Hydrologic model complexity depends on the magnitude of model parameters

Saket Pande

Delft University of Technology, Netherlands

s.pande@tudelft.nl

Abstract

Abstract: Three measures of complexities are computed for SIXPAR model on a MOPEX basin data set. One of the measures is based on an approximation of marginal likelihood function, while the other is based on measuring stability of system representation by SIXPAR. Results suggest that the 2 measures are equivalent, indicating that if the Bayesian measure is a valid measure of model complexity then the other is a valid measures of complexity as well. Results further suggest that SIXPAR complexity varies with the magnitude of its parameters.

1. Definitions

We define the set of observations of a variable of prediction interest such as streamflow by a vector $\vec{y}^0 = \{y^0(1), y^0(2), \ldots, y^0(N)\}$. Here N is the data size or the number of observations. It represents one realization of observations. Similarly, let forcing be represented by $\vec{x} = (x_1, x_2, \ldots, x_N)$ where x_1 may not be univariate, though assumed here to be univariate for simplicity. Let a model be represented by a parameter set α that for given forcing $\vec{x}$ simulates $\vec{y}(\vec{x};\alpha) = (y(1,\vec{x};\alpha), y(2,\vec{x};\alpha), .., y(N,\vec{x};\alpha))$.
Let $\|.\|$ measure the magnitude of vectors based on the metric used. For example $\|\vec{A}\| = d(\vec{p}_2, \vec{o}_2)$, where $d(\vec{p}_2, \vec{o}_2)$ measures the closeness between two points in the model output space, for example mean of absolute values or any other measure that satisfies the conditions of being a metric.
We define a model structure $\wedge$ is defined as a collection of models parameterised by αs, i.e. $\wedge = (\alpha_1, \alpha_2, ..)$.

2. SIXPAR model and the data set

The SIXPAR model structure, which is a conceptual simplification of the SAC-SMA (Sacramento Soil Moisture Accounting Model) with one upper and lower zone, ignores evaporation and the concept of tension water zones but retains the complex conceptualization of percolation. These models are run at daily time steps using input forcing from a selected MOPEX basin in this study.
The parameter ranges used in the study are given in the table below.

Table 1: *SIXPAR model structure parameter ranges used in the study. "HR" range contains high recession parameter values obtained from the "Reference" range. "LR" range similarly contains low recession parameter values from the "Reference" range. "HS/LR" has the same recession values as "LR" but it now has larger range for storage capacities than those in the "Reference" range.*

Parameter	"Reference"	"HR"	"LR"	"HS/LR"
UM [mm]	0-50	0-50	0-50	1-300
UK [day^{-1}]	0-1	0.75-1.00	0.10-0.25	0.10-0.25
BM [mm]	0-50	0-50	0-50	1-3000
BK [day^{-1}]	0-1	0.75-1.00	0.001-0.005	0.001-0.005
Z[-]	0-1	1-250	1-250	1-250
X[-]	0-10	1-5	1-5	1-5

3. Bayesian measure of model complexity

Marginal likelihood function ($m\left(\vec{y}^0, \vec{x}; \wedge\right)$) is the integral of likelihood function ($\ell\left(\vec{y}^0, \vec{x}; \alpha\right)$) with respect to parameters, $\alpha \in \wedge$. The integration is weighted by prior probabilities that are attached to the parameters, $p(\alpha)$, i.e. $m\left(\vec{y}^0, \vec{x}; \wedge\right) \propto \int_{\alpha\in\wedge} \ell\left(\vec{y}^0, \vec{x}; \alpha\right) p\left(\alpha\right) \partial\alpha$. This integral, in log space under certain assumptions such as a uniform prior and large N, can be approximated around a parameter, α^*, that maximizes the likelihood function, as

$$ll \log\left(\int_{\alpha\in\wedge} \ell\left(\vec{y}^0, \vec{x}; \alpha\right) p\left(\alpha\right) \partial\alpha\right) \approx \log\left(\ell\left(\vec{y}^0, \vec{x}; \alpha^*\right)\right) - \frac{1}{2}\log|\tilde{\mathcal{H}}^{-1}_{\alpha^*}|$$

