NPL Report CMSC 38/04

Report to the National Measurement
System Policy Unit, Department of Trade
and Industry

From the Software Support for Metrology
Programme

Parameter estimation methods for data fusion

By
R Boudjemaa and A B Forbes

February 2004

Parameter estimation methods for data fusion

R Boudjemaa and A B Forbes
Centre for Mathematics and Scientific Computing

February 2004

ABSTRACT

Many metrology systems involve more than one sensor and the analysis of the data produced by these systems has to take into account the characteristics of the data arising from the different sensors. For well-characterized systems in which the behaviour of the sensors is known *a priori*, appropriate methods for estimating the parameters of the system from measurement data can be derived according to maximum likelihood principles. For systems subject to unknown or unpredictable variations, estimation methods that can adapt to these variations are required. In this report, we show how a class of such methods based on Bayesian approaches can be defined and illustrate their behaviour on a number of examples relevant to metrology.

ISSN 1471–0005

Authorised by Dr Dave Rayner,
Head of the Centre for Mathematics and Scientific Computing

National Physical Laboratory,
Queens Road, Teddington, Middlesex, United Kingdom TW11 0LW

Contents

1 Introduction

In order to improve the accuracy of measurement systems more factors have to be included in the model and monitored through a number of sensors. We use the term *data fusion* [19] for the aggregation and analysis of the resulting multi-sensor measurement data. Data fusion problems can arise in metrology in a number of ways, including the following:

Fusion across sensors. In this situation, a number of sensors measure nominally the same quantity, as, for example, in the case of a number of temperature sensors measuring the temperature of an object.

Fusion across attributes. In this situation, a number of sensors measure different quantities associated with the same experimental situation, as, for example, in the measurement of air temperature, pressure and humidity to determine air refractive index.

Fusion across domains. In this situation, a number of sensors measure the same attribute over a number of different ranges or domains. This arises, for example, in the definition of the temperature and pressure scales, geodesy, triangulation, photogrammetry, and theodolite metrology.

Fusion across time. In this situation, current measurements are fused with historical information, for example, from an earlier calibration [12]. Often the current information is not sufficient to determine the system (accurately) and historical information has to be incorporated to determine the system (accurately).

Consider, for example, the measurement of the length of a gauge block by a laser-interferometric system. The displacement measured by the interferometer depends on the refractive index of air which in turn depends on air temperature, pressure and humidity. The length of the gauge block also depends on its own temperature so this also has to be monitored. Thus, in order to estimate the gauge block length, at least five sensors are required to measure interferometer fringe counts, air temperature, pressure, humidity and artefact temperature.

In this report we consider parameter estimation techniques for analyzing data associated with measurement systems subject to a number of different random effects. In determining the parameter estimates the nature of these random effects have to be taken into account. For well-characterized systems, it is usually possible to design an estimation algorithm that gives appropriate weight to the different types of data in order to arrive at these estimates. For systems that are only partially characterized, the way forward is less straightforward and it is possible that our prior expectations could lead to an analysis method that makes poor use of the data. We look instead for methods that take into account our prior expectations but are flexible enough to make adjustments in the light of the data to hand. As with all parameter estimation methods in metrology, we are required to provide uncertainties associated with the estimates of fitted parameters.

This report is organized as follows. In section 2, we give an overview of parameter estimation in the context of data fusion and describe the generalized maximum likelihood estimation approach (GMLE). In sections 3 to 6, we illustrate how the GMLE approach can be applied to problems relevant to metrology. In section 7, we discuss the algorithmic and software requirements for GMLE. In section 8, we present a summarizing discussion.

2 Data fusion and parameter estimation

In this section we give an overview of an approach to data fusion that generalizes standard approaches used in metrological data analysis. We suppose we are interested in a finite set of parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ and that we wish to provide estimates of $\boldsymbol{\alpha}$ and associated uncertainties derived from measurement information. We illustrate standard approaches to parameter estimation on a response calibration.

2.1 Classical least-squares analysis for linear models

Suppose we have a linear model in which the response $\eta = \phi(\xi, \boldsymbol{\alpha})$ depends on a variable ξ and parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)^{\mathrm{T}}$ with

$$\phi(\xi, \boldsymbol{\alpha}) = \sum_{j=1}^{n} \alpha_j \phi_j(\xi).$$

We assume that measurements y_i of $\eta_i = \phi(\xi_i, \boldsymbol{\alpha})$ are made corresponding to known values ξ_i, $i = 1, \ldots, m$, $m \geq n$. To model the random effects associated with the measurements of η, we suppose that

$$y_i = \eta_i + \epsilon_i,$$

where ϵ_i is an observation of a random variable E_i and $\boldsymbol{E} = (E_1, \ldots, E_m)^{\mathrm{T}}$ is such that its expectation $E(\boldsymbol{E}) = \mathbf{0}$ and it variance $\mathrm{Var}(\boldsymbol{E}) = \sigma^2 I$ [24]. Equivalently, y_i is an observation of the random variable Y_i with $Y_i = \eta_i + E_i$. Let C be the fixed, $m \times n$ observation matrix with $C_{ij} = \phi_j(\xi_i)$ and let $\mathbf{c}_i^{\mathrm{T}}$ be the ith row of C. The least-squares estimate $\mathbf{a}$ of $\boldsymbol{\alpha}$, given $\mathbf{y}$, minimizes

$$F(\boldsymbol{\alpha}|\mathbf{y}) = (\mathbf{y} - C\boldsymbol{\alpha})^{\mathrm{T}}(\mathbf{y} - C\boldsymbol{\alpha}) = \sum_{i=1}^{m} (y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2.$$

The solution estimate is $\mathbf{a} = C^\dagger \mathbf{y}$, where $C^\dagger = (C^{\mathrm{T}}C)^{-1}C^{\mathrm{T}}$ is the *pseudo-inverse* of C. If C has QR decomposition

$$C = QR = \left[\begin{array}{cc} Q_1 & Q_2 \end{array}\right] \left[\begin{array}{c} R_1 \\ \mathbf{0} \end{array}\right] = Q_1 R_1, \qquad (1)$$

where Q is an orthogonal matrix and R is upper-triangular [8, 17], then $C^\dagger = R^{-1}Q_1^{\mathrm{T}}$. We note that the estimate $\mathbf{a}$ does not depend on σ.

2.1.1 Uncertainty matrix associated with the least squares estimate

For fixed C and $\boldsymbol{Y}$ an m-vector of random variables, the equation $\boldsymbol{A} = C^\dagger \boldsymbol{Y}$ defines the n-vector $\boldsymbol{A}$ of random variables as linear combinations of $\boldsymbol{Y}$. Taking expectations, we have

$$E(\boldsymbol{A}) = E(C^\dagger \boldsymbol{Y}) = C^\dagger \boldsymbol{\eta} = C^\dagger C \boldsymbol{\alpha} = \boldsymbol{\alpha},$$

and the covariance matrix associated with $\boldsymbol{A}$ is

$$V_{\boldsymbol{A}} = \mathrm{Var}(\boldsymbol{A}) = C^\dagger \sigma^2 I (C^\dagger)^{\mathrm{T}} = \sigma^2 (C^{\mathrm{T}}C)^{-1}. \qquad (2)$$

We note here that $V_{\boldsymbol{A}}$ does not depend on the observations $\mathbf{y}$. If $\mathbf{y}$ is sampled from $\boldsymbol{Y}$ with $E(\boldsymbol{Y}) = C\boldsymbol{\alpha}$, $\mathrm{Var}(\boldsymbol{Y}) = \sigma^2 I$, then $\mathbf{a}$ is sampled from $\boldsymbol{A}$ with $E(\boldsymbol{A}) = \boldsymbol{\alpha}$, $\mathrm{Var}(\boldsymbol{A}) = V_{\boldsymbol{A}}$.

2.1.2 Posterior estimate of σ

If $X_i \sim N(0,1)$, $i = 1, \ldots, m$, are independent normal variates then $\sum_{i=1}^m X_i^2$ has a χ_m^2 distribution with $E(\chi_m^2) = m$ and $\mathrm{Var}(\chi_m^2) = 2m$. Let $\boldsymbol{R}$ be the random vector of residuals so that

$$\boldsymbol{R} = \boldsymbol{Y} - C\boldsymbol{A} = \boldsymbol{Y} - CC^\dagger \boldsymbol{Y} = (I - CC^\dagger)\boldsymbol{Y}.$$

If $C = Q_1 R_1$ as in (1), then $CC^\dagger = Q_1 Q_1^{\mathrm{T}}$ and $I - Q_1 Q_1^{\mathrm{T}} = Q_2 Q_2^{\mathrm{T}}$, so that

$$S^2 = \boldsymbol{R}^{\mathrm{T}}\boldsymbol{R} = \left(Q_2^{\mathrm{T}}\boldsymbol{Y}\right)^{\mathrm{T}} Q_2^{\mathrm{T}}\boldsymbol{Y}.$$

Now Q is orthogonal so setting $\tilde{\boldsymbol{Y}} = Q\boldsymbol{Y}$ we have $\mathrm{Var}(\tilde{\boldsymbol{Y}}) = I$ also. Therefore, $S^2 = \sum_{i=n+1}^m \tilde{Y}_i^2$ is a sum of squares of $m-n$ independent, normal variates and has a χ_ν^2 distribution with $\nu = m - n$ degrees of freedom, with $E(S^2) = \nu$, $\mathrm{Var}(S^2) = 2\nu$. From this analysis, we see that given a least squares solution $\mathbf{a}$, a posterior estimate of the input σ is s with

$$s^2 = \frac{\mathbf{r}^{\mathrm{T}}\mathbf{r}}{m-n}. \qquad (3)$$

While this estimate is derived under the assumption that the random effects are governed by a Gaussian distribution, it is likely to be a good approximation for distributions with similar features, e.g., unimodal (that is, having one peak). A posterior estimate of the uncertainty matrix associated with $\mathbf{a}$ is

$$V_{\mathbf{a}} = s^2 \left(C^{\mathrm{T}}C\right)^{-1}. \tag{4}$$

We note that this estimate, in contrast to $V_{\boldsymbol{A}}$ in (2), does depend on the observed data $\mathbf{y}$.

2.2 Maximum likelihood estimation

We now make the further assumption that E_i are normally distributed with $\boldsymbol{E} \in N(0, \sigma^2 I)$. Then, the probability $p(y_i|\boldsymbol{\alpha}, \sigma)$ of observing y_i given parameters $\boldsymbol{\alpha}$ and σ is

$$p(y_i|\boldsymbol{\alpha}, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right\},$$

and, since the random variables E_i are independently distributed, the probability of observing $\mathbf{y}$ is the *likelihood* function

$$\begin{aligned} l(\boldsymbol{\alpha}, \sigma|\mathbf{y}) &= p(\mathbf{y}|\boldsymbol{\alpha}, \sigma) = \prod_{i=1}^{m} p(y_i|\boldsymbol{\alpha}, \sigma) \\ &= \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y - i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right\}. \end{aligned} \tag{5}$$

(The notation suggests that the likelihood function is viewed as a function of $\boldsymbol{\alpha}$ and σ with the observed data regarded as fixed. The notation $p(y|\boldsymbol{\alpha}, \sigma)$ indicates the probability density function $p(Y|\boldsymbol{\alpha}, \sigma)$, a function of Y depending on parameters $\boldsymbol{\alpha}$ and σ, evaluated at $Y = y$.) The log likelihood function is given by

$$L(\boldsymbol{\alpha}, \sigma|\mathbf{y}) = -m \log \sigma - \frac{m}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2.$$

The likelihood is maximized by $\mathbf{a}$ and s if $\mathbf{a}$ minimizes

$$F(\boldsymbol{\alpha}|\mathbf{y}) = \sum_{i=1}^{m}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2,$$

and s is such that

$$s^2 = \sqrt{\frac{\mathbf{r}^{\mathrm{T}}\mathbf{r}}{m}}, \tag{6}$$

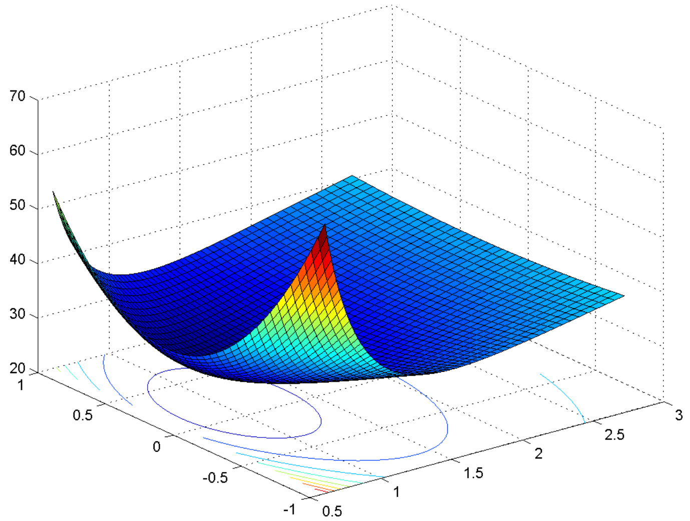

Figure 1: Log likelihood surface $-L(\alpha, \sigma|\mathbf{y})$ associated with the estimation of α and σ for data $y_i \in N(0, 1)$ with 20 data points, plotted as a function of α and $\log \sigma^2$.

where $\mathbf{r} = \mathbf{y} - C\mathbf{a}$ are the residuals at the solution. We note that ML estimate of $\mathbf{a}$ for normally distributed E_i is the same as the least squares estimate estimate while (6) differs (slightly) from that derived from the expectation of the χ^2 distribution in (3). Figures 1 and 2 graph the negative likelihood surfaces associated with determining a constant α and standard deviation σ from 20 and 100 data points sampled from a normal distribution. The surface for 20 points is flatter than that for 100, so that the minimum is less well defined for 20 points.

