Documentation on Uncertainty quantification routines

1 Fisher's Information and Cramér–Rao's bound

Fisher's information matrix (FIM) encodes the insight yielded by the observations on the parameters' values. A key motivation behind its use is Fréchet–Cramér–Rao's bound, also known as the information bound or Cramér–Rao's bound, which states that unbiased estimators necessarily have a variance greater than the inverse FIM, computed at the true set of parameters [Lehmann and Casella, 1998]. In linear regression with Gaussian noise setting, this lower bound is achieved when the number of observations exceeds the number of parameters.

Similar to the standard Bayesian framework, Fisher's information requires statistical modelling. Coherently with the way the observations were noised, the statistical model **??** is considered. In this setting, Fisher's information matrix is defined as

$$\mathrm{FIM}(\theta^*) = \sigma^{-2} J_{\mathrm{AD}}(\theta^*) J_{\mathrm{AD}}(\theta^*)^T \tag{1}$$

where $J_{\mathrm{AD}}(\theta)_{i,j} = \frac{\partial \log(\mathrm{AD})_j}{\partial \theta_i}$ is the Jacobian matrix of $\log \circ \mathrm{AD}$. Assuming the estimator $\hat{\theta}$ is unbiased, Fréchet–Cramér–Rao's bound states that its covariance is lower bounded by

$$\mathrm{Cov}(\hat{\theta}) \geq \mathrm{FIM}^{-1}(\theta^*) = \sigma^2 \left(J_{\mathrm{AD}}(\theta^*) J_{\mathrm{AD}}(\theta^*)^T \right)^{-1}.$$

Since neither the exact θ^* used to generate the data, nor the noise level σ are known, both are replaced with estimates. All in all, the uncertainty is encoded in the following covariance on the parameters

$$\widehat{\mathrm{Cov}}(\hat{\theta}) = \hat{\sigma}^2 (J_{\mathrm{AD}}(\hat{\theta}) J_{\mathrm{AD}}(\hat{\theta})^T)^{-1}. \tag{2}$$

The covariance was used to construct confidence regions shaped as ellipsoids,

$$CR_{\mathrm{FIM}}(\alpha) := \{\theta \mid (\theta - \hat{\theta}) \mathrm{FIM} (\theta - \hat{\theta}) \leq Q_{\chi^2(p)}(\alpha)\}$$

where $Q_{\chi^2(p)}$ is the quantile function of a $\chi^2(p)$ distribution. Under the hypothesis that the estimator have a Gaussian behavior, as would be the case in linear regression, these confidence regions have coverage α.

For AM2, the derivative of the AD model with respect to the distribution's parameters was obtained through the resolution of an ordinary differential equation (ODE), which can be efficiently performed while computing the predictions. Indeed, AM2 is fully described by an ODE, where the successive states $S(t)$ of the digester satisfy

$$\begin{cases} \frac{\mathrm{dAM2}}{\mathrm{d}t} = F(\theta, I_t, \mathrm{AM2}_t) \\ \mathrm{AM2}_{t=0} = S_0 \end{cases}$$

As such, the derivative of the state with respect to the parameters satisfy the following differential equation

$$\begin{cases} \frac{\mathrm{d}^2\mathrm{AM2}}{\mathrm{d}t\mathrm{d}\theta} = \mathrm{d}_\theta F + \mathrm{d}_S F \ \frac{\mathrm{dAM2}}{\mathrm{d}\theta} \\ \frac{\mathrm{dAM2}}{\mathrm{d}\theta}(t=0) = 0. \end{cases}$$

The second ODE can then be solved efficiently while solving the first ODE.

Such a method can not be used directly for ADM1 since corrections are applied to the ODE results. As such, for ADM1, the derivative was computed using a two-points symmetric finite difference scheme. This second approach occasionally failed due to instabilities in the behavior of the ODE solver, and the maximum time step of the ODE was manually adjusted until a suitable estimation of the derivative was found.

