QUICS: Quantifying Uncertainty in Integrated Catchment Studies

D1.2 ESR6 Open source computational libraries of efficient algorithms for adaptive MCMC sampling and/or posterior emulation.

Lead Partner: Eawag

Revision: 31.05.2017

Report Details

Title: Computational library in the open source language R of efficient algorithms for adaptive MCMC sampling and/or posterior emulation.

Deliverable Number (If applicable): D1.2

Author(s): Jörg Rieckermann, Sanda Dejanic

Dissemination Level: Public

Document History

Version	Date	Status	Submitted by	Checked by	Comment
v1.0	07/12/2016	Draft First	Jörg Rieckermann	Will Shepherd	
v1.1	28/05/2017	Draft Second	Sanda Dejanic	Jeroen Langeveld	
v1.2	31/05/2017	Final	Sanda Dejanic	Will Shepherd	

Acronyms and Abbreviations

MCMC	Markov Chain Monte Carlo

Acknowledgements

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 607000.

Executive Summary

One challenge in uncertainty estimation of environmental models is the high number of parameters, as well as model structure errors. The first make traditional optimization routines difficult due to co-linearities of model parameters. The second often require additional error terms which capture these structural deficits with stochastic processes.

Inference with the latter requires Bayesian inference methods for model calibration, which typically rely on sampling from the posterior. Obtaining a reliable estimate of the posterior usually requires many thousand iterative model runs and there is the danger that sampling a very complex posterior distribution is inadequate due to too few iterations.

Efficient samplers have been proposed to relax this problem by increasing the acceptance rate due to sophisticated proposal algorithms, such as adaptive sampling or parallelizing the inference with ensembles.

In this Deliverable we present three open source computational libraries of efficient algorithms for adaptive (1) and ensemble-based MCMC sampling (2). In addition, we explore and compare the effectiveness of each sampler in standardized inference problems.

Our results on the posterior sample of a Rosenbrock distribution suggest that the ensemble-sampler using a Stretch Move is more efficient than the traditional Differential Evolution move for sampling nonlinear correlated posteriors, which we expect to find in inference of environmental models. Also, our results suggest that Stretch Move algorithm is less sensitive to the ensemble size than the Differential Evolution move.

All three algorithms packages have been made available on GitHub, a web-based Git or version control repository. This offers the distributed version control and source code management functionality of Git, which means that the repositories can be distributed and managed in a concise way.

CONTENTS

1 Introduction

1.1 Partners Involved in Deliverable

Eawag

1.2 Deliverable Objectives

Computational library in the open source language R of efficient algorithms for adaptive MCMC sampling and/or posterior emulation.

2 Computational library for MCMC sampling

2.1 Problem

- One of the errors while making predictions in environmental studies comes from sampling a very complex posterior distribution with too few iterations.
- Having an insufficient number of iterations leads to building a prediction based on a set of samples that is not entirely representative.
- This leads to predictions with additional errors due to the numerical techniques that were used.

2.2 Solution

- Testing current techniques for Bayesian inference and choosing the most reliable one would lead to reducing this error
- Implementing methods from other sciences that have already proven to provide satisfying results on complex and highly dimensional models, such as hydrological models, would open up a whole new perspective.
- Adapting the best performing algorithms and making them accessible and easy to use would allow hydrologists, with QUICS fellows as pioneers, to directly benefit from them.
- Coding these algorithms from scratch will make the comparison clear and users can directly benefit from the most efficient algorithm, in the form of a new, user-friendly software.

2.3 Current progress

1. Computational tools for efficient sampling of high-dimensional distributions with nonlinear correlation structures, e.g. “banana” or “Coffee-Mug” shapes in R and Julia.

- R package available on GitHub with the affine invariant ensemble sampler and a sampler based on the differential evolution move.
  https://github.com/SandaD/MCMCEnsembleSampler
- Julia package on GitHub

https://github.com/SandaD/MCMCEnsembleSampler.jl

- R package for adaptive sampling with Vihola-algorithm
https://github.com/scheidan/adaptMCMC

2. Draft paper on “Appraisal of jump distributions in ensemble based sampling algorithms”, to be submitted by 01.07.2017. We explore properties of the MCMC chains obtained by two different algorithms, one commonly used in hydrology and another commonly used in astrophysics (Figure 1).

   Performance analysis is focused on:

- Estimating the convergence time with methods developed for analyzing ensembles of particles in the field of statistical physics
- Making a video showing the difference in algorithm performance for didactical purposes based on images showing the progress of both algorithms given same initial conditions (Figure 2)
- Testing and comparing Robustness – sensitivity to tuning parameters (Figure 3)

2.4 Next Steps

- Uploading the packages on CRAN with documentation.
- Publishing a paper and having environmental scientists as audience.