2.3 Bayesian formulation

Both least squares and maximum likelihood methods are based on a so-called classical approach to statistical inference. In this model, the parameters $\boldsymbol{\alpha}$ we are trying to determine are fixed but unknown. The measurements $\mathbf{y}$ are assumed to have been generated according to a statistical model whose behaviour depends on $\boldsymbol{\alpha}$. (We assume here that α represents all the relevant parameters, including σ.) On the basis of the measurements $\mathbf{y}$ estimates $\mathbf{a}$ are found for $\boldsymbol{\alpha}$. These estimates are regarded as a sample from a vector of random variables $\boldsymbol{A}$ and the uncertainty associated with $\mathbf{a}$ is determined from the distribution associated with this random vector.

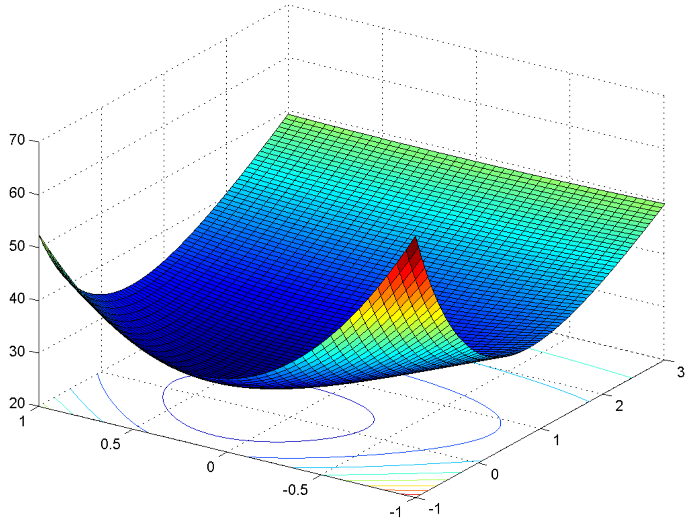

Figure 2: As Figure 1 but with 100 data points.

In a Bayesian formulation [3, 20, 25], our knowledge about $\boldsymbol{\alpha}$ is encoded in a probability distribution $p(\boldsymbol{\alpha}|I)$ derived from the information I we have to hand. As more information is gathered through measurement experiments, for example, these distributions are updated.

In the context of data analysis, we assume a *prior* distribution $p(\boldsymbol{\alpha})$ and that data $\mathbf{y}$ has been gathered according to a sampling distribution depending on $\boldsymbol{\alpha}$ from which we can calculate the probability $p(\mathbf{y}|\boldsymbol{\alpha})$ of observing $\mathbf{y}$. This probability is the same as the likelihood function $l(\boldsymbol{\alpha}|\mathbf{y})$ used in maximum likelihood estimation. Bayes Theorem states that the *posterior* distribution $p(\boldsymbol{\alpha}|\mathbf{y})$ for $\boldsymbol{\alpha}$ after observing $\mathbf{y}$ is related to the likelihood and the prior distribution by

$$p(\boldsymbol{\alpha}|\mathbf{y}) = kp(\mathbf{y}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}), \tag{7}$$

where the constant k is chosen to ensure that the posterior distribution integrates to 1, i.e,

$$\int p(\boldsymbol{\alpha}|\mathbf{y})\, d\boldsymbol{\alpha} = 1.$$

In this form, Bayes theorem says that the posterior distribution is the likelihood weighted by the prior distribution.

2.3.1 Generalized maximum likelihood estimate

The posterior distribution represents all the information about $\boldsymbol{\alpha}$ taking into account the measurement data $\mathbf{y}$ and the prior information. In practice, summary information about this distribution is required and in metrology it is usual to provide parameter estimates along with associated uncertainties. Ideally, this would be in the form of the mean (expectation) and variance of the posterior distribution. However, both these quantities require integration of multivariate functions and for problems involving even a modest number of parameters, 10 say, this integration is computationally expensive. For large problems it becomes impractical.

An alternative to providing estimates that require global knowledge of the distribution is to provide an approximation to the distribution on the basis of local knowledge. This is the approach taken in generalized maximum likelihood estimation (GMLE), also known as maximizing the posterior (MAP) [20]. The main idea is to determine a quadratic approximation to the logarithm $L(\boldsymbol{\alpha}) = \log p(\boldsymbol{\alpha}|\mathbf{y})$ of the posterior distribution about its mode $\mathbf{a}$:

$$L(\boldsymbol{\alpha}) \approx L(\mathbf{a}) + \frac{1}{2}(\boldsymbol{\alpha} - \mathbf{a})^{\mathrm{T}} H(\boldsymbol{\alpha} - \mathbf{a}), \qquad (8)$$

where

$$H_{jk} = \frac{\partial^2 L}{\partial \alpha_j \partial \alpha_k}$$

is the Hessian matrix of second partial derivatives of L evaluated at the maximum $\mathbf{a}$. (The linear term in this approximation is absent since $\partial L/\partial \alpha_j = 0$ at the maximum.) This approximation is sometimes referred to the Laplace approximation [20]. Taking exponentials of (8), we approximate the posterior distribution by

$$p(\boldsymbol{\alpha}|\mathbf{y}) \approx k \exp\left\{\frac{1}{2}(\boldsymbol{\alpha} - \mathbf{a})^{\mathrm{T}} H(\boldsymbol{\alpha} - \mathbf{a})\right\},$$

where k is a normalizing constant. Recognizing this as a multivariate normal distribution, setting $V = -H^{-1}$, we have

$$p(\boldsymbol{\alpha}|\mathbf{y}) \approx \frac{1}{|2\pi V|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\alpha} - \mathbf{a})^{\mathrm{T}} V^{-1}(\boldsymbol{\alpha} - \mathbf{a})\right\},$$

i.e., $\boldsymbol{\alpha} \sim N(\mathbf{a}, V)$. (The notation $|V|$ denotes the determinant of V.) This approach provides parameter estimates $\mathbf{a}$ and associated uncertainty matrix V using standard nonlinear optimization techniques. We note that we can determine these terms without knowing the constant of proportionality in (7). In practice, we minimize $-\log p(\boldsymbol{\alpha}|\mathbf{y})$ so that V is set to be the inverse of the Hessian matrix.

As with most approximating methods, this approach has to be used with some care. The multivariate normal distribution is unimodal and symmetric. If the true posterior distribution is multimodal or skewed, then the approximation could well provide poor information. (There may also be numerical difficulties in implementing the approach in these circumstances.)

2.3.2 GMLE method for linear responses

For the case of the linear response considered above in section 2.1, the likelihood is given by (5). The prior $p(\boldsymbol{\alpha}, \sigma)$ should reflect what is known before the experiment takes place. If nothing is known, then an *non-informative prior* should be assigned which is essentially constant so that the posterior distribution is proportional to the likelihood. In metrological examples it is likely that some prior information is available, based on nominal values or previous experience using the measuring instrument, for example. In these circumstances, we may propose a prior distribution for $\boldsymbol{\alpha}$ of the form $p(\boldsymbol{\alpha}) = N(\boldsymbol{\alpha}_0, \tau^2 I)$ and one for σ of the form

$$\log \sigma^2 \sim N(\log \sigma_0^2, (\log \rho)^2), \quad \rho \geq 1,$$

where $\boldsymbol{\alpha}_0$, τ, σ_0 and ρ are specified. Roughly, this says that we are 95% certain that $\sigma_0^2/\rho^2 \leq \sigma^2 \leq \sigma_0^2\rho^2$. (We might prefer to use, for example, a beta or gamma distribution instead of a log normal distribution to represent our knowledge about in σ, but the general approach would be essentially the same.) Assuming $\boldsymbol{\alpha}$ and σ are independently distributed, the logarithm of the prior distribution is given by

$$\begin{aligned} -\log p(\boldsymbol{\alpha}, \sigma) &= \frac{1}{2}\log\left(2\pi\tau^2\right) + \frac{1}{2\tau^2}\sum_{j=1}^{n}(\alpha_j - \alpha_{0,j})^2 + \\ & \frac{1}{2}\log\left(2\pi(\log\rho)^2\right) + \frac{1}{2(\log\rho)^2}(\log\sigma^2 - \log\sigma_0^2)^2. \end{aligned}$$

The generalized ML estimate is found by minimizing

$$\begin{aligned} -\log p(\boldsymbol{\alpha}, \sigma|\mathbf{y}) &= \frac{m}{2}\log\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2 + \\ & \frac{1}{2\tau^2}\sum_{j=1}^{n}(\alpha_j - \alpha_{0,j})^2 + \frac{1}{2(\log\rho)^2}(\log\sigma^2 - \log\sigma_0^2)^2, \end{aligned}$$

with respect to $\boldsymbol{\alpha}$ and σ.

2.4 GMLE applied to data fusion problems

In the following sections, we illustrate how generalized maximum likelihood estimation can be used for parameter estimation in data fusion on a number of examples. A common feature in all the examples is the requirement to weight different sources of information appropriately so that best use is made of the data and the prior information.

3 Multiple random effects

In many measuring instruments, the variance of the random effects associated with the measurements has a dependence on the response value. As an example, suppose the model is

$$\boldsymbol{\eta} = C\boldsymbol{\alpha}, \quad Y_i = \eta_i + E_i, \quad E_i \sim N(0, (\sigma_1 + \sigma_2\eta_i)^2), \tag{9}$$

with the random variables E_i independently distributed and that $\mathbf{y}$ is a set of observations of $\boldsymbol{Y}$. The likelihood of observing y_i given $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ is

$$p(y_i|\boldsymbol{\alpha}, \boldsymbol{\sigma}) = \left(\frac{\phi_i}{2\pi}\right)^{1/2} \exp\left[-\frac{\phi_i}{2}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right],$$

where $\phi_i = \phi_i(\boldsymbol{\alpha}, \boldsymbol{\sigma}) = 1/(\sigma_1 + \sigma_2\eta_i)^2$. The log likelihood $L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) = \log p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\sigma})$ is given by

$$\begin{aligned} -L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) &= -\sum_{i=1}^{m} \log p(y_i|\boldsymbol{\alpha}, \boldsymbol{\sigma}), \\ &= \frac{1}{2}\left\{-\sum_{i=1}^{m} \log \phi_i \; + \; \sum_{i=1}^{m} \phi_i(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right\}. \end{aligned}$$

For a prior distribution, we set

$$-\log p(\boldsymbol{\alpha}, \boldsymbol{\sigma}) = u^2(\log\sigma_1 - \log\sigma_{1,0})^2 + v^2(\sigma_2 - \sigma_{2,0})^2 \quad + \quad \text{Const.},$$

where u and v are weights that reflect our confidence in the prior estimates $\sigma_{k,0}$, $k = 1, 2$. This distribution reflects some prior information avout $\boldsymbol{\sigma}$ but none about $\boldsymbol{\alpha}$ since with $\boldsymbol{\sigma}$ fixed, $p(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ is constant. The use of a log normal prior distribution is intended to reflect our belief that the estimate $\boldsymbol{\sigma}_0$ is equally likely to be an under- or overestimate by a multiplicative factor. As defined, $p(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ is an improper distribution as its integral over $\boldsymbol{\alpha}$ is infinite. We could instead choose a prior which was zero outside some region $\Omega \subset \mathbb{R}^n$ of sufficiently large but finite volume. However, since our approximation to the posterior distribution is based only local information, both priors would lead to the same parameter estimates and uncertainties (so long as the region Ω contained the solution estimate of $\boldsymbol{\alpha}$).

Estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ are found by minimizing

$$F(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) = -L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) + u^2(\log\sigma_1 - \log\sigma_{1,0})^2 + v^2(\sigma_2 - \sigma_{2,0})^2,$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$. If H is the Hessian matrix at the solution $(\mathbf{a}, \mathbf{s})$ and $V = H^{-1}$ its inverse, then the standard uncertainties associated with the estimates of the fitted parameters are the square roots of the diagonal elements of V.

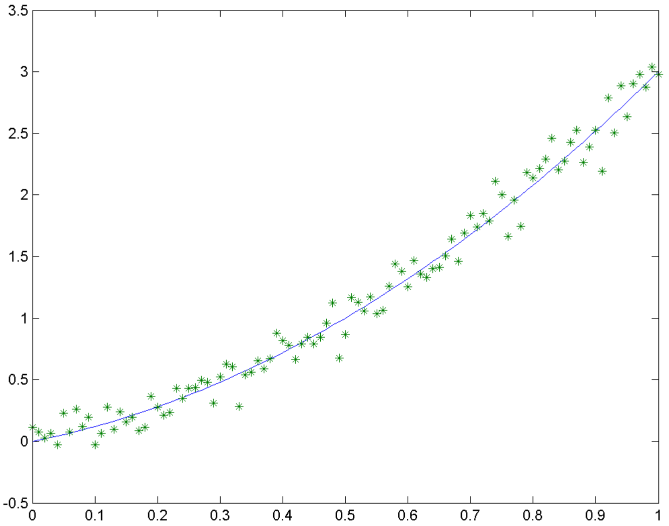

Figure 3: Data generated for a quadratic response and model (9) with $\boldsymbol{\alpha} = (0.0, 1.0, 2.0)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (0.10, 0.02)^{\mathrm{T}}$.

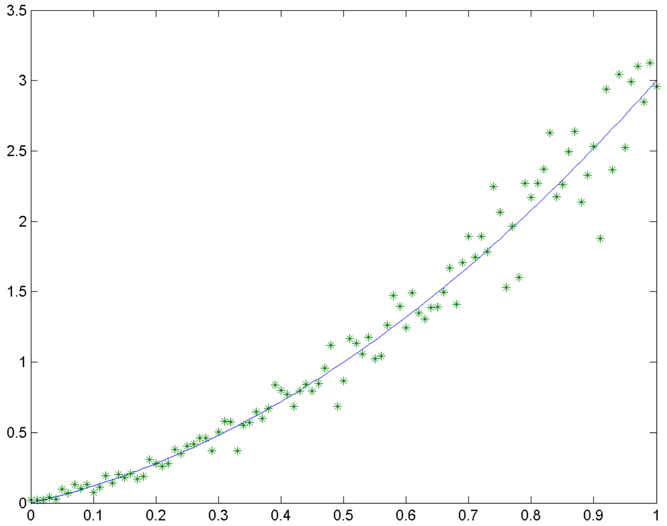

Figure 4: Data generated for a quadratic response and model (9) with $\boldsymbol{\alpha} = (0.0, 1.0, 2.0)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (0.02, 0.10)^{\mathrm{T}}$.

