Optimal design of experiment

Oliwer Sliczniuk

November 16, 2023

Notes about the optimal design of an experiment

1 What information do we have?

- Process model: $\dot{x} = f(t, x, p)$
- Measurement equation: $\dot{y} = g(x)$
- Likelihood function with respect to the dataset
- Parameter estimation
- Sensitivity equations: $\dot{S} = J_x \dot{x} + J_p$

2 Maximum likelihood

Let y^s be the vector of all experimental results (random vector) to be used to estimate p and $y^m(p)$. The vector of the corresponding quantities computed by the model $\dot{x} = f(t, x, p)$. For a parallel model, the parameters of which are to be estimated from measurements of an output vector $\dot{y} = g(x)$. We shall call any optimizer of j, i.e., any $\hat{p}$ that corresponds to an optimal value of the cost function j, an estimate of p in the sense of j.

The vector $\hat{p}_{ml}$ will be a maximum-likelihood estimate if it maximizes the cost function:

$$j_{ml} = \pi_y(y^s|p) \tag{1}$$

If p were fixed, $\pi_y(y^s|p)$ would be the probability density of the random vector y^s generated by a model with parameters p. Here, to the contrary, y^s is fixed and corresponds to the observations. Considered as a function of p, $\pi_y(y^s|p)$ is then called the likelihood of y^s. The maximum-likelihood method looks for the parameter vector p value that gives the observed data the highest likelihood. In practice, it is often easier to look for $\hat{p}_{ml}$ by maximizing the log-likelihood function, yielding the same estimate since the logarithm function is monotonically increasing.

$$j_{ml} = \ln(\pi_y(y^s|p)) \tag{2}$$

Assume that the observed outputs satisfy

$$y(t_i) = y^m(t_i, p^*) + \epsilon_i, \qquad i = 1, ..., n_t \tag{3}$$

where the vector $y^m(t_i.p^*)$ is the output of a deterministic model, p^* is the true value of the parameter vector, and ϵ_i belongs to a sequence of independent random variables with probability density $\pi_\epsilon(\epsilon_i)$. Since the ϵ_i are independent

$$\pi_\epsilon(\epsilon_1, \epsilon_2, ..., \epsilon_{n_t}) = \prod_{i=1}^{n_t} \pi_{\epsilon_i}(\epsilon_i) \tag{4}$$

Consider the output error

$$e^y(t_i, p) = y(t_i) - y^m(t_i, p) \tag{5}$$

For the true value of the parameters, it satisfies $e^y(t_i, p^*) = \epsilon_i$. and, since y^m is deterministic, $\pi_{y_i}(y^s(t_i)|p) = \pi_{\epsilon_i}(e^y(t_i)|p)$. The likelihood of the n_t observations can be written as

$$\pi_y(y^s|p) = \prod_{i=1}^{n_t} \pi_{\epsilon_i}(e^y(t_i, p)) = \prod_{i=1}^{n_t} \pi_{\epsilon_i}(y(t_i) - y^m(t_i, p)) \tag{6}$$

If the noise is assumed to follow the normal distribution:

$$\pi_y(y^s|p) = \prod_{i=1}^{n_t} \frac{1}{\sqrt{2\pi\sigma_{t_i}^2}} \exp\left(-\frac{1}{2}\left(\frac{y(t_i) - y^m(t_i, p)}{\sigma_{t_i}}\right)\right) \tag{7}$$

The associated log-likelihood can be written as

$$\ln(\pi_y(y^s|p)) = (\text{term independent of p}) - \frac{1}{2}\sum_{i=1}^{n_t}\left(\frac{y(t_i) - y^m(t_i, p)}{\sigma_{t_i}^2}\right) \tag{8}$$

Its gradient is thus

$$\frac{\partial}{\partial p}\ln(\pi_y(y^s|p)) = \frac{1}{2}\sum_{i=1}^{n_t}\left(\frac{y(t_i) - y^m(t_i, p)}{\sigma_{t_i}^2}\frac{\partial y^m(t_i, p)}{\partial p}\right) \tag{9}$$

3 Fisher information

The Fisher information is a way of measuring the amount of information that an observable random variable carries about an unknown parameter of a distribution that models the random variable. The Fisher information is related to the second derivative (or the curvature) of the log-likelihood function with respect to the parameter. This

relationship provides a measure of how "sensitive" the likelihood is to changes in the parameter value. The Fisher information matrix can be calculated as follow.

