Math Geosci
DOI 10.1007/s11004-011-9359-0

Dependence of Bayesian Model Selection Criteria and Fisher Information Matrix on Sample Size

Dan Lu · Ming Ye · Shlomo P. Neuman

Received: 11 June 2010 / Accepted: 26 August 2011

Abstract Geostatistical analyses require an estimation of the covariance structure of a random field and its parameters jointly from noisy data. Whereas in some cases (as in that of a Matérn variogram) a range of structural models can be captured with one or a few parameters, in many other cases it is necessary to consider a discrete set of structural model alternatives, such as drifts and variograms. Ranking these alternatives and identifying the best among them has traditionally been done with the aid of information theoretic or Bayesian model selection criteria. There is an ongoing debate in the literature about the relative merits of these various criteria. We contribute to this discussion by using synthetic data to compare the abilities of two common Bayesian criteria, *BIC* and *KIC*, to discriminate between alternative models of drift as a function of sample size when drift and variogram parameters are unknown. Adopting the results of Markov Chain Monte Carlo simulations as reference we confirm that *KIC* reduces asymptotically to *BIC* and provides consistently more reliable indications of model quality than does *BIC* for samples of all sizes. Practical considerations often cause analysts to replace the observed Fisher information matrix entering into *KIC* with its expected value. Our results show that this causes the performance of *KIC* to deteriorate with diminishing sample size. These results are equally valid for one and multiple realizations of uncertain data entering into our analysis. Bayesian theory indicates that, in the case of statistically independent and identically distributed data, posterior model probabilities become asymptotically insensitive to prior probabilities as sample size increases. We do not find this to be the case when working with samples taken from an autocorrelated random field.

D. Lu · M. Ye (✉)
Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA
e-mail: mye@fsu.edu

S.P. Neuman
Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA

Published online: 14 October 2011

Keywords Model uncertainty · Model selection · Variogram models · Drift models · Prior model probability · Asymptotic analysis

1 Introduction

Geostatistical analyses require estimating the covariance structure of a random field and its parameters jointly from noisy data. Whereas in some cases (as in that of a Matérn variogram; Matérn 1986; Marchant and Lark 2007) a range of structural models can be captured with one or a few parameters, in many other cases one must consider a discrete set of structural (such as drift and variogram) model alternatives (Hoeksema and Kitanidis 1985; McBratney and Webster 1986; Samper and Neuman 1989a, 1989b; Ye et al. 2004, 2005; Marchant and Lark 2004, 2007; Riva and Willmann 2009; Nowak et al. 2010; Nowak 2010; Singh et al. 2010; Riva et al. 2011). Ranking these alternatives and identifying the best among them has traditionally been done with the aid of information theoretic model selection (discrimination, information) criteria such as Akaike Information Criterion (*AIC*, Akaike 1974) and Corrected Akaike Information Criterion (*AICc*, Hurvich and Tsai 1989) or Bayesian criteria, most commonly Bayesian Information Criterion (*BIC*, Schwarz 1978; Rissanen 1978) and Kashyap Information Criterion (*KIC*, Kashyap 1982). There is an ongoing debate in the literature about the relative merits and demerits of these and related criteria (Poeter and Anderson 2005; Tsai and Li 2008a, 2008b, 2010; Ye et al. 2008a, 2010a, 2010b; Riva et al. 2011). We contribute to this discussion by using synthetic data coupled with Markov Chain Monte Carlo (*MCMC*) simulations to compare the abilities of *BIC* and *KIC* to discriminate between alternative models of drift as a function of sample size when drift and variogram parameters are unknown. As our comparison is based on Bayesian statistics, it does not apply directly to *AIC* and *AICc*, since these are derived based on other principles.

Consider a set of K alternative (drift and/or variogram) models M_k of an autocorrelated random function Y with N_k unknown parameters $\boldsymbol{\theta}_k$ (bold letters designating vectors), $k = 1, \ldots, K$. We wish to compare and rank these models after each has been calibrated by maximum likelihood (ML) against a common sample $\mathbf{D}$ of N measured Y values at various points in space and/or time. One way to do so is to associate the following criteria with each model

$$AIC_k = -2\ln\left[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})\right] + 2N_k, \tag{1}$$

$$AICc_k = -2\ln\left[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})\right] + 2N_k + \frac{2N_k(N_k+1)}{N-N_k-1}, \tag{2}$$

$$BIC_k = -2\ln\left[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})\right] + N_k \ln N, \tag{3}$$

$$KIC_k = -2\ln\left[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})\right] - 2\ln p(\hat{\boldsymbol{\theta}}_k) + N_k \ln(N/2\pi) + \ln|\bar{\mathbf{F}}_k| \tag{4}$$

where $\hat{\boldsymbol{\theta}}_k$ is the ML estimate of $\boldsymbol{\theta}_k$; $-\ln[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})]$ is the negative log-likelihood (*NLL*) function $-\ln[L(\boldsymbol{\theta}_k|\mathbf{D})]$ evaluated at $\hat{\boldsymbol{\theta}}_k$; $p(\hat{\boldsymbol{\theta}}_k)$ is the prior probability of $\boldsymbol{\theta}_k$ evaluated at $\hat{\boldsymbol{\theta}}_k$; and $\bar{\mathbf{F}}_k = \mathbf{F}_k/N$ is the normalized (by N) observed (implicitly con-

ditioned on $\mathbf{D}$ and evaluated at $\hat{\boldsymbol{\theta}}_k$) Fisher information matrix (FIM), $\mathbf{F}_k$, having elements (Kashyap 1982)

$$\bar{F}_{k,ij} = \frac{1}{N} F_{k,ij} = -\frac{1}{N} \frac{\partial^2 \ln[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})]}{\partial\theta_{ki}\partial\theta_{kj}}\bigg|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_k}. \tag{5}$$

This allows rewriting KIC_k in (4) as

$$KIC_k = -2\ln\big[L(\hat{\boldsymbol{\theta}}_k|\mathbf{D})\big] - 2\ln p(\hat{\boldsymbol{\theta}}_k) - N_k \ln(2\pi) + \ln|\mathbf{F}_k|, \tag{6}$$

a form known in Bayesian statistics as the Laplace approximation (Kass and Vaidyanathan 1992; Kass and Raftery 1995), the origin of which can be traced back to Jeffreys (1961) and Mosteller and Wallace (1964). The four criteria embody the principle of parsimony, penalizing (to various degrees) models having a relatively large number of parameters if this does not bring about a corresponding improvement in model fit. A comparative discussion of the four criteria, their underlying principles and ways to compute them can be found in Ye et al. (2008a).

Rather than selecting the highest ranking model, it may sometimes be advantageous to retain several dominant models and average their results. Here we consider an maximum likelihood (*ML*) version of Bayesian model averaging (*MLBMA*) proposed by Neuman (2003) and employed by Ye et al. (2004, 2005, 2008a). If Δ is the quantity one wants to predict, then Bayesian Model Averaging (BMA) consists of expressing the posterior probability of Δ as (Hoeting et al. 1999)

$$p(\Delta|\mathbf{D}) = \sum_{k=1}^{K} p(\Delta|M_k,\mathbf{D})\,p(M_k|\mathbf{D}) \tag{7}$$

where $p(\Delta|M_k,\mathbf{D})$ is the posterior probability of Δ under model M_k having posterior probability $p(M_k|\mathbf{D})$. Bayes' theorem implies that the latter is given by

$$p(M_k|\mathbf{D}) = \frac{p(\mathbf{D}|M_k)\,p(M_k)}{\sum_{l=1}^{K} p(\mathbf{D}|M_l)\,p(M_l)} \tag{8}$$

where $p(M_k)$ is the prior probability of model M_k and

$$p(\mathbf{D}|M_k) = \int p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)\,p(\boldsymbol{\theta}_k|M_k)\,d\boldsymbol{\theta}_k \tag{9}$$

is its integrated likelihood, $p(\boldsymbol{\theta}_k|M_k)$ being the prior probability density of $\boldsymbol{\theta}_k$ under model M_k and $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ the joint likelihood of M_k and $\boldsymbol{\theta}_k$. The integrated likelihood can be evaluated using various Monte Carlo (*MC*) simulation methods (Kass and Raftery 1995), most notably the Markov Chain Monte Carlo (*MCMC*) technique (Lewis and Raftery 1997; Draper 2007). Though we employ *MCMC* in this paper, we note that *MC* methods are computationally demanding and require specifying a prior probability distribution for each set of model parameters. When prior information about parameters is unavailable, vague or diffuse one can resort to *MLBMA* (Neuman 2003) according to which (Ye et al. 2004)

$$p(\mathbf{D}|M_k) \approx \exp\left(-\frac{1}{2}KIC_k\right). \tag{10}$$

Theory shows (Draper 1995; Raftery 1995; Ye et al. 2008a) that as the sample size N increases relative to N_k, *KIC* tends asymptotically to *BIC* and so

$$p(\mathbf{D}|M_k) \approx \exp\left(-\frac{1}{2}BIC_k\right). \tag{11}$$

Substituting (10) or (11) into (8) yields

$$p(M_k|\mathbf{D}) = \frac{\exp(-\frac{1}{2}\Delta IC_k)\,p(M_k)}{\sum_{l=1}^{K}\exp(-\frac{1}{2}\Delta IC_l)\,p(M_l)} \tag{12}$$

where $\Delta IC_k = IC_k - IC_{\min}$ and $IC_{\min} = \min_k\{IC_k\}$, *IC* being either *KIC* or *BIC*.

