SIAM Review
Vol. 57, No. 1, pp. 131–149

Statistics of Parameter Estimates: A Concrete Example*

Oscar Aguilar†
Moritz Allmaras‡
Wolfgang Bangerth§
Luis Tenorio¶

Abstract. Most mathematical models include parameters that need to be determined from measurements. The estimated values of these parameters and their uncertainties depend on assumptions made about noise levels, models, or prior knowledge. But what can we say about the validity of such estimates, and the influence of these assumptions? This paper is concerned with methods to address these questions, and for didactic purposes it is written in the context of a concrete nonlinear parameter estimation problem. We will use the results of a physical experiment conducted by Allmaras et al. at Texas A&M University [M. Allmaras et al., *SIAM Rev.*, 55 (2013), pp. 149–167] to illustrate the importance of validation procedures for statistical parameter estimation. We describe statistical methods and data analysis tools to check the choices of likelihood and prior distributions, and provide examples of how to compare Bayesian results with those obtained by non-Bayesian methods based on different types of assumptions. We explain how different statistical methods can be used in complementary ways to improve the understanding of parameter estimates and their uncertainties.

Key words. parameter estimation, Bayesian inference, frequentist inference, model validation, data analysis, residual analysis, maximum likelihood, nonlinear regression, surrogate models

AMS subject classifications. 34A55, 62F15, 62P35, 97M50

DOI. 10.1137/130929230

1. Introduction.

Mathematical models of physical phenomena usually depend on parameters that need to be estimated from measurement data. Such estimation can be done using a variety of methods, but since no single procedure works best for every problem, we try to select a method such that (i) it takes advantage of all relevant available information; (ii) it is computationally feasible; (iii) it makes no unreasonable assumptions; and (iv) it is consistent with the data and physical models. However, we are sometimes too eager to focus on (i) and (ii) and neglect (iii) and (iv), which are important because the validity of the results depends on satisfying the assumptions of the statistical model. In particular, reported uncertainties are meaningful to the

*Received by the editors July 15, 2013; accepted for publication (in revised form) April 10, 2014; published electronically February 5, 2015.
http://www.siam.org/journals/sirev/57-1/92923.html

†Department of Statistics, Iowa State University, Ames, IA 50011 (oaguilar@iastate.edu).
‡Siemens AG, Otto-Hahn-Ring 6, 81739 Munich, Germany (moritz.allmaras@gmail.com).
§Department of Mathematics, Texas A&M University, College Station, TX 77843 (bangerth@math.tamu.edu). The work of this author was partially supported by award KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

¶Department of Applied Mathematics and Statistics, Colorado School of Mines, CO 80401 (ltenorio@mines.edu). The work of this author was partially supported by NSF grant DMS-0914987.

extent that the assumptions are approximately satisfied. In this paper we describe different ways to check assumptions and explore the validity of parameter estimates and their uncertainties. This is done in the context of a concrete nonlinear parameter estimation problem. We use the results of the experiment published by Allmaras et al. in this journal in 2013 [1]. Thus, this article is a continuation of that tutorial, which we will henceforth refer to as "A13."

The experiment discussed in A13 consisted of recording the free fall of a box dropped from a known height. The fall was recorded by a video camera at a rate of thirty frames per second with the aim of determining the gravitational acceleration and air friction coefficient of the box. The distance traveled by the object was obtained by analyzing 31 frames of this video. t_0 is the time in seconds when the object starts to travel and was initially chosen to be the time corresponding to the first of the 31 frames, with t_i the times corresponding to the other frames. As was shown in A13, the time, t_0, when the object actually started to travel was impossible to determine accurately because any motion in the first frames is small compared with the resolution of the camera, and the second part of A13 therefore treats it as uncertain as well.

From the frames of the video showing the body falling past a tape measure, the authors determined the distance d_i the body has fallen at time t_i. But, of course, measuring d_i is subject to uncertainties. Even more than motion blur, the most important source of error in these data was the fact that the box rotated during its fall and the consequential uncertainty in how to define the location of the object. A13 defined the error, δ_i, to be half the distance between the lower and upper corners of the bottom of the box,[1] as depicted in Figure 1.1 (see also the photographs in Figure 3.2 of A13).

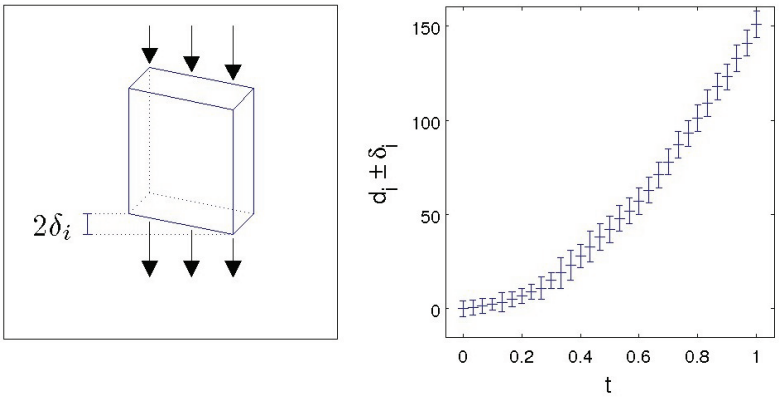

Fig. 1.1 *Left: Depiction of the falling box and the errors δ_i. See Figure* 3.2 *of* A13 *for an actual picture of the falling box in the experiments described there. Right: Plot of $d_i \pm \delta_i$.*

These values were used in A13 to define the support of the probability density of the errors associated with d_i. The right panel in Figure 1.1 shows the distance d_i (in inches) as a function of t_i (s) with error bars defined as $\pm\delta_i$.

[1]This might not be the best definition of the measurement error, given that the box can rotate around all three axes, but it is one that could be extracted from the individual frames of the video. As we will discuss below, their choice was overly conservative, and so the details of their definition of measurement error might not have mattered very much.

As explained in A13, Newton's equations of motion lead to an explicit expression of the distance $z(t)$ traveled by the object at time t:

$$z(t) = (1/C)\,\log\cosh\left[\sqrt{gC}(t-t_0)\right], \tag{1.1}$$

where g is the acceleration due to gravity and C is the specific coefficient of air resistance for this particular box. Although both parameters are unknown, our main interest is in identifying g, for which we have the reliable estimate

$$g \approx g_{\text{ref}} = 9.7935\ (\text{m/s}^2) \tag{1.2}$$

for the location where the experiment took place (College Station, TX; see A13). Here, as in A13, we are therefore primarily concerned with estimating g and comparing it against g_{ref} for validation. As in every experiment, we do not of course use this knowledge of a reference value in the estimation procedures. Since A13 determined that knowledge of the initial time t_0 was subject to a small error, it considered the case where t_0 is unknown. Hence, the vector of unknown parameters we use will be either $\boldsymbol{m} = (g, C)$ or $\boldsymbol{m} = (g, C, t_0)$.

To summarize, the data, d_i, consist of a noisy recording of $z(t)$ at 31 different times t_i: $d_i = z(t_i) + \varepsilon_i$, where ε_i is modeled as random noise. We write the data vector as

$$\boldsymbol{d} = \boldsymbol{z}(\boldsymbol{m}) + \boldsymbol{\varepsilon}, \quad z_i(\boldsymbol{m}) = (1/C)\,\log\cosh\left[\sqrt{gC}(t_i - t_0)\right]. \tag{1.3}$$

The objective of the inverse problem is to use the data $\boldsymbol{d}$ to estimate the unknown parameter vector $\boldsymbol{m}$. Our objective in this paper is to interpret and assess the reliability of the estimate and its associated uncertainty.

In the Bayesian approach for parameter estimation used by A13, $\boldsymbol{m}$ is modeled as a random variable with a specified prior probability distribution that encodes what the authors believed they knew about $\boldsymbol{m}$ independent of the measured data. The result of their inference is the posterior probability distribution of $\boldsymbol{m}$ conditional on $\boldsymbol{d}$. The likelihood and priors were selected using heuristic arguments that are typical in applications. In particular, they represented their best guesses at interpreting the measured data and a statistical description of their prior knowledge. In this paper we question whether these choices were appropriate, and we describe methods that can be used to validate these choices. We also describe alternative non-Bayesian inversion methods that make different types of assumptions on the statistical model. We will then explain how to compare their results to those from the Bayesian procedures.

The rest of the paper is organized as follows. In section 2 we summarize the work presented in A13. We start the validation analysis by checking the frequentist performance of the Bayesian procedures from A13 in section 3. We focus on the validation of the likelihood and priors in section 4. This analysis leads to a new likelihood and to a revision of the results in A13. The new likelihood is then used in what follows. Since the prior distributions for the experiment are not defined based on the physics of the problem, but on intuitive and convenient choices, it is worth comparing the Bayesian results to those obtained with non-Bayesian methods. In section 5 we find parameter estimates and confidence intervals using maximum likelihood and nonlinear least-squares. For these methods, $\boldsymbol{m}$ is no longer modeled as random. In section 6, we use a method that does not require parameter estimates to construct the confidence sets. We summarize and discuss all the results in section 7.

