DEFRA Project: Improved Methods for National Spatial-Temporal Rainfall and Evaporation Modelling for BSM

Department of Civil and Environmental Engineering, IC

and

Department of Statistical Science, UCL

in collaboration with

CEH Wallingford

Internal report, no. 7

Moment-based inference for stochastic-mechanistic models

August 29, 2004

Richard Chandler

Contents

1 Introduction

Within the current DEFRA project, a substantial portion of the research is devoted to the development and application of models for rainfall based on point processes. These models are 'stochastic-mechanistic', in the sense that they attempt to provide a simplified stochastic representation of the mechanics of the rainfall process. They are parameterised in terms of physically interpretable quantities (e.g. storm arrival rate and mean cell intensity). However, estimation of the parameters is difficult, mainly because they are related rather indirectly to observable properties of rainfall sequences. Likelihood-based inference is generally infeasible, owing to the difficulty in formulating a likelihood function (this is a consequence of the complex dependencies induced by the model specification). Moreover, it has been argued by Rodriguez-Iturbe et al. (1988, Section 3) that likelihood-based inference is not necessarily appropriate in any case, because the models are necessarily over-simplified so that the joint distribution of an observed rainfall sequence differs substantially in some respects from that implied by the models. For example, the rectangular profile of rain cells in a single-site model leads to short-term deterministic features in model realisations; these are not present in real rainfall. This argument is to some extent supported by experience with the 'spectral likelihood' approach, which attempts to formulate an approximate Gaussian likelihood based on collections of sample Fourier coefficients. This approximate likelihood only involves the second-order moment properties of the data (mean, variance and autocorrelations); models fitted using this method are very good at reproducing these properties of observed rainfall sequences, but poor when it comes to other properties of interest such as lengths of dry intervals. Informally, the problem is that the likelihood method tries too hard to achieve a good match between model and data at very short timescales, whereas in practice this is not to be expected.[1]

In the absence of a suitable likelihood-based approach, stochastic-mechanistic models are usually fitted using a generalised method of moments: select a set of properties of interest (e.g. mean, variance, autocorrelations and proportion of 'dry' intervals at various levels of aggregation) and choose parameter estimates that minimise some measure of discrepancy between model and data with respect to these properties. This measure is usually a (possibly weighted) sum of squared differences. A particular advantage of this approach is that the model parameterisation can be chosen to reproduce, as closely as possible, those properties that are deemed to be particularly important in any specific application. However, a major disadvantage (compared with, say, a likelihood-based approach) is that assessments of uncertainty are not readily available. This note is an attempt to summarise the available options for obtaining uncertainty estimates (e.g. confidence intervals) when model parameters are estimated using a generalised method of moments. The problem can be regarded as an application of the theory of estimating equations; the relevant aspects are summarised in Section 3 below. Before this, however, we review some standard theory of likelihood-based inference, by way of illustrating the general concepts. In Section 4, we present the moment-based estimation procedure within the estimating equation framework; and Section

[1] Although this is largely irrelevant to the present note, it occurs to me — I wonder if you could obtain better performance by omitting the higher frequencies from the spectral likelihood?

5 provides a concise summary along with some practical suggestions for implementation.

2 Inference in a likelihood-based setting

The relevant results from likelihood-based inference are most easily illustrated in the context of a problem in which a vector of observations $\mathbf{y} = (y_1 \ \ldots \ y_n)'$ has been generated from a joint probability distribution whose density has the form $f(\mathbf{y};\theta)$. The functional form of f is known but the exact value of θ is not. The LIKELIHOOD FUNCTION for θ given the data $\mathbf{y}$ is defined as

$$L(\theta|\mathbf{y}) = f\left(\mathbf{y};\theta\right) \ ,$$

and can be interpreted as the probability of obtaining the observed data for any given value of θ. The MAXIMUM LIKELIHOOD ESTIMATE (MLE) of θ is the value, $\hat{\theta}$ say, for which the likelihood function is maximised (i.e. the value that allocates the highest probability to the observations). Equivalently, it is the value for which the log-likelihood

$$\ell(\theta|\mathbf{y}) = \ln L(\theta|\mathbf{y})$$

is maximised. In well-behaved problems, the MLE therefore satisfies the equation

$$U\left(\hat{\theta}|\mathbf{y}\right) = 0 \ , \tag{1}$$

where $U\left(\theta|\mathbf{y}\right) = \partial\ell\left(\theta|\mathbf{y}\right)/\partial\theta$ is the SCORE FUNCTION. We assume here that (1) has a unique solution. Note that

$$U\left(\theta|\mathbf{y}\right) = \frac{\partial \ln f\left(\mathbf{y};\theta\right)}{\partial\theta} = \frac{1}{f\left(\mathbf{y};\theta\right)}\frac{\partial f\left(\mathbf{y};\theta\right)}{\partial\theta} \ . \tag{2}$$

2.1 Properties of the score

Since $\mathbf{y}$ has been generated from a probability distribution, it can be regarded as the realised value of a vector $\mathbf{Y}$ of random variables. Hence the score function $U\left(\theta|\mathbf{y}\right)$ is the realised value of a random variable $U_\theta = U\left(\theta|\mathbf{Y}\right)$. The properties of this random variable depend on the true value of θ; call this θ_0. For example, we have

$$\mathrm{E}\left(U_\theta\right) = \int U\left(\theta|\mathbf{y}\right) f\left(\mathbf{y};\theta_0\right) d\mathbf{y} = \int \frac{1}{f\left(\mathbf{y};\theta\right)}\frac{\partial f\left(\mathbf{y};\theta\right)}{\partial\theta} f\left(\mathbf{y};\theta_0\right) d\mathbf{y} \ ,$$

the last step following from (2). This expression is valid for all θ. In particular, the expected score at the *true* parameter value is

$$\mathrm{E}\left(U_{\theta_0}\right) = \int \frac{1}{f\left(\mathbf{y};\theta_0\right)} \left.\frac{\partial f\left(\mathbf{y};\theta\right)}{\partial\theta}\right|_{\theta=\theta_0} f\left(\mathbf{y};\theta_0\right) d\mathbf{y} = \int \left.\frac{\partial f\left(\mathbf{y};\theta\right)}{\partial\theta}\right|_{\theta=\theta_0} d\mathbf{y}. \tag{3}$$

Now, since $f(\mathbf{y};\theta)$ is a probability density for all values of θ, we have

$$\int f(\mathbf{y};\theta)d\mathbf{y} = 1 \quad \text{so that} \quad \frac{\partial}{\partial\theta}\int f(\mathbf{y};\theta)d\mathbf{y} = 0 \ .$$

In well-behaved problems we can interchange the order of differentiation and integration, to yield

$$\int \frac{\partial f(\mathbf{y};\theta)}{\partial \theta} d\mathbf{y} = 0 \ .$$

This identity holds for all values of θ, and in particular for $\theta = \theta_0$. Hence, from (3), we have

$$\mathrm{E}(U_{\theta_0}) = 0 \ . \tag{4}$$

We now turn to the variance of the score function. This can be related to the expected value of its derivative — or equivalently, of the second derivative of the log-likelihood. For, differentiating (2) with respect to θ, we obtain

$$\begin{aligned} \frac{\partial U(\theta|\mathbf{y})}{\partial \theta} = \frac{\partial^2 \ell(\theta|\mathbf{y})}{\partial \theta^2} &= -\frac{1}{f^2(\mathbf{y};\theta)} \left(\frac{\partial f(\mathbf{y};\theta)}{\partial \theta} \right)^2 + \frac{1}{f(\mathbf{y};\theta)} \frac{\partial^2 f(\mathbf{y};\theta)}{\partial \theta^2} \\ &= -U^2(\theta|\mathbf{y}) + \frac{1}{f(\mathbf{y};\theta)} \frac{\partial^2 f(\mathbf{y};\theta)}{\partial \theta^2} \ . \end{aligned} \tag{5}$$

As before, all of these quantities are the realised values of random variables, so we can consider replacing $\mathbf{y}$ with $\mathbf{Y}$ and taking expectations. In particular, let $H_\theta = \partial^2 \ell(\theta|\mathbf{Y}) / \partial \theta^2$. Then we obtain

$$\mathrm{E}(H_\theta) = -\mathrm{E}\left[U_\theta^2\right] + \int \frac{1}{f(\mathbf{y};\theta)} \frac{\partial^2 f(\mathbf{y};\theta)}{\partial \theta^2} f(\mathbf{y};\theta_0)\, d\mathbf{y} \ .$$

Evaluated at $\theta = \theta_0$, the last term here is zero and we obtain

$$\mathrm{E}\left[U_{\theta_0}^2\right] = -\mathrm{E}(H_{\theta_0}) = I(\theta_0) \ , \text{ say.}$$

But since $\mathrm{E}(U_{\theta_0}) = 0$, we must have $\mathrm{E}\left[U_{\theta_0}^2\right] = \mathrm{var}(U_{\theta_0})$. Thus we have shown that

$$\mathrm{var}(U_{\theta_0}) = I(\theta_0) \ . \tag{6}$$

$I(\theta_0)$ is called the (FISHER) INFORMATION.

