.

The Bayesian procedure is both appealing and rigorous, in principle, but we cannot avoid facing up to the question of how to deal in a practical way with parameter estimation problems which involve a substantial number of parameters, extensive data, and complex relationships between the observables and parameters. Brute force numerical integration, even when Monte–Carlo techniques are pursued, is scarcely an option to be considered. The logarithmic transformation works in a limited number of situations. What is to be done? We shall return to this question, but first let us explore a serious trap into which we may be easily seduced if we are not careful in our quest for convenient approximations.

We begin by examining what often happens in the "real world" of measurers and evaluators. It is the task of an evaluator to determine a best value for x. He refers to the literature and finds only two relevant experiments. Experiment 1 provides the measured result $x_1 = 1.0 \pm 0.22$. while experiment 2 yields $x_2 = 1.50 \pm 0.34$. If the evaluator is not very sophisticated, he may very well assume that these are totally independent measurements, accept the given errors at face value, and derive the result 1.148 ± 0.185 (Method 3), a simple weighted average, for his evaluated estimate. It is not a bad result from a numerical point of view, but we know that this approach is conceptually flawed for several reasons, as mentioned above. Let us suppose next that our evaluator is more thorough. He digs into the papers that document these two experiments and discovers that, in both works, the authors interpret x to be a/c. Each measures a independently with a 10% error, i.e., $a_1 = 1.5$ (± 10%) and $a_2 = 1.0$ (± 10%). But, the evaluator also discovers that both experiments use the same value of c, i.e., $c_0 = 1.0$ (± 20%), which they have drawn from a third literature source, as a normalization standard. What is the evaluator to do now? Before applications of covariance methodology became commonplace in the nuclear field, the normal procedure was to evaluate a first. The weighted average of a_1 and a_2 is just 1.154 ± 0.0832 (see above). Since c is, unambiguously, $c_0 = 1.0 \pm 0.2$, we are led to the evaluated result 1.154 ± 0.246 (Method 4), following simple division and an exercise in conventional error propagation. This result does not differ by much from the value obtained via completely naive weighted averaging. So, where is the reward for the evaluator's additional effort? Let us proceed even further and suppose that the evaluator wishes to take into account the obvious correlation introduced by a common normalization factor. He treats the quantities $x_1 = a_1/c_0 = 1.0$ and $x_2 = a_2/c_0 = 1.5$ as data, with an associated 2 × 2 covariance matrix $\underline{V}$ whose values are derived from the information given above. The matrix elements are: $V_{11} = (0.2236)^2 = 0.05$, $V_{12} = V_{21} = (0.2)(0.3) = 0.06$, and $V_{22} = (0.3354)^2 = 0.1125$. Note that in deriving these matrix elements the given data values and percent errors are used to obtain the absolute errors. More likely than not, he will then proceed to use the least–squares method, as discussed in Sec. 12.1.1, to obtain his evaluated result. Since we have been describing the Bayesian procedure, we will assume that he takes this approach (in this instance it makes no difference). He treats x as a location parameter. Since there is no prior information, the non–informative uniform prior distribution is employed. With these considerations in mind, he writes down the posterior distribution, namely,

$$p(x|x_1,x_2,\underline{V})dx \propto \exp[(-1/2)\delta\underline{x}^+ \underline{V}^{-1}\delta\underline{x}]dx,$$