Equation (12) does not apply formally to *AIC* and *AICc* as these are based on information theory, not on Bayesian statistics. The information theoretic equivalent of MLBMA has been to average model outputs using so-called Akaike weights computed via (Burnham and Anderson 2002, 2004; Poeter and Anderson 2005)

$$w_k = \frac{\exp(-\frac{1}{2}\Delta AIC_k)}{\sum_{l=1}^{K}\exp(-\frac{1}{2}\Delta AIC_l)} \quad \text{or} \quad w_k = \frac{\exp(-\frac{1}{2}\Delta AICc_k)}{\sum_{l=1}^{K}\exp(-\frac{1}{2}\Delta AICc_l)}. \tag{13}$$

Interpreting (13) within a Bayesian context would be equivalent to assigning an equal prior probability $p(M_k) = 1/K$ to each model. Another ad hoc option discussed by Burnham and Anderson (2002, 2004) and adopted by Poeter and Hill (2007) is to use (12) regardless of whether *IC* is information theoretic or Bayesian.

There is general consensus in the literature that *AICc* is more accurate than *AIC* when $N/N_k < 40$, converging rapidly to *AIC* as N/N_k increases beyond 40 (Burnham and Anderson 2002; Poeter and Anderson 2005). There is less consensus about the relative merits of *BIC* and *KIC* (Tsai and Li 2008a, 2010; Ye et al. 2008a, 2010a, 2010b) even though *KIC* is known to converge toward *BIC* as N/N_k goes up (Draper 1995; Raftery 1995; Kass and Raftery 1995; Ye et al. 2008a). Whereas Kass and Raftery (1995) found *BIC* to yield unsatisfactory approximations of integrated likelihood even for very large N/N_k, Tsai and Li (2008a) considered *KIC* to have yielded "unreasonable" posterior model probabilities. In light of contrary findings by Ye et al. (2008a) and several references therein, there appears to be a need to clarify the issue further, as we do below.

In this paper we use samples from a synthetic random field to compare the abilities of *BIC* and *KIC* to discriminate between alternative models of drift as a function of sample size when drift and variogram parameters are unknown. Our study focuses on model probabilities, not on predictive performance. It contributes theoretically and numerically to our understanding of roles played in model selection by observed and expected versions of the Fisher information matrix (Ye et al. 2010b). The study extends to one as well as multiple realizations of uncertain data employed in model calibration. To our knowledge, we are the first to compare the performance of model discrimination criteria against Markov Chain Monte Carlo simulations of various size replicate data sampled from a spatially correlated random field in more than one dimension. For reasons noted earlier, we exclude *AIC* and *AICc* from further consideration in this paper.

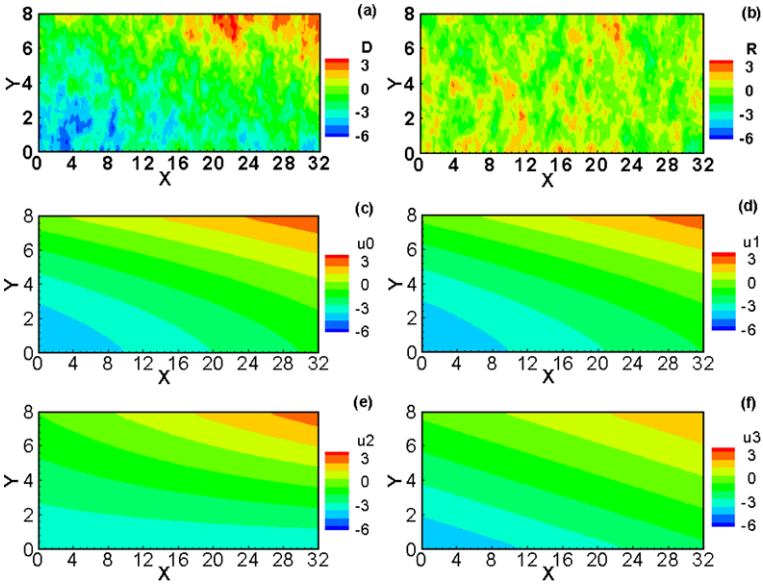

Fig. 1 Contours of true (synthetically generated) (**a**) data, **D**, (**b**) residual, **R**, (**c**) trend, $\boldsymbol{\mu}_0$, of true (data generating) model M_0 and (**d**) trend, $\boldsymbol{\mu}_1$, of model M_1, (**e**) trend, $\boldsymbol{\mu}_2$, of model M_2, (**f**) trend, $\boldsymbol{\mu}_3$, of model M_3 fitted to 1,000 true data points

2 Synthetic Example and Model Calibration

We decompose the sample **D** of random field measurements into a deterministic trend $\boldsymbol{\mu}$ and a spatially correlated random residual **R**,

$$\mathbf{D} = \boldsymbol{\mu} + \mathbf{R}. \tag{14}$$

Estimating trends is important in geostatistics (Pardo-Iguzquiza and Dowd 2001; Ortiz and Deutsch 2002; Nowak 2010), groundwater modeling (Pardo-Iguzquiza et al. 2009; Nowak et al. 2010; Riva et al. 2011) and other areas of the geosciences. However, the choice of trend model is rarely self-evident and often arbitrary (Journel and Rossi 1989; Kyriakidis and Journel 1999; Leuangthong and Deutsch 2004). As **R** in (14) depends on $\boldsymbol{\mu}$ and **D**, uncertainty about the drift translates into uncertainty about the structure (variogram) of **R** (Cressie 1993). To address this issue we consider a two-dimensional rectangular domain of length 32 and width 8 measured in arbitrary units. The domain is discretized into $160 \times 40 = 6400$ squares of uniform size 0.2×0.2. We then generate a random field over this grid having drift

$$M_0: \quad \boldsymbol{\mu}(x, y) = a_0 + a_1 x + a_2 y + a_3 y^2, \tag{15}$$

with coefficients $a_0 = -5.0, a_1 = 0.1, a_2 = 0.2$ and $a_3 = 0.05$ and designate this (true, data generating) model by M_0. We also consider three alternative drift models

$$M_1: \quad \boldsymbol{\mu}(x, y) = a_0 + a_1 x + a_2 y + a_3 y^2 + a_4 xy, \tag{16}$$

$$M_2: \quad \boldsymbol{\mu}(x, y) = a_0 + a_2 y + a_3 y^2 + a_4 xy, \tag{17}$$

$$M_3: \quad \boldsymbol{\mu}(x, y) = a_0 + a_1 x + a_2 y. \tag{18}$$

In comparison to M_0, model M_1 is over-parameterized, model M_3 is under-parameterized and model M_2 has elements of both. We then use sequential Gaussian simulation (SGSIM; Deutsch and Journel 1998) to generate stationary, zero-mean random residuals $\mathbf{R}$ about $\boldsymbol{\mu}$ having unit variance and an exponential variogram, $\gamma(h) = s[1 - \exp(-h/\lambda)]$, with sill (variance) $s = 1$ and integral scale $\lambda = 1$ where h is separation distance or lag. The corresponding covariance function of $\mathbf{R}$ is $Q(h) = s\exp(-h/\lambda)$. To keep the problem manageable we consider neither alternative variogram models nor measurement noise in this example. The generated (true) field of $\mathbf{D}$, residual, $\mathbf{R}$, and drift, $\boldsymbol{\mu}$ (from model M_0) are illustrated in Figs. 1(a)–1(c), respectively. Figures 1(d)–1(f) depict drifts generated after fitting (via ML, as described later) the drift models corresponding to models M_1, M_2 and M_3 to 1,000 true data points. Visual inspection suggests that drift model M_1 is closest and M_2 farthest from the true drift model M_0 in Fig. 1(c).

Let $\boldsymbol{\theta} = \{\mathbf{a}, \boldsymbol{\beta}\}$ where $\mathbf{a}$ is a vector of drift parameters and $\boldsymbol{\beta}$ a vector of variogram parameters. If $\mathbf{D}$ is multivariate Gaussian with mean $\boldsymbol{\mu}(\mathbf{a}) = E[\mathbf{D}]$ and covariance matrix $\mathbf{Q}(\boldsymbol{\beta}) = E[(\mathbf{D} - \boldsymbol{\mu})(\mathbf{D} - \boldsymbol{\mu})^T]$, the negative logarithm of the joint likelihood function (*NLL*) for any given model takes the form (after dropping the subscript k)

$$\begin{aligned} NLL &= -2\ln\big[L(\boldsymbol{\theta}|\mathbf{D})\big] = -2\ln\big[L(\mathbf{a}, \boldsymbol{\beta}|\mathbf{D})\big] = 2\ln p(\mathbf{D}|\mathbf{a}, \boldsymbol{\beta}) \\ &= N\ln 2\pi + \ln\big|\mathbf{Q}(\boldsymbol{\beta})\big| + \big(\mathbf{D} - \boldsymbol{\mu}(\mathbf{a})\big)^T \mathbf{Q}^{-1}(\boldsymbol{\beta})\big(\mathbf{D} - \boldsymbol{\mu}(\mathbf{a})\big). \end{aligned} \tag{19}$$

Minimizing *NLL* simultaneously with respect to $\mathbf{a}$ and $\boldsymbol{\beta}$ would yield biased estimates of $\boldsymbol{\beta}$ (Cressie 1993). To avoid this, we follow Ye et al. (2004) by first using adjoint state ML cross validation (ASMLCV, Samper and Neuman 1989a) in conjunction with universal kriging to obtain an ML estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, then obtaining an ML estimate $\hat{\mathbf{a}}$ of $\mathbf{a}$ by minimizing

$$-2\ln\big[L(\mathbf{a}, \hat{\boldsymbol{\beta}}|\mathbf{D})\big] = N\ln 2\pi + \ln\big|\mathbf{Q}(\hat{\boldsymbol{\beta}})\big| + \big(\mathbf{D} - \boldsymbol{\mu}(\mathbf{a})\big)^T \mathbf{Q}^{-1}(\hat{\boldsymbol{\beta}})\big(\mathbf{D} - \boldsymbol{\mu}(\mathbf{a})\big) \tag{20}$$

with the aid of generalized least squares. The *NLL* component constituting the first term of each model selection criterion in (1)–(6) is thus given by

$$\begin{aligned} -2\ln\big[L(\hat{\boldsymbol{\theta}}|\mathbf{D})\big] &= -2\ln\big[L(\hat{\mathbf{a}}, \hat{\boldsymbol{\beta}}|\mathbf{D})\big] \\ &= N\ln 2\pi + \ln\big|\mathbf{Q}(\hat{\boldsymbol{\beta}})\big| + (\mathbf{D} - \boldsymbol{\mu}(\hat{\mathbf{a}}))^T \mathbf{Q}^{-1}(\hat{\boldsymbol{\beta}})\big(\mathbf{D} - \boldsymbol{\mu}(\hat{\mathbf{a}})\big). \end{aligned} \tag{21}$$

