Model parameters uncertainty estimation based on Bayesian inference for activated sludge models under aerobic conditions: a comparison with a linear theory method

Z. J. Zonta[*,**], X. Flotats[*,**] and A. Magrí[*]
[*]GIRO Technological Centre, Rambla Pompeu Fabra 1. E-08100, Mollet del Vallès, Barcelona, Spain.
(E-mail: zivko.juznik@giroct.irta.cat; albert.magri@giroct.irta.cat)
[**]Dept. of Agrifood Engineering and Biotechnology, Universitat Politècnica de Catalunya (UPC), Parc Mediterrani de la Tecnologia, Edifici D-4. E-08860, Castelldefels, Barcelona, Spain.
(E-mail: xavier.flotats@upc.edu)

Abstract
The purpose of the study is to apply Bayesian inference in order to estimate the uncertainty in model parameters and predictions for environmental models. The analysis was based on a global optimization routine that finds good initial values for an adaptive Markov chains Monte Carlo (MCMC) algorithm that finally computes the posterior parameter distribution. A revised activated sludge model was used in order to perform a comparison between Bayesian and linear theory methods. It was observed that the linear theory method systematically underestimates the confidence intervals of the estimated model parameters because the multivariate normality assumption is violated and practical unidentifiability for some parameters occurs.
Keywords
Activated sludge models (ASM); Bayesian inference; global sensitivity analysis; Markov chains Monte Carlo (MCMC); uncertainty analysis

INTRODUCTION

In environmental modelling, the residuals independent and identically distributed (IID) hypothesis is generally difficult to fulfil because of model structure and measurement systematic errors (Neumann and Gujer, 2008). If the residual IID hypothesis do not fulfil, the effects of the systematic errors should be expressed explicitly in a statistical model (Bayarri *et al*., 2007). Moreover, the typically assumed asymptotic normal approximation to the posterior parameter distribution should be verified in order to use a linear theory method (Neumann and Gujer, 2008). Asymptotic inference is widely used in wastewater model parameter uncertainty estimation because of its low computational burden and ease implementation (Marsili-Libelli *et al.*, 2003; Sin *et al.*, 2005; Hoque *et al.*, 2009), since it is based on the evaluation of parameter estimation error covariance matrix trough the inverse of the Fisher Information Matrix (FIM) or the Hessian matrix (Dochain and Vanrolleghem, 2001). Nevertheless, in environmental model identification, this linear method should be used wisely: scarce data and nonlinear model structures could give numerical problems in the matrix inversion operation (Dochain and Vanrolleghem, 2001) or lead to unfulfilled assumptions (Seber and Wild, 1989).

In wastewater treatment process modelling applications, researchers typically assume the accuracy of the asymptotic normal approximation to the posterior distribution. The objective of the present work is to test such assumption by comparing the confidence intervals (CIs) obtained from the asymptotic inference and from the more general Bayesian inference approach.

METHODS

Bayesian inference

In the Bayesian inference framework the concept of probability (Pr) is defined as the “degree of belief” or the plausibility that a proposition is true and is quantified as a real, positive number in the range of [0, 1]. Suppose we want to determine the probability of a continuous parameter, $\theta = \theta_1 \ldots \theta_p$, given the available data, $D = d_1 \ldots d_n$, and considering the prior information on the parameter θ, I. Applying the Bayes’ theorem (Sivia and Skilling, 2006), we can write

$$\Pr_{\text{posterior}}\left(\theta \mid D,I\right)=\frac{\Pr_{\text{prior}}\left(\theta \mid I\right)\Pr_{\text{likelihood}}\left(D \mid \theta,I\right)}{\Pr_{\text{evidence}}\left(D \mid I\right)}=\frac{\Pr\left(\theta \mid I\right)\Pr\left(D \mid \theta,I\right)}{\Pr\left(D \mid I\right)}, \tag{1}$$

where:

i. Pr (θ | *D, I*) is the *posterior probability* of θ, conditional on *D* and *I*, which express the plausibility of θ "after" performing the experiment and obtaining the data.

ii. Pr (*D* | *I*) is the normalising constant called *evidence*,

$$\Pr\left(D \mid I\right)=\int_{\Theta}\Pr\left(D,\theta \mid I\right)\mathrm{d}\theta,\ \text{with}\ \Theta\subseteq\mathbb{R}. \tag{2}$$

iii. Pr (*D* | θ, *I*) is the *likelihood function*, a conditional probability in function of θ, with *D* held fixed.

iv. Pr (θ | *I*) is the *prior probability*, the belief over the parameter θ "before" data is observed.