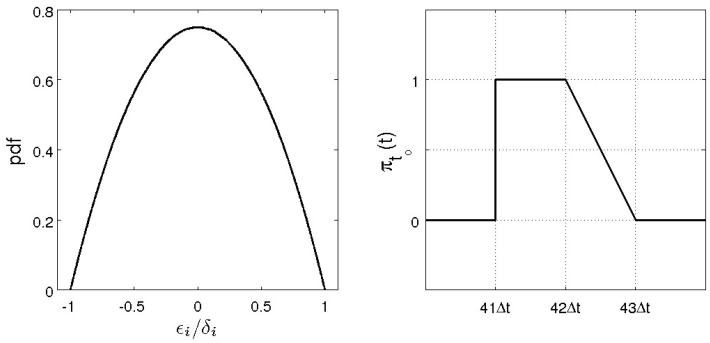

Fig. 2.1 *Left: PDF* (2.1) *of the noise* ε_i *plotted as a function of* ε_i/δ_i. *Right: Function* (2.3) *that defines the prior density* (2.3) *for* t_0.

2. Summary of A13. We start by summarizing the Bayesian framework used in A13. There, the noise variables, $\varepsilon_i = d_i - z_i(\mathbf{m})$, are assumed to be independent with a shifted/scaled beta$(2,2)$ distribution: for each i, $\varepsilon_i/\delta_i = 2U_i - 1$, where the U_i are independent and identically distributed (i.i.d.) beta$(2,2)$ random variables. Thus, the probability density function (PDF) of ε_i is (see the left panel in Figure 2.1)

$$f_i(\varepsilon_i) \propto \left(1 - \varepsilon_i^2/\delta_i^2\right) \mathbb{I}_{[-1,1]}(\varepsilon_i/\delta_i), \tag{2.1}$$

where $\mathbb{I}_A$ denotes the indicator function of the set A. For simplicity, normalization constants are omitted here and in what follows. Thus, the PDF of ε_i is symmetric about zero with support $[-\delta_i, \delta_i]$. Note that δ_i is not the standard deviation of ε_i but a scaling factor of it: $\mathrm{Std}(\varepsilon_i) = 4\delta_i^2\,\mathrm{Std}(\mathrm{beta}(2,2))$. This probability density was chosen by the authors of A13 to indicate that they truly believed that the "real" value should be in the interval $[d_i - \delta_i, d_i + \delta_i]$, and within this interval with larger probability in the vicinity of the center than near the boundaries. This PDF leads to the following likelihood of $\boldsymbol{m}$:

$$L(\boldsymbol{m};\boldsymbol{d}) \;\propto\; \prod_{i=1}^{31} \left[\,1 - (d_i - z_i(\boldsymbol{m}))^2/\delta_i^2\,\right] \mathbb{I}_{[-1,1]}(\,[\,d_i - z_i(\boldsymbol{m})\,]/\delta_i\,). \tag{2.2}$$

To define the prior distribution for $\boldsymbol{m}$, the parameters g, C, and t_0 are assumed to be independent with g and C uniformly distributed on the intervals $[0, 20]$ (m/s^2) and $[0, 0.5]$ (m^{-1}), respectively. The authors of A13 chose these as reflecting their best prior knowledge: both parameters certainly need to be positive, and an initial experiment may have provided plausible upper bounds of 20 and 0.5. When t_0 is also modeled as random (the three-parameter case), the authors of A13 chose the prior density for t_0 as shown in the right panel of Figure 2.1 and defined as

$$\pi_{t_o}(t_0) \propto \begin{cases} 1, & 41\Delta t \le t_0 \le 42\Delta t, \\ 1 - (t_0 - 42\Delta t)/\Delta t, & 42\Delta t \le t_0 \le 43\Delta t, \\ 0, & t_0 < 42\Delta t \;\text{ or }\; t_0 > 43\Delta t, \end{cases} \tag{2.3}$$

where $\Delta t = 1/30$ (s) and t is measured from the beginning of the recording, predating the time when the box started to fall. Again, this reflected the best heuristic interpretation of the data they obtained from their experiment: (i) they knew that the body

did not move before frame 41; (ii) they could not say for certain whether the body moved between frames 41 and 42; and (iii) it did move for certain in frame 43. In other words, π_{t_o} describes their beliefs about the time t_0 when the body might have started to move.

Given all this, the prior PDF of $\boldsymbol{m}$ is either

$$\pi(\boldsymbol{m}) \propto \mathbb{I}_{[0,20]}(g)\,\mathbb{I}_{[0,0.5]}(C), \tag{2.4}$$

when t_0 is fixed (the two-parameter case), or

$$\pi(\boldsymbol{m}) \propto \mathbb{I}_{[0,20]}(g)\,\mathbb{I}_{[0,0.5]}(C)\,\pi_{t_o}(t_0) \tag{2.5}$$

in the three-parameter case. In what follows, we will consider these as either separate cases or as two different priors: in the first case, the (very informative) prior distribution for the start time t_0 simply assumes that it is known.

In A13, moments of the posterior density $f_{\boldsymbol{m}|\boldsymbol{d}}(\boldsymbol{m} \mid \boldsymbol{d}) \propto L(\boldsymbol{m};\boldsymbol{d})\,\pi(\boldsymbol{m})$, in particular the mean and standard deviation, are evaluated by quadrature approximations of the integrals. This was possible because of the low-dimensional nature of the parameter estimation problem.[2] The posterior means of the parameters reported in A13 for priors (2.4) and (2.5) are, respectively,

$$\mathbb{E}(\,(g,C) \mid \boldsymbol{d}) = (8.82,\ 0.12) \quad \text{and} \quad \mathbb{E}(\,(g,C,t_0) \mid \boldsymbol{d}) = (8.64,\ 0.11,\ 1.40). \tag{2.6}$$

Using the fixed t_0 changes the posterior mean of g by 2% and that of C by 9%, but the difference in the results is more obvious in the posterior distributions. The posterior densities for g using either of the two priors for $\boldsymbol{m}$ are shown in Figure 2.2. The figure also shows the highest posterior density (HPD) 95% and 99% credible intervals for g (in m/s^2):

	Prior	95%	99%
(2.7a)	(2.4) :	[8.11, 9.53]	[7.97, 9.72]
(2.7b)	(2.5) :	[7.43, 9.93]	[7.33, 10.31].

Recall that credible intervals are not uniquely defined and, in particular, they can have very different lengths even if they have the same level $1-\alpha$ (as usual, we refer to the level of credible or confidence sets as $1-\alpha$ for some $\alpha \in (0,1)$). When possible, HPD intervals are used as they provide intervals of minimum length for a fixed level $1-\alpha$ [3]. We note that without the prior on t_0 (i.e., for the two-parameter case), neither the 99% nor the 95% credible interval contains g_{ref} and $\mathbb{P}(g \geq g_{\text{ref}} \mid \boldsymbol{d}) \approx 0.006$. Hence, g_{ref} seems to be an unlikely value under this posterior distribution. In contrast, the corresponding intervals both contain g_{ref} when using the prior on t_0. This prior adds variability to the posterior, which makes them right-skewed and yields intervals that are wide enough to include g_{ref}.

Throughout the rest of the paper, Bayesian results will be compared primarily through their posterior means, $\mathbb{E}(\,(g,C) \mid \boldsymbol{d})$ or $\mathbb{E}(\,(g,C,t_0) \mid \boldsymbol{d})$, and credible intervals. To compare these results to those obtained with non-Bayesian methods, we will use the frequentist coverage of credible intervals, which will be defined in section 3. Results from the various methods we consider are summarized and discussed in section 7.

[2] In more complex problems, one must approximate integrals using, for example, Markov chain Monte Carlo sampling, and in that case one would also need to check convergence of Markov chains in the validation process.

