Design of experiments in nonlinear models

Luc Pronzato

Université Côte d'Azur, CNRS, I3S, France

Outline I

1. DoE: objectives & examples
2. DoE based on asymptotic normality
3. Construction of (locally) optimal designs
4. Problems with nonlinear models
5. Small-sample properties
6. Nonlocal optimum design

1 DoE: objectives & examples

A/ Parameter estimation

<u>Ex1</u>: Weighing with a two-pan balance

☞ Determine the weights of 8 objets, with mass m_i, $i = 1, \ldots, 8$

i.i.d. errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

1 DoE: objectives & examples

A/ Parameter estimation

<u>Ex1</u>: Weighing with a two-pan balance

☞ Determine the weights of 8 objets, with mass m_i, $i = 1, \ldots, 8$

i.i.d. errors $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

Method a: weigh each objet successively

$\rightarrow y(i) = m_i + \varepsilon_i,\ i = 1, \ldots, 8$

$\rightarrow$ estimated weights : $\hat{m}_i = y_i \sim \mathcal{N}(m_i, \sigma^2)$

Repeat 8 times, average the results: $\hat{\hat{m}}_i \sim \mathcal{N}(m_i, \sigma^2/8)$ (<u>with 64 observations</u>. . .)

Method b: more sophisticated...

$$
\begin{aligned}
y_1 &= m_1 + m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + \varepsilon_1 \\
y_2 &= m_1 + m_2 + m_3 - m_4 - m_5 - m_6 - m_7 + m_8 + \varepsilon_2 \\
y_3 &= m_1 - m_2 - m_3 + m_4 + m_5 - m_6 - m_7 + m_8 + \varepsilon_3 \\
y_4 &= m_1 - m_2 - m_3 - m_4 - m_5 + m_6 + m_7 + m_8 + \varepsilon_4 \\
y_5 &= -m_1 + m_2 - m_3 + m_4 - m_5 + m_6 - m_7 + m_8 + \varepsilon_5 \\
y_6 &= -m_1 + m_2 - m_3 - m_4 + m_5 - m_6 + m_7 + m_8 + \varepsilon_6 \\
y_7 &= -m_1 - m_2 + m_3 + m_4 - m_5 - m_6 + m_7 + m_8 + \varepsilon_7 \\
y_8 &= -m_1 - m_2 + m_3 - m_4 + m_5 + m_6 - m_7 + m_8 + \varepsilon_8
\end{aligned}
$$

$$
\begin{aligned}
\to \quad \hat{m}_8 &= \frac{1}{8}\sum_{i=1}^{8} y_i \\
&= m_8 + \frac{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_6 + \varepsilon_7 + \varepsilon_8}{8} \\
&\sim \mathcal{N}(m_8, \sigma^2/8) \quad \text{(idem for all } m_j,\ j \leq 7)
\end{aligned}
$$

Method b: more sophisticated...

$$
\begin{aligned}
y_1 &= m_1 + m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + \varepsilon_1 \\
y_2 &= m_1 + m_2 + m_3 - m_4 - m_5 - m_6 - m_7 + m_8 + \varepsilon_2 \\
y_3 &= m_1 - m_2 - m_3 + m_4 + m_5 - m_6 - m_7 + m_8 + \varepsilon_3 \\
y_4 &= m_1 - m_2 - m_3 - m_4 - m_5 + m_6 + m_7 + m_8 + \varepsilon_4 \\
y_5 &= -m_1 + m_2 - m_3 + m_4 - m_5 + m_6 - m_7 + m_8 + \varepsilon_5 \\
y_6 &= -m_1 + m_2 - m_3 - m_4 + m_5 - m_6 + m_7 + m_8 + \varepsilon_6 \\
y_7 &= -m_1 - m_2 + m_3 + m_4 - m_5 - m_6 + m_7 + m_8 + \varepsilon_7 \\
y_8 &= -m_1 - m_2 + m_3 - m_4 + m_5 + m_6 - m_7 + m_8 + \varepsilon_8
\end{aligned}
$$

$$
\begin{aligned}
\to \quad \hat{m}_8 &= \frac{1}{8}\sum_{i=1}^{8} y_i \\
&= m_8 + \frac{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 + \varepsilon_5 + \varepsilon_6 + \varepsilon_7 + \varepsilon_8}{8} \\
&\sim \mathcal{N}(m_8, \sigma^2/8) \quad \text{(idem for all } m_j,\ j \leq 7)
\end{aligned}
$$

➠ **8 observations only, against 64 with method a!**

Here, selection of a good design = combinatorial problem

$y_k = \sum_{i=1}^{8} \mathbf{f}_{k\,i} m_i + \varepsilon_k = \mathbf{f}_k^\top \mathbf{m} + \varepsilon_k$,

(e.g., in Method b $\mathbf{f}_2 = [1\ 1\ 1\ -1\ -1\ -1\ -1\ 1]^\top$)

$\mathbf{y} = \mathbf{F}\mathbf{m} + \varepsilon$ with

method a: $\mathbf{F}_a = \mathbf{I}_8$

method b: $\mathbf{F}_b = 8 \times 8$ Hadamard matrix, $\mathbf{F}_b^\top \mathbf{F}_b = 8\,\mathbf{I}_8$

(= fractional factorial design with 2 levels)

$$\text{LS estimator } \hat{\mathbf{m}} = \arg\min_{\mathbf{m}} \sum_{k=1}^{n} [y_k - \mathbf{f}_k^\top \mathbf{m}]^2$$

$$= \left(\sum_{k=1}^{n} \mathbf{f}_k \mathbf{f}_k^\top \right)^{-1} \sum_{k=1}^{n} y_k \mathbf{f}_k = (\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{y}$$

Here, selection of a good design = combinatorial problem

$y_k = \sum_{i=1}^{8} \mathbf{f}_{k\,i} m_i + \varepsilon_k = \mathbf{f}_k^\top \mathbf{m} + \varepsilon_k$,

(e.g., in Method b $\mathbf{f}_2 = [1\ 1\ 1\ -1\ -1\ -1\ -1\ 1]^\top$)

$\mathbf{y} = \mathbf{F}\mathbf{m} + \varepsilon$ with

method a: $\mathbf{F}_a = \mathbf{I}_8$

method b: $\mathbf{F}_b = 8 \times 8$ Hadamard matrix, $\mathbf{F}_b^\top \mathbf{F}_b = 8\,\mathbf{I}_8$

(= fractional factorial design with 2 levels)

$$\text{LS estimator } \hat{\mathbf{m}} = \arg\min_{\mathbf{m}} \sum_{k=1}^{n} [y_k - \mathbf{f}_k^\top \mathbf{m}]^2$$

$$= \left(\sum_{k=1}^{n} \mathbf{f}_k \mathbf{f}_k^\top \right)^{-1} \sum_{k=1}^{n} y_k \mathbf{f}_k = (\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{y}$$

$\Longrightarrow$ Choose the $\mathbf{f}_k$'s such that $\mathbf{M}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{f}_k \mathbf{f}_k^\top = \frac{1}{n} \mathbf{F}^\top \mathbf{F}$ is nonsingular

Here, selection of a good design = combinatorial problem

$y_k = \sum_{i=1}^{8} \mathbf{f}_{k\,i} m_i + \varepsilon_k = \mathbf{f}_k^\top \mathbf{m} + \varepsilon_k$,

(e.g., in Method b $\mathbf{f}_2 = [1\ 1\ 1\ -1\ -1\ -1\ -1\ 1]^\top$)

$\mathbf{y} = \mathbf{F}\mathbf{m} + \varepsilon$ with

method a: $\mathbf{F}_a = \mathbf{I}_8$

method b: $\mathbf{F}_b = 8 \times 8$ Hadamard matrix, $\mathbf{F}_b^\top \mathbf{F}_b = 8\,\mathbf{I}_8$

(= fractional factorial design with 2 levels)

$$\text{LS estimator } \hat{\mathbf{m}} = \arg\min_{\mathbf{m}} \sum_{k=1}^{n} [y_k - \mathbf{f}_k^\top \mathbf{m}]^2$$

$$= \left(\sum_{k=1}^{n} \mathbf{f}_k \mathbf{f}_k^\top\right)^{-1} \sum_{k=1}^{n} y_k \mathbf{f}_k = (\mathbf{F}^\top \mathbf{F})^{-1} \mathbf{F}^\top \mathbf{y}$$

$\Longrightarrow$ Choose the $\mathbf{f}_k$'s such that $\mathbf{M}_n = \frac{1}{n} \sum_{k=1}^{n} \mathbf{f}_k \mathbf{f}_k^\top = \frac{1}{n} \mathbf{F}^\top \mathbf{F}$ is nonsingular

$\mathrm{E}\{\hat{\mathbf{m}}\} = \mathbf{m}$ (no bias)

$\mathrm{E}\{(\hat{\mathbf{m}} - \mathbf{m})(\hat{\mathbf{m}} - \mathbf{m})^\top\} = \frac{\sigma^2}{n} \mathbf{M}_n^{-1}$

$\Longrightarrow$ **minimize a scalar function of $\mathbf{M}_n^{-1}$**

In this particular situation: combinatorial problem (since $\mathbf{f}_{k\,i} \in \{-1, 0, 1\}$) [Fisher 1925 . . .]

More generally, when the design variables (*inputs*) are real numbers, optimum design for parameter estimation is obtained by optimization of a scalar function of the (asymptotic) covariance matrix of the estimator

Ex2: [D'Argenio 1981]: two-compartment model in pharmacokinetics
A product x is injected in blood ($\rightarrow$ input $u(t)$),
$x_C(t)$ (product in blood) moves to another tissue $\rightarrow x_P(t)$

Ex2: [D'Argenio 1981]: two-compartment model in pharmacokinetics
A product x is injected in blood ($\rightarrow$ input $u(t)$),
$x_C(t)$ (product in blood) moves to another tissue $\rightarrow x_P(t)$

$\rightarrow$ Linear differential equations:

$$\left\{ \begin{array}{rcl} \frac{dx_C(t)}{dt} & = & (-K_{EL} - K_{CP})x_C(t) + K_{PC}x_P(t) + u(t) \\ \frac{dx_P(t)}{dt} & = & K_{CP}x_C(t) - K_{PC}x_P(t) \end{array} \right.$$

we observe the concentration of x in blood: $y(t) = x_C(t)/V + \varepsilon(t)$,
the errors $\epsilon(t_i)$'s are i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.2\mu$g/ml

Ex2: [D'Argenio 1981]: two-compartment model in pharmacokinetics
A product x is injected in blood ($\rightarrow$ input $u(t)$),
$x_C(t)$ (product in blood) moves to another tissue $\rightarrow x_P(t)$

$\rightarrow$ Linear differential equations:

$$\begin{cases} \frac{dx_C(t)}{dt} = (-K_{EL} - K_{CP})x_C(t) + K_{PC}x_P(t) + u(t) \\ \frac{dx_P(t)}{dt} = K_{CP}x_C(t) - K_{PC}x_P(t) \end{cases}$$

we observe the concentration of x in blood: $y(t) = x_C(t)/V + \varepsilon(t)$,
the errors $\epsilon(t_i)$'s are i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.2\mu$g/ml

There are 4 unknown parameters $\theta = (K_{CP}, K_{PC}, K_{EL}, V)$
The profile of the input $u(t)$ is given (fast infusion 75 mg/min for 1 min, then 1.45 mg/min)

Ex2: [D'Argenio 1981]: two-compartment model in pharmacokinetics
A product x is injected in blood ($\rightarrow$ input $u(t)$),
$x_C(t)$ (product in blood) moves to another tissue $\rightarrow x_P(t)$

$\rightarrow$ Linear differential equations:

$$\begin{cases} \frac{dx_C(t)}{dt} = (-K_{EL} - K_{CP})x_C(t) + K_{PC}x_P(t) + u(t) \\ \frac{dx_P(t)}{dt} = K_{CP}x_C(t) - K_{PC}x_P(t) \end{cases}$$

we observe the concentration of x in blood: $y(t) = x_C(t)/V + \varepsilon(t)$,
the errors $\epsilon(t_i)$'s are i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.2\mu$g/ml

There are 4 unknown parameters $\theta = (K_{CP}, K_{PC}, K_{EL}, V)$
The profile of the input $u(t)$ is given (fast infusion 75 mg/min for 1 min, then 1.45 mg/min)
☞ **simulated experiments** with «true» parameter values

$$\bar{\theta} = (0.066\,\text{min}^{-1},\ 0.038\,\text{min}^{-1},\ 0.0242\,\text{min}^{-1},\ 30\,\text{l})$$

Experimental variables = sampling times t_i, $1 \leq t_i \leq 720$ min

- «conventional» design :

$$\mathbf{t} = (5, 10, 30, 60, 120, 180, 360, 720) \text{ (in min)}$$

Experimental variables = sampling times t_i, $1 \leq t_i \leq 720$ min

- «conventional» design :

$$\mathbf{t} = (5, 10, 30, 60, 120, 180, 360, 720) \text{ (in min)}$$

- «optimal» design (for $\bar{\theta}$) :

$$\mathbf{t} = (1, 1, 10, 10, 74, 74, 720, 720) \text{ (in min)}$$

(assumes that independent measurements at the same time are possible)

Experimental variables = sampling times t_i, $1 \leq t_i \leq 720$ min

- «conventional» design :

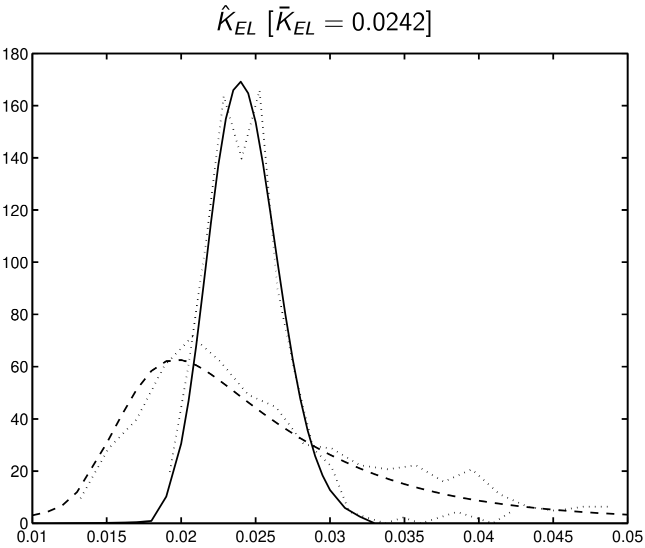
$$\mathbf{t} = (5, 10, 30, 60, 120, 180, 360, 720) \text{ (in min)}$$

