BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–9

A Bayesian Approach to Targeted Experiment Design (Supplement)

J. Vanlier [1,2*], C. A. Tiemann [1,2], P. A. J. Hilbers [1,2] and N. A. W. van Riel [1,2*]

[1]Department of BioMedical Engineering, Eindhoven University of Technology, Eindhoven 5612 AZ, The Netherlands
[2]Netherlands Consortium for Systems Biology, University of Amsterdam, Amsterdam 1098 XH, The Netherlands

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

1 DETAILS REGARDING THE MCMC SAMPLER

As mentioned in the paper Markov Chain Monte Carlo (MCMC) is a technique used to obtain samples from probability distributions known only up to a normalising factor. Given that $p(\vec{\theta})$ is a non-negative integrable function, the Metropolis Hastings algorithm will provide a sequence of samples (also known as a chain) whose equilibrium distribution is proportional to $p(\vec{\theta})$ using only evaluations of $p(\vec{\theta})$. If the chain is irreducible and aperiodic then the chain will settle down to a limiting (ergodic) distribution. Irreducibility requires that it must be possible to get from any two possible states of the system to the other in a finite number of steps. A sufficient condition to ensure that $p(\vec{\theta})$ is the limiting distribution is that the sampler satisfies detailed balance (1) where $\pi(.)$ corresponds to the invariant distribution and $p(x, y)$ corresponds to the distribution used for proposing the next step (the proposal distribution). This property ensures that the chain is reversible (two sides are equal).

$$\pi(x)p(x, y) = \pi(y)p(y, x) \tag{1}$$

The Metropolis algorithm ensured chain ergodicity by using only symmetric proposals (where the probability of a forward and backward move should be equal, i.e. the proposal distribution does not change). A generalisation by Hastings lead to the an additional term in the acceptance probability which ensures detailed balance for non-symmetric proposal distributions. The resulting algorithm, named the Metropolis-Hastings algorithm is generally considered as the workhorse of MCMC methods. The algorithm proceeds by iteratively performing a number of steps:

- 1. Generate a sample $\vec{\theta}_{new}$ by generating a sample taken from a proposal distribution based on the current state
- 2. Compute the likelihood of the data $L(\mathbf{y^D}|\vec{\theta}_{new})$ and calculate $P(\vec{\theta}_{new}|\mathbf{y^D}) = L(\mathbf{y^D}|\vec{\theta}_{new})P(\vec{\theta}_{new})$, where $P(\vec{\theta}_{new})$ refers to the prior density function.
- 3. Draw a random number γ from a uniform distribution between 0 and 1 and accept the new step if $\gamma < min\left(\frac{P(\vec{\theta}_{n+1}|\mathbf{y^D})Q(\vec{\theta}_{n+1}\rightarrow\vec{\theta}_n)}{P(\vec{\theta}_n|\mathbf{y^D})Q(\vec{\theta}_n\rightarrow\vec{\theta}_{n+1})}, 1\right)$.

The ratio of Q is known as the Hastings correction and ensures detailed balance, a sufficient condition for the Markov Chain to converge to the equilibrium distribution. It corrects for sampling biases resulting from non-symmetric proposal distributions. It corrects for the fact that the probability density going from parameter set $\vec{\theta}_n$ to $\vec{\theta}_{n+1}$ and $\vec{\theta}_{n+1}$ to $\vec{\theta}_n$ is unequal when the proposal distribution depends on the current parameter set. It is defined as the ratio between the proposal densities associated with going from n to $n + 1$ and $n + 1$ to n. The apparent simplicity of the algorithm makes it conceptually attractive. Note however that naive approaches can lead to MCMC samplers that converge slowly and/or stay in the local neighbourhood of a local mode (Calderhead and Girolami, 2009).

