Ecological Modelling 253 (2013) 107–116

Contents lists available at SciVerse ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Evaluation of a soil greenhouse gas emission model based on Bayesian inference and MCMC: Parameter identifiability and equifinality

Gangsheng Wang[a,b,*], Shulin Chen[a,*]

[a] Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120, USA
[b] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6301, USA

ARTICLE INFO

Article history:
Received 29 June 2012
Received in revised form 6 September 2012
Accepted 9 September 2012
Available online 22 November 2012

Keywords:
Bayesian inference
Condition number
Equifinality
Markov Chain Monte Carlo (MCMC)
Identifiability

ABSTRACT

Identifiability and equifinality are two interrelated concepts in mathematical modeling. The derivation of the Hessian matrix becomes crucial when the condition number is used as a diagnostic indicator for identifiability. The covariance-inverse (CI) method was proposed to derive the Hessian matrix via the inverse matrix of covariance. The covariance matrix is calculated directly from the posterior parameter samples. Compared with two existing methods, i.e., difference quotients (DQ) and quasi-analytical (QA), CI is more efficient and reliable. The CI method was then used for identifiability diagnosis on a soil greenhouse gas emission (SoilGHG) model. The model as a whole was poorly identified, but a reduced model with fewer parameters could become identifiable, which is called "conditionally identifiable" in this paper. The geometric mean condition numbers in terms of sorted singular values of the full Hessian matrix could be adopted as criteria to determine at most how many undetermined parameters might be included in an identifiable or weakly identifiable model. The combinations of parameters that made the model identifiable were also determined by the proposed diagnosis method. We addressed the importance of understanding both identifiability and equifinality in ecosystem modeling.

1. Introduction

Model identifiability is the problem of determining whether the parameters of a given mathematical model can be uniquely (globally or locally) recovered from data (Serban and Freeman, 2001). A parameter having a value of x_0 is globally identifiable if and only if no other x_1 within the feasible parameter space gives the same value of the likelihood function. If the value of the likelihood at x_0 is unique in a sub-space containing x_0, this parameter is said to be locally identifiable (Iskrev, 2007; Sorooshian and Gupta, 1985). Theoretically, the number of identifiable parameters is determined by the rank of the Hessian matrix (Sun, 2004; Viallefont et al., 1998). If the Hessian matrix is positive definite, then the model structure is locally identifiable. However, positive definite is often not true and other times impossible to determine due to computing errors (Sorooshian and Gupta, 1985). Generally, global identification analysis for non-linear models is infeasible, and the Hessian matrix approach is only applicable for local identification analysis (Iskrev, 2007). Many studies introduced the condition number to diagnose whether the Hessian matrix was ill conditioned or not (Dee and da Silva, 1999; Dee et al., 1999; Lebbe and Van Meir, 2000; Serban and Freeman, 2001; Sun, 2004; Viallefont et al., 1998). A problem with a low condition number is said to be well-conditioned, while a problem with a high condition number is said to be ill-conditioned (Iskrev, 2007; Viallefont et al., 1998). Xia (1989) stated that non-identifiability of a model could be caused by parameter interactions, the lack of data, or the limitations of parameter optimization methods.

Equifinality is a concept to reject the existence of a unique optimal parameter set for a data-driven model (Beven and Freer, 2001; Wang and Chen, 2012). In other words, there may coexist multiple choices of parameter sets or model structures, which can partially produce acceptable simulations (Wang et al., 2009b; Williams et al., 2009). A previous paper (Wang and Chen, in press) has provided evidence that equifinality did exist in the soil greenhouse gas emission (SoilGHG) model (see Appendix A for model description), although the equifinality could be reduced via uncertainty analysis. As defined earlier, a model having a unique optimal parameter set in a sub-space is locally identifiable. Thus, equifinality and identifiability show apparently opposite characteristics in model parameterization and assessment; however, they are two interdependent concepts. On one hand, the idea of attaining

* Corresponding authors at: Department of Biological Systems Engineering, Washington State University, Pullman, WA 99164-6120, USA.
E-mail addresses: wangg@ornl.gov (G. Wang), chens@wsu.edu (S. Chen).

0304-3800/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.ecolmodel.2012.09.011

identifiable parameters is part of the general working paradigm for modelers (Beven and Freer, 2001). On the other hand, a completely identified model is difficult to realize because of incorrect or defective descriptions of the real processes and characteristics in environmental science and engineering, which means confronting the phenomenon of equifinality is inevitable. The effort to strive for identifiable parameters is the recognition of equifinality at the same time. Equifinality cannot be completely eliminated, but can be attenuated in some sense.

The derivation of the Hessian matrix becomes crucial when the condition number is used as a diagnostic indicator of identifiability. Three methods were usually used to estimate the Hessian matrix: the analytical method, the quasi-analytical method and the difference quotients. The analytical method is limited to the conditions that the objective functions are explicitly derivable in second order. However, the objective functions usually cannot be obtained as an explicit function due to the implicit structure of most models. Therefore, the difference quotients (DQ) method and the quasi-analytical (QA) method with numerical approximation become useful in this circumstance. The DQ method approximately computes the second order derivatives at the optimal points of the model parameters (Viallefont et al., 1998; Xu and Vandewiele, 1995). Accordingly, an optimal parameter set is required prior to the implementation of DQ. Equifinality in a model structure could easily result in the failure of the DQ method. The QA method estimates the Hessian by means of least-squares fitting (Xia, 1989; Yam, 1997). The implementation of QA is complex because of the fitting of a large amount of data and variables. In addition, the significance and goodness (R^2) of the least-squares fitting directly influences the reliability of the Hessian matrix.

