Non-linear Gaussian estimation from orthogonal residuals

Granville Tunnicliffe Wilson*

March 22, 2019

Abstract

Keywords: Gaussian likelihood, score function, information function, non-linear estimation, iterative optimization

1 Gaussian estimation

In this document we present a procedure for estimating the coefficients (or parameters) θ_i, $i = 1 \dots K$ of a model for observed Gaussian distributed variables. The method is based on the transformation of the observations to a set of independent (or orthogonal) residuals (or innovations) e_t and their standard deviations σ_t, $t = 1 \dots T$. The transformation is linear but will generally depend non-linearly upon the model parameters. It is achieved in the time series modeling applications described in Tunnicliffe Wilson et al. (2015) by direct computations from the model or the less direct use of the Kalman filter. It can in theory be applied in the case of any model which can be used to compute the variance matrix of the observations, by applying the inverse Choleski factor of this matrix to the observations. When the observations are Gaussian, the parameter estimates are the maximum likelihood estimates. The same procedure may be applied when the Gaussian assumptions are relaxed, typically to the conditions that the residuals are orthogonal with the same cumulants to order 4 as Gaussian variables. The estimates will generally retain desirable properties such as consistency, and under slightly stronger conditions, large sample normality, though not optimal as maximum likelihood estimates. The term Gaussian estimation is intended to cover these somewhat wider conditions.

2 The log-likelihood, score and information functions

We will present a method of maximizing the (Gaussian) log-likelihood with respect to the model parameters θ_i. This is iterative, with the present, or local, set of model parameters updated using the local derivatives with respect to the parameters and an estimate of the local Hessian matrix. It may be viewed as an extension of non-linear least squares. On

*Department of Mathematics and Statistics, Lancaster University, Lancaster LA1 4YF, UK.

convergence, indicated by near zero changes in the parameter values (reflecting near-zero values of the score) the Hessian is used to form an estimate of the large sample dispersion matrix of the parameter estimates.

In practice we will minimize the deviance of the model defined as minus twice the log likelihood. The reason for this is that differences in the minimized deviance can be be referred to appropriate critical values of the chi-squared distribution to select between nested models, or used in the AIC, Akaike (1973), or similar criteria for model selection. The only quantities required to define the deviance are the orthogonal residuals e_t and their standard deviations σ_t, for $t = 1, \ldots, T$. We also introduce the normalized residual $f_t = e_t/\sigma_t$. At the true parameter values these satisfy

$$\mathrm{E}(e_t) = \mathrm{E}(f_t) = 0, \quad \mathrm{E}(e_t^2) = \sigma_t^2, \quad \mathrm{E}(f_t^2) = 1, \quad \mathrm{E}(e_t\, e_s) = \mathrm{E}(f_t\, f_s) = 0 \quad \text{for} \quad s \neq t. \tag{1}$$

The deviance is then

$$D = \sum_{t=1}^{T} \left\{ 2 \log \sigma_t + f_t^2 \right\}. \tag{2}$$

To develop the score and Hessian we will denote the derivatives wrt the model parameter θ_i by a parenthesized superscript and from $\sigma_t\, f_t = e_t$ derive

$$\sigma_t^{(i)} f_t + \sigma_t\, f_t^{(i)} = e_t^{(i)} \tag{3}$$

$$\sigma_t^{(i\,j)} f_t + \sigma_t^{(i)} f_t^{(j)} + \sigma_t^{(j)} f_t^{(i)} + \sigma_t\, f_t^{(i\,j)} = e_t^{(i\,j)}. \tag{4}$$

Differentiating (2) we obtain the score or gradient

$$D^{(i)} = 2 \sum_{t=1}^{T} \left\{ \frac{\sigma_t^{(i)}}{\sigma_t} + f_t\, f_t^{(i)} \right\} \tag{5}$$

and then the second derivative or Hessian

$$D^{(i\,j)} = 2 \sum_{t=1}^{T} \left\{ -\frac{\sigma_t^{(i)} \sigma_t^{(j)}}{\sigma_t^2} + \frac{\sigma_t^{(i\,j)}}{\sigma_t} + f_t^{(i)}\, f_t^{(j)} + f_t\, f_t^{(i\,j)} \right\}. \tag{6}$$

The remainder of this section is given to deriving a simplified approximation to the Hessian, which involves only first order derivatives. This is desirable, because numerical first derivatives, obtained from parameter perturbations, can be quite sufficiently accurate for iterative optimization, but numerical second derivatives can be poorly conditioned and add substantially to computational effort. The approximation is, strictly, only valid at the true parameter values, because it relies on the distributional properties in (1). But it does have the advantage of being always positive definite, which may fail with numerical second derivatives. After minimization of the deviance the likelihood information matrix may also be consistently estimated by half the approximated Hessian, and its inverse used to provide an estimate of the large sample variance matrix of the parameter estimates.