2.2 Large-sample properties of the MLE

The properties of the score function are fundamental to the development of asymptotic results for maximum likelihood estimators — in particular, to the construction of standard errors and confidence intervals for the parameters. The theory relies on the fact that, in well-behaved problems, as the sample size n tends to infinity the following two things happen:

1. The score function U_θ tends, when suitably normalised, to its expectation. For example, if the random variables in $\mathbf{Y}$ are independent and identically distributed (iid) then the log-likelihood is a sum of n independent contributions; it follows that the score function is also a sum of n independent contributions, and the law of large numbers dictates that $n^{-1}\left[U_\theta - \mathrm{E}(U_\theta)\right] \to 0$ as $n \to \infty$ in this case.

2. The distribution of $[U_\theta - \mathrm{E}(U_\theta)]/\sqrt{I(\theta)}$ tends to the standard normal distribution. Again, in the iid case this is easy to see: U_θ is a sum of independent terms, and the normality follows from the Central Limit Theorem.

Providing U_θ is continuous in θ, property 1 here implies that for large n, the score equation (1) has a solution in the neighbourhood of θ_0 (since, from (4), $\mathrm{E}(U_{\theta_0}) = 0$), and that this solution tends to θ_0 as $n \to \infty$. Hence, providing n is large enough, $\left|\hat{\theta} - \theta_0\right|$ will be small so that we can carry out a Taylor Series expansion for the score function in the neighbourhood of θ_0 and write

$$U_{\hat{\theta}} \approx U_{\theta_0} + \left(\hat{\theta} - \theta_0\right) H_{\theta_0} \tag{7}$$

(recall that H_θ is the second derivative of the log-likelihood at θ). But by definition, $U_{\hat{\theta}} = 0$, so that

$$\hat{\theta} - \theta_0 \approx -\frac{U_{\theta_0}}{H_{\theta_0}} = -\frac{U_{\theta_0}}{I(\theta_0)} \frac{I(\theta_0)}{H_{\theta_0}} .$$

As $n \to \infty$, H_θ tends to its expectation which is $-I(\theta)$, so that the second factor on the right-hand side here tends to -1. Hence we can approximate the estimation error $\hat{\theta} - \theta_0$ by $U_{\theta_0}/I(\theta_0)$. Strictly speaking, some care needs to be taken over the relative magnitudes of the various approximations here — for full details, see Cox and Hinkley (1974, Section 9.2).

Having expressed the estimation error in terms of the score, we can use property 2 above to deduce that for large samples, the estimation error has an approximate normal distribution. Specifically,

$$\sqrt{I(\theta_0)}\left(\hat{\theta} - \theta_0\right) \approx \frac{U_{\theta_0}}{\sqrt{I(\theta_0)}} = \frac{U_{\theta_0} - \mathrm{E}[U_{\theta_0}]}{\sqrt{I(\theta_0)}} \sim N(0,1) , \tag{8}$$

since $\mathrm{E}[U_{\theta_0}] = 0$. For practical purposes, an equivalent statement of this result is that for large n, the distribution of the MLE is approximately normal with mean θ_0 and variance $1/I(\theta_0)$. This can be used, for example, to construct approximate confidence intervals for θ_0: an approximate 95% interval is

$$\hat{\theta} \pm \frac{1.96}{\sqrt{I(\theta_0)}} . \tag{9}$$

Hypothesis tests based on (8) are referred to as WALD TESTS. As an alternative, inference could be based directly on the quantity $[U_\theta - \mathrm{E}(U_\theta)]/\sqrt{I(\theta)}$ at the right-hand side of (8), to yield a SCORE TEST. In general, the results from Wald and score tests will differ slightly due to the first approximation in (8).

A third possibility is to base inference on the log-likelihood function itself. A second-order Taylor expansion about the MLE yields, for some $\theta^\dagger$ between θ_0 and $\hat{\theta}$,

$$\begin{aligned} \ell(\theta_0|\mathbf{Y}) &= \ell\left(\hat{\theta}|\mathbf{Y}\right) + \left(\theta_0 - \hat{\theta}\right) \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\hat{\theta}} + \frac{1}{2}\left(\theta_0 - \hat{\theta}\right)^2 \left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\theta^\dagger} \\ &= \ell\left(\hat{\theta}|\mathbf{Y}\right) + \frac{1}{2}\left(\theta_0 - \hat{\theta}\right)^2 H_{\theta^\dagger} , \end{aligned} \tag{10}$$

as the first derivative of the log-likelihood is zero at $\hat{\theta}$ by definition. Since $\theta^\dagger$ is between θ_0 and $\hat{\theta}$ we have $H_{\theta^\dagger} \approx H_{\hat{\theta}}$. Moreover, again using the fact that $H_\theta \approx -I(\theta)$ for large n, we find

$$2\left[\ell\left(\hat{\theta}|\mathbf{Y}\right) - \ell\left(\theta_0|\mathbf{Y}\right)\right] \approx \left(\theta_0 - \hat{\theta}\right)^2 I\left(\hat{\theta}\right) = \left[\left(\theta_0 - \hat{\theta}\right)\sqrt{I\left(\hat{\theta}\right)}\right]^2 . \tag{11}$$

But from (8), the right hand side here is just the square of a standard normal random variable, and therefore has a chi-squared distribution with 1 degree of freedom. The asymptotic approximation

$$2\left[\ell\left(\hat{\theta}|\mathbf{Y}\right) - \ell\left(\theta_0|\mathbf{Y}\right)\right] \sim \chi^2_1 \tag{12}$$

can therefore be used to test hypotheses and construct confidence intervals. For example, a 95% confidence interval for θ consists of all values for which $2\left[\ell\left(\hat{\theta}|\mathbf{Y}\right) - \ell\left(\theta|\mathbf{Y}\right)\right]$ is less than the upper 95% point of a χ^2_1 distribution (which is 3.84).

Hypothesis tests based on (12) may be referred to as LIKELIHOOD RATIO TESTS. The three test procedures (Wald, score and likelihood ratio) are asymptotically equivalent, in the sense that their results will be very similar for large enough sample sizes. However, since the likelihood ratio test is derived from approximation (11) rather than from (7), its results will usually differ slightly, in finite samples, from both the Wald and score tests. There are grounds for preferring $2\left[\ell\left(\hat{\theta}|\mathbf{Y}\right) - \ell\left(\theta|\mathbf{Y}\right)\right]$ as a test statistic, although the accuracy of the χ^2 approximation is not guaranteed in finite samples.

2.3 More than one parameter

The theory above carries over straightforwardly to the case when there is more than one parameter. Specifically, denote the unknown parameter vector by $\boldsymbol{\theta} = (\theta_1 \ \ldots \ \theta_p)'$. Then the log-likelihood for $\boldsymbol{\theta}$ can be defined as previously, and the MLE satisfies the system of score equations

$$U_j\left(\hat{\boldsymbol{\theta}}|\mathbf{y}\right) = 0 \qquad (j = 1, \ldots, p)$$

where now $U_j\left(\boldsymbol{\theta}|\mathbf{y}\right) = \partial\ell\left(\boldsymbol{\theta}|\mathbf{y}\right)/\partial\theta_j$. These p equations can be written in vector form as

$$\mathbf{U}\left(\hat{\boldsymbol{\theta}}|\mathbf{y}\right) = \mathbf{0} . \tag{13}$$

$\mathbf{U}\left(\boldsymbol{\theta}|\mathbf{y}\right)$ is the SCORE VECTOR, and can be regarded as the realised value of a vector of random variables $\mathbf{U}_{\boldsymbol{\theta}}$. Denoting the true parameter by $\boldsymbol{\theta}_0$, we can show that

$$\mathrm{E}\left(\mathbf{U}_{\boldsymbol{\theta}_0}\right) = \mathbf{0} \quad \text{and} \quad \mathrm{var}\left(\mathbf{U}_{\boldsymbol{\theta}_0}\right) = \mathbf{I}\left(\boldsymbol{\theta}_0\right) = -\mathrm{E}\left(\mathbf{H}_{\boldsymbol{\theta}_0}\right) , \tag{14}$$

where $\mathbf{H}_{\boldsymbol{\theta}}$ is the Hessian matrix of second derivatives of the log-likelihood at $\boldsymbol{\theta}$. For large samples, $\mathbf{U}_{\boldsymbol{\theta}}$ again approaches its expectation and has an approximate normal distribution (this time in p dimensions):

$$\mathbf{U}_{\boldsymbol{\theta}_0} \sim MVN\left(\mathbf{0}, \mathbf{I}\left(\boldsymbol{\theta}_0\right)\right) . \tag{15}$$