To illustrate the GMLE approach we have generated data according to the model (9) for a quadratic response $\eta = \alpha_1 + \alpha_2\xi + \alpha_3\xi^2$ to data generated with $\boldsymbol{\alpha} = (0.0, 1.0, 2.0)^{\mathrm{T}}$ and firstly with $\boldsymbol{\sigma} = (0.10, 0.02)^{\mathrm{T}}$: see Figure 3. We have set prior estimates $\sigma_{k,0} = 0.05$, $k = 1, 2$, and weights 1) $u = v = 0.0001$ and 2) $u = v = 10000.0$, corresponding to weakly and strongly weighted prior information, respectively. Table 1 gives the resulting estimates $\mathbf{a}$ and $\mathbf{s}$ along with their associated uncertainties $\mathbf{u}$. Table 2 gives corresponding results for data generated with $\boldsymbol{\sigma} = (0.02, 0.10)^{\mathrm{T}}$, Figure 4, with all other parameters as above. The tables show that for the weakly weighted prior information, the posterior estimates of $\boldsymbol{\sigma}$ are reasonable while for the strongly weighted, the posterior estimates are close to the prior values, as to be expected.

To obtain an indication of the validity of the uncertainty estimates $\mathbf{u}$, we have repeated these numerical simulations N times, recording the estimates $\mathbf{a}_q$, $\mathbf{s}_q$ and $\mathbf{u}_q$, $q = 1, \ldots, N$, and then calculating the sample means $\bar{\mathbf{a}}$, $\bar{\mathbf{s}}$ and $\bar{\mathbf{u}}$ and sample standard deviations $s(\mathbf{a})$ and $s(\mathbf{s})$. At the same time we compare the behaviour of the GMLE algorithm with a weighted least squares algorithm (WLS) which we now describe.

Given a model of the form

$$\boldsymbol{\eta} = C\boldsymbol{\alpha}, \quad Y_i = \eta_i + E_i, \quad E_i \sim N(0, \sigma_i^2), \tag{10}$$

with σ_i known, the appropriate least squares estimate $\mathbf{a}$ of $\boldsymbol{\alpha}$ is found by solving the weighted linear least squares problem

$$\min_{\boldsymbol{\alpha}} \sum_{i=1}^{m} w_i^2 (y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2, \tag{11}$$

with $w_i = 1/\sigma_i$. The difficulty with applying (10) to the problem formulated by (9) is that the standard deviations $\sigma_1 + \sigma_2\eta_i$ depend on the unknowns $\boldsymbol{\alpha}$ through $\boldsymbol{\eta}$. However, we can use the observed y_i as an estimate of η_i and solves (11) with $\mathbf{c}_i = (1, x_i, x_i^2)$ and $w_i = 1/(\sigma_{1,0} + \sigma_{2,0}y_i)$ to provide a solution estimate $\mathbf{a}_{WLS}$.

For the N Monte Carlo trials we record estimates $\mathbf{a}_{WLS,q}$, sample mean $\bar{\mathbf{a}}_{WLS}$ and sample standard deviation $s(\mathbf{a}_{WLS})$. Tables 3 and 4 give the results for $N = 5000$ Monte Carlo trials for data generated with $\boldsymbol{\alpha} = (0.0, 1.0, 2.0)^{\mathrm{T}}$, $\sigma_{k,0} = 0.05$, $k = 1, 2$, $u = v = 0.0001$, and $\boldsymbol{\sigma} = (0.10, 0.02)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (0.02, 0.10)^{\mathrm{T}}$, respectively. The tables show i) the GMLE algorithm produces good estimates of both $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$, ii) the estimated uncertainties $\bar{\mathbf{u}}$ are in line with the sample standard deviations $s(\mathbf{a})$ and $s(\mathbf{s})$ and iii) on average, the GMLE algorithm produces better estimates of the parameters $\boldsymbol{\alpha}$ than the WLS algorithm.

For both types of dataset illustrated by Figures 3 and 4, the data has provided sufficient information from which to provide point estimates of

		$u, v = 0.0001$		$u, v = 10000.0$	
		a, **s**	**u**	**a**, **s**	**u**
α_1	0.00	0.0338	0.028	0.0471	0.018
α_2	1.00	0.8400	0.145	0.7673	0.115
α_3	2.00	2.1643	0.153	2.2375	0.138
σ_1	0.10	0.0874	0.012	0.0502	0.001
σ_2	0.02	0.0247	0.010	0.0547	0.006

Table 1: Estimates $\mathbf{a}$ and $\mathbf{s}$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ determined by the GMLE algorithm and associated uncertainties $\mathbf{u}$ for data generated with $\boldsymbol{\alpha} = (0.0, 1.0, 2.0)^{\mathrm{T}}$, $\boldsymbol{\sigma} = (0.10, 0.02)^{\mathrm{T}}$, prior estimates $\sigma_{k,0} = 0.05$, $k = 1, 2$, and weights 1) $u = v = 0.0001$ and 2) $u = v = 10000.0$.

		$u, v = 0.0001$		$u, v = 10000.0$	
		a, **s**	**u**	**a**, **s**	**u**
α_1	0.00	0.0142	0.008	0.0082	0.018
α_2	1.00	0.8899	0.078	0.9182	0.120
α_3	2.00	2.1537	0.119	2.1311	0.147
σ_1	0.02	0.0145	0.004	0.0499	0.001
σ_2	0.10	0.0997	0.010	0.0638	0.006

Table 2: As Table 1 but for data generated with $\sigma_1 = 0.02$, $\sigma_2 = 0.10$.

	α_1	α_2	α_3	σ_1	σ_2
$\boldsymbol{\alpha}$, $\boldsymbol{\sigma}$	0.0000	1.0000	2.0000	0.1000	0.0200
$\bar{\mathbf{a}}$, $\bar{\mathbf{s}}$	0.0001	0.9990	2.0011	0.0977	0.0198
$s(\mathbf{a})$, $s(\mathbf{s})$	0.0311	0.1572	0.1618	0.0131	0.0107
$\bar{\mathbf{u}}$	0.0303	0.1541	0.1589	0.0128	0.0104
$\bar{\mathbf{a}}_{WLS}$	-0.0187	1.0085	1.9990		
$s(\mathbf{a}_{WLS})$	0.0341	0.1794	0.1871		

Table 3: Results of 5000 Monte Carlo trials comparing GMLE and WLS algorithms on datasets generated with $\sigma_1 = 0.10$ and $\sigma_2 = 0.02$.

	α_1	α_2	α_3	σ_1	σ_2
$\boldsymbol{\alpha}$, $\boldsymbol{\sigma}$	0.0000	1.0000	2.0000	0.0200	0.1000
$\bar{\mathbf{a}}$, $\bar{\mathbf{s}}$	0.0000	1.0005	1.9990	0.0185	0.0998
$s(\mathbf{a})$, $s(\mathbf{s})$	0.0092	0.0882	0.1307	0.0051	0.0113
$\bar{\mathbf{u}}$	0.0086	0.0849	0.1259	0.0047	0.0109
$\bar{\mathbf{a}}_{WLS}$	-0.0007	0.9942	1.9566		
$s(\mathbf{a}_{WLS})$	0.0120	0.1142	0.1609		

Table 4: Results of 5000 Monte Carlo trials comparing GMLE and WLS algorithms on datasets generated with $\sigma_1 = 0.02$ and $\sigma_2 = 0.10$.

the parameters $\boldsymbol{\sigma}$. If we consider instead data as in Figure 5, the fact that the responses η_i are approximately constant means that there is little information from which to determine both σ_1 and σ_2. Increasing σ_1 has the same effect as increasing σ_2, for example. For this dataset, the results corresponding to Table 1 are presented in Table 5. For tje case of the weakly weighted prior information, the estimate $\mathbf{s}$ of $\boldsymbol{\sigma}$ differs significantly from the values used to generate the data but are consistent with the data. The correlation matrix associated with the estimates $\mathbf{s}$ of $\boldsymbol{\sigma}$ is

$$\left[\begin{array}{rr} 1.0000 & -0.9874 \\ -0.9874 & 1.0000 \end{array}\right]$$

showing that σ_1 is negatively correlated with σ_2.

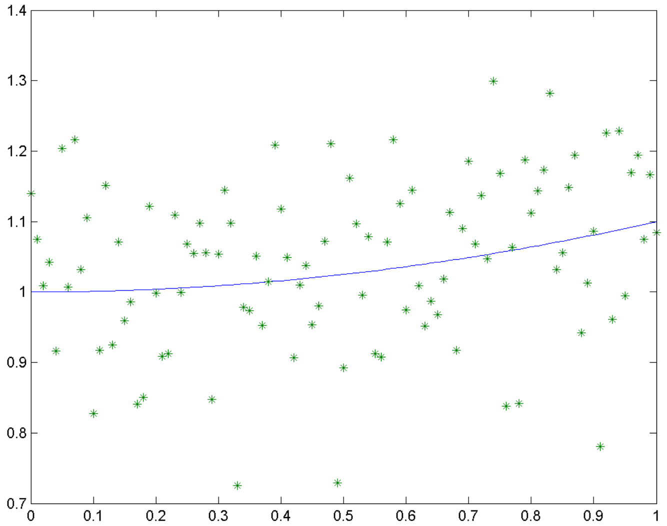

Figure 5: Data generated for a quadratic response and model (9) with $\boldsymbol{\alpha} = (1.0, 0.0, 0.1)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (0.10, 0.02)^{\mathrm{T}}$.

		$u, v = 0.0001$		$u, v = 10000.0$	
		a, **s**	**u**	**a**, **s**	**u**
α_1	1.00	1.0319	0.032	1.0330	0.031
α_2	0.00	-0.1340	0.149	-0.1362	0.146
α_3	0.10	0.2304	0.147	2.2319	0.143
σ_1	0.10	0.0011	0.050	0.0500	0.001
σ_2	0.02	0.1078	0.050	0.0571	0.005

Table 5: Estimates **a** and **s** of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ determined by the GMLE algorithm and associated uncertainties **u** for data generated with $\boldsymbol{\alpha} = (1.0, 0.0, 0.1)^{\mathrm{T}}$, $\boldsymbol{\sigma} = (0.10, 0.02)^{\mathrm{T}}$, prior estimates $\sigma_{k,0} = 0.05$, $k = 1, 2$, and weights 1) $u = v = 0.0001$ and 2) $u = v = 10000.0$.

4 Weighted least squares problems

In this section we consider the case where we wish to determine estimates of parameters $\boldsymbol{\alpha}$ from measurements arising from different sources. We assume a model of the form

$$\boldsymbol{\eta} = C\boldsymbol{\alpha}, \quad Y_i = \eta_i + E_i, \quad E_i \sim N(0, \sigma_i^2),$$

where C is an $m \times n$ matrix, the random variables E_i are independently distributed and $\mathbf{y}$ are a set of observations of $\boldsymbol{Y}$. Associated to each measurement y_i is an estimated standard uncertainty $\sigma_{i,0}$.

The likelihood of observing y_i given $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ is

$$p(y_i|\boldsymbol{\alpha}, \boldsymbol{\sigma}) = \left(\frac{1}{2\pi\sigma_i^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma_i^2}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2\right],$$

The log likelihood $L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y})$ is given by

$$\begin{aligned} -L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) &= -\sum_{i=1}^{m} \log p(y_i|\boldsymbol{\alpha}, \boldsymbol{\sigma}), \\ &= \frac{1}{2}\sum_i \log 2\pi\sigma_i^2 + \frac{1}{2}\sum_i \frac{1}{\sigma_i^2}(y_i - \mathbf{c}_i^{\mathrm{T}}\boldsymbol{\alpha})^2. \end{aligned}$$

Notice that the maximum likelihood estimate of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ is not well defined since there are $m+n$ parameters and only m observations. In order to arrive at such estimates it is necessary to add further information. This can be in the form of prior distributions for $\boldsymbol{\sigma}$. For example we can propose that

$$\log\sigma_i \sim N(\log\sigma_{i,0}, (\log\rho)^2), \quad \rho \geq 1.$$

The generalized ML estimate is found by minimizing

$$F(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) = -L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) + \frac{1}{2(\log\rho)^2)}\sum_{i=1}^{m}\left(\log\sigma_i - \log\sigma_{i,0}\right)^2,$$

with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$. We note that if the prior distribution does not involve $\boldsymbol{\alpha}$ the posterior estimate of $\boldsymbol{\alpha}$ is the same as the least squares estimate with weights $w_i = 1/s_i$, where s_i is the posterior estimate of $\boldsymbol{\sigma}$.

