Estimating functions and the generalized method of moments

Joao Jesus and Richard E. Chandler*

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

Estimating functions provide a very general framework for statistical inference, and are particularly useful when one is either unable or unwilling to specify a likelihood function. This paper aims to provide an accessible review of estimating function theory that has potential for application to the analysis and modelling of a wide range of complex systems. Assumptions are given in terms that can be checked relatively easily in practice, and some of the more technical derivations are relegated to an online supplement for clarity of exposition. The special case of the generalized method of moments is considered in some detail. The main points are illustrated by considering the problem of inference for a class of stochastic rainfall models based on point processes, with simulations used to demonstrate the performance of the methods.

Keywords: estimating equation; M-estimation; calibration; model selection; uncertainty

1. INTRODUCTION

This paper deals with the generic situation in which one wishes to carry out parametric inference on the basis of a vector $y$ of data, regarded as the realized value of a corresponding vector $Y$ of random variables. By ‘parametric inference’, we mean that interest lies in learning about a finite vector $\theta = (\theta_1 \ldots \theta_p)'$ of $p$ unknown quantities that are known to lie within some subset, $\Theta$, say, of $\mathbb{R}^p$. In a conventional statistical setting, the elements of $\theta$ would be parameters in a statistical model that specifies the joint distribution from which the data are considered to be generated: hence, we will refer to $\Theta$ as the parameter space. More generally, however, $\theta$ may contain just relevant descriptors (formally, functionals) of this joint distribution such as means, variances or specific probabilities of interest (such as the probability that an individual becomes infected during an epidemic). More generally still, one might wish to calibrate or ‘tune’ a model that is not explicitly probabilistic.

In settings involving a parametric statistical model that specifies the full joint distribution, $f(\cdot; \theta)$ say, of $Y$, the likelihood function forms the basis of most modern statistical techniques for inference. Given data $y$, the likelihood function is defined as

$$L(\theta; y) \propto f(y; \theta),$$

and plays a central role in Bayesian inference as well as in maximum likelihood and its variants. The likelihood can be regarded as measuring the relative fidelity of the model to the data under different parameter values.

However, particularly when studying complex systems, one may be unable or unwilling to formulate a likelihood function. For example, in non-Gaussian settings, there is a relative scarcity of tractable multivariate distributions; if there are dependencies among the elements of $Y$ therefore, one may be unable to specify a joint density that is both plausible and tractable.

Another feature of complex systems is that the available data often do not correspond directly with what is being modelled. A common scenario is that the fundamental mechanisms of the process under consideration operate at fine temporal and spatial scales, but only coarse-scale data are available. For example, many stochastic epidemic models attempt to represent the process of disease transmission between individuals but outbreak data, such as those from daily or monthly hospital reports, are typically aggregated both spatially and temporally. Likelihood construction for such models is complicated without knowing the exact times of individuals becoming infected (e.g. [1–3]).

Many stochastic models attempt to provide a simplified probabilistic representation of system structure and dynamics. Although potentially useful for obtaining insights into the system behaviour, a fully likelihood-based analysis of such simplified models may create problems by encouraging fidelity to features of the data that the model was never supposed to reproduce. An example of this type of problem is given in §2.

A final situation in which likelihood inference is infeasible is where the model is not probabilistic. This is the case, for example, with many deterministic models of system dynamics in application areas such as epidemic modelling [4]. Such models are, however,
routinely calibrated by minimizing some measure of discrepancy between model outputs and observations and, unless they provide a perfect fit to the observations, one is often interested in making some statements about model parameter uncertainty. In such settings, it is still conceptually possible to consider the data as realized values of random variables; any discrepancy between model and data can be described probabilistically if one is prepared to accept, for example, that two sets of observations, gathered under conditions that are identical in terms of the model inputs, will in general be different. In some sense, if the model at hand provides some information about the data $y$, then we can think of it as partially specifying some aspects of an underlying distribution. Clearly, however, likelihood-based analysis is not possible in such settings because this underlying distribution is not fully specified.

It is possible to develop specific techniques to address each of the problems outlined above; indeed, some of the cited references give explicit suggestions for doing so. As an alternative, however, in this paper we provide a review of a general theory of estimation that is applicable whenever a model is calibrated by solving an equation of the form

$$g(\theta; y) = 0.$$  

(1.1)

Here $g(\cdot; \cdot)$ is a vector-valued function of the unknown $\theta$ and data $y$, containing $p$ elements (recall that $p$ is the number of elements in $\theta$). In this context, the function $g(\cdot; \cdot)$ is called an \emph{estimating function} (EF), and an equation of the form (1.1) is an \emph{estimating equation}. EFs provide a unifying framework for many statistical techniques—for example, $g(\theta; y)$ may be the gradient of a log-likelihood or error sum of squares—but their applicability extends well beyond this.

\textbf{Example.} To help fix ideas, here and subsequently, suppose that $p = 1$ so that $\theta = \theta = \theta$ is scalar-valued. Suppose in addition that there are $n$ data items in total so that we can write $y = (y_1 \ldots y_n)\!^\top$; and consider the equation

$$\sum_{i=1}^n (y_i - \theta) = 0.$$  

(1.2)

This has the form of equation (1.1), with $g(\theta; y) = \sum_{i=1}^n (y_i - \theta)$. The solution to equation (1.2) is clearly $\theta = n^{-1} \sum_{i=1}^n y_i$: this is just the sample mean, which is often computed as an estimator of the population mean in a statistical model where the data are regarded as drawn independently from a common probability distribution. Thus, the sample mean can be regarded as the root of an estimating equation.

Much of the material presented here is perhaps well-known as ‘folklore’ within the statistical community. However, we are not aware of a source that collates and presents all of it in generally applicable terms. The extensive statistical literature on EFs is focused primarily on optimality in specific settings [5,6]. In parallel, a large body of econometrics literature, under the guise of the generalized method of moments (GMM) discussed in §4, deals mainly with regression-type models (see [7–9], ch. 9). Bera & Bilias [10] provide a good review of the historical development of, and connections between, several different inference techniques including EFs, GMM and likelihood methods. We aim here to provide an accessible summary of EF theory, with assumptions given in operational terms so that they can be checked relatively easily in applications. Our approach is similar in spirit to that of Bibby \textit{et al.} [11], who focus on inference for diffusion processes with a view towards applications in mathematical finance. There are also close connections with the method of indirect inference [12] (see [13] for a review of this area).

In §2, we discuss a class of stochastic models for rainfall time series that illustrate many of the issues discussed above. In §3, we review the general theory of EFs and in §4 we cover the important special case of the GMM: additional details for these sections can be found in the electronic supplementary material. In §5, we present the results of a simulation study to illustrate the performance of the theory in practice when applied to a simple stochastic rainfall model and §6 concludes.

\section{Motivating Example: Stochastic Rainfall Models}

Our use of EFs has been largely motivated by work with a class of stochastic models for rainfall time series. These models date back to Rodriguez-Iiturbe \textit{et al.} [14] and are widely used in the hydrological community to generate artificial rainfall time series, for purposes such as the assessment of flood risk and the impacts of climate change. A model of this type is used in the ‘weather generator’ provided as part of the latest suite of UK national climate change projections [15]. These models exemplify many of the issues summarized in §1. Here, we describe their structure, and highlight some of the difficulties that they pose for inference.

The models are based on a simplified conceptual representation of the physical structure of the rainfall process, in which rainfall is considered to be generated from ‘cells’ of activity in the atmosphere [16]. This conceptual representation is illustrated in figure 1. At the heart of such models is a point process, each event of which marks the arrival time of a cell. Each cell has a random duration, during which it deposits rain with a constant intensity (also random) so that its temporal profile is rectangular. At any point in time, the rainfall intensity is the sum of the intensities of all cells that are currently active.

In one of the simplest models of this type, the driving point process is Poisson and the cell durations are exponentially distributed random variables, realized independently and identically for each cell. Cell intensities have a two-parameter distribution and are again realized independently and identically for each cell; they are also independent of cell duration. This simple model has four parameters: the cell arrival rate ($\lambda$, say), mean cell duration ($\mu_d$) and the parameters of the cell intensity distribution which are conveniently
expressed via its mean ($\mu_X$) and coefficient of variation ($\sigma_X/\mu_X$). The model is stationary, and a variety of theoretical properties can be derived [14], such as the mean rainfall intensity (which is $\lambda \mu$, as might be expected), the autocorrelation function and the probability that no cells are active at a particular time instant (i.e. that it is not raining at that instant). In general, however, full distributional properties are hard to derive, even for this very simple model: this prohibits the construction of a likelihood for the model parameters, as it is possible to have an unrealistically high cell arrival rate and very short cells. This is an example of a model simplification that potentially renders likelihood-based analysis undesirable: maximum-likelihood parameter estimates might be expected to reflect the temporal resolution of the data rather than the underlying physical structures [19].

