Generalised Gauss-Markov regression

Alistair B Forbes, Peter M Harris and Ian M Smith

National Physical Laboratory, Teddington, Middlesex, TW11 0LW, UK.
alistair.forbes@npl.co.uk, peter.harris@npl.co.uk, ian.smith@npl.co.uk

Abstract

Experimental data analysis is an key activity in metrology, the science of measurement. It involves developing a mathematical model of the physical system in terms of mathematical equations involving parameters that describe all the relevant aspects of the system. The model specifies how the system is expected to respond to input data and the nature of the uncertainties in the inputs. Given measurement data, estimates of the model parameters are determined by solving the mathematical equations constructed as part of the model, and this requires developing an algorithm (or estimator) to determine values for the parameters that best explain the data. In many cases, the parameter estimates are given by the solution of a least-squares problem. This paper discusses how various uncertainty structures associated with the measurement data can be taken into consideration and describes the algorithms used to solve the resulting regression problems. Two applications from NPL are described which require the solution of generalised distance regression problems: the use of measurements of primary standard natural gas mixtures to estimate the composition of a new natural gas mixture, and the analysis of calibration data to estimate the effective area of a pressure balance.

1 Introduction

Many metrology experiments involve determining the behaviour of a response variable y as a function of a set of independent variables $\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$. Model building involves establishing the functional relationship between these quantities, usually involving a set of model parameters $\mathbf{a}$, i.e.,

$$y^* = \phi(\mathbf{x}^*, \mathbf{a}),$$

where y^* and $\mathbf{x}^*$ represent exact values of the variables. The terms $\mathbf{a}$ parametrize the range of possible response behaviour and the actual behaviour is specified by determining values for these parameters from measurement data. In practice, measurements are subject to error, and the error structure must be taken into account firstly in order to determine effective methods for obtaining parameter estimates and secondly in determining the uncertainty in the fitted model parameters. For a set of measurement data $\{\mathbf{x}_i, y_i\}_{i=1}^m$, the data analysis problem involves the accurate estimation of the parameters $\mathbf{a}$, taking into account knowledge of the uncertainties in $\{\mathbf{x}_i\}$ and/or $\{y_i\}$, and typically leads to a least-squares problem [4].

This paper describes the various uncertainty structures that arise and corresponding regressions problems for determining estimates of the model parameters. If the covari-

ance information associated with the measurements is structured so that only the ith set of measurement errors are correlated with each other, a generalised distance regression approach is appropriate. However, some applications have quite general correlation structure and a full Gauss-Markov estimation approach is required to make efficient use of the statistical model [7]. This leads to a *generalised Gauss-Markov regression* problem to take into account the errors in the variables and the general correlation structure. While the covariance structure may dictate which solution algorithms are to be employed, the information required of the model function ϕ is limited to the evaluation of the function and its derivatives with respect to $\mathbf{a}$ and $\mathbf{x}$. This means that solution algorithms can be based on a compact set of model-dependent modules and a generic set of harnessing routines that link the models to general purpose least-squares optimisation software.

The layout of the paper is as follows. In Section 2 we consider the various error structures and corresponding regression problems. Section 3 introduces two measurement problems encountered at NPL: the use of measurements of primary standard natural gas mixtures to estimate the composition of a new natural gas mixture; and the analysis of calibration data to estimate the effective area of a pressure balance. Although the functional models for these measurement systems are simple, taking the form of low-order polynomials, the statistical models need to account for (a) uncertainties in both the dependent and independent variables, and (b) possible correlations between measurements. These requirements lead us to solve generalised regression problems. An overview of solution algorithms for the various problems is given in Section 4. Concluding remarks are made in Section 5.

2 Error structures and regression problems

Within metrology, various error structures arise all of which can be taken into account. We now consider the main types.

2.1 Error in one variable only

2.1.1 *Ordinary (weighted) least squares*

The simplest type of error structure occurs when only one of the system variables is subject to error and there is no correlation between errors. The model is summarised by

$$y_i^* = \phi(\mathbf{x}_i^*, \mathbf{a}), \quad y_i = y_i^* + \epsilon_i, \quad \mathbf{x}_i = \mathbf{x}_i^*,$$

where it is assumed that

$$E(\epsilon_i) = 0, \quad \text{var}(\epsilon_i) = \sigma_i^2, \quad \text{cov}(\epsilon_i, \epsilon_j) = 0, \quad i \neq j. \tag{2.1}$$

Good estimates of $\mathbf{a}$ can be found by solving

$$\min_{\mathbf{a}} \sum_{i=1}^{m} w_i^2 [y_i - \phi(\mathbf{x}_i, \mathbf{a})]^2,$$

where $w_i = 1/\sigma_i$, $i = 1, \ldots, m$.