- «optimal» design (for $\bar{\theta}$) :

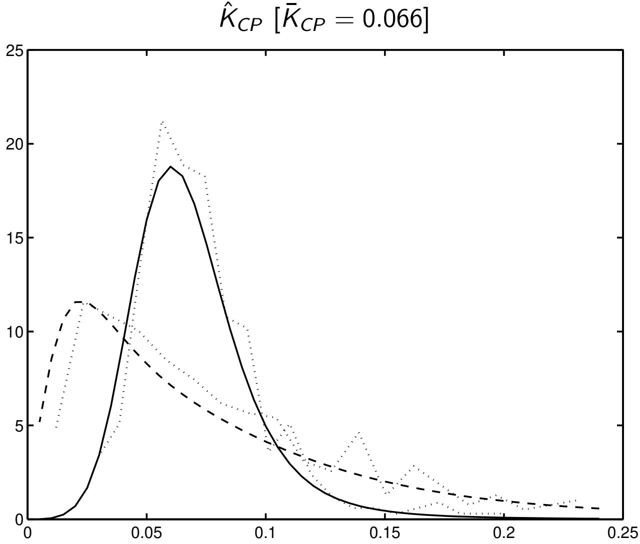
$$\mathbf{t} = (1, 1, 10, 10, 74, 74, 720, 720) \text{ (in min)}$$

(assumes that independent measurements at the same time are possible)
→ 400 simulations
→ 400 sets of 8 observations each, for each design
→ 400 parameter estimates (LS) for each design . . .
➔ histograms of $\hat{\theta}_i$
(and approximated marginals [Pázman & P 1996])

➠ «optimal» design gives more precise estimation

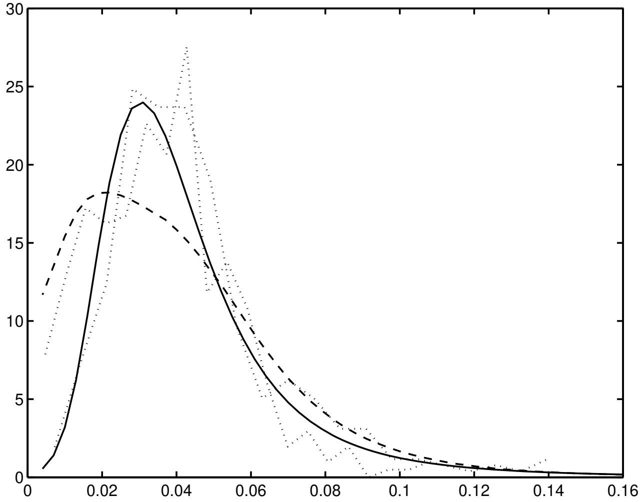


➠ «optimal» design gives more precise estimation

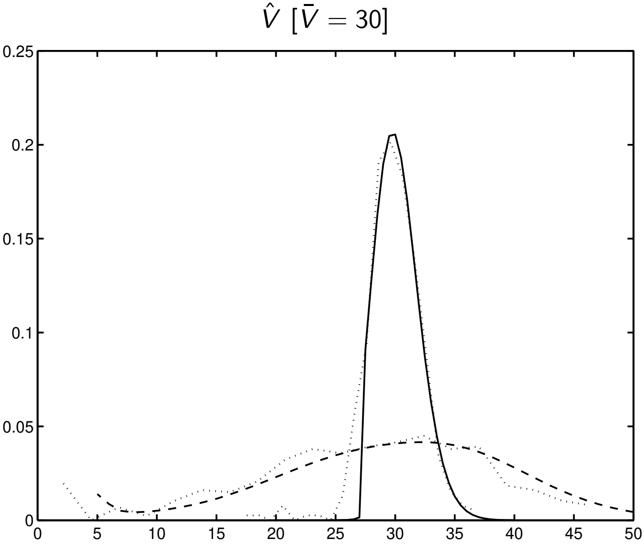


➠ «optimal» design gives more precise estimation

$\hat{K}_{PC}$ [$\bar{K}_{PC} = 0.038$]



➠ «optimal» design gives more precise estimation



B/ Model discrimination

Ex3: [Box & Hill 1967] Chemical reaction $A \to B$
2 design variables: $\mathbf{x} = (\text{time } t, \text{ temperature } T)$
reaction of 1st, 2nd, 3rd ou 4th order?
$\to$ 4 model structures are candidate:

$$\begin{aligned}
\eta^{(1)}(\mathbf{x}, \theta_1) &= \exp[-\theta_{11} t \exp(-\theta_{12}/T)] \\
\eta^{(2)}(\mathbf{x}, \theta_2) &= \frac{1}{1 + \theta_{21} t \exp(-\theta_{22}/T)} \\
\eta^{(3)}(\mathbf{x}, \theta_3) &= \frac{1}{[1 + 2\theta_{31} t \exp(-\theta_{32}/T)]^{1/2}} \\
\eta^{(4)}(\mathbf{x}, \theta_4) &= \frac{1}{[1 + 3\theta_{41} t \exp(-\theta_{42}/T)]^{1/3}}
\end{aligned}$$

Simulated experiment

Observations with 2nd structure («true»): $y(\mathbf{x}_j) = \eta^{(2)}(\mathbf{x}_j, \bar{\theta}_2) + \varepsilon_j$, with

➤ $\bar{\theta}_2 = (400, 5000)^\top$ the «true» value (unknown) of parameters in model 2

➤ (ϵ_j) i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.05$

Admissible experimental domain: $0 \leq t \leq 150$, $450 \leq T \leq 600$

Simulated experiment

Observations with 2nd structure («true»): $y(\mathbf{x}_j) = \eta^{(2)}(\mathbf{x}_j, \bar{\theta}_2) + \varepsilon_j$, with

➤ $\bar{\theta}_2 = (400,\ 5000)^\top$ the «true» value (unknown) of parameters in model 2

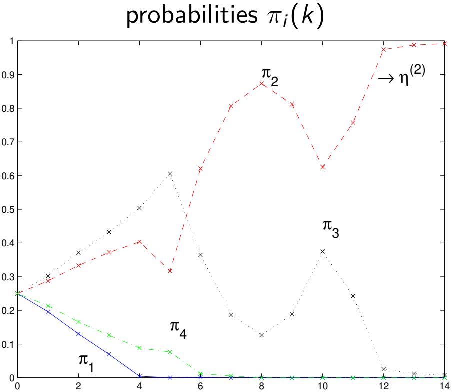
➤ (ϵ_j) i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.05$

Admissible experimental domain: $0 \leq t \leq 150$, $450 \leq T \leq 600$

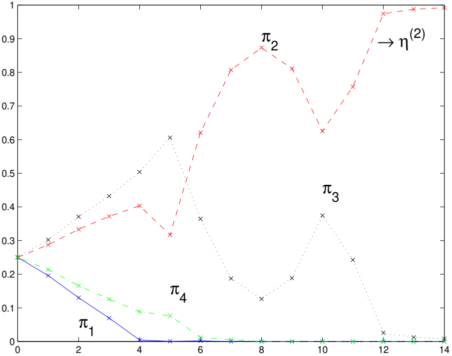
Sequential design: after the observation of $y(\mathbf{x}_j)$, $j = 1, \ldots, k$,

- estimate $\hat{\theta}_i^k$ (LS) for $i = 1, 2, 3, 4$
- compute posterior probability $\pi_i(k)$ that model i is correct for $i = 1, 2, 3, 4$

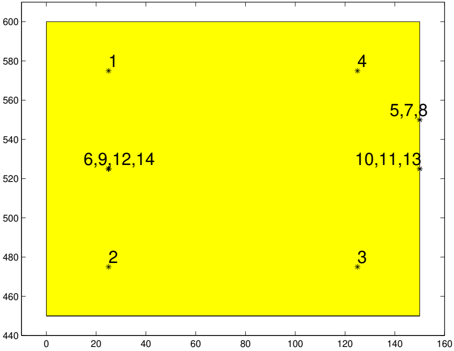
Initialization: $\pi_i(0) = 1/4$, $i = 1, \ldots, 4$ and $\mathbf{x}_1, \ldots, \mathbf{x}_4$ are given


probabilities $\pi_i(k)$
π_1
π_2
π_3
π_4
$\rightarrow \eta^{(2)}$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
2
4
6
8
10
12
14

probabilities $\pi_i(k)$



design points $\mathbf{x}_k$



Design for discrimination is not considered in the following

A simple sequential method for discriminating between two structures $\eta^{(1)}(\mathbf{x}, \theta_1)$, $\eta^{(2)}(\mathbf{x}, \theta_2)$ [Atkinson & Fedorov 1975]

- After observation of $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_k)$ estimate $\hat{\theta}_1^k$ and $\hat{\theta}_2^k$ for both models
- place next point $\mathbf{x}_{k+1}$ where $[\eta^{(1)}(\mathbf{x}, \hat{\theta}_1^k) - \eta^{(2)}(\mathbf{x}, \hat{\theta}_2^k)]^2$ is maximum
- $k \rightarrow k+1$, repeat

Design for discrimination is not considered in the following

A simple sequential method for discriminating between two structures $\eta^{(1)}(\mathbf{x}, \theta_1)$, $\eta^{(2)}(\mathbf{x}, \theta_2)$ [Atkinson & Fedorov 1975]

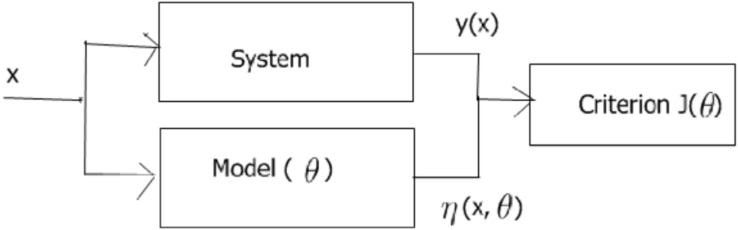
- After observation of $y(\mathbf{x}_1), \ldots, y(\mathbf{x}_k)$ estimate $\hat{\theta}_1^k$ and $\hat{\theta}_2^k$ for both models
- place next point $\mathbf{x}_{k+1}$ where $[\eta^{(1)}(\mathbf{x}, \hat{\theta}_1^k) - \eta^{(2)}(\mathbf{x}, \hat{\theta}_2^k)]^2$ is maximum
- $k \rightarrow k + 1$, repeat

If more than two models: estimate $\hat{\theta}_i^k$ for all of them, place next point using the two models with best and second best fitting
(see [Atkinson & Cox 1974; Hill 1978] for surveys)

2 DoE based on asymptotic normality

A/ Regression models

$y_i = y(x_i) = \eta(x_i, \theta) + \varepsilon_i$



- System: physical experimental device, experimental conditions x_i, $i = 1, 2, \ldots, n$
- Model(θ): mathematical equations, parameters $\theta = (\theta_1, \ldots, \theta_p)^\top$ (response $\eta(x, \theta)$ known explicitly or result of simulation of ODEs or PDEs)
- Criterion: similarity between $y_i = y(x_i)$ and $\eta(x_i, \theta)$, $i = 1, 2, \ldots, n$, e.g., $J(\theta) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \eta(x_i, \theta)]^2$ (LS)

Remarks:

- Model(θ) also provides derivatives $\partial\eta(x,\theta)/\partial\theta = (\partial\eta(x,\theta)/\partial\theta_1, \ldots, \partial\eta(x,\theta)/\partial\theta_p)^\top$
  — plus higher-order derivatives if necessary—
  via simulation of sensitivity functions, or automatic differentiation (adjoint code)

Remarks:

- Model(θ) also provides derivatives $\partial\eta(x,\theta)/\partial\theta = (\partial\eta(x,\theta)/\partial\theta_1, \ldots, \partial\eta(x,\theta)/\partial\theta_p)^\top$
  — plus higher-order derivatives if necessary—
  via simulation of sensitivity functions, or automatic differentiation (adjoint code)
- Criterion: other criteria than LS can be used, e.g., $J(\theta) = \frac{1}{n}\sum_{i=1}^{n} |y_i - \eta(x_i,\theta)|$ ($\rightarrow$ robust estimation), including Maximum-Likelihood (ML) estimation in more general settings ($J(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log\pi(y_i|\theta) \rightarrow$ max!)

Remarks:

- Model(θ) also provides derivatives $\partial\eta(x,\theta)/\partial\theta = (\partial\eta(x,\theta)/\partial\theta_1, \ldots, \partial\eta(x,\theta)/\partial\theta_p)^\top$ — plus higher-order derivatives if necessary— via simulation of sensitivity functions, or automatic differentiation (adjoint code)
- Criterion: other criteria than LS can be used, e.g., $J(\theta) = \frac{1}{n}\sum_{i=1}^n |y_i - \eta(x_i,\theta)|$ ($\rightarrow$ robust estimation), including Maximum-Likelihood (ML) estimation in more general settings ($J(\theta) = \frac{1}{n}\sum_{i=1}^n \log\pi(y_i|\theta) \rightarrow$ max!)
- We always assume independent observations $y(x_i)$ (independent ε_i in regression) — often much more difficult otherwise!

Remarks:

- Model(θ) also provides derivatives
  $\partial\eta(x,\theta)/\partial\theta = (\partial\eta(x,\theta)/\partial\theta_1, \ldots, \partial\eta(x,\theta)/\partial\theta_p)^\top$
  — plus higher-order derivatives if necessary—
  via simulation of sensitivity functions, or automatic differentiation (adjoint code)
- Criterion: other criteria than LS can be used, e.g.,
  $J(\theta) = \frac{1}{n}\sum_{i=1}^n |y_i - \eta(x_i,\theta)|$ ($\rightarrow$ robust estimation), including Maximum-Likelihood (ML) estimation in more general settings
  ($J(\theta) = \frac{1}{n}\sum_{i=1}^n \log \pi(y_i|\theta) \rightarrow$ max!)
- We always assume independent observations $y(x_i)$ (independent ε_i in regression) — often much more difficult otherwise!