In view of the infeasibility and potential undesirability of likelihood-based inference for this class of rainfall models, parameters are usually chosen to achieve as close a match as possible, according to a weighted least-squares criterion, between the observed and fitted values of selected statistical properties [18]. Specifically, let $T = (T_1, \ldots, T_p)$ be a vector of $k \geq p$ summary statistics computed from data, and denote by $\tau(\theta) = (\tau_1(\theta), \ldots, \tau_k(\theta))'$ the expected value of $T$ under the model. Then, $\theta$ is estimated by minimizing an objective function of the form

$$S(\theta; T) = \sum_{i=1}^{k} w_i [T_i - \tau_i(\theta)]^2$$

for some collection of weights $\{w_1, \ldots, w_k\}$. This raises questions such as: how to select the properties $\{T_i\}$; how to choose the weights $\{w_i\}$; how to calculate standard errors and confidence regions for the model parameters; and how to discriminate between competing models. In the hydrological literature, many authors have sought to address some of these questions empirically (e.g. [18, 20]). However, EF theory provides direct answers to all of them.

3. ESTIMATING FUNCTIONS: REVIEW OF THEORY

Most standard treatments of EF theory start by examining the properties of the EFs themselves: e.g. [6, 21]. See Desmond [22] for a clear and readable justification of this approach. In applications, however, practitioners typically will be more interested in the properties of the corresponding estimator of $\theta$. For most EFs, only asymptotic properties of the estimators can be derived, although these can still provide useful approximations in finite samples. Here, therefore, we focus on the asymptotic theory for EF estimators.

Throughout, we consider the data as a vector of random variables. For an asymptotic treatment, it is necessary to consider the properties of EF estimators as the sample size (i.e. the length of the data vector $Y$) tends to infinity. It is convenient to extend the earlier notation to make explicit this dependence on sample size: a set of $n$ data items will be denoted by $Y_n$, and the corresponding EF by $g_n(\cdot)$, say. Then, for a sample of size $n$, the EF estimator ($\hat{\theta}_n$, say) is the root of the equation

$$g_n(\theta; Y_n) = 0.$$
This implicitly assumes that equation (3.1) has a unique root. In practice, if there are multiple roots, then a common pragmatic solution is to restrict the search to ‘plausible’ regions of parameter space so as to guarantee a unique root: this is operationally equivalent to working with an amended version of the EF that takes some arbitrary non-zero value elsewhere. By contrast, if there are no roots, then it may be possible to work with an amended EF for which a root exists: as discussed below, this often does not affect the asymptotic properties of the estimator and hence may be considered as a legitimate device for inference, at least insofar as the asymptotics provide useful approximations in finite samples. We therefore proceed under the assumption that equation (3.1) has at least one root.

3.1. Target of the estimation

Our first task is to establish whether an estimating equation of the form (3.1) has a target value. By this, we mean a value, \( \theta_0 \), say, to which the associated estimator \( \hat{\theta} \) converges as the sample size increases; this is sometimes called the object of inference. In settings involving a fully specified statistical model, one would hope that \( \hat{\theta}_n \) is the value of the parameter vector corresponding to the distribution from which the data were generated. However, as noted in §1, one might equally be interested in quantities that would not conventionally be regarded as statistical model parameters. The notion of a target value provides a way of handling this. We proceed under the following assumption:

**Assumption 3.1.** There exists a sequence \((\eta_n)\) of \( p \times p \) matrices, independent of \( \theta \) and such that as \( n \to \infty \), \( \eta_n g_n(\theta; Y_n) \) converges uniformly in probability to a non-random function, \( g(\theta) \), say, with the following properties:

1. The equation \( g(\theta) = 0 \) has a unique root at \( \theta = \theta_0 \).
2. \( g(\theta) \) is bounded away from zero except in the neighbourhood of \( \theta_0 \) (formally: for any \( \epsilon > 0 \), we can find some \( \delta > 0 \) such that \( |g(\theta)| > \delta \) whenever \( |\theta - \theta_0| > \epsilon \)).

Under this assumption, notice first that as \( g(\hat{\theta}_n; Y_n) = 0 \) by definition, and since \( \eta_n \) does not vary with \( \theta \), we must also have \( \eta_n g(\hat{\theta}_n; Y_n) = 0 \). Next, consider an arbitrary neighbourhood of \( \theta_0 \). As \( n \to \infty \), requirement (2) in assumption 3.1 excludes the possibility that infinitesimal wiggles in \( \eta_n g(\theta; Y_n) \), arising from sampling variation, can produce roots to the estimating equation outside this neighbourhood. As the neighbourhood can be arbitrarily small, \( \theta_0 \) is the only value of \( \theta \) that is guaranteed to lie within it. We have already limited the discussion to situations in which the estimating equation (3.1) has at least one root; any roots of the equation must therefore converge to \( \theta_0 \). This provides an informal justification of the following result:

**Result 3.2.** Under the conditions of assumption 3.1, as \( n \to \infty \), the EF defines a unique estimator \( \hat{\theta}_n \) that converges in probability to \( \theta_0 \).

In practice, it is often technically challenging to verify the uniform convergence of \( \eta_n g_n(\theta; Y_n) \) in assumption 3.1: this requires that, with probability tending to 1 as \( n \to \infty \), the discrepancy \( |\eta_n g_n(\theta; Y_n) - g(\theta)| \) becomes arbitrarily small simultaneously at all candidate values of \( \theta \). It is often straightforward to establish point-wise convergence (i.e. the discrepancy vanishes at each value of \( \theta \) considered individually), but this does not guarantee uniform convergence without additional assumptions. Moreover, it is difficult to legislate mathematically for all the possible ways in which uniform convergence may fail. Fortunately, however, such failure is usually associated with pathological situations that are unlikely to be encountered in applications: in practice, it can usually be ensured via fairly innocuous requirements, for example by imposing some conditions of smoothness on the EFs \( \{g_n(\theta; \cdot)\} \) (see [23, ch. 6] for examples of this in the context of maximum-likelihood estimation). Some other possibilities are indicated in the examples below; for a more thorough discussion, see [24, ch. 5].

**Example (continued).** In the present notation, the EF in equation (1.2) is \( g_n(\theta; Y_n) = \sum_{i=1}^n (Y_i - \theta) \). Suppose that the \( \{Y_i\} \) are independent random variables with a common mean \( \mu \) and finite variances; and consider the sequence

\[
n^{-1} g_n(\theta; Y_n) = n^{-1} \sum_{i=1}^n (Y_i - \theta) = \overline{Y}_n - \theta,
\]

where \( \overline{Y}_n \) is the sample mean of the observations. Now, as \( n \to \infty \) the weak law of large numbers guarantees that \( \overline{Y}_n \) converges to \( \mu \) in probability; and therefore \( n^{-1} g_n(\theta; Y_n) \) converges to \( g(\theta) = \mu - \theta \) (which is the expectation of the scaled EF: this type of situation is common). The convergence is uniform, as

\[
|n^{-1} g_n(\theta; Y_n) - g(\theta)| = |(\overline{Y}_n - \theta) - (\mu - \theta)| = |\overline{Y}_n - \mu|
\]

which does not depend on \( \theta \). Moreover, the function \( g(\theta) \) has a unique zero at \( \theta = \mu \) and is bounded away from zero elsewhere. Hence, the conditions of assumption 3.1 are satisfied by taking \( \eta_n = n^{-1} \); and we conclude that the target value of the estimator defined by equation (1.2) is \( \theta_0 = \mu \), the common mean of the \( \{Y_i\} \).

To illustrate the role of requirement (2) in assumption 3.1, consider now modifying the EF in this example to \( g_n(\theta; Y_n) = \exp(-\theta^2) \sum_{i=1}^n (Y_i - \theta) \). When multiplied by \( n^{-1} \) this modified EF converges uniformly to \( \hat{g}(\theta) = \exp(-\theta^2)(\mu - \theta) \), which again has a unique zero at \( \theta = \mu \) but also approaches zero as \( |\theta| \to \infty \). In this case, for any finite sample size there will be values of \( \theta \) for which \( |\hat{g}(\theta)| \) is small relative to the sampling variation in \( n^{-1} g_n(\theta; Y_n) \); hence, this modified EF will typically yield multiple (non-convergent) roots to the estimating equation tending to \( \pm \infty \) as the sample size increases, in addition to the ‘obvious’ target \( \mu \).

