Maximizing identifiability

(Unit 3 of 4)

Eric Walter

Eric.Walter@lss.supelec.fr

Laboratoire des Signaux et Systèmes

Parameter Estimation in Physiological Models, Euro Summer School, Lipari, September 2009.

Abstract:

Once a measure of identifiability has been chosen, it becomes possible to maximize it with respect to experimental conditions.

This pertains to *experiment design*.

Various approaches are presented and illustrated by simple examples.

Specific difficulties raised by knowledge-based models are addressed, and approaches for tackling them are proposed.

More on the material presented here can be found in

Eric Walter and Luc Pronzato, *Identification of Parametric Models from Experimental Data*, Springer, London, 1997.

Despite their name, the data may not be a given of the problem!

We assume we have some degrees of freedom to specify, *e.g.*,

- location of sensors and actuators,
- input shape,
- measurement times

Otherwise, experiment design stops here...

Poorly designed experiments may render data useless...

EXAMPLE

We wish to estimate the weights w_1^*, w_2^* and w_3^* of three objects O_1, O_2 and O_3 with four measurements on a spring balance.

Random measurement errors ε assumed i.i.d. $\mathcal{N}(0, \sigma^2)$, plus unknown systematic error w_0^*.

Intuitive approach:

First use the balance with no object to estimate systematic error: $y(0) = w_0^* + \varepsilon(0)$.

Then weigh the three objects successively: $y(i) = w_0^* + w_i^* + \varepsilon(i)$, $i = 1, 2, 3$.

Easy to show that the estimates $\hat{w}_i = y(i) - y(0)$ of the weights w_i^* $(i = 1, 2, 3)$ are unbiased, with $\mathrm{var}(\hat{w}_i) = 2\sigma^2$ and $\mathrm{cov}(\hat{w}_i, \hat{w}_j) = \sigma^2$, $i \neq j$.

Alternative approach:

First use the balance with all objects: $y(0) = w_0^* + w_1^* + w_2^* + w_3^* + \varepsilon(0)$.

Then, as before, weigh the three objects successively: $y(i) = w_0^* + w_i^* + \varepsilon(i)$, $i = 1, 2, 3$.

The estimates of the weights are then

$$\hat{w}_i = \frac{y(0) + y(i) - y(j) - y(k)}{2}, \qquad i = 1, 2, 3,\ i \neq j,\ i \neq k,\ j \neq k.$$

They are unbiased, with $\mathrm{var}(\hat{w}_i) = \sigma^2$ and $\mathrm{cov}(\hat{w}_i, \hat{w}_j) = 0$, $i \neq j$.

Much better for same price of four measurements:

- more accurate estimates (variances reduced by a factor of two),
- independent estimates.

Designing of an optimal experiment requires:

- choosing an optimality criterion,
- specifying the set of all feasible experiments,
- optimizing this criterion over this set.

This is constrained optimization (unlike parameter estimation).

We assume here that the purpose is maximizing some measure of parameter identifiability.

Optimality criterion thus related to how uncertainty in the parameters is characterized.

Let ξ^i describes the experimental conditions of the ith scalar observation.

Simplest case is when ξ^i is a scalar, *e.g.*, measurement time.

When n_t such observations are taken, the concatenation of their ξ^i's yields the vector

$$\Xi = (\xi^{1T}, \xi^{2T}, \ldots, \xi^{n_t T})^T,$$

which characterizes all experimental conditions to be optimized.

A number of constraints on Ξ must be taken into account, *e.g.* on the duration of the experiments, the energy or amplitude of the inputs, the minimum time between samples, the total number of samples, working hours…

The resulting set of feasible values for Ξ is $[\![\Xi]\!]$.

Optimal experiment Ξ^* then defined as

$$\Xi^* = \arg\operatorname{opt}_{\Xi \in \mathbb{\Xi}} j(\Xi),$$

were $j(\cdot)$ is some appropriate cost function.

Most often Ξ^* lies on the boundary of $\mathbb{\Xi}$, so a proper definition of $\mathbb{\Xi}$ is essential.

Cost evaluation must be simple enough to allow easy optimization

$\Rightarrow$ uncertainty in $\mathbf{p}$ characterized via FIM $\mathbf{F}(\mathbf{p}, \Xi)$.

Next section describes commonly used criteria for optimal experiment design.

Criteria

Under technicalities, maximum-likelihood estimator is *asymptotically* $\mathcal{N}(\mathbf{p}^*, \mathbf{F}^{-1}(\mathbf{p}^*, \Xi))$.

Optimizing some scalar function of the FIM $\mathbf{F}(\mathbf{p}, \Xi)$] then amounts to optimizing a scalar measure of the asymptotic covariance of $\mathbf{p}$.

Recall, however, that characterizing parameter uncertainty in a non-LP model by inverse of FIM involves daring approximations, to which few alternatives exist.

For the time being, we neglect the obvious and crucial problem of dependency in $\mathbf{p}$.

Most commonly used criterion is D-optimality:

$$\Xi_D = \arg\min_{\Xi \epsilon [\Xi]} \det \mathbf{F}^{-1}(\mathbf{p}, \Xi)$$

$$= \arg\max_{\Xi \epsilon [\Xi]} \det \mathbf{F}(\mathbf{p}, \Xi)$$

$$= \arg\max_{\Xi \epsilon [\Xi]} \ln \det \mathbf{F}(\mathbf{p}, \Xi)$$

D-optimal experiment thus minimizes the volume of asymptotic confidence ellipsoids for **p**.