Here, $|\tilde{\mathcal{H}}^{-1}_{\alpha^*}|$ is the determinant of the inverse of the Hessian of $\log\left(\ell(\vec{y}^0, \vec{x}; \alpha)\right)$ evaluated at α^*. The Hessian is a matrix whose element in i^{th} row and j^{th} column is,

$$\tilde{\mathcal{H}}_{i,j} = \frac{\partial^2 \log\left(\ell(\vec{y}^0, \vec{x}; \alpha)\right)}{\partial\alpha_i \partial\alpha_j}$$

The approximation demonstrates that marginal likelihood of a model structure $\wedge$ is composed of maximum log-likelihood value and inverse of the Hessian evaluated at the optimum. The larger the determinant of the inverse of Hessian for a given value of log-likelihood value, the smaller is the marginal likelihood of a model structure. This implies that a larger determinant (of Hessian inverse) value makes a model structure less likely to represent the underlying system for a given level of maximum log-likelihood value. This is synonymous to the notion of model complexity presented in this paper in the sense that there is a trade off between model complexity and model performance (log-likelihood). The Hessian measures the curvature of the likelihood function around its maximum value. Thus, the Bayesian notion of model complexity favors those models whose log-likelihood function is more sensitive to perturbations in parameters.
We here assume that a likelihood function is a function of $\eta = \|\vec{y}^0 - \vec{y}(\vec{x};\alpha)\|$. That is, $\ell(\eta) = \ell(\vec{y}^0, \vec{x}; \alpha)$. Thus we assume that the likelihood function is based on the similarity of a simulated response to observed time series. Assuming mean of absolute values as the metric, $\|\vec{y}^0 - \vec{y}(\vec{x};\alpha)\|$ is Mean Absolute Error (or mean of absolute residuals), i.e. $\eta = \sum_{t=1}^{N} \frac{|y^0(t) - y(t,\vec{x};\alpha)|}{N}$. Then it can be shown that when evaluated at the optimum,

$$\tilde{\mathcal{H}}_{i,j} = \frac{1}{\ell(\eta)} \frac{\partial^2 \ell(\eta)}{\partial \eta^2} \sum_{t=1}^{N} \frac{\partial y(t,\vec{x};\alpha)}{\partial \alpha_i} \sum_{t=1}^{N} \frac{\partial y(t,\vec{x};\alpha)}{\partial \alpha_j}. \quad (1)$$

This shows that the Hessian depends on the relative curvature of the likelihood function (ratio of second derivative of the likelihood function and the likelihood value at the optimum) and the sensitivity of model output to perturbations in parameters. The latter is specific to the model and the parameter magnitudes used. For SIXPAR model, parameters corresponding to storage capacities and recession coefficients are used (UM, UK, BM and BK; see Table 1).
Since the Bayesian measure of complexity is the log determinant of Hessian inverse ($\log(|\tilde{\mathcal{H}}^{-1}_{\alpha^*}|)$), we define our measure of complexity as $\log(1/\kappa_{\mathcal{H}})$, where $\kappa_{\mathcal{H}}$ is the condition number of the matrix $\mathcal{H}$ with elements $\mathcal{H}_{i,j} = \sum_{t=1}^{N} \frac{\partial y(t,\vec{x};\alpha)}{\partial \alpha_i} \sum_{t=1}^{N} \frac{\partial y(t,\vec{x};\alpha)}{\partial \alpha_j}$. We here note that $1/\kappa_{\mathcal{H}}$ is similar to taking the determinant of the inverse of $\mathcal{H}$. Such a measure thus only considers the contribution of model parameter magnitudes to model complexity and excludes $\frac{1}{\ell(\eta)} \frac{\partial^2 \ell(\eta)}{\partial \eta^2}$.