Equation (2) implies that $\widehat{\mathrm{Cov}}$ is non negative. However, since Fisher's information involves parameters whose typical values differ by several order of magnitude, the ratio between the highest and lowest eigenvalues (conditioning number) is usually large ($> 10^{16}$), and therefore, numerical approximations can result in estimated small, negative eigenvalues. As a precaution, all eigenvalues of FIM were raised to 10^{-8} at least. This implies that the covariance has all its eigenvalues upper bounded by 10^8.

Code for FIM's procedure can be found in the 'aduq' library. (function 'fim' in 'aduq.uncertainty' module).

2 Beale's criteria

Beale's criteria (see Beale [1960] and Dochain and Vanrolleghem [2001], Section 6.7.3, for a complete description) describes uncertainty on the parameter values in the form of confidence regions which can have any shape. These are defined by considering all the sets of parameters whose score are below a statistically motivated threshold. This threshold is constructed by considering the linear regression with Gaussian noise case. Considering as the score function the sum of squared residuals S^2, and noting $\hat{\theta}$ the

minimizer of the score, the statistic $\frac{N-p}{p}\frac{S^2(\theta^*)-S^2(\hat{\theta})}{S^2(\hat{\theta})}$ follows an $F(p, N-p)$ distribution. This implies that

$$CR_{\text{Beale}}(\alpha) := \left\{ \theta \mid S^2(\theta) \leq T_\alpha, T_\alpha := S^2(\hat{\theta}) \left(1 + \frac{p}{N-p} Q_{F(p,N-p)}(\alpha) \right) \right\} \tag{3}$$

where $Q_{F(p,N-p)}$ is the quantile function of the $F(p, N-p)$ distribution, is a confidence region of level α. Beale's UQ consists in using these confidence regions, even in nonlinear setting.

In order to estimate the confidence region, the line search procedure advocated in Dochain and Vanrolleghem [2001] was implemented and performed on the log parameters. The covariance obtained through the inverse FIM was used to obtain an initial guess of the uncertainty region. To avoid numerical abberations, parameters whose uncertainty in log was larger than 6 were set to the optimized value, and uncertainty was not ascertained on them. An initial sample of points $(\theta_i)_i$ on the boundary of FIM's confidence region was drawn at random (2048 points for ADM1, 5120 points for AM2). A line search algorithm was used to solve $S^2(\lambda(\theta_i - \hat{\theta}) = T_\alpha$ for each θ_i. If the line search algorithm failed to find a parameter achieving the threshold with a precision of less than $0.01 * (T_\alpha - S^2(\hat{\theta})$ in less than 20 steps, the point was removed from the set. The final set $(\lambda_i\theta_i + (1-\lambda_i)\hat{\theta})_i$ was used as the approximative boundary of Beale's confidence region.

Code for Beale's procedure can be found in the 'aduq' library ('beale_boundary' in 'aduq.uncertainty' module).

3 Residual bootstrapping

Residual bootstrapping is designed to perform UQ when the noise structure is unknown. After calibration, the residuals are assumed to be representative of the noise, and samples of the noise are used to generate new observations. These are in turn transformed into samples of sets of parameters through recalibration. As such, the residual bootstrapping procedure describes uncertainty in the form of a sample of sets of parameters.

The bootstrap procedure used is similar to the one used in Gonzalez-Gil et al. [2018] and used in the context of AD modeling by Regueira et al. [2021]. Log-residuals of the calibrated model are drawn with replacement and added to the log of the calibrated model output to generate new observations. New calibration procedures with these bootstrapped observations are then performed to compute new sets of parameters. 512 bootstrapped sets of parameters were constructed for each dataset.

In order to limit the computational cost of the procedure, the new calibration procedures were restarted from the initial calibrated set of parameters, with less stringent convergence criteria (tolerance in score of $2e-4$ compared to $e-8$ for the initial calibration procedure, and a maximum of 30 optimisation steps compared to 250).

While this decreases the ability of the bootstrap procedure to construct sets of parameters far from the calibrated model, it proved necessary to control the computational cost.

This procedure was assessed only on the AM2 model, since its computational cost would have been prohibitive for ADM1.

Code for the bootstrap procedure can be found in the 'aduq' library ('bootstrap' in 'aduq.uncertainty' module).