Advantage

A D-optimal experiment is invariant under any nonsingular reparametrization that does not depend on the experiment.

Thus, the optimal experiment does not depend on the units in which the parameters are expressed.

This seems natural to ask for, and is one of the reasons for the popularity of D-optimality.

Drawback

A D-optimal experiment may correspond to a very elongated confidence ellipsoid, with large confidence intervals for the parameters.

This is why other criteria may be preferred, *e.g.*,

A-optimality:

$$\Xi_A = \arg\min_{\Xi \epsilon \Xi} \text{trace } \mathbf{F}^{-1}(\mathbf{p}, \Xi),$$

which minimizes sum of squares of lengths of axes of asymptotic confidence ellipsoids.

E-optimality:

$$\Xi_E = \arg\max_{\Xi \epsilon \Xi} \text{smallest eigenvalue of } \mathbf{F}(\mathbf{p}, \Xi),$$

which minimizes length of largest axis of asymptotic confidence ellipsoids.

In both cases, invariance under reparametrization is lost.

In what follows, we concentrate on D-optimality and friends.

D_s-optimality

Sometimes, only some of the parameters are of interest.

Partition $\mathbf{p}$ into $(\mathbf{p}_1^T, \quad \mathrm{p}_2^T)^T$, with

- $\mathbf{p}_1$ the parameters of interest (dim $\mathbf{p}_1 = s$)
- $\mathbf{p}_2$ the others (called nuisance parameters).

Partition $\mathbf{F}(\mathbf{p}, \Xi)$ accordingly into

$$\mathbf{F}(\mathbf{p}, \Xi) = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix} .$$

$\mathbf{F}^{-1}(\mathbf{p}, \Xi)$ then given by

$$\begin{bmatrix} (\mathbf{F}_{11} - \mathbf{F}_{12}\mathbf{F}_{22}^{-1}\mathbf{F}_{21})^{-1} & -(\mathbf{F}_{11} - \mathbf{F}_{12}\mathbf{F}_{22}^{-1}\mathbf{F}_{21})^{-1}\mathbf{F}_{12}\mathbf{F}_{22}^{-1} \\ -\mathbf{F}_{22}^{-1}\mathbf{F}_{21}(\mathbf{F}_{11} - \mathbf{F}_{12}\mathbf{F}_{22}^{-1}\mathbf{F}_{21})^{-1} & \mathbf{F}_{22}^{-1}+\mathbf{F}_{22}^{-1}\mathbf{F}_{21}(\mathbf{F}_{11} - \mathbf{F}_{12}\mathbf{F}_{22}^{-1}\mathbf{F}_{21})^{-1}\mathbf{F}_{12}\mathbf{F}_{22}^{-1} \end{bmatrix},$$

where $\mathbf{F}_{22}$ is assumed not to be singular. Since only $\mathbf{p}_1$ is of interest, we wish to minimize a scalar function of $(\mathbf{F}_{11} - \mathbf{F}_{12}\mathbf{F}_{22}^{-1}\,\mathbf{F}_{21})^{-1}$.

D_s-optimal design then maximizes

$$j_{\mathrm{Ds}}(\Xi) = \det\,(\mathbf{F}_{11} - \mathbf{F}_{12}\mathbf{F}_{22}^{-1}\mathbf{F}_{21}).$$

EXAMPLE OF D-OPTIMAL DESIGN: NICE CASE

Consider again the FIR model

$$y_{\mathrm{m}}(k, \mathbf{p}) = p_1 u(k-1) + p_2 u(k-2),$$

and assume

$$y(k) = y_{\mathrm{m}}(k, \mathbf{p}^*) + n(k), k = 1, \dots, 10,$$

with n i.i.d. $\mathcal{N}(0, \sigma^2)$.

We have shown in Unit 2 that

$$\mathbf{F}(\mathbf{p}, \boldsymbol{\Xi}) = \frac{1}{\sigma^2} \begin{bmatrix} \sum_{i=0}^{9} u^2(i) & \sum_{i=0}^{9} u(i)u(i-1) \\ \sum_{i=0}^{9} u(i)u(i-1) & \sum_{i=0}^{9} u^2(i-1) \end{bmatrix}.$$

A D-optimal sequence of inputs is to be computed, under the constraint

$$-u_{\max} \leq u(k) \leq u_{\max}, \quad k = 1, \ldots, 10$$

We must maximize det **F**(**p**, Ξ), or equivalently

$$\sum_{i=0}^{9} u^2(i) \cdot \sum_{i=0}^{9} u^2(i-1) - \left[\sum_{i=0}^{9} u(i)u(i-1)\right]^2.$$

An intuitive solution is $u(i) = \pm u_{\max}$,

with $++\ --\ ++\ --\ ++$

so as to ensure $\sum_{i=0}^{9} u(i)u(i-1) = 0.$

Another solution is PRBS, which can be used for higher-order FIRs. Neither depends on **p**.

What do we gain?

Take $u_{\max} = 1$ and compare unit step input and D-optimal input.

