Calibration: deterministic and bayesian methods PRACE training on High Performance Computing in Uncertainty Quantification

M. Baudin

February 4, 2020

Contents

Deterministic code calibration

We consider a computer model $\boldsymbol{h}$ (i.e. a deterministic function) to calibrate:

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{\theta}),$$

where

- $\boldsymbol{x} \in \mathbb{R}^{d_x}$ is the input vector;
- $\boldsymbol{z} \in \mathbb{R}^{d_z}$ is the output vector;
- $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$ are the unknown parameters of $\boldsymbol{h}$ to calibrate.

Let $n \in \mathbb{R}$ be the number of observations. The standard hypothesis of the probabilistic calibration is:

$$\boldsymbol{Y}^i = \boldsymbol{z}^i + \varepsilon^i,$$

for $i = 1, ..., n$ where ε^i is a random measurement error such that:

$$E(\varepsilon) = \mathbf{0} \in \mathbb{R}^{d_z}, \qquad Cov(\varepsilon) = \Sigma \in \mathbb{R}^{d_z \times d_z},$$

where Σ is the error covariance matrix.

The goal of calibration is to estimate $\boldsymbol{\theta}$, based on observations of n inputs $(\boldsymbol{x}^1, \ldots, \boldsymbol{x}^n)$ and the associated n observations of the output $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n)$.
In other words, the calibration process reduces the discrepancy between

- the observations $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n)$ and
- the predictions $\boldsymbol{h}(\boldsymbol{\theta})$.

Given that $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n)$ are realizations of a random variable, the estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is also a random variable.
Hence, the secondary goal of calibration is to estimate the distribution of $\hat{\boldsymbol{\theta}}$ representing the uncertainty of the calibration process.

The standard observation model makes the hypothesis that the covariance matrix of the error is diagonal, i.e.

$$\Sigma = \sigma^2 \mathbf{I}$$

where $\sigma^2 \in \mathbb{R}$ is the constant observation error variance.
In the remaining of this section, the input $\boldsymbol{x}$ is not involved anymore in the equations.
This is why we simplify the equation into:

$$\boldsymbol{z} = \boldsymbol{h}(\boldsymbol{\theta}).$$

Least squares

The residuals is the difference between the observations and the predictions:

$$\boldsymbol{r}^i = \boldsymbol{y}^i - \boldsymbol{h}(\boldsymbol{\theta})^i$$

for $i = 1, ..., n$. The method of least squares minimizes the square of the Euclidian norm of the residuals. This is why the least squares method is based on the cost function C defined by:

$$C(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{\theta})\|^2 = \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}^i - \boldsymbol{h}(\boldsymbol{\theta})^i\right)^2,$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.

The least squares method minimizes the cost function C:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d_h}}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{\theta})\|^2.$$

The unbiased estimator of the variance is:

$$\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{\theta})\|^2}{n - d_h}.$$

Notice that the previous estimator is not the maximum likelihood estimator (which is biased).

Linear least squares

In the particular case where the deterministic function $\boldsymbol{h}$ is linear with respect to the parameter $\boldsymbol{\theta}$, then the method reduces to the linear least squares.
Let $J \in \mathbb{R}^{n \times d_h}$ be the Jacobian matrix made of the partial derivatives of $\boldsymbol{h}$ with respect to $\boldsymbol{\theta}$:

$$J(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}.$$

Let $\boldsymbol{\mu} \in \mathbb{R}^{d_h}$ be a reference value of the parameter $\boldsymbol{\theta}$.
Let us denote by $J = J(\boldsymbol{\mu})$ the value of the Jacobian at the reference point $\boldsymbol{\mu}$.
Since the function is, by hypothesis, linear, the Jacobian is independent of the point where it is evaluated. Since $\boldsymbol{h}$ is linear, it is equal to its Taylor expansion:

$$\boldsymbol{h}(\boldsymbol{\theta}) = \boldsymbol{h}(\boldsymbol{\mu}) + J(\boldsymbol{\theta} - \boldsymbol{\mu}),$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.

The corresponding linear least squares problem is:

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d_h}}{\arg\min} \frac{1}{2} \|\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{\mu}) + J(\boldsymbol{\theta} - \boldsymbol{\mu})\|^2.$$

The Gauss-Markov theorem applied to this problem states that the solution is:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu} + \left(J^T J\right)^{-1} J^T (\boldsymbol{y} - \boldsymbol{h}(\boldsymbol{\mu})).$$

The previous equations are the *normal equations*.
Notice, however, that the previous linear system of equations is not implemented as is, i.e. we generally do not compute and invert the Gram matrix $J^T J$.
Alternatively, various orthogonalization methods such as the QR or the SVD decomposition can be used to solve the linear least squares problem so that potential ill-conditioning of the normal equations is mitigated.