In this example, uniform convergence of \( \eta_n g_n(\theta; Y_n) \) is guaranteed owing to its being a continuous function of a single statistic \( \overline{Y}_n \), which itself converges in probability.
to \( \mu \). The same principle applies whenever the (normalized) EF is a continuous function of a finite vector of summary statistics, say \( T_n(Y_n) = (T_1(Y_n) \ldots T_k(Y_n)) \); and where \( T_n(Y_n) \) itself converges in probability to some limiting value. This is the case for an EF constructed by differentiating an objective function such as equation (2.1), for example.

The conditions in assumption 3.1 are sufficient to guarantee the existence of a unique target value for the estimation procedure, and can be verified reasonably intuitively in most applications. However, they are not necessary. For example, requirement (2) can be relaxed if the EF is obtained by differentiating some objective function (the corresponding estimator is sometimes called an M-estimator; see [24, ch. 5]). In this case, it may be that the normalized EF itself has multiple zeroes or that it approaches zero away from \( \theta_0 \); but that these zeroes can be eliminated from consideration by considering the corresponding values of the (similarly normalized) objective function.

As in the example above, usually the sequence \( \{\eta_n\} \) tends to the zero matrix as \( n \to \infty \). This gives some flexibility for modifying the EF without changing the target value, for example by adding some quantity \( \kappa_n(\theta; Y_n) \) such that \( \eta_n, \kappa_n(\theta; Y_n) \) converges uniformly in \( \theta \) to a vector of zeroes as \( n \to \infty \). This in turn allows us to work with modified EFs that have desirable properties in finite samples, for example so that at least one root to equation (3.1) is guaranteed or to reduce estimator bias [25,26].

### 3.2. Limiting distribution of the estimator

Having established that a particular EF has a unique target value \( \theta_0 \), it is of interest to characterize the uncertainty in the estimator \( \hat{\theta}_n \) by providing an approximate expression for its covariance matrix and if, for its joint distribution. To proceed with this, we need to consider the \( p \times p \) partial derivative matrices \( \{\partial g_n/\partial \theta \} \), say (which are assumed to exist). We make the following additional assumption:

**Assumption 3.3.** The matrices \( \{G_n(\theta)\} \) are invertible. Moreover, there exist sequences \( \{\gamma_n\} \) and \( \{\delta_n\} \) of invertible \( p \times p \) matrices that do not depend on \( \theta \) and are such that:

1. At \( \theta = \theta_0 \), the covariance matrix of the normalized EF \( \tilde{g}_n(\theta_0; Y_n) = \gamma_n g_n(\theta_0; Y_n) \) exists and converges to a limiting positive definite matrix, \( \Sigma \), say, as \( n \to \infty \).
2. Defining \( \tilde{G}_n(\theta) = \partial g_n/\partial \theta \), within a neighbourhood of \( \theta_0 \), the matrix \( \tilde{G}_n(\theta) \) converges uniformly in probability to an invertible matrix \( M(\theta) \) with elements that are continuous functions of \( \theta \).

The following result can now be stated (a sketch proof is provided in the electronic supplementary material):

**Result 3.4.** Under the conditions of assumptions 3.1 and 3.3, as \( n \to \infty \), the expected value of \( \theta_n \) converges to \( \theta_0 \) and the covariance matrix of \( \delta_n^{-1} \theta_n \) converges to \( C = M_0^{-1} \Sigma (M_0^{-1})' \), where \( M_0 = M(\theta_0) \).

If, in addition, the normalized EF \( \tilde{g}_n(\theta; Y_n) \) has a limiting multivariate normal (MVN) distribution, then the distribution of \( \delta_n^{-1} \theta_n \) is itself multivariate normal in the limit.

Notice that the exact covariance matrix of \( \tilde{g}_n(\theta_0; Y_n) \) (defined in the first requirement of assumption 3.3) can be written as \( \gamma_n \tilde{G}_n \), say, where \( \tilde{G}_n \) is the covariance matrix of the original EF \( g_n(\theta_0; Y_n) \). Moreover, if the expectation of \( G_n(\theta_0) \) exists and is equal to \( G_0 \), then clearly \( \tilde{G}_n \) in assumption 3.3 is just the limiting expectation of \( \gamma_n \tilde{G}_n \).

In this case, an operational interpretation of result 3.4 is as follows:

**Result 3.4 (operational version).** Let \( \gamma_n \) denote the covariance matrix of \( g_n(\theta_0; Y_n) \). Then, if assumptions 3.1 and 3.3 hold, and if \( G_0 = E[\tilde{g}_n(\theta_0; Y_n)] \) exists, the distribution of \( \theta_n \) is approximately MVN(\( \theta_0, G_0^{-1} \gamma_n \tilde{G}_n(\nu_n)^{-1} \)) in large samples.

The conditions of assumption 3.3 hold in a wide variety of applications and are easy to check. A continuation of the earlier example gives a flavour of their nature.

**Example (continued).** Suppose the \( \{Y_i\} \) are independent random variables with a common mean \( \mu \). In this case, \( \theta_0 = \mu \) is the target of the estimation procedure as demonstrated earlier. Suppose in addition that for all \( i \), \( \var{Y_i} = \sigma^2 \), such that \( n^{-1} \sum_{i=1}^{n} \sigma_i^2 \) converges to a constant value as \( n \to \infty \). Then, the EF \( g_n(\theta; Y_n) = \sum_{i=1}^{n} (Y_i - \theta) \) satisfies the conditions of assumption 3.3 with \( \gamma_n = n^{-1/2} \). For this case, we have \( G_n(\theta) = \partial g_n(\theta)/\partial \theta = -n \), meeting the differentiability and invertibility condition. Next,

\[
\var{\gamma_n g_n(\theta_0; Y_n)} = \left[n^{-1/2} \sum_{i=1}^{n} (Y_i - \theta)\right]^2
\]

which converges to a constant so that requirement (1) is satisfied. Finally, \( \tilde{G}_n(\theta) = \gamma_n G_n(\theta) \) \( \delta_n = n^{-1/2} \). \( n^{-1/2} \omega 1 = M(\theta) \), which fulfils requirement (2).

Recall that the root of the estimating equation in this example is \( \theta_n = Y_n \). In this case, therefore, result 3.4 states that \( n^{-1/2}(Y_n - \mu) \) has asymptotic mean zero and variance \( \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \sigma_i^2 \). Moreover, one can appeal to a central limit theorem [27, §2.5] to claim that \( n^{-1} \sum_{i=1}^{n} (Y_i - \mu) \) has a limiting normal distribution. In the limit as \( n \to \infty \) therefore, we have

\[
n^{-1/2}(\theta_n - \mu) \sim N(0, \lim_{n \to \infty} n^{-1} \sum_{i=1}^{n} \sigma_i^2) .
\]

The operational version of result 3.4 restates this by saying that in large samples, \( \theta_n \sim N(\mu, n^{-2} \sum_{i=1}^{n} \sigma_i^2) \) approximately.

For this particular example, the reasoning as presented above seems laboured: indeed, the application of result 3.4 here relies on the properties of \( Y_n \) so that the reasoning is arguably circular. The main purpose, however, is to illustrate the nature of the sequences \( \{\gamma_n\} \) and \( \{\delta_n\} \) in a simple example, and to show the steps needed to verify the requirements of assumption 3.3. The bottom line is to choose \( \gamma_n \) so as to normalize the variance of the EFs themselves; and then to choose \( \delta_n \) to
normalize the expectation of their derivatives. As here, it is often appropriate to take both sequences proportional to $n^{-1/2}$ (when $\Theta$ has $p > 1$ elements, the choice $\gamma_n = \delta_n = n^{-1/2}\mathbb{I}_{p^2}$ is often appropriate, where $\mathbb{I}_{p^2}$ is the $p \times p$ identity matrix). This is not always the case, however: particularly in situations involving a combination of stationary and non-stationary stochastic processes, a mixture of normalizations may be needed. A simple example of this occurs when regressing a time series $Y_n$ upon the time index: here, to stabilize the variance of the least-squares EF, the component corresponding to the regression slope must be normalized by a factor $n^{-3/2}$ (this can be deduced from the results given in [28, p. 72]).

As with assumption 3.1, the requirements of assumption 3.3 are not strictly necessary to obtain useful asymptotic results for EF estimators. For example, Sweeting [29] provides a generalization of result 3.4 without assuming the existence of expectations or covariance matrices. Another extension is to processes for which the $\theta_n$ converge in distribution to a random matrix $\Theta_n$, rather than to a fixed value: in this case, inference about $\theta_0$ can be considered as conditional upon the realized value of $\Theta_n$ [30]. This occurs, for example, when regressing a time series upon a random walk covariate: in this case, conditioning upon $\Theta_n$ is equivalent to conditioning on the observed covariate trajectory, which seems reasonable in applications (see [28, §5.2] for more discussion). Such extensions ensure, in effect, that result 3.4 is applicable in a very broad range of settings.