With unit step input, covariance of estimation error is

$$\mathbf{P} = \sigma^2 \left(\mathbf{R}^{\mathrm{T}}\mathbf{R}\right)^{-1} = \frac{\sigma^2}{9}\begin{bmatrix} 9 & -9 \\ -9 & 10 \end{bmatrix},$$

whereas with D-optimal input, it becomes

$$\mathbf{P} = \sigma^2 \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{10} \end{bmatrix}.$$

Errors are no longer correlated and standard deviations are divided by $\approx \sqrt{10}$.

EXAMPLE OF D-OPTIMAL DESIGN: NOT SO NICE CASE

Consider again

$$y(t) = p_1^* \exp(-p_2^* t) + \varepsilon(t),$$

with ε i.i.d. $\mathcal{N}(0, \sigma^2)$.

Assume two observations $y(t_1)$ and $y(t_2)$ are to be made.

The problem is to choose measurement times t_1 and t_2, with $t_2 > t_1 \geq 0$, so as to estimate $\mathbf{p} = (p_1, p_2)^T$ as accurately as possible.

Here,

$$\boldsymbol{\Xi} = (t_1, t_2)^T.$$

We have shown in Unit 2 that

$$\mathbf{F}(\mathbf{p}, \Xi) = \frac{1}{\sigma^2} \sum_{k=1}^{2} \begin{bmatrix} \exp(-2p_2 t_k) & -p_1 t_k \exp(-2p_2 t_k) \\ -p_1 t_k \exp(-2p_2 t_k) & p_1^2 t_k^2 \exp(-2p_2 t_k) \end{bmatrix}.$$

$$\Rightarrow \det \mathbf{F}(\mathbf{p}, \Xi) = \frac{1}{\sigma^4} p_1^2 (t_2 - t_1)^2 \exp[-2p_2(t_1 + t_2)],$$

and

$$\Xi_D = \begin{pmatrix} 0 & \frac{1}{p_2} \end{pmatrix}^T.$$

[Without the constraints, “solution” would involve negative measurement times...]

Optimal experiment depends on the values of the parameters to be estimated!

This problem is common to non-LP structures, and thus to most knowledge-based models ☹.

The most classical (and simplest) approach then takes $\mathbf{p} = \mathbf{p}^0$, where $\mathbf{p}^0$ is some (hopefully) reasonable numerical value for $\mathbf{p}$. This is called

Local design

Optimization is most often iterative.

The type of algorithm to be used then depends on dim Ξ.

When dim Ξ is not too large, classical nonlinear programming methods may be used.

The cost function $j_D(\Xi) = \det \mathbf{F}(\mathbf{p}, \Xi)$ generally has several local optimizers, so a global optimization method is recommended.

When the dimension of Ξ is large, it is preferable to use dedicated algorithms. The principles of some of them are now very briefly presented.

Exact design

Optimization is with respect to the variables defining the experiment to be performed.

Exchange algorithms

Let Ξ^k be the estimate of Ξ at iteration k. Assume Ξ^k is not degenerate (*i.e.*, det $\mathbf{F}(\mathbf{p}, \Xi^k) \neq 0$).

At each iteration, one of the support points ξ^i of Ξ^k is replaced in Ξ^{k+1} by ξ^*, with ξ^i and ξ^* chosen so that

$$\det \mathbf{F}(\mathbf{p}, \Xi^{k+1}) > \det \mathbf{F}(\mathbf{p}, \Xi^k).$$

Exchange algorithms differ in the selection of ξ^i and in the construction of ξ^*.

DETMAX algorithm

Assume Ξ^k is not degenerate. If an additional observation were allowed, characterized by ξ^{n_t+1}, one would choose $\xi^{n_t+1} = \xi^*$, with ξ^* such that det $\mathbf{F}(\mathbf{p}, \Xi^{k+})$ be maximized, where

$$\Xi^{k+} = \begin{bmatrix} \Xi^k \\ \xi^{n_t+1} \end{bmatrix} .$$

If ξ^{n_t+1} were to replace one of the support points of Ξ^k, one should obtain Ξ^{k+1} from Ξ^{k+} by removing ξ^{i*} such that det $\mathbf{F}(\mathbf{p}, \Xi^{k+1})$ remains as large as possible.

This augmentation of the number of support points, followed by the removal of some of them so as to keep the number of observations equal to n_t, is called an *excursion*.

The introduction of several support points before the removal of as many of them may be considered (excursion of length ≥ 2).

Note: algorithms for exact design can only increase the value of $j_D(\Xi)$, but convergence to optimal experiment is not guaranteed.

In contrast, the methods presented now yield a global optimum, sometimes at the price of approximation to make the designed experiment implementable.

Distribution of experimental effort

When experiments are repeated, the number n_e of distinct ξi's is less than the total number of observations n_t. FIM can then be written as

$$\mathbf{F}(\mathbf{p}, \Xi) = \sum_{i=1}^{n_e} r_i \frac{1}{w(\xi i)} \mathbf{s}_y(\xi i, \mathbf{p})\mathbf{s}_y^T(\xi i, \mathbf{p}) ,$$

with r_i the number of repetitions of measurements under conditions ξi, so $\sum_{i=1}^{n_e} r_i = n_t$.