This estimator can be proved to be the best linear unbiased estimator, the *BLUE*, that is, among the unbiased linear estimators, it is the one which minimizes the variance of the estimator.
Assume that the random observations are gaussian:

$$\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Therefore, the distribution of $\hat{\boldsymbol{\theta}}$ is:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 J^T J).$$

Non Linear Least squares

In the general case where the function $\boldsymbol{h}$ is non linear with respect to the parameter $\boldsymbol{\theta}$, then the resolution involves a non linear least squares optimization algorithm. Instead of directly minimizing the squared Euclidian norm of the residuals, most implementations rely on the residual vector, which lead to an improved accuracy.
The difficulty in the nonlinear least squares is that, compared to the linear situation, the theory does not provide the distribution of $\hat{\boldsymbol{\theta}}$ anymore.
There are two practical solutions to overcome this limitation.

- bootstrap (randomly resample within the observations),
- linearization (in the neighborhood of $\hat{\boldsymbol{\theta}}$).

Link with likelihood maximization

Assume that the observation noise is gaussian with zero mean and constant variance σ^2:

$$\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}),$$

where $\sigma > 0$ et $\mathbf{I} \in \mathbb{R}^{n \times n}$.
This implies that the observations are independent.
The likelihood of the i-th observation is:

$$\ell(\mathbf{y}_i | \theta, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - H_i(\theta))^2}{2\sigma^2} \right)$$

for $i = 1, ..., n$.

Since the observations are independent, the likelihood of the observations is the product:

$$\ell(\mathbf{y}|\theta,\sigma^2) = \prod_{i=1}^{n} \ell(\mathbf{y}_i|\theta,\sigma^2)$$

for $i = 1, ..., n$.
This implies:

$$\log(\ell(\mathbf{y}|\theta,\sigma^2)) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\|\mathbf{y} - H(\theta)\|_2^2}{2\sigma^2}$$

for any $\theta \in \mathbb{R}^p$ and $\sigma > 0$.
We maximize the likelihood with:

$$\hat{\theta} = \mathit{argmin}_{\theta\in\mathbb{R}^p} \frac{1}{2}\|H(\theta) - \mathbf{y}\|_2^2$$

and:

$$\hat{\sigma}^2 = \frac{1}{n}\|H(\hat{\theta}) - \mathbf{y}\|_2^2.$$

The bayesian calibration framework is based on two hypotheses.
The first hypothesis is that the parameter $\boldsymbol{\theta}$ has a known distribution, called the *prior* distribution, and denoted by $p(\boldsymbol{\theta})$.
The second hypothesis is that the output observations $(\boldsymbol{y}^1, \ldots, \boldsymbol{y}^n)$ are sampled from a known conditional distribution denoted by $p(\boldsymbol{y}|\boldsymbol{\theta})$.
For any $\boldsymbol{y} \in \mathbb{R}^{d_z}$ such that $p(\boldsymbol{y}) > 0$, the Bayes theorem implies that the conditional distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y}$ is:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.

The denominator of the previous Bayes fraction is independent of $\boldsymbol{\theta}$, so that the posterior distribution is proportional to the numerator:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.
In the Gaussian calibration, the two previous distributions are assumed to be Gaussian.
More precisely, we make the hypothesis that the parameter $\boldsymbol{\theta}$ has the Gaussian distribution:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}, B),$$

where $\boldsymbol{\mu} \in \mathbb{R}^{d_h}$ is the mean of the Gaussian prior distribution, which is named the *background* and $B \in \mathbb{R}^{d_h \times d_h}$ is the covariance matrix of the parameter.

Secondly, we make the hypothesis that the output observations have the conditional gaussian distribution:

$$\boldsymbol{y}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{h}(\boldsymbol{\theta}), R),$$

where $R \in \mathbb{R}^{d_z \times d_z}$ is the covariance matrix of the output observations.