Result 3.4 provides means of comparing different EFs with the same target value, via their limiting covariance matrices. In particular, suppose that EFs $\mathbf{g}^{(A)}_n$ and $\mathbf{g}^{(B)}_n$ have the same target value and share the same normalizing sequence $(\delta_n)$ in assumption 3.3; and denote the limiting covariance matrices of the corresponding normalized estimators by $\mathbf{C}^{(A)}$ and $\mathbf{C}^{(B)}$. If the matrix $\mathbf{C}^{(B)} - \mathbf{C}^{(A)}$ is non-negative definite for all $\Theta \in \Theta$ then, in the limit, any linear transformation of $\Theta$ is estimated at least as precisely (in the sense of having no larger variance) when using $\mathbf{g}^{(B)}_n$ as when using $\mathbf{g}^{(A)}_n$. All other things being equal therefore, in this case $\mathbf{g}^{(A)}_n$ is preferable to $\mathbf{g}^{(B)}_n$. For example, it is known that in regular problems maximum-likelihood estimators (for which the EF is the derivative of the log-likelihood function, as noted in §1) are asymptotically optimal within the class of all EFs that have the same target value: in [23, §6.5; 31, ch. 8]. In settings where likelihood-based inference is infeasible or undesirable however, one might instead choose to work with a more restricted class of EFs: for example, all EFs defined by expressions of the form (2.1) with different values for the weights $\{w_l\}$. In this case, if $\mathbf{g}^{(A)}_n$ is a member of the chosen class and if the non-negative definite condition holds for any other choice $\mathbf{g}^{(B)}_n$ from the same class, then $\mathbf{g}^{(A)}_n$ can be thought of as (asymptotically) optimal in that class. The idea of finding the optimal EF within a class will be revisited in §4.

The proof of result 3.4 relies upon a Taylor expansion for the estimation error $\theta_n - \theta_0$ (see electronic supplementary material, equation S2). This expansion takes the form $\theta_n - \theta_0 = \Lambda_n \tilde{\theta}_n - \tilde{\theta}_0$, where $\Lambda_n$ is a matrix of partial derivatives of the EF and $\tilde{\theta}_n$ converges, upon normalizing, to $\tilde{\Theta}_n$, as defined in result 3.4. Of course, the EF could be multiplied by any non-singular matrix without changing the estimator itself. In particular, the estimator is unaffected if the EF is multiplied by $\mathbf{A}_n$, or by something asymptotically equivalent to it (such as its expectation, where this exists). Under this standardization, the estimation error and EF have the same limiting distribution. From a practitioner’s perspective, this can be seen as providing some justification for the use of optimality criteria based on the standardized EFs themselves, rather than upon the estimators that they define (see references given earlier).

### 3.3. Model comparison

It is often of interest to investigate whether the data are consistent with specific restrictions on $\Theta$, potentially involving several elements simultaneously, that have a relevant subject-matter interpretation. For example, in the context of the Poisson rainfall model considered in §2, one may wish to determine whether the parameter $\sigma^2 / \mu_1$ is equal to 1, as this might suggest that cell intensities could be modelled using an exponential distribution. More generally, in statistical practice, such restrictions often involve elements of $\Theta$ being set to zero to indicate that the corresponding model components or processes have no effect upon the data $\mathbf{Y}$. Other restrictions of interest might involve setting two elements of $\Theta$ equal to each other, or requiring that some subset of the elements should sum to zero. Any such restriction can be regarded as defining a new model that is a special case of the one initially envisaged: to determine whether the restriction is supported by the data is therefore to compare two competing models.

The types of restriction outlined above can all be defined via a set of $q \leq p$ linear constraints: $\Xi \Theta = \xi_0$, say, where $\Xi$ is a $q \times p$ matrix of rank $q$. Thus, to represent the restriction $\theta_j = 0$, we could write $\Xi = (1 \ 0 \ldots \ 0)$ and $\xi_0 = 0$; to represent the restriction $\theta_j = \theta_k$, we could write $\Xi = (1 - 1 \ldots 0)$ and $\xi_0 = 0$ (note that $\xi_0$ is a scalar in both of these cases as there is only one restriction). To represent a set of $q > 1$ such restrictions, simply stack the defining equations on top of each other.

Clearly, given an estimator $\tilde{\theta}_n$, of $\Theta$, a natural estimator of $\Xi \Theta$ is $\tilde{\xi}_n = \Xi \tilde{\Theta}_n$. Suppose, as in the operational version of result 3.2 above, that $\tilde{\Theta}_n$ has approximately a multivariate normal distribution with mean $\Theta_0$ and covariance matrix

$$\mathbf{G}^{-1}_0 \Sigma_n \mathbf{G}^{-1}_0' = \Gamma_n, \quad \text{say.} \quad (3.2)$$

Then, we have

$$\tilde{\xi}_n \sim \text{MVN}(\Xi \Theta_0; \Xi \Gamma_n \Xi) \quad (3.3)$$

approximately. Using standard properties of the multivariate normal distribution therefore, if $\Xi \Theta_0$ is indeed equal to $\xi_0$, then the quadratic form

$$(\tilde{\xi}_n - \xi_0)' \Xi \Gamma_n \Xi (\tilde{\xi}_n - \xi_0)' \quad (3.4)$$

has approximately a $\chi^2$ distribution with $q$ degrees of freedom; otherwise, it will tend to be larger. Thus, the hypothesis $H_0 : \Xi \Theta_0 = \xi_0$ can be tested at the $100(1 - \alpha)$% level by comparing (3.4) with the $100(1 - \alpha)$% point
of the $\chi^2$ distribution: if equation (3.4) exceeds this value, the hypothesis is rejected. This procedure is sometimes called a quasi-Wald test.

An alternative approach to testing restrictions is based on the limiting distribution of the EF itself; this is analogous to the use of score tests in a likelihood-based setting. If the EF is the gradient vector of some objective function, then the procedure is to optimize the objective function subject to the restrictions of interest, and then to calculate a test statistic as a quadratic form in the original EF, which is slightly unsatisfactory that the objective function is minimized to obtain the estimator $\hat{\theta}$, (situations involving maximization are easily handled in the same framework: just work with $-Q_{n}(\hat{\theta}; Y_{n})$ instead). In this case, denote by $Q_{0}(\theta; Y_{n})$ the value of $Q_{n}(\theta; Y_{n})$ under the restriction $\hat{\theta} = \xi_{0}$. By definition, we must have $Q_{0}(\xi_{0}; Y_{n}) \geq Q_{n}(\hat{\theta}; Y_{n})$. However, it seems reasonable to expect that if the restriction holds, then the difference $Q_{n}(\hat{\theta}; Y_{n}) - Q_{0}(\xi_{0}; Y_{n})$ will be small. The following result, which can be regarded as generalizing that of Kent [33] on the distribution of likelihood ratios in mis-specified models, makes this precise:

**Result 3.5.** Suppose that assumptions 3.1 and 3.3 hold and that $E[G_{0}(\theta_{0})] = G_{0}$ exists. Then, if $\Xi \theta_{0} = \xi_{0}$, the quantity $2\left[Q_{n}(\hat{\theta}; Y_{n}) - Q_{n}(\hat{\theta}; Y_{n})\right]$ has approximately the same distribution as $Z^{TA_{-}Z}$, where $Z \sim \text{MVN}(0, \Xi \Gamma_{n} \Xi')$ and $A = \Xi G_{0}^{-1} \Xi'$.

A derivation can be found in the electronic supplementary material.