4.1 Example: measurement of the Newtonian gravitational constant

We wish to determine a (combined) estimate of the Newtonian constant of gravitation G from ten measurements as considered by Weise and Wöger

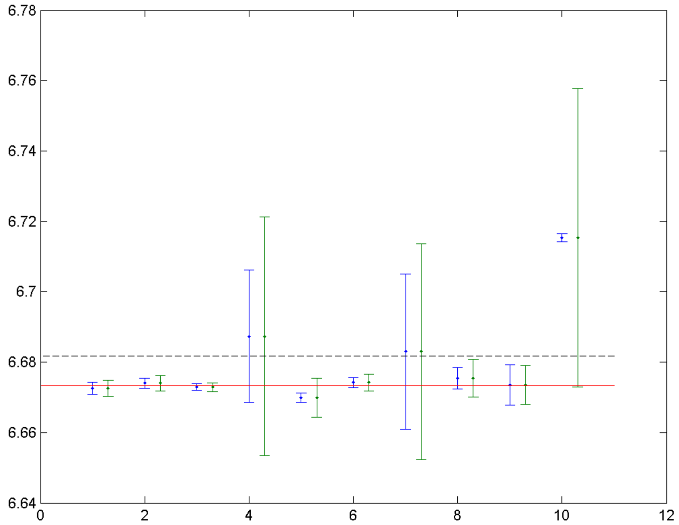

Figure 6: Data y_i along with intervals $y_i \pm \sigma_{i,0}$ and $y_i \pm s_i$, where $\sigma_{i,0}$ and s_i are the prior and posterior estimates for σ_i, associated with measurements of G. The dashed line represents the weighted least squares estimate determined from $\boldsymbol{\sigma}_0$ while the solid line is that determined from $\mathbf{s}$.

[26]. Figure 6 plots the estimates y_i and, on the left, the intervals $y_i \pm 2\sigma_{i,0}$. Observations 4 and 7 have relatively large prior estimates of σ_i. Observation 10 has a small estimate of σ_{10} but a value considerably different from the other measurements. The dashed line corresponds to the weighted least squares estimate of G based on the data $\mathbf{y}$ and weights $w_i = 1/\sigma_{i,0}$. It is seen that observation 10 has skewed the least squares estimate away from the majority of the data. In order to obtain a more consistent view of the data we regard $\sigma_{i,0}$ not as exact but as defining the mean for a log normal distribution for σ_i. The righthand set of intervals are $y_i \pm 2s_i$ where s_i is the posterior estimate of σ_i. It is seen that the posterior estimate associated with observation 10 is greatly increased. The solid line represents the posterior estimate of G and is seen to be consistent with $\mathbf{y}$ and $\mathbf{s}$. The issue of adjusting uncertainty matrices V associated with the measurement data in order to obtain data and model conformity is discussed in [6].

5 Random effects associated with more than one variable: generalized regression

We now consider the case where two sensors are required to measure the x and y variables. Both sensor measurements are subject to random effects independently distributed according to a Gaussian model so that the model equations are of the form

$$\left.\begin{array}{rcl} \eta_i & = & \phi(\xi_i, \boldsymbol{\alpha}), \\ x_i & = & \xi_i + \delta_i, \quad y_i = \eta_i + \epsilon_i, \\ \boldsymbol{\delta} & \in & N(0, \sigma_1^2 I), \quad \boldsymbol{\epsilon} \in N(0, \sigma_2^2 I). \end{array}\right\} \tag{12}$$

The probability $p(x_i, y_i|\xi_i, \boldsymbol{\alpha}, \boldsymbol{\sigma})$ of observing data point (x_i, y_i), given ξ_i, $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ is

$$p(x_i, y_i|\xi_i, \boldsymbol{\alpha}, \boldsymbol{\sigma}) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{1}{2\sigma_1^2}(x_i - \xi_i)^2 - \frac{1}{2\sigma_2^2}(y_i - \phi(\xi_i, \boldsymbol{\alpha}))^2\right].$$

The log likelihood function L associated with $X = \{(x_i, y_i)\}$ is given by

$$\begin{array}{rcl} -L(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\sigma}|X) & = & m \log 2\pi + m \log \sigma_1 + m \log \sigma_2 + \\ & & \displaystyle\frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i - \xi_i)^2 \quad + \quad \frac{1}{2\sigma_2^2}\sum_{i=1}^{m}(y_i - \phi(\xi_i, \boldsymbol{\alpha}))^2. \end{array}$$

If $\boldsymbol{\sigma}$ is known, then least squares estimates $\mathbf{a}$ of $\boldsymbol{\alpha}$ are found by solving the nonlinear optimization problem

$$\min_{\boldsymbol{\xi}, \boldsymbol{\alpha}} \left\{ \frac{1}{2\sigma_1^2}\sum_{i=1}^{m}(x_i - \xi_i)^2 \quad + \quad \frac{1}{2\sigma_2^2}\sum_{i=1}^{m}(y_i - \phi(\xi_i, \boldsymbol{\alpha}))^2 \right\}.$$

The MLE is found by minimizing $-L(\boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\sigma}|X)$ to determine estimates $\mathbf{x}^*$, $\mathbf{a}$ and $\mathbf{s}$. The solution $\mathbf{a}$ and $\mathbf{x}^*$ minimize

$$\frac{1}{s_1^2}\sum_{i=1}^{m}(x_i - x_i^*)^2 + \frac{1}{s_2^2}\sum_{i=1}^{m}(y_i - \phi_i^*)^2,$$

where $\phi_i^* = \phi(x_i^*, \mathbf{a})$, and depend on the ratio s_1/s_2. We also have

$$s_1^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - x_i^*)^2, \quad s_2^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \phi_i^*)^2.$$

In a GMLE approach, we incorporate prior information about the variables $\boldsymbol{\xi}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$.

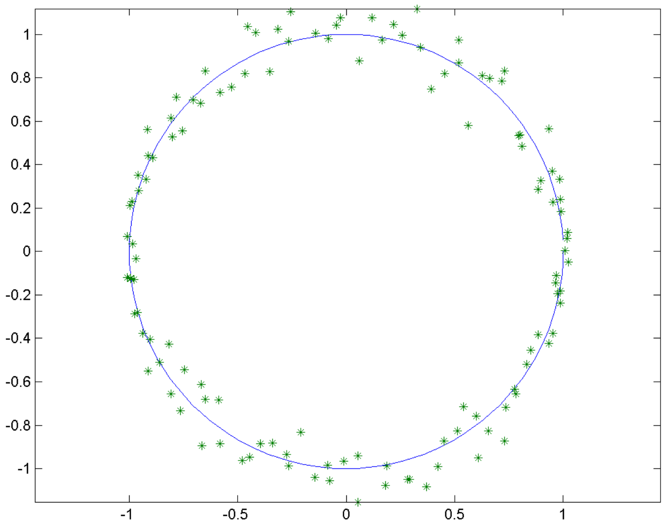

Figure 7: Data generated according to model (13) with $\boldsymbol{\alpha} = (0.0, 0.0, 1.0)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (0.02, 0.08)^{\mathrm{T}}$.

5.1 Example: circle fitting

To illustrate the GMLE approach, we consider the case of fitting a circle specified by centre coordinates (α_1, α_2) and radius α_3 to xy-data in which the $x-$ and $y-$ coordinates are subject to random effects drawn from Gaussian distributions with potentially different standard deviations. The model is

$$\left.\begin{array}{rcl} \xi_i & = & \alpha_1 + \alpha_3 \cos\theta_i, \\ \eta_i & = & \alpha_2 + \alpha_3 \sin\theta_i, \\ x_i & = & \xi_i + \delta_i, \quad y_i = \eta_i + \epsilon_i, \\ \boldsymbol{\delta} & \in & N(0, \sigma_1^2 I), \quad \boldsymbol{\epsilon} \in N(0, \sigma_2^2 I). \end{array}\right\} \tag{13}$$

Figure 7 illustrates 120 data points generated with $\boldsymbol{\alpha} = (0.0, 0.0, 1.0)^{\mathrm{T}}$ and $\boldsymbol{\sigma} = (0.02, 0.08)^{\mathrm{T}}$. The log likelihood $L(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}|X)$ associated with data points $X = \{(x_i, y_i)\}$ is therefore

$$\begin{array}{rcl} -L(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}|X) & = & m \log 2\pi + m \log \sigma_1 + m \log \sigma_2 + \\ & & \displaystyle\frac{1}{2\sigma_1^2} \sum_{i=1}^{m} (x_i - x_i^*)^2 \quad + \quad \frac{1}{2\sigma_2^2} \sum_{i=1}^{m} (y_i - y_i^*)^2, \end{array}$$

where $x_i^* = \alpha_1 + \alpha_3 \cos\theta_i$ and $y_i^* = \alpha_2 + \alpha_3 \sin\theta_i$. We choose a prior distribution $p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\sigma})$ of the form

$$-\log p(\boldsymbol{\theta}, \boldsymbol{\alpha}, \boldsymbol{\sigma}) = u^2(\log\sigma_1 - \log\sigma_{1,0})^2 + v^2(\log\sigma_2 - \log\sigma_{2,0})^2 \ + \ \mathrm{Const.},$$

reflecting prior knowledge about $\boldsymbol{\sigma}$ which is weighted by u and v. GMLE estimates $\mathbf{t}$, $\mathbf{a}$ and $\mathbf{s}$ of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ are determined by minimizing

$$\begin{aligned} F(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}|X) &= -L(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma}|X) + \\ & u^2(\log\sigma_1 - \log\sigma_{1,0})^2 + v^2(\log\sigma_2 - \log\sigma_{2,0})^2, \end{aligned}$$

with respect to $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$. If H is the Hessian matrix at the solution $(\mathbf{t},\mathbf{a},\mathbf{s})$, $V = H^{-1}$ its inverse, then the standard uncertainties associated with the estimates of the fitted parameters are the square roots of the diagonal elements of V.

Table 6 shows the results of applying the GMLE approach to data generated as for Figure 7 with $\boldsymbol{\sigma}_0 = (0.04, 0.04)^{\mathrm{T}}$ and different weights u and v. For $u = v = 100$, the posterior estimate of $\boldsymbol{\sigma}$ is essentially the same as the prior estimates $\boldsymbol{\sigma}_0$. As u and v become smaller the ratio s_2/s_1 becomes larger and for $u, v = 5.0$ this ratio is close to that $\sigma_2/\sigma_1 = 4.0$ used to generate the data. For $u, v = 2.5$, however, the posterior estimate of σ_1 becomes very small, 0.0003 compared to the value of 0.02 used to generate the data. If we look at the solution estimates $\mathbf{t}$ for this case we find that they are such that x_i^* is very close to x_i so that the estimate of σ_1 can become small without the likelihood becoming small. We will return to this point below.

Up to now we have found GMLE parameter estimates on the basis of prior information that is accorded a fixed weight. We can introduce further flexibility by assigning prior distributions rather than fixed values to these weights. We can therefore consider a prior distribution of the form

$$\begin{aligned} -\log p(\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\sigma},\boldsymbol{\omega}) &= \omega_1^2(\log\sigma_1 - \log\sigma_{1,0})^2 + \omega_2^2(\log\sigma_2 - \log\sigma_{2,0})^2 \;+ \\ & u^2(\log\omega_1 - \log\omega_{1,0})^2 + v^2(\log\omega_2 - \log\omega_{2,0})^2 \\ & +\text{Const.}, \end{aligned}$$

where the weights $\boldsymbol{\omega}$ themselves have a log normal prior distribution. The last two lines of Table 6 give the parameter estimates and their associated uncertainties obtained from using this augmented prior distribution with $\boldsymbol{\omega}_0 = (10.0, 10.0)^{\mathrm{T}}$ and $u = v = 5.0$. The posterior estimates $\mathbf{w}$ of $\boldsymbol{\omega}$ were $\mathbf{w} = (8.70, 8.47)^{\mathrm{T}}$.

One problem we have encountered is that if the prior information is given too small a weight, the posterior estimate of σ_1 can be unrealistically small. This suggests that the prior distribution does not encapsulate all our prior expectations about $\boldsymbol{\sigma}$. Suppose that in addition to our prior estimates $\boldsymbol{\sigma}_0$ we also have minimum and maximum values for $\boldsymbol{\sigma}$: $\boldsymbol{\sigma}_{min} \leq \boldsymbol{\sigma} \leq \boldsymbol{\sigma}_{max}$. We use the beta distribution to encode this information.

The probability density function for a beta distribution $B(\phi, \psi, a, b)$ defined

a_1	a_2	a_3	s_1	s_2
$u, v = 100.0$				
0.0043	-0.0035	1.0085	0.0400	0.0401
0.0052	0.0052	0.0037	0.0002	0.0002
$u, v = 20.0$				
0.0040	-0.0036	1.0079	0.0391	0.0423
0.0051	0.0054	0.0037	0.0010	0.0010
$u, v = 10.0$				
0.0032	-0.0037	1.0062	0.0345	0.0484
0.0049	0.0058	0.0038	0.0018	0.0020
$u, v = 5.0$				
0.0033	-0.0016	1.0047	0.0154	0.0672
0.0033	0.0069	0.0030	0.0016	0.0035
$u, v = 2.5$				
0.0073	0.0028	1.0166	0.0003	0.0933
0.0007	0.0085	0.0007	0.0001	0.0053
$\boldsymbol{\omega} = (10.0, 10.0)^{\mathrm{T}}$, $u, v = 5.0$				
0.0028	-0.0037	1.0056	0.0320	0.0513
0.0047	0.0060	0.0037	0.0030	0.0029

Table 6: Estimates $\mathbf{a}$ and $\mathbf{s}$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ determined by the GMLE algorithm and associated uncertainties $\mathbf{u}$ for data generated with $\boldsymbol{\alpha} = (0.0, 0.0, 1.0)^{\mathrm{T}}$, $\boldsymbol{\sigma} = (0.02, 0.08)^{\mathrm{T}}$, prior estimates $\sigma_{k,0} = 0.04$, $k = 1, 2$, and different weights u and v for the prior information. For each set of weights, the first row gives the parameter estimates and the second the associated standard uncertainties.