The FIM per sample is then

$$\mathbf{F}_{\mathrm{ps}}(\mathbf{p}, \Xi) = \sum_{i=1}^{n_{\mathrm{e}}} \frac{r_i}{n_{\mathrm{t}}} \frac{1}{w(\xi^i)} \mathbf{s}_{\mathrm{y}}(\xi^i, \mathbf{p}) \mathbf{s}_{\mathrm{y}}^{\mathrm{T}}(\xi^i, \mathbf{p}) ,$$

with

$$\mathbf{s}_{\mathrm{y}}(\xi^i, \mathbf{p}) = \frac{\partial y_{\mathrm{m}}}{\partial \mathbf{p}}(\xi^i, \mathbf{p}).$$

The proportion

$$\mu_i = \frac{r_i}{n_t}$$

of observations at ξ^i can be considered as a *percentage of experimental effort*.

Experiment Ξ thus represented as a discrete distribution m_D,

$$m_D = \begin{Bmatrix} \xi_1 & \dots & \xi_{n_e} \\ \mu_1 & \dots & \mu_{n_e} \end{Bmatrix},$$

which is normalized, since

$$\sum_{i=1}^{n_e} \mu_i = 1.$$

The μ_i's are rational numbers ($\mu_i = r_i/n_t$, with n_t fixed).

Kiefer and Wolfowitz showed that removing this constraint and extending design to any normalized measure on ξ drastically simplifies design.

This is the basis for the algorithms to be presented now.

Consider thus a normalized measure m on ξ, satisfying $\int_{\xi} m(\mathrm{d}\boldsymbol{\xi}) = 1$.

FIM per sample then takes the form

$$\mathbf{F}_{\mathrm{ps}}(\mathbf{p}, m) = \int_{\xi} \frac{1}{w(\boldsymbol{\xi})} \mathbf{s}_{\mathrm{y}}(\boldsymbol{\xi}, \mathbf{p}) \mathbf{s}_{\mathrm{y}}^{\mathrm{T}}(\boldsymbol{\xi}, \mathbf{p}) \; m(\mathrm{d}\boldsymbol{\xi}).$$

Any matrix of this form can be obtained with a discrete measure. We can thus restrict attention to *discrete* experiments, associated with normalized discrete measures on ξ.

A discrete experiment usually cannot be implemented exactly for a given number n_{t} of observations, since the weights must satisfy $\mu_i = r_i/n_{\mathrm{t}}$.

In most cases, the μ_i's must thus be approximated by rational numbers r_i/n_t.

This is why the approach is called *approximate* design.

Properties of optimal experiments

Kiefer-Wolfowitz equivalence theorem.

For D-optimality, design amounts to the minimization of the convex function $-\ln\det\mathbf{F}$ over the convex set of symmetric non-negative definite matrices $\mathbf{F}_{ps}(\mathbf{p}, m)$.

The matrix $\mathbf{F}_{ps}(\mathbf{p}, m_D)$ associated with a D-optimal design measure m_D is thus unique.

This does not imply uniqueness of m_D.

The optimum can be characterized by first-order stationarity conditions.

Design measure m_1 is D-optimal if and only if det $\mathbf{F}_{ps}(\mathbf{p}, m_1)$ is a maximum, or, equivalently, if and only if for any measure m_2,

$$\frac{\partial \ln \det \mathbf{F}_{ps}[\mathbf{p}, (1-\alpha)m_1 + \alpha m_2]}{\partial \alpha}\Big|_{\alpha=0} \leq 0.$$

Since

$$\frac{\partial \ln \det \mathbf{F}}{\partial \alpha} = \text{trace}\left(\mathbf{F}^{-1}\frac{\partial \mathbf{F}}{\partial \alpha}\right),$$

a necessary and sufficient condition for D-optimality of m_1 is

$$\text{trace}\,\{\mathbf{F}_{ps}^{-1}(\mathbf{p}, m_1)\,[\mathbf{F}_{ps}(\mathbf{p}, m_2) - \mathbf{F}_{ps}(\mathbf{p}, m_1)]\} \leq 0, \text{ for any } m_2,$$

or equivalently

$$\text{trace}\,[\mathbf{F}_{\text{ps}}^{-1}(\mathbf{p},\, m_1)\,\mathbf{F}_{\text{ps}}(\mathbf{p},\, m_2)] \leq \dim\ \mathbf{p}, \text{ for any } m_2.$$

In particular, this must be true when m_2 is the discrete measure with unit weight at a single support point $\boldsymbol{\xi} \in \Xi$.

A necessary condition for D-optimality of m_1 is thus

$$d(\boldsymbol{\xi},\, m_1) \leq \dim \mathbf{p}, \text{ for any } \boldsymbol{\xi} \in \Xi,$$

with

$$d(\boldsymbol{\xi}, m) = \frac{1}{w(\boldsymbol{\xi})}\,\mathbf{s}_{\text{y}}^{\text{T}}(\boldsymbol{\xi}, \mathbf{p})\mathbf{F}_{\text{ps}}^{-1}(\mathbf{p}, m)\mathbf{s}_{\text{y}}(\boldsymbol{\xi}, \mathbf{p}).$$

One can easily check that this necessary condition is also sufficient.

Note that $w(\boldsymbol{\xi})$ need not be known to compute $d(\boldsymbol{\xi}, m)$.