2.1.2 *Gauss-Markov regression*

If instead of (2.1), the measurement errors are correlated so that

$$E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad \text{var}(\boldsymbol{\epsilon}) = V,$$

with V full rank, then an estimate of $\mathbf{a}$ can be found by solving

$$\min_{\mathbf{a}} [\mathbf{y} - \boldsymbol{\phi}(\mathbf{a})]^{\mathrm{T}} V^{-1} [\mathbf{y} - \boldsymbol{\phi}(\mathbf{a})], \tag{2.2}$$

where the ith element of $\boldsymbol{\phi}(\mathbf{a})$ is $\phi(\mathbf{x}_i, \mathbf{a})$.

2.2 Errors in more than one variable

In many metrological applications more than one of the measured variables is subject to error, and this must be taken into account in order to determine estimates of the model parameters which are statistically efficient and free from major bias.

2.2.1 *Orthogonal distance regression*

The simplest case arises when the covariance matrix associated with the ith set of measurements is a multiple of the identity matrix and there is no correlation between any of the errors, summarised by the model

$$y_i^* = \phi(\mathbf{x}_i^*, \mathbf{a}), \quad y_i = y_i^* + \epsilon_i, \quad \mathbf{x}_i = \mathbf{x}_i^* + \boldsymbol{\delta}_i,$$

with

$$E(\boldsymbol{\eta}_i) = \mathbf{0}, \quad \text{var}(\boldsymbol{\eta}_i) = \rho_i^2 I, \tag{2.3}$$

where $\boldsymbol{\eta}_i = (\epsilon_i, \boldsymbol{\delta}_i^{\mathrm{T}})^{\mathrm{T}}$. In this case, appropriate estimates of the parameters are determined by the solution of

$$\min_{\{\mathbf{x}_i^*\}, \mathbf{a}} \sum_{i=1}^{m} v_i^2 \{(\mathbf{x}_i - \mathbf{x}_i^*)^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_i^*) + (y_i - \phi(\mathbf{x}_i^*, \mathbf{a}))^2\},$$

where $v_i = 1/\rho_i$, $i = 1, \ldots, m$.

Note that this optimisation problem involves m sets of parameters $\mathbf{x}_i^*$ as well as the parameters $\mathbf{a}$ specifying the model $y = \phi(\mathbf{x}, \mathbf{a})$.

2.2.2 *Generalised distance regression*

If we assume that the errors $\boldsymbol{\eta}_i$ are correlated with $\text{var}(\boldsymbol{\eta}_i) = V_i$ with V_i full rank, but that $\text{cov}(\boldsymbol{\eta}_i, \boldsymbol{\eta}_j) = \mathbf{0}$, $i \neq j$, then the appropriate regression problem is

$$\min_{\{\mathbf{x}_i^*\}, \mathbf{a}} \sum_{i=1}^{m} \begin{bmatrix} y_i - \phi(\mathbf{x}_i^*, \mathbf{a}) \\ \mathbf{x}_i - \mathbf{x}_i^* \end{bmatrix}^{\mathrm{T}} V_i^{-1} \begin{bmatrix} y_i - \phi(\mathbf{x}_i^*, \mathbf{a}) \\ \mathbf{x}_i - \mathbf{x}_i^* \end{bmatrix}. \tag{2.4}$$

2.2.3 *Generalised Gauss-Markov regression*

The most complicated error structure arises when all variables are subject to measurement error and there is general correlation between the errors. If $\boldsymbol{\xi}$ ($\boldsymbol{\xi}^*$) is the vector of measurements $\{\mathbf{x}_i\}$ (variables $\{\mathbf{x}_i^*\}$), then the corresponding regression problem is

$$\min_{\boldsymbol{\xi}, \mathbf{a}} \begin{bmatrix} \mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\xi}, \mathbf{a}) \\ \boldsymbol{\xi} - \boldsymbol{\xi}^* \end{bmatrix}^{\mathrm{T}} V^{-1} \begin{bmatrix} \mathbf{y} - \boldsymbol{\phi}(\boldsymbol{\xi}, \mathbf{a}) \\ \boldsymbol{\xi} - \boldsymbol{\xi}^* \end{bmatrix}, \tag{2.5}$$

where the ith element of $\boldsymbol{\phi}(\boldsymbol{\xi}, \mathbf{a})$ is $\phi(\mathbf{x}_i^*, \mathbf{a})$.