The objective of this study was to evaluate the model identifiability and equifinality by developing a new method called covariance-inverse (CI). The main characteristic of CI was to estimate the Hessian matrix by inversing the covariance matrix, which could be calculated from the parameter samples using the multinormal proposal distribution (MPD) as described in Wang and Chen (in press). The CI method was also based on Bayesian inference and Markov Chain Monte Carlo (MCMC) technique (Knorr and Kattge, 2005; Wang and Chen, in press). The same model input and GHG data (see Appendix B) as described in Wang and Chen (in press) and Wang et al. (2012a) were used. The results from the CI method were also compared with those from DQ and QA.

2. Methods

2.1. Diagnosis of model identifiability

The Hessian matrix of an objective function is not the only parameter analysis tool available, but it is a useful diagnostic to immediately present the identifiability of parameters (Bostick et al., 2007; Kavetski et al., 2006; Viallefont et al., 1998). A common expression of the objective function in modeling is the sum of squared errors (SSE):

$$F(\Theta) = \sum_{i=1}^{n} [y_i - \hat{y}_i]^2 \tag{1}$$

where y_i and $\hat{y}_i$ are observation data and simulation data, respectively, at time i; and $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is the parameter set including k parameters.

Table 1
Diagnosis of structural identifiability by condition number (after Xia, 1989).

Condition number (η)	$1/\eta$	Identifiability
$1 \le \eta \le 10$	$0.1 \le 1/\eta \le 1$	Identifiable
$10 < \eta \le 200$	$0.005 \le 1/\eta < 0.1$	Weakly identifiable
$\eta > 200$	$0 \le 1/\eta < 0.005$	Non-identifiable

The Hessian matrix (H) is the square matrix of second-order partial derivatives of the objective function (F):

$$H_{k\times k} = \begin{bmatrix} \dfrac{\partial^2 F}{\partial \theta_1^2} & \dfrac{\partial^2 F}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 F}{\partial \theta_1 \partial \theta_k} \\ \dfrac{\partial^2 F}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 F}{\partial \theta_2^2} & \cdots & \dfrac{\partial^2 F}{\partial \theta_2 \partial \theta_k} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 F}{\partial \theta_k \partial \theta_1} & \dfrac{\partial^2 F}{\partial \theta_k \partial \theta_2} & \cdots & \dfrac{\partial^2 F}{\partial \theta_k^2} \end{bmatrix} \tag{2}$$

The model identifiability can be measured by the matrix condition number (η) (Iskrev, 2007) defined as the ratio of the largest to the smallest singular value of the matrix (Edelman, 2005; Iskrev, 2007; Sorooshian and Gupta, 1985).

$$\eta = \frac{\lambda_{\max}}{\lambda_{\min}} \tag{3}$$

The diagnosis of model identifiability was based on the magnitude of the condition number as shown in Table 1 (Xia, 1989).

2.2. Relationships between Hessian, covariance, and correlation matrices

The covariance and correlation matrices of the parameters can be estimated from the Hessian matrix (Sorooshian and Gupta, 1985; Vandewiele et al., 1992; Xu and Vandewiele, 1995):

$$\sum\nolimits_{k\times k} = 2\sigma^2 H_{k\times k}^{-1} \tag{4}$$

$$\sigma = \sqrt{\frac{F(\Theta)_{\min}}{n-k}} \tag{5}$$

$$\rho(\theta_i, \theta_j) = \frac{\sum(\theta_i, \theta_j)}{\sigma(\theta_i)\cdot\sigma(\theta_j)} \tag{6}$$

$$\sigma(\theta_i) = \sqrt{\sum(\theta_i, \theta_i)} \tag{7}$$

where $F(\Theta)_{\min}$ is the minimum $F(\Theta)$; $H_{k\times k}$ is the Hessian of $F(\Theta)_{\min}$; $\sum_{k\times k}$ is the covariance matrix; $\rho(\theta_i,\theta_j)$ is the correlation coefficient between θ_i and θ_j; σ is the model standard deviation; $\sigma(\theta_i)$ is the standard deviation of θ_i; n is the number of terms in Eq. (1); and k is the number of parameters.

2.3. Estimation of Hessian matrix

If $F(\Theta)$ in Eq. (1) can be explicitly expressed as a function of Θ, and $F(\Theta)$ is derivable in second order, the Hessian matrix can be calculated directly by Eq. (2). As previously mentioned, this is the analytical method. However, it is impossible to analytically derive the second-order derivatives for a complex non-linear model. That is why the DQ and QA methods have been developed. In this section, the DQ and QA were introduced first, and then the CI method was proposed.

2.3.1. Difference quotients method

The Hessian can be approximated at mesh nodes by DQ (Viallefont et al., 1998). Such an approximation can be a forward,