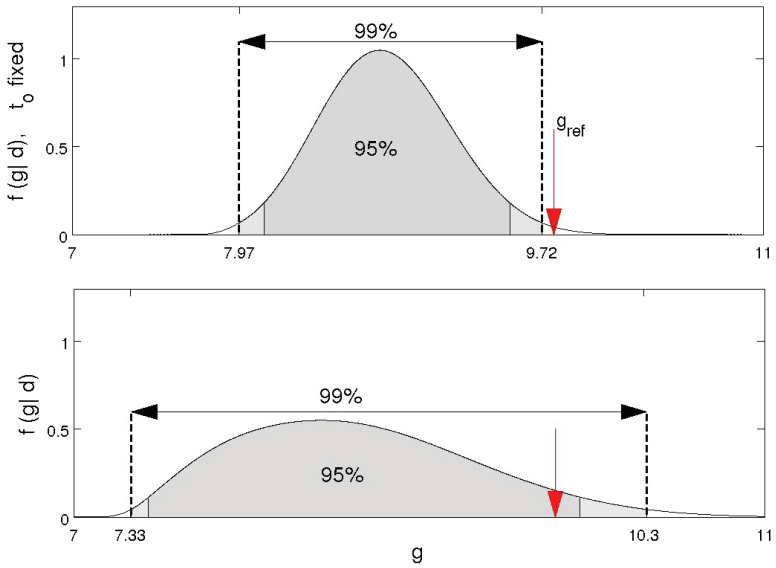

Fig. 2.2 *Top: Posterior of g corresponding to likelihood* (2.2) *and prior* (2.4). *The shaded areas define the* 95% *and* 99% *HPD credible intervals for g. Bottom: Corresponding results for prior* (2.5). *The value g_{ref}* (1.2) *is shown for reference and marked by the vertical arrows.*

Let us now conclude our summary of A13. There, the final result was the posterior density for g under prior (2.5), and corresponding credible intervals. One of the points of this paper is a validation analysis. Such checks are necessary to make sure the statistical model used is appropriate and the estimates are meaningful.

3. A First Validation Check. Assume for the moment that the likelihood function (2.2) is appropriate. If we asked the Texas A&M team to repeat the experiment 100 times with exactly the same setup, and if for each experiment they constructed 95% credible intervals for g as above, would they find that approximately 95 of the intervals contain the "true" g_{ref}? In statistics this is known as a check of the frequentist coverage of the credible intervals constructed above. If the frequentist coverage matches the target 95%, then we may conclude that despite the ad-hoc choice of priors, the Bayesian credible intervals can be interpreted as frequentist 95% confidence intervals. This provides one possible interpretation of the posterior credible intervals, and is reassuring to those who are concerned with subjectivity in the selection of the priors. For further discussion and examples of frequentist performance of Bayesian procedures, see [3, 10, 20].

We use simulations to obtain information about the frequentist coverage and mean length of the credible intervals using the procedure described in Algorithm 1. The basic idea is to generate synthetic data from a fixed $\widehat{\boldsymbol{m}}$, which we assume to be the "true" $\boldsymbol{m}$, and to repeat the estimation process many times. This method is known as parametric bootstrap [5, 8] because the noise is sampled from a parametric family (a beta distribution in this case). If the likelihood is correct, the simulations can be thought of as synthetic repetitions of the experiment with a fixed known $\boldsymbol{m}$.

The results of the simulations for 95% credible intervals are shown in Table 3.1. The table shows the usual 95% Agresti–Coull intervals [2] for the coverage and 95%

Algorithm 1 Simulations to estimate the frequentist coverage and mean length of $1-\alpha$ intervals $I_\alpha(g)$ for g.

fix: $\boldsymbol{m} = \widehat{\boldsymbol{m}}$, or $\boldsymbol{m} = (g_{\text{ref}}, \widehat{C}, t_0)$, or $\boldsymbol{m} = (g_{\text{ref}}, \widehat{C}, \widehat{t}_0)$ as the "true" parameters
for $k = 1 : N$ **do**
  generate: $U_1, \ldots, U_{31}$ i.i.d. beta$(2,2)$
  define $\varepsilon_i^* = \delta_i(2U_i - 1)$, $\boldsymbol{d}^* = \boldsymbol{z}(\boldsymbol{m}) + \boldsymbol{\varepsilon}^*$
  construct $1-\alpha$ credible interval $I_\alpha(g)$ for g by evaluating $\boldsymbol{f}_{\boldsymbol{m}|\boldsymbol{d}}(\boldsymbol{m}|\boldsymbol{d}^*)$
end for
Frequentist coverage is estimated by the proportion of intervals that contain the "true" g.

Table 3.1 *Frequentist coverage and mean length of the* 95% *credible intervals based on likelihood* (2.2) *and the two different priors on* $\boldsymbol{m}$, (2.4) *and* (2.5).

	$\widehat{g} = \mathbb{E}(g \mid \boldsymbol{d})$		$\widehat{g} = g_{\text{ref}}$	
$\pi(\boldsymbol{m})$	Coverage	Average length	Coverage	Average length
(2.4)	95.13 ± 0.01	0.83 ± 0.01	94.63 ± 0.01	0.84 ± 0.01
(2.5)	97.12 ± 0.01	2.03 ± 0.02	97.02 ± 0.01	2.20 ± 0.02

large-sample confidence intervals for the mean interval length [16]. We see that the intervals based on the prior with t_0 fixed have approximately 95% frequentist coverage but those based on the prior (2.5) are much wider and more conservative. Judging from this frequentist performance, we may question whether the prior for t_0 adds variability (thus making the intervals wide enough to cover g_{ref}) that may not be justified. We will address this question in the next section by checking the likelihood function.

4. Model Validation.

The results of the previous section have revealed possible problems with the credible intervals for three parameters as derived in A13. We will show in this section that the problem stems from an overestimation of the measurement uncertainties.

4.1. Checking the Noise Level.

We start by validating the assumed distribution of the noise whose PDF is given by (2.1). As mentioned before, this probability density was chosen by physical intuition, but is the distribution of the measurement noise assumed by the authors of A13 appropriate? The idea of this section is to "get to the noise" $\boldsymbol{\varepsilon}$ by somehow subtracting $\boldsymbol{z}(\boldsymbol{m})$ from $\boldsymbol{d}$. Since in this case estimating $\boldsymbol{m}$ is not the goal, we may take advantage of a surrogate model for $\boldsymbol{z}(\boldsymbol{m})$ that leads to residuals that are easier to analyze. We first note that a cubic fit to $z_i(\boldsymbol{m})$ (i.e., polynomial regression [6]) provides a good approximation. For example, the left panel in Figure 4.1 shows values of $\log\cosh\left[\sqrt{gC}\,t\right]/C$ (circles) at different times, and its cubic fit (lines) for three different choices of g and C. In each case there is a reasonably good agreement (the maximum difference is of order 10^{-3}). Therefore, to study the noise distribution, we model the data as $\boldsymbol{d}_\delta = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\delta$, where $\boldsymbol{d}_\delta = \boldsymbol{D}\boldsymbol{d}$, $\boldsymbol{\varepsilon}_\delta = \boldsymbol{D}\boldsymbol{\varepsilon}$, $\boldsymbol{D} = \text{diag}\{1/\delta_1, \ldots, 1/\delta_{31}\}$, and the 31×4 matrix $\boldsymbol{X}$ has columns $(1/\delta_i)$, (t_i/δ_i), (t_i^2/δ_i), and (t_i^3/δ_i). Note that $\boldsymbol{\varepsilon}_\delta$ is unit-less. The elements of the vector $\boldsymbol{\beta}$ we want to estimate are then the expansion coefficients of the cubic polynomial we want to fit to the data. The entries of the vector of differences between data and cubic model,

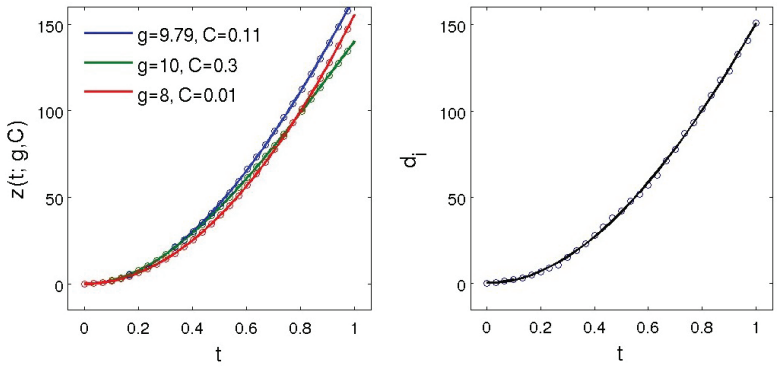

Fig. 4.1 *Left: Function $z(\boldsymbol{m})$ (1.3) as a function of time for three different choices of g and C. The lines show the cubic approximations. Right: Cubic fit (line) to the data (circles).*

$\boldsymbol{\varepsilon}_\delta$, are independent, zero-mean, with

$$(\boldsymbol{\varepsilon}_\delta)_i = 2U_i - 1, \quad U_i \text{ i.i.d. beta}(2,2). \tag{4.1}$$

We use least-squares to estimate $\boldsymbol{\beta}$: we find the vector $\widehat{\boldsymbol{\beta}}$ that minimizes $\|\boldsymbol{d}_\delta - \boldsymbol{X}\boldsymbol{\beta}\|^2$ over $\boldsymbol{\beta} \in \mathbb{R}^4$, in other words, a weighted least-squares fit to $\boldsymbol{d}$, where each point is weighted by $1/\delta_i$. The fit is shown in the right panel of Figure 4.1. The residual vector of the fit, $\boldsymbol{r} = \boldsymbol{d}_\delta - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, has zero mean and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{r}} = \sigma^2\boldsymbol{Q}$, where $\sigma^2 = \mathbb{V}\text{ar}(2U) = 1/5$ and $\boldsymbol{Q}$ is the projection matrix onto the orthogonal complement of the column space of $\boldsymbol{X}$ (see, for example, [6]): $\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$.