Our strategy for simplifying (6) is to switch between sample and expected values, evaluated at the true parameters. We first use (4) to simplify the last term in the sum in (6) as

$$\mathrm{E}\left\{ f_t\, f_t^{(i\,j)} \right\} = \mathrm{E}\left\{ f_t \left[\frac{e_t^{(i\,j)}}{\sigma_t} - \frac{\sigma_t^{(i\,j)}\, f_t}{\sigma_t} - \frac{\sigma_t^{(i)}\, f_t^{(j)}}{\sigma_t} - \frac{\sigma_t^{(j)}\, f_t^{(i)}}{\sigma_t} \right] \right\}. \tag{7}$$

Now $e_t = x_t - X_{t-1}$ where X_{t-1} is a linear function of x_{t-k} for $k > 0$, which are all uncorrelated with e_t. Hence all first and second derivatives of e_t are linear functions of the same set and uncorrelated with e_t. Thus on expanding the product in (7), the first term has expected value zero. From (1) the expected value of the second term in this product is $-\sigma_t^{(ij)}/\sigma_t$ which cancels the second term in (6).

From (3), the third term in the product in (7) becomes

$$-f_t \frac{\sigma_t^{(i)} f_t^{(j)}}{\sigma_t} = -f_t \frac{\sigma_t^{(i)} e_t^{(j)}}{\sigma_t^2} + f_t^2 \frac{\sigma_t^{(i)} \sigma_t^{(j)}}{\sigma_t^2}. \tag{8}$$

The first term on the right again has expected value zero and the second term expected value $\sigma_t^{(i)} \sigma_t^{(j)}/\sigma_t^2$ which cancels with the first term in (6). The fourth term in the product in (7) is similar and has the same expected value, resulting in a net positive contribution of $\sigma_t^{(i)} \sigma_t^{(j)}/\sigma_t^2$ in (6).

We now turn to the third term in the sum in (6) which using (3) again is

$$f_t^{(i)} f_t^{(j)} = \left(\frac{e_t^{(i)}}{\sigma_t} - \frac{\sigma_t^{(i)} f_t}{\sigma_t} \right) \left(\frac{e_t^{(j)}}{\sigma_t} - \frac{\sigma_t^{(j)} f_t}{\sigma_t} \right). \tag{9}$$

On expanding, the expectations of the cross terms are zero, but we retain the sample values of the product residual derivatives to give

$$\frac{e_t^{(i)} e_t^{(j)}}{\sigma_t^2} + \frac{\sigma_t^{(i)} \sigma_t^{(j)}}{\sigma_t^2}. \tag{10}$$

The final Hessian approximation is then

$$D^{(ij)} \approx \tilde{D}^{(ij)} = 2 \sum_{t=1}^{T} \left\{ \frac{e_t^{(i)} e_t^{(j)}}{\sigma_t^2} + 2 \frac{\sigma_t^{(i)} \sigma_t^{(j)}}{\sigma_t^2} \right\}. \tag{11}$$

The disappearance of second derivative from this expression is not simply fortuitous; the information matrix can always be expressed in terms of first derivatives of the log likelihood.

3 Iterative minimization of the deviance

The vector g with elements $g_i = D^{(i)}/2$ and matrix M with elements $M_{ij} = \tilde{D}^{(ij)}/2$ are respectively the gradient and approximate Hessian of the half deviance $D/2$, evaluated at a given vector θ of parameter values .

Starting from an initial vector θ^0, an iterative scheme for locating the vector that minimizes the (half) deviance is to form the sequence $\theta^s = \theta^{s-1} - \delta^s$ where the parameter correction vector δ at any given iteration is found by solving the equations

$$M\delta = g. \tag{12}$$

The motivation for this scheme is that when $D(\theta)$ is quadratic in θ it achieves the minimum in one step at θ_1. When D can be locally well approximated by a quadratic, the iterates may be expected to converge to the minimum. However, this is not guaranteed and it is

advisable to have a strategy by which a check is made that the next iterate does achieve a reduction in $D(\theta)$, and if not a reduction made in the size of the step δ. We achieve this reduction by modifying the equation (12) in the manner recommended in Marquardt (1963) for non-linear least squares:

$$(M + \lambda N)\delta = g, \tag{13}$$

where $\lambda > 0$ and N is the diagonal of M. For sufficiently large λ this results in a small iterative step in the direction of the scaled gradient, ensuring a reduction in D. The modifying scalar λ is adjusted at successive iterations, being increased after a successful step in which D is decreased, to allow the next step to be larger. However, if D is not decreased, the same equations are solved with successively increasing values of λ until a decrease in D is achieved.

