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

M. Baudin

January 2, 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:

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

for $i = 1, ..., n$ where $\boldsymbol{\varepsilon}^i$ 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:

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

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:

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

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-conditionning 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 neighbourhood 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:

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

and:

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