Proposals Regarding the proposal distribution, we employ an adaptive Gaussian proposal distribution whose covariance matrix is based on a quadratic approximation to the cost function (Gutenkunst *et al.*, 2007). This matrix is computed by taking the inverse of an approximation to the Hessian matrix of the model under consideration. Such adaptation to the local geometry of the problem results in taking larger steps in directions where the cost function does not change much (improving efficiency of the sampler). Such a Gaussian distribution is characterised by a positive definite covariance matrix $\mathbf{\Sigma}$, the number of dimensions d and the vector of mean values μ:

$$\frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{|\mathbf{\Sigma}|}}e^{-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)} \tag{2}$$

Sampling from such a distribution is straightforward. Compute a decomposition such that $\mathbf{RR}^T = \mathbf{\Sigma}$. Subsequently draw a vector $\vec{z}$ of N independent normal variates. Since in our case, the normal distribution is centered around the current point, the expression for the next point becomes $\vec{x}_{new} = \vec{x} + \mathbf{R}\vec{z}$. If the proposal distribution depends on the current state (asymmetric proposals), then the proposal needs to be corrected for the imbalance in proposal densities in the two directions using the Hastings term, which can be calculated for a multivariate Gaussian proposal distribution:

$$Q(\vec{\theta}_n \rightarrow \vec{\theta}_{n+1}) = \frac{1}{\sqrt{|\mathbf{\Sigma_a}|}}e^{-\frac{1}{2}(\vec{\theta}_{n+1}-\vec{\theta}_n)\mathbf{\Sigma_a}^{-1}(\vec{\theta}_{n+1}-\vec{\theta}_n)^T} \tag{3}$$

*to whom correspondence should be addressed

Calculating the true Hessian is costly and numerically challenging, which is why the approximation based on the model sensitivities is used. Depending on the model, these can either be computed by solving the sensitivity equations, or by means of finite differencing (for which *strict tolerances* are required to ensure reliable derivatives). The Hessian approximation is subsequently decomposed using the singular value decomposition (4), where **S** is a diagonal matrix containing the singular values.

$$\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^T \tag{4}$$

The decomposition delivers a matrix of singular vectors U and a diagonal matrix with associated singular values. Large singular values of the Hessian matrix correspond to well constrained directions, while low values correspond to poorly constrained directions in parameter space. In practical cases, some directions in parameter space can be so poorly constrained that this leads to a (near) singular Hessian (some singular values near zero). As a result, the proposal distribution will become extremely elongated in these directions, leading to proposals where parameters take on extreme values and acceptance ratios decline due to integration failures or rejections due to low probability density in these regions. One approach to avoid such numerical difficulties is to set singular values below a certain cut-off to a specific minimal value (prior to inversion) or to make use of a trust region approach. Rather than setting a fixed cutoff, a trust region approach (5) involves an adaptive mechanism for shrinking the sampling kernel.

$$\mathbf{H} = \mathbf{J}^T\mathbf{J} + \lambda\mathbf{I} \tag{5}$$

Here **J** corresponds to the sensitivity matrix while **I** corresponds to an identity matrix. Since the Hessian approximation corresponds to a quadratic approximation of the cost function, we can estimate what the cost would be at a certain point prior to sampling it. If this cost deviates more than a certain allowed threshold, we increase λ, making the proposal more circular, otherwise we decrease λ. Additionally, we include non-uniform priors (when available) in the Hessian approximation, by including their respective derivatives in the approximation of the sensitivity matrix **S**.

The avoid the computation of costly inverses, we compute the sampling matrix directly from the SVD using

$$\mathbf{R} = \frac{s\sqrt{T}}{\sqrt{N_{dim}}}\mathbf{V}\sqrt{\mathbf{S}^{-1}} \tag{6}$$

Here s corresponds to a problem specific (tuned) scaling factor, T to the temperature (see section on multimodality) N_{dim} to the number of parameters. The inverse required for the Hastings correction can subsequently be computed as:

$$\mathbf{\Sigma}^{-1} = \frac{N_{dim}}{s^2T}\mathbf{V}\mathbf{S}\mathbf{V}^{\mathbf{T}} \tag{7}$$

Since the determinant only appears in ratios, the linear scaling needs not be explicitly calculated since it will cancel out due to the dimensionality of the problem remaining constant:

$$det(\mathbf{\Sigma}) = \left(\frac{N_{dim}}{s^2T}\right)^{-N_{dim}} \prod \frac{1}{S_{ii}} \tag{8}$$

Therefore the ratio is computed as a product of the ratios of the singular values.