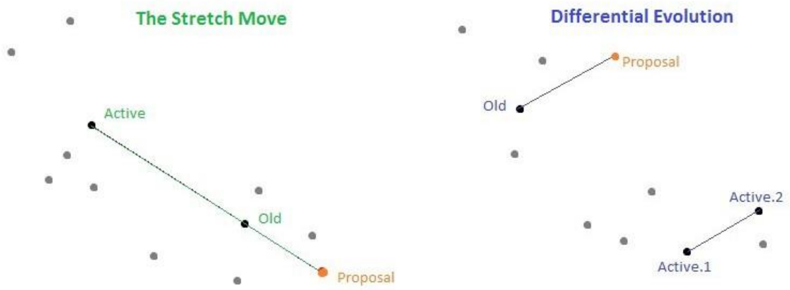

Figure 1: illustration of the MCMC sampler with a stretch move (left), in comparison to the popular differential evolution algorithm (right). The hypothesis is that the stretch move can more efficiently sample nonlinear correlated posteriors, which we expect in hydrological applications and integrated catchment studies.

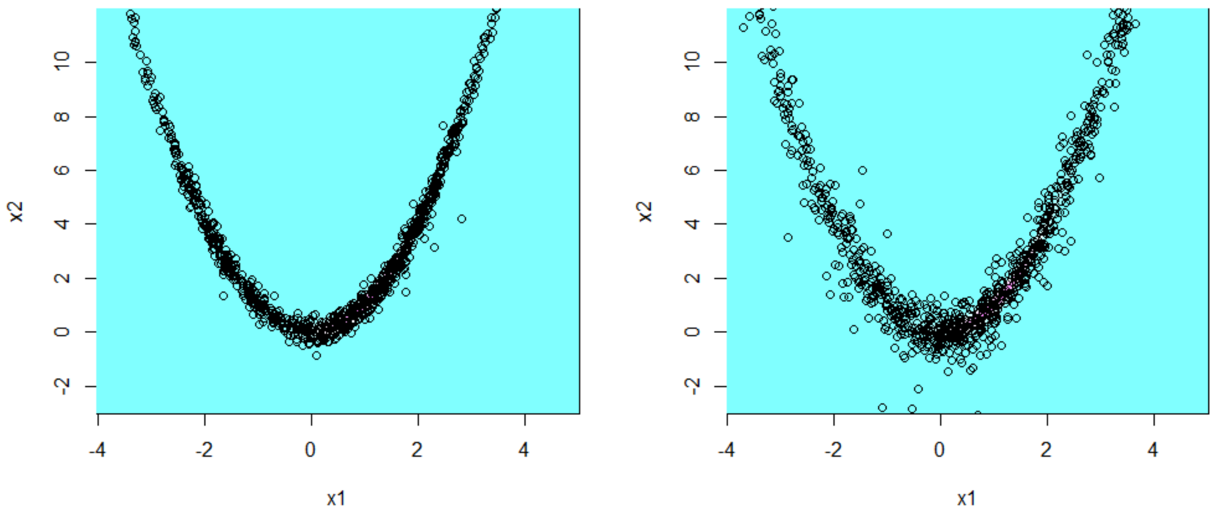

Figure 2: The posterior sample of a Rosenbrock distribution (n=10E4) shows that the stretch move (left) is more efficient than the Differential evolution move (right) for sampling nonlinear correlated posteriors.

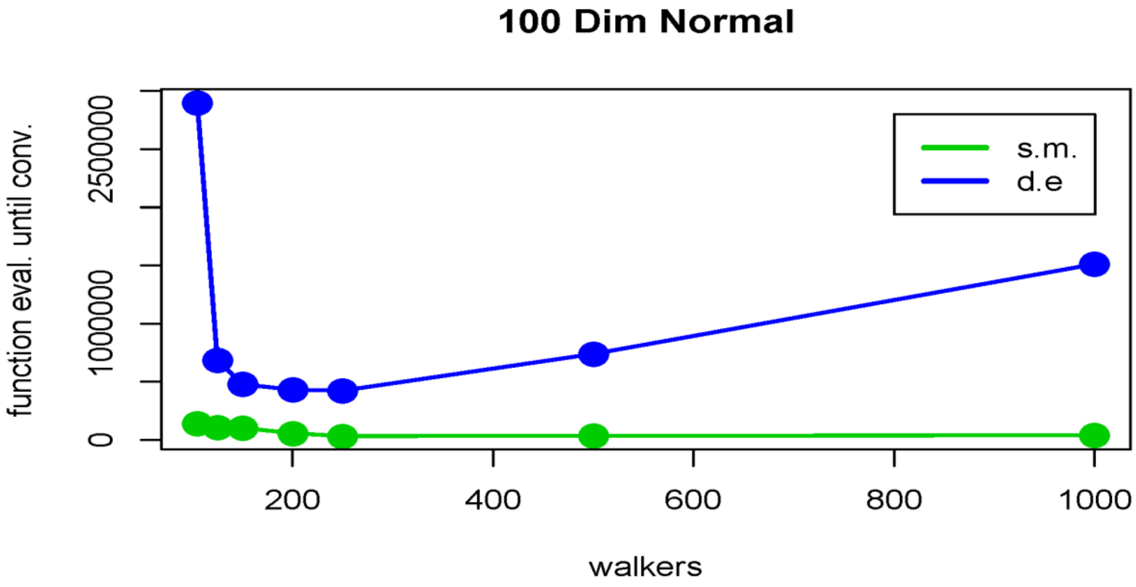

Figure 3: Robustness plot shows that the stretch move (green) is less sensitive to the number of ensemble members than the Differential evolution move (blue). Here, we show the results for sampling a 100 dimensional Normal distribution.