We start by drawing a sample $\mathbf{D}$ of 50 randomly situated data from the 6,400 synthetic random field values in Fig. 1(a), and then increase the sample size to 100. Subsequently, the sample size is gradually increased in 12 increments of 100–200 additional randomly selected data till the total reaches 1,700. We then use each of these 14 samples $\mathbf{D}$ to calibrate the variogram and all alternative drift models in the manner just described. Corresponding estimates of sill s and integral scale λ associated with each model are plotted in Fig. 2. The estimates of both parameters are seen

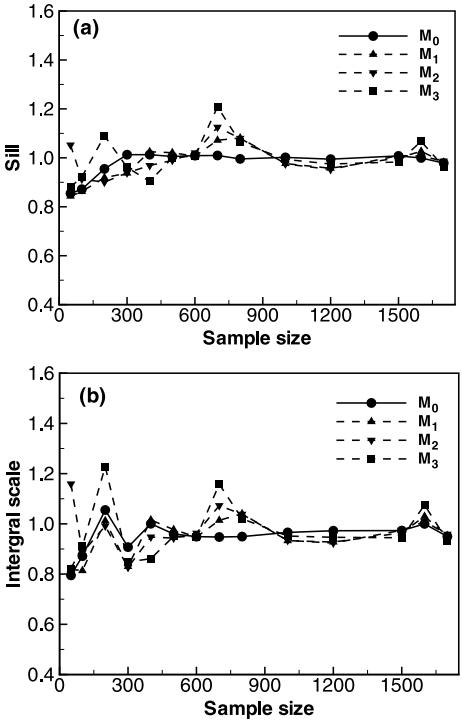

Fig. 2 Estimates of variogram parameters (**a**) sill and (**b**) integral scale for four alternative drift models and fourteen sample sizes

to improve (or approach their true values) with sample size regardless of which drift model one uses, the best estimates being associated with the true model M_0 and the worst with the under-parameterized model M_3, as one would expect.

Fluctuations in parameter estimates and their statistics reflect randomness of the generated field and of samples drawn from it. We explore this effect below for selected sample sizes (50, 100, 300, 500, 800 and 1,000) by extracting 100 replicate samples of each size and calibrating the models against each replicate in the manner described earlier. The results of one and replicate samples are equally important. The one-sample situation mimics what could happen in an actual situation; the replicates indicate the average properties of different diagnostics. When results obtained from one sample are ambiguous or conflict with expectation, one may wish to collect more data and/or supplement them with subjective expert judgment (Ye et al. 2005, 2008b).

Equation (19) is the negative logarithm of a corresponding likelihood function $p(\mathbf{D}|\boldsymbol{\theta}, M)$. The latter is associated with an observed Fisher information matrix (FIM) given analytically by (Kitanidis and Lane 1985)

$$
\begin{aligned}
F_{ij} &= -\frac{\partial^2 \ln p(\mathbf{D}|\boldsymbol{\theta}, M)}{\partial\theta_i \partial\theta_j} \\
&= -\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial\theta_i}\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial\theta_j}\right) + \frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\frac{\partial^2 \mathbf{Q}}{\partial\theta_i \partial\theta_j}\right) \\
&\quad + \frac{\partial \mu^T}{\partial\theta_i}\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial\theta_j}\mathbf{Q}^{-1}(\mathbf{D}-\boldsymbol{\mu}) + (\mathbf{D}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial\theta_i}\mathbf{Q}^{-1}\frac{\partial \mathbf{Q}}{\partial\theta_j}\mathbf{Q}^{-1}(\mathbf{D}-\boldsymbol{\mu})
\end{aligned}
$$

$$-\frac{1}{2}(\mathbf{D}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}\frac{\partial^2\mathbf{Q}}{\partial\theta_i\partial\theta_j}\mathbf{Q}^{-1}(\mathbf{D}-\boldsymbol{\mu})$$

$$+\frac{\partial\boldsymbol{\mu}^T}{\partial\theta_i}\mathbf{Q}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_j}+(\mathbf{D}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}\frac{\partial\mathbf{Q}}{\partial\theta_i}\mathbf{Q}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\theta_j}-(\mathbf{D}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}\frac{\partial^2\boldsymbol{\mu}}{\partial\theta_i\partial\theta_j}. \quad (22)$$

To avoid computing second-order derivatives of $\boldsymbol{\mu}$ and $\mathbf{Q}$ one often approximates the observed FIM by the more popular expected FIM (Kitanidis and Lane 1985)

$$\langle F_{ij}\rangle=\frac{1}{2}\mathrm{Tr}\left(\mathbf{Q}^{-1}\frac{\partial\mathbf{Q}}{\partial\boldsymbol{\theta}_i}\mathbf{Q}^{-1}\frac{\partial\mathbf{Q}}{\partial\boldsymbol{\theta}_j}\right)+\frac{\partial\boldsymbol{\mu}^T}{\partial\boldsymbol{\theta}_i}\mathbf{Q}^{-1}\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}_j}. \quad (23)$$

Equations (22) and (23) yield two different values of *KIC* which we designate below, respectively, by KIC_{obs} and KIC_{exp}.

To set a standard against which the accuracies of *BIC* and *KIC* are measured we evaluate the integrated likelihood in (9) by Markov Chain Monte Carlo (*MCMC*) simulations. Treating the prior distribution of each parameter, $p(\boldsymbol{\theta}_k|M_k)$, as if it was uniform and independent between specified lower and upper bounds, the conventional Metropolis-Hastings algorithm (Hastings 1970) is used to implement the *MCMC* process. Our implementation of the algorithm is similar to that of Rojas et al. (2008, 2010a, 2010b, 2010c); more advanced *MCMC* techniques are discussed by Chib (1995) and Marshall et al. (2004, 2005). The ranges of the uniform distributions are initially set to 50% and 150% of the ML parameter estimates and modified by trial-and-error during the sampling process to obtain an acceptance rate of 20–40%. Three chains are launched and their convergence of *MCMC* is evaluated using the $\hat{R}$ statistics of Gelman et al. (1995). A total of 50,000 samples are generated from each chain; the first 1,000 samples are discarded (burn-in) after $\hat{R}$ is 1.0002 for all the parameters. Random parameter samples from *MCMC* allow calculating a joint negative log-likelihood function according to (19), which in turn is used to approximate the integrated likelihood function (9) numerically via

$$\hat{p}(\mathbf{D}|M_k)=\frac{1}{l}\sum_{i=1}^{l}p\left(\mathbf{D}|\boldsymbol{\theta}_k^{(i)},M_k\right) \quad (24)$$

where l is the number of parameter samples, $p(\mathbf{D}|\boldsymbol{\theta}_k^{(i)},M_k)$ being the joint likelihood function of the ith parameter sample, $\boldsymbol{\theta}_k^{(i)}$. We then adopt the *MCMC* results as reference against which the accuracies of *BIC* and *KIC* are judged. Note that we could not do so for *AIC* and *AICc* because these are not associated with integrated likelihood functions.

3 Results and Discussion

In this section we compare values of *BIC* and *KIC* computed for each of the four alternative drift models and fourteen sample sizes. We also compare values of *KIC* computed using observed and expected FIM, integrated likelihood functions computed by *BIC*, *KIC* and *MCMC*, and posterior model probabilities obtained using the

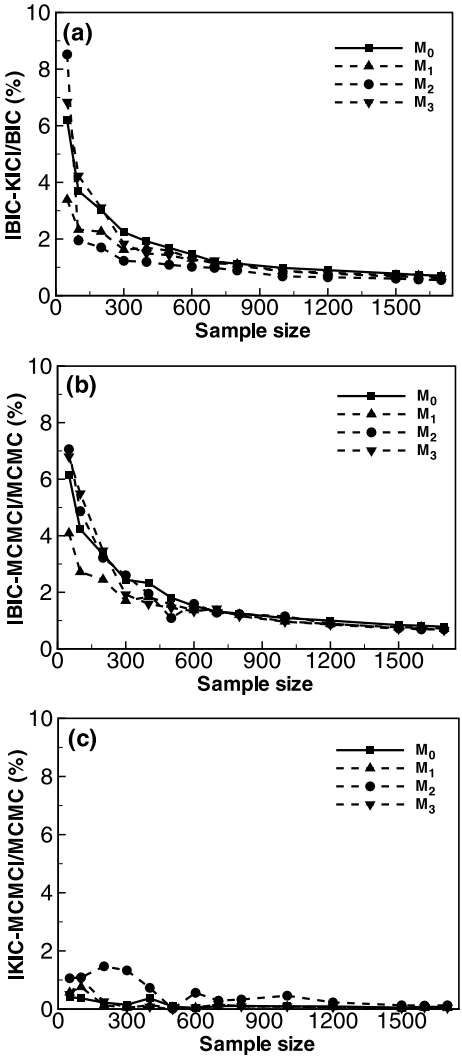

Fig. 3 Percent relative difference between (**a**) *BIC* and KIC_{obs}, (**b**) *BIC* and *MCMC*, and (**c**) KIC_{obs} and *MCMC* for each drift model and sample size based on one sample

Bayesian criteria and *MCMC*. We start by presenting results based on a single sample, followed by those based on 100 replicates.