Parameter representation In order to deal with the large difference in scales, certain parameters can be considered in log-space. Note however, that the prior distribution is generally not invariant of the way the model is parameterised. The transformation between parameters can be described by the matrix of partial derivatives with respect to the equations which transform the parameters from one parameterisation to another (the Jacobian of the transformation). In order to calculate a prior distribution that is equivalent in terms of inference under a different parameterisation, one needs to compute the absolute value of the determinant of the Jacobian of the transformation. This corrects for the stretching and compression of the distribution due to the reparameterisation (9).

$$P(f(\vec{\theta})) = P(\vec{\theta})\left|\left|\begin{bmatrix} \frac{df(\theta_1)}{d\theta_1} & \cdots & \frac{df(\theta_1)}{d\theta_n} \\ \vdots & \ddots & \vdots \\ \frac{df(\theta_n)}{d\theta_1} & \cdots & \frac{df(\theta_n)}{d\theta_n} \end{bmatrix}\right|\right| \tag{9}$$

In the case where we perform the MCMC in logarithmic space, we obtain the following expression to be included in the acceptance probability:

$$\frac{|J(\theta^a)|}{|J(\theta^b)|} = \prod_i^{N_{pars}} \left(\frac{\theta_i^b}{\theta_i^a}\right) \tag{10}$$

For the Hessian based approach, the proposals can subsequently be generated using the equation in:

$$\vec{\theta}_{n+1} = \vec{\theta}_n e^{N(0,\mathbf{\Sigma_{log}})} \tag{11}$$

Where the Hessian approximation in log space is computed by applying the chain rule:

$$\frac{\delta^2 L}{\delta log\theta_i \delta log\theta_j} = \frac{\delta^2 L}{\delta\theta_i\delta\theta_j}\theta_i\theta_j \tag{12}$$

Multi modality In some cases, posterior distributions can be multi modal and the sampler is unable to leave a local mode within a reasonable number of iterations. One option to improve mixing is to start a number of parallel chains using the data at different 'temperatures' T:

$$P_T(\mathbf{y^D}|\vec{\theta}, T) = P(\mathbf{y^D}|\vec{\theta})^{\frac{1}{T}} \tag{13}$$

Since the cost function will flatten out for higher temperatures, chains at higher temperatures are able to traverse the solution space more freely. Exploiting this property, we use Metropolis-Coupled MCMC (Calderhead and Girolami, 2009). Here multiple MCMC chains are started where samples from the higher temperatures are exchanged with samples at lower temperatures using switch moves. These are performed by randomly selecting two adjacent temperatures and computing an Metropolis-Hastings step using the acceptance ratio given by:

$$\alpha < min\left(\frac{P(\mathbf{y^D}|\vec{\theta}_{new})^{\frac{1}{T_n}} P(\mathbf{y^D}|\vec{\theta}_n)^{\frac{1}{T_{new}}}}{P(\mathbf{y^D}|\vec{\theta}_n)^{\frac{1}{T_n}} P(\mathbf{y^D}|\vec{\theta}_{new})^{\frac{1}{T_{n+1}}}}\right) \tag{14}$$

Here one defines a joint probability space, where multiple instances of the parameter vector are concatenated. Similarly, the objective function is copied and concatenated with each copy corresponding to a different temperature. Updates are performed per

parameter group, while switch moves enable the sampler to switch the parameters between two groups. Since we are interested in the distribution at $T = 1$ we only use samples from this respective chain.

COMPUTATIONAL METHODS

All of the algorithms were implemented in Matlab (Natick, MA). Numerical integration was performed using compiled MEX files using numerical integrators from the SUNDIALS CVode package (Lawrence Livermore National Laboratory, Livermore, CA). Absolute and relative tolerances were set to 10^{-8} and 10^{-9} respectively. Integration time for a single integration was allowed to be 10 seconds at most after which an integration is assumed to fail and a large error is returned. Throughout the analysis integration failures were carefully monitored.

In order to attain an adequate acceptance rate and good mixing, the proposal scaling was determined during an initial tuning stage. This tuning was performed by running many short chains (100 iterations each), targeting an acceptance rate between 0.2 and 0.4. If the acceptance rate was high, 10% was added to the scale, while 10% was substracted in the case where the acceptance was too low. Interestingly the chains at higher temperatures had very similar acceptance rates once started. For the MCMC method, no cutoff was necessary in the case of the uniform priors, while the cut off was set to 10^{-6} in the case of log uniform priors. It was observed that this greatly affected the acceptance rate.

