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

J. Pelamatti, M. Baudin

April 25, 2022

Contents

Deterministic code calibration

We consider a deterministic model g to calibrate:

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

where

- $p \in \mathbb{N}$ is the number of parameters,
- $m \in \mathbb{N}$ is the number of inputs,
- $\mathbf{x} \in \mathbb{R}^m$ is the input vector;
- $z \in \mathbb{R}$ is the output;
- $\boldsymbol{\theta} \in \mathbb{R}^p$ are the unknown parameters of g to calibrate.

Introduction

Let $n \in \mathbb{N}$ be the number of observations. We assume that $m \geq p$, i.e. the problem is over-determined. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^m$ be the observed inputs.
Therefore,

$$z_i = g(\mathbf{x}_i, \boldsymbol{\theta}),$$

for $i = 1, ..., n$. The vector of predictions is $\mathbf{z} \in \mathbb{R}^n$.
Let $\mathbf{h} : \mathbb{R}^p \to \mathbb{R}^n$ be the function:

$$h_i(\boldsymbol{\theta}) = g(\mathbf{x}_i, \boldsymbol{\theta})$$

for $i = 1, \ldots, n$. We write it in the vector form:

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

where $\mathbf{h}(\boldsymbol{\theta}) = (h_1(\boldsymbol{\theta}), \ldots, h_n(\boldsymbol{\theta}))^T$.

The standard hypothesis of the probabilistic calibration is the following. There is a *true*, but unknown, value of the vector $\boldsymbol{\theta}$, denoted by $\boldsymbol{\theta}^\star$ such that:

$$\mathbf{y} = \mathbf{h}(\boldsymbol{\theta}^\star) + \boldsymbol{\varepsilon},$$

for $i = 1, ..., n$ where $\boldsymbol{\varepsilon}$ is a random measurement error vector such that:

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0} \in \mathbb{R}^n, \qquad \mathbf{Cov}(\boldsymbol{\varepsilon}) = R^\star \in \mathbb{R}^{n \times n},$$

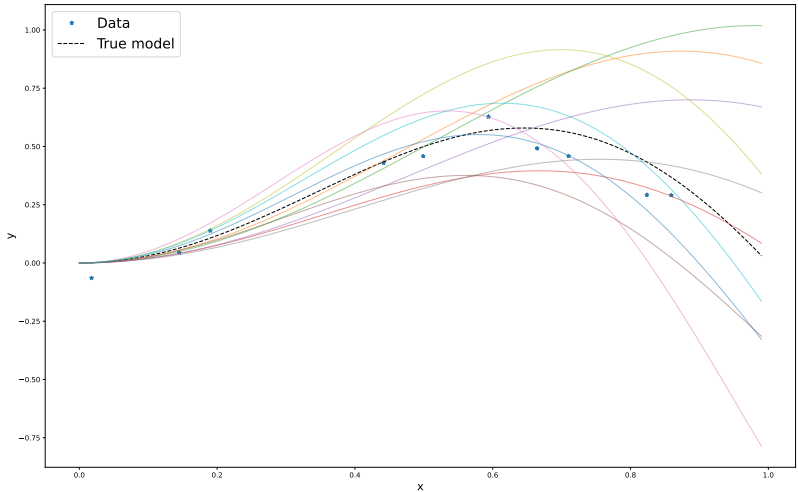
where $R^\star$ is an unknown semi-positive definite covariance matrix.

The goal of calibration is to estimate $\boldsymbol{\theta}$, based on observations of n inputs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and the associated n observations of the output $\mathbf{y}$.
In other words, the calibration process reduces the discrepancy between

- the observations $\mathbf{y}$ and
- the predictions $\mathbf{h}(\boldsymbol{\theta})$.

Given that $\mathbf{y}$ is the realization of a random vector, the estimate of $\boldsymbol{\theta}$, denoted by $\hat{\boldsymbol{\theta}}$, is also a random variable.
The secondary goal of calibration is to estimate the distribution of $\hat{\boldsymbol{\theta}}$ representing the uncertainty of the calibration process.

Example : 2 parameters model



The standard observation model makes the hypothesis that the observation error is Gaussian, and that every observation has an independent error:

$$\varepsilon = \mathcal{N}\left(\mathbf{0}, (\sigma^\star)^2 \mathbf{I}\right)$$

where $\sigma^\star > 0$ is the (unknown) constant observation error variance.

Least squares

The residuals vector is made of the differences between the observations and the predictions:

$$\boldsymbol{r}(\boldsymbol{\theta}) = \mathbf{y} - \mathbf{h}(\boldsymbol{\theta}),$$

for any $\boldsymbol{\theta} \in \mathbb{R}^p$.
When the parameter $\boldsymbol{\theta}$ is the *true* $\boldsymbol{\theta}^\star$, the residual vector is:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{h}\left(\boldsymbol{\theta}^\star\right).$$

The method of least squares minimizes the square of the Euclidian norm of the residuals [3]:

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

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

The least squares method minimizes the cost function c:

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

The unbiased estimator of the observation noise variance is:

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

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 **h** is linear with respect to the parameter $\boldsymbol{\theta}$, then the method reduces to the **linear least squares**.