The Taylor expansion corresponding to (7) is now

$$\mathbf{U}_{\hat{\boldsymbol{\theta}}} \approx \mathbf{U}_{\boldsymbol{\theta}_0} + \mathbf{H}_{\boldsymbol{\theta}_0}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \ , \tag{16}$$

so that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \approx -\mathbf{H}^{-1}{}_{\boldsymbol{\theta}_0}\mathbf{U}_{\boldsymbol{\theta}_0} \approx \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\,\mathbf{U}_{\boldsymbol{\theta}_0}$. Hence $\mathrm{E}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \approx 0$ and $\mathrm{var}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \approx \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\,\mathrm{var}\left(\mathbf{U}_{\boldsymbol{\theta}_0}\right)\mathbf{I}^{-1}(\boldsymbol{\theta}_0) = \mathbf{I}^{-1}(\boldsymbol{\theta}_0)$. For large samples we therefore have, approximately,

$$\hat{\boldsymbol{\theta}} \sim MVN\left(\boldsymbol{\theta}_0, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\right) \ . \tag{17}$$

In the multiparameter case, the equivalent of (10) is

$$\ell(\boldsymbol{\theta}_0|\mathbf{Y}) = \ell\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right) + \frac{1}{2}\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right)'\mathbf{H}_{\boldsymbol{\theta}^\dagger}\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right)$$

for some $\boldsymbol{\theta}^\dagger$ between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$. Now for large n, the elements of the matrix $\mathbf{H}_{\boldsymbol{\theta}^\dagger} - \mathrm{E}\left[\mathbf{H}_{\boldsymbol{\theta}^\dagger}\right]$ are order $n^{1/2}$ in probability. Also, since $\left(\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\dagger\right)$ is order $n^{-1/2}$, the elements of $\mathrm{E}\left[\mathbf{H}_{\boldsymbol{\theta}^\dagger}\right] - \mathrm{E}\left[\mathbf{H}_{\boldsymbol{\theta}_0}\right]$ are themselves $\left(o_p\left(n^{1/2}\right)\right.$. Therefore we can write $\mathbf{H}_{\boldsymbol{\theta}^\dagger} = -\mathbf{I}(\boldsymbol{\theta}_0) + \mathbf{E}$, where the elements of $\mathbf{E}$ are $O_p\left(n^{1/2}\right)$. Hence

$$\begin{aligned} 2\left[\ell\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right) - \ell(\boldsymbol{\theta}_0|\mathbf{Y})\right] &= \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right)'\mathbf{I}(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right) + O_p\left(n^{-1/2}\right) \\ &= \left[\mathbf{A}(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right)\right]'\mathbf{A}(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right) + O_p(n^{-1/2}) \ , \end{aligned} \tag{18}$$

where $\mathbf{A}(\boldsymbol{\theta}_0)$ is a matrix such that $\mathbf{A}'(\boldsymbol{\theta}_0)\,\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0)$. $\mathbf{A}(\boldsymbol{\theta}_0)$ is not uniquely defined but could be, for example, the Cholesky square root of $\mathbf{I}(\boldsymbol{\theta}_0)$ (which is guaranteed to exist since $\mathbf{I}(\boldsymbol{\theta}_0)$ is a covariance matrix and is therefore positive definite). Now, since $\hat{\boldsymbol{\theta}} \sim MVN\left(\boldsymbol{\theta}_0, \mathbf{I}^{-1}(\boldsymbol{\theta}_0)\right)$, we must have

$$\mathbf{A}(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right) \sim MVN\left(\mathbf{0}, \mathbf{A}(\boldsymbol{\theta}_0)\,\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\,\mathbf{A}'(\boldsymbol{\theta}_0)\right)$$

approximately. Now, assuming the information matrix $\mathbf{I}(\boldsymbol{\theta}_0)$ is nonsingular, we must have

$$\begin{aligned} \mathbf{A}(\boldsymbol{\theta}_0)\,\mathbf{I}^{-1}(\boldsymbol{\theta}_0)\,\mathbf{A}'(\boldsymbol{\theta}_0) &= \mathbf{A}(\boldsymbol{\theta}_0)\left[\mathbf{A}'(\boldsymbol{\theta}_0)\,\mathbf{A}(\boldsymbol{\theta}_0)\right]^{-1}\mathbf{A}(\boldsymbol{\theta}_0)' \\ &= \mathbf{A}(\boldsymbol{\theta}_0)\left[\mathbf{A}(\boldsymbol{\theta}_0)\right]^{-1}\left[\mathbf{A}'(\boldsymbol{\theta}_0)\right]^{-1}\mathbf{A}(\boldsymbol{\theta}_0)' = \mathbf{1}_{p\times p} \ , \end{aligned}$$

the $p \times p$ identity matrix. Together with (18), this shows that $2\left[\ell\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right) - \ell(\boldsymbol{\theta}_0|\mathbf{Y})\right]$ is approximately a sum of squares of p standard normal random variables. Asymptotically therefore,

$$2\left[\ell\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right) - \ell(\boldsymbol{\theta}_0|\mathbf{Y})\right] \sim \chi^2_p \ . \tag{19}$$

This is the multiparameter equivalent of (12).

Either (15) or (17) can be used to construct confidence intervals for individual parameters, as well as confidence *regions* for subsets of the parameters. Moreover, (19) allows the construction of a confidence region for the entire parameter vector — for example, an approximate 95% confidence region for $\boldsymbol{\theta}$ consists of all values such that $2\left[\ell\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right) - \ell(\boldsymbol{\theta}|\mathbf{Y})\right]$ is less than the 95% point of χ^2_p. As it stands however, (19) does not allow the construction of confidence regions for subsets of the parameter vector. We now address this problem.

2.4 Profile likelihood

Suppose now that the parameter vector is partitioned into two subsets: $\boldsymbol{\theta} = (\boldsymbol{\psi}' \ \boldsymbol{\lambda}')'$, with target value $\boldsymbol{\theta}_0 = (\boldsymbol{\psi}_0' \ \boldsymbol{\lambda}_0')'$. Write $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\psi}, \boldsymbol{\lambda})$ for the log-likelihood,

$$\mathbf{U}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{U}_\psi(\boldsymbol{\psi}, \boldsymbol{\lambda}) \\ \mathbf{U}_\lambda(\boldsymbol{\psi}, \boldsymbol{\lambda}) \end{pmatrix} \text{ for the score vector,}$$

$$\mathbf{I} = \begin{pmatrix} \mathbf{I}_{\psi\psi} & \mathbf{I}_{\psi\lambda} \\ \mathbf{I}_{\lambda\psi} & \mathbf{I}_{\lambda\lambda} \end{pmatrix} \text{ for } \operatorname{var}[\mathbf{U}(\boldsymbol{\theta}_0)], \quad \text{and } \mathbf{H} = \begin{pmatrix} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\lambda} \\ \mathbf{H}_{\lambda\psi} & \mathbf{H}_{\lambda\lambda} \end{pmatrix} \text{ for } \mathrm{E}\left[\partial^2\ell/\partial\boldsymbol{\theta}^2_{|\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right].$$

Suppose also that $\boldsymbol{\psi}$ is held fixed, and that the likelihood is maximised with respect to $\boldsymbol{\lambda}$ alone for this value of $\boldsymbol{\psi}$. In general, the resulting estimate of $\boldsymbol{\lambda}$ will depend on $\boldsymbol{\psi}$, so call it $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$. The value of the resulting maximised likelihood, $\ell\left(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})\right)$ will also depend on $\boldsymbol{\psi}$; this is called the PROFILE LIKELIHOOD for $\boldsymbol{\psi}$.