We transform the residuals to correct for their correlation and heterogeneous variances. Since $\boldsymbol{Q}$ is a projection matrix with a four-dimensional nullspace, we can write its SVD as

$$\boldsymbol{Q} = \boldsymbol{U}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{U}^\top = \begin{pmatrix}\boldsymbol{U}_1 & \boldsymbol{U}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\begin{pmatrix}\boldsymbol{U}_1^\top\\ \boldsymbol{U}_2^\top\end{pmatrix} = \boldsymbol{U}_1\boldsymbol{U}_1^\top,$$

where $\boldsymbol{U}$ is an orthogonal matrix, $\boldsymbol{I}$ is the 27×27 identity matrix, and $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ have orthonormal columns. Furthermore, since $\boldsymbol{r} = \boldsymbol{Q}\boldsymbol{\varepsilon}_\delta$, it follows that

$$\widetilde{\boldsymbol{r}} \equiv \boldsymbol{U}_1^\top\boldsymbol{r} = \widetilde{\boldsymbol{\varepsilon}} \equiv \boldsymbol{U}_1^\top\boldsymbol{\varepsilon}_\delta. \tag{4.2}$$

In particular, $\widetilde{\boldsymbol{r}}$ and $\widetilde{\boldsymbol{\varepsilon}}$ have the same distribution, which helps us check the posited distribution of $\boldsymbol{\varepsilon}$. To compare distributions, we will use empirical cumulative distribution functions (CDFs), which are often used for validation and do not require the choice of bins as do histograms (see, for example, [16]). We simulate `nsims`=10,000 noise vectors $\boldsymbol{\varepsilon}_\delta$ with distribution (4.1). For each realization of $\boldsymbol{\varepsilon}_\delta$, we compute the empirical CDF, $\widehat{F}_{\widetilde{\boldsymbol{\varepsilon}}}(x)$, of $\widetilde{\boldsymbol{\varepsilon}}$ on a uniform grid of values, x, of the random variable. The solid red line in the left panel in Figure 4.2 is the sample mean approximation of $\mathbb{E}[\,\widehat{F}_{\widetilde{\boldsymbol{\varepsilon}}}(x)\,]$ (i.e., $\texttt{mean}(\widehat{F}_{\widetilde{\boldsymbol{\varepsilon}}}(x))$), and the dashed lines define 95% confidence intervals for $\mathbb{E}[\,\widehat{F}_{\widetilde{\boldsymbol{\varepsilon}}}(x)\,]$ given by $\texttt{mean}(\widehat{F}_{\widetilde{\boldsymbol{\varepsilon}}}(x)) \pm 1.96\,\texttt{std}(\widehat{F}_{\widetilde{\boldsymbol{\varepsilon}}}(x))/\sqrt{\texttt{nsims}}$, where `mean` and `std` denote the usual sample mean and sample standard deviation (over the `nsims` simulations). The empirical CDF of $\widetilde{\boldsymbol{r}}$ is shown in blue. The plot shows that the assumed distribution of the noise is not consistent with that of the noise in the data: the true measurement noise appears to be much more narrowly centered around zero than the

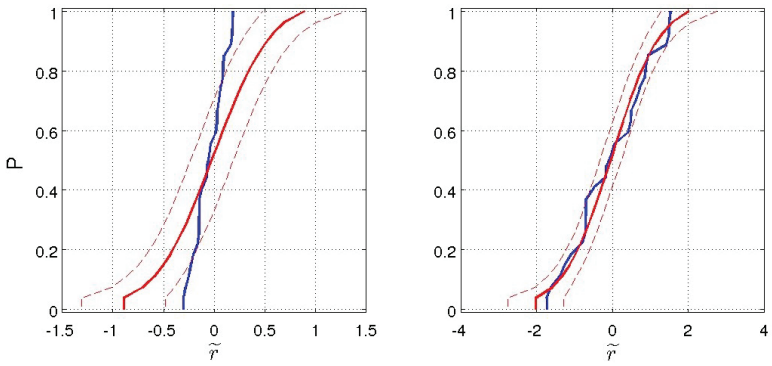

Fig. 4.2 *Left: Empirical CDF of $\widetilde{\boldsymbol{r}}$ (blue) and the mean and standard deviation of the empirical CDF of $\widetilde{\boldsymbol{\varepsilon}}$ (red) when the noise is assumed to have PDF* (2.1). *Right: Empirical CDF of $\widetilde{\boldsymbol{r}}$ (blue) and empirical CDF of a standardized sequence of Gaussian random variables (red).*

modeled uncertainty described by the noise PDF. In fact, even the standard deviations are not consistent: our estimate of the standard deviation of d_i is $0.15\delta_i$ as compared to the assumed $\sqrt{1/5}\,\delta_i \approx 0.45\delta_i$. It turns out that we can use a Gaussian distribution to model these data. To see this, note that if $\boldsymbol{\varepsilon}$ is multivariate Gaussian, then so are $\boldsymbol{\varepsilon}_\delta$ and $\boldsymbol{U}_1^\top \boldsymbol{\varepsilon}_\delta$, Moreover, the entries of $\boldsymbol{U}_1^\top \boldsymbol{\varepsilon}_\delta$ are i.i.d. Gaussian because the columns of $\boldsymbol{U}_1$ are orthonormal. This means that the entries of $\widetilde{\boldsymbol{r}}$ are also i.i.d. Gaussian and therefore each of the corrected entries of $\widetilde{\boldsymbol{r}}$, $T_i = (\widetilde{r}_i - \texttt{mean}(\widetilde{\boldsymbol{r}}))/\texttt{std}(\widetilde{\boldsymbol{r}})$, has a Student t_{26} distribution. Since these centered residuals are correlated, we conduct simulations to estimate the pointwise mean and standard deviation of the empirical CDF of 27 i.i.d. Gaussian variables that are corrected by subtracting their sample mean and dividing by their sample standard deviation. The right panel in Figure 4.2 shows the results. The blue line corresponds to the empirical CDF of the variables T_i. We see that this Gaussian distribution seems more reasonable than the scaled/shifted beta$(2, 2)$ distribution. Other distributions could be used but the Gaussian is convenient, especially for the estimation methods we will use below.

The primary conclusion of this analysis is that while the authors of A13 might have had good reasons to choose their noise model (2.1) with measured values for δ_i, the noise in the data does not match this assumption; the true distribution has a smaller standard deviation. In other words, their distributional choice was overly cautious, a fact already acknowledged in A13. The consequence is that their credible intervals are unnecessarily large and not backed up by the actual data.

The analysis above leads us to assume instead that the variables ε_i/δ_i are i.i.d. $N(0, \sigma^2_{\text{cubic}})$, where σ^2_{cubic} is the standard least-squares estimate of σ^2 (unit-less) obtained from the cubic fit. Thus, the likelihood is now

$$L(\boldsymbol{m}; \boldsymbol{d}) \;\propto\; \prod_{i=1}^{31} \exp\left[-\frac{(d_i - z_i(\boldsymbol{m}))^2}{2\,\delta_i^2 \sigma^2_{\text{cubic}}}\right], \tag{4.3}$$

and we can repeat the analysis to obtain

$$\mathbb{E}(\,(g, C) \mid \boldsymbol{d}) = (8.77,\ 0.11) \quad \text{and} \quad \mathbb{E}(\,(g, C, t_0) \mid \boldsymbol{d}) = (8.47,\ 0.09,\ 1.39). \tag{4.4}$$

The largest difference between these results and those of A13 (stated in (2.6)) is in the posterior mean of g. This is more evident when we compare the posterior distributions.