3.1 Accuracies of *BIC* and *KIC*

Figure 3(a) plots percent relative difference, $(BIC - KIC_{\text{obs}})/BIC$, between *BIC* and KIC_{obs} (computed using observed FIM) corresponding to four drift models and fourteen sizes of a single sample. Figures 3(b) and 3(c) do the same for $(BIC - MCMC)/MCMC$ and $(KIC_{\text{obs}} - MCMC)/MCMC$, respectively. Figure 3(a) shows that *BIC* and KIC_{obs} differ from each other by up to about 9% of *BIC* when sample size is small ($N = 50$), the difference diminishing asymptotically to be-

Table 1 Posterior model probabilities (%) of the three alternative models obtained using *BIC*, KIC_{obs} and *MCMC* for $N = 50$, 100, 300, 500 based on one sample. The largest model probabilities are in bold

N	*BIC*			KIC_{obs}			*MCMC*		
	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
50	**66.84**	0.30	32.86	12.71	3.98	**83.30**	24.24	1.11	**74.65**
100	**96.69**	0.26	3.05	41.83	0.27	**57.90**	34.29	0.02	**65.69**
300	**63.18**	0.00	36.82	44.15	0.00	**55.85**	43.41	0.01	**56.58**
500	43.29	0.00	**56.71**	**65.70**	0.00	34.30	**58.41**	0.02	41.57

low 1% with sample size (confirming that KIC_{obs} reduces asymptotically to *BIC*). Figures 3(b) and 3(c) show that, using *MCMC* results as reference, *BIC* provides a less accurate approximation of integrated likelihood than does KIC_{obs}. Whereas the quality of *BIC* improves with sample size, its relative deviation from the reference does not fall much below 1% regardless of sample size. The corresponding deviation of KIC_{obs} is consistently less than 1%, becoming much smaller when N is large. The same is true for KIC_{exp} (computed using expected FIM) even though KIC_{exp} is less accurate than KIC_{obs} (as shown in Sect. 3.2). Small differences between *BIC* and KIC_{obs} may result in large differences between corresponding posterior model probabilities by virtue of their exponentiation in (10) and (11). Table 1 demonstrates that the less accurate *BIC* values lead to inaccurate posterior model probabilities and result in the selection of inferior models. In Table 1, *BIC* is seen to assign the largest posterior probability to an inferior model when $N = 50$, 100, 300 and 500. In contrast, model probabilities based on KIC_{obs} are similar to those based on *MCMC* regardless of sample size.

Figure 4 summarizes via box plots several statistics of *BIC*, KIC_{obs}, *MCMC* and model probabilities for 100 replicate samples of size $N = 100$ including median, lower and upper quartiles, minimum and maximum. Whereas the statistics of KIC_{obs} and *MCMC* are quite similar, those of *BIC* are significantly different, as was the case for a single sample in Table 1. With reference to *MCMC*, KIC_{obs} evidently approximates the integrated likelihood more accurately than does *BIC*. This observation is further discussed in Sect. 3.4 below. The wide spread of model probabilities in Figs. 4(e) and 4(f) is due to fluctuations in model calibration data. This implies some ambiguity in model selection. For example, whereas KIC_{obs} and *MCMC* in Fig. 4(f) prefer model M_3 in more than 75% of the replicates (model probabilities corresponding to the lower, 25th percentile, quartiles exceed 60%), in some other cases they prefer model M_1. *BIC*, on the other hand, prefers model M_3 in about 50% of the replicates (the median indicates that at least 50% of probability values are close to 60%). When ambiguity associated with objective model selection criteria is large, one may wish to supplement these criteria with subjective expert judgment (Ye et al. 2005, 2008b) or to collect more data and re-evaluate model probabilities based on these. The potential benefit of collecting additional data is discussed, in the context of Bayesian model averaging, by Neuman et al. (2011).

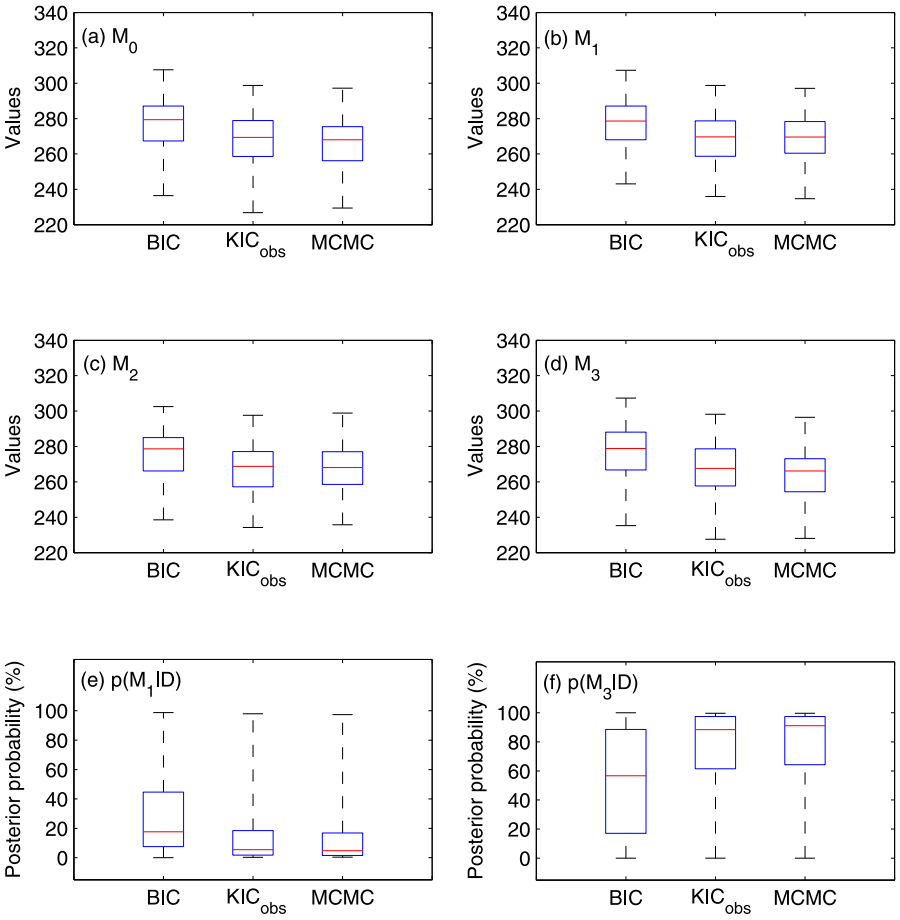

Fig. 4 Box plots of (**a**)–(**d**) *BIC*, KIC_{obs} and *MCMC* values for four models and (**e**)–(**f**) posterior probabilities of models M_1 and M_3 for 100 replicate samples of size $N = 100$

3.2 Differences between Observed and Expected Fisher Information Matrices (FIM)

Figure 5 plots differences between $\ln|\mathbf{F}_k|$ values based on observed and expected FIM for each drift model and size of a single sample according to (22) and (23), respectively. Although observed and expected FIM can be validly interchanged for large i.i.d. (independent and identically distributed) samples, Fig. 5 indicates that, in our case of correlated data, the two may differ substantially when sample size is small ($N < 700$). Due to exponentiation, such differences translate into large differences between posterior model probabilities, as illustrated for M_1 in Fig. 6a plotted for one sample with different sample sizes; since the posterior probability of M_2 is close to zero, results corresponding to M_3 are similar. One notes in Fig. 6a that absolute differences between posterior probabilities based on KIC_{exp} and the *MCMC* reference are consistently larger than those based on KIC_{obs}. Hence observed FIM leads to more accurate estimation of model probabilities than does expected FIM. It follows

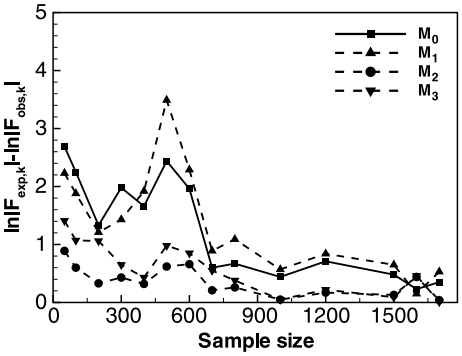

Fig. 5 Differences between $\ln|\mathbf{F}_k|$ values based on expected and observed Fisher information matrices (FIM) for each drift model and size of a single sample

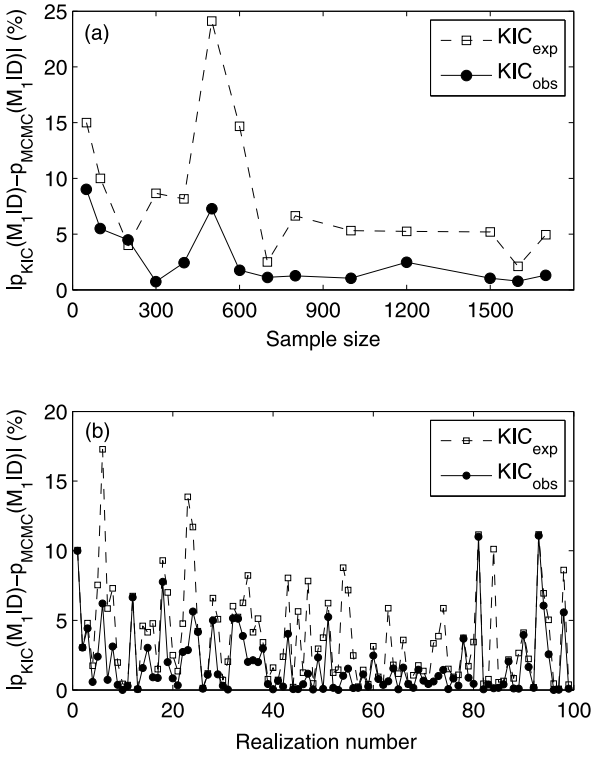

Fig. 6 Absolute differences between posterior probabilities of M_1 based on KIC_{obs} and KIC_{exp} relative to *MCMC* for (**a**) each size of a single sample and (**b**) 100 replicate samples of size $N = 100$. As posterior probability of M_2 is almost zero, absolute differences corresponding to M_3 are similar

that KIC_{obs} and KIC_{exp} may, in some cases, prefer different models as observed in the numerical experiments when $N = 500$ and 1200. The poor performance of KIC_{exp} relative to KIC_{obs} stems from the following assumption behind (23) (Kitanidis and Lane 1985),

$$\begin{aligned} E[\mathbf{D} - \boldsymbol{\mu}] &= \mathbf{0} \\ E\left[(\mathbf{D} - \boldsymbol{\mu})(\mathbf{D} - \boldsymbol{\mu})^T\right] &= \mathbf{Q} \end{aligned} \tag{25}$$