At this point, the uncertainty in the future model predictions $y = f(\theta)$ can be inferred as

$$\Pr\left(y \mid D,I\right)=\int_{\Theta}\Pr\left(y,\theta \mid D,I\right)\mathrm{d}\theta=\int_{\Theta}\Pr\left(y \mid \theta\right)\Pr\left(\theta \mid D,I\right)\mathrm{d}\theta, \tag{3}$$

where *y* and *D* are assumed to be conditionally independent given the value of θ.

Now, if we assume an additive model for the IID observations *D*,

$$d = f\left(\theta\right)+\varepsilon,\ \ \varepsilon \sim N\left(0,\sigma^2\right), \tag{4}$$

where the term ε is an independent Gaussian error; the corresponding *likelihood function* for this model of the data takes the form

$$\Pr\left(D \mid \theta,\sigma^2\right)=\left(2\pi\sigma^2\right)^{-\frac{n}{2}}\exp\left\{-\frac{1}{2\sigma^2}SS\left(\theta\right)\right\}. \tag{5}$$

The sum of the squares function (*SS*) is calculated as

$$SS\left(\theta\right)=\sum_{i=1}^{n}\left(d_i - f_i\left(\theta\right)\right)^2. \tag{6}$$

Historically, the main problem with the Bayes' formula (Eq. 1) was the costly computation of the integral in the denominator, especially for the multidimensional case. Nowadays, the problem is tractable by the increasing CPU capabilities and the use of Markov chains Monte Carlo (MCMC) methods. The last are a class of algorithms for sampling from probability distributions based on constructing a Markov chain that has the desired distribution as its target distribution, Pr (θ | *D, I*). The state of the chain after a large number of steps is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of steps.

Asymptotic inference

If we are largely ignorant about the *prior probability*, then we might indicate this naively with a uniform, or flat probability distribution function (pdf). With this simplification, the logarithm *L* of the posterior pdf is simply given by

$$L=\ln\left\{\Pr\left(\theta \mid D,I\right)\right\}=const-\frac{n}{2}\ln\left(2\pi\sigma^2\right)-\frac{\chi^2}{2}, \tag{7}$$

where χ^2 is the chi-square statistic, and the constant term (*const*) accounts for the flat prior pdf probability. Minimizing χ^2 yields the optimal solution θ^{opt}, which is usually called the *last-squares estimate*. In the case where $f(\theta)$ is a linear model, the posterior pdf is completely defined by θ^{opt} and its estimated parameter error covariance matrix *C*. The last can be related to the *Hessian matrix, H*, trough the relation

$$C = \frac{\chi^2\left(\theta^{opt}\right)}{n-p} 2H\left(\theta^{opt}\right)^{-1}. \qquad (8)$$

Given the estimated covariance matrix *C* and $n >> p$, the individual parameter confidence interval $CI_i \mid i = 1 \dots p$ is estimated as

$$CI_i = \pm t_{n-p}^{1-(\alpha/2)} \sqrt{C_{ii}}, \qquad (9)$$

where $t_{n-p}^{1-(\alpha/2)}$ is the two-tails Student's *t* distribution for the given confidence level α and ***n-p*** degrees of freedom (Marsili-Libelli *et al.*, 2003).

As we have seen, the last-square procedure is only a reduced Bayesian approach, since linear methods are a good or bad approximation of the Bayesian analysis according to the simplification assumptions considered. If the asymptotic inference fail to provide consistent results, it is still possible to formulate a better statistical prescription for the case in study (Sivia and Skilling, 2006).

Conditions to be fulfilled in order to correctly apply the linear approach for the CIs estimation according to Seber and Wild (1989) are:

i. The prior probability Pr $(\theta \mid I)$ is uninformative.
ii. The measurement and model error ε is $N \sim (0, \sigma^2 I_n)$ and $n \to \infty$.
iii. The last-square θ^{opt} is asymptotically θ^* (the true value of θ).
iv. The posterior probability Pr $(\theta \mid D, I)$ is $N \sim (\theta^{opt}, \sigma^2 FIM^{-1})$, where $FIM = [f_\theta^T f_\theta]$ at θ^{opt}.