Now, using the fact that

$$\dim \mathbf{p} = \text{trace}\,[\mathbf{F}_{\text{ps}}^{-1}(\mathbf{p},\, m_1)\, \mathbf{F}_{\text{ps}}(\mathbf{p},\, m_1)] = \sum_{i=1}^{n_e} \mu_i d(\xi^i,\, m_1)\,,$$

where the μ_i's and ξ^i's are respectively the weights and support points of the D-optimal design measure m_1, the following theorem is obtained.

EQUIVALENCE THEOREM (Kiefer and Wolfowitz)

The following properties are equivalent:

- the design measure m_D is D-optimal,
- the maximum of $d(\xi,\, m_D)$ for $\xi \in \Xi$ is equal to dim $\mathbf{p}$,
- m_D is the measure m that minimizes the maximum of $d(\xi,\, m)$ for $\xi \in \Xi$.

EXAMPLE (continued)

Consider again

$$y(t) = p_1^* \exp(-p_2^* t) + \varepsilon(t),$$

with ε i.i.d. $\mathcal{N}(0, \sigma^2)$. The number of observations to be made is no longer specified, and a D-optimal design measure is sought, with an admissible domain $[0, \infty[$ for the measurement times. It can readily be checked that the design measure

$$m_D = \begin{Bmatrix} 0 & 1/p_2 \\ 1/2 & 1/2 \end{Bmatrix}$$

is D-optimal. Indeed,

$$d(t, m_D) = 2 \exp(-2p_2 t)\,(1 - 2p_2 t + p_2^2\, t^2 e^2 + p_2^2\, t^2)\ .$$

Since $d(0, m_D) = d(1/p_2, m_D) = 2 = \dim \mathbf{p}$, and $d(t, m_D) \leq 2$ for all $t \geq 0$, m_D is D-optimal.

Algorithms based on equivalence theorem

Consider a non-degenerate design measure m^k. From equivalence theorem, m^k is D-optimal if and only if $d(\xi, m^k) \leq \dim \mathbf{p}$ for any $\xi \in \Xi$.

Structure of algorithms then as follows.

Step 1: Choose a discrete non-degenerate initial design measure m^1 (a normalized discrete distribution with at least n_p support points, such that $\det \mathbf{F}_{ps}(\mathbf{p}, m^1) \neq 0$). Choose some positive tolerance $\delta << 1$. Set $k = 1$.

Step 2: Find $\xi^* = \arg\max_{\xi \in \Xi} d(\xi, m^k)$. If $d(\xi^*, m^k) < \dim \mathbf{p} + \delta$, stop.

Step 3: Set $m^{k+1} = (1 - \alpha_k) m^k + \alpha_k m_{\xi^*}$, with m_{ξ^*} a measure with unit mass at ξ^*, increment k by one and go to Step 2.

Step size α_k remains to be chosen.

Fedorov uses the optimal value

$$\alpha_k^* = \frac{d(\xi^*, m^k) - \dim \mathbf{p}}{\dim \mathbf{p}\,(d(\xi^*, m^k) - 1)},$$

while Wynn uses a predefined sequence $\{\alpha_k\}$ that satisfies

$$\alpha_k > 0, \quad \lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty,$$

for instance

$$\alpha_k = \frac{1}{k+1}.$$

Provided a global maximizer ξ^* is computed at Step 2, both algorithms converge to a D-optimal measure, a considerable improvement over the algorithms for exact design.

No support point is ever removed from the design measure. Easy to modify the algorithms to avoid this drawback.

Applications

Optimal measurement times

When the number of observations is small, measurement times can be individually optimized.

When the number of observations is large, optimization of measurement times will often be restricted to determination of optimal sampling frequencies.

Optimal inputs

For dynamical systems, the result of each observation depends on previous inputs, so the experimental conditions ξ^i for observation i depend on those for others. The search will be for the optimal input from a predefined class of admissible inputs, which may be parametric (*e.g.*, weighted superposition of inputs with simple shapes) or not.

Parametric inputs

In biology, the admissible inputs are often the superposition of basic signals with given shapes. For instance, one may consider the superposition of

- rectangular pulses $a_i[H(t - t_i) - H(t - T_i - t_i)]$, where H is Heaviside's step function,
- impulses $b_i\delta(t - \tau_i)$, where δ is Dirac's distribution,
- polynomials $\sum_{k=0}^{n_k} d_k\, t^k$.

The parameters Ξ that characterize an input u are then respectively

- the starting time t_i, amplitude a_i and duration T_i of each rectangular pulse,
- the time τ_i and area b_i of each impulse (b_i may correspond to a dose administered),
- the coefficients d_k of the polynomials.

The design problem then amounts to a nonlinear optimization problem with respect to Ξ.

EXAMPLE

Consider the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\, y_{\mathrm{m}}(t, p^*) = -p^* y_{\mathrm{m}}(t, p^*) + u(t), \quad y_{\mathrm{m}}(0_-, p^*) = 0, \qquad y(t) = y_{\mathrm{m}}(t, p^*) + \varepsilon(t),$$

where p^* is the (scalar) parameter to be estimated, with ε i.i.d. $\mathcal{N}(0, \sigma^2)$.