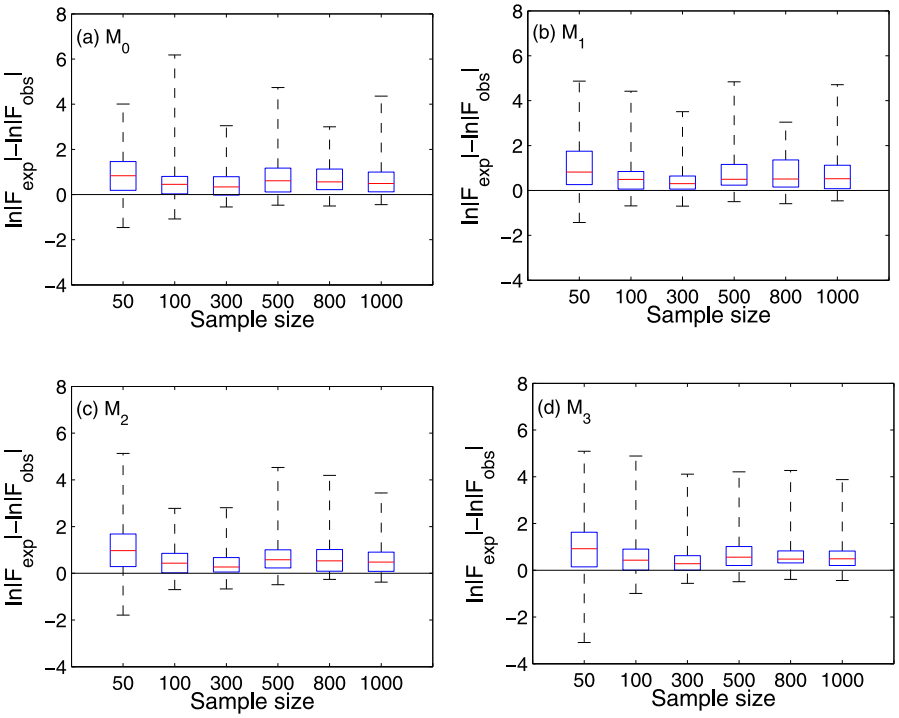

Fig. 7 Box plots of differences between $\ln|\mathbf{F}_k|$ values based on expected and observed Fisher information matrices (FIM) for 100 replicates of various size samples

that $\boldsymbol{\mu}$ is an unbiased estimate of $\mathbf{D}$ and that $\mathbf{Q}$ represents the covariance of the residuals exactly. In reality, Q is uncertain due to lack of certainty about the functional form and parameters of $\boldsymbol{\mu}$ as well as the parameters of the corresponding variogram. It suggests that the higher accuracy of observed FIM is general for any samples. This is confirmed in Fig. 6b that plots the absolute differences for the 100 replicate samples of size $N = 100$. For all the 100 replicates, the absolute differences based on KIC_{exp} are consistently larger than those based on KIC_{obs}; the deviation is small for some replicates. This indicates that KIC_{obs} is more accurate than KIC_{exp} over the replicate samples.

To investigate the difference between observed and expected FIM under data uncertainty, $\ln|\mathbf{F}_k|$ values based on observed and expected FIM are evaluated for 100 replicates of samples including $N = 50$, 100, 300, 500, 800 and 1,000 data points. The differences are summarized statistically with the aid of box diagrams in Fig. 7. Median differences are seen to be nearly zero and interquartile ranges (IQR) to be smaller than 2 for all sample sizes. Actual differences may be substantial and lie within relatively large ranges that do not decrease in any consistent manner as sample size increases.

3.3 Randomness of Observed and Expected Fisher Information Matrices (FIM)

The large ranges depicted in Fig. 7 between observed and expected FIM stem from the random natures of calibration data and associated parameter estimates. The observed FIM is evidently random due to its dependence on random calibration data **D**. The expected FIM is theoretically nonrandom but its approximation, obtained by calibrating each model against each random sample and then applying (23), remains random due to limited sample size and estimation errors associated with corresponding parameter estimates. To our knowledge, the effect of randomness in observed and expected FIM has not been previously analyzed in the manner we do here.

Effects of data and parameter estimation uncertainty are illustrated in Figs. 8(a)–8(d) where $\ln|\mathbf{F}_k|$ associated with expected and observed FIM based on a single sample are plotted against sample size. Fluctuation in Fig. 8(a) reflects uncertainty in $\ln|\mathbf{F}_k|$ as computed on the basis of (23). This uncertainty is removed for the true model in Fig. 8(b) (the curve becoming smooth) by computing $\ln|\mathbf{F}_k|$ using true drift and variogram parameters. Figures 8(c) and 8(d) are similar to Figs. 8(a) and 8(b) but correspond to observed FIM. Even though $\ln|\mathbf{F}_k|$ in Fig. 8(d) is computed using true drift model with true drift and variogram parameters, it nevertheless fluctuates. This is due to data uncertainty which does not impact Fig. 8(b). Averaging out data uncertainty as described below leads to a smooth result in Fig. 8(d). Comparing Figs. 8(a) and 8(c) with Figs. 8(b) and 8(d), respectively, shows that effects of data and parameter estimation uncertainty on the estimation of FIM is large. In addition, the effect of data uncertainty is larger than that of parameter estimation uncertainty. When fluctuations in model calibration data cause ambiguity in model selection, one may wish to rely on expert judgment or collect additional data as discussed in Sect. 3.1. Figures 8(e) and 8(f) depict $\ln|\mathbf{F}_k|$ based on expected and observed FIM averaged over 100 replicate samples for various sample sizes. Comparing the two figures with their counterparts in Figs. 8(a) and 8(c) confirms that averaging reduces randomness in data and parameter estimates. More importantly, the averaged $\ln|\mathbf{F}_k|$ in Figs. 8(e) and 8(f) are almost identical. This does not conflict with the implication of Fig. 7 that KIC_{obs} is more accurate than KIC_{exp} for one sample. Instead, it suggests that averaging over many replicates reduces differences between observed and expected *KIC*.

3.4 Role of FIM in Model Selection

Figure 8 shows that $\ln|\mathbf{F}_1| > \ln|\mathbf{F}_2| > \ln|\mathbf{F}_0| > \ln|\mathbf{F}_3|$ regardless of sample size. This consistent order reflects a tendency of $\ln|\mathbf{F}_k|$, and hence *KIC*, to impose increasing penalty on models having greater complexity (number of parameters). Although model M_1 contains more information per datum (as measured by $\ln|\mathbf{F}_k|$) than does model M_3, its goodness-of-fit measure (see next section) is only slightly better. Ye et al. (2010b) pointed out that "all else being equal, if increasing the expected information content of a model fails to improve its performance relative to another model, then selecting a model with greater expected information content would, according to *KIC*, be unjustified". This is supported by the fact that *KIC* favor M_3 in

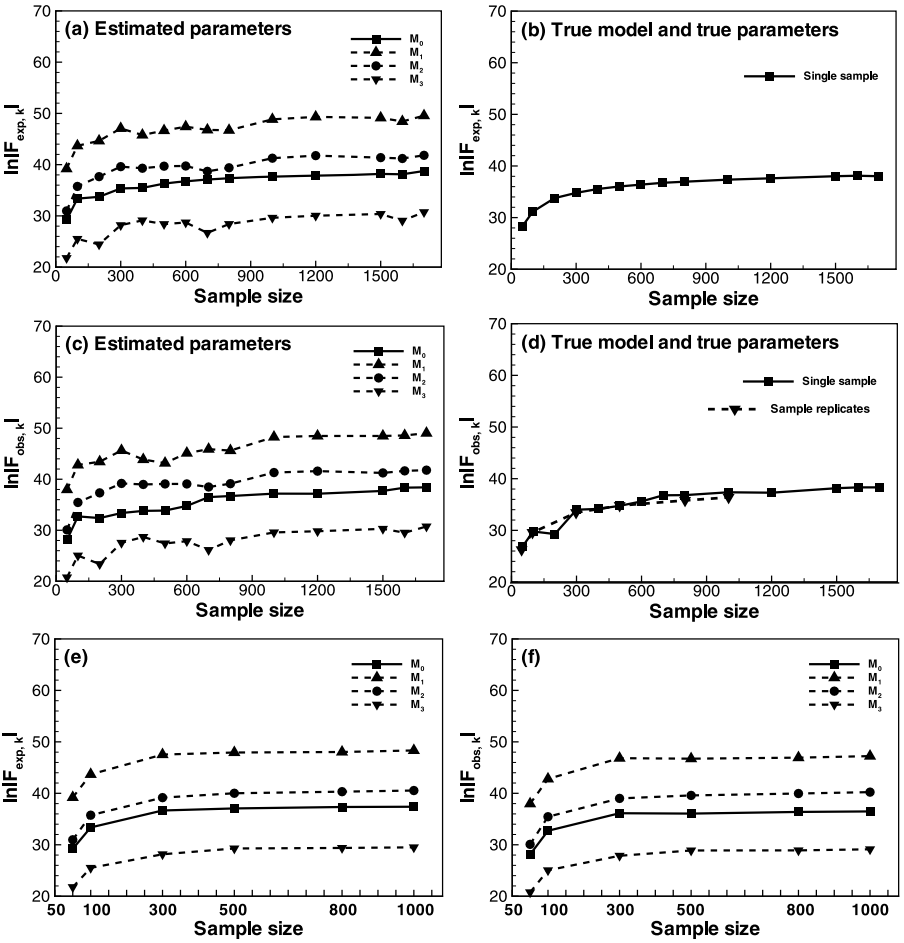

Fig. 8 Values of $\ln|\mathbf{F}_k|$ associated with approximation of expected FIM based on (**a**) variogram parameters corresponding to each model estimated against each random sample and (**b**) known parameters corresponding to true model, as functions of sample size. (**c**) and (**d**) show the same for observed FIM; *dashed curve* in (**d**) is average of *solid curve* over 100 sample replicates. (**e**) and (**f**) correspond to (**a**) and (**c**), respectively, but are based on 100 sample replicates

Figs. 4(e) and 4(f). The order $\ln|\mathbf{F}_2| > \ln|\mathbf{F}_0|$ of two models having an equal number of parameters demonstrates the ability of *KIC* to differentiate between disparate model structures regardless of sample size. This ability is not shared by other model selection criteria, whether *BIC* or information theoretic. Figure 8 shows that information content per datum increases faster with N when the latter is small than when it is large, indicating a diminishing incremental gain in information with sample size. This suggests that FIM is able to incorporate data correlation into model selection, another unique feature of *KIC* not shared by *BIC* or information theoretic criteria that do not include FIM.