In principle, result 3.5 yields a straightforward procedure for testing the hypothesis $H_{0}: \Xi \theta = \xi_{0}$: accept if

$$2\left[Q_{n}(\hat{\theta}; Y_{n}) - Q_{n}(\hat{\theta}; Y_{n})\right] \quad \text{is less than the appropriate percentile of the distribution of } Z^{TA_{-}Z} \text{ and reject otherwise. Equivalently, the procedure defines an objective function threshold below which candidate values of } \xi_{0} \text{ are accepted and above which they are rejected. The corresponding set of } \xi_{0}-values \text{ defines a confidence region for } \Xi \theta_{0} \quad \text{(of course, the quasi-Wald and quasi-score tests can also be used to generate confidence intervals in this way).}$$

Unfortunately, distributions of quadratic forms such as $Z^{TA_{-}Z}$ are intractable in general. Notice, however, that in the present context $G_{0}$ is the expectation of a Hessian matrix containing the second derivatives of $Q_{n}(.: Y_{n})$ at $\theta_{0}$: it is therefore symmetric. Thus, $A = \Xi G_{0}^{-1} \Xi'$ is also symmetric, as is $A^{-1}$. In this case, the distribution of $Z^{TA_{-}Z}$ can be approximated with a scaled and shifted $\chi^{2}$ distribution [34, p. 88]. Specifically, the distribution is approximated by that of $\alpha X + c$ where

$$X \sim \chi^{2}_{a}, \quad a = \frac{|\kappa_{1}|}{4\kappa_{2}}, \quad b = \frac{8\kappa_{2}^{2}}{\kappa_{3}^{2}} \quad \text{and}$$

$$c = \kappa_{1} - ab,$$

(3.6)

$$\kappa_{r} = 2^{-1}(r-1)!\text{tr}[(\Xi \Gamma_{n} \Xi')^{r}](r = 1, 2, 3), \quad \text{and tr(\cdot)} \text{ denotes the trace operator.}$$

An important special case of result 3.5 occurs when $q = 1$ so that $\xi = \xi_{0}$, $A = A$ and $Z = Z$ are all scalars. In this case, the quantity $Z^{TA_{-}Z}$ becomes $Z^{2}/A$, which is equal to $\text{var}(Z)/A \times \chi^{2}_{a}$, where $\chi^{2}_{a}$ has a $\chi^{2}$ distribution with one degree of freedom. Thus, denoting by $\chi_{a}^{2}(1 - \alpha)$ the $100(1 - \alpha)\%$ point of the $\chi^{2}_{a}$ distribution, a $100(1 - \alpha)\%$ confidence interval for $\xi_{0}$ can be constructed as the set of values for which the objective function can be made smaller than the threshold

$$Q_{n}(\theta_{0}; Y_{n}) + \frac{\text{var}(Z)}{2A} \chi_{a}^{2}(1 - \alpha).$$

(3.7)

In particular, if $\Xi$ is a $1 \times p$ matrix containing zeroes everywhere except for a 1 in position $i$ so that $\Xi \theta = \xi_{0}$ is the $i$th component of $\theta$, then $\text{var}(Z)$ is just the approximate variance of $\xi_{0}$, as given by result 3.4, and $A$ is the corresponding diagonal element of $G_{0}^{-1}$. Another special case is when $\Xi$ is the $p \times p$ identity matrix. Here, the result provides a means of calculating simultaneous confidence regions for all elements of $\theta$ based on the values of the objective function: this can be seen as a formal way to determine which values are ‘near-optimal’. Notice that in this case $Z \sim \text{MVN}(0, \Sigma)$ and $A = G_{0}^{-1}$ so that $Z^{TA_{-}Z} = Z^{2}G_{0}Z$.

Finally, note that if $G_{0} = \Sigma_{n}$ (i.e. the expected Hessian of the objective function is equal to the covariance matrix of its gradient vector), then $\Sigma_{n} = G_{0}^{-1}$ (from equation (3.2), coupled with the symmetry of $G_{0}$). In this case, we have $Z \sim \text{MVN}(0, \Xi G_{0}^{-1} \Xi') = \text{MVN}(0, A)$, and the distribution of $Z^{TA_{-}Z}$ is exactly $\chi^{2}_{a}$ with $q$ degrees of freedom. The equality $G_{0} = \Sigma_{n}$ is known as the second Bartlett identity and holds, for example, when $Q_{n}(.: Y_{n})$ is the negative log-likelihood for a correctly specified model for $Y_{n}$; this is the basis for the standard chi-squared reference distribution used in likelihood ratio tests.

For the confidence intervals in the $q = 1$ case above, it is slightly unsatisfactory that the objective function threshold (3.7) varies depending on the matrix $\Xi$: in particular, different thresholds are required for different elements of $\theta$. If the second Bartlett identity holds, however, all confidence intervals are defined via the single threshold $Q_{n}(\theta_{0}; Y_{n}) + \frac{1}{2} \chi_{a}^{2}(1 - \alpha)$: this seems much more intuitive. It also has computational advantages, as approximations such as equation (3.6) are not required. Sometimes therefore, it may be convenient to adjust the objective function to recover the second Bartlett identity. This also has theoretical benefits, as
described by Chandler & Bate [35] in the context of inference based on a mis-specified log-likelihood.

3.4. Estimation of $G_0$ and $\Sigma_n$

To use results 3.4 and 3.5 in practice, it is necessary to estimate $G_0$ and $\Sigma_n$ (or equivalently $M_0$ and $\Sigma$ as defined in result 3.4 and assumption 3.3 respectively). In fact, in large samples it suffices to use any estimator for which, with probability tending to 1 as $n \to \infty$, the estimation error is negligible compared with the quantity being estimated. One possibility is to plug the estimation error is negligible compared with the quantity being estimated. One possibility is to plug the estimate $\theta_0$ into analytical expressions for $G_0$ and $\Sigma_n$ if these are available. For $M_0$, another option is numerical differentiation of $g_n(\cdot; \cdot)$ at $\theta_0$.

In the absence of an analytical expression for $\Sigma_n$, often an empirical estimator can be used instead. For example, if $g_n(\cdot; \cdot)$ can be written as a sum of uncorrelated contributions from $m$ distinct groups of observations:

$$g_n(\theta; Y) = \sum_{k=1}^{m} g(k)^{(i)}(\theta; Y_k), \text{ say},$$

and if $m \to \infty$ as $n \to \infty$, then

$$\Sigma_n = E\left\{ \sum_{k=1}^{m} g(k)^{(i)}(\theta_0; Y_k)g(k)^{(i)}(\theta_0; Y_k)^T \right\}$$

can be estimated to the required accuracy as

$$\hat{\Sigma}_n = \sum_{k=1}^{m} g(k)^{(i)}(\theta_0; Y_k)g(k)^{(i)}(\theta_0; Y_k)^T. \quad (3.8)$$

A potential difficulty here is that the $p(p-1)/2$ distinct elements of $\Sigma_n$ are estimated individually. Unless $p$ is rather small, in finite samples the associated estimation errors can accumulate to degrade many of the approximations upon which result 3.4 is based. It has been noted, for example, that estimators of the form (3.8) can lead to liberal tests of hypotheses [36], and in such situations it can pay to think more carefully about how to estimate $\Sigma_n$. We see an example of this in §5.

4. THE GENERALIZED METHOD OF MOMENTS

Given data $Y_n$, one way to construct an estimating function is to consider a vector $T_n = T_n(Y_n)$ of $k \geq p$ summary statistics with expectation $\tau(\theta) = E[T_n]$ under the model, so that

$$0 = E[T_n - \tau(\theta)] = E[h_n(\theta; Y_n)], \text{ say.} \quad (4.1)$$

This suggests that $\theta$ could be estimated by equating $\tau(\theta)$ with $T_n$. However, sampling variation in $T_n$ will usually render this infeasible, particularly when $k > p$. As an alternative therefore, we could minimize an expression of the form

$$Q_n^{\theta}(\theta; Y_n) = \left[ h_n(\theta; Y_n)^T W_n h_n(\theta; Y_n) \right]. \quad (4.2)$$

where $W_n$ is a $k \times k$ matrix which may depend on the data but which converges in probability to a positive definite matrix $W$, say. Notice that the expression (2.1), used to fit the stochastic rainfall models reviewed in §2, can be written in the form (4.2); in this case, $W = \text{diag}(w_1 w_2 \ldots w_k)$ is a diagonal matrix.

In fact, for a general treatment, equation (4.2) must be modified slightly. We will assume the existence of a sequence $(\gamma_n)$ of $k \times k$ matrices that do not depend on $\theta$ and are such that, at $\theta = \theta_0$, the covariance matrix of $h_n(\theta; Y_n) = \gamma_n h_n(\theta; Y_n)$ tends to a limiting positive definite matrix $S$. Under this assumption, instead of minimizing equation (4.2) we will minimize

$$Q_n(\theta; Y_n) = \left[ h_n(\theta; Y_n)^T W_n h_n(\theta; Y_n) \right]. \quad (4.3)$$

The rationale for this is given in the electronic supplementary material, section S4.2. Clearly, equations (4.2) and (4.3) will deliver identical results if $\gamma_n$ is proportional to the $k \times k$ identity matrix: $\gamma_n = \gamma_n I_{k \times k}$ say. This is the case in many applications (see §5). However, the variances of different components of $T_n$ may change at different rates with $n$; in this case, the theory below applies only to estimators derived from equation (4.3).

This approach to inference was formalized by Hansen [8] and is referred to as the generalized method of moments. The closely related area of indirect inference (see §1) also involves optimizing an expression of the form (4.3); here however, the quantity $h_n(\theta; Y_n)$ is itself derived from an approximating model. In this section, we examine GMM estimation from an estimating functions perspective, and also discuss the choice of matrix $W_n$.