We wish to determine a D-optimal input of unit area (*i.e.*, a unit dose), in the family

$$u(t) = (1 - \alpha)\delta(t) + \frac{\alpha}{T}\,[H(t) - H(t - T)], \;\; T > 0,\; 0 \le \alpha \le 1.$$

Only one observation is to be made. We also wish to find the optimal measurement time t, with the constraint $0 \le t \le T$. The experiment to be performed is thus characterized by

$$\boldsymbol{\Xi} = (\alpha, T, t)^{\mathrm{T}}.$$

The response of the model for $0 \le t \le T$ is

$$y_{\mathrm{m}}(t, p) = \frac{\alpha}{pT} + \exp(-pt)\left[1 - \alpha\left(1 + \frac{1}{pT}\right)\right].$$

The FIM, here a scalar, can be written as

$$\mathbf{F}(p, \Xi) = \frac{1}{\sigma^2}\left\{\exp(-pt)\left[\frac{\alpha}{p^2T} - t + \alpha t\left(1 + \frac{1}{pT}\right)\right] - \frac{\alpha}{p^2T}\right\}^2$$

and the D-optimal experiment is

$$\Xi_{\mathrm{D}} = \left(0, T, \frac{1}{p}\right)^{\mathrm{T}},$$

with T arbitrary. The D-optimal input in the family considered is thus the unit impulse.

Nonparametric inputs

The restriction to parametric inputs must sometimes be relaxed. Determining an optimal parametric input may serve to initialize input design in a larger class. Some constraints on the shapes of admissible inputs, easily be taken into account for parametric inputs, may raise difficulties for nonparametric inputs.

Simultaneous choice of inputs and sampling times

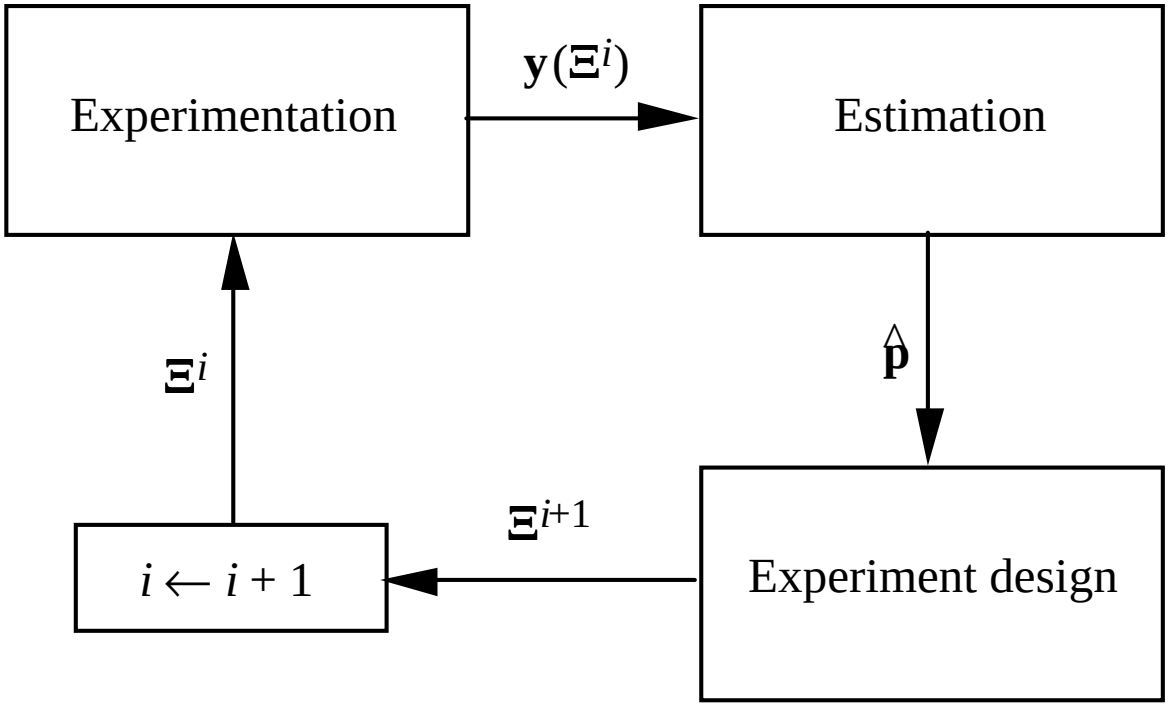
When the input is parametric, the components of the vector Ξ that characterizes the experiment may consist of the n_t sampling times and the parameters defining the input signal.

When the input signal is nonparametric, the construction of an optimal sampling schedule becomes difficult.

Robust design

Can we avoid assuming that we know the value of the parameters **p** that we wish to estimate?

Sequential design



What one should not do:

Estimate $\hat{\mathbf{p}}(i)$ from observations $\mathbf{y}(\Xi^i)$, for $i = 1, \ldots, k$,
Estimate $\hat{\mathbf{p}}^k$ as the average of the $\hat{\mathbf{p}}(i)$'s:

$$\hat{\mathbf{p}}^k = \frac{1}{k} \sum_{i=1}^{k} \hat{\mathbf{p}}(i) \, .$$

Usually, this does not converge to $\mathbf{p}^*$ as k tends to infinity.

To guarantee convergence of $\hat{\mathbf{p}}^k$ to $\mathbf{p}^*$, the estimation of $\hat{\mathbf{p}}^k$ should make use of all previous observations (*i.e.*, $\mathbf{y}(\Xi^1), \ldots, \mathbf{y}(\Xi^k)$).