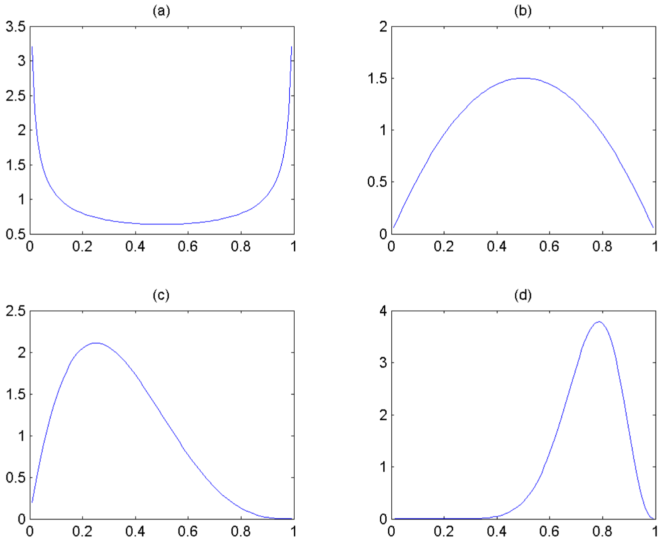

Figure 8: Graphs of beta probability density functions with $a = 0$, $b = 1$ and different values of the shape parameters (ϕ, ψ): a) (0.5,0.5), b) (2.0,2.0), c) (2.0,4.0) and d) (12.0,4.0).

on an interval $[a,\ b]$ is

$$p(x|\phi,\psi,a,b) = \frac{(x-a)^{\phi-1}(b-x)^{\psi-1}}{(b-a)^{\phi+\psi-1}}\frac{\Gamma(\phi+\psi)}{\Gamma(\psi)\Gamma(\psi)}, \quad a \le x \le b, \quad \alpha,\beta > 0,$$

If $X \sim B(\phi,\psi,a,b)$ then

$$E(X) = \mu = a + (b-a)\frac{\phi}{\phi+\psi},$$

and

$$\text{Var}(X) = \sigma^2 = (b-a)^2\frac{\phi\psi}{(\phi+\psi)^2(\phi+\psi+1)}.$$

Conversely, given a, b, mean μ and standard deviation σ, it is possible to calculate the corresponding ϕ and ψ values. Figure 8 graphs the beta probability density functions for a number of values of ϕ and ψ. As ϕ and ψ get large the beta distribution approximates a Gaussian. Figure 9 compares a beta pdf and Gaussian pdf both with mean $\mu = 0$ and standard deviation $\sigma = 0.1$. On the righthand side is the logarithm of the pdfs showing that the beta distribution approaches zero more quickly as x approaches $a = -0.5$ or $b = 0.5$.

Given $\boldsymbol{\sigma}_0$, $\boldsymbol{\sigma}_{min}$ and $\boldsymbol{\sigma}_{max}$ and standard deviations $\boldsymbol{\rho}$, we can assign a prior

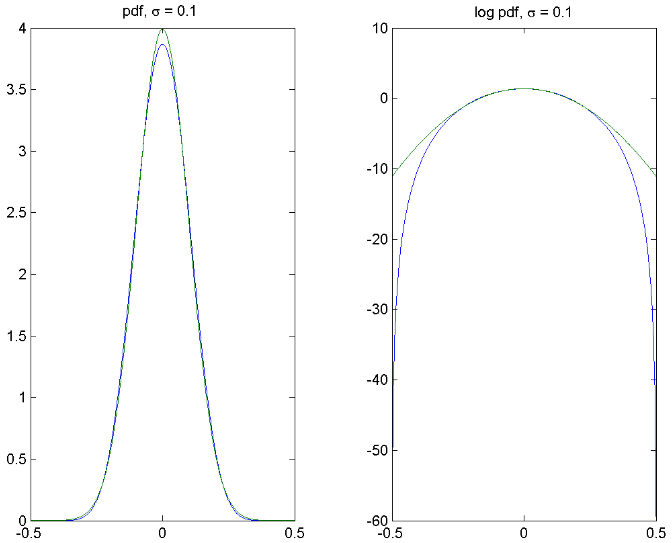

Figure 9: On the left, graphs of a beta probability density function with $a = -0.5$, $b = 0.5$, $\mu = 0.0$ and $\sigma = 0.1$ and a Gaussian pdf with the same mean and standard deviation. The Gaussian has the higher peak. On the right, are graphs of the logarithms of these pdfs. The beta function approaches zero more quickly as x approaches ± 0.5.

distribution to σ_k of the form

$$\log \sigma_k \sim B(\phi_k, \psi_k, \log \sigma_{k,min}, \log \sigma_{k,max}),$$

where ϕ_k and ψ_k are such that the expected value of $\log \sigma_k$ is $\log \sigma_{k,0}$ and its standard deviation is ρ_k.

Table 7 shows the parameter estimates $\mathbf{a}$ and $\mathbf{s}$ for the circle fitting problem corresponding to prior estimates $\boldsymbol{\sigma}_0 = (0.04, 0.04)^{\mathrm{T}}$, $\boldsymbol{\sigma}_{min} = (0.01, 0.01)^{\mathrm{T}}$, $\boldsymbol{\sigma}_{max} = (0.16, 0.16)$ and values of standard deviations ρ corresponding to the same choices of weights u and v as in Table 6. Comparing the two tables, we see that for small standard deviations/large weights the two types of prior information lead to very similar parameter estimates $\mathbf{a}$ and $\mathbf{s}$ but as the weight given to the prior information decreases, the estimate s_1 tends to the lower bound $\sigma_{1,min}$ for the beta prior distribution rather than to zero as for the case of the log normal prior.

As for the case of log normal prior distributions, instead of fixing the standard deviations ρ_k for the prior beta distributions, we can assign prior distributions for them. For the choice of $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}_{min}$ and $\boldsymbol{\sigma}_{max}$, the associated beta distributions are necessarily symmetric with $\phi_k = \psi_k$. The results in Table 8 are for the case where $\log \phi_k \sim N(\log \phi_{k,0}, \nu^2)$ where $\phi_{k,0}$ corresponds to a standard deviation of 0.1, i.e., $\phi_{k,0} \approx 383.9$. The table gives the estimates $\mathbf{a}$, $\mathbf{s}$ and $\mathbf{p}$ of $\boldsymbol{\alpha}$, $\boldsymbol{\sigma}$ and $\boldsymbol{\phi}$, respectively. As ν gets larger, the parameters ϕ_k are more free to move from the prior value of 383.9. The results show that the posterior estimates of ϕ_k ($= \psi_k$) become smaller as ν becomes larger, indicating that the prior value for $\sigma_{k,0} = 0.04$ is accorded less weight.

a_1	a_2	a_3	s_1	s_2
$\rho = 1/100.0$				
0.0043	-0.0035	1.0085	0.0400	0.0401
0.0052	0.0052	0.0037	0.0002	0.0002
$\rho = 1/20.0$				
0.0040	-0.0036	1.0079	0.0391	0.0423
0.0051	0.0054	0.0037	0.0010	0.0010
$\rho = 1/10.0$				
0.0032	-0.0037	1.0063	0.0346	0.0483
0.0049	0.0058	0.0037	0.0018	0.0019
$\rho = 1/5.0$				
0.0025	-0.0027	1.0045	0.0200	0.0638
0.0037	0.0067	0.0033	0.0014	0.0032
$\rho = 1/2.5$				
0.0044	0.0004	1.0059	0.0122	0.0787
0.0032	0.0077	0.0029	0.0005	0.0046

Table 7: Estimates $\mathbf{a}$ and $\mathbf{s}$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ determined by the GMLE algorithm and associated uncertainties $\mathbf{u}$ for data generated with $\boldsymbol{\alpha} = (0.0, 0.0, 1.0)^{\mathrm{T}}$, $\boldsymbol{\sigma} = (0.02, 0.08)^{\mathrm{T}}$, prior estimates $\sigma_{k,0} = 0.04$, $k = 1, 2$, and different standard deviations ρ for the prior information. For each standard deviation, the first row gives the parameter estimates and the second the associated standard uncertainties.

a_1	a_2	a_3	s_1	s_2	p_1	p_2
$\rho = 0.1$, $\nu = 0.1$						
0.0032	-0.0037	1.0062	0.0344	0.0485	376.3775	371.0084
0.0049	0.0058	0.0038	0.0019	0.0020	26.9774	26.3756
$\rho = 0.1$, $\nu = 0.2$						
0.0031	-0.0037	1.0060	0.0338	0.0494	349.7008	332.1989
0.0048	0.0059	0.0038	0.0021	0.0023	52.9476	47.9455
$\rho = 0.1$, $\nu = 0.3$						
0.0028	-0.0037	1.0056	0.0318	0.0514	278.1896	264.1674
0.0047	0.0060	0.0037	0.0030	0.0029	79.0253	58.8185
$\rho = 0.1$, $\nu = 0.4$						
0.0037	-0.0010	1.0049	0.0134	0.0655	32.1921	115.0501
0.0030	0.0066	0.0028	0.0011	0.0037	7.7530	26.8746

Table 8: Estimates $\mathbf{a}$ and $\mathbf{s}$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$ determined by the GMLE algorithm and associated uncertainties $\mathbf{u}$ for data generated with $\boldsymbol{\alpha} = (0.0, 0.0, 1.0)^{\mathrm{T}}$, $\boldsymbol{\sigma} = (0.02, 0.08)^{\mathrm{T}}$, prior estimates $\sigma_{k,0} = 0.04$, $k = 1, 2$, and different standard deviations ρ for the prior information. For each standard deviation, the first row gives the parameter estimates and the second the associated standard uncertainties.

6 Laser tracker case study

A laser tracker is a co-ordinate measuring instrument that determines the location of a target in three dimensions from measurements of azimuth and elevation angles and radial distance. The radial distance measurement is made using a laser interferometric transducer that records changes in optical displacement. In order to arrive at a geometric displacement, it is necessary to correct for the refractive index of the air. Both angle and interferometric measurements are subject to random effects due to changes in the refractive index of the air along the light path. A uniform increase in the refractive index increases the optical distance measurement but has no effect on the angle measurement. On the other hand, a uniform gradient in the refractive index orthogonal to the light path will cause the light path to bend affecting the angle measurements significantly with little effect on the displacement measurement. In order to determine the appropriate correction factor for the interferometric measurements it is necessary to estimate the average refractive index along the light path. For angle measurements, to make the appropriate correction it is necessary to know the refractive index at every point along the light path. In practice, the refractive index is estimated from pressure and temperature measurements made at a small number of locations and so the lack of knowledge about the refractive index is a major source of uncertainty. Furthermore, because the environmental conditions affect the angle and displacement measurements in different it ways it is very difficult to assign *a priori* estimates of the relative uncertainties for a given set of environmental conditions.

In order to improve the accuracy of such measurements NPL, and others, have assembled multi-station systems, combining measurements from up to four laser tracker systems. While trackers may have similar measuring characteristics, their performance will depend on their calibration history and on their position within the measuring environment, leading to different uncertainties associated with their measurements. Again, it is difficult to assign these uncertainties *a priori.*

We look for a GMLE approach that will help assign appropriate uncertainties on the basis of the measurements themselves along with any prior information.

6.1 Model for a two-dimensional system

In order to describe the basic approach more compactly, we consider a two dimensional system involve measurements of angle and distance. If a tracker is located at $\mathbf{p} = (p, q)^{\mathrm{T}}$ and the target is at $\boldsymbol{\xi} = (\xi, \eta)^{\mathrm{T}}$, the angle θ is given

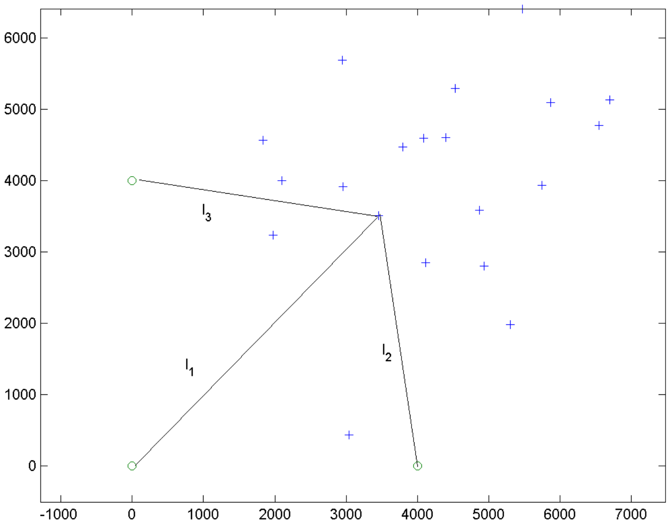

Figure 10: Example target '+' and station 'o' locations for a three station laser tracker system.

by

$$\theta = \tan^{-1}\left(\frac{\eta - q}{\xi - p}\right),$$

and the distance is

$$\lambda = \|\boldsymbol{\xi} - \mathbf{p}\|.$$

We assume that the tracker records measurements $\mathbf{u} = (t, l)^{\mathrm{T}}$ with

$$t = \theta + \delta, \quad l = \lambda + \epsilon, \quad \delta \in N(0, \sigma_1^2), \quad \epsilon \in N(0, (\sigma_2 + \sigma_3\lambda)^2).$$

Therefore, the likelihood of observing $\mathbf{u}$ given $\boldsymbol{\xi}$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)^{\mathrm{T}}$ is

$$\begin{aligned} p(\mathbf{u}|\boldsymbol{\xi}, \boldsymbol{\sigma}) &= \frac{1}{(2\pi\sigma_1^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma_1^2}(t-\theta)^2\right\} \\ &\times \frac{1}{(2\pi(\sigma_2+\sigma_3\lambda)^2)^{1/2}} \exp\left\{-\frac{1}{2(\sigma_2+\sigma_3\lambda)^2}(l-\lambda)^2\right\}. \end{aligned}$$

We assume we have n_T trackers at locations $\mathbf{p}_k$, $k = 1, \ldots, n_T$, that take measurements $U = \{\mathbf{u}_i, i = 1, \ldots, m\}$ associated with targets $\{\boldsymbol{\xi}_j, j = 1, \ldots, n_X\}$. We associate to the ith measurement indices $(j(i), k(i))$ that specify the corresponding target and station. Figure 10 shows example station and target locations.