$$F(p) = -\mathop{\mathbb{E}}_{y^s|p}\left[\frac{\partial^2 \ln(\pi_y(y^s|p))}{\partial p \partial p^T}\right] = \mathop{\mathbb{E}}_{y^s|p}\left[\frac{\partial \ln(\pi_y(y^s|p))}{\partial p}\frac{\partial \ln(\pi_y(y^s|p))}{\partial p^T}\right] \tag{10}$$

$$= \mathop{\mathbb{E}}_{y^s|p}\left[\sum_{k=1}^{n_t}\left(\frac{y(t_k) - y^m(t_k,p)}{\sigma_{t_k}^2}\frac{\partial y^m(t_k,p)}{\partial p}\right) \times \sum_{i=1}^{n_t}\left(\frac{y(t_i) - y^m(t_i,p)}{\sigma_{t_i}^2}\frac{\partial y^m(t_i,p)}{\partial p^T}\right)\right] \tag{11}$$

Since

$$\mathop{\mathbb{E}}_{y^s|p}\left[\left(y(t_i) - y^m(t_k,p)\right)\left(y(t_i) - y^m(t_i,p)\right)\right] = \sigma_{t_i}^2 \delta_{ik} \tag{12}$$

one gets

$$F(p) = \sum_{i=1}^{n_t}\left(\frac{1}{\sigma_{t_i}^2}\frac{\partial y^m(t_i,p)}{\partial p}\frac{\partial y^m(t_i,p)}{\partial p^T}\right) \tag{13}$$

If every experiment at time t_i is independent and characterized by its own σ_{t_i}, then the Fisher information matrix can be presented in the more compact way:

$$F(p) = \frac{\partial y^m(t,p)}{\partial p}\mathrm{diag}\left(\frac{1}{\sigma_t^2}\right)\frac{\partial y^m(t,p)}{\partial p^T} \tag{14}$$

4 Cramer-Rao inequality

Let $\hat{p}$ be an (absolutely) unbiased estimator p^*, i.e. such that which amounts to saying that if it were possible to replicate the same experiment and estimate $\hat{p}$ an infinite number of times, the mean of the estimates would coincide with the true value. Let P be the covariance matrix of this estimator. Since $\hat{p}$ is unbiased, P can be written as

$$P = \mathop{\mathbb{E}}_{y^s|p^*}\left[\left(\hat{p}(y^s) - p^*\right)\left(\hat{p}(y^s) - p^*\right)^T\right] \tag{15}$$

which quantifies how the estimates are spread around the true value p^*. One would like the estimates to be as concentrated as possible around this true value, of course. An estimator $\hat{p}_1$ with covariance matrix P_1 is said to be more efficient than an estimator $\hat{p}_2$ with covariance matrix P_2 if $P_1 < P_2$, that is if $P_2 - P_1$ is positive-definite (i.e.if all the eigenvalues of $P_2 - P_1$ arc strictly positive). Since estimators with high efficiency are desirable, a natural request is to make P as small as possible. The Cramer-Rao inequality provides a lower bound to what can be achieved.

Under the hypotheses that:

- the set of all data vectors y^s with $\pi_y(y^s|p) > 0$ does not depend on p

- $\frac{\partial \pi_y(y^s|p)}{\partial p_i}$ $(i = 1, 2, ..., n_p)$ is absolutely integrable
- $\mathbb{E}_{y^s|p}\left[\frac{\partial \ln(\pi_y(y^s|p))}{\partial p}\frac{\partial \ln(\pi_y(y^s|p))}{\partial p^T}\right]$ exists and is invertible

the covariance of any absolutely unbiased estimator satisfies

$$P \geq F^{-1}(p^*) \tag{16}$$

In other words, the precision to which we can estimate p is fundamentally limited by the Fisher information of the likelihood function. Based on the Cramer-Rao inequality, the Fisher information matrix can used to calculate the covariance matrices associated with maximum-likelihood estimates.