JAK-STAT MODEL EQUATIONS

Model equations for the JAK-STAT model were specified as (15). The model is driven by an external input u_1, which is based on a spline interpolation to phosphorylated EpoR data. See Raue *et al.* (2009) for further details.

$$
\begin{aligned}
\dot{x_1} &= 2\frac{V_{nucleus}}{V_{cyto}}(p_4 x_{13}) - p_1 x_1 u_1 & \dot{x_8} &= p_4 x_7 - p_4 x_8 \\
\dot{x_2} &= p_1 x_1 u_1 - 2p_2 x_2^2 & \dot{x_9} &= p_4 x_8 - p_4 x_9 \\
\dot{x_3} &= p_2 x_2^2 - p_3 x_3 & \dot{x_{10}} &= p_4 x_9 - p_4 x_{10} \\
\dot{x_4} &= \frac{V_{cyto}}{V_{nucleus}}(p_3 x_3) - p_4 x_4 & \dot{x_{11}} &= p_4 x_{10} - p_4 x_{11} \\
\dot{x_5} &= p_4 x_4 - p_4 x_5 & \dot{x_{12}} &= p_4 x_{11} - p_4 x_{12} \\
\dot{x_6} &= p_4 x_5 - p_4 x_6 & \dot{x_{13}} &= p_4 x_{12} - p_4 x_{13} \\
\dot{x_7} &= p_4 x_6 - p_4 x_7
\end{aligned}
\tag{15}
$$

2 GAUSSIAN MEASUREMENT ERRORS

In the case of additive Gaussian measurement noise one can compute the PPD directly from the PD as follows:

$$P(y_n|\theta) = \tag{16}$$

$$\int_{-\infty}^{\infty} K_1 P(y_n|y_p) P(y_p|\vec{\theta}) dy_p = \tag{17}$$

$$\int_{-\infty}^{\infty} K_1 e^{-\frac{(y_n-y_p)^2}{2\sigma^2}} e^{-\frac{(y_p-y(\vec{\theta}))^2}{2\sigma^2}} dy_p = \tag{18}$$

$$K_2 e^{-\frac{(y_n-y(\vec{\theta}))^2}{4\sigma^2}} \tag{19}$$

Here K_1 and K_2 denote different normalization constants independent of $\vec{\theta}$. $P(y_p|\theta)$ is the posterior probability of predictions while $P(y_n|\theta)$ represents the probability of observing y_n in a new measurement. Furthermore $P(y_t|y_p)$ refers to the error model of the new measurement. Since the sampling is self-normalising these are not required. Therefore, the prediction noise can be taken into consideration by multiplying the measurement noise in the error model by a factor of $\sqrt{2}$. Including this step avoids the requirement to add simulated Gaussian noise to the PD. Additionally, this will reduce the number of samples required for an accurate estimation of the variance reduction.

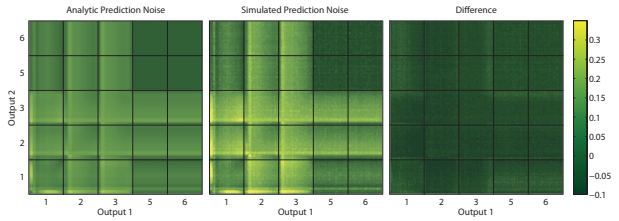

Fig. 1. Expected variance reduction compared with true variance reduction after the experiment has been performed.

3 SAMPLING BIAS

The most critical component of the experiment design strategy is the sampling step. Since every point of the MCMC is treated as a potential measurement result, this also includes samples farther from the high density region of the posterior. Since the density of samples is lower here, the number of samples to estimate the new posterior variance from is small. In the most extreme case (say the outermost sample), the expected mean after incorporating the new measurement would be biased towards the high density region, while the variance will be underestimated. This worst case scenario is illustrated in Figure 2 for a 1D distribution.