Data analysis procedure

The uncertainty analysis procedure was the following:

Step1. The selected model structure was the revised version of the activated sludge model ASM3 for simultaneous storage and growth processes under aerobic conditions (Sin *et al.*, 2005). Parameter definitions and nomenclature used here were taken from that reference. The model was implemented in Matlab (Mathworks Inc., USA) as a Simulink S-function Cmex code block. The notation used is: τ: first-order time constant (min); K_S: S_S affinity constant (mg COD/L); K_1: regulation constant of X_H as function of X_{STO}/X_H (mg COD/mg COD); K_2: a lumped parameter related to the affinity of X_H towards X_{STO}/X_H (mg COD/mg COD); f_{STO}: fraction of S_S used for storage (mg COD/mg COD); q_{MAX}: maximum S_S uptake rate (1/d); δ: efficiency of oxidative phosphorylation (mol/mol); k_{STO}: maximum storage rate of X_H (1/d); $\mu_{MAX,S}$: maximum growth rate of X_H on S_S (1/d). Y_{STO}: yield coefficient for storage on S_S (mg COD/mg COD); $Y_{H,S}$: yield coefficient for growth on S_S (mg COD/mg COD); $Y_{H,STO}$: yield coefficient for growth on X_{STO} (mg COD/mg COD). X_{STO}: storage products (mg COD/L); X_H: biomass (mg COD/L); S_S: substrate (mg COD/L).

Step2. The importance ranking of the parameters influencing the chi-square statistic (χ^2) was performed by probabilistic sensitivity analysis (Oakley and O'Hagan, 2004).

Step3. The last-square estimate for parameter (θ) was performed by a global optimization routine (Rodriguez-Fernandez ***et al***., 2006). This estimate was used in the next step for the MCMC sampler initialization.

Step4. Gaussian likelihood function and a uniform prior pdf were assumed for Bayesian inference in order to make it comparable with the asymptotic inference results. An additive error model was considered, with an uninformative conjugate prior defined by an inverse gamma distribution for ε. The posterior probability Pr $(\theta \mid D, I)$ was approximated with a sample size of 20000, obtained after convergence of the Delayed Rejection Adaptive MCMC (DRAM) sampler (Laine, 2008). Only positive parameter values were allowed in order to preserve their physical meaning. The mean and the variance for the individual parameter θ_i were computed by integration over the estimated posterior distribution. Finally, the corresponding CI_{Bayes} was computed at 5% confidence level. The posterior probability Pr $(\theta \mid D, I)$ were tested for normality with the Jarque-Bera test (Bera and Jarque, 1981). The test rejects the null hypothesis that the sample comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution, at 5% significance level.

Step5. The asymptotic inference CI_{asym} was based on an estimate of the Hessian matrix at the last-square estimate, θ^{ppt}. The estimation of the Hessian matrix was performed by a finite difference high order estimation method, based on Romberg extrapolation with an adaptive method for the determination of the step size perturbation parameters. A free Matlab toolbox called "Adaptive Robust Numerical Differentiation" was used. The accuracy of the results for matrix inversion was given by the *condition number* measure (Dochain and Vanrolleghem, 2001), computed through the singular value decomposition of the matrix in question.

Data sources

The experimental data consisted of three oxygen uptake rate (OUR) profiles, obtained from batch experiments: (A) two acetate pulses of 40 mg COD/L for an optimal experimental design (Sin *et al*., 2005) - COD: Chemical Oxygen Demand -; (B) single acetate pulse of 50 mg COD/L (Hoque *et al*., 2009); (C) single acetate pulse of 40 mg COD/L, where the sludge employed came from a lab-scale sequencing batch reactor, fed with raw leachate under intermittent aeration.