3.5 Consistency and Asymptotic Behavior of *BIC* and *KIC*

BIC (and, by implication, *KIC*) is known to be theoretically consistent in that the associated probability of selecting the true model when the latter is included in the set of alternative models converges to one with sample size N. Figure 9 plots fractions of 100 replicates among which *BIC* and KIC_{obs} select the true model versus sample size. Whereas in the case of KIC_{obs} this fraction reaches 1 at $N = 800$, in the case of *BIC* the fraction remains less than 1 at $N = 1{,}000$. Hence in our example KIC_{obs} is seen to identify the true model faster than does *BIC*, being therefore more consistent.

As noted by Ye et al. (2008a), the components of *KIC* in (4) exhibit asymptotic behaviors of the following orders as N approaches infinity: $-2\ln p(\boldsymbol{\theta}|M_k)$ is $O(N_k)$, $-N_k \ln 2\pi$ is $O(N_k)$, $\ln|\bar{\mathbf{F}}_k|$ is $O(\ln N_k)$, $N_k \ln N$ is $O(\ln N)$ and $-2\ln p(\mathbf{D}|\boldsymbol{\theta}, M_k)$ is $O(N)$. Figure 10 verifies that $-2\ln p(\mathbf{D}|\boldsymbol{\theta}, M_k)$ is indeed asymptotically linear in sample size, N, based on either a single sample (Fig. 10(a)) or 100 replicates (Fig. 10(c)). Similarly, Figs. 10(b) and 10(d) demonstrate that, when N is large, $\ln|\bar{\mathbf{F}}_k|$ (based on observed FIM) is negligibly small in comparison to $-2\ln p(\mathbf{D}|\boldsymbol{\theta}, M_k)$. Hence terms of $O(N)$ dominate and KIC_{obs} approaches *BIC*. Figure 10d is smoother than Fig. 10b because averaging over 100 replicates smoothes out fluctuations caused by randomness of data. Even though disregarding terms of $O(N_k)$ or lesser order results in an error of $O(1)$ that does not vanish regardless of how large N is, Raftery (1995) and Tsai and Li (2010) argue that this error does not affect posterior model probability and thus model selection. Ye et al. (2010b) point out that, theoretically, this is not the case: ignoring terms of $O(N_k)$ or lesser order is not appropriate because model selection depends on differences between the *KIC* (or *BIC*) values of different models, not on the actual values of these criteria. Their point is demonstrated numerically below.

Figure 11 plots differences between values of *BIC*, KIC_{obs}, KIC_{exp}, and their components corresponding to models M_1 and M_3, respectively, as functions of sample size for one and 100 replicate samples. Model M_2 is not considered as its probability is close to zero. Figures 11(a1) and 11(a2) show that differences between values of $-2\ln p(\mathbf{D}|\boldsymbol{\theta}, M_k)$ corresponding to model M_1 with 5 parameters and model M_3 with

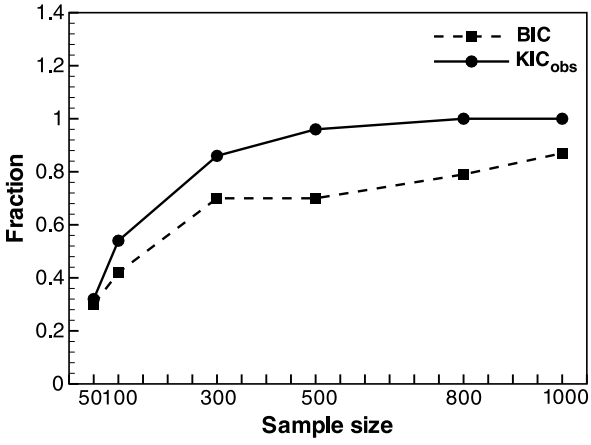

Fig. 9 Fractions of 100 replicates among which *BIC* and KIC_{obs} select the true model versus sample size

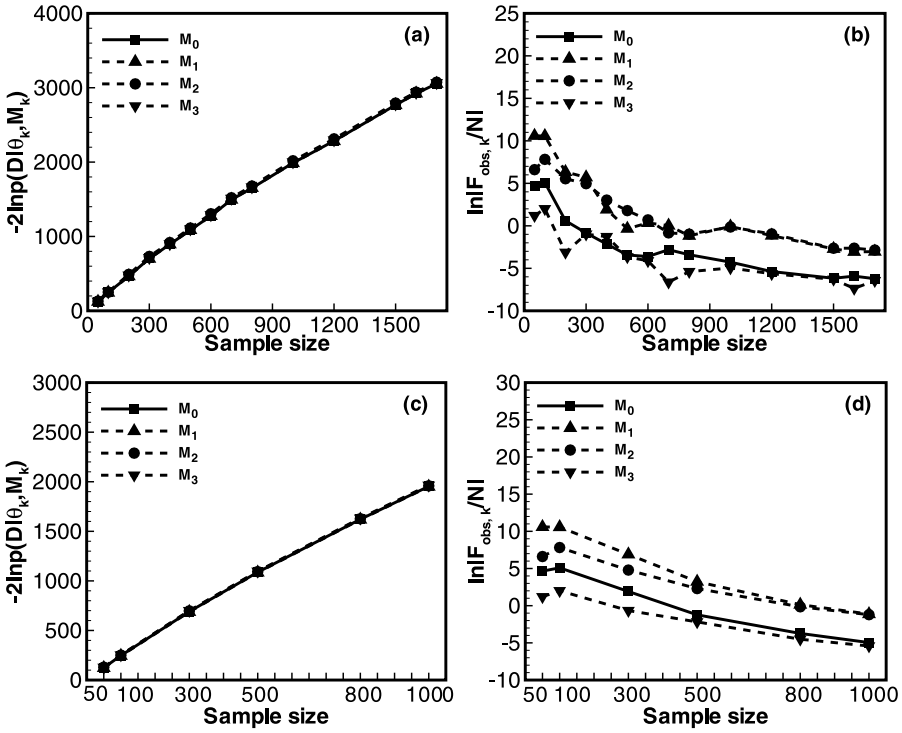

Fig. 10 Variations of (**a**) $-2\ln p(\mathbf{D}|\boldsymbol{\theta}, M_k)$ and (**b**) $\ln|\bar{\mathbf{F}}_k|$ observed with sample size for single sample. (**c**) and (**d**) are corresponding averages, respectively, over 100 replicates

3 parameters are negative and become more so as sample size increases. This is so because model M_1 can be fitted to data more closely than model M_3. Differences between values of $N_k \ln N$ corresponding to these two models increases monotonically with sample size, reflecting an increasing penalty for model complexity with sample size. *BIC* is seen to prefer model M_1 when the $-2\ln p(\mathbf{D}|\hat{\boldsymbol{\theta}}_k, M_k)$ difference exceeds the $N_k \ln N$ difference and model M_3 when the opposite is true.

Figures 11(b1) and 11(b2) plot differences between values of *KIC* and its three components (*BIC*, $-\ln p(\hat{\boldsymbol{\theta}}_k) - N_k \ln(2\pi)$ and $\ln|\bar{\mathbf{F}}_k| = \ln|\mathbf{F}_k/N|$) corresponding to models M_1 and M_3, respectively, versus sample size based on observed FIM; Figs. 11(c1) and 11(c2) do the same based on expected FIM. *BIC* and *KIC* prefer different models when the differences between their values, corresponding to the two models, have opposite signs. Differences in values of $-\ln p(\hat{\boldsymbol{\theta}}_k) - N_k \ln(2\pi)$ are negative (due to two extra parameters in M_1) and constant. The role played by $-\ln p(\hat{\boldsymbol{\theta}}_k) - N_k \ln(2\pi)$ in model selection is discussed by Neath and Cavanaugh (1997). Since differences associated with $-\ln p(\hat{\boldsymbol{\theta}}_k) - N_k \ln(2\pi)$ are constant while those associated with $\ln|\mathbf{F}_k/N|$ decrease with sample size, the signs of differences associated with *KIC* are controlled asymptotically by those associated with $\ln|\mathbf{F}_k/N|$. As $\ln|\mathbf{F}_{\text{obs},k}/N|$ differs from $\ln|\mathbf{F}_{\text{exp},k}/N|$, KIC_{obs} and KIC_{exp} may prefer different models. In the cases of a single sample, KIC_{obs} and KIC_{exp} are seen to prefer dif-