where $\delta\underline{x}$ is a vector with components $\delta x_i = x_i - x$, for i = 1,2. This probability function is normal in x. Determination of $\langle x \rangle$ and var(x) is straightforward (see Secs. 2.1.1, 7.1.10 and 12.1.1) . However, the analysis yields $\langle x \rangle = 0.8824$ and var(x) = 0.0476 (Method 5), a result which is quite alarming! The only way this could be a correct result would be if x_1, x_2, and $\underline{V}$ were the only available information, and we did not know that x = a/c, with data given for a and c. This is not the case here. Actually, it was this strange, non-intuitive result which led Peelle to bring the whole matter to the attention of several of his colleagues, in the form of PPP [Pee87]. As a way out of this dilemma, and desiring all the while to preserve the least-squares method as a viable approach for dealing with problems of this nature, Zhao and Perey [Zha90, Per90] suggested that the problem could be dealt with by altering the manner in which the covariance matrix $\underline{V}$ is evaluated. They treat $x_1 = 1.0$ and $x_2 = 1.5$ as data, but argue that the computation of the elements of $\underline{V}$ should be carried out using $c_0 = 1.0$ for c, and the weighted average of a_1 and a_2 (namely, 1.154) for a, not the directly measured values $a_1 = 1.0$ and $a_2 = 1.5$ (as was the case for Method 5). This is a very rational suggestion. With this assumption, the elements of $\underline{V}$ are: $V_{11} = 0.07575$, $V_{12} = V_{21} = 0.05325$, and $V_{22} = 0.06325$. Application of the least squares method (see Sec. 12.1.1) then leads to the much more reasonable result 1.154 ± 0.245 (Method 6) for the best estimate of x and its error. Finally, let us assume that there exists a true value, x_0, but we do not know exactly what it is. However, two experimental attempts to determine x_0 have produced the results $x_1 = 1.0$ and $x_2 = 1.5$. Furthermore, our statement of errors indicates that each of these measurements involved a 10% random error and a 20% fully correlated error. Consequently, we make the assumption that the absolute errors are computed in terms of these percentage errors and the true (unknown) value which we shall call x_0. We shall refer to this approach (suggested by Chiba [Chi90] and this author) as

Method 7. The result for the covariance matrix $\underline{V}$ is: $V_{11} = 0.05\ x_0^2$, $V_{12} = V_{21} = 0.04\ x_0^2$, and $V_{22} = 0.05\ x_0^2$. It happens, for this simple example, that the unknown factor, x_0^2, in the elements of $\underline{V}$ actually cancels when the least–squares condition is invoked. Therefore, we are led to the solution $\langle x \rangle = 1.250 \pm 0.265$. This is fairly close to the result from Method 1. Method 7 can also be invoked in problems involving more than two random variables. However, it is then necessary to seek a solution through an iterative procedure.

Why should those seven distinct approaches (Methods $1 - 7$) to solving this simple problem generally lead to noticeably different answers, including one which appears to defy common sense? The reason is really quite simple and should come as no surprise: Each method is unique in concept and each treats the available experimental information differently. One should expect that different answers would emerge. In PPP, the differences are exaggerated because the data errors are quite large (especially that for the normalization factor, c_0) and the measured data for a are seriously discrepant. In short, the posterior probability distribution is quite skewed.

Let us summarize the lessons to be learned from PPP (Ex. 11.17): Knowledge of the mean values and their uncertainties (through a covariance matrix) for observables enables us to write down a posterior probability distribution which is normal, provided that the parameters in question are location parameters (or, for practical purposes, can be treated as such so that non–informative priors can be used), and that there is either a direct or, at the very worst, a linear relationship between the observables and the parameters to be estimated. Then, the least–squares, maximum likelihood, and full Bayesian estimation techniques are completely equivalent. However, more complex problems call for the use of approximation methods if the determination of requisite expectation values with respect to the posterior distributions is prohibitively difficult. Great care must be taken in applying these approximate methods, since they can lead readily to rejection of valuable information, with disastrous consequences. There are differences of opinion on this issue which need to be resolved eventually. In other words, a consensus must be sought on how one ought to proceed with making these approximations.

With the exception of Method 5, the results produced for PPP by these methods tend to fall into two distinct groups. Methods 1 and 2 are linked to the true Bayesian approach. Therefore, the best estimate given in each case is essentially the mean value of a skewed, non-Gaussian posterior probability distribution, i.e., $\langle x \rangle \approx 1.22$. The other methods (including Method 5) all yield estimates of the most probable value, $x_{mp} \approx 1.15$. According to decision theory, the most probable value is not the best estimate, but for skewed distributions it may be easier to determine. We can take the following general approach, known as *saddle–point integration*, and avoid the pitfall illustrated by the Method 5 solution to PPP (see Ex. 11.17): We approximate the true posterior distribution by a Gaussian with the same maximum location and curvature (second–order terms in a Taylor series expansion). This surrogate distribution is intended to be a valid approximation only in the vicinity of the most probable solution. In short, we abandon our quest for the correct mean values, as a mathematical expedient. If the posterior distribution is based on measured data (consisting of estimates of mean values and their covariance matrix), and we assume a non–informative prior distribution, i.e., the parameters are (for all practical purposes) treatable as location parameters, we note that the posterior probability can be written in the form $p(\underline{\theta}|\ \mathscr{D})d\underline{\theta} = \exp(-\chi^2/2)d\underline{\theta}$. Saddle–point integration then involves maximizing the exponential function and thus minimizing the expression χ^2 (thereby establishing a link to the usual least–squares procedures discussed in Sec. 11.2.2 and Chap. 12). The only point to remember is that we must not be overzealous in our attempts to simplify χ^2 by throwing away those terms of a Taylor's series approximation that contain valuable information, thus ending up with a situation akin to that which we encountered in the Method 5 approach to PPP.