Let $\hat{\boldsymbol{\psi}}$ be the overall MLE for $\boldsymbol{\psi}$; then the overall MLE for $\boldsymbol{\lambda}$ is $\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)$. By definition, $\ell\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)\right)$ cannot be less than the maximised log-likelihood at any other value of $\boldsymbol{\psi}$. Therefore the likelihood ratio statistic

$$\Lambda(\boldsymbol{\psi}) = 2\left[\ell\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)\right) - \ell\left(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})\right)\right] \tag{20}$$

is always positive-valued, although we would expect $\Lambda(\boldsymbol{\psi}_0)$ to be 'small' in general, if $\hat{\boldsymbol{\psi}}$ is close to $\boldsymbol{\psi}_0$. This suggests that when $\boldsymbol{\psi}$ is unknown, a confidence region could be determined as the set of values for which $\Lambda(\boldsymbol{\psi})$ is less than some threshold — or equivalently, as the set of values for which the profile likelihood exceeds a corresponding threshold. An appropriate threshold can be determined by considering the distribution of $\Lambda(\boldsymbol{\psi}_0)$. We have

$$\Lambda(\boldsymbol{\psi}_0) = 2\left\{\left[\ell\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)\right) - \ell(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0)\right] - \left[\ell\left(\boldsymbol{\psi}_0, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)\right) - \ell(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0)\right]\right\} . \tag{21}$$

Now, using essentially the same argument as that given in the previous section we find that the term $2\left[\ell\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)\right) - \ell(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0)\right]$ can be written as

$$\begin{aligned}
& -\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right)' \mathbf{H}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) + O_p\left(n^{-1/2}\right) \\
= & -\left(\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \ \left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right)'\right)' \begin{pmatrix} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\lambda} \\ \mathbf{H}_{\lambda\psi} & \mathbf{H}_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0 \end{pmatrix} + O_p\left(n^{-1/2}\right) \\
= & -\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \mathbf{H}_{\psi\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) - \left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right)' \mathbf{H}_{\lambda\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) \\
& -\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \mathbf{H}_{\psi\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right) - \left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right)' \mathbf{H}_{\lambda\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right) + O_p\left(n^{-1/2}\right) .
\end{aligned} \tag{22}$$

For the second term in (21), the analysis can repeated as though $\boldsymbol{\psi}_0$ is known and $\boldsymbol{\lambda}$ is the unknown parameter vector, to yield

$$2\left[\ell\left(\boldsymbol{\psi}_0, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0)\right) - \ell(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0)\right] = -\left(\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0) - \boldsymbol{\lambda}_0\right)' \mathbf{H}_{\lambda\lambda}\left(\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi}_0) - \boldsymbol{\lambda}_0\right) + O_p\left(n^{-1/2}\right) . \tag{23}$$

We now substitute (22) and (23) into (21). This requires a relationship between $\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)$ and $\hat{\boldsymbol{\lambda}}\left(\boldsymbol{\psi}_0\right)$. To find this we use (16), which we now write as

$$\begin{pmatrix} \mathbf{U}_{\psi}\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)\right) \\ \mathbf{U}_{\lambda}\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)\right) \end{pmatrix} = \begin{pmatrix} \mathbf{U}_{\psi}\left(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0\right) \\ \mathbf{U}_{\lambda}\left(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0\right) \end{pmatrix} + \begin{pmatrix} \mathbf{H}_{\psi\psi} & \mathbf{H}_{\psi\lambda} \\ \mathbf{H}_{\lambda\psi} & \mathbf{H}_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0 \\ \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0 \end{pmatrix} + o_p\left(n^{1/2}\right) \ . \tag{24}$$

If $\boldsymbol{\psi}_0$ is known so that only $\boldsymbol{\lambda}$ is being estimated, the corresponding expansion is

$$\mathbf{U}_{\lambda}\left(\boldsymbol{\psi}_0, \hat{\boldsymbol{\lambda}}\left(\boldsymbol{\psi}_0\right)\right) = \mathbf{U}_{\lambda}\left(\boldsymbol{\psi}_0, \boldsymbol{\lambda}_0\right) + \mathbf{H}_{\lambda\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\boldsymbol{\psi}_0\right) - \boldsymbol{\lambda}_0\right) + o_p\left(n^{1/2}\right) \ . \tag{25}$$

Since the left hand sides of both (24) and (25) are zero by definition, we can equate (25) with the bottom row of (24) to obtain

$$\mathbf{H}_{\lambda\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\boldsymbol{\psi}_0\right) - \boldsymbol{\lambda}_0\right) = \mathbf{H}_{\lambda\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + \mathbf{H}_{\lambda\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right) + o_p\left(n^{1/2}\right) \ ,$$

so that

$$\hat{\boldsymbol{\lambda}}\left(\boldsymbol{\psi}_0\right) - \boldsymbol{\lambda}_0 = \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0 + \mathbf{H}_{\lambda\lambda}^{-1}\mathbf{H}_{\lambda\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + o_p\left(n^{1/2}\right) \ ,$$

the order of magnitude of the error following from the fact that the elements of $\mathbf{H}_{\lambda\lambda}$ are $O_p(n)$. The quadratic term in (23) can now be written as

$$\begin{aligned} &-\left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right)' \mathbf{H}_{\lambda\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right) - \left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right)' \mathbf{H}_{\lambda\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) \\ &-\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \mathbf{H}_{\psi\lambda}\left(\hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right) - \boldsymbol{\lambda}_0\right) - \left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \mathbf{H}_{\psi\lambda}\mathbf{H}_{\lambda\lambda}^{-1}\mathbf{H}_{\lambda\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + o_p(1) \ . \end{aligned} \tag{26}$$

We now combine (21), (22), (23) and (26), and find

$$\begin{aligned} \Lambda\left(\boldsymbol{\psi}_0\right) &= -\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \mathbf{H}_{\psi\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + \left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \mathbf{H}_{\psi\lambda}\mathbf{H}_{\lambda\lambda}^{-1}\mathbf{H}_{\lambda\psi}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + o_p(1) \\ &= -\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \left[\mathbf{H}_{\psi\psi} - \mathbf{H}_{\psi\lambda}\mathbf{H}_{\lambda\lambda}^{-1}\mathbf{H}_{\lambda\psi}\right]\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + o_p(1) \ . \end{aligned} \tag{27}$$

Next, observe that $\left[\mathbf{H}_{\psi\psi} - \mathbf{H}_{\psi\lambda}\mathbf{H}_{\lambda\lambda}^{-1}\mathbf{H}_{\lambda\psi}\right]^{-1}$ is the submatrix of $\mathbf{H}^{-1}$ corresponding to $\boldsymbol{\psi}$ (this is not immediately obvious, but is a standard result in matrix algebra — see, for example, Horn and Johnson 1985, page 18). For notational convenience therefore, if we write $\mathbf{H}^{-1}$ as

$$\mathbf{H}^{-1} = \begin{pmatrix} \mathbf{H}^{(\psi\psi)} & \mathbf{H}^{(\psi\lambda)} \\ \mathbf{H}^{(\lambda\psi)} & \mathbf{H}^{(\lambda\lambda)} \end{pmatrix} \ ,$$

the likelihood ratio statistic (27) can be written as

$$\Lambda\left(\boldsymbol{\psi}_0\right) = -\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right)' \left[\mathbf{H}^{(\psi\psi)}\right]^{-1}\left(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\right) + o_p(1) \ . \tag{28}$$

But from (17), the distribution of $\hat{\boldsymbol{\theta}}$ is approximately $MVN\left(\boldsymbol{\theta}_0, \mathbf{I}^{-1}\right) = MVN\left(\boldsymbol{\theta}_0, -\mathbf{H}^{-1}\right)$, so that the distribution of $\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0$ is approximately $MVN\left(\mathbf{0}, -\mathbf{H}^{(\psi\psi)}\right)$. The development from (18) to (19) can now be repeated, to conclude that in large samples

$$\Lambda\left(\boldsymbol{\psi}_0\right) \sim \chi^2_q \tag{29}$$

approximately, where q is the dimension of $\boldsymbol{\psi}$. Therefore, if $\boldsymbol{\psi}_0$ is unknown, a confidence region can be determined as the set $\{\boldsymbol{\psi} : \Lambda\left(\boldsymbol{\psi}\right) < c\}$, where c is the appropriate percentile of the χ^2_q distribution.

3 Estimating equations

In the likelihood setting above, the asymptotic results depend on the following properties of the score function:

1. The expected score is zero at the true parameter value.
2. The variance of the score can be calculated at the true parameter value.
3. The score, when suitably normalised, tends to its expectation as $n \to \infty$.
4. The score is a continuous function of the parameter, in the neighbourhood of the true parameter value.
5. As $n \to \infty$, the distribution of the score vector tends to the multivariate normal.
6. The second derivative of the score is bounded in the neighbourhood of the true parameter value (this was not made explicit in the discussion above, but is necessary to control the magnitudes of the various approximations leading to (8) and (17)).