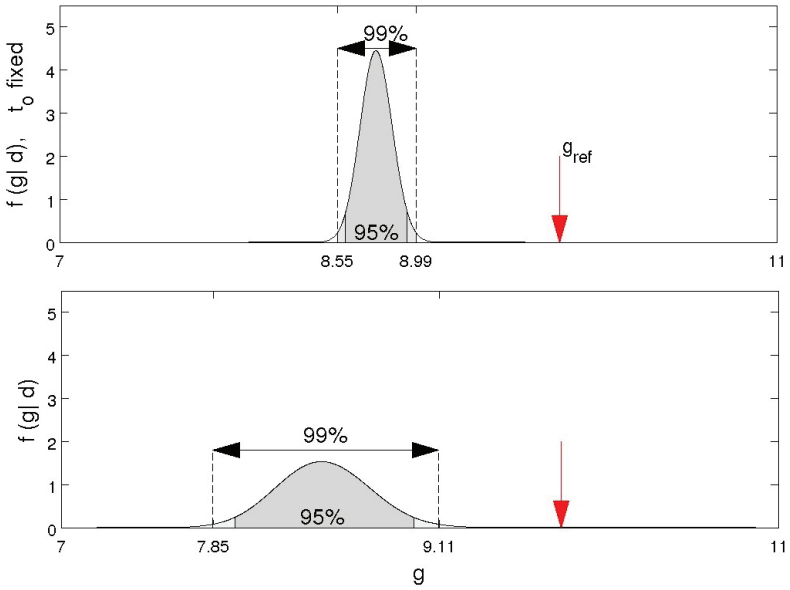

Fig. 4.3 *Same as Figure* 2.2 *but using the Gaussian likelihood* (4.3).

Table 4.1 *Same as Table* 3.1 *but using the Gaussian likelihood* (4.3).

	$\widehat{g} = \mathbb{E}(g \mid \boldsymbol{d})$		$\widehat{g} = g_{\text{ref}}$	
$\pi(\boldsymbol{m})$	Coverage	Average length	Coverage	Average length
(2.4)	95.03 ± 0.01	0.34 ± 0.01	95.13 ± 0.01	0.35 ± 0.01
(2.5)	94.13 ± 0.01	0.99 ± 0.01	94.23 ± 0.01	1.01 ± 0.01

Figure 4.3 shows that the new posteriors are indeed tighter and more symmetric than before, and that g_{ref} is an even more unlikely value than under likelihood (2.2), whether or not we use the prior on t_0. The 95% and 99% HPD credible intervals for g are now (compare these with (2.7))

	Prior	95%	99%
(4.5a)	(2.4) :	[8.60, 8.94]	[8.55, 8.99]
(4.5b)	(2.5) :	[7.97, 8.97]	[7.85, 9.11].

The frequentist coverage of these intervals is given in Table 4.1. The new coverage for prior (2.4) is closer to 95% and it is achieved with intervals that are, on average, less than half the size. The coverage for (2.5) is also closer to the target 95%, again with intervals half the size. This shows that the new likelihood, in addition to being more consistent with the noise in the data, leads to Bayesian results that have good frequentist performance.

4.2. Reconsidering the Use of Individual Uncertainty Measures δ_i. It may be intuitively appealing to control the uncertainty at every point with the uncertainty factors δ_i, as is done in A13, but is it necessary? One can argue that these uncertainties reflect the rotation of the falling box and errors caused by the resolution of the camera,

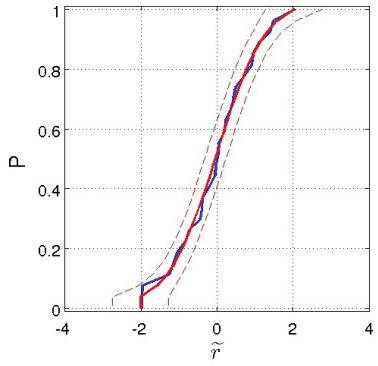

Fig. 4.4 *Same as in the right panel in Figure* 4.2 *but using the likelihood* (4.6) *without* δ_i.

but we believe that the uncertainty in d_i may not change much in time and that it may not be necessary to include different δ_i in the likelihood. In this case, we could compute the cubic approximation $\boldsymbol{\beta}$ weighing all points equally. Figure 4.4 is the equivalent of the right panel in Figure 4.2 but using this equally weighted cubic fit. The results seem to indicate that the residuals are also consistent with the noise when we use the likelihood with equal δ_i. In fact, this plot looks better than the one that uses the δ_i. In the spirit of using the simplest model consistent with the data, we shall assume henceforth that the errors ε_i are i.i.d. $N(0, \sigma^2_{\text{cubic}})$, where σ^2_{cubic} is now obtained from the cubic fit without δ_i (this time σ_{cubic} has the same units as d_i). Thus, the likelihood that will be used in the rest of the paper is:

$$L(\boldsymbol{m}; \boldsymbol{d}) \;\propto\; \prod_{i=1}^{31} \exp\left[-\frac{(d_i - z_i(\boldsymbol{m}))^2}{2\,\sigma^2_{\text{cubic}}}\right], \tag{4.6}$$

which leads to the posterior means

$$\mathbb{E}(\,(g, C) \mid \boldsymbol{d}) = (8.77,\ 0.11) \quad \text{and} \quad \mathbb{E}(\,(g, C, t_0) \mid \boldsymbol{d}) = (8.47,\ 0.09,\ 1.39). \tag{4.7}$$

These are almost the same as (4.4), which were obtained using δ_i. The corresponding posterior distributions are shown in Figure 4.5. The new credible intervals for g are

	Prior	95%	99%
(4.8a)	(2.4) :	[8.61, 8.94]	[8.56, 8.99]
(4.8b)	(2.5) :	[7.87, 9.07]	[7.72, 9.24].

There are some small differences with (4.5), but to compare them, it is better to look at their frequentist coverage and mean length. The estimated frequentist coverage of the 95% credible intervals is shown in Table 4.2. The coverage and mean length of the intervals under prior (2.4) are similar to those that used different δ_i. The coverage under (2.5) is a little above the target (compared to slightly below with the δ_i) and the intervals are slightly wider. However, overall, the use of δ_i seems to make little difference.

4.3. Prior and Posterior Predictive Checks. If the likelihood function and prior distribution provide an appropriate model for the data, we should be able to generate

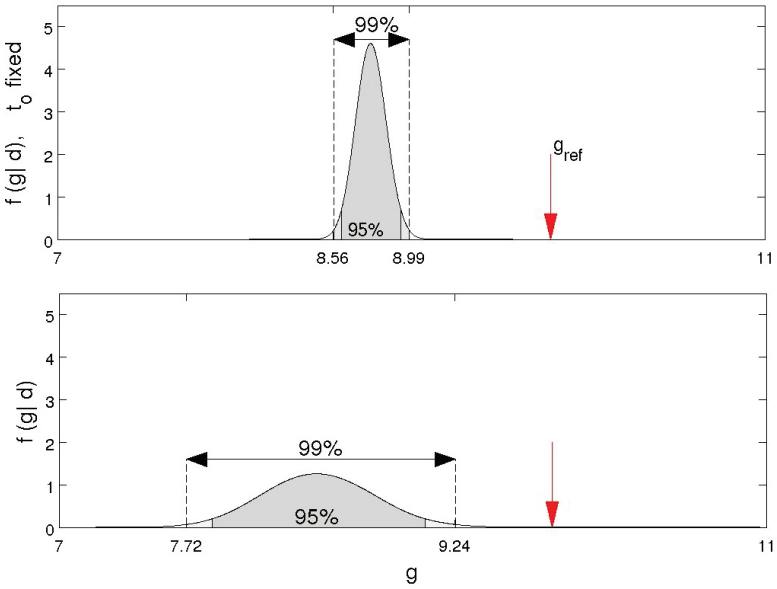

Fig. 4.5 *Same as Figure* 2.2 *but using the likelihood* (4.6) *that does not use* δ_i.

Table 4.2 *Same as Table* 3.1 *but using the likelihood* (4.6) *that does not use* δ_i.