Posterior distribution

Denote by $\|\cdot\|_B$ the Mahalanobis distance associated with the matrix B :

$$\|\boldsymbol{\theta} - \boldsymbol{\mu}\|_B^2 = (\boldsymbol{\theta} - \boldsymbol{\mu})^T B^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}),$$

for any $\boldsymbol{\theta}, \boldsymbol{\mu} \in \mathbb{R}^{d_h}$. Denote by $\|\cdot\|_R$ the Mahalanobis distance associated with the matrix R :

$$\|\boldsymbol{y} - H(\boldsymbol{\theta})\|_R^2 = (\boldsymbol{y} - H(\boldsymbol{\theta}))^T R^{-1} (\boldsymbol{y} - H(\boldsymbol{\theta})).$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$ and any $\boldsymbol{y} \in \mathbb{R}^{d_z}$. Therefore, the posterior distribution of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$ is :

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left(-\frac{1}{2}\left(\|\boldsymbol{y} - H(\boldsymbol{\theta})\|_R^2 + \|\boldsymbol{\theta} - \boldsymbol{\mu}\|_B^2\right)\right)$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.

MAP estimator

The maximum of the posterior distribution of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$ is reached at :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^{d_h}}{\arg\min} \frac{1}{2} \left(\|\boldsymbol{y} - H(\boldsymbol{\theta})\|_R^2 + \|\boldsymbol{\theta} - \boldsymbol{\mu}\|_B^2 \right).$$

It is called the *maximum a posteriori* estimator or *MAP* estimator.

Regularity of solutions of the Gaussian Calibration

The gaussian calibration is a tradeoff, so that the second expression acts as a *spring* which pulls the parameter θ closer to the background μ (depending on the "spring constant" B, meanwhile getting as close a possible to the observations.
Depending on the matrix B, the computation may have better regularity properties than the plain non linear least squares problem.

Non Linear Gaussian Calibration : 3DVAR

The cost function of the gaussian nonlinear calibration problem is :

$$C(\boldsymbol{\theta}) = \frac{1}{2}\|\boldsymbol{y} - H(\boldsymbol{\theta})\|_R^2 + \frac{1}{2}\|\boldsymbol{\theta} - \boldsymbol{\mu}\|_B^2$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.
The goal of the non linear gaussian calibration is to find the value of $\boldsymbol{\theta}$ which minimizes the cost function C. In general, this involves using a nonlinear unconstrained optimization solver.
Let $J \in \mathbb{R}^{n \times d_h}$ be the Jacobian matrix made of the partial derivatives of $\boldsymbol{h}$ with respect to $\boldsymbol{\theta}$:

$$J(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{\theta}}.$$

If the covariance matrix B is positive definite, then the Hessian matrix of the cost function is positive definite (no matter of the rank of J).

Linear Gaussian Calibration : bayesian BLUE

We make the hypothesis that h is linear with respect to $\boldsymbol{\theta}$, i.e., for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$, we have :

$$h(\boldsymbol{\theta}) = h(\boldsymbol{\mu}) + J(\boldsymbol{\theta} - \boldsymbol{\mu}),$$

where J is the constant Jacobian matrix of h.
Let A be the matrix:

$$A^{-1} = B^{-1} + J^T R^{-1} J.$$

We denote by K the Kalman matrix:

$$K = A J^T R^{-1}.$$

The maximum of the posterior distribution of $\boldsymbol{\theta}$ given the observations $\boldsymbol{y}$ is:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\mu} + K(\boldsymbol{y} - H(\boldsymbol{\mu})).$$

It can be proved that:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left(\frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T A^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})\right)$$

for any $\boldsymbol{\theta} \in \mathbb{R}^{d_h}$.
This implies:

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, A)$$

General bayesian calibration

If the full posterior distribution is to be known, the general bayesian framework may be used.
This allows to:

- use whatever (proper) prior distribution (no need to restrict to gaussian prior),
- use whatever computer code $\boldsymbol{h}$ (no need to restrict to linear code).

One method is to create a Monte-Carlo Markov Chain ; the classical algorithm is then the Metropolis-Hastings algorithm.
However, sampling from the posterior distribution requires to generate a large sample size, which is impractical when the $\boldsymbol{h}$ is costly.
In this case, using a surrogate model is *mandatory*.

- N. H. Bingham and John M. Fry (2010). Regression, Linear Models in Statistics, Springer Undergraduate Mathematics Series. Springer.
- S. Huet, A. Bouvier, M.A. Poursat, and E. Jolivet (2004). Statistical Tools for Nonlinear Regression, Springer.
- C.E. Rasmussen and C. K. I. Williams (2006), Gaussian Processes for Machine Learning, The MIT Press.