Data sets A and B were digitalized from the original authors paper. In the case of data set C, the composition of the raw leachate used was equivalent to 9.81 g COD-VFA/L (VFA: Volatile Fatty Acid; 25% acetate, 9% propionate, 50% n-butyrate, 2% iso-butyrate, 11% n-valerate, and 3% iso-valerate), 48% COD-VFA/COD, and 1.01 g N/L. Test was carried out in a LFS respirometer with flowing gas and static liquid (Spanjers *et al*., 1998). The OUR signal was estimated from dissolved oxygen measurements by applying an optimal local polynomial filtration paradigm called Lazy Learning (Bontempi and Birattari, 1999). The initial content of storage products in biomass, $X_{STO}(0)$, was assumed 7.6 mg COD/L. The initial concentration of biomass, $X_H(0)$, was calculated from the endogenous OUR as 214 mg COD/L.

RESULTS AND DISCUSSION

All the MCMC estimated posteriors for the three data sets (Figure 1) presented a clear evidence for rejecting the null hypothesis according to the Jarque-Bera test. Anyway, a visual inspection of the posteriors for data set A and B for parameters other than K_1, K_2 and K_S, suggested that those posteriors could be approximated reasonably well by a normal distribution. The main reason of failing the Jarque-Bera test for normality resided in the tails of the posteriors pdf. On the other side, the posterior pdf conditional on data set C was bi-modal, which means that it was still possible to have the same goodness of fit for OUR data with two different sets of parameters. Unfortunately, the only way to solve this problem is to collect more information from the system or to change the model structure. We could be temped to compute the mean of the bi-modal distribution as the "true" parameter value: note that in doing so, the corresponding variance would be very large if the modes were far apart from each other and if their corresponding masses (integrals) were similar.

The estimated posteriors distribution of the standard deviation of the additive error model *std* = σ for all the data sets were quite narrow and with small mean values (Figure 1). This result reflects the good model fit to the measurement data. Anyway, it is interesting to note that *std* do not reflect random measurement noise, but rather the model's structure systematic error (data were filtered).

Taking data set B as example, from the probabilistic sensitivity analysis resulted that the output variance of the χ^2 statistic was highly explained by δ, with 60%, whereas $\mu_{MAX,S}$ explained 10% and τ only 3.6%. The rest of the χ^2 variance was explained by the sum of the remaining main parameters effects (~1.6%) and joint parameters effects (~22.8%). From this analysis it is reasonable to expect that at least δ, $\mu_{MAX,S}$ and τ are practically identifiable. Approximately the same results were obtained for data set A and C.