4.1. The generalized method of moments estimating functions

Differentiating equation (4.3) with respect to $\theta$ yields, apart from an irrelevant factor of 2, the EF

$$g_n(\theta; Y_n) = \hat{H}_n^{\theta}(\theta) W_n \hat{h}_n(\theta; Y_n). \quad (4.4)$$

where $\hat{H}_n^{\theta}(\theta) = \partial \hat{h}_n(\theta)/\partial \theta = -\gamma_n \partial \tau/\partial \theta$ (the existence of this $k \times p$ matrix is assumed). To translate the theory of §3 to the GMM context therefore, it is necessary to ensure that assumptions 3.1 and 3.3 hold for equation (4.4). For operational purposes, the requirements are most helpfully stated in terms of the summary statistics $T_n$ and their theoretical counterparts $\tau(\theta)$. In fact, as noted in the earlier discussion of result 3.2, the conditions of assumption 3.1 are trivially satisfied if the elements of $T_n$ converge to $\tau(\theta_0)$, and if $\tau(\theta)$ is bounded away from this limiting value outside the neighbourhood of $\theta_0$.

To establish the conditions for assumption 3.3, the first task is to ensure the existence and invertibility of the derivative matrix $G_n(\theta) = \partial g_n/\partial \theta$. In the electronic supplementary material, it is shown that $G_n(\theta)$ exists if $\tau(\theta)$ is twice differentiable with respect to $\theta$. $G_n(\theta)$ will then be invertible if it is of full rank. This is the case if the functions $\left\{ \tau_{ij}(\theta); i = 1, \ldots, k \right\}$ are linearly independent (this can be seen by inspection of equations (S9) and (S10) in the electronic supplementary material, together with equation (4.1)). In practice, this means that $T_n$ should not contain redundant elements (i.e. elements that add nothing to the information contained elsewhere): this is not a serious restriction.
To progress further, we need to define $\Omega^{(r)}_m = \partial^2 \tilde{h}_m / \partial \theta \partial \theta^\prime$, where $\tilde{h}_m$ is the rth element of $\tilde{h}_m(\theta; Y_n)$ ($r = 1, \ldots, k$). With this definition, the requirements of assumption 3.3 in the GMM context can now be stated equivalently as follows:

**Assumption 4.1.** There exists a sequence $(\tilde{\sigma}_n)$ of invertible $p \times p$ matrices that do not depend on $\theta$ and are such that:

1. $\lim_{n \to \infty} \tilde{\sigma}_n = 0$, the zero matrix;
2. Within a neighbourhood of $\theta_0$, for all $r \in \{1, \ldots, k\}$ the matrices $\tilde{H}_n(\theta) \tilde{\sigma}_n$ and $\Omega^{(r)}(\theta) \tilde{\sigma}_n$ converge to limiting matrices $H(\theta)$ and $\Omega^{(r)}(\theta)$, respectively, all elements of which are continuous functions of $\theta$ and where $H(\theta)$ is invertible.

The application in §5 provides some insights into the nature of assumption 4.1. However, the continuity requirements are not strong. If $H(\theta)$ exists, then its continuity is guaranteed by the earlier assumption that $\tilde{h}_m(\theta; Y_n)$ is twice differentiable; and continuity of $\Omega^{(r)}(\theta)$ corresponds to that of the second derivatives of $\tau(\theta)$.

The GMM version of result 3.4 can now be stated as follows (for a proof, see the electronic supplementary material):

**Result 4.2.** Assuming the existence of a sequence $(\gamma_n)$ such that $\tilde{h}_m(\theta; Y_n)$ has a limiting covariance matrix $S$ as described earlier, let $\theta_0$ be the value of $\theta$ minimizing $Q_n(\theta; Y_n)$ in equation (4.3). Then, under assumption 4.1, the expected value of $\theta$ converges to $\theta_0$ and the covariance matrix of $\tilde{\sigma}_n^{-1} \tilde{\theta}_n$ converges to $C = [M(\theta_0)]^{-1}\Sigma [M(\theta_0)]^{-1}$, where

$$\Sigma = H(\theta_0)^T W_n S W_n^T H(\theta_0) \quad (4.5)$$

and

$$M(\theta) = H(\theta)^T W \Sigma^{-1} H(\theta_0) \quad (4.6)$$

Furthermore, if $T_n$ has a limiting multivariate normal distribution after standardization, the limiting distribution of $\tilde{\sigma}_n^{-1} \tilde{\theta}_n$ is itself multivariate normal.

As in the general EF setting, to use this result in practice the quantities $\Sigma$ and $M(\theta_0)$ must be estimated. As suggested in §3, this can be performed via plug-in or empirical estimation based on the EF (4.4). However, for the GMM, it is often simpler (and, in our experience, computationally more stable) to work directly with $T_n$ and $\tau(\theta)$, and then to use equations (4.5) and (4.6) to derive estimates of $\Sigma$ and $M(\theta_0)$. Specifically, $H(\theta_0)$ can be estimated as $-\gamma_n \partial \tau / \partial \theta^{\prime \prime}\tilde{\sigma}_n^{\prime \prime}$ and $S$ can be estimated via an empirical or plug-in estimator of the covariance matrix of $T_n$. Once again however, some care is required with empirical covariance matrix estimators: indeed, there are now $k(k-1)/2$ separate elements of $S$ to estimate rather than $p(p-1)/2$ elements of $\Sigma$. This can lead to poor performance if $T_n$ contains too many elements (e.g. [37]). This is another argument for careful choice of $T_n$: ideally, it will contain a small number of summary statistics that contain most of the information about $\theta$.

### 4.2. Optimal weighting matrices

Until now, we have not considered in detail the choice of weighting matrix $W_n$ in equation (4.3). However, it is clear that this choice will affect the properties of the estimator, and in particular its covariance matrix. In §3, the concept of optimality within a class of EFs was introduced. In the present context, this leads to the question of whether, given a specific set of summary statistics $T_n$, there is an optimal weighting matrix $W_n$ in equation (4.3). Recalling that $S$ is the limiting covariance matrix of $\tilde{h}(\theta_0; Y_n) = \gamma_n [T_n - \tau(\theta_0)]$, the answer to this question is given by the following:

**Result 4.3.** Under the conditions set out above, the smallest attainable limiting covariance matrix of $\tilde{\sigma}_n^{-1} \tilde{\theta}_n$ is $C_{\text{inv}} = [H(\theta_0) S^{-1} H(\theta_0)]^{-1}$, and is achieved by setting $W_n = S^{-1}$ in equation (4.3).

The proof follows exactly the steps given in [38, §3.6], and is therefore omitted.

As well as its theoretical optimality, the choice $W_n = S^{-1}$ is appealing from another perspective. Specifically, using the results in §S4 of the electronic supplementary material it may be shown that with this choice, asymptotically the objective function (4.3) satisfies the second Bartlett identity. Hypotheses involving linear restrictions on $\theta$ can therefore be tested by referring test statistics of the form (3.5) directly to the quantiles of the appropriate chi-squared distribution.

### 4.3. Extensions

In the discussion above, the matrices $\tilde{H}_n(\theta) \tilde{\sigma}_n$ and $\Omega^{(r)}(\theta)$ contained derivatives of $\tau(\theta)$ and, in particular, were non-random. This is because of the form (4.1) of the quantities $h_m(\theta; Y_n)$ used to construct the GMM objective function (4.3). More generally, however, the GMM idea can be applied with any choice of $h_m(\theta; Y_n)$ for which $E[h_m(\theta; Y_n)] = 0$. For most choices of $h_m(\theta; Y_n)$, $\tilde{h}_m(\theta) \tilde{\sigma}_n$ and $\Omega^{(r)}(\theta)$ will be random matrices: in this case, the theory outlined above can be adapted by requiring that $\tilde{H}_n(\theta) \tilde{\sigma}_n$ and $\Omega^{(r)}(\theta)$ converge uniformly in probability to $H(\theta)$ and $\Omega^{(r)}(\theta)$, respectively. See the study of Lindsay & Qu [30] for an accessible overview of this area, and its links to many other inference methods.

### 5. THEORY IN PRACTICE: A SIMULATION STUDY

We now illustrate the theory outlined above, using a simulation study involving the four-parameter Poisson rainfall model from §2. As described previously, such models are often fitted by minimizing an expression of the form (2.1); this is a special case of the GMM in which the weighting matrix $W_n$ is diagonal. Typically, the minimization must be carried out numerically; our experience is that this can be challenging, and that working with the logarithms of the model parameters can stabilize the procedure. Accordingly, for the purposes of this study the model parameter vector is taken as $\theta = (\log A \log \mu_X \log \sigma_X^2 / \mu_X \log \mu_Y')^T$. The computations are all carried out in the R environment [40] using routines that are available.
5.1. Simulation set-up

To account for seasonality in rainfall data, these models are usually fitted separately to data for each calendar month [18,19]: data from different years are considered to provide independent replicates of the rainfall process. Each of our simulations represents 20 years’ worth of data for a single calendar month: specifically, \( m = 20 \) independent sets of 30 days’ worth of hourly values are generated, each from a model with parameter vector \( \theta_0 = (-3.5\ 0\ 0\ 1.1)' \). These values were obtained by rounding estimates obtained by fitting the model to January data from Birmingham airport in the UK for the period 1950–1997 (with time measured in hours and in mm h\(^{-1}\)), and can be considered typical of January rainfall in the UK. The 20 year record length is towards the low end of what might be considered desirable in the hydrological literature; however, long hourly rainfall records are relatively scarce so that practitioners are often forced to use shorter records such as those simulated here.