The designed experiment will then tend to the optimal experiment for $\mathbf{p}^*$.

Repetition of experiments often impossible and a one-shot experiment must be designed.

Even if design is sequential, each design step should make use of all information available.

Non-sequential design approaches that determine a single experiment taking all prior uncertainty in **p** into account are thus useful.

Two types of robust-design procedure will be considered.

They differ in

- how prior information is characterized,
- the importance attached to the risk of designing an experiment badly suited to some rare parameter values.

Average optimality

Relies on a probabilistic description of the prior uncertainty in **p**.

Prior distribution $\pi_{\mathbf{p}}(\mathbf{p})$ may have been inferred from previous observations collected on similar processes or individuals in a population.

Criteria

We only consider cost functions related to D-optimality, but others could be treated similarly.

Using the prior distribution $\pi_{\mathbf{p}}(\mathbf{p})$ makes it possible to remove the dependence on **p** by considering the expectation of the original cost function.

Whereas cost functions $-\det \mathbf{F}(\mathbf{p}, \boldsymbol{\Xi})$, $-\ln \det \mathbf{F}(\mathbf{p}, \boldsymbol{\Xi})$ and $1/\det \mathbf{F}(\mathbf{p}, \boldsymbol{\Xi})$ lead to identical D-optimal designs, the introduction of expectations makes these approaches different.

EID-optimal design minimizes

$$j_{\mathrm{EID}}(\Xi) = \underset{\mathbf{p}}{\mathrm{E}} \ \{1/\det \mathbf{F}(\mathbf{p}, \Xi)\}.$$

ELD-optimal design maximizes

$$j_{\mathrm{ELD}}(\Xi) = \underset{\mathbf{p}}{\mathrm{E}} \ \{\ln \det \mathbf{F}(\mathbf{p}, \Xi)\}.$$

EID-optimal experiment depends on parametrization, contrary to ELD-optimal experiment.

EXAMPLE (continued)

Consider again the system:

$$y(t) = p_1^* \ \exp(-p_2^* \, t) + \varepsilon(t),$$

with ε i.i.d. $\mathcal{N}(0, \sigma^2)$. We look for two optimal sampling times, to estimate p_1^* and p_2^* .

For a nominal value p_2^0

$$\Xi_D = (0,\ 1/p_2^0\ )^T.$$

If $\pi_p(p_2)$ uniform over [1, 10],

$$\Xi_{EID} = (0,\quad 0.139)^T.$$

If $\pi_p(p_2)$ $\mathcal{N}(5.5,\ 1.5^2)$,

$$\Xi_{EID} = (0,\quad 0.161)^T.$$

For the two densities above,

$$\Xi_{ELD} = (0,\quad 0.182)^T.$$

Except in very simple situations, an optimal experiment cannot be found analytically, and numerical procedures are required.

Algorithms

Exact design.

Stochastic approximation allows a cost function like

$$j_{\mathrm{E}}(\Xi) = \underset{\mathbf{p}}{\mathrm{E}} \; \{j(\mathbf{p}, \Xi)\}$$

to be optimized without having to evaluate expectations.

Simplest version is the stochastic gradient algorithm

$$\Xi^{k+1} = \Xi^{k} - \lambda_k \frac{\partial j(\mathbf{p}^k, \Xi)}{\partial \Xi} \Big| \; \Xi = \Xi^{k}.$$

At each iteration k, a value $\mathbf{p}^k$ is randomly generated according to $\pi_{\mathrm{p}}(\mathbf{p})$.

The sequence of scalar steps λ_k must satisfy

$$\lambda_k > 0, \quad \sum_{k=1}^{\infty} \lambda_k = \infty, \quad \sum_{k=1}^{\infty} \lambda^2{}_k < \infty,$$

and the most popular choice is the harmonic sequence

$$\lambda_k = \frac{\alpha}{k+1}, \quad \alpha > 0.$$

Speed of convergence is very sensitive to the choice of the scalar α in the sequence λ_k.

Componentwise normalization of the gradient allows easier choice of a suitable value for α.

The algorithm thus modified becomes

$$\Xi^{k+1} = \Xi^{k} - \lambda_k \, \Lambda_k \, \frac{\partial j(\mathbf{p}^k, \Xi)}{\partial \Xi} \Big| \, \Xi = \Xi^k,$$

where Λ_k is a diagonal matrix, the ith diagonal entry of which is

$$\Lambda_{k_{ii}} = \frac{\Xi_{i_{\max}} - \Xi_{i_{\min}}}{\left[\frac{1}{k} \sum_{n=1}^{k} \left(\frac{\partial j(\mathbf{p}^n, \Xi)}{\partial \Xi_i} \Big| \Xi = \Xi^n\right)^2\right]^{1/2}},$$

where $\Xi_{i_{\max}}$ and $\Xi_{i_{\min}}$ are upper and lower bounds on the possible values of Ξ_i.

If $\lambda_k = \alpha/(k + 1)$, this implies

$$\Xi_i^1 - \Xi_i^0 = \pm\, \alpha(\Xi_{i_{\max}} - \Xi_{i_{\min}}).$$

The scalar α is then the relative length of the first step. A typical choice is $\alpha = 0.1$.