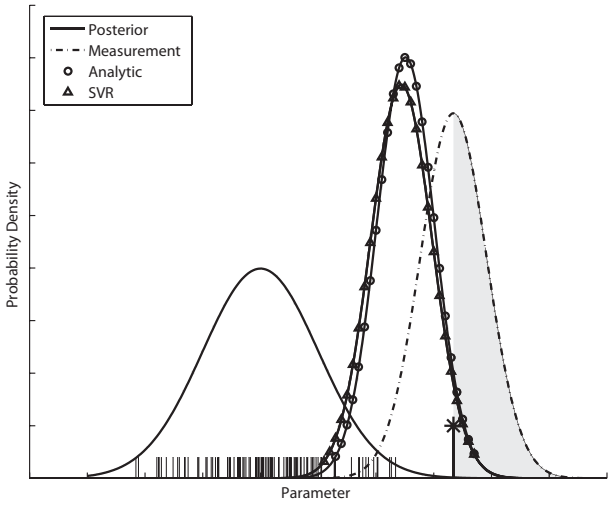

Fig. 2. Illustration of the effect of a new datapoint on the Posterior Predictive Distribution (PPD). The stems indicate the samples from the MCMC, where the one selected as new outcome value for the measurement is denoted with a star. The lines referring to the different distributions correspond to analytic probability density functions based on computed means and variances. Shown are the posterior before incorporating the new datapoint, the distribution of the measurement, the distribution based on the analytic mean and variance of the new posterior and the distribution based on the mean and variance estimated from resampling as performed in the proposed method

We are however not interested in a single variance, but rather in the expected value for the variance over the entire posterior. Given that most of the resampling will take place in the high density region it is postulated that the estimation error will not be very large. It is expected that the method will show slight bias for low numbers of included samples, but that the bias will quickly decrease as the sample size increases. Several factors play a role in this sampling. Some of these are: the number of points included in the sampling step, the dimensionality of the problem, the difference between the variance of the posterior and the new measurement and the amount of correlation between measurement and quantity of interest. We've investigated this sampling step by performing tests using multidimensional Gaussians. One example of such a test is shown in Figure 4. Here it can be observed that the bias of the sampling approach is indeed more pronounced for smaller sample sizes. Interestingly, low correlations (associated with low variance reductions) result in slightly more bias. Furthermore, the linear method is unbiased, even at low sample sizes. Note however that this method depends on the fact

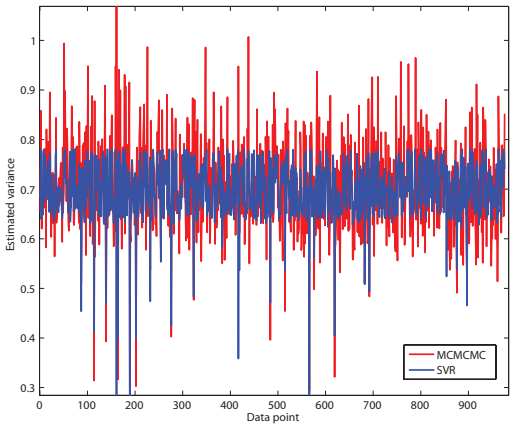

Fig. 3. Comparison of nested MCMC approach to resampling technique for the banana function.

that the PPD is Gaussian, an assumption which in this experiment holds by design, but is questionable for real PPDs.

The effect of the different variances were also investigated. There are three variances that might play a role. The variance of the old posterior, on the side of the quantity of interest, the variance of the posterior where the measurement will take place, and the variance associated with the uncertainty of the new measurement. Interestingly, numerical experiments showed that the absolute bias was unaffected by the variance of the quantity of interest.

To investigate these effects in the non-linear case we performed an analysis on a 2D banana function. The residual vector used in this analysis was defined as:

$$\vec{r}(\vec{x}) = \left[\sqrt{10}\left(x_2 - x_1^2\right), \sqrt{2} - x_1\right] \tag{20}$$

with x_1 and x_2 as the parameters. The associated probability density function was given by:

$$C(\vec{x}) = e^{\Sigma_i r_i^2} \tag{21}$$

Here the variances for each sample of the MCMC chain were computed in two ways. First by means of running a new MCMC for each sample of the previous posterior (MCMCMC) as well as using the Sampling Variance Reduction (SVR). The results are shown in Figure 3. What can be observed here is that although the extremely high and low values do not agree well, most of the chain results in the same values for the variance and the mean is still well estimated. This can also be observed when considering the actual predicted variance for different values of σ (see Figure 5).