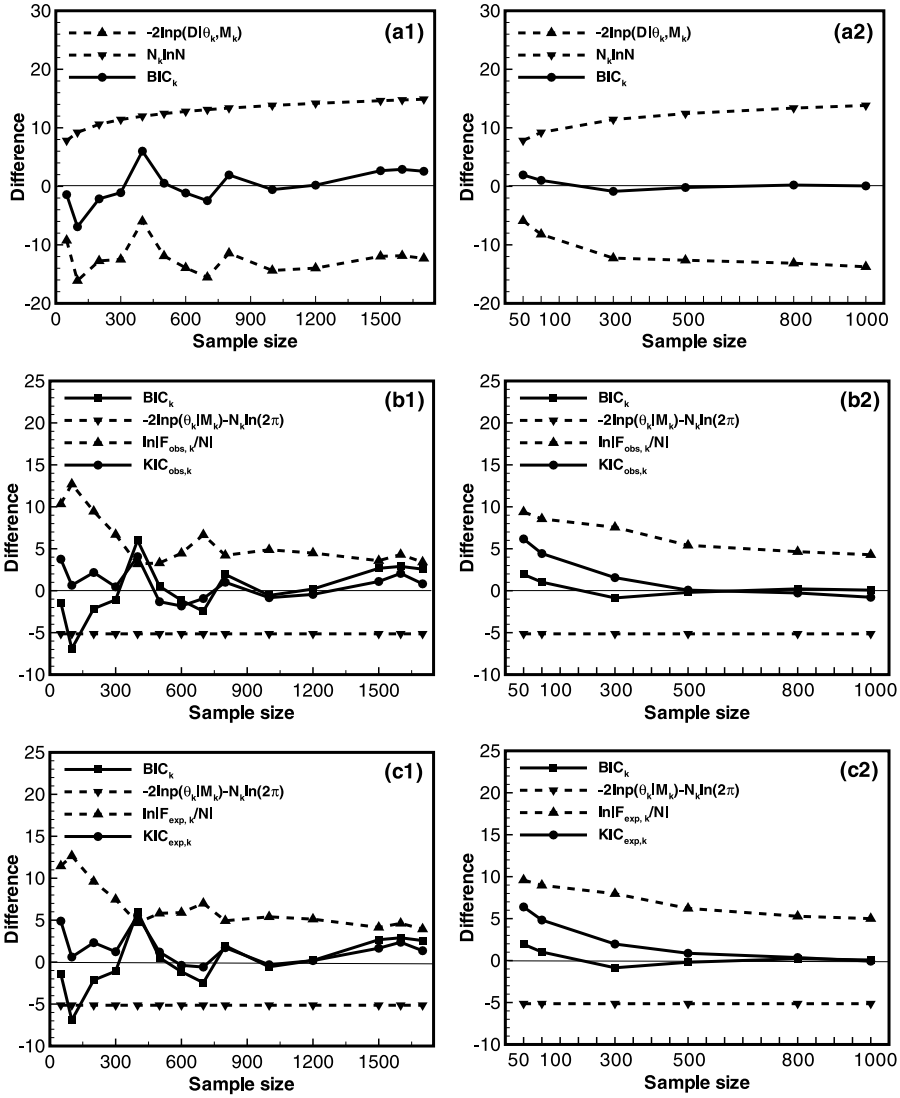

Fig. 11 Differences between values of (**a1**)–(**a2**) *BIC*, (**b1**)–(**b2**) KIC_{obs}, (**c1**)–(**c2**) KIC_{exp} and their components corresponding to models M_1 and M_3, respectively, as functions of sample size for single sample (*left*) and 100 replicates (*right*)

ferent models when $N = 500$ and 1200 (Figs. 11(b1) and 11(c1)); in the case of 100 replicates, this happens when $N = 800$ (Figs. 11(b2) and 11(c2)).

3.6 Sensitivity of Posterior to Prior Model Probability

Ye et al. (2005) found the sensitivity of posterior to prior model probability for a fixed data set to be significant, especially for models with large likelihoods. Here

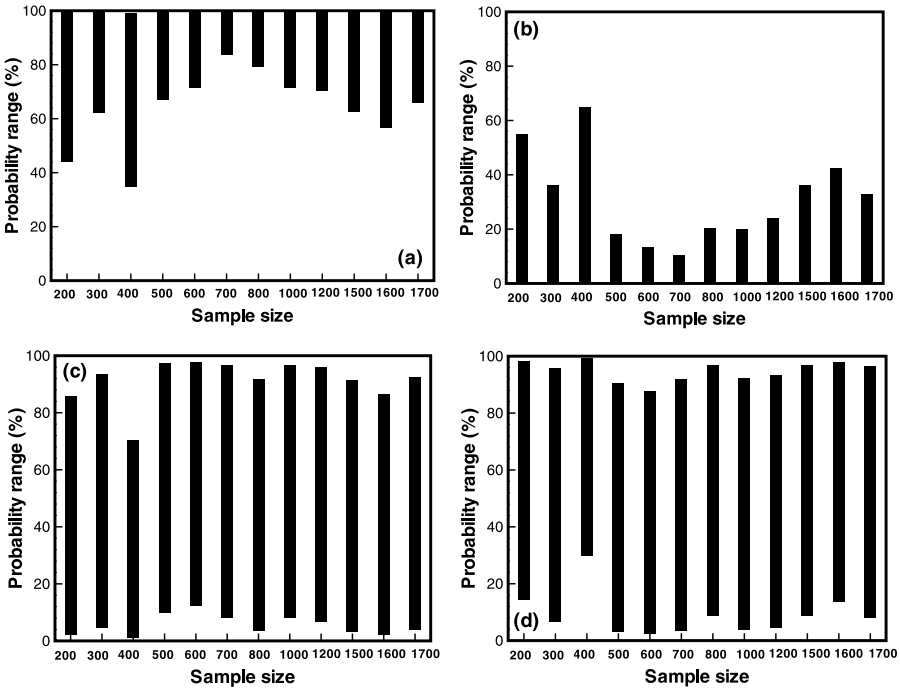

Fig. 12 Ranges of KIC_{obs}-based posterior probabilities for models (**a**) M_0 and (**b**) M_3 when M_0 is included and for (**c**) M_1 and (**d**) M_3 when M_0 is excluded

we explore, based on a single sample, the manner in which this sensitivity varies with sample size when the true model M_0 is included and excluded from the set of alternatives. Figures 12(a) and 12(b) plot ranges of KIC_{obs}-based posterior probabilities for models M_0 and M_3 corresponding to all twelve sample sizes when M_0 is included; Figs. 12(c) and 12(d) do so for models M_1 and M_3 when M_0 is excluded. Theory indicates that, for i.i.d. data, the sensitivity of posterior to prior model probability decreases asymptotically as sample size increases (Berk 1966; Bernardo and Smith 1994). Figure 12 does not display this type of behavior. According to Bernardo and Smith (1994), in the case of i.i.d. data, the sensitivity of posterior to prior model probabilities vanishes as $N \to \infty$ because the log ratio $\ln[p(\mathbf{D}|M_k)/p(\mathbf{D}|M_0)]$ between the likelihoods of any alternative model and the true model tends to negative infinity. Figure 13(a) plots this log ratio for all three alternative drift models in our example when the likelihood function is approximated using observed FIM and the true model, M_0, is included. The log ratios do not vary much with sample size; one notes a correlation between the graph corresponding to M_3 in Fig. 13(a) and the bar graph in Fig. 12(b). Figure 13(b) plots log ratios between the likelihood of M_1 and M_3 and between those of M_2 and M_3 for the case where M_0 is excluded. The graphs correlate with the corresponding bar graphs in Figs. 12(b) and 12(c), respectively, showing no reduction in sensitivity with increasing sample size.

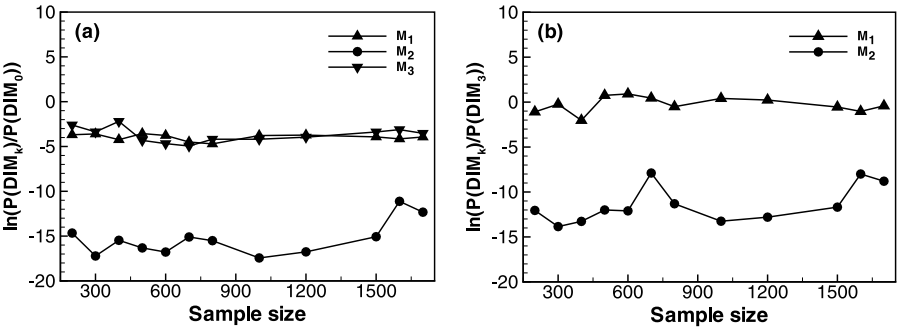

Fig. 13 Logarithms of KIC_{obs}-based likelihood ratios (**a**) between alternative models and true model when M_0 is included, and (**b**) between models M_1 and M_3 when M_0 is excluded

4 Conclusions

Geostatistical analysis requires estimating the covariance structure of a random field and its parameters jointly from data. In many cases, one must consider a discrete set of structural (drift and variogram) model alternatives. Ranking these alternatives and identifying the best among them has traditionally been done with the aid of information theoretic or Bayesian model selection (discrimination, information) criteria. There is an ongoing debate in the literature about the relative merits of these various criteria. We have contributed to this discussion by using synthetic data to compare the abilities of two common Bayesian criteria, *BIC* and *KIC,* to discriminate between alternative models of drift as a function of sample size when drift and variogram parameters are unknown. The results of this study are equally valid for one and multiple realizations of uncertain data entering into our analysis. Our analysis is based on Bayesian statistics and does not include information criteria derived using other principles.

It has been shown that using *MCMC* results as a reference, *KIC* yields more accurate approximations of integrated likelihood and posterior model probability than does *BIC*. Although *KIC* reduces asymptotically to *BIC*, *KIC* provides consistently more reliable indications of model quality for a range of sample sizes; *BIC* selects inferior models more often than does *KIC*. We have demonstrated that the Fisher information term in *KIC* allows (a) imposing a more severe penalty on models having greater complexity (number of parameters) and (b) differentiating more accurately between models of disparate structures than is possible with other criteria that do not include a Fisher term. In the case of correlated data, information content per datum (as measured by $\ln|\mathbf{F}_k|$) increases faster with sample size when the latter is small than when it is large. This implies a diminishing incremental gain in information content with sample size. Theory indicates that, for i.i.d. data, the sensitivity of posterior to prior model probability decreases asymptotically as sample size increases. This does not happen in our example, possibly due to autocorrelation between our synthetically generated data.

