Geophysical Research Abstracts
Vol. 13, EGU2011-12761, 2011
EGU General Assembly 2011

Likelihood under autocorrelated residuals using effective sample size

Sjur Kolberg (1), Kolbjørn Engeland (2), and Ingelin Steinsland (3)

(1) SINTEF Energy Research, Trondheim, Norway (Sjur.Kolberg@sintef.no, 47 73947250), (2) SINTEF Energy Research, Trondheim, Norway (Kolbjorn.Engeland@sintef.no, 47 73947250), (3) NTNU Dept of Mathematics, Trondheim, Norway (ingelins@math.ntnu.no)

For estimating the uncertainty of hydrological model parameters and the predictions they generate, two classes of performance measures are well known in hydrologic literature. One is the additive, sum of squares based approach, here represented by the Nash-Sutcliffe efficiency (eq.1). The other is a formal, multiplicative likelihood function, here represented by a Normal likelihood for independent data (eq.2).

$$R_{eff} = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i(\theta))^2}{Var[x_i]} \quad \text{Eq. .1}$$

$$L(\theta|X) \propto \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \hat{x}(\theta)_i\right)^2\right] \quad \text{Eq. .2}$$

The two measures largely differ in their coherence; which determines how they respond to more data being added. Equation 1 is technically incoherent, since R_{eff} is independent of data set size n. Equation 2 is coherent, but ignores the high autocorrelation of residuals always present in hydrological modelling, and hence underestimates parameter variance.

The traditional way to avoid this over-conditioning is to formulate an autoregressive error model, for instance by assuming an AR(1) process. However, this approach integrates the hydrological model with a statistical error model which relies on the last time step's residual as an input variable. This combined model is only validated for known residuals, and quickly deteriorates for longer lead times. In addition, the likelihood tends to maximise the error persistence, rather than minimising the errors themselves.

Instead, we propose the concept of effective sample size to quantify the information content in a data set. We rewrite the Normal log-likelihood by combining equations 1 and 2:

$$L(\theta|X) \propto \exp\left[(R_{eff} - 1) \cdot n \cdot \frac{Var[X]}{\sigma^2}\right] \quad \text{Eq. .3}$$

In the log-likelihood expression (inside the brackets) only the factor n depends on the amount of data. The first factor measures the goodness of fit, whereas the third is a simple scaling, independent of both of these. We assume that a proper transformation (Box-Cox or similar) of the residuals is performed so that the Normal assumption is reasonable. Our proposal is now to simply replace the nominal n with an effective sample size n_{eff} reflecting the autocorrelation in the data. Using $n_{eff} = n$implicitly assumes that all error terms are independent. By lowering n_{eff} , Eq. 3 can be directly applied to an autocorrelated series of residuals. The error covariance matrix provides the relation of n_{eff} to the nominal sample size n, hence n_{eff} can be calculated for a given covariance model. For instance, it has been shown that an AR(1) model yields:

$$n_{eff} = n\frac{1-a}{1+a} \quad \text{Eq. .4}$$

where a is the lag 1 autocorrelation.

Estimating n_{eff} can be extended beyond the AR(1) process without interfering with the likelihood function. Even when no residual covariance model can be constructed from realistic assumptions, the modeller may still subjectively choose a value for n_{eff} . This value precisely quantifies the modeller's assessment of the information content in the data, it is intuitively pervieved by readers and reviewers, and thus aids in communicating and discussing the uncertainty of parameters and their resulting predictions.

Preliminary analyses suggest that (1) n_{eff} approach can be used for regional parameter estimation in a distributed model, (2) The results are sensitive to the value of n_{eff} , (3) A simple iterative routine can be used to assess n_{eff} , (4) Reducing n_{eff} by 50% increases confidence interval width by approx. 30%, (5) The parameter uncertainty alone cannot explain the model residuals.