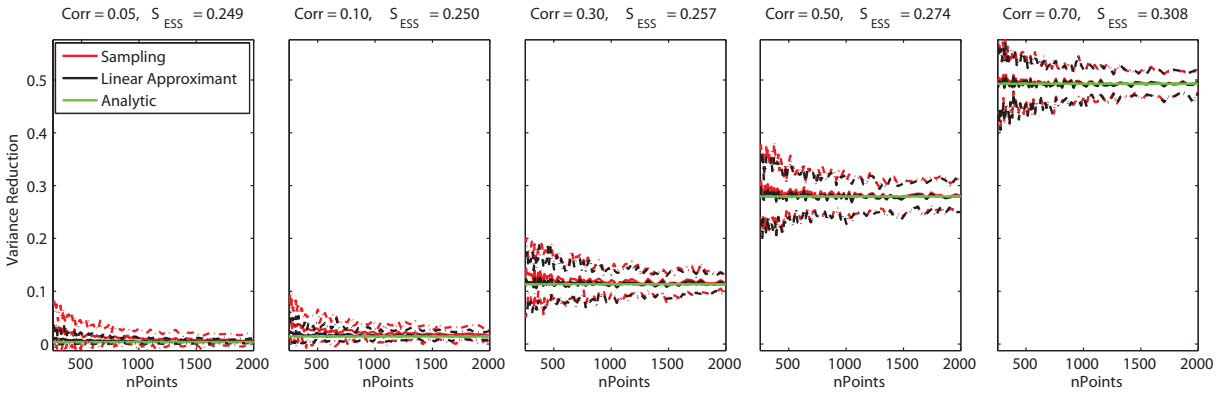

Fig. 4. Estimated variance reduction as a function of number of points included in the analysis. For PPD a multivariate Gaussian distribution was used, with output standard deviation 5, and measurable standard deviations 4 and 3. The measurement accuracy of the new measurement was assumed to be Gaussian with a standard deviation of 1. All correlation coefficients were set to the same value. Each experiment was repeated 50 times. Shown in red, black and green are the mean variance reductions based on sampling, the linear approximation (which is exact for a Gaussian) and the true analytical solution. Dashed lines indicate 95 percentile bounds. Figure titles indicate used correlation and estimated slope of the Effective Sample Size as a function of the number of sample points.

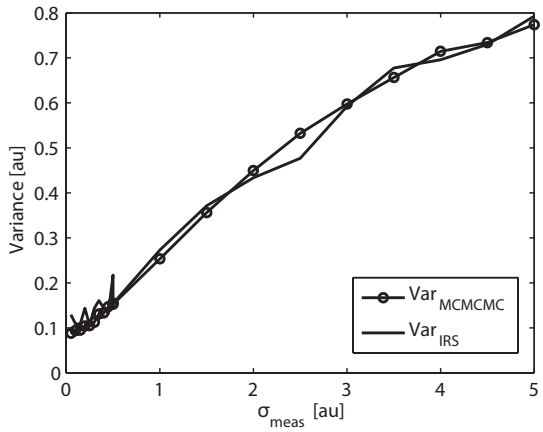

Fig. 5. Comparison of nested MCMC approach to resampling technique for the banana function.

4 ADDITIONAL MODEL RESULTS

The PPDs for all states are shown in Figure 6. Additionally, we performed a similar analysis for the decay time of state 2. This decay time is defined as the time point where the simulated response has decreased to 50% of the maximum value. As shown in figure 7, this distribution is far from Gaussian and contains multiple modes.

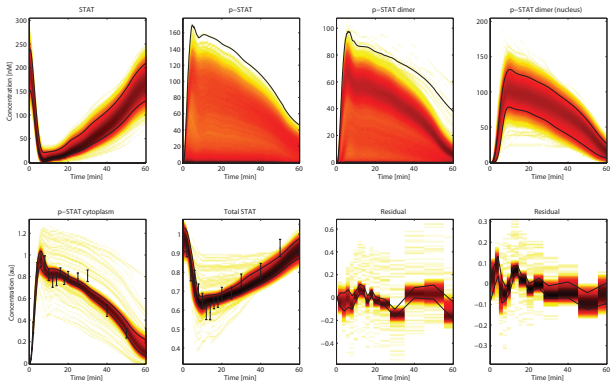

Fig. 6. Posterior predictive distribution of model predictions (colours) with 95% credible intervals (black lines). Top: Unmeasured internal model predictions. Bottom: Measured model output, data ± standard deviation and residual distributions.