Most of the following can be found in [P & Pázman: Design of Experiments in Nonlinear Models, Springer, 2013]

B/ LS estimation

$$\hat{\theta}^n = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} [y_i - \eta(x_i, \theta)]^2$$

B/ LS estimation

$$\boxed{\hat{\theta}^n = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n [y_i - \eta(x_i,\theta)]^2}$$

Linear model: $\eta(x,\theta) = \mathbf{f}^\top(x)\theta \rightarrow \boxed{\hat{\theta}^n = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{y}}$,
with $\mathbf{y} = (y_1,\ldots,y_n)^\top$ and $\mathbf{F}^\top = (\mathbf{f}(x_1),\ldots,\mathbf{f}(x_n))^\top$

$\Longrightarrow$ choose the x_i such that $\mathbf{M}_n = \frac{1}{n}\mathbf{F}^\top\mathbf{F}$ has full rank
($\mathbf{M}_n$ = normalized information matrix)

B/ LS estimation

$$\boxed{\hat{\theta}^n = \arg\min_\theta \frac{1}{n}\sum_{i=1}^n [y_i - \eta(x_i,\theta)]^2}$$

Linear model: $\eta(x,\theta) = \mathbf{f}^\top(x)\theta \rightarrow \boxed{\hat{\theta}^n = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{y}}$,
with $\mathbf{y} = (y_1,\ldots,y_n)^\top$ and $\mathbf{F}^\top = (\mathbf{f}(x_1),\ldots,\mathbf{f}(x_n))^\top$

$\Longrightarrow$ choose the x_i such that $\mathbf{M}_n = \frac{1}{n}\mathbf{F}^\top\mathbf{F}$ has full rank
($\mathbf{M}_n$ = normalized information matrix)

Since $y_i = \mathbf{f}^\top(x_i)\bar{\theta} + \varepsilon_i$ for some $\bar{\theta}$ and $\mathrm{E}\{\varepsilon_i\} = 0$ for all i, $\mathrm{E}\{\mathbf{y}\} = \mathbf{F}\bar{\theta}$
and $\mathrm{E}\{\hat{\theta}^n\} = \bar{\theta}$
Also, $\mathrm{Var}(\hat{\theta}^n) = \mathrm{E}\{(\hat{\theta}^n - \bar{\theta})(\hat{\theta}^n - \bar{\theta})^\top\} = \sigma^2(\mathbf{F}^\top\mathbf{F})^{-1} = \frac{\sigma^2}{n}\mathbf{M}_n^{-1}$ when the ε_i are i.i.d. with finite variance σ^2
$\Longrightarrow$ choose the x_i to minimize a scalar function of $\mathbf{M}_n^{-1}$
(see Example 1: weighing with a two-pan balance)

Nonlinear model: $\eta(x, \theta)$
Under «standard» assumptions ($\theta \in \Theta$ compact, $\eta(x, \theta)$ continuous in θ for all $x \ldots$) and for a suitable sequence (x_i)

$\boxed{\hat{\theta}^n \stackrel{\text{a.s.}}{\to} \bar{\theta} \text{ as } n \to \infty}$ (strong consistency)

Nonlinear model: $\eta(x,\theta)$

Under «standard» assumptions ($\theta \in \Theta$ compact, $\eta(x,\theta)$ continuous in θ for all $x \ldots$) and for a suitable sequence (x_i)

$\boxed{\hat{\theta}^n \stackrel{\text{a.s.}}{\to} \bar{\theta} \text{ as } n \to \infty}$ (strong consistency)

Moreover, under «standard» regularity assumptions ($\eta(x,\theta)$ twice continuously differentiable in θ for all $x \ldots$), for i.i.d. errors ε_i with finite variance σ^2, for a suitable sequence (x_i)

$\boxed{\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \stackrel{\text{d}}{\to} \mathcal{N}(\mathbf{0}, \sigma^2\, \mathbf{M}^{-1}) \text{ as } n \to \infty}$ (asymptotic normality)

with $\mathbf{M} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \eta(x_i,\theta)}{\partial \theta}\Big|_{\bar{\theta}} \frac{\partial \eta(x_i,\theta)}{\partial \theta^\top}\Big|_{\bar{\theta}}$ (information matrix)

Nonlinear model: $\eta(x,\theta)$
Under «standard» assumptions ($\theta \in \Theta$ compact, $\eta(x,\theta)$ continuous in θ for all $x\ldots$) and for a suitable sequence (x_i)

$$\boxed{\hat{\theta}^n \overset{\text{a.s.}}{\to} \bar{\theta} \text{ as } n \to \infty}$$ (strong consistency)

Moreover, under «standard» regularity assumptions ($\eta(x,\theta)$ twice continuously differentiable in θ for all $x\ldots$), for i.i.d. errors ε_i with finite variance σ^2, for a suitable sequence (x_i)

$$\boxed{\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \overset{\text{d}}{\to} \mathcal{N}(\mathbf{0}, \sigma^2\,\mathbf{M}^{-1}) \text{ as } n \to \infty}$$ (asymptotic normality)

with $\mathbf{M} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \eta(x_i,\theta)}{\partial \theta}\Big|_{\bar{\theta}} \frac{\partial \eta(x_i,\theta)}{\partial \theta^\top}\Big|_{\bar{\theta}}$ (information matrix)

$\Longrightarrow$ choose the x_i (design) to minimize a scalar function of $\mathbf{M}^{-1}$,
or maximize a function $\Phi(\mathbf{M})$

= **classical approach for DoE**
(see Example 2 with a two-compartment model)

Remarks:

- Weighted LS: suppose heteroscedastic errors
  $\mathrm{var}\{\varepsilon_i\} = \mathrm{E}\{\varepsilon_i^2\} = \mathrm{E}\{\varepsilon^2(x_i)\} = \sigma^2(x_i)$

Weighted LS estimator $\hat{\theta}^n_{WLS}$ minimizes $J_{WLS}(\theta) = \frac{1}{n}\sum_{i=1}^n w(x_i)\,[y_i - \eta(x_i,\theta)]^2$

Remarks:

- Weighted LS: suppose heteroscedastic errors
  $\mathrm{var}\{\varepsilon_i\} = \mathrm{E}\{\varepsilon_i^2\} = \mathrm{E}\{\varepsilon^2(x_i)\} = \sigma^2(x_i)$
  Weighted LS estimator $\hat{\theta}^n_{WLS}$ minimizes $J_{WLS}(\theta) = \frac{1}{n}\sum_{i=1}^n w(x_i)\,[y_i - \eta(x_i,\theta)]^2$

  Strong consistency and asymptotic normality $\sqrt{n}(\hat{\theta}^n_{WLS} - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \mathbf{C})$ as $n \to \infty$, where
  $\mathbf{C} = \mathbf{M}_a^{-1}\mathbf{M}_b\mathbf{M}_a^{-1}$ and
  $\mathbf{M}_a = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n w(x_i)\, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\bar{\theta}} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\bar{\theta}}$
  $\mathbf{M}_b = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n w^2(x_i)\,\sigma^2(x_i)\, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\bar{\theta}} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\bar{\theta}}$

Remarks:

- Weighted LS: suppose heteroscedastic errors
  $\mathrm{var}\{\varepsilon_i\} = \mathrm{E}\{\varepsilon_i^2\} = \mathrm{E}\{\varepsilon^2(x_i)\} = \sigma^2(x_i)$
  Weighted LS estimator $\hat{\theta}^n_{WLS}$ minimizes $J_{WLS}(\theta) = \frac{1}{n}\sum_{i=1}^n w(x_i)\,[y_i - \eta(x_i,\theta)]^2$

  Strong consistency and asymptotic normality $\sqrt{n}(\hat{\theta}^n_{WLS} - \bar{\theta}) \stackrel{\mathrm{d}}{\to} \mathcal{N}(\mathbf{0},\mathbf{C})$ as $n \to \infty$, where
  $\mathbf{C} = \mathbf{M}_a^{-1}\mathbf{M}_b\mathbf{M}_a^{-1}$ and
  $\mathbf{M}_a = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n w(x_i)\, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\bar{\theta}} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\bar{\theta}}$
  $\mathbf{M}_b = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n w^2(x_i)\,\sigma^2(x_i)\, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\bar{\theta}} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\bar{\theta}}$

  ➠ $\mathbf{C} \succeq \mathbf{M}^{-1}$
  $\mathbf{M} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma^2(x_i)}\, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\bar{\theta}} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\bar{\theta}}$

  ➠ $\mathbf{C} = \mathbf{M}^{-1}$ when $w(x) \propto \sigma^{-2}(x)$

$\Longrightarrow$ choose the best estimator, then the best design

Remarks (continued):

- One may also consider the case $\text{var}\{\varepsilon_i\} = \sigma^2(x_i, \theta)$ (errors with parameterized variance)

  ➠ Use two-stage LS: 1/ use $w(x) \equiv 1 \rightarrow \hat{\theta}^n_{(1)}$; 2/ use $w(x) = \sigma^{-2}(x, \hat{\theta}^n_{(1)})$
  or use iteratively-reweighted LS (i.e., go on with more stages), or penalized LS…

Remarks (continued):

- One may also consider the case $\text{var}\{\varepsilon_i\} = \sigma^2(x_i, \theta)$ (errors with parameterized variance)
  ➠ Use two-stage LS: 1/ use $w(x) \equiv 1 \rightarrow \hat{\theta}^n_{(1)}$; 2/ use $w(x) = \sigma^{-2}(x, \hat{\theta}^n_{(1)})$
  or use iteratively-reweighted LS (i.e., go on with more stages), or penalized LS. . .
- Similar asymptotic results for ML estimation $\sqrt{n}(\hat{\theta}^n_{ML} - \bar{\theta}) \xrightarrow{\text{d}} \mathcal{N}(\mathbf{0}, \sigma^2 \, \mathbf{M}_F^{-1})$ as $n \rightarrow \infty$, with $\mathbf{M}_F$ = Fisher information matrix

Remarks (continued):

- One may also consider the case $\mathrm{var}\{\varepsilon_i\} = \sigma^2(x_i, \theta)$ (errors with parameterized variance)
  ➠ Use two-stage LS: 1/ use $w(x) \equiv 1 \rightarrow \hat{\theta}^n_{(1)}$; 2/ use $w(x) = \sigma^{-2}(x, \hat{\theta}^n_{(1)})$
  or use iteratively-reweighted LS (i.e., go on with more stages), or penalized LS...
- Similar asymptotic results for ML estimation $\sqrt{n}(\hat{\theta}^n_{ML} - \bar{\theta}) \xrightarrow{\mathrm{d}} \mathcal{N}(\mathbf{0}, \sigma^2\, \mathbf{M}_F^{-1})$ as $n \rightarrow \infty$, with $\mathbf{M}_F$ = Fisher information matrix
- Model(θ) = linear ODE, experimental design = system input $u(t)$
  ➠ ($\approx$ simple) analytic expression for $\mathbf{M}$
  ➠ optimal input design $\Leftrightarrow$ optimal control problem (frequency domain $\rightarrow$ optimal combination of sinusoidal signals) [Goodwin & Payne 1977; Zarrop 1979; Ljung 1987; Walter & P 1994, 1997]

C/ Design based on the information matrix

Maximize $\Phi(\mathbf{M})$, but which $\Phi(\cdot)$?

➠ There are many possibilities!

LS estimation in linear regression with i.i.d. errors $\mathcal{N}(0, \sigma^2)$

$$\mathcal{R}(\hat{\theta}^n, \alpha) = \{\theta \in \mathbb{R}^p : (\theta - \hat{\theta}^n)^\top \mathbf{M}_n(\theta - \hat{\theta}^n) \leq \frac{\sigma^2}{n} \chi^2_p(1-\alpha)\}$$

$=$ confidence region (ellipsoid) at level α: $\text{Prob}\{\bar{\theta} \in \mathcal{R}(\hat{\theta}^n, \alpha)\} = \alpha$

(**asymptotically true** in nonlinear situations — e.g., nonlinear regression)

C/ Design based on the information matrix

Maximize $\Phi(\mathbf{M})$, but which $\Phi(\cdot)$?

➠ There are many possibilities!

LS estimation in linear regression with i.i.d. errors $\mathcal{N}(0, \sigma^2)$

$$\mathcal{R}(\hat{\theta}^n, \alpha) = \{\theta \in \mathbb{R}^p : (\theta - \hat{\theta}^n)^\top \mathbf{M}_n(\theta - \hat{\theta}^n) \leq \frac{\sigma^2}{n}\chi_p^2(1-\alpha)\}$$

$=$ confidence region (ellipsoid) at level α: $\text{Prob}\{\bar{\theta} \in \mathcal{R}(\hat{\theta}^n, \alpha)\} = \alpha$

(**asymptotically true** in nonlinear situations — e.g., nonlinear regression)

➠ Most criteria can be related to geometrical properties of $\mathcal{R}(\hat{\theta}^n, \alpha)$

C/ Design based on the information matrix

Maximize $\Phi(\mathbf{M})$, but which $\Phi(\cdot)$?

➠ There are many possibilities!