We have avoided mentioning a chi–square test for the solution (Sec. 11.3.2). It is strictly valid only if the posterior probability distribution for the estimated parameters is normal. The approximation procedures we have described are predicated (either explicitly or implicitly) on substituting a normal distribution for the true distribution whenever it is not normal. Therefore, it makes sense to apply the chi–square test in practical applications, to determine the consistency of the data and solution. In PPP (Ex. 11.17), each solution method led to a value of chi–square per degree of freedom much larger than unity due to the measured data discrepancies.

Reading List for Chapter 11

Samples and sample statistics:
[Ash70, Bev69, Bur68, Coc63, Coo69, Dav82, Dem50, Fel50, Fis63, Fre62, Kha76, Lew75, Lyo70, Mar71, Mun51, Ney50, Par60, Sch69, Smi88b, Wal86, YK50, Zeh70]

Sampling distributions for normal populations:
[Ash70, Bev69, Bje73, Bro60, Bur68, Coo69, Dem50, Fis63, Fre62, Gir77, Hem67, Kha76, Mar71, Men67, Mil75, Mor68, Par60, Par61, Sch69, Shc65, Wal86, YK50, Zeh70, Zij87]

Asymptotic behavior of samples and sampling distributions:
[Ash70, Bas66, Bee58, Bro60, Cla75, Cra70, Dem50, Fis63, Kha76, Mor68, MS73, Mun51, Ney50, Par60, Smi88b, Tuc67, YR86, Zac71, Zeh70]

Point estimators and their properties:
[Ash70, Bas66, Bje73, Bro60, Bur68, Coo69, DB34, Dem50, Ead+71, Fis63, Fre62, Gir77, Hem67, Lyo70, Mar71, Men67, Mil75, Mor68, Par61, Sch69, Wal86, Was70,YK50, YR86, Zac71, Zeh70]

Classical methods for determining point estimators:
[Ash70, Bje73, Bro60, Bur68, BW47, Coo69, DB34, Dem50, Ead+71, Fis63, Fre62, Gir77, Hem67, Lyo70, Mar71, Mil75, Mui88, Mui89, Par61, Sch69, Wal86, Was70, Zac71, Zeh70]

Interval estimation:
[Bas66, Bje73, Bur68, Coo69, Dem50, Fis63, Fre62, Mar71, Men67, Par61, Sch69, Wal86, Was70, Zac71, Zeh70]

Linear regression models:
[Bas66, Bev69, Bje73, Bro60, Bur68, BW47, Coc63, Coo69, Dav82, Ead+71, Fis63, Fre62, Hem67, Kha76, Lyo70, Mar71, Men67, Mil75, Par61, Sch69, Shc65, Wal86, Was70, YK50, YR86, Zac71, Zeh70, Zij87]

Chi–square test:
[Bas66, Bev69, Bro60, Bur68, Coo69, Dav82, Dem50, Ead+71, Fis63, Fre62, Gir77, Hem67, Kha76, Lyo70, Mar71, Men67, Mil75, Mor68, MS73, Par61, Sch69, Wal86, YK50, YR86, Zeh70]

Decision theory:
[Ash70, Bje73, Bro60, Bur68, Dem50, Fis63, Fre62, Men67, Sch69, Wal86, Was70, Zac71, Zeh70]

Bayesian parameter estimation:
[Cox46, Fro86, Fro87, Sha48, SW49, Wal50]

Fundamental considerations in the choice of Bayesian a priori probabilities:
Bur68, Fro86, Fro87, Jay68, Jay73, Jay76, Jay78, Jay80, Mar71, Per82, Sch69, Sha48, Shc65, Smi88b, SW49, WM89, Zac71, Zeh70]