	$\widehat{g} = \mathbb{E}(g \mid \boldsymbol{d})$		$\widehat{g} = g_{\text{ref}}$	
$\pi(\boldsymbol{m})$	Coverage	Average length	Coverage	Average length
(2.4)	95.23 ± 0.01	0.33 ± 0.01	95.03 ± 0.01	0.34 ± 0.01
(2.5)	95.72 ± 0.01	1.18 ± 0.01	95.53 ± 0.01	1.23 ± 0.01

data that are similar to the original $\boldsymbol{d}$. For example, we can simulate data $\mathcal{D}_{\text{pri}} = \{\boldsymbol{d}_1^*, \ldots, \boldsymbol{d}_m^*\}$ by drawing m samples $\boldsymbol{m}_1^*, \ldots, \boldsymbol{m}_m^*$ from the prior distribution and then generating $\boldsymbol{d}_i^*$ from the conditional distribution $F_{\boldsymbol{d}|\boldsymbol{m}}(\boldsymbol{d} \mid \boldsymbol{m}_i^*)$ using the forward model (1.1). The distribution of the samples in $\mathcal{D}_{\text{pri}}$ is called a prior predictive distribution. If instead we sample $\boldsymbol{m}_i^*$ from the posterior distribution of $\boldsymbol{m}$ conditional on $\boldsymbol{d}_i^*$, then the distribution of the sample, $\mathcal{D}_{\text{post}}$, is called a posterior predictive distribution [3, 13]. If the prior and likelihood are reasonable, we would expect the characteristics of $\boldsymbol{d}$ to be similar to those of the samples in $\mathcal{D}_{\text{pri}}$ or $\mathcal{D}_{\text{post}}$. In particular, if $\widehat{\boldsymbol{\beta}}$ are the coefficients of the cubic fit to $\boldsymbol{d}$ and $\widehat{\boldsymbol{\beta}}_i^*$ those of the cubic fit to $\boldsymbol{d}_i^*$, then $\boldsymbol{\beta}$ should be consistent with the distribution of the $\widehat{\boldsymbol{\beta}}_i^*$. To check this we can, for example, compute the principal components (PCs) [14] of the data matrix $[\,\widehat{\boldsymbol{\beta}}_1^* \cdots \widehat{\boldsymbol{\beta}}_m^*\,]$ and plot the first two coefficients of $\widehat{\boldsymbol{\beta}}$ (white circle) and $\widehat{\boldsymbol{\beta}}_i^*$ (blue) with respect to the PCs (which account for more that 98% of the total variability). To sample from the posterior predictive distributions, we have used the Metropolis–Hastings algorithm [13]. Figure 4.6 shows the results. We would question the choice of priors if the white circles were far from the bulk of the blue points. In this case, we see that the PC coefficients of $\widehat{\boldsymbol{\beta}}$ are consistent with realizations of the PC coefficients of $\widehat{\boldsymbol{\beta}}^*$ under the assumed model with two or three parameters. The same happens for the prior predictive distributions.

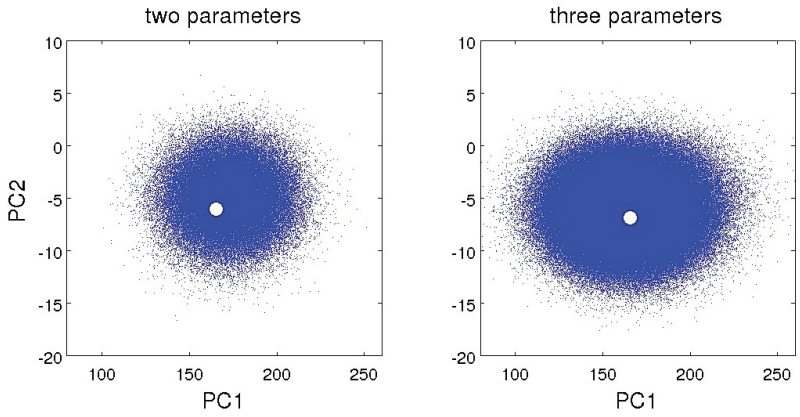

Fig. 4.6 *Left: Plot of $\boldsymbol{\beta}$ (white) and $\boldsymbol{\beta}_i^*$ (blue) in the first two PCs for prior* (2.4)*. Right: The same for prior* (2.5)*.*

To provide an example, we have used only one function of the data to check the priors. Usually we try to use functions that reduce the dimensionality of $\boldsymbol{d}$; it was reduced to two using the first two PCs. Many other functions, $T(\boldsymbol{d})$, can be used to compare to characteristics of $T(\boldsymbol{d}_i^*)$. Of course, we could find functions for which we see some discrepancies, but the idea is to choose a function that is relevant to the problem. More examples on the use of predictive distributions for validation can be found in [3, 13], and also in [10, 21] in the context of inverse problems.

5. Maximum Likelihood and Nonlinear Least-Squares. It is important to know how to use a variety of estimation methods that make different assumptions and incur different computational costs. Since different methods used soundly should lead to results that are consistent, comparing their results is one way to look for possible problems. Up to now we have only considered the Bayesian inference of $\boldsymbol{m}$. This parameter vector was modeled as random with a prior probability distribution, and the results of the inference were conditional on the data. But parameter estimates can also be obtained without assuming that $\boldsymbol{m}$ is random. Inferential methods in which the unknown parameter is fixed (not random) and whose performance is based on their distribution induced by different realizations of the data are called frequentist. An introduction to Bayesian and non-Bayesian methods for inverse problems can be found in [10]. A more general reference is [17].

5.1. Maximum Likelihood Estimates. Let us first consider the method of maximum likelihood (ML) to obtain parameter estimates without using prior distributions. It is based only on $\boldsymbol{d}$ and the corresponding likelihood function defined by (4.6). A value, $\widehat{\boldsymbol{m}}_{\text{ML}}$, that maximizes the likelihood, $L(\boldsymbol{m};\boldsymbol{d})$, is called an ML estimate of $\boldsymbol{m}$. To find $\widehat{\boldsymbol{m}}_{\text{ML}}$, we evaluate $L(\boldsymbol{m};\boldsymbol{d})$ on a grid of parameter values.[3] The ML estimates for the two cases $\boldsymbol{m} = (g, C)$ and $\boldsymbol{m} = (g, C, t_0)$ are, respectively,

$$(\widehat{g}, \widehat{C}\,)_{\text{ML}} = (8.77,\ 0.11) \quad \text{and} \quad (\widehat{g}, \widehat{C}, \widehat{t_0}\,)_{\text{ML}} = (8.41,\ 0.09,\ 1.39). \tag{5.1}$$

We observe that these ML estimates are almost identical to the corresponding posterior means under priors (2.4) and (2.5). This is not surprising given the flat priors for g and C and the "almost" flat prior for t_0.

[3] For consistency, we use the same grid of parameter values used to compute the integrals in the Bayesian calculations conducted by A13.

In addition to parameter estimates, ML can be used to obtain approximate confidence intervals, which are derived using the asymptotic theory of ML estimators. Under regularity conditions, ML estimators are asymptotically unbiased and Gaussian with covariance matrix equal to the inverse Fisher information matrix [4, 9]. The Fisher information matrix of a $k \times 1$ parameter vector $\boldsymbol{m}$ is the $k \times k$ matrix with entries

$$\boldsymbol{I}(\boldsymbol{m})_{ij} = \mathbb{E}\left[\frac{\partial^2(-\ln L(\boldsymbol{m};\boldsymbol{d}))}{\partial m_i \partial m_j}\right]. \tag{5.2}$$

An approximate $1-\alpha$ confidence interval for m_j is given by [4]

$$(\widehat{\boldsymbol{m}}_{\mathrm{ML}})_j \pm z_{\alpha/2}\sqrt{[\boldsymbol{I}(\widehat{\boldsymbol{m}}_{\mathrm{ML}})^{-1}]_{jj}}\,, \tag{5.3}$$

where $z_{\alpha/2}$ is the usual $1-\alpha/2$ quantile of the $N(0,1)$ distribution. When the expectation in (5.2) cannot be easily computed, it may be approximated using sample averages. This leads to the observed Fisher information matrix $\boldsymbol{J}(\boldsymbol{m})$,

$$\boldsymbol{J}(\boldsymbol{m})_{ij} = \sum_{\ell=1}^{K} \frac{\partial^2(-\ln L(\boldsymbol{m};d_\ell))}{\partial m_i \partial m_j}, \tag{5.4}$$

and approximate $1-\alpha$ confidence intervals $(\widehat{\boldsymbol{m}}_{\mathrm{ML}})_j \pm z_{\alpha/2}\sqrt{[\boldsymbol{J}(\widehat{\boldsymbol{m}}_{\mathrm{ML}})^{-1}]_{jj}}$ [7]. Since we are using a Gaussian likelihood, it is easy enough to use $\boldsymbol{I}(\widehat{\boldsymbol{m}}_{\mathrm{ML}})$ instead of $\boldsymbol{J}(\widehat{\boldsymbol{m}}_{\mathrm{ML}})$. The approximate 95% and 99% confidence intervals (5.3) for g for the two- and three-parameter cases are

	$\boldsymbol{m}$	95%	99%
(5.5a)	(g, C) :	[8.60, 8.94]	[8.55, 8.99]
(5.5b)	(g, C, t_0) :	[7.78, 9.03]	[7.59, 9.23].

The intervals for the two-parameter case are almost identical to the corresponding credible intervals (4.8). The intervals for the three-parameter case seem to be slightly wider than the corresponding credible intervals, but since the two types of intervals have different interpretations, we compare them through their frequentist coverage and mean lengths. Also, since the ML intervals are based on asymptotic approximations, we should assess their actual coverage with $n = 31$. To this end, we conduct simulations using the following slight modification of Algorithm 1: In the construction step, we compute ML intervals using (5.3) instead of credible intervals. The results are shown in Table 5.1. We see that the coverage and mean length of the ML intervals and credible intervals are very similar, although the latter seem to be slightly shorter.