These observations suggest that, as an alternative to likelihood-based inference, we may consider obtaining parameters by solving the equation

$$\mathbf{g}\left(\hat{\boldsymbol{\theta}}|\mathbf{y}\right) = \mathbf{0}\ , \tag{30}$$

where $\mathbf{g}$ is a function such that the associated vector of random variables $\mathbf{g}_{\boldsymbol{\theta}}$ has properties 1–6 above. Any such equation is called an ESTIMATING EQUATION; we will call $\mathbf{g}_{\boldsymbol{\theta}}$ an ESTIMATING FUNCTION. If $\hat{\boldsymbol{\theta}}$ solves an estimating equation of the form (30), then the arguments of the previous section can be repeated. Let $\mathbf{H}\left(\boldsymbol{\theta}\right)$ denote the expected value of the Hessian matrix $\mathbf{H}_{\boldsymbol{\theta}} = \partial \mathbf{g}_{\boldsymbol{\theta}} / \partial \boldsymbol{\theta}$, and let $\mathbf{J}\left(\boldsymbol{\theta}\right)$ be the covariance matrix of $\mathbf{g}_{\boldsymbol{\theta}}$. Then for large n,

$$\hat{\boldsymbol{\theta}} \sim MVN\left(\boldsymbol{\theta}_0, \mathbf{V}\left(\boldsymbol{\theta}_0\right)\right) \text{ approximately,} \tag{31}$$

where $\mathbf{V}\left(\boldsymbol{\theta}\right) = \left[\mathbf{H}\left(\boldsymbol{\theta}\right)\right]^{-1} \mathbf{J}\left(\boldsymbol{\theta}\right) \mathbf{H}\left(\boldsymbol{\theta}\right)^{-1}$.

Many numerical optimisation methods (for example, those based around Newton-Raphson iterative schemes) compute the Hessian $\mathbf{H}_{\hat{\boldsymbol{\theta}}}$ as a by-product of the optimisation procedure.

We may therefore use this to approximate $\mathbf{H}(\boldsymbol{\theta}_0)$ in the calculation of $\mathbf{V}(\boldsymbol{\theta}_0)$. Further simplification is possible if we can choose $\mathbf{g}$ in such a way that $\mathbf{H}(\boldsymbol{\theta}) = -\mathbf{J}(\boldsymbol{\theta})$ (from (14), this is the case for score-based estimation), since in this case $\mathbf{V}(\boldsymbol{\theta}) = -[\mathbf{H}(\boldsymbol{\theta})]^{-1}$.

As in the case of likelihood-based inference, tests of hypotheses can be based either on (31) (which is the equivalent of a Wald test) or on the multivariate normal distribution of the estimating function itself (the equivalent of a score test):

$$\mathbf{g}_{\boldsymbol{\theta}_0} \sim MVN(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0)) \ . \tag{32}$$

Much of the literature on estimating equations takes (30) as its starting point. In this case, the resulting estimate is not necessarily the maximiser (or minimiser) of a function such as the log-likelihood, whence there is no obvious equivalent to the likelihood ratio test. However, a generalisation *is* possible if $\mathbf{g}_{\boldsymbol{\theta}}$ is the gradient vector of some objective function. Specifically, suppose that the estimating equations arise from minimising a measure of discrepancy between data and model, $S(\boldsymbol{\theta}|\mathbf{y})$ say. In this case, a confidence region could be defined as the set of all values for which $S(\boldsymbol{\theta}|\mathbf{y})$ is less than some threshold. We now establish what this threshold should be.

A second-order Taylor expansion of S about $\hat{\boldsymbol{\theta}}$ yields

$$2\left[S(\boldsymbol{\theta}_0|\mathbf{Y}) - S\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right)\right] \ = \ \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right)' \mathbf{H}(\boldsymbol{\theta}_0)\left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}\right) \ + o_p(1).$$

This is dominated by the first term, which is a quadratic form in normal random variables. Distributions of quadratic forms are difficult to compute exactly. However, it is common (e.g. Bowman and Azzalini 1997, p.88) to approximate their quantiles with those of a scaled and shifted χ^2 distribution. The shift, scale and degrees of freedom of the approximating distribution are chosen to match the first three moments (or equivalently, cumulants) of the quadratic form. The rth cumulant is given (Kuonen, 1999, Section 2) by

$$\kappa_r = 2^{r-1}\Gamma(r)\mathrm{tr}\left\{[\mathbf{V}(\boldsymbol{\theta}_0)\mathbf{H}(\boldsymbol{\theta}_0)]^r\right\} \ , \tag{33}$$

with tr() denoting the trace operator. The distribution of the quadratic form is then approximated by that of $aX + c$, where $X \sim \chi^2_b$ and

$$a = \frac{|\kappa_3|}{4\kappa_2} \qquad b = \frac{8\kappa_2^3}{\kappa_3^2} \qquad c = \kappa_1 - ab \ . \tag{34}$$

In practice, it is necessary to replace $\boldsymbol{\theta}_0$ in (33) with $\hat{\boldsymbol{\theta}}$. Since $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right) = \left[\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1} \mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)\left[\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right)\right]^{-1}$, we therefore compute

$$\kappa_r = 2^{r-1}\Gamma(r)\mathrm{tr}\left\{\left[\mathbf{H}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)\right]^r\right\} \ , \tag{35}$$

This then yields a reasonably straightforward procedure for constructing confidence regions based on the values of the objective function. For example, a 95% region consists of the set of values for which

$$a^{-1}\left\{2\left[S(\boldsymbol{\theta}_0|\mathbf{Y}) - S\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right)\right] - c\right\}$$

is less than the 95th percentile of the χ^2_b distribution.

Confidence regions for subsets of the parameters can be constructed in a similar way, following the theory outlined in Section 2.4. Specifically, suppose a profile objective function is calculated for a subvector $\boldsymbol{\psi}$, by holding this subvector fixed and maximising over the remaining parameters $\boldsymbol{\lambda}$, say. With notation as in Section 2.4, define a profile test statistic as

$$\Lambda(\boldsymbol{\psi}) = 2\left[S\left(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})\,|\mathbf{Y}\right) - S\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}}\left(\hat{\boldsymbol{\psi}}\right)|\mathbf{Y}\right)\right], \tag{36}$$

which is positive-valued by definition. Expansion (28) holds here, as in the likelihood setting; the difference is, once again, that for estimating equations the covariance matrix of $\hat{\boldsymbol{\psi}}$ is not directly related to the Hessian. It should be clear, however, that the same procedure can be applied as in the case of the full parameter vector above. All that is required is to replace $\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right)$ with the estimated value of $\left[\mathbf{H}^{(\psi\psi)}\right]^{-1}$, and to extract the submatrix of $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)$ corresponding to $\boldsymbol{\psi}$. In fact, once the full matrices $\mathbf{H}^{-1}$ and $\mathbf{V}$ have been estimated, the modification simply consists of extracting the elements corresponding to $\boldsymbol{\psi}$ from each of these matrices when calculating the κs in (33). Notice that, in general, different choices of $\boldsymbol{\psi}$ will lead to different thresholds. Notice also that the cancellation leading to (35) does not hold in general, when considering subsets of the parameter vector.

If $\boldsymbol{\psi}$ consists of a single parameter ψ, the procedure outlined above is particularly simple. In this case $\left[\mathbf{H}^{(\psi\psi)}\right]^{-1}$ is a scalar, as is the corresponding submatrix of $\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)$. Denote these scalars by h and v respectively; then direct calculation shows that the constants defined in (34) are given by $a = hv$, $b = 1$ and $c = 0$. A confidence interval for ψ can therefore be defined as the set values for which the profile test statistic (36) is less than hv times the appropriate percentage point of a χ^2_1 distribution.

There is a substantial body of theory on the use of estimating equations. However, for current purposes there is no need to go beyond what has been presented above.

4 Estimating equations for the method of moments

We now return to the problem of parameter estimation for stochastic-mechanistic models, using a generalised method of moments. To formalise the problem, it is necessary to establish some notation. Specifically:

- Let $\mathbf{y}$ be a vector of observations as previously; this is regarded as the realised value of a vector $\mathbf{Y}$ of random variables.
- Let $\boldsymbol{\theta} = (\theta_1 \ \ldots \ \theta_p)'$ be a vector of unknown parameters in the model.
- Let $\mathbf{T}(\mathbf{y}) = (T_1(\mathbf{y}) \ \ldots \ T_k(\mathbf{y}))'$ be a vector of summary statistics computed from the observations. $\mathbf{T}(\mathbf{y})$ is the realised value of a random vector $\mathbf{T} = (T_1 \ \ldots \ T_k)$ say. Denote the expected value of this random vector by $\mathrm{E}_{\boldsymbol{\theta}}(\mathbf{T}) = \boldsymbol{\tau}(\boldsymbol{\theta}) = (\tau_1(\boldsymbol{\theta}) \ \ldots \ \tau_k(\boldsymbol{\theta}))'$.