We assign beta prior distributions for the logarithm of the parameters $\boldsymbol{\sigma}_k = (\sigma_{k,1}, \sigma_{k,2}, \sigma_{k,3})^{\mathrm{T}}$ associated with the kth station in terms of maxima $\boldsymbol{\sigma}_{k,max}$, minima $\boldsymbol{\sigma}_{k,min}$, prior estimates $\boldsymbol{\sigma}_{k,0}$ and standard deviations $\boldsymbol{\rho}_k$.

6.2 Numerical simulations

Below we report the results of experiments with $n_X = 10$ targets with data generated with $\boldsymbol{\sigma}_k = (1.0e-5, 1.0e-3, 1.0e-6)^{\mathrm{T}}$. The prior beta distributions were specified by $\boldsymbol{\sigma}_{k,0} = (5.0e-6, 2.0e-3, 2.0e-6)^{\mathrm{T}}$ which represents an over-weighting of the angle measurements relative to the length measurements by a factor of four, $\boldsymbol{\sigma}_{k,max} = 20.0 \times \boldsymbol{\sigma}_{k,0}$ and $\boldsymbol{\sigma}_{k,min} = \boldsymbol{\sigma}_{k,0}/10.0$.

In Table 9 we present the solution estimates $\mathbf{s}_k$ of $\boldsymbol{\sigma}_k$, $k = 1,2,3$, for various values of the standard deviation $\boldsymbol{\rho}$. Each group of nine rows holds the estimates $\mathbf{s}_k$, $k = 1,2,3$. For small values of $\boldsymbol{\rho}$, the solution estimates $\mathbf{s}_k$ are close to the prior estimates $\boldsymbol{\sigma}_{k,0}$ as is to be expected. As $\boldsymbol{\rho}$ becomes larger the posterior estimates become more in line with the values $\boldsymbol{\sigma}_k$ used to generate the data. For large values of $\boldsymbol{\rho}$, the values of $\mathbf{s}_k$ associated with the length measurements using trackers 2 and 3 become small relative to the values of the other parameters. As in the case of generalized regression, if the prior information is accorded too low a weight, the optimization scheme finds a local minimum that corresponds to regarding two of the sensors as very accurate and all others inaccurate.

We also examine in Table 10 the uncertainties associated with the estimated target locations $\mathbf{x}_i$ for different values of $\boldsymbol{\rho}$. The pairs of columns correspond to the uncertainties associated with the estimates of the $x-$ and $y-$ coordinates. The groups of the rows correspond to the ten targets. As $\boldsymbol{\rho}$ increases, the GMLE has more flexibility to make better use of the data and the uncertainties associated with the targets reduce. For large values, however, these estimates become unrealistically small since the posterior estimates $\mathbf{s}_k$ correspond to having two very accurate length measurements which are sufficient to determine the targets accurately.

According a low weight to the prior information leads, in the example considered here, to estimates in which some of the sensors are assigned an unrealistically small uncertainty. This is an unwelcome feature and we would like to improve on this. The first point to make is that the maximum and minimum values for $\boldsymbol{\sigma}_k$ in this example corresponded to a very broad band with $\boldsymbol{\sigma}_{k,max} = 200\boldsymbol{\sigma}_{k,min}$. In practice, we would reduce the range by an order of magnitude. This would ensure that any posterior estimate has some feasibility. The second point is that while we may expect one tracker to be perhaps twice as accurate as another, we do not expect there to be an order of magnitude difference as in the case $\rho = 0.20$ in Table 9. Our prior information, as it is currently formulated, does not reflect this expectation and the unrealistic posterior estimates are the consequence of this omission. In Table 11, we record the estimates $\mathbf{s}$ of $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)^{\mathrm{T}}$

$\boldsymbol{\sigma}_k$	$\boldsymbol{\sigma}_{k,0}$	$\boldsymbol{\rho}$			
		0.001	0.01	0.05	0.07
1.0e-5	5.0e-6	5.0034e-6	6.5506e-6	7.5366e-6	8.1740e-6
1.0e-3	2.0e-3	2.0000e-3	1.9465e-3	1.8461e-3	1.7556e-3
1.0e-6	2.0e-6	1.9997e-6	1.7336e-6	1.4062e-6	1.1959e-6
1.0e-5	5.0e-6	5.0014e-6	5.9259e-6	6.9450e-6	7.7306e-6
1.0e-3	2.0e-3	1.9999e-3	1.8834e-3	1.6592e-3	1.3459e-3
1.0e-6	2.0e-6	1.9996e-6	1.6694e-6	1.2546e-6	8.9005e-7
1.0e-5	5.0e-6	5.0031e-6	6.4846e-6	7.4682e-6	8.0889e-6
1.0e-3	2.0e-3	1.9999e-3	1.9292e-3	1.7514e-3	1.4601e-3
1.0e-6	2.0e-6	1.9997e-6	1.7251e-6	1.2867e-6	8.8979e-7
		0.09	0.11	0.15	0.20
1.0e-5	5.0e-6	8.5921e-6	8.8678e-6	9.1424e-6	9.3329e-6
1.0e-3	2.0e-3	1.7279e-3	1.7351e-3	1.7907e-3	2.0796e-3
1.0e-6	2.0e-6	1.1272e-6	1.1082e-6	1.0799e-6	1.0079e-6
1.0e-5	5.0e-6	8.2217e-6	8.5158e-6	8.8272e-6	9.0055e-6
1.0e-3	2.0e-3	1.0194e-3	7.5801e-4	4.6114e-4	3.0231e-4
1.0e-6	2.0e-6	6.2604e-7	4.5961e-7	3.0379e-7	2.3484e-7
1.0e-5	5.0e-6	8.4750e-6	8.7304e-6	8.9719e-6	9.1478e-6
1.0e-3	2.0e-3	1.1226e-3	8.4065e-4	5.1032e-4	3.2332e-4
1.0e-6	2.0e-6	6.1521e-7	4.5091e-7	2.9993e-7	2.3317e-7

Table 9: Estimates $\mathbf{s}_k$ of the parameters $\boldsymbol{\sigma}_k$ determined from simulated laser tracker measurement data using beta prior distributions with different standard deviations ρ.

$\boldsymbol{\rho}$							
0.001		0.01		0.05		0.07	
0.0072	0.0083	0.0073	0.0085	0.0062	0.0076	0.0049	0.0059
0.0081	0.0076	0.0081	0.0076	0.0065	0.0058	0.0051	0.0043
0.0107	0.0102	0.0108	0.0103	0.0088	0.0081	0.0066	0.0062
0.0078	0.0072	0.0078	0.0072	0.0065	0.0054	0.0052	0.0041
0.0075	0.0080	0.0076	0.0081	0.0059	0.0066	0.0045	0.0053
0.0100	0.0116	0.0101	0.0117	0.0082	0.0099	0.0063	0.0076
0.0079	0.0078	0.0079	0.0078	0.0062	0.0059	0.0047	0.0045
0.0082	0.0113	0.0082	0.0115	0.0068	0.0109	0.0053	0.0086
0.0068	0.0058	0.0068	0.0058	0.0060	0.0043	0.0048	0.0032
0.0114	0.0113	0.0117	0.0115	0.0100	0.0092	0.0073	0.0069
0.09		0.11		0.15		0.20	
0.0037	0.0044	0.0028	0.0033	0.0019	0.0021	0.0014	0.0016
0.0038	0.0031	0.0029	0.0023	0.0019	0.0015	0.0014	0.0011
0.0048	0.0046	0.0035	0.0034	0.0023	0.0022	0.0017	0.0017
0.0039	0.0030	0.0029	0.0022	0.0019	0.0014	0.0014	0.0010
0.0033	0.0040	0.0024	0.0030	0.0015	0.0019	0.0011	0.0014
0.0046	0.0055	0.0035	0.0041	0.0023	0.0027	0.0017	0.0020
0.0035	0.0033	0.0026	0.0025	0.0017	0.0016	0.0013	0.0012
0.0039	0.0064	0.0030	0.0048	0.0020	0.0031	0.0015	0.0024
0.0037	0.0023	0.0029	0.0017	0.0019	0.0011	0.0015	0.0008
0.0052	0.0050	0.0038	0.0037	0.0025	0.0024	0.0019	0.0018

Table 10: Estimates of the uncertainty in the target locations determined from simulated laser tracker measurement data using beta prior distributions with different standard deviations ρ.

$\boldsymbol{\sigma}$	$\boldsymbol{\sigma}_0$	$\boldsymbol{\rho}$		
		0.001	0.01	0.05
1.0e-5	5.0e-6	5.0014e-6	5.1364e-6	6.7309e-6
1.0e-3	2.0e-3	2.0000e-3	1.9956e-3	1.8635e-3
1.0e-6	2.0e-6	1.9998e-6	1.9821e-6	1.5709e-6
		0.09	0.15	0.20
1.0e-5	5.0e-6	7.9462e-006	8.7853e-6	9.0942e-6
1.0e-3	2.0e-3	1.5522e-003	1.1174e-3	8.7857e-4
1.0e-6	2.0e-6	1.0688e-006	7.2120e-7	6.0605e-7

Table 11: Estimates $\mathbf{s}$ of the parameters $\boldsymbol{\sigma}$ determined from simulated laser tracker measurement data using beta prior distributions with different standard deviations ρ for the case in which all three trackers are constrained to have the same uncertainty characteristics.

for the model in which we constrain all the trackers to have the same behaviour, i.e., $\boldsymbol{\sigma}_k = \boldsymbol{\sigma}$. In this situation, increasing ρ to relatively large values has no unexpected effect on the estimates $\mathbf{s}$. This shows that having prior distributions that better encode the prior information leads to better posterior estimates. Table 12 gives the uncertainties associated with the calculated target locations. As ρ becomes larger, the uncertainties become smaller but remain realistic. It should be noted that those associated with weaker prior information are approximately half those for the case where the prior information is taken to be exact. In the context of large scale co-ordinate metrology, this gain in performance is very significant.

In practice, one would want to constrain all three trackers to have similar rather than exactly the same performance and it is straightforward to specify prior distributions to encode this information.

The importance of prior information reflects the fact that there is little redundancy in the measurements. To every two target parameters there are only six observations and there is only a limited amount of information from which to determine estimates of the statistical parameters $\boldsymbol{\sigma}_k$. In practice, instead of taking one set of measurements per target, we are likely to take repeat measurements. In Table 13 we record the estimates $\mathbf{s}_k$ of $\boldsymbol{\sigma}_k$ using the same simulation scheme as for Table 9 but with four repeat measurements per target. The tables show that even for relatively large values of $\boldsymbol{\rho}$, the estimates $\mathbf{s}_k$ are consistent with those used to generate the data. Table 14 gives the corresponding estimates of the uncertainty associated with the target locations. We note that the uncertainties associated with the weak prior information are approximately 60% of those with the strongly weighted prior information. The fact that there are repeated measurements means that the data itself is providing strong information about the statistical parameters $\boldsymbol{\sigma}_k$.

$\boldsymbol{\rho}$					
0.001		0.01		0.05	
0.0072	0.0083	0.0073	0.0084	0.0069	0.0084
0.0081	0.0076	0.0081	0.0076	0.0074	0.0068
0.0107	0.0102	0.0107	0.0102	0.0099	0.0093
0.0078	0.0072	0.0078	0.0072	0.0071	0.0063
0.0075	0.0080	0.0075	0.0081	0.0067	0.0073
0.0100	0.0116	0.0101	0.0117	0.0093	0.0113
0.0079	0.0078	0.0079	0.0078	0.0070	0.0068
0.0082	0.0113	0.0082	0.0114	0.0076	0.0119
0.0068	0.0058	0.0068	0.0058	0.0066	0.0050
0.0114	0.0113	0.0115	0.0114	0.0107	0.0105
0.09		0.15		0.2	
0.0053	0.0068	0.0038	0.0049	0.0032	0.0041
0.0055	0.0051	0.0039	0.0035	0.0032	0.0029
0.0075	0.0070	0.0053	0.0049	0.0044	0.0041
0.0053	0.0047	0.0037	0.0032	0.0030	0.0027
0.0050	0.0055	0.0035	0.0038	0.0029	0.0032
0.0070	0.0088	0.0049	0.0062	0.0041	0.0052
0.0051	0.0050	0.0036	0.0035	0.0030	0.0029
0.0058	0.0098	0.0041	0.0072	0.0035	0.0060
0.0050	0.0038	0.0035	0.0026	0.0029	0.0021
0.0080	0.0079	0.0056	0.0055	0.0047	0.0046

Table 12: Estimates of the uncertainty in the target locations determined from simulated laser tracker measurement data using beta prior distributions with different standard deviations ρ for case in which all three trackers are constrained to have the same uncertainty characteristics.

$\boldsymbol{\sigma}_k$	$\boldsymbol{\sigma}_{k,0}$	$\boldsymbol{\rho}$			
		0.01	0.05	0.10	0.20
1.0e-5	5.0e-6	6.0138e-6	8.9814e-6	9.6153e-6	9.8261e-6
1.0e-3	2.0e-3	1.9758e-3	1.6309e-3	1.4450e-3	1.1192e-3
1.0e-6	2.0e-6	1.8675e-6	1.0285e-6	8.7105e-7	8.7067e-7
1.0e-5	5.0e-6	6.1129e-6	9.6504e-6	1.0450e-5	1.0710e-5
1.0e-3	2.0e-3	1.9661e-3	1.5023e-3	1.1769e-3	6.5035e-4
1.0e-6	2.0e-6	1.8932e-6	1.1320e-6	1.0125e-6	1.0853e-6
1.0e-5	5.0e-6	5.7342e-6	8.1178e-6	8.6089e-6	8.7683e-6
1.0e-3	2.0e-3	1.9751e-3	1.6359e-3	1.4843e-3	1.2417e-3
1.0e-6	2.0e-6	1.8904e-6	1.1117e-6	9.6795e-7	9.6945e-7

Table 13: Estimates $\mathbf{s}_k$ of the parameters $\boldsymbol{\sigma}_k$ determined from simulated laser tracker measurement data using beta prior distributions with different standard deviations ρ and four repeat measurements per target.