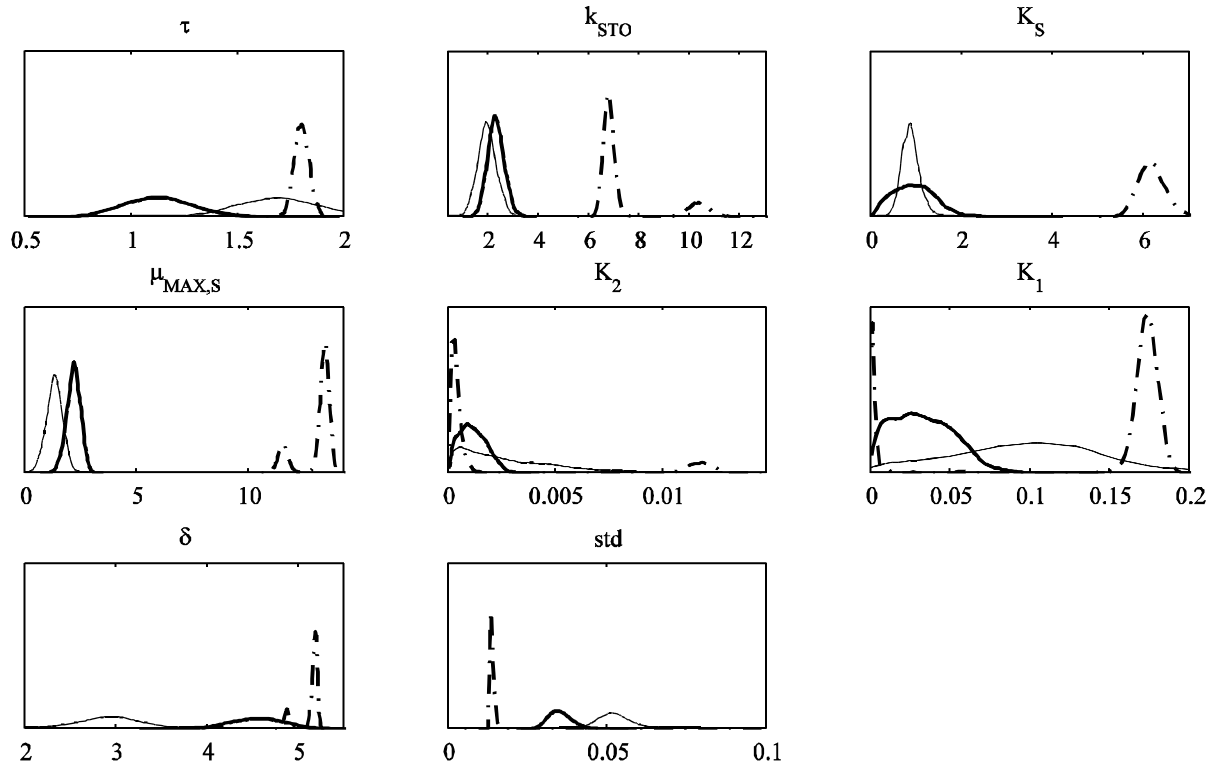

Figure 1. Model parameters marginal density distributions smoothed by a Gaussian kernel for the data set A (bold solid line), the data set B (thin solid line) and the data set C (dash-dot line). The parameter *std* is the unknown standard deviation of the Gaussian distribution error term of the IID additive error model.

The Bayesian mean values from data set A and B (Table 1) were close to the values initially reported by the respective authors (results not shown). The only exemption was the first-order time constant τ in the case of data set A, where the Bayesian mean value was 1.70 ±0.36 min, while the value estimated by Sin *et al*. (2005) was 0.51 ±0.07 min, which falls out the Bayesian 95% probability CI. It was found that Bayesian CIs were approximately one order of magnitude higher than the asymptotic CIs estimated by Sin *et al*. (2005) and Hoque *et al.* (2009). The authors' CIs estimated values and the Bayesian CIs values for the same data sets are reported in Table 1. In order to explain the difference between the Bayesian and asymptotic evaluations of the CIs, we must first note that formally they should have given the same results if all the four listed conditions in section "methods: asymptotic inference" were fulfilled. Assuming that the Bayesian approach is correct, since it is more general, one way to identify which conditions were violated is to remove the parameters that potentially created problems to the asymptotic estimation and to check if the results improve. If Bayesian inference is used, uninformative data over a parameter will simply imply that the posterior of that parameter will not be updated and will remain the same as its prior pdf. In the case asymptotic inference, the absence of information lead to unidentifiability, which means that the matrix inversion operation is impossible or badly conditioned, and possibly leading to high approximation errors of its uncertainty estimates.

The estimated parameters from data set C were not presented because of the bi-modal posterior pdf characteristics. In order to estimate the parameters for data set C, a more informative experimental design should be planned or a new model structure has to be proposed.

Table 1. Comparison between parameters uncertainty estimations from the Bayesian inference and literature results for OUR experiments with acetate substrate for the data set A and B. Parameters in bold font were estimated, while those in normal font were calculated by Monte Carlo propagation.

Parameters	**Data set A**				**Data set B**			
	$mean_{Bayes}$	CI_{Bayes}	CI_{asym} *	R_{CI} a	$mean_{Bayes}$	CI_{Bayes}	CI_{asym} **	R_{CI} a
τ	1.70	0.361	0.07	5.2	1.14	0.339	0.024	14.1
K_S	0.88	0.365	0.1	3.7	1.00	0.883	0.056	15.8
K_2	2.57E-3	3.75E-3	3.00E-3	1.3	1.07E-3	1.13E-3	4.80E-7	2354.2
K_1	0.0944	0.088	0.012	7.3	0.0322	0.037	0.0033	11.2
f_{STO}	0.53	0.207	0.03	6.9	0.46	0.128	0.029	4.4
q_{MAX}	4.56	0.185	0.03	6.2	5.52	0.305	0.023	13.3
δ	2.92	0.586	0.08	7.3	4.58	0.576	0.132	4.4
k_{STO}	1.99	0.713	-	-	2.25	0.624	-	-
$\mu_{MAX,S}$	4.56	0.684	-	-	5.52	0.58	-	-
Y_{STO}	0.83	0.036	-	-	0.89	0.014	-	-
$Y_{H,S}$	0.61	0.059	-	-	0.73	0.029	-	-
$Y_{H,STO}$	0.71	0.044	-	-	0.80	0.021	-	-

[*] CI at 5% confidence level estimated by Sin *et al.* (2005).
[**] CI at 5% confidence level estimated by Hoque *et al.* (2009).
[a] R_{CI} is the rate between CI_{Bayes} and CI_{asym}.