Let $J \in \mathbb{R}^{n \times p}$ be the Jacobian matrix made of the partial derivatives of **h** with respect to $\boldsymbol{\theta}$:

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

Let $\boldsymbol{\mu} \in \mathbb{R}^p$ 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 $\boldsymbol{\theta}$.

Since $\mathbf{h}$ is linear, it is equal to its Taylor expansion:

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

for any $\boldsymbol{\theta} \in \mathbb{R}^p$.
The corresponding linear least squares problem is:

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

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

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

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 [3].

When the problem is rank-deficient, a way to proceed is to truncate the SVD [4].

This estimator can be proved to be the best linear unbiased estimator, that is, among the unbiased linear estimators, it is the one which minimizes the variance of the estimator [2, 7]. Assuming that the random observations are gaussian:

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

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

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}\left(\boldsymbol{\theta}, \sigma^2 (J^T J)^{-1}\right).$$

Non Linear Least squares

In the general case where the function $\mathbf{h}$ is non linear with respect to the parameter $\boldsymbol{\theta}$, then the resolution involves a non linear least squares optimization algorithm.

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:

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

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:

$$\ell(y_i|\boldsymbol{\theta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - h_i(\boldsymbol{\theta}))^2}{2\sigma^2}\right)$$

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

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

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

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

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

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

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min} \frac{1}{2}\|\mathbf{h}(\boldsymbol{\theta}) - \mathbf{y}\|_2^2$$

and:

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

Bayesian Calibration

The bayesian calibration framework is based on two hypotheses [8, 1] :

- The parameter $\boldsymbol{\theta}$ has a known distribution, called the *prior* distribution, and denoted by $p(\boldsymbol{\theta})$.
- The output observations $\mathbf{y}$ are sampled from a known conditional distribution denoted by $p(\mathbf{y}|\boldsymbol{\theta})$.

For any $\mathbf{y} \in \mathbb{R}^n$ such that $p(\mathbf{y}) > 0$, the Bayes theorem implies that the conditional distribution of $\boldsymbol{\theta}$ given $\mathbf{y}$ is:

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

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

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

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

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

In the Gaussian calibration, the two previous distributions are assumed to be Gaussian.
We make the hypothesis that the parameter $\boldsymbol{\theta}$ follows a Gaussian distribution:

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

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

We make the hypothesis that the output observations follow the conditional gaussian distribution:

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

where $R \in \mathbb{R}^{n \times n}$ 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 $\boldsymbol{\theta},\boldsymbol{\mu}\in\mathbb{R}^p$. Denote by $\|\cdot\|_R$ the Mahalanobis distance associated with the matrix R :

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

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

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

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

MAP estimator

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

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

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 $\boldsymbol{\theta}$ closer to the background $\boldsymbol{\mu}$ (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.
This is similar to Tikhonov regularization, with identity matrix B and zero parameter $\boldsymbol{\mu}$ [5] or to the truncated SVD [4], for some regularization parameter value: the significant difference is that the Bayesian framework considers that the regularization is an hypothesis with known parameter, while regularization methods produce the minimum required regularization parameter values on output.

Linear Gaussian Calibration

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

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

Let $J \in \mathbb{R}^{n \times p}$ be the Jacobian matrix made of the partial derivatives of $\mathbf{h}$ with respect to $\boldsymbol{\theta}$:

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

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).
Let A be the matrix:

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

We denote by K the Kalman matrix:

$$K = AJ^T R^{-1}.$$

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

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

It can be proved that (e.g. [6], p.9):

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

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

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

Non Linear Gaussian Calibration : 3DVAR

The cost function of the gaussian nonlinear calibration problem is :

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

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

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 **h** (no need to restrict to linear code).

We generally do not know the posterior distribution, but we can try to generate a sample from it.
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 **h** is costly.
In this case, using a surrogate model is *mandatory*.

Conclusion

The following table presents a hierarchy of methods depending on the prior and the linearity of the function **h**.

	Linear	**Non linear**
Without *prior*	Linear Least squares	Non Linear Least Squares
Prior: Gaussian	Linear Gaussian	3DVAR
Prior: Non Gaussian	MCMC	MCMC

The previous methods define in-between solutions,

- more complex than the least squares criterion, but more robust with respect to rank deficiency,
- less powerful than the MCMC simulation, but with less computational cost.

Bibliography I


Mark Asch, Marc Bocquet, and Maëlle Nodet.
Data assimilation. Methods, algorithms and applications.
SIAM, 2016.


N. H. Bingham and John M. Fry.
Regression, Linear Models in Statistics.
Springer Undergraduate Mathematics Series. Springer, 2010.


Ake Björck.
Numerical Methods for Least Squares Problems.
Society for Industrial Applied Mathematics, 1996.


P. C. Hansen.
The l-curve and its use in the numerical treatment of inverse problems.
In *in Computational Inverse Problems in Electrocardiology, ed. P. Johnston, Advances in Computational Bioengineering*, pages 119–142. WIT Press, 2000.

Bibliography II


Jennifer L Mueller and Samuli Siltanen.
Linear and nonlinear inverse problems with practical applications.
SIAM, 2012.


C. E. Rasmussen and C. K. I. Williams.
Gaussian Processes for Machine Learning.
The MIT Press, 2006.


Ashish Sen and Muni Srivastava.
Regression analysis.
Springer, 1990.


Albert Tarantola.
Inverse problem theory.
SIAM, 2005.