Given the structure of the equation (13), the better conditioned numerical method of determining their solution uses the QR algorithm. To apply this first use (3) to express

$$g_i = \sum_{t=1}^{T} \left\{ \frac{\sigma_t^{(i)}}{\sigma_t}(1 - f_t^2) + \frac{e_t\, e_t^{(i)}}{\sigma_t^2} \right\}. \tag{14}$$

We may then write

$$g = A'Y, \quad M + \lambda N = A'A \tag{15}$$

where the vector Y of length $2T + K$ and matrix A of size $(2T + K) \times K$ have the partitioned forms

$$Y = \begin{pmatrix} Y_e \\ Y_\sigma \\ Y_\lambda \end{pmatrix}, \quad A = \begin{pmatrix} A_e \\ A_\sigma \\ A_\lambda \end{pmatrix} \tag{16}$$

with the partitioned components having respective lengths T, T and K, and elements

$$Y_{e,\,t} = \frac{e_t}{\sigma_t}, \quad Y_{\sigma,\,t} = \frac{1 - f_t^2}{\sqrt{2}}, \quad Y_{\lambda,\,i} = 0, \tag{17}$$

and

$$A_{e,\,t\,i} = \frac{e_t^{(i)}}{\sigma_t}, \quad A_{\sigma,\,t\,i} = \frac{\sqrt{2}\sigma_t^{(i)}}{\sigma_t}, \quad A_{\lambda,\,i\,i} = \sqrt{\lambda N_{i\,i}}. \tag{18}$$

The equation (13) for the parameter step δ is then expressed

$$A'A\delta = A'Y \tag{19}$$

which is of the form of standard least squares equation. On factorizing $A = QR$ with Q orthonormal and R upper triangular, we may then solve for δ from

$$R\,\delta = Q'\,Y. \tag{20}$$

4 The psuedo-deviance

Consider again the deviance (2) written in terms of the residuals as

$$D = \sum_{t=1}^{T} \left\{ 2\log\sigma_t + \frac{e_t^2}{\sigma_t^2} \right\}. \tag{21}$$

Assume now, as is true for our particular applications, that for any model with parameters that furnishes a particular set of the quantities e_t and σ_t, there is a model with parameters that furnishes the identical set of values of the residuals e_t, but scaled residual standard errors $\alpha\sigma_t$ for any scalar $\alpha > 0$. For our applications this occurs by scaling all the variances and covariances by α^2. In this case the model may be re-parameterized so that $\sigma_t = \sigma\tau_t$ where σ is one of the free parameters. The deviance then becomes

$$D = 2T \log \sigma + \sum_{t=1}^{T} \{2 \log \tau_t\} + \sigma^{-2} \sum_{t=1}^{T} \left\{ \frac{e_t^2}{\tau_t^2} \right\}. \tag{22}$$

This may be explicitly minimized with respect to σ by setting

$$\sigma^2 = T^{-1} \sum_{t=1}^{T} \left\{ \frac{e_t^2}{\tau_t^2} \right\} \tag{23}$$

so that the minimum deviance is

$$\begin{aligned} D &= T \log \sum_{t=1}^{T} \left\{ \frac{e_t^2}{\tau_t^2} \right\} - T \log T + \sum_{t=1}^{T} \{2 \log \tau_t\} + T \qquad (24) \\ &= T \log \sum_{t=1}^{T} \left\{ \frac{e_t^2}{\sigma_t^2} \right\} - T \log T + \sum_{t=1}^{T} \{2 \log \sigma_t\} + T, \qquad (25) \end{aligned}$$

where in the last line we have simply subtracted $T \log \sigma^2$ from the first sum and added it to the second. The deviance is now a monotonic function of

$$pD = \sum_{t=1}^{T} \left\{ \frac{e_t^2}{\tau_t^2} \right\} \left\{ \prod_{1}^{T} \tau_t^2 \right\}^{\frac{1}{T}} = \sum_{t=1}^{T} \left\{ \frac{e_t^2}{\sigma_t^2} \right\} \left\{ \prod_{1}^{T} \sigma_t^2 \right\}^{\frac{1}{T}}. \tag{26}$$

We will use the second form as it is explicit in the quantities e_t and σ_t which we stated initially were those needed to define the likelihood. The first form in e_t and τ_t motivates our use of the term psuedo-deviance because in the simple regression context $\tau_t \equiv 1$ and the psuedo-deviance is the residual sum of squares. We consider pD defined in (26) to be a natural generalization of the residual sum of squares. It is positive and under the assumptions we have made it reduces as the deviance reduces and is a useful quantity for monitoring the progress of iterative estimation. Upon convergence it is a simple function of the deviance, $D = T \log pD + T - T \log T$and we use $T \log pD$ in place of D in expressions for the AIC.

References

H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AC–19(2):716–723, 1973.

D. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *SIAM J. Appl. Math.*, 11:431–441, 1963.

G. Tunnicliffe Wilson, M. Reale, and J. Haywood. *Models for dependent time series*. New York, CRC Press, 2015.