LS estimation in linear regression with i.i.d. errors $\mathcal{N}(0,\sigma^2)$

$$\mathcal{R}(\hat{\theta}^n,\alpha) = \{\theta \in \mathbb{R}^p : (\theta - \hat{\theta}^n)^\top \mathbf{M}_n (\theta - \hat{\theta}^n) \leq \frac{\sigma^2}{n} \chi_p^2(1-\alpha)\}$$

= confidence region (ellipsoid) at level α: Prob$\{\bar{\theta} \in \mathcal{R}(\hat{\theta}^n,\alpha)\} = \alpha$

(**asymptotically true** in nonlinear situations — e.g., nonlinear regression)

➠ Most criteria can be related to geometrical properties of $\mathcal{R}(\hat{\theta}^n,\alpha)$

➠ Nonlinear model $\Longrightarrow$ $\mathbf{M} = \mathbf{M}(\theta)$ depends on the θ where $\eta(x,\theta)$ is linearized:

for the moment use a nominal value θ^0

➠ **locally optimum design**

A few choices for $\Phi(\cdot)$

- *A*-optimality: maximize $-\mathrm{trace}[\mathbf{M}^{-1}] \Leftrightarrow$ maximize $1/\mathrm{trace}[\mathbf{M}^{-1}]$
$\Leftrightarrow$ minimize the sum of lengths^2 of axes of (asymptotic) confidence ellipsoids $\mathcal{R}(\hat{\theta}^n, \alpha)$

A few choices for $\Phi(\cdot)$

- *A*-optimality: maximize $-\mathrm{trace}[\mathbf{M}^{-1}] \Leftrightarrow$ maximize $1/\mathrm{trace}[\mathbf{M}^{-1}]$
  $\Leftrightarrow$ minimize the sum of lengths^2 of axes of (asymptotic) confidence ellipsoids $\mathcal{R}(\hat{\theta}^n, \alpha)$
- *E*-optimality: maximize $\lambda_{\min}(\mathbf{M})$
  $\Leftrightarrow$ minimize the longest axis of $\mathcal{R}(\hat{\theta}^n, \alpha)$

A few choices for $\Phi(\cdot)$

- *A*-optimality: maximize $-\mathrm{trace}[\mathbf{M}^{-1}] \Leftrightarrow$ maximize $1/\mathrm{trace}[\mathbf{M}^{-1}]$
  $\Leftrightarrow$ minimize the sum of lengths^2 of axes of (asymptotic) confidence ellipsoids $\mathcal{R}(\hat{\theta}^n, \alpha)$
- *E*-optimality: maximize $\lambda_{\min}(\mathbf{M})$
  $\Leftrightarrow$ minimize the longest axis of $\mathcal{R}(\hat{\theta}^n, \alpha)$
- *D*-optimality: maximize $\log\det\mathbf{M}$
  $\Leftrightarrow$ minimize volume of $\mathcal{R}(\hat{\theta}^n, \alpha)$ (proportional to $1/\sqrt{\det\mathbf{M}}$)
  Very much used:
  - a *D*-optimum design is invariant by reparametrization

$$\det\mathbf{M}'(\beta(\theta)) = \det\mathbf{M}(\theta)\det{}^{-2}\left(\frac{\partial\beta}{\partial\theta^\top}\right)$$

  - often leads to repeat the same experimental conditions (replications)
    (remember Ex2: $\dim(\theta) = 4 \rightarrow 4$ different sampling times, several observations at each)

- D_s-optimality: only $s < p$ parameters of interest (and $p - s$ «nuisance» parameters) $\rightarrow \theta^\top = (\theta_1^\top, \theta_2^\top)$, with θ_1 the vector of s parameters of interest

$$\mathbf{M}(\theta) = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}, \ \mathbf{M}^{-1}(\theta) = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

with

$$\begin{aligned} \mathbf{A}_{11} &= [\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1} \\ \mathbf{A}_{12} &= -[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1}\mathbf{M}_{12}\mathbf{M}_{22}^{-1} \\ \mathbf{A}_{21} &= -\mathbf{M}_{22}^{-1}\mathbf{M}_{21}[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1} \\ \mathbf{A}_{22} &= \mathbf{M}_{22}^{-1} + \mathbf{M}_{22}^{-1}\mathbf{M}_{21}[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1}\mathbf{M}_{12}\mathbf{M}_{22}^{-1} \end{aligned}$$

$\rightarrow$ maximize $\Phi_{D_s}[\mathbf{M}] = \det[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]$

- D_s-optimality: only $s < p$ parameters of interest (and $p - s$ «nuisance» parameters) $\rightarrow \theta^\top = (\theta_1^\top, \theta_2^\top)$, with θ_1 the vector of s parameters of interest

$$\mathbf{M}(\theta) = \begin{pmatrix} \mathbf{M}_{11} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{pmatrix}, \; \mathbf{M}^{-1}(\theta) = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$$

with

$$\begin{aligned} \mathbf{A}_{11} &= [\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1} \\ \mathbf{A}_{12} &= -[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1}\mathbf{M}_{12}\mathbf{M}_{22}^{-1} \\ \mathbf{A}_{21} &= -\mathbf{M}_{22}^{-1}\mathbf{M}_{21}[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1} \\ \mathbf{A}_{22} &= \mathbf{M}_{22}^{-1} + \mathbf{M}_{22}^{-1}\mathbf{M}_{21}[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]^{-1}\mathbf{M}_{12}\mathbf{M}_{22}^{-1} \end{aligned}$$

$\rightarrow$ maximize $\Phi_{D_s}[\mathbf{M}] = \det[\mathbf{M}_{11} - \mathbf{M}_{12}\mathbf{M}_{22}^{-1}\mathbf{M}_{21}]$

➠ Useful for **model discrimination**:
if $\eta^{(2)}(x, \theta_2) = \eta^{(1)}(x, \theta_1) + \delta(x, \theta_{2\setminus 1})$ (nested models),
estimate $\theta_{2\setminus 1}$ in $\eta^{(2)}$ to decide whether $\eta^{(1)}$ or $\eta^{(2)}$ is more appropriate, see [Atkinson & Cox 1974]

3 Construction of (locally) optimal designs

A/ Exact design

n observations at $X_n = (x_1, \ldots, x_n)$ in a regression model (for simplicity)
Each design point x_i can be anything, e.g. a point in a subset $\mathscr{X}$ of $\mathbb{R}^d$

☞ maximize $\Phi(\mathbf{M}_n)$ w.r.t. X_n with $\mathbf{M}_n = \mathbf{M}(X_n, \theta^0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \eta(x_i, \theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i, \theta)}{\partial \theta^\top}\Big|_{\theta^0}$

3 Construction of (locally) optimal designs

A/ Exact design

n observations at $X_n = (x_1, \ldots, x_n)$ in a regression model (for simplicity)
Each design point x_i can be anything, e.g. a point in a subset $\mathscr{X}$ of $\mathbb{R}^d$

☞ maximize $\Phi(\mathbf{M}_n)$ w.r.t. X_n with $\mathbf{M}_n = \mathbf{M}(X_n, \theta^0) = \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\theta^0}$

• If problem dimension $n \times d$ not too large $\rightarrow$ standard algorithm (but with constraints, local optimas…)

3 Construction of (locally) optimal designs

A/ Exact design

n observations at $X_n = (x_1, \ldots, x_n)$ in a regression model (for simplicity)
Each design point x_i can be anything, e.g. a point in a subset $\mathscr{X}$ of $\mathbb{R}^d$

☞ maximize $\Phi(\mathbf{M}_n)$ w.r.t. X_n with $\mathbf{M}_n = \mathbf{M}(X_n, \theta^0) = \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\theta^0}$

- If problem dimension $n \times d$ not too large → standard algorithm (but with constraints, local optimas. . .)
- Otherwise, use an algorithm that takes the particular form of the problem into account

Exchange methods: at iteration k, exchange **one** *support point* x_j with a better one x^* in $\mathscr{X}$ (design space) — better for $\Phi(\cdot)$

$$X_n^k = (x_1, \ldots, \boxed{x_j \updownarrow x^*}, \ldots, x_n)$$

- [Fedorov 1972]: consider all n possible exchanges successively, each time starting from X_n^k, retain the «best» one among these $n \rightarrow X_n^{k+1}$

$$X_n^k = \begin{pmatrix} x_1 & ,\ldots, & x_j & ,\ldots, & x_n \\ \updownarrow & & \updownarrow & & \updownarrow \\ x_1^* & & x_j^* & & x_n^* \end{pmatrix}$$

- [Fedorov 1972]: consider all n possible exchanges successively, each time starting from X_n^k, retain the «best» one among these $n \to X_n^{k+1}$

$$X_n^k = (\begin{array}{c} x_1 \\ \updownarrow \\ x_1^* \end{array} , \ldots , \begin{array}{c} x_j \\ \updownarrow \\ x_j^* \end{array} , \ldots , \begin{array}{c} x_n \\ \updownarrow \\ x_n^* \end{array})$$

One iteration $\to$ n optimizations of dimension d followed by ranking n criterion values

- [Mitchell, 1974]: DETMAX algorithm
  If one additional observation were allowed $\rightarrow$ optimal choice

$$X_n^{k+} = (x_1, \ldots, x_j, \ldots, x_n, x_{n+1}^*)$$

Then, remove one support point to return to a n-points design:

➔ consider all $n+1$ possible cancellations,
retain the less penalizing in the sense of $\Phi(\cdot)$

- [Mitchell, 1974]: DETMAX algorithm
  If one additional observation were allowed $\rightarrow$ optimal choice

$$X_n^{k+} = (x_1, \ldots, x_j, \ldots, x_n, x_{n+1}^*)$$

Then, remove one support point to return to a n-points design:

➔ consider all $n+1$ possible cancellations,
retain the less penalizing in the sense of $\Phi(\cdot)$

➔ globally, exchange some x_j with x_{n+1}^*
[= excursion of length 1, longer excursions are possible. . .]
One iteration $\rightarrow$ 1 optimization of dimension d followed by ranking $n+1$ criterion values

- DETMAX has simpler iterations than Fedorov, but usually requires more iterations

- DETMAX has simpler iterations than Fedorov, but usually requires more iterations
- dead ends are possible:
  - DETMAX: the point to be removed is x_{n+1}
  - Fedorov: no possible improvement when optimizing **one** x_i at a time

- DETMAX has simpler iterations than Fedorov, but usually requires more iterations
- dead ends are possible:
  - DETMAX: the point to be removed is x_{n+1}
  - Fedorov: no possible improvement when optimizing **one** x_i at a time
- ▲ both give local optima only ▲

- DETMAX has simpler iterations than Fedorov, but usually requires more iterations
- dead ends are possible:
  - DETMAX: the point to be removed is x_{n+1}
  - Fedorov: no possible improvement when optimizing **one** x_i at a time
- ▲ both give local optima only ▲
- Other methods:
  - Branch and bound: guaranteed convergence, but complicated [Welch 1982]
  - Rounding an optimal design measure (support points x_i and associated weights w_i^*, $i = 1, \ldots, m$, presented next in B/):
    choose n integers r_i (r_i= nb. of replications of observations at x_i) such that $\sum_{i=1}^m r_i = n$ and $r_i/n \approx w_i^*$
    (e.g., maximize $\min_{i=1,\ldots,m} r_i/w_i^*$ = Adams apportionment, see [Pukelsheim & Reider 1992])

B/ Design measures: approximate design theory

[Chernoff 1953; Kiefer & Wolfowitz 1960, Fedorov 1972; Silvey 1980, Pázman 1986, Pukelsheim 1993, Fedorov & Leonov 2014...]
(nonlinear) regression, n observations at $X_n = (x_1, \ldots, x_n)$ with i.i.d. errors:

$$\mathbf{M}(X_n, \theta^0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \eta(x_i, \theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i, \theta)}{\partial \theta^\top}\Big|_{\theta^0}$$

(the additive form is essential — related to the independence of observations)

B/ Design measures: approximate design theory

[Chernoff 1953; Kiefer & Wolfowitz 1960, Fedorov 1972; Silvey 1980, Pázman 1986, Pukelsheim 1993, Fedorov & Leonov 2014...]

(nonlinear) regression, n observations at $X_n = (x_1, \ldots, x_n)$ with i.i.d. errors:

$$\mathbf{M}(X_n, \theta^0) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \eta(x_i, \theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i, \theta)}{\partial \theta^\top}\Big|_{\theta^0}$$

(the additive form is essential — related to the independence of observations)

Suppose that several x_i's coincide (replications): only $m < n$ different x_i's

$$\mathbf{M}(X_n, \theta^0) = \sum_{i=1}^{m} \frac{r_i}{n} \frac{\partial \eta(x_i, \theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i, \theta)}{\partial \theta^\top}\Big|_{\theta^0}$$

$\frac{r_i}{n}$ = proportion of observations collected at x_i

= «percentage of experimental effort» at x_i

= weight w_i of support point x_i

$$\mathbf{M}(X_n,\theta^0) = \sum_{i=1}^{m} w_i \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\theta^0}$$

➔ design $X_n \Leftrightarrow \left\{ \begin{array}{ccc} x_1 & \cdots & x_m \\ w_1 & \cdots & w_m \end{array} \right\}$ with $\sum_{i=1}^m w_i = 1$

➔ normalized discrete distribution on the x_i,

with $\boxed{\text{constraints } r_i/n = w_i}$

$$\mathbf{M}(X_n,\theta^0) = \sum_{i=1}^{m} w_i \frac{\partial \eta(x_i,\theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i,\theta)}{\partial \theta^\top}\Big|_{\theta^0}$$

➔ design $X_n \Leftrightarrow \left\{ \begin{array}{ccc} x_1 & \cdots & x_m \\ w_1 & \cdots & w_m \end{array} \right\}$ with $\sum_{i=1}^m w_i = 1$

➔ normalized discrete distribution on the x_i,
with constraints $r_i/n = w_i$

➠ Release the constraints: only enforce $w_i \geq 0$ with $\sum_{i=1}^m w_i = 1$

➔ ξ = discrete probability measure on $\mathscr{X}$ (= design space)
support points x_i and associated weights w_i
= «approximate design»

$$\mathbf{M}(X_n,\theta^0) = \sum_{i=1}^{m} w_i \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\theta^0}$$

➔ design $X_n \Leftrightarrow \left\{ \begin{array}{ccc} x_1 & \cdots & x_m \\ w_1 & \cdots & w_m \end{array} \right\}$ with $\sum_{i=1}^m w_i = 1$

➔ normalized discrete distribution on the x_i,

with constraints $r_i/n = w_i$

➠ Release the constraints: only enforce $w_i \geq 0$ with $\sum_{i=1}^m w_i = 1$

➔ ξ = discrete probability measure on $\mathscr{X}$ (= design space)

support points x_i and associated weights w_i

= «approximate design»

More general expression: ξ = any probability measure on $\mathscr{X}$

$$\mathbf{M}(\xi) = \mathbf{M}(\xi,\theta^0) = \int_{\mathscr{X}} \frac{\partial\eta(x,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x,\theta)}{\partial\theta^\top}\Big|_{\theta^0} \xi(\mathrm{d}x)\,,\ \int_{\mathscr{X}} \xi(\mathrm{d}x) = 1$$

$\mathbf{M}(\xi) \in$ convex closure of $\mathcal{M}$= set of rank 1 matrices

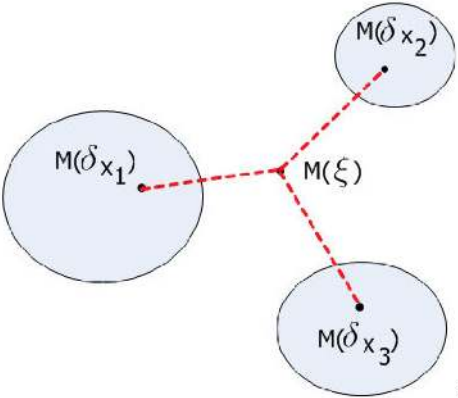
$$\mathbf{M}(\delta_x) = \frac{\partial \eta(x,\theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x,\theta)}{\partial \theta^\top}\Big|_{\theta^0}$$

$\mathbf{M}(\xi)$ is symmetric $p \times p$: $\in$ q-dimensional space, $q = \frac{p(p+1)}{2}$

$\mathbf{M}(\xi) \in$ convex closure of $\mathcal{M}$= set of rank 1 matrices

$$\mathbf{M}(\delta_x) = \frac{\partial \eta(x,\theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x,\theta)}{\partial \theta^\top}\Big|_{\theta^0}$$

$\mathbf{M}(\xi)$ is symmetric $p \times p$: $\in q$-dimensional space, $q = \frac{p(p+1)}{2}$



$\xi = w_1\delta_{x_1} + w_2\delta_{x_2} + w_3\delta_{x_3}$
(3 points are enough for $q = 2$)

Caratheodory Theorem:
$\mathbf{M}(\xi)$ can be written as the linear combination of at most $q+1$ elements of $\mathcal{M}$:

$$\mathbf{M}(\xi) = \sum_{i=1}^{m} w_i \, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\theta^0} \,, \quad m \leq \frac{p(p+1)}{2} + 1$$

$\Rightarrow$ consider discrete probability measures with $\frac{p(p+1)}{2} + 1$ support points at most
(true in particular for the optimum design!)