$\boldsymbol{\rho}$							
0.01		0.05		0.10		0.20	
0.0037	0.0044	0.0028	0.0036	0.0025	0.0032	0.0024	0.0030
0.0041	0.0038	0.0029	0.0027	0.0026	0.0023	0.0025	0.0022
0.0055	0.0052	0.0040	0.0037	0.0035	0.0033	0.0034	0.0032
0.0040	0.0036	0.0027	0.0024	0.0024	0.0021	0.0023	0.0020
0.0038	0.0040	0.0026	0.0029	0.0023	0.0025	0.0022	0.0025
0.0052	0.0061	0.0037	0.0047	0.0033	0.0041	0.0032	0.0040
0.0040	0.0039	0.0027	0.0026	0.0024	0.0023	0.0023	0.0022
0.0042	0.0062	0.0030	0.0052	0.0027	0.0047	0.0026	0.0045
0.0035	0.0028	0.0027	0.0020	0.0024	0.0017	0.0023	0.0015
0.0058	0.0058	0.0042	0.0041	0.0038	0.0037	0.0037	0.0036

Table 14: Estimates of the uncertainty in the target locations determined from simulated laser tracker measurement data using beta prior distributions with different standard deviations ρ and four repeat measurments per target.

7 Algorithmic and software requirements for GMLE

The main algorithmic requirement of the GMLE approach in the minimization of a function of several variables. This topic has received much attention and researchers have developed reliable and effective computational tools to solve a range of function minimization problems; see, for example, [10, 16, 21]. In this section we review the features particular to GMLE.

7.1 General form of the GMLE objective function

As we have seen, the posterior distribution $p(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y})$ given a set of data $\mathbf{y}$ satisfies

$$p(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\sigma})p(\boldsymbol{\alpha}, \boldsymbol{\sigma})$$

where $p(\mathbf{y}|\boldsymbol{\alpha}, \boldsymbol{\sigma}) = l(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y})$ is the likelihood function and $p(\boldsymbol{\alpha}, \boldsymbol{\sigma})$ is the prior distribution for $\boldsymbol{\alpha}$ and $\boldsymbol{\sigma}$. Taking logarithms, we minimize a function of the form

$$F(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) = -L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) - \log p(\boldsymbol{\alpha}, \boldsymbol{\sigma})$$

where $L = \log l(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y})$ is the log likelihood function that depends on the data.

7.2 Minimizing a function of several variables

In this section we overview the main approaches to minimizing a function of several variables, i.e., solving

$$\min_{\mathbf{a}} F(\mathbf{a})$$

with respect to $\mathbf{a} = (a_1, \ldots, a_n)^{\mathrm{T}}$. We will assume that F has continuous first and second partial derivatives.

7.2.1 Newton's algorithm

A first order necessary condition for $\mathbf{a}$ to be a minimum of F is that $\mathbf{g} = \nabla_{\mathbf{a}} F = \mathbf{0}$, that is $g_j = \partial F/\partial a_j = 0$, $j = 1, \ldots, n$. Given an approximate solution $\mathbf{a}$, linearizing $\mathbf{g}$ about $\mathbf{a}$ we have $\mathbf{g}(\mathbf{a}+\mathbf{p}) \approx \mathbf{g} + H\mathbf{p}$ where H is the Hessian matrix of second partial derivatives evaluated at $\mathbf{a}$, i.e.,

$$H_{jk} = \frac{\partial^2 F}{\partial a_j \partial a_k}.$$

The requirement that $\mathbf{g}(\mathbf{a}+\mathbf{p})$ should be zero leads to the equation

$$H\mathbf{p} = -\mathbf{g}. \tag{14}$$

Given the solution $\mathbf{p}$ an update of the form $\mathbf{a} := \mathbf{a} + t\mathbf{p}$ is made where t is chosen to ensure a sufficient decrease in F. Near a local minimum of a well behaved function, H will be strictly positive definite and t can be chosen to be unity. If the Hessian matrix in strictly positive definite then it has a Cholesky decomposition $H = TT^{\mathrm{T}}$ where T is lower triangular and the Newton step $\mathbf{p}$ can be found by solving, in sequence two triangular systems of equations,

$$T\mathbf{q} = -\mathbf{g}, \quad T^{\mathrm{T}}\mathbf{p} = \mathbf{q}.$$

It is not uncommon for H to have negative eigenvalues away from the solution and optimization algorithms have to take appropriate action in this case.

We note here that if $F(\mathbf{a})$ is a sum of squares

$$F(\mathbf{a}) = \frac{1}{2}\sum_{i=1}^{m} f_i^2(\mathbf{a}),$$

then

$$\mathbf{g} = J^{\mathrm{T}}\mathbf{f}, \qquad H = J^{\mathrm{T}}J + \sum_{i=1}^{m} f_i H_i,$$

where J is the Jacobian matrix with $J_{ij} = \partial f_i/\partial a_j$ and H_i is the matrix of second partial derivatives of f_i. In the Gauss-Newton algorithm, H is approximated by $J^{\mathrm{T}}J$ and the Gauss-Newton step is found by solving

$$J^{\mathrm{T}}J\mathbf{p} = -J^{\mathrm{T}}\mathbf{f}.$$

These are the normal equations associated with the linear least squares problem

$$\min_{\mathbf{p}} \|J\mathbf{p} + \mathbf{f}\|^2,$$

which can be solved stably using QR factorization techniques [8, 17]. Therefore, estimates of $\mathbf{a}$ can be found by solving a sequence of linear least squares systems. If the functions f_i are linear in $\mathbf{a}$, e.g., $\mathbf{f} = \mathbf{y} - C\mathbf{a}$, then the solution $\mathbf{a}$ is found by solving the associated linear least squares problem.

In the case of GMLE (or MLE), even if the model is linear and a Gaussian distribution is assumed, the log likelihood function is necessarily nonlinear and cannot in general be formulated as a sum of squares objective function. This means that more general optimization approaches are required.

Many optimization algorithms do not assume that the Hessian matrix is available explicitly because it is often burdensome to calculate. Some algorithms approximate it using finite differences. In a quasi-Newton algorithm an approximation to H or its inverse is built up from first order information as the iterations progress. The use of automatic differentiation techniques [1, 2] to provide derivatives has made pure Newton approaches more accessible. For problems with a large number of variables storing and solving equations involving the Hessian may be computationally impractical. In these cases large scale optimization techniques based on conjugate gradients can be applied [9].

7.3 Likelihood functions arising from Gaussian models for data

If the data is associated with a Gaussian model so that

$$\mathbf{y} \in N(\boldsymbol{\phi}(\boldsymbol{\alpha}), V(\boldsymbol{\sigma}))$$

where $V(\boldsymbol{\sigma})$ is an $m \times m$ uncertainty matrix of full rank, then the log likelihood function is given by

$$-L(\boldsymbol{\alpha}, \boldsymbol{\sigma}|\mathbf{y}) = \frac{1}{2}\left\{\log|2\pi V(\boldsymbol{\sigma})| + (\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha}))^{\mathrm{T}} V^{-1}(\boldsymbol{\sigma})\,(\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha}))\right\}.$$

Here, $|V|$ denotes the determinant of V. If the matrix V is diagonal, the calculations involving V will generally be straightforward. For the case where V is full, in order to use standard optimization algorithms we need to be able to calculate L and its derivatives with respect to σ_q. Any symmetric positive definite matrix V has a Cholesky factorization of the form $V = TT^{\mathrm{T}}$, where T is a lower triangular matrix. The matrix V can similarly be factored as $V = TT^{\mathrm{T}}$ where T is an upper-triangular matrix. These factorizations can be computed in a numerically stable way using a simple algorithm. For example, consider the factorization

$$TT^{\mathrm{T}} = \left[\begin{array}{cc} T_{11} & \mathbf{t}_{12} \\ & t_{22} \end{array}\right]\left[\begin{array}{cc} T_{11}^{\mathrm{T}} & \\ \mathbf{t}_{12}^{\mathrm{T}} & t_{22} \end{array}\right] = \left[\begin{array}{cc} V_{11} & \mathbf{v}_{12} \\ \mathbf{v}_{12}^{\mathrm{T}} & v_{22} \end{array}\right] = V.$$

Equating terms we have $t_{22}^2 = v_{22}$, $t_{22}\mathbf{t}_{12} = \mathbf{v}_{12}$ and $T_{11}T_{11}^{\mathrm{T}} = V_{11} - \mathbf{t}_{12}\mathbf{t}_{12}^{\mathrm{T}}$. The problem is now reduced to finding the factorization of the modified submatrix $V_{11} - \mathbf{t}_{12}\mathbf{t}_{12}^{\mathrm{T}}$. The complete factorization can be achieved by repeating this step:

I Set $T(i,j) = V(i,j)$, for all $i \geq j$.

II For $k = n : -1 : 1$, set $T(k,k) = T(k,k)^{1/2}$ and

$$T(1 : k-1, k) = T(1 : k-1, k)/T(k,k).$$

```
    II.j For $j = k-1:-1:1$, set
            $T(1:j,j) = T(1:j,j) - T(1:j,k)T(j,k)$.
```

Suppose now that $V = V(\boldsymbol{\sigma}) = T(\boldsymbol{\sigma})T^{\mathrm{T}}(\boldsymbol{\sigma})$, and $\dot{V} = \frac{\partial V}{\partial \sigma_q}$ and $\dot{T} = \frac{\partial T}{\partial \sigma_q}$. The matrix $\dot{T}$ satisfies $\dot{T}T^{\mathrm{T}} + T\dot{T}^{\mathrm{T}} = \dot{V}$ and can be determined by differentiating the algorithm above to compute T:

```
I Set $\dot{T}(i,j) = \dot{V}(i,j)$, for all $i \geq j$.
II For $k = n:-1:1$, set $\dot{T}(k,k) = \dot{T}(k,k)/(2T(k,k))$ and
        $\dot{T}(1:k-1,k) = (\dot{T}(1:k-1,k) - \dot{T}(k,k)T(1:k-1,k))/T(k,k)$.
    II.j For $j = k-1:-1:1$, set
            $\dot{T}(1:j,j) = \dot{T}(1:j,j) - T(1:j,k)\dot{T}(j,k) - \dot{T}(1:j,k)T(j,k)$.
```

This algorithm can be easily vectorized to compute all the partial derivatives of $T(\boldsymbol{\sigma})$ simultaneously.

With $V(\boldsymbol{\sigma}) = T(\boldsymbol{\sigma})T(\boldsymbol{\sigma})^T$, then

$$\frac{1}{2}\log|V(\sigma)| = \sum_{i=1}^{m}\log t_{ii}, \quad \frac{\partial}{\partial \sigma_q}\left(\frac{1}{2}\log|V(\sigma)|\right) = \sum_{i=1}^{m}\frac{\dot{t}_{ii}}{t_{ii}},$$

where t_{ii} ($\dot{t}_{ii}$) is the ith diagonal element of $T(\boldsymbol{\sigma})$ ($\dot{T}(\boldsymbol{\sigma})$).

Also,

$$(\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha}))^{\mathrm{T}} V^{-1}(\boldsymbol{\sigma})(\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha})) = \sum_{i=1}^{m} f_i^2(\boldsymbol{\alpha}, \boldsymbol{\sigma}),$$

with

$$T^{\mathrm{T}}(\boldsymbol{\sigma})\mathbf{f}(\boldsymbol{\alpha}, \boldsymbol{\sigma}) = (\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha})).$$

We note the partial derivatives of the inverse of $T(\boldsymbol{\sigma})$ can be found by solving triangular systems of equations of the form

$$T\left(\frac{\partial}{\partial \sigma_q}T^{-1}\right) = -\frac{\partial}{\partial \sigma_q}TT^{-1}.$$

As an alternative to working with the inverse of T, which could introduce numerical instabilities if T is ill-conditioned, we can introduce a new set of variables $\boldsymbol{\tau}$ which satisfy

$$\mathbf{y} = \boldsymbol{\phi}(\boldsymbol{\alpha}) + T(\boldsymbol{\sigma})\boldsymbol{\tau}, \tag{15}$$

so that

$$(\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha}))^{\mathrm{T}} V^{-1}(\boldsymbol{\sigma})\,(\mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\alpha})) = \boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\tau}.$$

With this approach, the generalized ML estimates are found by solving

$$\min_{\boldsymbol{\alpha},\boldsymbol{\sigma}} \sum_{i=1}^{m} \log t_{ii}(\boldsymbol{\sigma}) + \frac{1}{2}\boldsymbol{\tau}^{\mathrm{T}}\boldsymbol{\tau} - \log p(\boldsymbol{\alpha},\boldsymbol{\sigma})$$

subject to the constraints (15). This leads to better numerical properties and simpler derivative calculations at the expense of introducing nonlinear equality constraints. However, standard general purpose optimization software such as that implemented in NAG subroutine E04UCF [22] is able to deal effectively with this type of problem.

We note that for Gauss-Markov least squares regression problems [9, 14] the use of the generalized QR factorization [18, 23] allows us to cope with rank deficient uncertainty matrices V. For MLE, the inclusion of a term corresponding to the determinant of V requires V to be full rank.

7.4 Exploiting structure in the optimization problem

For many least squares problems, including generalized regression (section 5, [11]) and the analysis of data gathered by multi-station co-ordinate metrology systems [4], the observation matrix has a well-defined sparsity structure that can be exploited by the solution algorithms [7, 13]. We would like similar gains in efficiency to be possible for GMLE approaches. In general, this is possible. For example, both the generalized regression problem and the laser tracker problem produce Hessian matrices which have an 'arrow head' structure; see Figures 11 and 12. For such matrices, the formation of the Cholesky factor of the Hessian can be performed in $O(m)$ steps where m is the number of data points/targets. Similarly, determination of the Newton step can also be made in $O(m)$ steps. Without exploiting structure in some way, a Newton approach would take $O(m^3)$ steps making the GMLE approach impractical for large data sets.