Data set B was chosen for the comparison in the evaluation of the confidence intervals considering both Bayesian and asymptotic inference. The sensitivity analysis ranking for the parameters indicated that the regulation constant of biomass, K_1, and the lumped parameter K_2, were the less important variables in explaining the misfit function variance. Because the marginal posteriors for the parameters K_1 and K_2 were far from a normal distribution and moreover would imply a high conditional number (order of 10^8) of the Hessian matrix, they were omitted. A visual comparison between the asymptotic and the Bayesian confidence ellipses (Figure 2) showed that after K_1 and K_2 are removed from the analysis, the differences diminish. The conditional number of the Hessian matrix improved, decreasing to 10^2.The highest mismatch was observed for the parameter K_S, which was probably due to the assumed positivity constrains for the parameters and to the not valid normality assumption.

CONCLUSIONS

If a nonlinear model is considered, there is no guarantee that the asymptotic normal approximation to the likelihood function will hold: posterior distributions may be far from Gaussian, have a nonlinear correlation structure between the unknowns and be multi-modal.

In the current case study, if unidentifiable parameters were considered, it was found that asymptotic inference CIs were approximately one order of magnitude lower than the Bayesian CIs.

Bayesian uncertainty estimation has one important advantage over the classical linear method: in the case where data contains no information on a parameter, the posterior of that parameter is simply not updated and will remain very close to its prior, providing credible intervals instead of confidence intervals. On the other side, since asymptotic inference methods are based on the matrix inversion operation, the absence of information lead to unidentifiability, which means that matrix inversion is impossible or has very high approximation errors. The last case is the most dangerous if the conditional number of the matrix has not been checked. In our case, if only practically identifiable parameters were considered, the linear uncertainty analysis was still considered reliable. This is because the Hessian matrix inversion operation resulted more reliable and the normality assumption was not heavily violated.

In any case, since we could not know a priori if the normality assumption is going to hold for nonlinear models, it is advisable to rely on methods that explore the whole support of the posterior parameter probability distribution Pr $(\theta \mid D, I)$, not just at a single base-line values of the last-square estimate parameter θ^{pt}. This will make little difference in the conclusion for some simple models and where uniform IID priors assumed. However, for complex nonlinear models there is no guarantee that the posterior will be Gaussian (e.g., multi-modal pdf with nonlinear correlation structure).