The idea here is that $\mathbf{T}$ is a vector of data properties (means, variances, autocorrelations etc.) and that $\boldsymbol{\tau}(\boldsymbol{\theta})$ is the corresponding set of theoretical properties derived from the model. The generalised method of moments seeks to minimise some measure of disagreement between $\mathbf{T}$ and $\boldsymbol{\tau}(\boldsymbol{\theta})$. Following the notation above, denote this measure by $S(\boldsymbol{\theta}|\mathbf{y})$. In practice, this is invariably a (possibly weighted) sum of squares:

$$S(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{k} w_i(\boldsymbol{\theta}) \left[T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})\right]^2 \ . \tag{37}$$

for some collection of positive weights $\{w_i(\boldsymbol{\theta}) : i = 1, \ldots, k\}$. For the moment, we allow the possibility that these may be parameter-dependent, although we will see below that this is actually a bad idea. In well-behaved problems, the minimiser of this function satisfies the vector equation $\mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) = \partial S/\partial\boldsymbol{\theta} = \mathbf{0}$. We have

$$\mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{k} \left\{ \frac{\partial w_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \left[T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})\right]^2 - 2w_i(\boldsymbol{\theta}) \frac{\partial\tau_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \left[T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})\right] \right\} \ , \tag{38}$$

so that the parameter estimate $\hat{\boldsymbol{\theta}}$ satisfies $\mathbf{g}\left(\hat{\boldsymbol{\theta}}|\mathbf{y}\right) = \mathbf{0}$, as in (30).

To apply the theory of estimating equations here, we need to ensure that properties 1–6 in Section 3 are satisfied by the random variables $\mathbf{g}_{\boldsymbol{\theta}}$ whose values are given by (38). Properties 4 and 6 ($\mathbf{g}_{\boldsymbol{\theta}}$ is continuous in $\boldsymbol{\theta}$ with bounded second derivatives) are unlikely to cause problems. The remainder require some thought.

4.1 Zero mean

For $\mathbf{g}_{\boldsymbol{\theta}}$ to have zero mean, we require

$$\sum_{i=1}^{k} \left\{ \frac{\partial w_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}}\left[T_i - \tau_i(\boldsymbol{\theta})\right]^2 - 2w_i(\boldsymbol{\theta}) \frac{\partial\tau_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \mathrm{E}_{\boldsymbol{\theta}}\left[T_i - \tau_i(\boldsymbol{\theta})\right] \right\} \ = \ \mathbf{0}$$

at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Since $E(\mathbf{T}) = \boldsymbol{\tau}(\boldsymbol{\theta}_0)$, this reduces to the requirement that

$$\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \mathrm{var}_{\boldsymbol{\theta}}(T_i) = \mathbf{0} \ ,$$

which is trivially true providing the weights are independent of $\boldsymbol{\theta}$. If the weights depend on $\boldsymbol{\theta}$, however, the requirement is not fulfilled in general. In particular, it is not fulfilled if $w_i(\boldsymbol{\theta})$ is set proportional to $1/\mathrm{var}_{\boldsymbol{\theta}}(T_i)$ (which is a natural weighting scheme to consider, given the received wisdom that 'in least squares problems with unequal variances, observations should be weighted according to the inverse of their variances', and that in such problems, the weighted least squares estimates are known to be unbiased). To see this, consider any collection of weights satisfying

$$\sum_{i=1}^{k} w_i(\boldsymbol{\theta}) \, \mathrm{var}_{\boldsymbol{\theta}}(T_i) = \text{constant, independent of } \boldsymbol{\theta}.$$

Differentiating both sides with respect to θ_j yields

$$\sum_{i=1}^{k} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j} \mathrm{var}(T_i) + \sum_{i=1}^{k} w_i(\boldsymbol{\theta}) \frac{\partial \mathrm{var}_{\boldsymbol{\theta}}(T_i)}{\partial \theta_j} = 0$$

in which case, the first term can only be zero if the second is also. But since the ws are positive, the second term can only be zero if $\mathrm{var}_{\boldsymbol{\theta}}(T_i)$ is independent of θ_j for each i. Also, since each element of $\mathbf{g}_{\boldsymbol{\theta}}$ must have zero expectation, the result must hold for all j: hence no collection of weights satisfying the constraint above will yield a valid estimating equation, unless $\mathrm{var}_{\boldsymbol{\theta}}(T_i)$ is independent of $\boldsymbol{\theta}$ for each i.

At first sight, this appears to contradict the 'standard' theory of weighted least squares in regression problems. The resolution of the problem appears to lie in the fact that in regression problems, the weights do not depend on the regression parameters (which are the θs in the present context) — hence $\partial w_i/\partial \boldsymbol{\theta} = 0$ in such problems. I suspect that the difficulty, when the weights depend on $\boldsymbol{\theta}$, is related to the known problems of bias in estimating equations when nuisance parameters are present (Liang and Zeger, 1995), although the current setting is slightly different.

The upshot of all this is that if we want to weight the fitting properties, the weights should not depend on $\boldsymbol{\theta}$; otherwise the resulting estimates will be biased (at least, for the kind of weighting scheme that may be considered in practice). In many situations, it is likely that the bias will tend to zero as the sample size (i.e. dimension of $\mathbf{Y}$) increases. However, as a first step in obtaining sampling distributions for moment-based estimators, it seems reasonable to restrict ourselves to estimators that are exactly unbiased. Hence the objective function (37) becomes

$$S(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{k} w_i \left[T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})\right]^2 \tag{39}$$

and the corresponding estimating equation becomes

$$\mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) = -2\sum_{i=1}^{k} w_i \frac{\partial \tau_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left[T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})\right] = \mathbf{0}\ . \tag{40}$$

The factor of -2 is retained here to avoid confusion later on.

4.1.1 $\boldsymbol{\theta}$-dependent weights — a cunning plan

The problems above, regarding the use of weights depending on $\boldsymbol{\theta}$, can be resolved completely if we modify the objective function (37) slightly, to

$$\sum_{i=1}^{k} \left\{ w_i(\boldsymbol{\theta}) \left[T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})\right]^2 - \ln w_i(\boldsymbol{\theta}) \right\}\ . \tag{41}$$

If we do this, the estimating function becomes

$$\mathbf{g}_{\boldsymbol{\theta}} = \sum_{i=1}^{k} \left\{ \frac{\partial w_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left[\left[T_i - \tau_i(\boldsymbol{\theta})\right]^2 - \frac{1}{w_i(\boldsymbol{\theta})} \right] - 2w_i(\boldsymbol{\theta}) \frac{\partial \tau_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \left[T_i - \tau_i(\boldsymbol{\theta})\right] \right\} = \mathbf{0}\ ,$$

which clearly has zero expectation if we set $w_i(\boldsymbol{\theta}) = 1/\mathrm{var}_{\boldsymbol{\theta}}(T_i)$. We do not pursue this any further here; however, it may be prove useful in the future.

4.2 Asymptotic normality

From (40), it is clear that the estimating function $\mathbf{g}_{\boldsymbol{\theta}}$ will have an approximate normal distribution if either of the following conditions hold:

1. k is large, and the components of $\mathbf{T}$ are not too strongly dependent. For in this case, $\mathbf{g}_{\boldsymbol{\theta}}$ is a sum of a large number of terms and the Central Limit Theorem applies.
2. $\mathbf{T}$ itself has an approximate multivariate normal distribution.

In practice, providing the elements of $\mathbf{T}$ are chosen appropriately, condition 2 is likely to be satisfied for large datasets, since most statistics of interest have an approximate normal distribution in the limit. Obviously, the closer this approximation, the better will be the normal approximation to the distribution of $\mathbf{g}_{\boldsymbol{\theta}}$. This suggests that we should seek fitting properties with distributions that are 'as normal as possible', for example by transformation.

4.3 Consistency

The moment estimator will be consistent if $\mathrm{E}[\mathbf{g}_{\boldsymbol{\theta}_0}] = \mathbf{0}$ and, when suitably normalised, $\mathbf{g}_{\boldsymbol{\theta}}$ converges in probability to its expectation as $n \to \infty$. Again from (40), this convergence will occur if $\mathbf{T}$ converges to $\boldsymbol{\tau}(\boldsymbol{\theta})$; and again, most statistics of interest *do* converge to their expectations in the required sense.

4.4 Variance calculation

To complete the estimating equation framework it is necessary to calculate, or at least estimate, $\mathbf{J}(\boldsymbol{\theta}_0) = \mathrm{var}[\mathbf{g}_{\boldsymbol{\theta}_0}]$, since this is required for the calculation of $\mathbf{V}(\boldsymbol{\theta}_0)$ in (31) and (33). A number of options are available here:

1. Find an analytical expression for $\mathbf{J}(\boldsymbol{\theta})$, and use $\mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)$ as an estimate of $\mathbf{J}(\boldsymbol{\theta}_0)$.
2. Obtain an empirical estimate of $\mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)$, and use this to estimate $\mathbf{J}(\boldsymbol{\theta}_0)$.
3. If possible, set up the estimating equation in such a way that $\mathbf{J}(\boldsymbol{\theta}_0) \propto \mathbf{H}(\boldsymbol{\theta}_0)$. In this case, $\mathbf{V}(\boldsymbol{\theta}_0) \propto [\mathbf{H}(\boldsymbol{\theta})]^{-1}$ and we can use the observed Hessian to estimate $\mathbf{V}(\boldsymbol{\theta}_0)$ without ever needing to calculate $\mathbf{J}(\boldsymbol{\theta}_0)$.