Table 1: Overview of UQ routines

	Bootstrap	FIM	Beale	VarBUQ
Confidence region shape	Any	Ellipse	Any*	User chosen**
Output	Samples	Covariance	Samples	Distribution
Prerequisite	None	Statistical Model	Mean squared error calibration	Prior
Theory	None	Cramer-Rao	Beale	PAC-Bayes
Error hypothesis	None	Gaussian***	Gaussian	Bounded
Model hypothesis	None	Linear***	Linear	None

* The algorithm approximating Beale's confidence regions assumes they are connex.
** Through probability distribution class.
**** While Cramer-Rao's result holds for any statistical model, it only provides an equality for Gaussian noise and linear model.

4 Computation of p-values

For FIM UQ, the p-value is computed using the statistic $(\hat{\theta}-\theta^*)\mathrm{Cov}_{\mathrm{FIM}}^{-1}(\hat{\theta}-\theta^*)$, which, under the null hypothesis in linear regression setting, is distributed as a $\chi^2(p)$ where p is the number of parameters calibrated. Since VarBUQ considers gaussian posteriors on the log parameters, the similar statistic $(\hat{\pi}[\log(\theta)]-\log(\theta^*))\mathrm{Cov}_{\log\hat{\pi}}^{-1}(\hat{\pi}[\log(\theta)]-\log(\theta^*))$ was used, where $\mathrm{Cov}_{\log\hat{\pi}}$ denotes the covariance of $\log(\theta)$ where θ is drawn from $\hat{\pi}$.

For Beale's criteria, the p-values are computed using the theoretical rather than estimated confidence regions. As such, the key statistic is $\frac{N-p}{p}(S^2(\theta^*)-S^2(\hat{\theta}))/S^2(\hat{\theta})$ which, under the null hypothesis, should be distributed as $F(p,N-p)$.

For the bootstrap procedure, the computation of the p-value is slightly more complicated. An anomaly score function A is learned on the first half of the sample using Isolation Forest algorithm [Liu et al., 2008], then evaluating the anomaly score on the second half of the bootstrapped sample, resulting in a sample of scores $A(\theta_i)_{i>n/2}$. The p-value is obtained by considering the quantile achieved by the anomaly score of the true set of parameters, $A(\hat{\theta})$, i.e $\hat{p}=\frac{2}{n}\sum_{i>\frac{n}{2}}1\{A(\hat{\theta})<A(\theta_i)\}$. For a sufficiently large sample, the central limit theorem implies that this quantity does converge to the p-value for a family of tests assessing whether a parameter is drawn from the bootstrapped sample. Tight high confidence upper bounds for the p-value given infinite sample size can be computed by noting that $A(\hat{\theta})$ is a mean of independent Bernoulli of mean p, and that

$$\left[0,\hat{p}^+:=\sup\left\{\tilde{p}\mid\ \sum_{i=0}^{n*\hat{p}}\binom{n}{i}\tilde{p}^i(1-\tilde{p})^{n-i}\geq\alpha\right\}\right]$$

is a confidence interval of p of level α, i.e. $p \leq \hat{p}^+$ with probability α.

Code for the computation of the p-values can be found in 'aduq' library (functions 'fim_pval', 'beale_pval', 'sample_pval'. 'sample_pval' is used for Bootstrap. Note that 'fim_pval' function can also be used to compute the p-value for VarBUQ, using as arguments the log parameter values, mean of the log parameter and covariance computed in log space).

5 Construction of confidence regions

For each UQ routine, the confidence region of level α is constructed as the set of parameters which would be given a p-value higher than $1-\alpha$ when testing whether they generated the observations.

For FIM, this can be easily computed as the set of parameters θ such that $(\hat{\theta} - \theta)\text{Cov}_{\text{FIM}}^{-1}(\hat{\theta} - \theta) \leq q_{\chi 2(p)}(\alpha)$ where p is the number of parameters fitted. Similarly, the confidence regions for VarBUQ are the sets of parameters θ such that $(\hat{\pi}[\log(\theta)] - \log(\theta^*))\text{Cov}_{\log \hat{\pi}}^{-1}(\hat{\pi}[\log(\theta)] - \log(\theta^*)) < q_{\chi 2(p)}(\alpha)$.