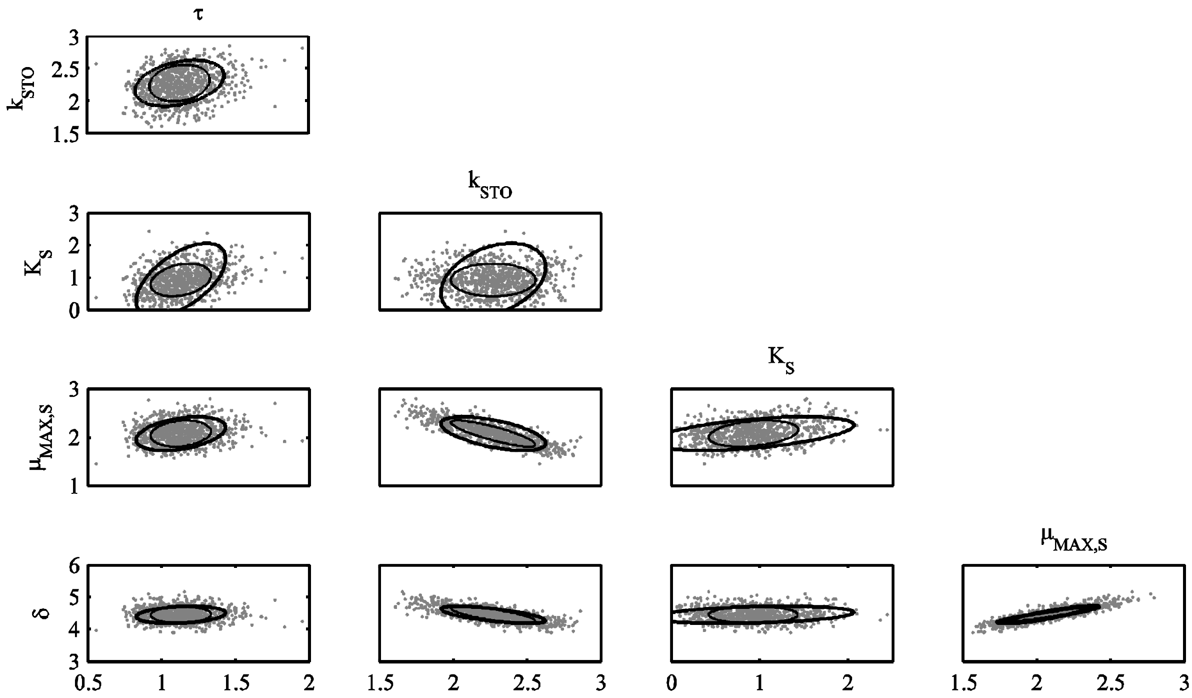

Figure 2. Bayesian inference scatter plot for the chain pairs (grey dots) for data set B, with the asymptotic 95% probability confidence ellipse (thin line), and the DRAM estimated covariance matrix 95% probability confidence ellipse (bold line).

Acknowledgements

This work was partially supported by the Spanish Ministry of Science and Innovation (Project CTM2010-18212).

REFERENCES

Bayarri M. J., Berger J. O., Paulo R., Sacks J., Cafeo J. A., Cavendish J., Lin C. H. and Tu J. (2007). A framework for validation of computer models. *Technometrics,* **49**(2), 138-154.

Bera A. K. and Jarque C. M. (1981). Efficient tests for normality, heteroskedasticity and serial independence of regression residuals: Monte Carlo evidence. *Econ. Lett.*, **7**(4), 313-318.

Bontempi G. and Birattari M. (1999). Lazy learning for local modeling and control design. *Int. J. Control*, **72**(7/8), 643-658.

Dochain D. and Vanrolleghem P. A. (2001). *Dynamical Modelling and Estimation in Wastewater Treatment Processes.* IWA Publishing, UK.

Hoque M. A., Aravinthan V. and Pradhan N. M. (2009). Assessment on activated sludge models for acetate biodegradation under aerobic conditions. *Wat. Sci. Technol.,* **60**(4), 983-994.

Laine M. (2008). *Adaptive MCMC Methods With Applications In Environmental And Geophysical Models*. PhD thesis, Department of Mathematics and Physics, Lappeenranta University of Technology, Finland.

Marsili-Libelli S., Guerrizio S. and Checchi N. (2003). Confidence regions of estimated parameters for ecological systems. *Ecol. Mod.*, **165**(2-3), 127-146.

Neumann M. B. and Gujer W. (2008). Underestimation of uncertainty in statistical regression of environmental models: influence of model structure uncertainty. *Environ. Sci. Technol.,* **42**(11), 4037-4043.

Oakley J. E. and O'Hagan A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J. R. Stat. Soc. B,* **66**(3), 751-769.

Rodriguez-Fernandez M., Egea J. A. and Banga J. R. (2006). Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. *BMC Bioinformatics*, **7**(1), 483.

Seber G. A. F. and Wild C. J. (1989). *Nonlinear Regression.* John Wiley and Sons Ltd., Canada.

Sin G., Guisasola A., De Pauw D. J. W., Baeza J. A., Carrera J. and Vanrolleghem P. A. (2005). A new approach for modelling simultaneous storage and growth processes for activated sludge systems under aerobic conditions. *Biotechnol. Bioeng.,* **92**(5), 600-613.

Sivia D. and Skilling J. (2006). *Data Analysis: A Bayesian Tutorial*, Oxford University Press, USA.

Spanjers H., Vanrolleghem P. A., Olsson G. and Dold P. L. (1998). *Respirometry in Control of the Activated Sludge Process: Principles*. IAWQ. Scientific and Technical Report, 7. UK.