3 Examples from metrology

3.1 Preparation of primary standard natural gas mixtures

Within the Centre for Optical and Analytical Measurement at NPL, one part of the work of the Environmental Standards Group is to prepare primary standard natural gas mixtures. These are cylinders containing natural gas prepared gravimetrically to contain known compositions of each of the 11 constituent components (methane, ethane, propane, l-butane, n-butane, l-pentane, n-pentane, neo-pentane, hexane, nitrogen and carbon dioxide). Mixtures are prepared to cover various concentration ranges, e.g., methane: $64\% - 98\%$. These primary standard mixtures are used as the basis for determining the composition of a new mixture and hence its calorific value.

Given a number of primary standard natural gas mixtures containing known concentrations of one of the constituent components (e.g., CO_2), the detector response for each mixture and the detector response for the new mixture, we wish to determine the concentration of CO_2 in the new mixture.

An approach to solving this problem is firstly to use the calibration data (relating to the primary gas mixtures) to calibrate the detector and, secondly, to use the calibration curve so constructed with the new measurement to predict the concentration in the new mixture.

Errors to be accounted for are:

- the calibration data is known inexactly. The process of preparing the primary standards means that they are known inexactly, and indeed the errors in the standards may be correlated (this is a consequence of the gravimetric process used to prepare the standard mixtures which involves comparing on a balance each standard mixture at each stage of preparation against calibrated masses selected from a common set of masses),
- the data returned by the detector (which is based on the analytical technique of chromatography) is subject to measurement error.

Consequently, we wish our data analysis to account for the inexactness of the measurement data and to quantify the resulting uncertainty associated with the final measurement result.

Figure 1 shows a sample set of measurement data, with the ellipses around the calibration points illustrating the errors in the concentrations and detector responses. (The error ellipses have been magnified greatly for illustrative purposes.) The figure also shows a calibration curve which is used to estimate the concentration of the component for which the detector response (and its uncertainty) is known.

3.2 Calibration of pressure balances

The principal role of the Pressure and Vacuum Section in the Centre for Mechanical and Acoustical Metrology at NPL is the development and maintenance of primary measurement standards for pressure and vacuum and their dissemination to industry. Pressure balances are pressure generators and consist essentially of finely-machined pistons moun-

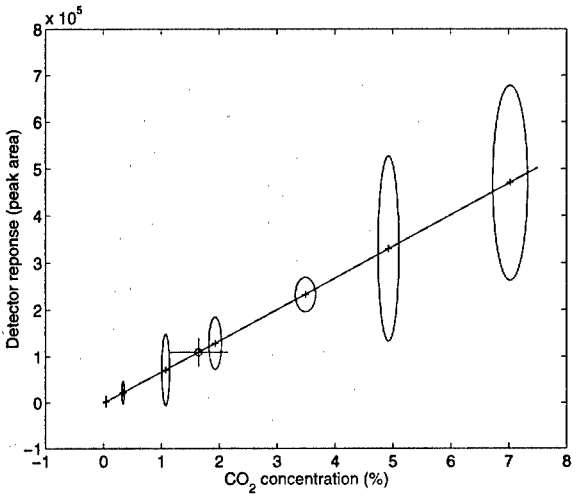

FIG. 1. Sample data (+), fitted calibration curve and predicted measurement (o).

ted vertically in close-fitting cylinders. The pressure required to support a piston and associated ring-weights depends on the mass of the piston and ring-weights and the cross-sectional area of the piston [5]. Due to various fluid dymamic effects, the effective area $A(p, \mathbf{a})$ of the piston-cylinder assembly is a function of pressure, usually taken to be a linear function $A(p, \mathbf{a}) = a_1 + a_2 p$. Many other factors such as temperature and air buoyancy have to be taken into account but for our purposes here, the pressure generated satisfies