Caratheodory Theorem:

$\mathbf{M}(\xi)$ can be written as the linear combination of at most $q+1$ elements of $\mathcal{M}$:

$$\mathbf{M}(\xi) = \sum_{i=1}^{m} w_i \, \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}\Big|_{\theta^0} , \quad m \leq \frac{p(p+1)}{2} + 1$$

$\Rightarrow$ consider discrete probability measures with $\frac{p(p+1)}{2}+1$ support points at most (true in particular for the optimum design!)

[Even better: for many criteria $\Phi(\cdot)$, if ξ^* is optimal (maximizes $\Phi[\mathbf{M}(\xi)]$) then $\mathbf{M}(\xi^*)$ is on the boundary of the convex closure of $\mathcal{M}$ and $\frac{p(p+1)}{2}$ support points are enough]

Caratheodory Theorem:

$\mathbf{M}(\xi)$ can be written as the linear combination of at most $q+1$ elements of $\mathcal{M}$:

$$\mathbf{M}(\xi) = \sum_{i=1}^{m} w_i \, \frac{\partial \eta(x_i,\theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i,\theta)}{\partial \theta^\top}\Big|_{\theta^0} \,, \quad m \leq \frac{p(p+1)}{2} + 1$$

$\Rightarrow$ consider discrete probability measures with $\frac{p(p+1)}{2} + 1$ support points at most

(true in particular for the optimum design!)

[Even better: for many criteria $\Phi(\cdot)$, if ξ^* is optimal (maximizes $\Phi[\mathbf{M}(\xi)]$) then $\mathbf{M}(\xi^*)$ is on the boundary of the convex closure of $\mathcal{M}$ and $\frac{p(p+1)}{2}$ support points are enough]

Suppose we found an optimal $\xi^* = \sum_{i=1}^{m} w_i^* \delta_{x_i}$

☞ for a given n, choose the r_i so that $\frac{r_i}{n} \simeq w_i^*$ optimum

$\rightarrow$ *rounding of an approximate design*

Caratheodory Theorem:

$\mathbf{M}(\xi)$ can be written as the linear combination of at most $q+1$ elements of $\mathcal{M}$:

$$\mathbf{M}(\xi) = \sum_{i=1}^{m} w_i \frac{\partial \eta(x_i,\theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i,\theta)}{\partial \theta^\top}\Big|_{\theta^0}, \quad m \leq \frac{p(p+1)}{2} + 1$$

$\Rightarrow$ consider discrete probability measures with $\frac{p(p+1)}{2} + 1$ support points at most
(true in particular for the optimum design!)

[Even better: for many criteria $\Phi(\cdot)$, if ξ^* is optimal (maximizes $\Phi[\mathbf{M}(\xi)]$) then $\mathbf{M}(\xi^*)$ is on the boundary of the convex closure of $\mathcal{M}$ and $\frac{p(p+1)}{2}$ support points are enough]

Suppose we found an optimal $\xi^* = \sum_{i=1}^{m} w_i^* \delta_{x_i}$

☞ for a given n, choose the r_i so that $\frac{r_i}{n} \simeq w_i^*$ optimum
$\rightarrow$ *rounding of an approximate design*

☞ Sometimes, ξ^* can be implemented *without any approximation*: $\xi =$ *power spectral density* of an input signal
$\rightarrow$ design of optimal input for ODE model in the frequency domain

Caratheodory Theorem:

$\mathbf{M}(\xi)$ can be written as the linear combination of at most $q+1$ elements of $\mathcal{M}$:

$$\mathbf{M}(\xi) = \sum_{i=1}^{m} w_i \, \frac{\partial \eta(x_i,\theta)}{\partial \theta}\Big|_{\theta^0} \frac{\partial \eta(x_i,\theta)}{\partial \theta^\top}\Big|_{\theta^0} , \quad m \leq \frac{p(p+1)}{2} + 1$$

$\Rightarrow$ consider discrete probability measures with $\frac{p(p+1)}{2}+1$ support points at most
(true in particular for the optimum design!)

[Even better: for many criteria $\Phi(\cdot)$, if ξ^* is optimal (maximizes $\Phi[\mathbf{M}(\xi)]$) then $\mathbf{M}(\xi^*)$ is on the boundary of the convex closure of $\mathcal{M}$ and $\frac{p(p+1)}{2}$ support points are enough]

Suppose we found an optimal $\xi^* = \sum_{i=1}^{m} w_i^* \delta_{x_i}$

☞ for a given n, choose the r_i so that $\frac{r_i}{n} \simeq w_i^*$ optimum
$\rightarrow$ *rounding of an approximate design*

☞ Sometimes, ξ^* can be implemented *without any approximation*: ξ = *power spectral density* of an input signal
$\rightarrow$ design of optimal input for ODE model in the frequency domain

Why design measures are interesting?
How does it simplify the optimization problem?

C/ Optimal design measures

➠ Maximize $\Phi(\cdot)$ concave w.r.t. $\mathbf{M}(\xi)$ in a convex set

Ex: D-optimality: $\forall\ \mathbf{M}_1 \succ \mathbf{O}$, $\mathbf{M}_2 \succeq \mathbf{O}$, with $\mathbf{M}_2 \not\propto \mathbf{M}_2$, $\forall \alpha$, $0 < \alpha < 1$,

$$\log\det[(1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2] > (1-\alpha)\log\det\mathbf{M}_1 + \alpha\log\det\mathbf{M}_2$$

$\Rightarrow \log\det[\cdot]$ is (strictly) concave

convex set + concave criterion $\Rightarrow$ one unique optimum!

C/ Optimal design measures

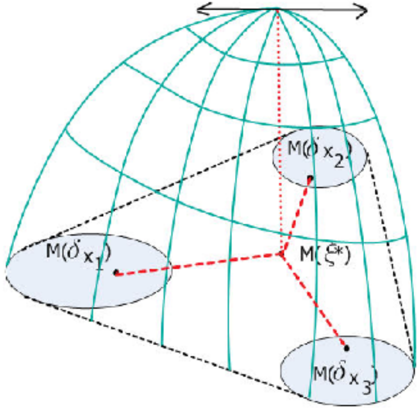
➠ Maximize $\Phi(\cdot)$ concave w.r.t. $\mathbf{M}(\xi)$ in a convex set

Ex: *D*-optimality: $\forall\, \mathbf{M}_1 \succ \mathbf{O}$, $\mathbf{M}_2 \succeq \mathbf{O}$, with $\mathbf{M}_2 \not\propto \mathbf{M}_2$, $\forall \alpha$, $0 < \alpha < 1$,

$\log\det[(1-\alpha)\mathbf{M}_1 + \alpha\mathbf{M}_2] > (1-\alpha)\log\det \mathbf{M}_1 + \alpha \log\det \mathbf{M}_2$

$\Rightarrow \log\det[\cdot]$ is (strictly) concave

convex set + concave criterion $\Rightarrow$ one unique optimum!



ξ^* is optimal $\Leftrightarrow$ directional derivative ≤ 0 in all directions

$\Longrightarrow$ "Equivalence Theorem" [Kiefer & Wolfowitz 1960]

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave, $\phi(\xi) = \Phi[\mathbf{M}(\xi)]$

$F_\phi(\xi;\nu) = \lim_{\alpha\to 0^+} \frac{\phi[(1-\alpha)\xi+\alpha\nu]-\phi(\xi)}{\alpha}$
= directional derivative of $\phi(\cdot)$ at ξ in direction ν

Equivalence Theorem: $\boxed{\xi^* \text{ maximizes } \phi(\xi) \Leftrightarrow \max_{\nu\in\Xi} F_\phi(\xi^*;\nu) \leq 0}$

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave, $\phi(\xi) = \Phi[\mathbf{M}(\xi)]$

$F_\phi(\xi;\nu) = \lim_{\alpha\to 0^+} \frac{\phi[(1-\alpha)\xi+\alpha\nu]-\phi(\xi)}{\alpha}$
$\quad$ = directional derivative of $\phi(\cdot)$ at ξ in direction ν

Equivalence Theorem: $\boxed{\xi^* \text{ maximizes } \phi(\xi) \Leftrightarrow \max_{\nu\in\Xi} F_\phi(\xi^*;\nu) \leq 0}$

➔ Takes a simple form when $\Phi(\cdot)$ is differentiable

$$\boxed{\xi^* \text{ maximizes } \phi(\xi) \Leftrightarrow \max_{x\in\mathscr{X}} F_\phi(\xi^*;\delta_x) \leq 0}$$

☞ Check optimality of ξ^* by plotting $F_\phi(\xi^*;\delta_x)$

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave, $\phi(\xi) = \Phi[\mathbf{M}(\xi)]$

$F_\phi(\xi;\nu) = \lim_{\alpha\to 0^+} \frac{\phi[(1-\alpha)\xi+\alpha\nu]-\phi(\xi)}{\alpha}$

$=$ directional derivative of $\phi(\cdot)$ at ξ in direction ν

Equivalence Theorem: $\boxed{\xi^* \text{ maximizes } \phi(\xi) \Leftrightarrow \max_{\nu\in\Xi} F_\phi(\xi^*;\nu) \leq 0}$

➔ Takes a simple form when $\Phi(\cdot)$ is differentiable

$$\boxed{\xi^* \text{ maximizes } \phi(\xi) \Leftrightarrow \max_{x\in\mathscr{X}} F_\phi(\xi^*;\delta_x) \leq 0}$$

☞ Check optimality of ξ^* by plotting $F_\phi(\xi^*;\delta_x)$

Ex: D-optimal design

- ξ_D^* maximizes $\log\det[\mathbf{M}(\xi)]$ w.r.t. $\xi \in \Xi$
- $\Leftrightarrow \max_{x\in\mathscr{X}} d(\xi_D^*, x) \leq p$
- $\Leftrightarrow \xi_D^*$ minimizes $\max_{x\in\mathscr{X}} d(\xi, x)$ w.r.t. $\xi \in \Xi$

where $d(\xi, x) = \frac{\partial\eta(x,\theta)}{\partial\theta^\top}\Big|_{\theta^0} \mathbf{M}^{-1}(\xi) \frac{\partial\eta(x,\theta)}{\partial\theta}\Big|_{\theta^0}$

Moreover, $d(\xi_D^*, x_i) = p = \dim(\theta)$ for any x_i = support point of ξ_D^*

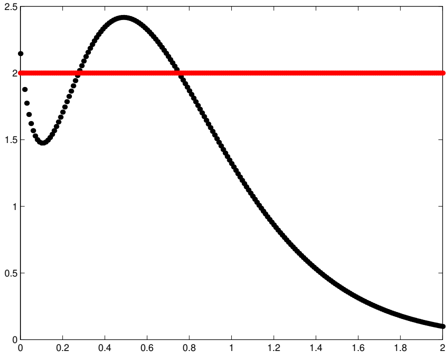
Ex: $\eta(x,\theta) = \theta_1 \exp(-\theta_2 x)$ ($p = 2$) i.i.d. errors, $\mathscr{X} = \mathbb{R}^+$
$[\theta_2^0 = 2]$

$\rightarrow d(\xi, x)$ as a function of x

Ex: $\eta(x,\theta) = \theta_1 \exp(-\theta_2 x)$ ($p=2$) i.i.d. errors, $\mathscr{X} = \mathbb{R}^+$
$[\theta_2^0 = 2]$

$\rightarrow d(\xi, x)$ as a function of x

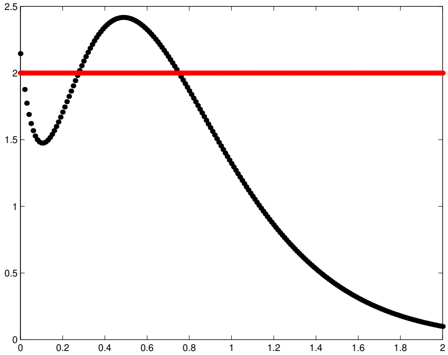
$$\xi_2 = \left\{ \begin{array}{cc} 0.01 & 0.75 \\ 1/2 & 1/2 \end{array} \right\}$$

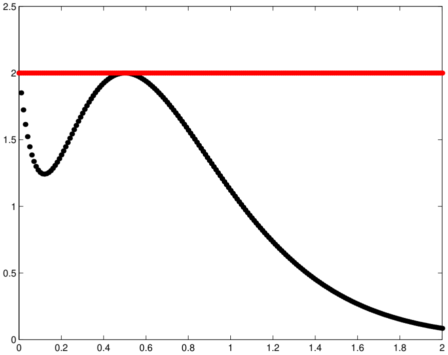


Ex: $\eta(x,\theta) = \theta_1 \exp(-\theta_2 x)$ ($p = 2$) i.i.d. errors, $\mathscr{X} = \mathbb{R}^+$
$[\theta_2^0 = 2]$

$\rightarrow d(\xi, x)$ as a function of x

$$\xi_2 = \left\{ \begin{array}{cc} 0.01 & 0.75 \\ 1/2 & 1/2 \end{array} \right\} \qquad \xi_D^* = \left\{ \begin{array}{cc} 0 & 1/\theta_2 = 0.5 \\ 1/2 & 1/2 \end{array} \right\}$$





KW Eq. Th. relates optimality in θ space to optimality in y space (i.i.d. errors)

$$n\,\text{var}[\eta(x,\hat{\theta}^n)] \to \sigma^2 \, \frac{\partial\eta(x,\theta)}{\partial\theta^\top}\Big|_{\bar{\theta}} \mathbf{M}^{-1}(\xi,\bar{\theta}) \frac{\partial\eta(x,\theta)}{\partial\theta}\Big|_{\bar{\theta}} = \sigma^2 \, d(\xi,x)\Big|_{\bar{\theta}},\ n\to\infty$$

D-optimality $\Leftrightarrow$ G-optimality

➟ ξ_D^* minimizes the maximum value of prediction variance over $\mathscr{X}$

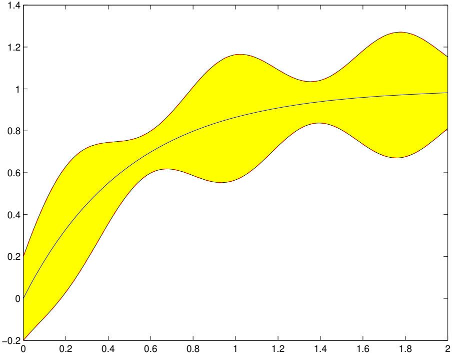
KW Eq. Th. relates optimality in θ space to optimality in y space (i.i.d. errors)

$$n\,\mathrm{var}[\eta(x,\hat{\theta}^n)] \to \sigma^2 \left.\frac{\partial\eta(x,\theta)}{\partial\theta^\top}\right|_{\bar{\theta}} \mathbf{M}^{-1}(\xi,\bar{\theta}) \left.\frac{\partial\eta(x,\theta)}{\partial\theta}\right|_{\bar{\theta}} = \sigma^2\, d(\xi,x)\big|_{\bar{\theta}},\ n\to\infty$$

D-optimality $\Leftrightarrow$ G-optimality

➠ ξ_D^* minimizes the maximum value of prediction variance over $\mathscr{X}$

$\eta(x,\bar{\theta})$, $\eta(x,\bar{\theta}) \pm$ 2 st.d.