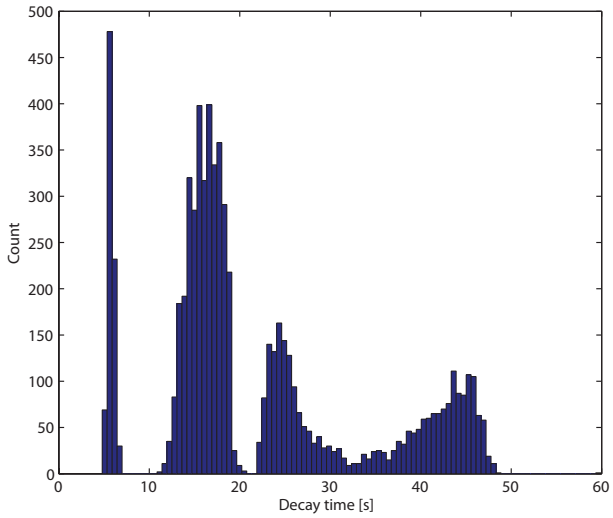

Fig. 7. Distribution of decay times (time until the 50% mark is crossed) of state two for the entire ensemble.

Subsequently a similar analysis was performed to attempt to reduce the variance of the decay time of state two. Interestingly, despite the clear non-Gaussian nature of the PPD, the results from the LVR and sampling method agreed reasonably well (see Figure 8). Subsequently, a more dense sampling was performed using the LVR (since it is orders of magnitude faster), resulting in Figure 9. It can clearly be seen that also in this case, measuring additional time points in the states that were already observed, would not reduce the variance appreciably. It can also be observed that the most beneficial experiment would be an early measurement of state 3 and a late measurement of state 1.

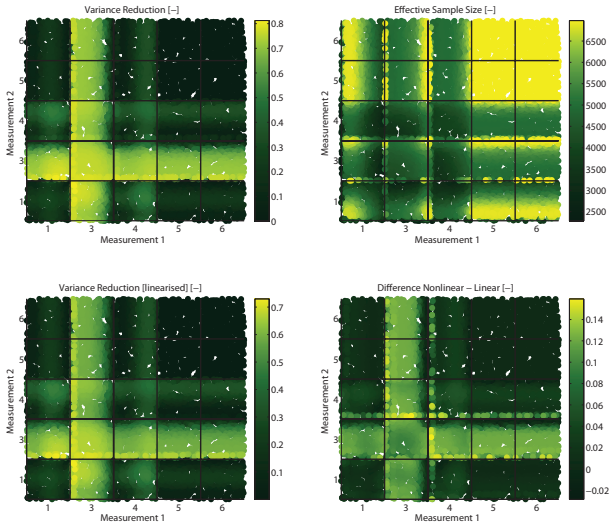

Fig. 8. Top left: Expected variance reduction based on the sampling method. Top right: Median Effective Sample Size for a specific experimental combination. Bottom left: Expected variance reduction based on LVR. Bottom right: Difference between LVR and sampling method.

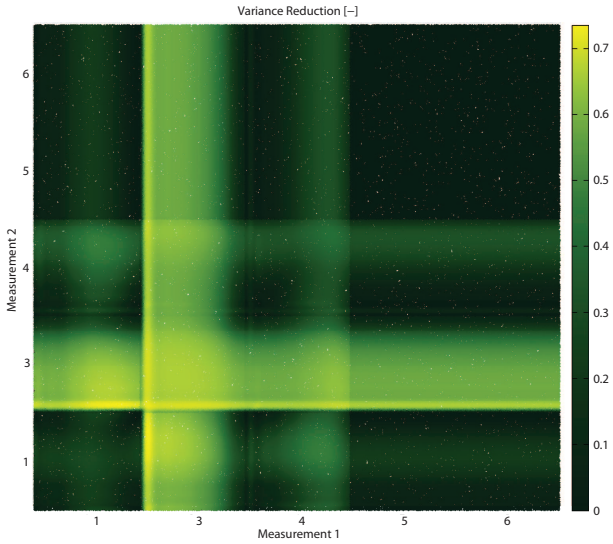

Fig. 9. Expected variance reduction based on LVR.

4.1 Leave one out experiment

In order to test our approach, we performed OED using only a subset of the data. In order to do this, we computed the posterior distribution omitting data corresponding to the total amount of cytoplasmic STAT. Of this observable only the first data point was included.

Subsequently we performed experiment design for the time points omitted earlier. Hence computing expected variance reductions were these to be included. As measurement accuracy we used the standard deviations of the omitted experiments. The results of this analysis are shown in Figure 10. We can see that the expected Variance Reductions agree well with the actual reductions obtained.