4. Complexity measure based on relative condition number (stability of system representation by SIXPAR)

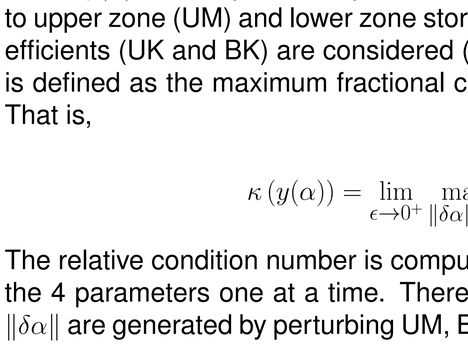
Consider two data sets of same size N that have been sampled from the same underlying distribution. Model parameters that are selected on the first data set by an ill-conditioned model selection problem will be different from model parameters that are selected on the second data set. A more ill-conditioned model selection problem will select models with more diverse parameters. The inverse of condition number that quantifies how much variation in parameters is possible per unit perturbation in model response is therefore used as a corresponding measure of model complexity.
We use relative condition number, $\kappa\left(y(\vec{x};\alpha)\right)$ that quantifies the sensitivity of a hydrological model $y(\alpha)$ with respect to its parameters. The model that is used is SIXPAR and sensitivities to upper zone (UM) and lower zone storage (BM) capacities and corresponding recession coefficients (UK and BK) are considered (see Table 1). For a given parameter value α, $\kappa\left(y(\alpha)\right)$ is defined as the maximum fractional change in model output to any fractional change in α. That is,

$$\kappa\left(y(\alpha)\right) = \lim_{\epsilon\to 0^+} \max_{\|\delta\alpha\|\leq\epsilon} \frac{\|\vec{y}\left(\vec{x};\alpha+\delta\alpha\right) - \vec{y}\left(\vec{x};\alpha\right)\|}{\|\vec{y}\left(\vec{x};\alpha\right)\|} \frac{\|\alpha\|}{\|\delta\alpha\|} \quad (2)$$

The relative condition number is computed for $\epsilon = 0.01$. Perturbations $\|\delta\alpha\|$ are considered in the 4 parameters one at a time. Therefore, for a given SIXPAR parameter value 4 different $\|\delta\alpha\|$ are generated by perturbing UM, BM, UK and BK parameters by 0.01 one at a time. The metric used is mean of absolute values, i.e. $\|\vec{y}\left(\vec{x};\alpha\right)\| = \sum_{t=1}^{N} \frac{|y(t,\vec{x};\alpha_0)|}{N}$ and $\|\alpha\| = \sum_{j=1}^{J} \frac{|\alpha_j|}{J}$, where J is the number of parameters and α_j is the j^{th} parameter.
The measure of complexity based on this condition number is then given by $\log\left(1/\kappa(y(\vec{x};\alpha))\right)$, which then measures the stability of system representation by SIXPAR.

5. Results

Figure 1: *Measures of complexity for SIXPAR based on condition numbers for 'ME' basin for various parameter ranges as described in table 1. These quantify how ill conditioned corresponding model selection problems are. $\kappa\left(y(\alpha)\right)$ is the relative condition number of a model ($y(\alpha)$ = SIXPAR) with respect to its parameters (α_j = storage capacities and recession parameters), while $\kappa_{\mathcal{H}}$ is the condition number of a matrix $\mathcal{H}$ whose elements $\mathcal{H}_{i,j}$ are given by $\sum_{t=1}^{N} \frac{\partial y(t,\vec{x};\alpha)}{\partial \alpha_i} \sum_{t=1}^{N} \frac{\partial y(t,\vec{x};\alpha)}{\partial \alpha_j}$.*



If the Bayesian measure is a valid measure of model complexity then the measure based on relative condition number is a valid measures of complexity as well. Both the measures of complexity suggest that complexity increases with higher values of recession coefficients and lower values of storage capacities, indicating that model complexity depends on parameter magnitudes as well.

References

[1] S. Pande et al. Hydrological model parameter dimensionality is a weak measure of prediction uncertainty. *Hydrology and Earth System Sciences Discussion*, 2015.