We have generated 1000 simulated datasets with the structure described above. For each dataset, GMM estimates have been computed from a ‘typical’ set of statistics, we have var\((\mathbf{T})\) = \( S/m \), where \( S \) is the covariance matrix of a single set of summary statistics. Thus, the sequence (\( \gamma_n \)) defined in 4.1 can be taken as \( m^{1/2}\lambda_{\text{asy}} \). With this normalization, equations (4.2) and (4.3) yield identical results (see the discussion following equation (4.3)). Moreover, setting \( \delta_n = m^{-1/2}\lambda_{\text{asy}} \) (which tends to the zero matrix as \( m \to \infty \)), as required by assumption 4.1) ensures that the matrices \( H(\theta_0)\delta_n \) and \( \mathbf{D}(\theta_0)\delta_n \) converge to the appropriate derivatives of \( \mathbf{T}(\theta_0) \). It follows that \( \mathbf{H}(\theta) \) in equations (4.5) and (4.6) is equal to \( \partial\mathbf{T}/\partial \theta \), and that \( \mathbf{S} \) in equation (4.5) is equal to \( m/2\text{var}(\mathbf{T}) \). We estimate \( \mathbf{H}(\theta_0) \) by numerical differentiation of \( \mathbf{T}(\theta) \) at \( \theta_0 \), and estimate \( \mathbf{S} \) as \( m/2\text{var}(\mathbf{T}) \). We note finally that \( \mathbf{T} \) might be expected to have approximately a multivariate normal distribution owing to central limit-type results for the individual replicates \( \{\mathbf{T}^{(i)}\} \), as well as to the averaging of these replicates. Thus, result 4.2 should hold in its entirety. The simulation exercise therefore provides an opportunity to determine whether the asymptotics provide a reasonable approximation when \( m = 20 \) (i.e. when 20 years’ worth of data are available).

As well as examining the validity of the asymptotics, we compare different choices of weighting matrix \( \mathbf{W} \) in equation (4.3) (for notational convenience, we drop the \( n \) subscript here). Result 4.3 gives the optimal choice as \( \mathbf{S}^{-1} \). Apart from an irrelevant constant of proportionality, this can be estimated as \( \mathbf{S}^{-1} = \mathbf{W}_6 \), say, with \( \text{var}(\mathbf{T}) \) defined at equation (5.1). To investigate the benefits that accrue from the use of \( \mathbf{W}_6 \) for each simulated dataset we also compute estimates using the following alternatives:

\[
\mathbf{W}_1 = \mathbf{I}_m; \text{ equal weights.}
\]

\[
\mathbf{W}_2; \text{ unequal weights. This is the same as } \mathbf{W}_1, \text{ except that the entries corresponding to the hourly summary statistics are set to 100. This has been suggested as a means of ensuring that the fitted models reproduce the most hydrologically important properties of observed rainfall sequences [18].}
\]

\[
\mathbf{W}_3; \text{ inverse variances. This matrix is again diagonal, but the entries are now the inverses of the corresponding diagonal elements of } \text{var}(\mathbf{T}). \text{ This can be regarded as a crude approximation to the ‘optimal’ weighting matrix } \mathbf{W}_6 \text{ in which correlations between the summary statistics are ignored.}
\]

5.2. Performance measurement criteria

A simple way to compare the performance of the different weighting matrices is to estimate the biases and covariance matrices of the corresponding estimators. With 1000 simulated datasets, these quantities can be
estimated directly using the sample mean and covariance matrix of each estimator. We refer to the resulting covariance matrix estimates as ‘empirical’.

Another question of interest is whether the asymptotic covariance matrix given in result 4.2 can be regarded as an adequate approximation to the true covariance matrix. To assess this, for each dataset and weighting matrix we estimate the covariance matrix of the estimator by substituting estimates of $H(\theta_0)$ and $S$ (as described above) into result 4.2 via equations (4.5) and (4.6); and for each weighting matrix we calculate the sample mean of these estimated covariance matrices as a summary measure. We refer to the results as ‘theoretical’ covariance matrices: one way to check the asymptotic theory is to compare these with the empirical versions described above, for example, by extracting the standard errors corresponding to each parameter and comparing these. It is also of interest, of course, to compare the standard errors between different weighting schemes.

Finally, we assess the accuracy of confidence intervals and regions computed using the asymptotic theory. Specifically, the limiting multivariate normal distribution from result 4.2 provides a means of calculating individual confidence intervals for each parameter in the model: if the theory holds in practice, then a $100(1 - \alpha)$% confidence interval should contain the true parameter value roughly $100(1 - \alpha)$% of the time. This can be assessed by computing confidence intervals for each simulated dataset, and recording the proportion of these that contain the true value. Similarly, confidence regions for the entire parameter vector can be computed based on objective function thresholds as in result 3.5. To check the validity of such regions, for each dataset we estimate the threshold by plugging in estimates of the required quantities as usual; and then evaluate the objective function at the true parameter value. If the asymptotic theory is valid, the result should be less than the computed threshold roughly $100(1 - \alpha)$% of the time.

5.3. Results

Figure 2 shows, for each choice of weighting matrix, the distributions of estimation error obtained from the 1000 simulated datasets. The variance-based weighting schemes, $W_3$ and $W_0$, perform similarly and deliver much less variable estimators than the fixed schemes $W_1$ and $W_2$, especially for the parameter log $m_L$. Both of the fixed schemes produce several outlying estimates (this is especially true for $W_1$ when estimating log $\sigma_x/\mu$, which suggests that inference based on asymptotic normality of the estimators may not be suitable for these weighting schemes. Interestingly, the use of fixed weighting schemes to fit this type of model is standard in the hydrological literature, and numerical optimization of the GMM objective function frequently must be constrained to ensure that parameter estimates are physically realistic (e.g. [20]). The results of figure 2, together with similar results for other models (not shown here), suggest that the use of variance-based weighting schemes may alleviate this problem to a large extent.

Another conclusion from figure 2 is that, for all weighting matrices, any bias in the estimators is negligible when compared with the sampling variability. This is reassuring, and suggests that any bias in the summary statistics themselves is relatively unimportant for the purposes of parameter estimation for this particular model.
We next consider the variances of the estimators. Table 1 shows the parameter standard errors for each weighting scheme, computed from both the empirically and theoretically estimated covariance matrices as described above. The differences between the empirical standard errors for the different weighting schemes merely reflect the differences in variability noted in figure 2: in particular, there seems little to choose between weighting scheme $W_3$ and the optimal scheme $W_0$. Of more interest is to compare the theoretical standard errors with the empirical ones. In general, the theoretical standard errors in table 1 are lower than those observed: exceptions are for $\log \sigma X/\mu X$ with weighting schemes $W_1$ and $W_2$ (for which the accuracy of the computed theoretical covariance matrices was often suspect owing to ill-conditioned matrix calculations); and the results for weighting scheme $W_3$ where the two sets of standard errors are more or less in agreement. Overall, these results suggest that the asymptotic theory tends to underestimate variability and to make the estimators look more precise than they actually are.

The discrepancies between the empirical and theoretical standard errors suggest that confidence intervals and regions, constructed using the theoretical covariance matrices, will be inaccurate. This is indeed the case: table 2 shows the proportion of simulations for which confidence intervals and regions, constructed using the asymptotic theory, included the true parameter value. In general, the coverages are well below their nominal levels. This is particularly true for the confidence regions for $\theta$ constructed under the optimal weighting scheme $W_0$. The results for scheme $W_3$ seem rather better, although the coverages are still slightly too low.

Clearly, the results in tables 1 and 2 suggest that—as implemented here, at least—the asymptotic theory yields unreliable assessments of estimator variability for this sample size. The problems are at least partly associated with poorly estimated covariance matrices, as indicated by the comparison of empirical and theoretical standard errors in table 1. For the estimators based on $W_1$ and $W_2$, it is likely that failure of the normal approximation also contributes. This failure can be seen from the box plots in figure 2. It has also been verified using quantile–quantile plots, which however show that the distribution of estimators from $W_3$ and $W_0$ is indeed very close to normal.