$$a_1 p + a_2 p^2 = y(m),$$

where $\mathbf{a}$ are the instrument parameters and $y(m)$ is a simple function of the applied load m. This equation determines p implicitly as a function of m and $\mathbf{a}$. Suppose a reference pressure balance has been calibrated so that estimates of the instrument parameters $\mathbf{a}$ and their uncertainties are known. The reference balance can be used to calibrate a test balance in a cross-floating experiment in the following way. A load m_i is applied to the reference balance to generate pressure $p_i = p(m_i, \mathbf{a})$. A load n_i is applied to the test balance so that the pressures generated are matched. The test calibration curve is determined from a best fit to the data (n_i, p_i)

$$b_1 p_i^* + b_2 (p_i^*)^2 = y(n_i^*), \quad p_i = p_i^* + \epsilon_i, \quad n_i = n_i^* + \epsilon_i,$$

where δ_i and ϵ_i represent measurement error associated with the pressures and masses, respectively. However, the following must be taken into account. Firstly, the pressures p_i all depend on the common estimates $\mathbf{a}$ of the instrument parameters of the reference balance, leading to correlation of the measurement errors δ_i. Secondly, the masses n_i and m_i are made up from the same ensemble of masses $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_M)^{\mathrm{T}}$ so that

$$n_i = \mathbf{n}_i^{\mathrm{T}} \boldsymbol{\mu}, \quad m_i = \mathbf{m}_i^{\mathrm{T}} \boldsymbol{\mu},$$

where $\mathbf{n}_i$ and $\mathbf{m}_i$ are binary coefficient vectors. This means that measurement errors associated with the masses μ_k give rise to (further) correlation between δ_i and ϵ_i. Taking this general correlation into account, estimates of the the instrument parameters $\mathbf{b}$, are found from solving

$$\min_{\mathbf{b},\mathbf{p}^*} \begin{bmatrix} \mathbf{y}-\boldsymbol{\phi} \\ \mathbf{p}-\mathbf{p}^* \end{bmatrix}^{\mathrm{T}} V^{-1} \begin{bmatrix} \mathbf{y}-\boldsymbol{\phi} \\ \mathbf{p}-\mathbf{p}^* \end{bmatrix}, \tag{3.1}$$

where the ith elements of $\boldsymbol{\phi}$ and $\mathbf{y}$ are $b_1 p_i^* + b_2 (p_i^*)^2$ and $y(n_i)$, respectively, and V is the appropriate covariance matrix determined from the dependence of $\mathbf{y}$ and $\boldsymbol{\phi}$ on $\mathbf{a}$ and $\boldsymbol{\mu}$. This is a generalised Gauss-Markov regression problem.

4 Algorithms for generalised regression

Algorithms for ordinary least squares problems of the form $\min_{\mathbf{a}} \sum_i f_i^2(\mathbf{a})$ are well known and include QR factorisation methods for linear models or the Gauss-Newton algorithm for non-linear models; see, e.g., [2, 6]. The latter algorithm requires the user to supply a software module to evaluate the vector of function values $\mathbf{f}(\mathbf{a})$ and the Jacobian matrix J of partial derivatives

$$J_{ij} = \frac{\partial f_i}{\partial a_j}.$$

If $f_i(\mathbf{a}) = y_i - \phi(\mathbf{x}_i, \mathbf{a})$ as considered above, the user has to supply a module to calculate $\phi(\mathbf{x}, \mathbf{a})$ and $\partial\phi/\partial a_j$.

If V is symmetric and strictly positive definite, the Gauss-Markov regression problem (2.2) can be formulated as an ordinary least squares problem. If $V = LL^{\mathrm{T}}$ where L is lower-triangular, then the problem becomes

$$\min_{\mathbf{a}} \tilde{f}_i^2(\mathbf{a}),$$

where $\tilde{\mathbf{f}} = L^{-1}\mathbf{f}$. The associated Jacobian matrix is $\tilde{J} = L^{-1}J$. If the matrix V is well-conditioned, matrix operations with V or L^{-1} should not lead to unnecessary loss of precision. However, explicit calculations involving V can be avoided by using the generalised QR factorisation [2, 8, 9], leading to solution algorithms with good numerical properties.