For the first two options, it may be useful to note that $\mathbf{g}_{\boldsymbol{\theta}}$ can be written in matrix form as

$$\mathbf{g}_{\boldsymbol{\theta}} = -2\left[\mathbf{W}(\boldsymbol{\theta})\right]'\left(\mathbf{T} - \boldsymbol{\tau}(\boldsymbol{\theta})\right) \ ,$$

where $\mathbf{W}(\boldsymbol{\theta})$ is a $k \times p$ matrix whose (i,j)th element is $w_i \partial \tau_i(\boldsymbol{\theta}) / \partial \theta_j$. Standard results for covariance matrices then give us

$$\mathbf{J}(\boldsymbol{\theta}) = 4\text{var}\left\{ [\mathbf{W}(\boldsymbol{\theta})]' (\mathbf{T} - \boldsymbol{\tau}(\boldsymbol{\theta})) \right\} = 4 [\mathbf{W}(\boldsymbol{\theta})]' \text{var}(\mathbf{T}) \mathbf{W}(\boldsymbol{\theta}) \ . \tag{42}$$

Hence $\mathbf{J}(\boldsymbol{\theta})$ can be calculated from the covariance matrix of $\mathbf{T}$. A specific suggestion for estimating this covariance matrix empirically is given in Section 5 below. In practice, the derivatives of $\boldsymbol{\tau}(\boldsymbol{\theta})$ appearing in $\mathbf{W}(\boldsymbol{\theta})$ can be evaluated numerically if necessary.

4.4.1 Variance calculation using the Hessian

In the third option above, the idea is to define the objective function in such a way that $\mathbf{J}(\boldsymbol{\theta}_0) \propto \mathbf{H}(\boldsymbol{\theta}_0) = \text{E}[\partial \mathbf{g}_{\boldsymbol{\theta}} / \partial \boldsymbol{\theta}]$. We now investigate how to achieve this. The starting point is the zero-mean requirement for the estimating function, which implies that

$$\int \mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) f(\mathbf{y};\boldsymbol{\theta}) \, d\mathbf{y} = \mathbf{0} \ .$$

Differentiating both sides with respect to $\boldsymbol{\theta}$ yields

$$\int \left[\frac{\partial \mathbf{g}(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} f(\mathbf{y};\boldsymbol{\theta}) + \mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) \left(\frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right] d\mathbf{y} = \mathbf{0} \ ,$$

so that

$$\begin{aligned} \text{E}\left[\frac{\partial \mathbf{g}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}} \right] &= -\int \mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) \left(\frac{1}{f(\mathbf{y};\boldsymbol{\theta})} \frac{\partial f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' f(\mathbf{y};\boldsymbol{\theta}) \, d\mathbf{y} \\ \text{i.e.} \quad \text{E}[\mathbf{H}_{\boldsymbol{\theta}}] &= -\text{E}\left[\mathbf{g}_{\boldsymbol{\theta}} \left(\frac{\partial \ln f(\mathbf{Y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right] \ . \end{aligned}$$

At $\boldsymbol{\theta}_0$, we require the right-hand side here to be proportional to $\mathbf{J}(\boldsymbol{\theta}) = \text{var}[\mathbf{g}_{\boldsymbol{\theta}}] = \text{E}[\mathbf{g}_{\boldsymbol{\theta}} \mathbf{g}_{\boldsymbol{\theta}}']$. It is not obvious that this is the case, unless $\mathbf{g}(\boldsymbol{\theta}|\mathbf{y})$ is proportional to the score function

$$\mathbf{g}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\partial \ln f(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \ .$$

However, note that $\mathbf{g}_{\boldsymbol{\theta}}$ is a function of $\mathbf{T}$, which in turn is a function of $\mathbf{Y}$. Hence we can obtain an equivalent development to the one above, by taking expectations with respect to $\mathbf{T}$; the only difference is that $f(\mathbf{y};\boldsymbol{\theta})$ will be replaced by the density of $\mathbf{T}$, $f_{\mathbf{T}}$ say, throughout.

This shows that $\mathbf{J}(\boldsymbol{\theta}_0) \propto \mathbf{H}(\boldsymbol{\theta}_0)$, as required, if $\mathbf{g}(\boldsymbol{\theta}|\mathbf{y})$ is the $\boldsymbol{\theta}$-derivative of the log density for $\mathbf{T}$. But in Section 4.2 above we argued that for large samples, $\mathbf{T}$ is likely to have an approximate multivariate normal distribution. Suppose, for the sake of argument, that we can choose $\mathbf{T}$ in such a way that (i) its elements are mutually uncorrelated (ii) $\text{var}[T_i]$ is independent of $\boldsymbol{\theta}$ for each i. In this case the $\boldsymbol{\theta}$-derivative of the log density has jth element

$$\frac{\partial \ln f_{\mathbf{T}}}{\partial \theta_j} = \sum_{i=1}^{n} \frac{1}{\text{var}[T_i]} \frac{\partial \tau_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} [T_i(\mathbf{y}) - \tau_i(\boldsymbol{\theta})]$$

which, if we take the weight $w_i = 1/\text{var}[T_i]$, is equal to $-\frac{1}{2}\mathbf{g}(\boldsymbol{\theta}|\mathbf{y})$ in (40). In this case, therefore, $\text{var}\left[\partial \ln f_{\mathbf{T}}/\partial\theta\right] = \frac{1}{4}\text{var}[\mathbf{g}_{\boldsymbol{\theta}}]$, and $\text{E}\left[\partial^2 \ln f_{\mathbf{T}}/\partial\theta^2\right] = -\frac{1}{2}\text{E}[\partial\mathbf{g}_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}]$. Since the left-hand sides here differ by a factor of -1, we must have $\mathbf{J}(\boldsymbol{\theta}) = \text{var}[\mathbf{g}_{\boldsymbol{\theta}}] = 2\text{E}[\partial\mathbf{g}_{\boldsymbol{\theta}}/\partial\boldsymbol{\theta}] = 2\mathbf{H}(\boldsymbol{\theta})$.

Of course, it is unrealistic to expect that the elements of $\mathbf{T}$ should be uncorrelated and that their variances should be independent of $\boldsymbol{\theta}$. The argument above does suggest, however, that if we choose $\mathbf{T}$ in such a way that as many components as possible have variances that are independent of $\boldsymbol{\theta}$; and to set the weights for the remaining components to a 'ballpark' figure that roughly reflects their uncertainty, $\mathbf{J}(\boldsymbol{\theta}_0)$ should be approximated reasonably by $2\mathbf{H}(\boldsymbol{\theta}_0)$ so that $\mathbf{V}(\boldsymbol{\theta}_0)$ can be calculated as $2\left[\mathbf{H}(\boldsymbol{\theta}_0)\right]^{-1}$.

In passing, it is also worth noting that $\boldsymbol{\theta}$-dependent weights can be accommodated within this framework, by changing the objective function to (41). In this case, the resulting estimating function is exactly the $\boldsymbol{\theta}$-derivative of a normal density for uncorrelated Ts.

5 Summary, and implications

The main points to emerge from the discussion above are the following:

1. Using a generalised method of moments, unbiased estimators can be obtained by minimising an expression of the form

$$S(\boldsymbol{\theta}|\mathbf{Y}) = \sum_{i=1}^{k} w_i \left[T_i(\mathbf{Y}) - \tau_i(\boldsymbol{\theta})\right]^2$$

   where the Ts are properties of the data and the τs are their expected values under the model.

2. The weights $\{w_i\}$ must *not* depend on the model parameters (or on the data!). If parameter-dependent weights are used, the objective function must be modified to that given in (41).

3. Under fairly general conditions, the estimator resulting from the above minimisation has a multivariate normal distribution. This can be used, for example, to construct approximate confidence intervals for the model parameters. The mean of the distribution is $\boldsymbol{\theta}_0$ (the true parameter vector), and its covariance matrix is $\mathbf{V}(\boldsymbol{\theta}_0)$ where $\mathbf{V}(\boldsymbol{\theta}) = \left[\mathbf{H}(\boldsymbol{\theta})\right]^{-1}\mathbf{J}(\boldsymbol{\theta})\,\mathbf{H}(\boldsymbol{\theta})^{-1}$. Here, $\mathbf{H}(\boldsymbol{\theta})$ is the expected second derivative of the objective function, which can be estimated from the Hessian output of a numerical minimisation routine. $\mathbf{J}(\boldsymbol{\theta})$ is the covariance matrix of the objective function derivatives.