Overall, the analyses in this section suggest that the variance-dependent weighting schemes $W_3$ and $W_0$ offer a substantial improvement in estimator...
performance over the fixed schemes $\mathbf{W}_0$ and $\mathbf{W}_i$. We have obtained similar results in other simulation settings. This seems like a strong argument for using these variance-dependent schemes. It is, however, of interest to understand in more detail the reasons for the poor performance of the asymptotic theory in the assessment of estimator variability, and if possible to find practically feasible ways of improving this performance. We address this next.

### 5.4. Improved inference

In our simulations, we have seen that the theoretical covariance matrix calculations tend to underestimate the real variability of the estimators. This must be associated either with the estimation of $\mathbf{H}(\theta_0)$ and $\mathbf{S}$, or with a failure of the asymptotic theory to provide reasonable approximations on the basis of 20 replicates of 30 days’ data. An extensive set of analyses (not reported here) has been undertaken to identify the cause of the problem: the conclusion was that it is primarily associated with inaccurate estimation of $\mathbf{S}$, the (normalized) covariance matrix of the vector $\mathbf{T}$ of summary statistics.

This conclusion is perhaps unsurprising, given the known problems associated with empirical covariance matrix estimation as discussed previously: here, $\mathbf{S}$ is an $8 \times 8$ symmetric matrix to be estimated from $m = 20$ independent observation vectors. One way to avoid the problem is to obtain an initial estimate of $\theta$ using a sub-optimal weighting matrix, and then to plug this initial estimate into an analytical expression for $\mathbf{S}$. This effectively reduces the degrees of freedom of the covariance matrix estimator from $k(k-1)/2$ ($= 28$ in our application) to $p (=4)$. As a result, we might expect that the use of a plug-in estimator will reduce the cumulative effect of sampling variation in the estimation of $\mathbf{S}$.

In the present context however, and for many other problems involving complex system structures, analytical expressions for $\mathbf{S}$ are not available. For example, the variance of second-order sample properties depends on fourth-order theoretical properties of the aggregated rainfall process: even for the relatively simple model considered here, the amount of work involved to derive these properties is prohibitive. We therefore propose an alternative that can be used whenever analytical evaluation of $\mathbf{S}$ is impractical, but where simulation is possible. Specifically, we suggest that an initial estimate of $\theta$ is used to simulate a large artificial dataset from which an improved estimate of $\mathbf{S}$ can be obtained. Of course, apart from the simulation error (which can be reduced by increasing the size of the artificial dataset), this is equivalent to parametrizing the covariance matrix as a function of $\theta$. In a second stage, $\theta$ can be re-estimated using the optimal weights based on this improved estimate of $\mathbf{S}$.

We have applied this idea to our simulations. For each of our 1000 datasets, the initial estimate of $\theta$ was obtained using weighting scheme $\mathbf{W}_3$, which was shown in the previous section to be a reasonable estimator. One thousand independent 30 day replicates of the process were then simulated using this pilot estimate; and these replicates were used to construct a new estimate of $\mathbf{S}$. Revised estimates of $\theta$ were then obtained using weighting schemes $\mathbf{W}_i$ and $\mathbf{W}_0$ from the previous section, but with this new estimate of $\mathbf{S}$ substituted for the original, empirical estimate. To discriminate between the different procedures, we denote the resulting weighting matrices by $\mathbf{W}_i$ and $\mathbf{W}_0$, respectively.

The empirical and theoretical standard errors for this two-step procedure are compared in the bottom two rows of table 1. The results for $\mathbf{W}_i$ are very similar to those of the original version $\mathbf{W}_0$. However, there is now much better agreement between the empirical and theoretical versions for the optimal weighting scheme $\mathbf{W}_0$. Moreover, of all the schemes considered, this one now delivers the smallest empirical standard errors for all parameters except $\log(\mathbf{W}_i)$ (where $\mathbf{W}_i$ yields a slightly smaller value, although the difference is easily within sampling variation).

The bottom two rows in table 2 show an improvement in the coverages obtained from both $\mathbf{W}_3$ and $\mathbf{W}_0$, compared with those from $\mathbf{W}_i$ and $\mathbf{W}_0$; the improvement for $\mathbf{W}_0$ is particularly striking. Coupled with the smaller standard errors of the estimators, this makes the two-step optimal procedure potentially attractive for carrying out inference using this model with relatively modest sample sizes. The main drawback is the computing time which, owing to the need for a large simulation to estimate $\mathbf{S}$, can be several times greater than a one-step procedure using, say, $\mathbf{W}_3$.

It may be argued that, even with the use of the two-step procedure, the coverages from $\mathbf{W}_0$ in table 2 still tend to be less than their nominal values. However, the difficulty of the inference task on the basis of a relatively small sample of data should not be underestimated. We have conducted additional simulation experiments with more complex and realistic models, with results similar to those obtained above. We also find that increasing the number of replicates improves the coverages still further, as might be expected. The take-home message for the prudent analyst is that the asymptotic theory can deliver reasonable, but approximate, inference in difficult situations on the basis of relatively small amounts of data.

### 6. Discussion

We have attempted to summarize some inferential procedures that can be used whenever quantities of interest are estimated by solving an equation of the form (1.1). This is often the case for standard techniques such as maximum-likelihood and least-squares estimation, but is much more widely applicable in general. The theory reviewed here is particularly useful in situations where standard techniques are not available: the procedures are straightforward providing the covariance matrix and derivative of the estimating function can be estimated. For accurate inference, however, some care is required to obtain a reasonable estimate of the covariance matrix.

We have focused in particular upon the special case of GMM estimation, because this provides a general purpose—and, in practical terms, relatively painless—way to exploit links between the theoretical properties
of models and observable properties of data. As such, the GMM seems, along with the related method of indirect inference, similar in spirit to the technique of approximate Bayesian computation (ABC; [42,43]) in which the aim is to derive a posterior distribution for quantities of interest on the basis of selected data summaries. Indeed, the combination of ABC with indirect inference has recently been applied to the modelling of macroparasite populations [44].

Although the theory outlined here holds under rather mild conditions, they are not always satisfied. Situations in which the theory is not applicable include: procedures involving the optimization of an objective function where the optimum value does not satisfy (1.1) (as in maximum-likelihood estimation of the unknown endpoint of a distribution—see [45, §7.2]); changepoint problems in which one or more of the model parameters changes abruptly at a particular value of some covariate (often time—for some insight into why these problems are irregular, see [46]); situations where the parameter space \( \Theta \) is a closed set outside which the EF is undefined, and where \( \theta_0 \) lies on the boundary of \( \Theta \) so that derivatives are ill-behaved there (this can occur in models involving multiple variance components when some variances are zero); and problems where there are fundamental differences in the behaviour of the covariance or derivative matrix of the EF as \( \theta \) varies (as when one moves from the stationary to the non-stationary region of parameter space in many time-series models). It is also worth noting that even when the conditions do hold, the target \( \theta_0 \) of the estimation procedure is not guaranteed to be the value of substantive interest. Perhaps, the best-known example of this is the Neyman–Scott or incidental parameters problem in maximum-likelihood estimation [45, §7.6.2]: here, the total number of model parameters increases with the sample size and the maximum-likelihood estimator of a low-dimensional parameter of interest is inconsistent (in this review, we have only considered situations in which \( p \), the dimension of \( \Theta \), is fixed for all \( n \)). Another example involves the application of generalized estimating equations to mis-specified models for clustered data [47].

The techniques for model comparison in §3 assume that one model is nested within the other by imposing some linear restrictions on the parameter vector. In fact, essentially the same result holds for nonlinear restrictions under some mild conditions on their nature (see [31, §8.6] for a discussion of this in the context of maximum-likelihood estimation). However, we have not discussed techniques for comparing non-nested models. For likelihood-based inference, criteria such as the Akaike information criterion and Bayesian information criterion can be used for this purpose [46, §4.7], although even these are considerably less accurate when the models are not nested. Similar criteria seem less well-developed for general EF's. Various suggestions have been made in specific situations however (see, for example, [48–50]), and most of these follow a very similar structure: this points towards the possibility of formulating a general purpose criterion for the comparison of non-nested models under a wide range of different estimation techniques. See Cox [45] for a more general discussion of non-nested model comparison.

We conclude with a caveat. The theory presented here is general in nature, and relatively easy to apply in practice. However, in models of complex systems, it will often be the case that improved inference can be obtained from specialized methods that are more difficult to implement but are specifically designed for the particular problem at hand. In practice, therefore, it will often be necessary to weigh up the convenience of the methods presented here against the potential benefits of a specialist technique.

Joao Jesus’ research was funded by a PhD studentship from the Engineering and Physical Sciences Research Council. We thank the two anonymous reviewers for their constructive and supportive comments.

REFERENCES