The generalised distance regression problem (2.4) can be solved efficiently by making use of the fact that the parameters $\mathbf{x}_i^*$ appear only in the ith summand. The associated Jacobian matrix has a block-angular structure that can be exploited effectively in the QR factorisation stage [2, 3]. Alternatively, a separation-of-variables approach can be adopted in which the parameters $\mathbf{x}_i^*(\mathbf{a})$ are first determined as functions of $\mathbf{a}$ specified by the solution of the corresponding footpoint problem

$$\min_{\mathbf{x}_i^*} \begin{bmatrix} y_i - \phi(\mathbf{x}_i^*, \mathbf{a}) \\ \mathbf{x}_i - \mathbf{x}_i^* \end{bmatrix}^{\mathrm{T}} V_i^{-1} \begin{bmatrix} y_i - \phi(\mathbf{x}_i^*, \mathbf{a}) \\ \mathbf{x}_i - \mathbf{x}_i^* \end{bmatrix}$$

and the problem formulated as a non-linear least squares problem in $\mathbf{a}$ [1, 4]. Either approach yields an algorithm requiring $O(mn^2)$ flops while a full matrix approach requires $O(m^3)$ flops.

The generalised Gauss Markov problem (2.5) can be solved as a Gauss-Markov problem problem in the variables $\{\mathbf{x}_i^*\}$ and $\mathbf{a}$, but ideally, we would like to develop algorithms that exploit problem structure as in generalised distance regression algorithms. In particular, while the covariance matrix V may well be full, in many situations it is constructed from smaller matrices and for which more efficient algorithms could be developed.

From the user's point of view, all the regression algorithms discussed here require only the calculation of the model function ϕ and its derivatives $\frac{\partial \phi}{\partial x_k}$ and $\frac{\partial \phi}{\partial a_j}$. Thus, a wide range of regression problems can be solved using standard optimisation modules along with generic harness modules that perform the conversion without input from the user over and above the calculation of ϕ and its derivatives. For example, we have implemented a generalised Gauss-Markov solver to solve problems such as (3.1) for any explicit model $y = \phi(x, \mathbf{a})$. However, issues of efficiency and numerical stability need to be taken into account. As part of the UK Department of Trade and Industry's Software Support for Metrology programme, NPL is developing and making available to metrologists a suite of routines for the generalised regression problems discussed above. By combining structure exploiting linear algebra and numerically stable components such as the orthogonal factorisation, it is hoped that metrologists will be able to use these routines with the same confidence and effectiveness that they currently experience with standard, well-engineered regression modules available in numerical libraries.

5 Concluding remarks

In metrology, we are interested in the determination of accurate estimates of the parameters that describe a physical process. It is imperative that knowledge of the measurement system should be used to describe the error structure as accurately as possible. We have described the five types of regression problems that can occur in metrology depending on the error structures that are assumed. In all cases it is important that we employ efficient, numerically stable algorithms and exploit any structure in both the Jacobian and covariance matrices.

Acknowledgements. This work has been supported by the Department of Trade and Industry's National Measurement System Software Support for Metrology Programme and undertaken by a project team at the Centre for Mathematics and Scientific Software, National Physical Laboratory. The authors are particularly grateful to Paul Holland (Centre for Optical and Analytical Measurement) and the Pressure and Vacuum Section for their contributions.

Bibliography

1. M. Bartholomew-Biggs, B. P. Butler, and A. B. Forbes, Optimisation algorithms for generalised distance regression in metrology, in *Advanced Mathematical and Computational Tools in Metrology IV*, P. Ciarlini, A. B. Forbes, F. Pavese and D. Richter (eds), 21–31, World Scientific, Singapore, 2000.
2. A. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

3. M. G. Cox, The least-squares solution of linear equations with block-angular observation matrix, in *Advances in Reliable Numerical Computation*, M. G. Cox and S. Hammerling (eds), 227–240, Oxford University Press, 1989.
4. M. G. Cox, A. B. Forbes, and P. M. Harris, *Software Support for Metrology Best Practice Guide 4: Modelling Discrete Data*, National Physical Laboratory, Teddington, 2000.
5. A. B. Forbes, and P. M. Harris, Estimation algorithms in the calculation of the effective area of pressure balances, *Metrologia*, 36(6): 689–692, 1999.
6. G. H. Golub and C. F. Van Loan, Matrix Computations, John Hopkins University Press, Baltimore, third edition, 1996.
7. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.
8. C. C. Paige, Fast numerically stable computations for generalized least squares problems, *SIAM J. Numer. Anal.*, 16:165–171, 1979.
9. SIAM, Philadelphia, *The LAPACK Users' Guide*, third edition, 1999.