$\to$ put next observation where $d(\xi,x)$ is large

Remark:

Eq. Th. $=$ stationarity condition $=$ NS condition for optimality

$\neq$ duality property!

Remark:

Eq. Th. $=$ stationarity condition $=$ NS condition for optimality
$\neq$ duality property!

Dual problem to D-optimum design:

Define $\mathcal{S} = \left\{ \frac{\partial \eta(x,\theta)}{\partial \theta}\Big|_{\theta^0} ,\ x \in \mathscr{X} \right\}$ [$\mathcal{S} \cup -\mathcal{S} =$ Elfving's set]

$\mathcal{E}^* =$ minimum-volume ellipsoid centered at $\mathbf{0}$ that contains $\mathcal{S}$

Lagrangian theory $\Rightarrow \mathcal{E}^* = \{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{M}_F^{-1}(\xi_D^*)\mathbf{z} \leq p\}$ where ξ_D^* is D-optimum
support points of $\xi_D^* =$ contact between $\mathcal{E}^*$ and $\mathcal{S}$

Remark:

Eq. Th. $=$ stationarity condition $=$ NS condition for optimality
$\neq$ duality property!

Dual problem to *D*-optimum design:

Define $\mathcal{S} = \left\{ \frac{\partial\eta(x,\theta)}{\partial\theta}\Big|_{\theta^0} ,\ x \in \mathscr{X} \right\}$ [$\mathcal{S} \cup -\mathcal{S} =$ Elfving's set]

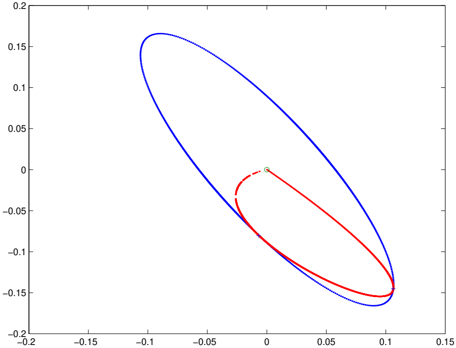
$\mathcal{E}^* =$ minimum-volume ellipsoid centered at $\mathbf{0}$ that contains $\mathcal{S}$

Lagrangian theory $\Rightarrow \mathcal{E}^* = \{\mathbf{z} \in \mathbb{R}^p : \mathbf{z}^\top \mathbf{M}_F^{-1}(\xi_D^*)\mathbf{z} \leq p\}$ where ξ_D^* is D-optimum
support points of $\xi_D^* =$ contact between $\mathcal{E}^*$ and $\mathcal{S}$

In general, few contact points $\rightarrow$ repeat observations at the same place (see [Yang 2010, Dette & Melas 2011])

There exist dual problems for other criteria $\Phi(\cdot)$
($=$ one of the main topics in [Pukelsheim 1993])

Ex: $\eta(x,\theta) = \frac{\theta_1}{\theta_1-\theta_2}[\exp(-\theta_2 x) - \exp(-\theta_1 x)]$
$\theta = (1,5)$, $\mathscr{X} = \mathbb{R}^+$



$\mathcal{E}^*$ and $\mathcal{S}$

$\Rightarrow$ D optimum design ξ_D^* supported on two points

D/ Construction of an optimal design measure

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave and differentiable,
$\phi(\xi) = \Phi[\mathbf{M}(\xi)]$
Concavity $\Longrightarrow$ $\boxed{\text{for any } \xi \in \Xi,\ \phi(\xi^*) \leq \phi(\xi) + \max_{x\in\mathscr{X}} F_\phi(\xi;\delta_x)}$

D/ Construction of an optimal design measure

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave and differentiable,
$\phi(\xi) = \Phi[\mathbf{M}(\xi)]$
Concavity $\Longrightarrow$ for any $\xi \in \Xi$, $\phi(\xi^*) \leq \phi(\xi) + \max_{x\in\mathscr{X}} F_\phi(\xi;\delta_x)$

Fedorov–Wynn Algorithm: sort of steepest ascent

- 1 : Choose ξ^1 not degenerate ($\det \mathbf{M}(\xi^1) > 0$)
- 2 : Compute $x_k^* = \arg\max_{\mathscr{X}} F_\phi(\xi^k;\delta_x)$
  If $F_\phi(\xi^k;\delta_{x_k^+}) < \epsilon$, stop: ξ^k is ϵ-optimal
- 3 : $\xi^{k+1} = (1-\alpha_k)\xi^k + \alpha_k\delta_{x_k^*}$ (delta measure at x_k^*)
  [*Vertex Direction*]
  $k \to k+1$, return to Step 2

D/ Construction of an optimal design measure

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave and differentiable, $\phi(\xi) = \Phi[\mathbf{M}(\xi)]$

Concavity $\Longrightarrow$ for any $\xi \in \Xi$, $\phi(\xi^*) \leq \phi(\xi) + \max_{x\in\mathscr{X}} F_\phi(\xi; \delta_x)$

Fedorov–Wynn Algorithm: sort of steepest ascent

- 1 : Choose ξ^1 not degenerate ($\det \mathbf{M}(\xi^1) > 0$)
- 2 : Compute $x_k^* = \arg\max_{\mathscr{X}} F_\phi(\xi^k; \delta_x)$
  If $F_\phi(\xi^k; \delta_{x_k^+}) < \epsilon$, stop: ξ^k is ϵ-optimal
- 3 : $\xi^{k+1} = (1 - \alpha_k)\xi^k + \alpha_k \delta_{x_k^*}$ (delta measure at x_k^*)
  [Vertex Direction]
  $k \to k + 1$, return to Step 2

Step-size α_k?

➠ $\alpha_k = \arg\max \phi(\xi^{k+1})$ [$= \frac{d(\xi^k, x_k^*) - p}{p[d(\xi^k, x_k^*) - 1]}$ for D-optimal design [Fedorov 1972]]

➔ monotone convergence

D/ Construction of an optimal design measure

Ξ = set of probability measures on $\mathscr{X}$, $\Phi(\cdot)$ concave and differentiable, $\phi(\xi) = \Phi[\mathbf{M}(\xi)]$

Concavity $\Longrightarrow$ for any $\xi \in \Xi$, $\phi(\xi^*) \leq \phi(\xi) + \max_{x\in\mathscr{X}} F_\phi(\xi;\delta_x)$

Fedorov–Wynn Algorithm: sort of steepest ascent

- 1 : Choose ξ^1 not degenerate ($\det \mathbf{M}(\xi^1) > 0$)
- 2 : Compute $x_k^* = \arg\max_{\mathscr{X}} F_\phi(\xi^k;\delta_x)$
  If $F_\phi(\xi^k;\delta_{x_k^+}) < \epsilon$, stop: ξ^k is ϵ-optimal
- 3 : $\xi^{k+1} = (1-\alpha_k)\xi^k + \alpha_k\delta_{x_k^*}$ (delta measure at x_k^*)
  [*Vertex Direction*]
  $k \to k+1$, return to Step 2

Step-size α_k?

➠ $\alpha_k = \arg\max \phi(\xi^{k+1})$ [$= \frac{d(\xi^k,x_k^*)-p}{p[d(\xi^k,x_k^*)-1]}$ for D-optimal design [Fedorov 1972]]
➔ monotone convergence

➠ $\alpha_k > 0$, $\lim_{k\to\infty}\alpha_k = 0$, $\sum_{i=1}^{\infty}\alpha_k = \infty$ [[Wynn 1970] for D-optimal design]

Remarks:

- Consider sequential design, one x_i at a time enters $\mathbf{M}(X)$
$$\mathbf{M}(X_{k+1}) = \frac{k}{k+1}\,\mathbf{M}(X_k) + \frac{1}{k+1}\,\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta}\Big|_{\theta^0}\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta^\top}\Big|_{\theta^0}$$
with $x_{k+1} = \arg\max_{\mathscr{X}} F_\phi(\xi^k;\delta_x)$
$\Leftrightarrow$ Wynn algorithm with $\alpha_k = \frac{1}{k+1}$

Remarks:

- Consider sequential design, one x_i at a time enters $\mathbf{M}(X)$
  $$\mathbf{M}(X_{k+1}) = \frac{k}{k+1}\,\mathbf{M}(X_k) + \frac{1}{k+1}\,\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta}\Big|_{\theta^0}\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta^\top}\Big|_{\theta^0}$$
  with $x_{k+1} = \arg\max_{\mathscr{X}} F_\phi(\xi^k;\delta_x)$
  $\Leftrightarrow$ Wynn algorithm with $\alpha_k = \frac{1}{k+1}$
- **Guaranteed convergence to the optimum**

Remarks:

- Consider sequential design, one x_i at a time enters $\mathbf{M}(X)$

$$\mathbf{M}(X_{k+1}) = \frac{k}{k+1}\,\mathbf{M}(X_k) + \frac{1}{k+1}\,\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta}\Big|_{\theta^0}\,\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta^\top}\Big|_{\theta^0}$$

  with $x_{k+1} = \arg\max_{\mathscr{X}} F_\phi(\xi^k;\delta_x)$

  $\Leftrightarrow$ Wynn algorithm with $\alpha_k = \frac{1}{k+1}$

- **Guaranteed convergence to the optimum**
- There exist faster methods:
  - remove support points from ξ^k ($\approx$ allow α_k to be < 0) [Atwood 1973; Böhning 1985, 1986]
  - combine with gradient projection (or a second-order method) [Wu 1978]
  - use a multiplicative algorithm [Titterington 1976; Torsney 1983–2009; Yu 2010] [for D or A optimal design, far from the optimum]
  - combine different methods [Yu 2011]
  - Still an active topic. . .

Remarks: Usually, $\mathscr{X}$ = compact subset of $\mathbb{R}^d$ (e.g., the probability simplex for mixture experiments)

$\rightarrow$ discretized into $\mathscr{X}_\ell$ with ℓ elements (a grid — or better, a low-discrepancy sequence, see [Niederreiter 1992])

➔ the algorithms above may be slow when ℓ is large

Remarks: Usually, $\mathscr{X}$ = compact subset of $\mathbb{R}^d$ (e.g., the probability simplex for mixture experiments)

$\rightarrow$ discretized into $\mathscr{X}_\ell$ with ℓ elements (a grid — or better, a low-discrepancy sequence, see [Niederreiter 1992])

➔ the algorithms above may be slow when ℓ is large

➠ Combine continuous search for support points in $\mathscr{X}$ with optimization of a design measure with few support points, say $m \ll \ell$

➠ Exploit guaranteed (and fast) convergence of algorithms for m small
+ use Eq. Th. to check optimality [Yang *et al.*, 2013, P & Zhigljavsky, 2014]

Ex: *D*-optimal design for

$$\eta(x,\theta) = \theta_0 + \theta_1 \exp(-\theta_2 x_1) + \frac{\theta_3}{\theta_3 - \theta_4} [\exp(-\theta_4 x_2) - \exp(-\theta_3 x_2)]$$

with $x = (x_1, x_2) \in \mathscr{X} = [0,2] \times [0,10]$ (and $p = 5$, $\theta_2^0 = 2$, $\theta_3^0 = 0.7$, $\theta_4^0 = 0.2$)

Ex: D-optimal design for

$$\eta(x,\theta) = \theta_0 + \theta_1 \exp(-\theta_2 x_1) + \frac{\theta_3}{\theta_3 - \theta_4} [\exp(-\theta_4 x_2) - \exp(-\theta_3 x_2)]$$

with $x = (x_1, x_2) \in \mathscr{X} = [0,2] \times [0,10]$ (and $p = 5$, $\theta_2^0 = 2$, $\theta_3^0 = 0.7$, $\theta_4^0 = 0.2$)

Additive model [Schwabe 1995]: ξ_D^* = tensor product of optimal designs for

$$\beta_0^{(1)} + \beta_1^{(1)} \exp(-\beta_2^{(1)} x_1)$$

and

$$\beta_0^{(2)} + \beta_1^{(2)} [\exp(-\beta_2^{(2)} x_2) - \exp(-\beta_1^{(2)} x_2)]/(\beta_1^{(2)} - \beta_2^{(2)})$$

Ex: *D*-optimal design for

$$\eta(x,\theta) = \theta_0 + \theta_1 \exp(-\theta_2 x_1) + \frac{\theta_3}{\theta_3 - \theta_4} [\exp(-\theta_4 x_2) - \exp(-\theta_3 x_2)]$$

with $x = (x_1, x_2) \in \mathscr{X} = [0,2] \times [0,10]$ (and $p = 5$, $\theta_2^0 = 2$, $\theta_3^0 = 0.7$, $\theta_4^0 = 0.2$)