Table 5.1 *Interval length and coverage results for the approximate* 95% *ML confidence intervals using the Gaussian likelihood* (4.6) *with two or three parameters.*

	$\widehat{g} = \widehat{g}_{\mathrm{ML}}$		$\widehat{g} = g_{\mathrm{ref}}$	
$\boldsymbol{m}$	Coverage	Average length	Coverage	Average length
(g, C)	94.83 ± 0.01	0.34 ± 0.01	94.33 ± 0.01	0.35 ± 0.01
(g, C, t_0)	94.93 ± 0.02	1.26 ± 0.01	95.15 ± 0.01	1.29 ± 0.01

5.2. Nonlinear Least-Squares. One advantage of least-squares (LS), linear or nonlinear, is that no distributional assumptions are required to obtain the estimates. We need only find a vector of parameters, $\widehat{\boldsymbol{m}}_{\rm NR}$, that minimizes $\|\boldsymbol{d}-\boldsymbol{z}(\boldsymbol{m})\|^2$. Knowledge of the noise distribution is needed, however, to construct confidence intervals. In the nonlinear Gaussian case, confidence sets are defined by applying the Gaussian theory to a linearization around the estimate. In the case when $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, the distribution of $\widehat{\boldsymbol{m}}_{\rm NR}$ is approximately $N(\boldsymbol{m}, \sigma^2[\,\boldsymbol{F}(\boldsymbol{m})^\top \boldsymbol{F}(\boldsymbol{m})\,]^{-1})$, where $\boldsymbol{F}(\boldsymbol{m})$ is the Jacobian matrix of the forward model with entries $\boldsymbol{F}(\boldsymbol{m})_{ij} = \partial z_i(\boldsymbol{m})/\partial m_j$; see [18]. An approximate $1-\alpha$ confidence interval for m_j is given by

$$(\widehat{\boldsymbol{m}}_{\rm NR})_j \pm t_{31-p}(\alpha/2)\,\widehat{\sigma}\sqrt{[(\,\boldsymbol{F}(\widehat{\boldsymbol{m}}_{\rm NR})^\top \boldsymbol{F}(\widehat{\boldsymbol{m}}_{\rm NR})\,)^{-1}]_{jj}}\,, \tag{5.6}$$

where the estimate of the variance, σ^2, is $\widehat{\sigma}^2 = \|\boldsymbol{d}-\boldsymbol{z}(\widehat{\boldsymbol{m}}_{\rm NR})\|^2/(31-p)$ with $p=2$ or $p=3$, depending on whether the fit is done to two or three parameters. We again use Algorithm 1 to assess the actual coverage of these approximate confidence intervals (using (5.6) in the construction step).

In the Gaussian likelihood case, ML and nonlinear LS estimates should be exactly the same. In our case there are small differences because the ML estimates were obtained using a grid (so as to be able to compare to the posterior results in A13), while the nonlinear LS fit was done using Levenberg–Marquadt search method [18]. The square root in (5.6) should also be the same as that in (5.3), but for the nonlinear LS estimates used here, σ^2 is estimated each time using the residuals of the nonlinear fit, not the residuals of the cubic fit to the surrogate model. Therefore, we expect small differences in the coverage and length of the confidence intervals. The reason for using these different procedures is to check consistency of the results under different but reasonable estimation methods.

The nonlinear LS estimates for the two- and three-parameter cases are

$$(\,\widehat{g}, \widehat{C}\,) = (8.77,\ 0.11) \quad \text{and} \quad (\,\widehat{g}, \widehat{C}, \widehat{t_0}\,) = (8.45,\ 0.09,\ 1.39). \tag{5.7}$$

The 95% and 99% confidence intervals for g are given by

	$\boldsymbol{m}$	95%	99%
(5.8a)	(g, C) :	[8.60, 8.94]	[8.54, 9.01]
(5.8b)	(g, C, t_0) :	[7.80, 9.10]	[7.57, 9.33].

Table 5.2 shows their estimate coverages and mean lengths. Again we get results consistent with ML and with the credible intervals. However, the nonlinear LS intervals tend to be wider than those from the other two methods. This is not surprising, as σ is fixed in the latter while it is estimated in each nonlinear fit.

If we had no idea about the distribution of the noise, we could still use nonlinear LS and we could find approximate confidence intervals using nonparametric bootstrap

Table 5.2 *Interval length and coverage results for the* 95% *nonlinear LS confidence intervals of* g *using likelihood* (4.6) *with two or three parameters.*

	$\widehat{g} = \widehat{g}_{\rm NR}$		$\widehat{g} = g_{\rm ref}$	
$\boldsymbol{m}$	Coverage	Average length	Coverage	Average length
(g, C)	94.33 ± 0.01	0.35 ± 0.01	94.33 ± 0.01	0.36 ± 0.01
(g, C, t_0)	95.23 ± 0.01	1.31 ± 0.01	95.13 ± 0.01	1.34 ± 0.01

methods. For example, in Algorithm 1, we can draw noise samples by sampling with replacements from a corrected version of the residuals of the surrogate model [5, 8, 10].

We can summarize this section by observing that both ML and nonlinear LS estimates and confidence intervals are very similar to those obtained in previous sections. We take this as additional validation that our results are consistent.

6. Confidence Intervals without Inversion. The confidence intervals used in section 5 are of the form

$$\text{estimate} \pm k_\alpha \times \text{ (approximate standard error of estimate)},$$

where the constant k_α is chosen to approximately achieve the desired confidence level $1-\alpha$. This requires solving the inverse problem. In this section we describe a method to construct approximate confidence intervals directly, without first finding an estimate, $\widehat{\boldsymbol{m}}$, and its standard error. The basic tool is a duality between hypothesis tests and confidence sets, which allows us to find a confidence set by inverting a test (e.g., [4, 15]). Some applications of this method to inverse problems can be found in [19].

For the likelihood (4.6), and under the null hypothesis that $\boldsymbol{m}$ is the true vector of parameters that generated $\boldsymbol{d}$, we have a χ^2 random variable with 31 degrees of freedom: $\|\boldsymbol{d}-\boldsymbol{z}(\boldsymbol{m})\|^2/\sigma^2 \sim \chi^2_{31}$. Therefore, the set

$$A_\alpha(\boldsymbol{m}) = \{\boldsymbol{d} \,:\, \|\boldsymbol{d}-\boldsymbol{z}(\boldsymbol{m})\|^2 \le \sigma^2\, \chi^2_{31}(\alpha)\}$$

defines an acceptance region of significance level α. A $1-\alpha$ confidence region, $C_\alpha(\boldsymbol{d})$, for $\boldsymbol{m}$ is defined by inverting the test: $C_\alpha(\boldsymbol{d}) = \{\boldsymbol{m} \,:\, \boldsymbol{d} \in A_\alpha(\boldsymbol{m})\}$. We use σ_{cubic} from section 4.2 to approximate σ. Figure 6.1 shows the boundary of the confidence region for the two-parameter case and $\alpha = 0.05$. It is approximately an ellipsoid that results from the intersection of the plane $h(\boldsymbol{m}) = \sigma^2_{\text{cubic}}\chi^2_{31}(\alpha)$ with the surface $h(\boldsymbol{m}) = \|\boldsymbol{d}-\boldsymbol{z}(\boldsymbol{m})\|^2$. By projecting this region onto each of the axes, we obtain conservative $1-\alpha$ confidence intervals for each parameter. The right panel in Figure 6.1 shows the 95% projected intervals for g and C. The 95% and 99% projected

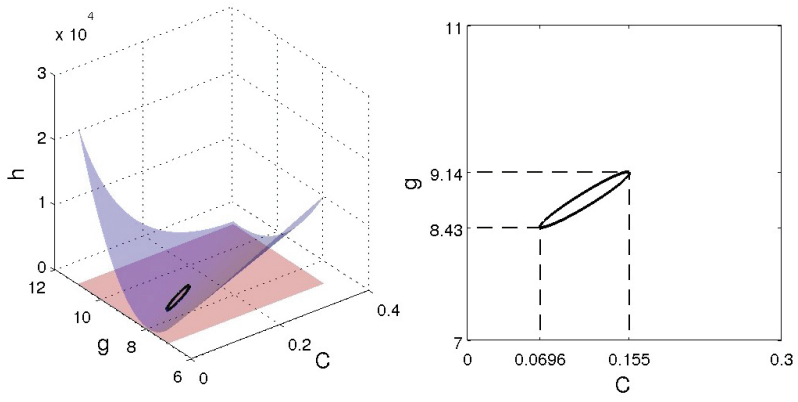

Fig. 6.1 *95% confidence region that results from the intersection of the plane* $h_1(\boldsymbol{m}) = \sigma^2_{\text{cubic}}\chi^2_{31}(0.05)$ *with the surface* $h_2(\boldsymbol{m}) = \|\boldsymbol{d}-\boldsymbol{z}(\boldsymbol{m})\|^2$ *for the two-parameter case.*

Table 6.1 *Interval length and coverage results for the projected 95% intervals of g for the two- and three-parameter cases and likelihood* (4.6).