4. An alternative way to construct confidence regions uses the objective function itself. Specifically, an approximate confidence region at a specified level consists of all points $\boldsymbol{\theta}$ such that

$$a^{-1}\left\{2\left[S(\boldsymbol{\theta}|\mathbf{Y}) - S\left(\hat{\boldsymbol{\theta}}|\mathbf{Y}\right)\right] - c\right\}$$

is less than the appropriate percentile of a chi-squared distribution with b degrees of freedom. The constants a, b and c are given by

$$a = \frac{|\kappa_3|}{4\kappa_2} \qquad b = \frac{8\kappa_2^3}{\kappa_3^2} \qquad c = \kappa_1 - ab \ ,$$

with $\kappa_r = 2^{r-1}\Gamma(r)\mathrm{tr}\left\{\left[\mathbf{V}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{H}\left(\hat{\boldsymbol{\theta}}\right)\right]^r\right\} = 2^{r-1}\Gamma(r)\mathrm{tr}\left\{\left[\mathbf{H}^{-1}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)\right]^r\right\}$.

Confidence regions for subsets of parameters can be constructed using profile objective functions, defined for a subset of parameters $\boldsymbol{\psi}$ as $S(\boldsymbol{\psi}) = S\left(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})|\mathbf{Y}\right)$ where $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ minimises the objective function for a fixed value of $\boldsymbol{\psi}$. The procedure is exactly the same as for the full parameter vector, except that the κs are calculated from the appropriate submatrices of $\mathbf{H}^{-1}$ and $\mathbf{V}$. In the case of a single parameter, let v be the appropriate diagonal element of $\mathbf{V}$, and h^{-1} the corresponding element of $\mathbf{H}^{-1}$; then $a = hv$, $b = 1$ and $c = 0$ in this case.

5. A final way to carry out tests uses the fact that at the true parameter value $\boldsymbol{\theta}_0$, the objective function gradient vector is distributed as $MVN(\mathbf{0}, \mathbf{J}(\boldsymbol{\theta}_0))$. Any $\boldsymbol{\theta}$ where the gradient is 'large' according to this distribution is therefore not supported by the data. This does not require calculation of the Hessian, which may be seen as a potential advantage.

6. The matrix $\mathbf{J}(\boldsymbol{\theta}_0)$ can be estimated in any of three ways:

   (a) Find an analytical expression for the covariance matrix of the fitting properties under the model; then estimate $\mathbf{J}(\boldsymbol{\theta}_0)$ as $4\left[\mathbf{W}\left(\hat{\boldsymbol{\theta}}\right)\right]' \mathrm{var}_{\hat{\boldsymbol{\theta}}}(\mathbf{T})\mathbf{W}\left(\hat{\boldsymbol{\theta}}\right)$, where $\mathbf{W}(\boldsymbol{\theta})$ is a $k \times p$ matrix whose (i,j)th element is $w_i \partial\tau_i(\boldsymbol{\theta})/\partial\theta_j$. If necessary, use numerical differentiation to evaluate $\partial\tau_i(\boldsymbol{\theta})$.

   (b) Calculate an empirical estimate of $\mathrm{var}(\mathbf{T})$, and use $4\left[\mathbf{W}\left(\hat{\boldsymbol{\theta}}\right)\right]' \widehat{\mathrm{var}}(\mathbf{T})\mathbf{W}\left(\hat{\boldsymbol{\theta}}\right)$ as an estimate of $\mathbf{J}(\boldsymbol{\theta}_0)$. For example, if $n > 1$ years of data are available, fitting properties $\mathbf{T}_1, \ldots, \mathbf{T}_n$ can be computed separately for each year: $\mathbf{T}$ can then be taken as the mean over all years, and $\widehat{\mathrm{var}}(\mathbf{T})$ as $n^{-2}\sum_{i=1}^n (\mathbf{T}_i - \mathbf{T})(\mathbf{T}_i - \mathbf{T})'$. This suggestion follows Rodriguez-Iturbe et al. (1988). Notice, however, that some components of each $\mathbf{T}_i$, in particular those relating to daily data, will be computed using relatively small samples. It is therefore important to use estimators that are, as far as possible, unbiased in small samples. This applies particularly to estimators of autocorrelation coefficients, for example — standard estimators can suffer from serious bias problems in small samples. Methods for correcting this are given by Kendall and Ord (1990, page 79), for example.

   (c) Choose fitting properties $\{T_i\}$ in such a way that (i) the chosen properties are approximately uncorrelated (ii) as many components as possible have variances that are independent of $\boldsymbol{\theta}$. Make an educated guess as to the variances of the remaining properties. Then, in the objective function, set $w_i = 1/\mathrm{var}[T_i]$. Throw the result

at a nonlinear minimisation routine that returns the Hessian as a by-product. Multiply this Hessian by 2, and take the result as an estimate of $\mathbf{J}(\boldsymbol{\theta}_0)$; invert this to obtain an estimate of $\mathbf{V}(\boldsymbol{\theta}_0)$ without any further matrix multiplication.

For practical purposes, $\mathbf{J}(\boldsymbol{\theta}_0)$ and $\mathbf{J}\left(\hat{\boldsymbol{\theta}}\right)$ are interchangeable.

The guidelines in (6c) above, regarding choice of fitting properties, apply more generally — indeed, lack of correlation was one of the criteria given by Rodriguez-Iturbe et al. (1988) for choosing fitting properties. The theory outlined in the preceding sections also suggests the following considerations:

1. The chosen statistics should be unbiased for the corresponding theoretical properties i.e. $\mathrm{E}[T_i] = \tau_i$.

2. The chosen statistics should have a normal distribution to a reasonable degree of approximation. This might involve, for example, taking logarithms of quantities that are essentially positive (another suggestion of Rodriguez-Iturbe et al. 1988), or applying a z-transformation to autocorrelations as in Wheater et al. (2000, Section 2.8.5).

3. The chosen statistics should have variances that are as small as possible. This is intuitively obvious; in terms of the mathematics, it is easiest to see in the case of a single parameter, so that all matrices become scalars. In this case the variance of the parameter estimate is proportional to a weighted sum of variances of fitting properties.

4. The chosen statistics should vary rapidly with respect to the model parameters. Mathematically, this requirement corresponds to large values of the Hessian matrix (i.e. the matrix of derivatives of fitting properties with respect to parameters). For a single parameter, the variance is inversely proportional to this Hessian.

The theory outlined here represents an alternative to the approach suggested at the bottom of page 290 of Rodriguez-Iturbe et al. (1988). Instead of calculating the covariance matrix $\mathbf{V}(\boldsymbol{\theta}_0)$, they suggested perturbing each of the fitting properties by a small amount, and re-estimating the model parameters at each of the perturbed configurations. This determines an approximate linear transformation from fitting properties to parameter estimates, which can be combined with an estimate of $\mathrm{var}(\mathbf{T})$ to estimate the covariance matrix. The difference here is that we avoid refitting the model many times by transforming in the opposite direction (from $\boldsymbol{\theta}$ to $\boldsymbol{\tau}$ rather than from $\mathbf{T}$ to $\hat{\boldsymbol{\theta}}$) and using an analytical (or numerical) linearisation of the transformation in the matrix $\mathbf{W}\left(\hat{\boldsymbol{\theta}}\right)$.

A further development is the ability to judge parameter sets on the basis of the objective function itself. This can be used, for example, to identify the region of the parameter space for which the objective function is 'almost' optimal.

References

Bowman, A. and Azzalini, A. (1997). *Applied smoothing techniques for data analysis — the kernel approach with S-Plus illustrations*, volume 18 of *Oxford Statistical Science series*. Oxford University Press, Oxford.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman & Hall, London.

Horn, R. and Johnson, C. (1985). *Matrix Analysis*. Cambridge University Press, Cambridge.

Kendall, M. and Ord, J. (1990). *Time Series (third edition)*. Edward Arnold.

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika*, 86(4):929–935.

Liang, K.-Y. and Zeger, S. (1995). Inference based on estimating functions in the presence of nuisance parameters. *Statistical Science*, 10:158–173.

Rodriguez-Iturbe, I., Cox, D., and Isham, V. (1988). A point process model for rainfall: further developments. *Proc. R. Soc. Lond.*, A417:283–298.

Wheater, H., Isham, V., Onof, C., Chandler, R., Northrop, P., Guiblin, P., Bate, S., Cox, D., and Koutsoyiannis, D. (2000). Generation of spatially consistent rainfall data. Report to the Ministry of Agriculture, Fisheries and Food (2 volumes). Also available as Research Report no. 204, Department of Statistical Science, University College London (`http://www.ucl.ac.uk/Stats/research/abstracts.html`).