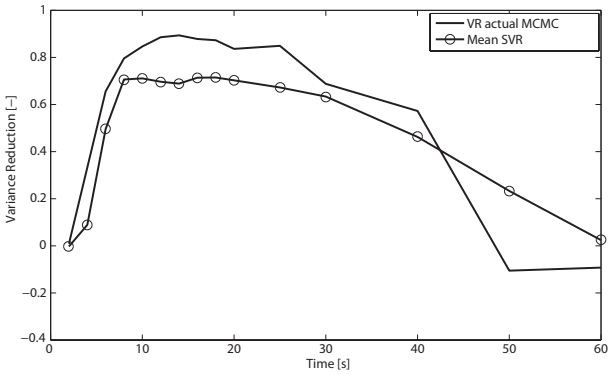

Fig. 10. Expected variance reduction compared with true variance reduction after the experiment has been performed.

4.2 Prior dependence

Since parameter two and three were non-identifiable from the data, we had to assume a bounded prior distribution for these. What we used was a log-uniform prior bounded between two values. In order to test the prior dependence, we extended the range of the log uniform priors for parameters two and three (from $[10^{-8}, 10^{2}]$ and $[10^{-8}, 10^{1.5}]$ to $[10^{-8}, 10^{3}]$ and $[10^{-8}, 10^{3}]$). As shown in Figure 11 the measurement of state 1 at an early and late time point is sensitive to the choice of prior. The expected variance reductions obtained when measuring state 2 or 3 in combination with state 1 were more robust.

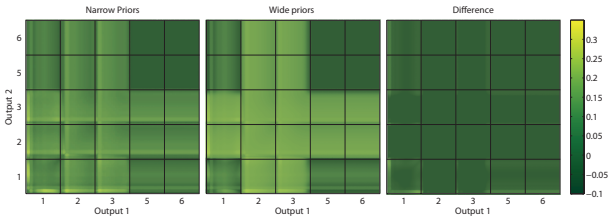

Fig. 11. Expected variance reductions considering narrow and wide ranges for parameters two and three.

5 OPENCL IMPLEMENTATION

Profiling of the targeted experiment design code revealed that the biggest computational burden resided in the computation of distances between the particles (even in the fully vectorised case). However, since this computation is the same for a large number of particles, this could straightforwardly be outsourced to hardware designed for parallel processing. For the OpenCL implementation of the sampling based approach a MEX file was written that computes a single estimated variance reduction. The sequence of samples is stored in global memory, while all the computations are performed using local registers. In order to avoid the overhead of having to build the OpenCL code into GPU runnable binaries at each point, we return the binary code coming from the graphics driver back to MATLAB, so that it can be used as an input in subsequent calls to the function. This approach gave us considerable speedups even on modest graphics hardware (see Figure 12).

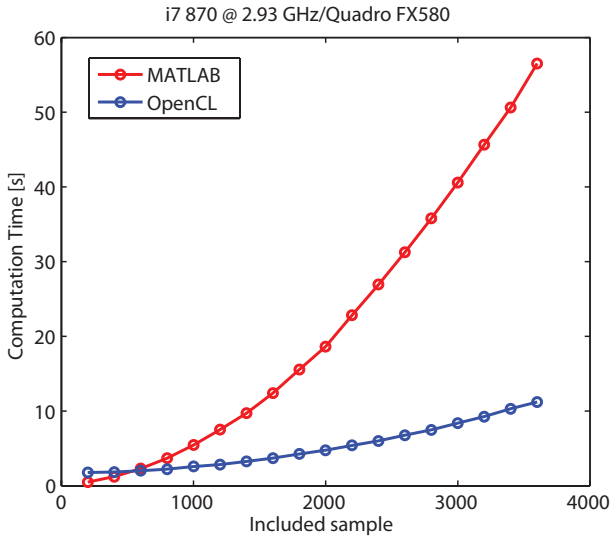

Fig. 12. Comparison of implementation of vectorised MATLAB versus OpenCL

REFERENCES

Calderhead, B. and Girolami, M. (2009). Estimating Bayes factors via thermodynamic integration and population MCMC. *Computational Statistics & Data Analysis*, **53**(12), 4028–4045.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., and Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, **3**(10), e189.

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., and Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, **25**(15), 1923.