	$\widehat{g} = \widehat{g}_{\mathrm{ML}}$		$\widehat{g} = g_{\mathrm{ref}}$	
$\boldsymbol{m}$	Coverage	Average length	Coverage	Average length
(g, C)	97.02 ± 0.01	0.66 ± 0.01	96.92 ± 0.01	0.68 ± 0.01
(g, C, t_0)	97.72 ± 0.01	2.60 ± 0.04	97.52 ± 0.01	2.60 ± 0.04

confidence intervals for g are

	$\boldsymbol{m}$	95%	99%
(6.1a)	(g, C) :	[8.43, 9.14]	[8.37, 9.21]
(6.1b)	(g, C, t_0) :	[7.24, 9.98]	[7.04, 10.31].

Table 6.1 shows the corresponding coverage results again obtained using again a modification of Algorithm 1. As expected, these intervals are more conservative and wider than all the others. However, an advantage of this method is that we only need to use the forward model and find a rejection region defined by the noise distribution, and so there is no need to solve the inverse problem. Note also that one could use simulations to tune the value of α so that the confidence level of the projected intervals is closer to 95%.

7. Discussion. The authors of A13 used real experimental data to obtain estimates of the value of g at the location where their experiment took place, College Station, TX. It was reassuring that the credible intervals under the three-parameter prior in A13 actually included g_{ref}. However, our validation analysis revealed that the spread of the posterior distribution, which allowed the inclusion of g_{ref} in the intervals, was actually caused by an artificially wide likelihood produced by an overcautious estimate of the measurement noise level. This is a reminder of the importance of checking the choice of likelihood function in addition to that of prior distributions. Once a better likelihood was selected, the credible intervals no longer covered g_{ref}. The prior distributions were chosen as is often done in practice; inequality constraints were enforced using flat priors. Given such ad-hoc choices, it is important to make sure the results are driven by the data and not by artificial information introduced by the priors. Prior and posterior predictive checks with the new likelihood did not reveal significant problems. In addition, the frequentist coverage of the credible intervals was near the target level. Thus, the priors do not seem to dominate the inference.

To validate the results, we also constructed confidence intervals using ML, nonlinear LS, and test inversion. Figure 7.1 provides a summary of all the 95% intervals for g found in the previous sections. Once the new likelihood function is used, the Bayesian and non-Bayesian approaches give consistent results as judged by their frequentist behavior and even by the similarity of the actual intervals; excluding the intervals (6.1), which are very conservative, the other intervals are quite similar. Although not shown here, we also obtained similar results using a uniform or a Gaussian distribution with both different and equal δ_i, in all cases modeled as symmetric about zero with a standard deviation that matches that of the actual noise in the data. Such agreement is expected in simple problems with only a few parameters to be estimated. It is in line with the Bernstein–von Mises theorem, which roughly states that if the amount of data (here $n = 31$) is large compared to the number of unknown parameters (here $p = 3$), then Bayesians and frequentists agree (see, for example,

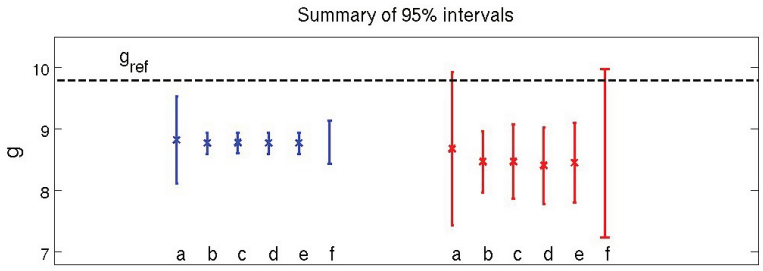

Fig. 7.1 95% *intervals for* g *for the case of two- (left group) and three-parameter (right group) vector* $\boldsymbol{m}$*:* (a) *credible intervals in* A13*;* (b) *and* (c) *are credible intervals* (4.5) *and* (4.8)*, respectively;* (d) *ML intervals* (5.5)*;* (e) *nonlinear LS intervals* (5.8)*;* (f) *projected intervals* (6.1).

[9, 11]). This is not necessarily true for more complex large-scale inverse problems [12, 10].

In addition to using frequentist methods to explore the validity of Bayesian procedures, we have seen how Bayesian methods can be used to derive procedures that have good frequentist properties. That is, just as ML provides a recipe to find estimators, we can often use Bayes theorem to do the same. Thus, Bayesian and frequentist methods of inference can be used in complementary ways.

It may be disappointing that our analysis here showed that the credible intervals are significantly smaller than those shown in A13 (primarily due to the fact that the measurement uncertainties are much smaller than assumed there) and do not include g_{ref}. However, this can be viewed in different ways. First, given that we know g_{ref}, this could indicate the presence of a systematic bias in the experiment that would not have been detectable using the original analysis without validation. Second, in practice parameters are of course unknown and the validation process shown here points out the importance of conducting a careful systematic error analysis to determine potential bias (and possible corrections). If the experiment described in A13 had been the first ever attempt to estimate g at College Station, the results would have underestimated its value, but we would have been unaware of this until a new, better experiment came along with a more accurate estimate. This is not unusual in science. Uncertainty estimates provide some guidance, but replication of the results with independent, better experiments is what scientists expect in order to accept a new result.

REFERENCES

[1] M. Allmaras, W. Bangerth, J. M. Linhart, J. Polanco, F. Wang, K. Wang, J. Webster, and S. Zedler, *Estimating parameters in physical models through Bayesian inversion: A complete example*, SIAM Rev., 55 (2013), pp. 149–167.

[2] L. D. Brown, T. T. Cai, and A. DasGupta, *Interval estimation for a binomial proportion*, Statist. Sci., 16 (2001), pp. 101–133.

[3] B. P. Carlin and T. A. Louis, *Bayesian Methods for Data Analysis*, 3rd ed., CRC Press, Boca Raton, FL, 2008.

[4] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed., Duxbury, Pacific Grove, CA, 2002.

[5] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and Their Applications*, Cambridge University Press, Cambridge, UK, 1997.

[6] N. R. Draper and H. Smith, *Applied Regression Analysis*, 3rd ed., Wiley, New York, 1998.

[7] B. Efron and D. V. Hinkley, *Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information*, Biometrika, 65 (1978), pp. 457–482.
[8] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London, 1993.
[9] T. S. Ferguson, *A Course in Large Sample Theory*, Chapman & Hall, London, 1996.
[10] C. Fox, Y. Marzouk, and L. Tenorio, *An Introduction to Data Analysis and Uncertainty Quantification for Inverse Problems*, in preparation.
[11] D. Freedman, *Some issues in the foundation of statistics*, Found. Sci., 1 (1995), pp. 19–39.
[12] D. Freedman, *On the Bernstein-von Mises theorem with infinite-dimensional parameters*, Ann. Statist., 27 (1999), pp. 1119–1140.
[13] A. Gelman, J. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, 2nd ed., Chapman & Hall, London, 2003.
[14] R. A. Johnson and D. W. Wichern, *Applied Multivariate Analysis*, 6th ed., Prentice-Hall, Englewood Cliffs, NJ, 2007.
[15] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypotheses*, Springer, New York, 2005.
[16] J. Rice, *Mathematical Statistics and Data Analysis*, 3rd ed., Duxbury, Belmont, CA, 2007.
[17] F. J. Samaniego, *A Comparison of the Bayesian and Frequentist Approaches to Estimation*, Springer, New York, 2010.
[18] G. A. F. Seber and C. J. Wild, *Nonlinear Regression*, Wiley, New York, 2003.
[19] P. B. Stark, *Inference in infinite-dimensional inverse problems: Discretization and duality*, J. Geophys. Res., 97 (1992), pp. 14055–14082.
[20] P. B. Stark and L. Tenorio, *A primer of frequentist and Bayesian inference* in Large-Scale Inverse Problems and Quantification of Uncertainty, Wiley, 2011, pp. 9–32.
[21] L. Tenorio, F. Andersson, M. de Hoop, and P. Ma, *Data analysis tools for uncertainty quantification of inverse problems*, Inverse Problems, 27 (2011), 045001.