5 Optimal experimental design

The optimal design of experiments (DOE) is a statistical concept that refers to the process of planning an experiment, which allow parameters to be estimated without bias and with minimum variance. Optimal design ensures that the experiment can provide the most informative data possible. This often involves balancing the study of main effects and interactions between factors. Moreover, by efficiently planning experiments, optimal design aims to reduce the overall resources required, such as time, materials, and manpower.

The methodology for data to estimate the parameters of a specific model is influenced by a series of qualitative decisions made throughout the experimental and modelling process, such as: a model structure, a location of sensors or an equipment. Once these choices have been made, the experimenter still has some freedom to specify the quantitative experimental conditions (such as temperature, pressure, sampling times, etc.). Experiment design aims to determine experimental conditions adapted to the final purpose of the modeling.

Let's consider that each scalar observation in our study can be expressed as $y(\xi_i)$, where the n_ξ-dimensional vector ξ_i representing the specific experimental conditions (such as the timing of measurements, operating conditions, etc.) under which the i'th observation is gathered. When collecting n_t such observations, the assembly of these ξ_i vectors forms the matrix $\Xi = (\xi_1, \xi_2, ..., \xi_{n_t})$, which combine all the experimental conditions that need optimization. In order to align the design of the experiment with practical realities, it's important to take into account various constraints, such as the total duration of the experiments, the maximum temperature of the inlet stream, and the minimum interval between sampling events. The set of all possible combinations for $\bar{\Xi}$ that adhere to these constraints is denoted as Ξ.

The formulation of a cost function j allows for the framing of optimal experiment design as a problem of constrained optimization. In this context, the optimal experiment, denoted as Ξ^*:

$$\Xi^* = \arg \underset{\Xi \in \bar{\Xi}}{\text{opt}}\ j\left(\Xi\right) \tag{17}$$

The cost function should describe the amount of information from an experiment. For that purpose it can be assumed a function ϕ can be related to the Fisher information obtained an arbitrary operating conditions.

$$j(\Xi) = \phi\left[F(p, \Xi)\right] \tag{18}$$

A general class of optimality criteria for DOE is given by

$$\begin{aligned} \phi_k(F) &= \left[\frac{1}{n_p}\text{trace}\left(QF^{-1}Q^T\right)^k\right]^{1/k} && if\ \det F \neq 0 \\ \phi_k(F) &= \infty && if\ \det F = 0 \end{aligned} \tag{19}$$

where Q is a weighting matrix. The special case $k = 1$ corresponds to the L-optimality cost function,

$$j_L(\Xi) = \text{trace}\left[Q^T Q F^{-1}(p, \Xi)\right] \tag{20}$$

and choosing $Q = \mathbf{I}_{n_p}$ then corresponds to the A-optimality cost function. An A-optimal experiment minimizes the sum of the squares of the lengths or the axes of asymptotic confidence ellipsoids. Choosing Q diagonal, with $[Q]_{ii} = 1/p_i$ corresponds to C-optimality, which is connected with the relative precision of estimates. Taking Q to be a row vector leads to c-optimality. Taking $Q = \mathbf{I}_{n_p}$ and $k = \infty$ corresponds to E-optimality; E-optimal design maximizes the smallest eigenvalues of the Fisher information matrix and thus minimizes the length of the largest axis of the asymptotic confidence ellipsoids. The most widely used optimality criterion has $k = 0$, $Q = \mathbf{I}_{n_p}$, requiring minimization of $\det F^{-1}(p, \Xi)$, or, equivalently, maximization of

$$j_D(\Xi) = \det F(p, \Xi) \tag{21}$$

A D-optimal experiment minimizes the volume of the asymptotic confidence ellipsoids for the parameters. The graphical representation of the different optimality conditions is shown in Figure 1.

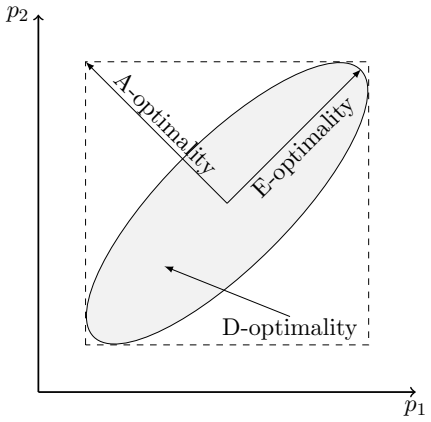

Figure 1: Graphical representation of score functions