Additive model [Schwabe 1995]: ξ_D^* = tensor product of optimal designs for

$$\beta_0^{(1)} + \beta_1^{(1)} \exp(-\beta_2^{(1)} x_1)$$

and

$$\beta_0^{(2)} + \beta_1^{(2)} [\exp(-\beta_2^{(2)} x_2) - \exp(-\beta_1^{(2)} x_2)]/(\beta_1^{(2)} - \beta_2^{(2)})$$

Use the Equivalence Th. to construct ξ_D^* (with arbitrary precision — Maple)
[weight 1/9 at (0, 0.46268527927, 2) $\otimes$ (0, 1.22947139883, 6.85768905493)]

➡ 7 iterations of the algorithm in [P & Zhigljavsky, 2014] yield ξ such that $\max_{x \in \mathscr{X}} F_\phi(\xi; \delta_x) < 10^{-5}$

What if $\Phi(\cdot)$ not differentiable? (e.g., maximize $\Phi(\mathbf{M}) = \lambda_{\min}(\mathbf{M})$)

$\Phi(\cdot)$ concave, $\mathscr{X}$ discretized into $\mathscr{X}_\ell$, ℓ not too large

➔ optimal design $\Longleftrightarrow$ optimal vector of weights $\mathbf{w} \in \mathbb{R}^\ell$

$w_i \geq 0$, $\sum_{i=1}^{\ell} w_i = 1$

What if $\Phi(\cdot)$ not differentiable? (e.g., maximize $\Phi(\mathbf{M}) = \lambda_{\min}(\mathbf{M})$)
$\Phi(\cdot)$ concave, $\mathscr{X}$ discretized into $\mathscr{X}_\ell$, ℓ not too large
➔ optimal design $\Longleftrightarrow$ optimal vector of weights $\mathbf{w} \in \mathbb{R}^\ell$
$w_i \geq 0$, $\sum_{i=1}^{\ell} w_i = 1$

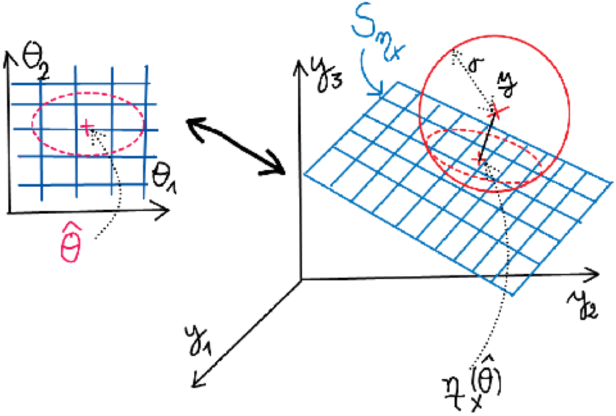
- subgradients ($\leftrightarrow$ directional derivatives)

What if $\Phi(\cdot)$ not differentiable? (e.g., maximize $\Phi(\mathbf{M}) = \lambda_{\min}(\mathbf{M})$)
$\Phi(\cdot)$ concave, $\mathscr{X}$ discretized into $\mathscr{X}_\ell$, ℓ not too large
➔ optimal design $\Longleftrightarrow$ optimal vector of weights $\mathbf{w} \in \mathbb{R}^\ell$
$w_i \geq 0$, $\sum_{i=1}^{\ell} w_i = 1$

- subgradients ($\leftrightarrow$ directional derivatives)
- general method for non-differentiable optimization (cutting plane method [Kelley 1960], level method [Nesterov 2004]), see Chap. 9 of [P & Pázman 2013]

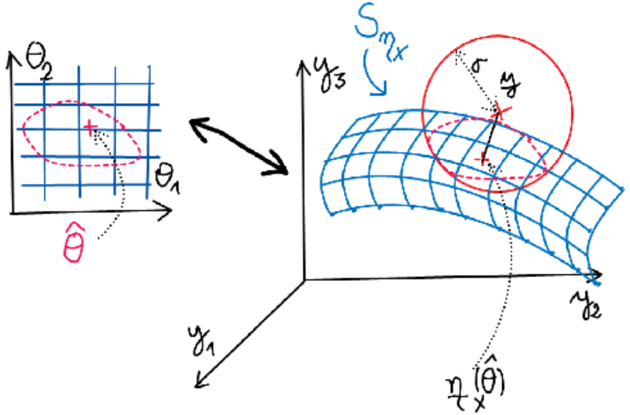
4 Problems with nonlinear models

Linear regression: easy



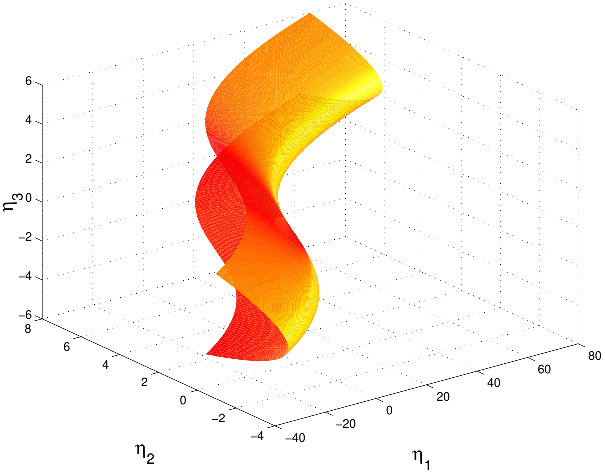
The expectation surface $\mathbb{S}_\eta = \{\eta(\theta) = (\eta(x_1,\theta),\ldots,\eta(x_n,\theta))^\top : \theta \in \mathbb{R}^p\}$ is flat and linearly parameterized

Nonlinear regression: may be a bit tricky



$\mathbb{S}_\eta$ is curved (intrinsic curvature) and nonlinearly parameterized (parametric curvature) [Bates & Watts 1980]

Ex: $\eta(\mathbf{x},\theta) = \theta_1\{\mathbf{x}\}_1 + \theta_1^3(1-\{\mathbf{x}\}_1) + \theta_2\{\mathbf{x}\}_2 + \theta_2^2(1-\{\mathbf{x}\}_2)$
$X = (\mathbf{x}_1,\mathbf{x}_2,\mathbf{x}_3)$, $\mathbf{x}_1 = (0\ 1)$, $\mathbf{x}_2 = (1\ 0)$, $\mathbf{x}_3 = (1\ 1)$, $\theta \in [-3,4]\times[-2,2]$



Two important difficulties:

❶ Asymptotically ($n \to \infty$) — or if σ^2 small enough — all seems fine
(use linear approximations),

but the distribution of $\hat{\theta}^n$ may be far from normal for small n (or for σ^2 large)
➠ small-sample properties

Two important difficulties:

❶ Asymptotically ($n \to \infty$) — or if σ^2 small enough — all seems fine (use linear approximations),

but the distribution of $\hat{\theta}^n$ may be far from normal for small n (or for σ^2 large)

➠ small-sample properties

❷ Everything is local (depends on θ): if we linearize, where do we linearize? (choice of a nominal value θ^0)

➠ nonlocal optimum design

5 Small-sample properties

A/ A classification of regression models

Suppose that

$$y_i = y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i \text{ with } \mathsf{E}\{\varepsilon_i\} = 0 \text{ and } \mathsf{E}\{\varepsilon_i^2\} = \sigma^2(x_i) \text{ for all } i$$

Divide y_i and $\eta(x_i, \bar{\theta})$ by $\sigma(x_i)$ ➔ one may suppose that $\sigma^2(x) = \sigma^2$ for all x

Denote

$$\mathbf{y} = (y_1, \ldots, y_n)^\top \text{ and } \eta(\theta) = (\eta(x_i, \theta), \ldots, \eta(x_n, \theta))^\top$$

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \text{ so that } \mathsf{E}\{\varepsilon\} = \mathbf{0} \text{ and } \mathsf{Var}(\varepsilon) = \sigma^2 \, \mathbf{I}_n$$

We suppose $\eta(x, \theta)$ twice continuously differentiable w.r.t. θ for any x

5 Small-sample properties

A/ A classification of regression models

Suppose that

$$y_i = y(x_i) = \eta(x_i, \bar{\theta}) + \varepsilon_i \text{ with } \mathrm{E}\{\varepsilon_i\} = 0 \text{ and } \mathrm{E}\{\varepsilon_i^2\} = \sigma^2(x_i) \text{ for all } i$$

Divide y_i and $\eta(x_i, \bar{\theta})$ by $\sigma(x_i)$ ➔ one may suppose that $\sigma^2(x) = \sigma^2$ for all x

Denote

$$\mathbf{y} = (y_1, \ldots, y_n)^\top \text{ and } \eta(\theta) = (\eta(x_i, \theta), \ldots, \eta(x_n, \theta))^\top$$

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^\top \text{ so that } \mathrm{E}\{\varepsilon\} = \mathbf{0} \text{ and } \mathrm{Var}(\varepsilon) = \sigma^2\, \mathbf{I}_n$$

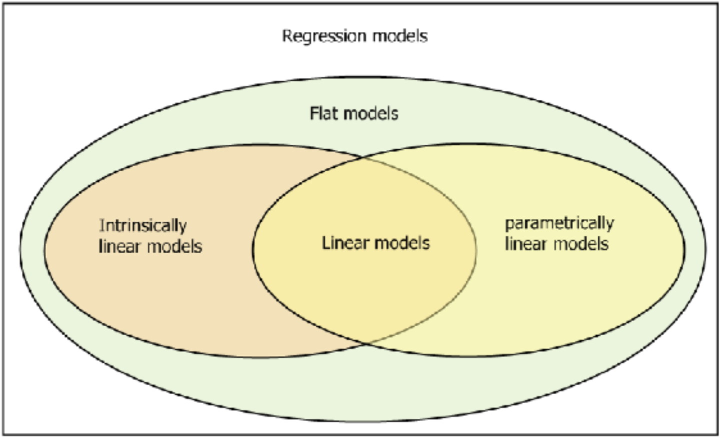
We suppose $\eta(x, \theta)$ twice continuously differentiable w.r.t. θ for any x

- Expectation surface: $\mathbb{S}_\eta = \{\eta(\theta) : \theta \in \mathbb{R}^p\}$
- Orthogonal projector onto the tangent space to $\mathbb{S}_\eta$ at $\eta(\theta)$:

$$\mathbf{P}_\theta = \frac{1}{n} \frac{\partial \eta(\theta)}{\partial \theta^\top} \mathbf{M}^{-1}(X, \theta) \frac{\partial \eta(\theta)}{\partial \theta} \text{ (a } n \times n \text{ matrix)}$$

(both depend on X)

A classification of regression models [Pázman 1993]



Intrinsically linear models

➤ The expectation surface $\mathbb{S}_\eta = \{\eta(\theta) : \theta \in \mathbb{R}^p\}$ is flat (plane) — intrinsic curvature $\equiv 0$

➤ A reparameterization (continuously differentiable) exists that makes the model linear

➤ $\mathbf{P}_\theta \mathbf{H}_{\cdot ij}(\theta) = \mathbf{H}_{\cdot ij}(\theta)$, where $\mathbf{H}_{\cdot ij}(\theta) = \frac{\partial^2 \eta(\theta)}{\partial \theta_i \partial \theta_j}$

Intrinsically linear models

➤ The expectation surface $\mathbb{S}_\eta = \{\eta(\theta) : \theta \in \mathbb{R}^p\}$ is flat (plane) — intrinsic curvature $\equiv 0$

➤ A reparameterization (continuously differentiable) exists that makes the model linear

➤ $\mathbf{P}_\theta \mathbf{H}_{ij}^{\cdot}(\theta) = \mathbf{H}_{ij}^{\cdot}(\theta)$, where $\mathbf{H}_{ij}^{\cdot}(\theta) = \frac{\partial^2 \eta(\theta)}{\partial\theta_i \partial\theta_j}$

Observing at p different x_i only (replications) makes the model intrinsically linear

Parametrically linear models

- $\mathbf{M}(X,\theta) = \text{constant}$
- $\mathbf{P}_\theta \mathbf{H}_{\boldsymbol{\cdot}ij}(\theta) = \mathbf{0}$ — parametric curvature $\equiv 0$

Parametrically linear models

- $\mathbf{M}(X, \theta) =$ constant
- $\mathbf{P}_\theta \mathbf{H}_{\cdot ij}(\theta) = \mathbf{0}$ — parametric curvature $\equiv 0$

Linear models

- $\eta(x, \theta) = \mathbf{f}^\top(x)\theta + c(x)$
- the model is intrinsically and parametrically linear

Parametrically linear models

- $\mathbf{M}(X,\theta)$ = constant
- $\mathbf{P}_\theta \mathbf{H}_{ij}^{\cdot}(\theta) = \mathbf{0}$ — parametric curvature $\equiv 0$

Linear models

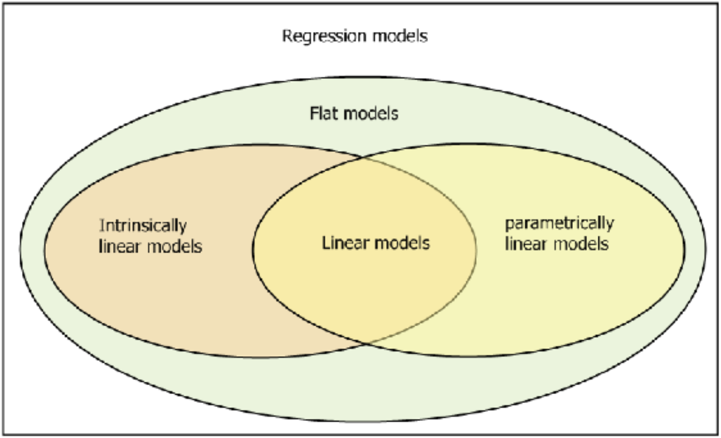
- $\eta(x,\theta) = \mathbf{f}^\top(x)\theta + c(x)$
- the model is intrinsically and parametrically linear

Flat models

- A reparameterization exists that makes the information matrix constant
- Riemannian curvature tensor $\equiv 0$ $R_{hijk}(\theta) = T_{hjik}(\theta) - T_{hkij}(\theta) \equiv 0$ where $T_{hjik}(\theta) = [\mathbf{H}_{hj}^{\cdot}(\theta)]^\top [\mathbf{I}_n - \mathbf{P}_\theta]\mathbf{H}_{ik}^{\cdot}(\theta)$