Note that convergence to a global optimum is not guaranteed.

Approximate design

The equivalence theorem can extended to average-optimal design.

Global convergence can be guaranteed, but calculations will be extremely heavy unless the prior distribution for **p** is discrete.

Maximin optimality

Sometimes, the best experiment in the worst circumstances should be preferred to the best one on average.

This depends on the importance attached to (unlikely) parameter values whose estimation with an average-optimal experiment might be very inaccurate.

Maximin optimal design requires the definition of a set $\mathbb{P}$ of prior admissible values for $\mathbf{p}$.

Using $\mathbb{P}$ makes it possible to remove the dependence on $\mathbf{p}$ by considering the worst possible value of the original cost function.

We consider a criterion based on D-optimality, but this is applicable to others.

MMD-optimal design maximizes

$$j_{\mathrm{MMD}}(\Xi) = \min_{\mathbf{p}\in\mathbb{P}} \det \mathbf{F}(\mathbf{p}, \Xi).$$

EXAMPLE (continued)

Consider again

$$y(t) = p_1^* \exp(-p_2^* t) + \varepsilon(t),$$

where ε is i.i.d. $\mathcal{N}(0, \sigma^2)$. We search for two sampling times to estimate the parameters p_1^* and p_2^* , when the prior admissible values for p_2 belong to

$$\mathbb{P}_2 = [p_{2\min}, p_{2\max}].$$

The MMD-optimal experiment then coincides with the D-optimal experiment for $p_2 = p_{2\max}$:

$$\Xi_{\mathrm{MMD}} = \left(0, \ \frac{1}{p_{2\max}}\right)^{\mathrm{T}}.$$

The maximin-optimal experiment sometimes coincides with a D-optimal experiment for a particular value of the model parameters that can be determined *a priori*. This greatly facilitates computation.

When the maximin-optimal experiment cannot be found analytically and does not coincide with a particular D-optimal experiment, specific optimization algorithms must be used.

Algorithms

The goal is to find $\boldsymbol{\Xi}_{\mathrm{MMD}}$ (in $\mathbb{\Xi}$) that maximizes

$$j_{\mathrm{MMD}}(\boldsymbol{\Xi}) = \min_{\mathbf{p}\in\mathbb{P}} \det \mathbf{F}(\mathbf{p}, \boldsymbol{\Xi}).$$

What one should not do: alternate maximization with respect to $\boldsymbol{\Xi}$ and minimization with respect to $\mathbf{p}$ of $\det \mathbf{F}(\mathbf{p}, \boldsymbol{\Xi})$. This may cycle forever.

The most straightforward approach uses brute force and maximizes j_{MMD} by a general-purpose nonlinear-programming algorithm, each evaluation of $j_{\mathrm{MMD}}(\boldsymbol{\Xi})$ being the result of a minimization with respect to $\mathbf{p}$ using a second nonlinear programming algorithm.

May require a huge amount of computation.

Relaxation algorithm (for exact design).

Problem redefined as finding $\boldsymbol{\Xi}_{\mathrm{MMD}}$ in $\mathbb{\Xi}$ that maximizes the scalar α under the constraints

$$j_{\mathrm{MMD}}(\boldsymbol{\Xi}) \geq \alpha, \forall\, \mathbf{p} \in \mathbb{P}.$$

This has an infinite number of constraints. Relaxation introduces a finite number of them:

Step 1: Choose initial value $\mathbf{p}^1$ in $\mathbb{P}$, and define a first set of representative values $\mathbb{S}^1 = \{\mathbf{p}^1\}$. Set $k = 1$.

Step 2: Find $\boldsymbol{\Xi}^k = \arg\max_{\boldsymbol{\Xi}\epsilon\mathbb{\Xi}} \min_{\mathbf{p}\epsilon\mathbb{S}^k} \det \mathbf{F}(\mathbf{p}, \Xi)$

Step 3: Find $\mathbf{p}^{k+1} = \arg \min_{\mathbf{p}\epsilon\mathbb{P}} \det \mathbf{F}(\mathbf{p}, \Xi^k)$.

Step 4: If $j(\mathbf{p}^{k+1}, \boldsymbol{\Xi}^k) \geq \min_{\mathbf{p}\epsilon\mathbb{S}^k} \det \mathbf{F}(\mathbf{p}, \Xi^k) - \delta$, where δ is some positive tolerance, accept $\boldsymbol{\Xi}^k$ as an approximate solution. Else include $\mathbf{p}^{k+1}$ in $\mathbb{S}^k$, increment k by one and go to Step 2.

Stops after a finite number of steps. A global optimization algorithm is recommended.

Conclusions

Experiments are always constrained, and the constraints must be taken into account.

Evaluation of the FIM is simple enough to allow constrained optimization.

Knowledge-based models are most often not LP, so their FIM depends on the value of the parameters to be estimated, which is a very serious difficulty.

Whereas local design assumes a known prior nominal value for **p**, the methods presented allow uncertainty in this nominal value to be taken into account.

The dependence of the optimal experiment on any other quantity (*e.g.*, a nuisance parameter) with an unreliable prior value could be treated in the same way.

The first three units have assumed that $M(\cdot)$ could be made identifiable (at least locally).

Unit 4 will show how the concept of identifiability can be bypassed.