For Beale, the theoretical confidence regions are the sets of parameters θ such that $\frac{N-p}{p}(S^2(\theta) - S^2(\hat{\theta}))/S^2(\hat{\theta}) < q_{F(p,N-p)}(\alpha)$.

For Bootstrap, the confidence regions of level α consists of sets of parameters θ whose anomaly score (computed using the methodology described in the previous section) is lower than the quantile $1-\alpha$ of the anomaly scores on the validation set of bootstrapped samples.

6 Representation of confidence regions

Representation of a confidence region is obtained by projecting it on 2 dimensions. For FIM and VarBUQ, exact uncertainty region could be computed using 95% confidence ellipses for two dimensional Gaussians. For Bootstrap and Beale, the two dimensional confidence regions were approximated by projecting the sample and using alpha shapes to obtain a boundary [Edelsbrunner, 2011]. For the bootstrap method, 5% of the sample with highest anomaly score - as computed using Isolation Forest - are removed before constructing the boundary in order to obtain an approximative 95% confidence region.

Code for two dimensional representations of the confidence region can be found in 'aduq' library (module 'aduq.uncertainty.plot').

7 Evaluation of the uncertainty on predictions

To limit the risk of underestimating the uncertainty in the early phase of the test set, all calls to the AD models are started from day 166, 30 days before the beginning of the test set, using as initial description of the digester the result of the simulation at that

day for the calibrated set of parameters (the optimised set of parameters for residual based UQ methods and the mean of the posterior for VarBUQ).

From a statistical viewpoint, the methodology used to extrapolate uncertainty from confidence region constructs confidence intervals on the predictions with matching level. However, the transfer of uncertainty guarantees that all predictions should be inside their confidence intervals with probability at least α, and not that a fraction α of the unnoised signal should be in the confidence intervals. As such, it can not be assumed that 95% confidence regions should result in 95% coverage for a good UQ procedure - all the more so as neither the UQ methods hypothesis are met, nor is the statistical model supporting them correct. Still, since confidence regions with high confidence levels are assessed, the coverage indicator of properly working UQ methods should be high, the target coverage is set informally at 95%.

References

E. M. L. Beale, Confidence regions in non-linear estimation, Journal of the Royal Statistical Society: Series B (Methodological) 22 (1960) 41–76. URL: https://doi.org/10.1111/j.2517-6161.1960.tb00353.x.

D. Dochain, P. Vanrolleghem, Dynamical Modelling and Estimation in Wastewater Treatment Processes, IWA Publishing, 2001. URL: https://doi.org/10.2166/9781780403045.

H. Edelsbrunner, Alpha shapes-a survey, in: Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings, 2011. URL: http://graphics.stanford.edu/courses/cs268-14-fall/Handouts/AlphaShapes/2010-B-01-AlphaShapes.pdf.

L. Gonzalez-Gil, M. Mauricio-Iglesias, M. Carballa, J. M. Lema, Why are organic micropollutants not fully biotransformed? a mechanistic modelling approach to anaerobic systems, Water Research 142 (2018) 115–128. URL: http://dx.doi.org/10.1016/j.watres.2018.05.032.

E. L. Lehmann, G. Casella, Theory of Point Estimation, 2nd ed., Springer New York, NY, 1998. URL: https://doi.org/10.1007/b98854.

F. T. Liu, K. M. Ting, Z.-H. Zhou, Isolation forest, in: 2008 Eighth IEEE International Conference on Data Mining, IEEE, 2008. URL: https://doi.org/10.1109/icdm.2008.17.

A. Regueira, R. Bevilacqua, M. Mauricio-Iglesias, M. Carballa, J. Lema, Kinetic and stoichiometric model for the computer-aided design of protein fermentation into volatile fatty acids, Chemical Engineering Journal 406 (2021) 126835. URL: https://doi.org/10.1016/j.cej.2020.126835.