There are two variants of *KIC* evaluated using the observed and expected FIM. KIC_{obs} based on observed FIM is more reliable than KIC_{exp} based on an approximation of expected FIM, the difference between the two increasing with diminishing

sample size. Computing these using multiple replicate samples reduces the difference between KIC_{obs} and KIC_{exp}. Difference in KIC_{obs} and KIC_{exp} causes the difference in their corresponding model probabilities. KIC_{obs}-based estimates of integrated model likelihood and model probability are close to but not identical with corresponding *MCMC*-based estimates. However, *MCMC* simulation may require a much larger computational effort.

Acknowledgements The authors thank David Draper and Bruno Mendes for their helpful comments and advices. The first two authors were supported in part by NSF-EAR grant 0911074 and DOE-SBR grant DE-SC0002687. The third author was supported in part by the US Department of Energy through a contract between Vanderbilt University and the University of Arizona under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III.

References

Akaike H (1974) New look at statistical model identification. IEEE Trans Autom Control AC-19:716–722
Berk R (1966) Limiting behavior of posterior distributions when the model is incorrect. Ann Math Stat 37:51–58
Bernardo JM, Smith AFM (1994) Bayesian theory. Wiley, Chichester
Burnham KP, Anderson DR (2002) Model selection and multiple model inference: a practical information-theoretical approach, 2nd edn. Springer, New York
Burnham KP, Anderson DR (2004) Multimodel inference—understanding AIC and BIC in model selection. Sociol Methods Res 33(2):261–304 conditions: 3. Application to synthetic and field data. Water Resour Res 22(2):228–242
Chib S (1995) Marginal likelihood from the Gibbs output. J Am Stat Assoc 90:1313–1321
Cressie N (1993) Statistics of spatial data. Wiley, New York
Deutsch CV, Journel AG (1998) GSLIB: Geostatistical software library and user's guide, 2nd edn. Oxford Univ Press, New York
Draper D (1995) Assessment and propagation of model uncertainty. J R Stat Soc B 57(1):45–97
Draper D (2007) Bayesian multilevel analysis and MCMC. In: de Leeuw J (ed) Handbook of Quantitative Multilevel Analysis. Springer, New York, Chapter 3
Gelman A, Carlin JB, Stern HS, Rubin DB (1995) Bayesian data analysis, 1st edn. Chapman & Hall, USA
Hastings W (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57(1):97–109
Hoeksema RJ, Kitanidis PK (1985) Analysis of the spatial structure of properties of selected aquifers. Water Resour Res 21(4):563–572
Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. Stat Sci 14(4):382–417
Hurvich CM, Tsai CL (1989) Regression and time series model selection in small sample. Biometrika 76(2):99–104
Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University Press, Oxford
Journel AG, Rossi ME (1989) When do we need a trend model in kriging? Math Geol 21(7):715–739
Kashyap RL (1982) Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Trans Pattern Anal Mach Intell 4(2):99–104
Kass RE, Vaidyanathan SK (1992) Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. J R Stat Soc, Ser B, Stat Methodol 54(1):129–144
Kass RE, Raftery AE (1995) Bayesian factor. J Am Stat Assoc 90:773–795
Kitanidis PK, Lane RW (1985) Maximum likelihood parameter estimation of hydrologic spatial processes by the Gaussian-Newton method. J Hydrodyn 79:53–71
Kyriakidis PC, Journel AG (1999) Geostatistical space–time models: a review. Math Geol 31(6):651–684
Lewis SM, Raftery AE (1997) Estimating Bayes factor via posterior simulation with the Laplace–Metropolis estimator. J Am Stat Assoc 92(438):648–655
Leuangthong O, Deutsch CV (2004) Transformation of residuals to avoid artifacts in geostatistical modeling with a trend. Math Geol 36(3):287–305
Marchant BP, Lark RM (2004) Estimating variogram uncertainty. Math Geol 36(8):867–898

Marchant BP, Lark RM (2007) The Matern variogram model: implications for uncertainty propagation and sampling in geostatistical surveys. Geoderma 140:337–345
Marshall L, Nott D, Sharma A (2004) A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling. Water Resour Res 40:W02501. doi:10.1029/2003WR002378
Marshall L, Nott D, Sharma A (2005) Hydrological model selection: A Bayesian alternative. Water Resour Res 41:W10422. doi:10.1029/2004WR003719
Matérn B (1986) Spatial variation. Springer, Berlin
McBratney AB, Webster R (1986) Choosing functions for semi-variogram of soil properties and fitting them to sampling estimates. J Soil Sci 37:617–639
Mosteller F, Wallace DL (1964) Inference and disputed authorship: the federalist. Addison-Wesley, Reading
Neath AA, Cavanaugh JE (1997) Regression and time series model selection using variants of the Schwarz information criterion. Commun Stat, Theory Methods 26:559–580
Neuman SP (2003) Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. Stoch Environ Res Risk Assess 17(5):291–305
Neuman SP, Xue L, Ye M, Lu D (2011) Bayesian analysis of data-worth considering model and parameter uncertainties. Adv Water Resour doi:10.1016/j.advwaters.2011.02007
Nowak W (2010) Measures of parameter uncertainty in geostatistical estimation and geostatistical optimal design. Math Geosci 42(2):199–221
Nowak W, de Barros FPJ, Rubin Y (2010) Bayesian geostatistical design: task—driven optimal site investigation when the geostatistical model is uncertain. Water Resour Res 46:W03535. doi:10.1029/2009WR008312
Ortiz CJ, Deutsch CV (2002) Calculation of uncertainty in the variogram. Math Geol 34(2):169–183
Pardo-Iguzquiza E, Dowd P (2001) Variance-covariance matrix of the experimental variogram: assessing variogram uncertainty. Math Geol 33(4):397–419
Pardo-Iguzquiza E, Chico-Olmo M, Garcia-Soldado MJ, Luque-Espinar JA (2009) Using semivariogram parameter uncertainty in hydrogeological applications. Ground Water 47(1):25–34
Poeter EP, Anderson DA (2005) Multimodel ranking and inference in ground water modeling. Ground Water 43(4):597–605
Poeter EP, Hill MC (2007) MMA, A computer code for multi-model analysis. US Geological Survey Techniques and Methods TM6-E3
Raftery AE (1995) Bayesian model selection in social research. Sociol Method 25:111–163
Riva M, Willmann M (2009) Impact of log-transmissivity variogram structure on groundwater flow and transport predictions. Adv Water Resour 32:1311–1322
Riva M, Panzeri M, Guadagnini A, Neuman SP (2011) Role of model selection criteria in geostatistical inverse estimation of statistical data- and model-parameters. Water Resour Res 47:W07502. doi:10.1029/2011WR010480
Rissanen J (1978) Modeling by shortest data description. Automatica 14:465–471
Rojas R, Feyen L, Dassargues A (2008) Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. Water Resour Res 44:W12418. doi:10.1029/2008WR006908
Rojas R, Batelaan O, Feyen L, Dassargues A (2010a) Assessment of conceptual model uncertainty for the regional aquifer Pampa del Tamarugal–North Chile. Hydrol Earth Syst Sci 14(2):171–192
Rojas R, Kahunde S, Peeters L, Batelaan O, Feyen L, Dassargues A (2010b) Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling. J Hydrol 394(3–4):416–435
Rojas R, Feyen L, Batelaan O, Dassargues A (2010c) On the value of conditioning data to reduce conceptual model uncertainty in groundwater modeling. Water Resour Res 46:W08520. doi:10.1029/2009WR008822
Samper FJ, Neuman SP (1989a) Estimation of spatial covariance structures by adjoint state maximum likelihood cross-validation: 1. Theory. Water Resour Res 25(3):351–362
Samper FJ, Neuman SP (1989b) Estimation of spatial covariance structures by adjoint state maximum likelihood cross-validation: 2. Synthetic experiments. Water Resour Res 25(3):363–371
Singh A, Walker DD, Minsker BS, Valocchi AJ (2010) Incorporating subjective and stochastic uncertainty in an interactive multi-objective groundwater calibration framework. Stoch Environ Res Risk Assess. doi:10.1007/s00477-010-0384-1
Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464

Tsai FTC, Li X (2010) Reply to comment by Ming Ye et al. on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window". Water Resour Res 46:W02802. doi:10.1029/2009WR008591

Tsai FTC, Li X (2008a) Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window. Water Resour Res 44:W09434. doi:10.1029/2007WR006576

Tsai FTC, Li X (2008b) Multiple parameterization for hydraulic conductivity identification. Ground Water 46(6):851–864

Ye M, Neuman SP, Meyer PD (2004) Maximum Likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour Res 40:W05113. doi:10.1029/2003WR002557

Ye M, Neuman SP, Meyer PD, Pohlmann KF (2005) Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. Water Resour Res 41:W12429. doi:10.1029/2005WR004260

Ye M, Meyer PD, Neuman SP (2008a) On model selection criteria in multimodel analysis. Water Resour Res 44:W03428. doi:10.1029/2008WR006803

Ye M, Pohlmann KF, Chapman JB (2008b) Expert elicitation of recharge model probabilities for the Death Valley regional flow system. J Hydrol 354:102–115. doi:10.1016/j.jhydrol.2008.03.001

Ye M, Pohlmann KF, Chapman JB, Pohll GM, Reeves DM (2010a) A model-averaging method for assessing groundwater conceptual model uncertainty. Ground Water. doi:10.1111/j.1745-6584.2009.00633.x

Ye M, Lu D, Neuman SP, Meyer PD (2010b) Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li. Water Resour Res 46:W02801. doi:10.1029/2009WR008501