7.5 Numerical evaluation of statistical functions

The objective functions for GMLE often involve the evaluation and differentiation of the logarithm of moderately complex statistical distributions such as the gamma or beta distributions. These have to be evaluated with care if problems with under- or overflow are to be avoided. It is best to use library software that have been specifically designed to operate in a numerical stable and robust way.

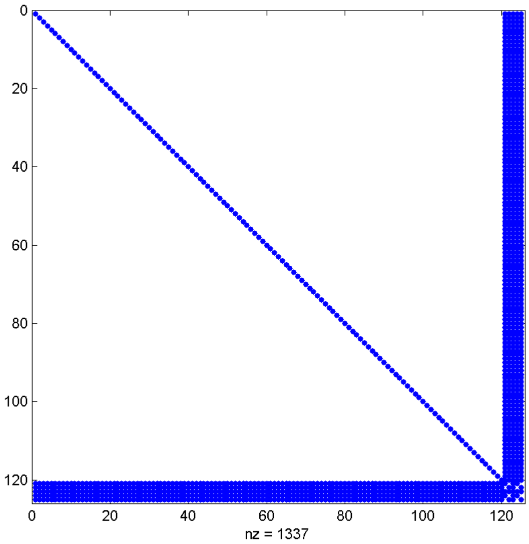

Figure 11: Location of non-zero elements in the Hessian matrix associated with generalized regression with a circle.

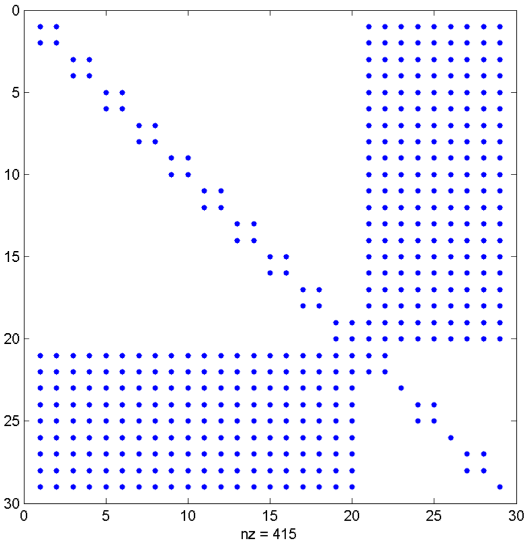

Figure 12: Location of non-zero elements in the Hessian matrix associated with the laser tracker case study.

(ϕ,ψ)	μ	ν	σ	s
(4.0,2.0)	0.6667	0.7500	0.1782	0.2165
(8.0,4.0)	0.6667	0.7000	0.1307	0.1449
(16.0,8.0)	0.6667	0.6818	0.0943	0.0961

Table 15: Values of i) mode ν, ii) standard deviation σ of the beta distribution, and iii) standard deviation s of the approximating Gaussian for different values of (ϕ, ψ).

8 Discussion on the GMLE approach

8.1 Approximation of the posterior distribution

As pointed out in section 2.3.1, the GMLE method approximates the posterior distribution by a Gaussian determined from information obtained at its mode (or at least at a local maximum). If the true posterior distribution is far from Gaussian, then this approximation may be poor. Figure 13 shows the Gaussian approximations (dashed lines) so determined to beta distributions (solid lines) on the interval [0,1] for values of (ϕ, ψ) = a) (4.0,2.0) b) (8.0,4.0) and c) (16.0,8.0). As ϕ and ψ increases, corresponding to increasing the number of observations, the approximation becomes better.

In Table 15, we give the corresponding values of i) mode ν, ii) standard deviation σ of the beta distribution, and iii) standard deviation s of the approximating Gaussian. All the beta distributions have mean $\mu = 0.6667$.

It is generally straightforward to perform a limited test on the quality of the Gaussian approximation. Using Bayes theorem it is possible evaluate the actual posterior distribution from the known likelihood function and prior distributions and this can be compared at a number of points with the Gaussian approximation. In practice the posterior distribution will only be known up to a multiplicative constant as in (7) but this does not prevent us making a meaningful comparison.

If we believe that that Gaussian approximation is not adequate, what other approaches are possible? The Gaussian approximation is derived from a quadratic approximation to the logarithm of the posterior distribution at its mode. We can instead take a higher order approximations, say, a cubic approximation. This will involve calculating or estimating all third order partial derivatives which may be feasible, particularly if automatic differentiation tools are available.

Other approaches are based on sampling from the posterior distribution in order to build up a more comprehensive information, in particular estimating

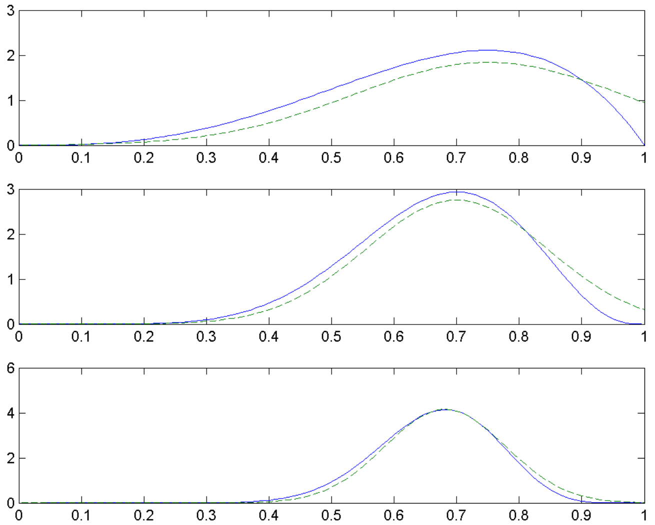

Figure 13: Gaussian approximations (dashed lines) to beta distributions (solid lines) on the interval [0,1] for values of (ϕ, ψ) = a) (4.0,2.0), top b) (8.0,4.0), middle, and c) (16.0,8.0), bottom.

the mean and standard deviations rather than properties associated with the mode. Standard Monte Carlo simulation approaches [5] can be applied and will be effective for problems involving a small number of parameters. For larger problems, the technique of choice would seem to be Markov chain Monte Carlo simulation (MCMS) which potentially delivers posterior estimates of mean and standard deviations (and even marginal distributions) for all the parameters; see e.g. [15]. The advantage of MCMS over standard Monte Carlo simulation is that the computational effort required, while significant, is largely independent of the number of parameters so that the analysis of large problems is feasible. It is likely that MCMS will be used more widely once the appropriate computational tools become more widespread and could be used to validate approximate methods.

8.2 The role of prior distributions

In situations where the measured data contains strong information about all the parameters, the likelihood function dominates and the role of the prior distribution is minimal. In this case, a maximum likelihood approach where no prior information is used (or, equivalently, the prior distribution employed is non-informative) will be effective. In situations in which the measurement data provides only weak information about some of the parameters, the prior distribution comes into play more strongly. For example, we have seen in the case of generalized regression (section 5) and the laser tracker case study (section 6) that the choice of the prior distribution can have a significant effect on the parameter estimates. For these examples, the ratio of number of observations to number of parameters is modest, approximately 2 in the case of the former, 3 in the latter. This means that the quantity of information to determine statistical behaviour of the system that generated the data is limited. It is therefore important in these cases that the prior distributions reflect all the information to hand so that unrealistic estimates are precluded on the basis of the prior information rather than relying on the likelihood function.

With Bayesian approaches, one has to weight the prior information relative to the measurement data. If this is done, as closely as possible, on a probabilistic basis, the output results should be satisfactory. However, it is not always possible to quantify vague prior information. In this case a hierarchical approach in which the weights for the prior information is not regarded as fixed but are themselves associated with probability distributions provides additional flexibility. The GMLE approach can then arrive at a balance between the prior and the likelihood. The additional flexibility has to be carefully designed so that any unrealistic behaviour is precluded.

8.3 Summary of the GMLE approach

8.3.1 Advantages

- The posterior distribution is summarized in terms of point estimates and covariance matrices from information that is straightforward to calculate using standard optimization techniques.

- The analysis is based on probabilistic models and statistical inference. It is natural generalization of well-known, least-squares methods.

- Prior information can be incorporated with measurement data in order to make maximal use of the information available.

- The method is flexible with respect to the type of likelihood functions and prior distributions. We have concentrated on likelihood functions based on Gaussian models but other models can be catered for easily.

- Optimization algorithms can be adapted to take into account structure in the matrix equations that need to be solved.

8.3.2 Disadvantages

- The optimization problems are nonlinear, even for linear models.
- The summary of the posterior distribution is based on a Gaussian approximation which may not be appropriate in all cases.
- For data providing weak information about the model parameters, the prior distributions have to chosen with care.

Acknowledgements. This work has been supported by the National Measurement System Directorate of the UK Department of Trade and Industry as part of its NMS Software Support for Metrology programme. The authors are grateful to Martyn Byng and Geoff Morgan (NAG) and colleagues Maurice Cox, Jeff Flowers and Peter Harris for useful discussions and comments. In particular, Peter Harris provided helpful review comments on a draft of this report.

References

[1] R. Boudjemaa, M. G. Cox, A. B. Forbes, and P. M. Harris. Automatic differentiation and its applications to metrology. Technical Report CMSC 26/03, National Physical Laboratory, June 2003.

[2] R. Boudjemaa, M. G. Cox, A. B. Forbes, and P. M. Harris. Automatic differentiation and its applications to metrology. In *Advanced Mathematical and Computational Tools in Metrology VI*, 2004. To appear.

[3] G. E. P. Box and G. C. Tiao. *Bayesian inference in statistical analysis.* Wiley, New York, Wiley Classics Library Edition 1992 edition, 1973.

[4] M. G. Cox, M. P. Dainton, A. B. Forbes, P. M. Fossati, P. M. Harris, and I. M. Smith. Modelling multi-lateration co-ordinate measuring systems. Technical Report CBTLM S12, National Physical Laboratory, February 2000.

[5] M. G. Cox, M. P. Dainton, and P. M. Harris. Software Support for Metrology Best Practice Guide No. 6: Uncertainty and Statistical Modelling. Technical report, National Physical Laboratory, Teddington, 2001.

[6] M. G. Cox, A. B. Forbes, J. Flowers, and P. M. Harris. Least squares adjustment in the presence of discrepant data. In *Advanced*

Mathematical and Computational Tools in Metrology VI, 2004. To appear.

[7] M. G. Cox, A. B. Forbes, P. M. Fossati, P. M. Harris, and I. M. Smith. Techniques for the efficient solution of large scale calibration problems. Technical Report CMSC 25/03, National Physical Laboratory, Teddington, May 2003.

[8] M. G. Cox, A. B. Forbes, and P. M. Harris. Software Support for Metrology Best Practice Guide 4: Modelling Discrete Data. Technical report, National Physical Laboratory, Teddington, 2000.

[9] M. G. Cox, A. B. Forbes, P. M. Harris, and I. M. Smith. Classification and solution of regression problems for calibration. Technical Report CMSC 24/03, National Physical Laboratory, May 2003.

[10] R. Fletcher. *Practical Optimization, Vol. 2.* John Wiley and Sons, Chichester, 1981.

[11] A. B. Forbes. Generalised regression problems in metrology. *Numerical Algorithms*, 5:523–533, 1993.

[12] A. B. Forbes. Fusing prior calibration information in metrology data analysis. In P. Ciarlini, A. B. Forbes, F. Pavese, and D. Richter, editors, *Advanced Mathematical and Computational Tools in Metrology IV*, pages 98–108, Singapore, 2000. World Scientific.

[13] A. B. Forbes. Efficient algorithms for structured self-calibration problems. In J. Levesley, I. Anderson, and J. C. Mason, editors, *Algorithms for Approximation IV*, pages 146–153. University of Huddersfield, 2002.

[14] A. B. Forbes, P. M. Harris, and I. M. Smith. Generalised Gauss-Markov Regression. In J. Levesley, I. Anderson, and J. C. Mason, editors, *Algorithms for Approximation IV*, pages 270–277. University of Huddersfield, 2002.

[15] D. Gamerman. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference.* Chapman&Hall/CRC, Boca Raton, Fl, 1997.

[16] P. E. Gill, W. Murray, and M. H. Wright. *Practical Optimization.* Academic Press, London, 1981.

[17] G. H. Golub and C. F. Van Loan. *Matrix Computations.* John Hopkins University Press, Baltimore, third edition, 1996.

[18] S. Hammarling. The numerical solution of the general Gauss-Markov linear model. Technical Report TR2/85, Numerical Algorithms Group, Oxford, 1985.

[19] G. P. Kelly. Report on: data fusion special interest group. In P. Ciarlini, A. B. Forbes, F. Pavese, and D. Richter, editors, *Advanced Mathematical Tools in Metrology, IV*, pages 297–301, Singapore, 2000. World Scientific.

[20] H. S. Migon and D. Gamerman. *Statistical inference: an integrated approach.* Arnold, London, 1999.

[21] J. J. Moré and S. J. Wright. *Optimization Software Guide.* SIAM, Philadelphia, 1993.

[22] The Numerical Algorithms Group Limited, Wilkinson House, Jordan Hill Road, Oxford, OX2 8DR. *The NAG Fortran Library, Mark 20, Introductory Guide*, 2002. *http://www.nag.co.uk/.*

[23] C. C. Paige. Fast numerically stable computations for generalized least squares problems. *SIAM J. Numer. Anal.*, 16:165–171, 1979.

[24] J. A. Rice. *Mathematical Statistics and Data Analysis.* Duxbury Press, Belmont, CA, second edition, 1995.

[25] D. S. Sivia. *Data analysis: a Bayesian tutorial.* Clarendon Press, Oxford, 1996.

[26] K. Weise and W. Woeger. Removing model and data non-conformity in measurement evaluation. *Measurement Science and Technology*, 11:1649–1658, 2000.