If all parameters but one appear linearly, then the model is flat

A classification of regression models [Pázman 1993] (bis)



B/ Density of the LS estimator

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

Intrinsically linear models (in particular, repetitions at p points):

$\rightarrow$ exact distribution $\boxed{\hat{\theta}^n \sim q(\theta|\bar{\theta}) = \frac{n^{p/2} \det^{1/2} \mathbf{M}(X,\theta)}{(2\pi)^{p/2} \sigma^p} \exp\left\{-\frac{1}{2\sigma^2} \|\eta(\theta) - \eta(\bar{\theta})\|^2\right\}}$

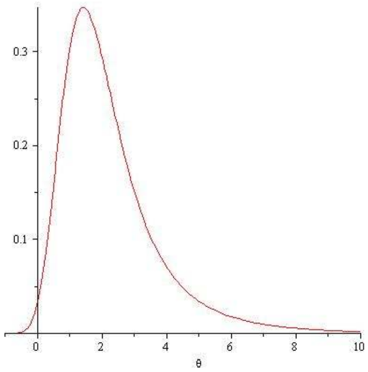
B/ Density of the LS estimator

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

Intrinsically linear models (in particular, repetitions at p points):

$\rightarrow$ exact distribution $\hat{\theta}^n \sim q(\theta|\bar{\theta}) = \frac{n^{p/2} \det^{1/2} \mathbf{M}(X,\theta)}{(2\pi)^{p/2} \sigma^p} \exp\left\{-\frac{1}{2\sigma^2} \|\eta(\theta) - \eta(\bar{\theta})\|^2\right\}$

Ex: $\eta(x,\theta) = \exp(-\theta x)$, $\bar{\theta} = 2$, 15 observations at the same $x = 1/2$ ($\sigma^2 = 1$)



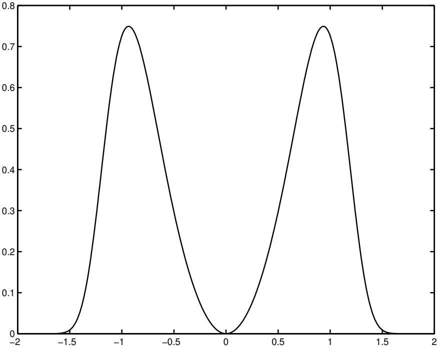
B/ Density of the LS estimator

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

Intrinsically linear models (in particular, repetitions at p points):

$\to$ exact distribution $\boxed{\hat{\theta}^n \sim q(\theta|\bar{\theta}) = \frac{n^{p/2} \det^{1/2} \mathbf{M}(X,\theta)}{(2\pi)^{p/2}\, \sigma^p} \exp\left\{-\frac{1}{2\sigma^2} \|\eta(\theta) - \eta(\bar{\theta})\|^2\right\}}$

Ex: $\eta(x, \theta) = x\,\theta^3$, $\bar{\theta} = 0$, all observations at the same $x \neq 0$



Flat models: approximate density of $\hat{\theta}^n$

$$q(\theta|\bar{\theta}) = \frac{\det[\mathbf{Q}(\theta,\bar{\theta})]}{(2\pi)^{p/2}\,\sigma^p\, n^{p/2}\, \det^{1/2}\mathbf{M}(X,\theta)} \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{P}_\theta[\eta(\theta)-\eta(\bar{\theta})]\|^2\right\}$$

where $\{\mathbf{Q}(\theta,\bar{\theta})\}_{ij} = \{n\,\mathbf{M}(X,\theta)\}_{ij} + [\eta(\theta)-\eta(\bar{\theta})]^\top[\mathbf{I}_n - \mathbf{P}_\theta]\mathbf{H}_{:ij}(\theta)$

Flat models: approximate density of $\hat{\theta}^n$

$q(\theta|\bar{\theta}) = \frac{\det[\mathbf{Q}(\theta,\bar{\theta})]}{(2\pi)^{p/2}\,\sigma^p\,n^{p/2}\,\det^{1/2}\mathbf{M}(X,\theta)}\,\exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{P}_\theta[\eta(\theta)-\eta(\bar{\theta})]\|^2\right\}$

where $\{\mathbf{Q}(\theta,\bar{\theta})\}_{ij} = \{n\,\mathbf{M}(X,\theta)\}_{ij} + [\eta(\theta)-\eta(\bar{\theta})]^\top[\mathbf{I}_n - \mathbf{P}_\theta]\mathbf{H}_{ij}(\theta)$

Remarks:

- This approximation coincides with the saddle-point approximation of Hougaard (1985)
- Other approximations (more complicated) for models with $R_{hijk}(\theta) \not\equiv 0$ (non-flat)
- An approximation of the density of the penalized LS estimator $\arg\min_{\theta\in\Theta}\left\{\|\mathbf{y}-\eta(\theta)\|^2 + 2w(\theta)\right\}$ (which includes the case of Bayesian estimation) is also available
- We also know the (approximate) marginal densities of the LS estimator $\hat{\theta}^n$ [Pázman & P 1996]

C/ Confidence regions

Suppose $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, define $\mathbf{e}(\theta) = \mathbf{y} - \eta(\theta)$

➔ $\mathbf{e}^\top(\bar{\theta})\, \mathbf{P}_{\bar{\theta}}\, \mathbf{e}(\bar{\theta})/\sigma^2 \sim \chi^2_p$

➔ $\mathbf{e}^\top(\bar{\theta})\, [\mathbf{I}_n - \mathbf{P}_{\bar{\theta}}]\, \mathbf{e}(\bar{\theta})/\sigma^2 \sim \chi^2_{n-p}$

and they are independent

C/ Confidence regions

Suppose $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, define $\mathbf{e}(\theta) = \mathbf{y} - \eta(\theta)$

➔ $\mathbf{e}^\top(\bar{\theta})\, \mathbf{P}_{\bar{\theta}}\, \mathbf{e}(\bar{\theta})/\sigma^2 \sim \chi^2_p$

➔ $\mathbf{e}^\top(\bar{\theta})\, [\mathbf{I}_n - \mathbf{P}_{\bar{\theta}}]\, \mathbf{e}(\bar{\theta})/\sigma^2 \sim \chi^2_{n-p}$

and they are independent

➠ exact confidence regions at level α

$$\left\{\theta \in \mathbb{R}^p : \mathbf{e}^\top(\theta)\, \mathbf{P}_\theta\, \mathbf{e}(\theta)/\sigma^2 < \chi^2_p[1-\alpha]\right\} \text{ (if } \sigma^2 \text{ known)}$$

$$\left\{\theta \in \mathbb{R}^p : \frac{n-p}{p} \frac{\mathbf{e}^\top(\theta)\, \mathbf{P}_\theta\, \mathbf{e}(\theta)}{\mathbf{e}^\top(\theta)\, [\mathbf{I}_n - \mathbf{P}_\theta]\, \mathbf{e}(\theta)} < F_{p,n-p}[1-\alpha]\right\} \text{ (if } \sigma^2 \text{ unknown)}$$

(but they are not of minimum volume, maybe composed of disconnected subsets. . .)

C/ Confidence regions

Suppose $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, define $\mathbf{e}(\theta) = \mathbf{y} - \eta(\theta)$

➔ $\mathbf{e}^\top(\bar{\theta})\, \mathbf{P}_{\bar{\theta}}\, \mathbf{e}(\bar{\theta})/\sigma^2 \sim \chi^2_p$

➔ $\mathbf{e}^\top(\bar{\theta})\, [\mathbf{I}_n - \mathbf{P}_{\bar{\theta}}]\, \mathbf{e}(\bar{\theta})/\sigma^2 \sim \chi^2_{n-p}$
and they are independent

➠ exact confidence regions at level α

$$\left\{\theta \in \mathbb{R}^p : \mathbf{e}^\top(\theta)\, \mathbf{P}_\theta\, \mathbf{e}(\theta)/\sigma^2 < \chi^2_p[1-\alpha]\right\} \text{ (if } \sigma^2 \text{ known)}$$

$$\left\{\theta \in \mathbb{R}^p : \frac{n-p}{p}\, \frac{\mathbf{e}^\top(\theta)\, \mathbf{P}_\theta\, \mathbf{e}(\theta)}{\mathbf{e}^\top(\theta)\, [\mathbf{I}_n - \mathbf{P}_\theta]\, \mathbf{e}(\theta)} < F_{p,n-p}[1-\alpha]\right\} \text{ (if } \sigma^2 \text{ unknown)}$$

(but they are not of minimum volume, maybe composed of disconnected subsets. . .)

➠ approximate confidence regions based on likelihood ratio (usually connected):

$$\left\{\theta \in \mathbb{R}^p : \|\mathbf{e}(\theta)\|^2 - \|\mathbf{e}(\hat{\theta})\|^2 < \sigma^2 \chi^2_p[1-\alpha]\right\} \text{ (if } \sigma^2 \text{ known)}$$

$$\left\{\theta \in \mathbb{R}^p : \|\mathbf{e}(\theta)\|^2 / \|\mathbf{e}(\hat{\theta})\|^2 < 1 + \frac{p}{n-p} F_{p,n-p}[1-\alpha]\right\} \text{ (if } \sigma^2 \text{ unknown)}$$

D/ Design based on small-sample properties

3 main ideas (exact design only) based on:

a) (approximate) volume of (approximate) confidence regions (not necessarily of minimum volume) [Hamilton & Watts 1985; Vila 1990; Vila & Gauchi 2007]
(ellipsoidal approximation $\rightarrow$ D-optimality)

D/ Design based on small-sample properties

3 main ideas (exact design only) based on:

a) (approximate) volume of (approximate) confidence regions (not necessarily of minimum volume) [Hamilton & Watts 1985; Vila 1990; Vila & Gauchi 2007]
(ellipsoidal approximation → D-optimality)

b) (approximate or exact) density of $\hat{\theta}^n$
e.g., minimize $\int \|\theta - \bar{\theta}\|^2 q(\theta|\bar{\theta})\, \mathrm{d}\theta$ w.r.t. X (using stochastic approximation, [Pázman & P 1992, Gauchi & Pázman 2006])

D/ Design based on small-sample properties

3 main ideas (exact design only) based on:

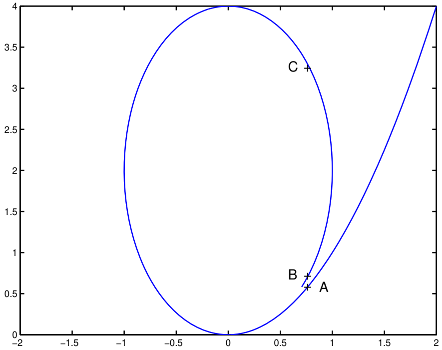
a) (approximate) volume of (approximate) confidence regions (not necessarily of minimum volume) [Hamilton & Watts 1985; Vila 1990; Vila & Gauchi 2007]
   (ellipsoidal approximation $\rightarrow$ D-optimality)

b) (approximate or exact) density of $\hat{\theta}^n$
   e.g., minimize $\int \|\theta - \bar{\theta}\|^2 q(\theta|\bar{\theta})\, \mathrm{d}\theta$ w.r.t. X (using stochastic approximation, [Pázman & P 1992, Gauchi & Pázman 2006])

c) higher-order approximation of optimality criteria,
   using $\varphi(\mathbf{y}|X,\bar{\theta}) = \mathcal{N}(\eta(\bar{\theta}), \sigma^2 \mathbf{I}_n)$
   minimize MSE $\int \|\hat{\theta}^n(\mathbf{y}) - \bar{\theta}\|^2 \varphi(\mathbf{y}|X,\bar{\theta})\, \mathrm{d}\mathbf{y}$ [Clarke 1980]
   minimize entropy $-\int \log[q(\hat{\theta}^n(\mathbf{y})|\bar{\theta})] \varphi(\mathbf{y}|X,\bar{\theta})\, \mathrm{d}\mathbf{y}$ [P & Pázman 1994]
   (usual normal approximation for $q(\cdot|\bar{\theta})$ $\rightarrow$ D-optimality)
   $\rightarrow$ explicit (but rather complicated) expressions (depend on 3rd-order derivatives of $\eta(x,\theta)$ w.r.t. θ)

E/ One additional difficulty

Overlapping of $\mathbb{S}_\eta$, local minimizers...



▲ Important and difficult problem, often neglected!

What can we do at the design stage?

➠ extensions of usual optimality criteria, e.g.

$$\text{maximize } \phi_{eE}(X) = \min_{\theta} \frac{\|\eta(\theta) - \eta(\theta^0)\|^2}{\|\theta - \theta^0\|^2}$$

or

$$\text{maximize } \phi_{eE}(\xi) = \min_{\theta} \frac{\int [\eta(x,\theta) - \eta(x,\theta^0)]^2 \, \xi(\mathrm{d}x)}{\|\theta - \theta^0\|^2}$$

➔ corresponds to E-optimal design if the model is linear (maximize $\lambda_{\min}\mathbf{M}(\xi)$), see Chap. 7 of [P & Pázman 2013] and [Pázman & P, 2014]

What can we do at the design stage?

➠ extensions of usual optimality criteria, e.g.

$$\text{maximize } \phi_{eE}(X) = \min_{\theta} \frac{\|\eta(\theta) - \eta(\theta^0)\|^2}{\|\theta - \theta^0\|^2}$$

or

$$\text{maximize } \phi_{eE}(\xi) = \min_{\theta} \frac{\int [\eta(x,\theta) - \eta(x,\theta^0)]^2 \, \xi(\mathrm{d}x)}{\|\theta - \theta^0\|^2}$$

➔ corresponds to E-optimal design if the model is linear (maximize $\lambda_{\min}\mathbf{M}(\xi)$), see Chap. 7 of [P & Pázman 2013] and [Pázman & P, 2014]

▲ All approaches presented so far are local

(the optimal design depends on $\bar{\theta}$ unknown $\leftarrow \theta^0$)

6 Nonlocal optimum design

Ex: $\eta(x,\theta) = \exp(-\theta x)$, $y_i = \eta(x_i,\bar{\theta}) + \varepsilon_i$, $\theta > 0$, $x \in \mathscr{X} = [0,\infty)$
→ $M(\xi,\theta^0) = \int_{\mathscr{X}} x^2 \exp(-2\theta^0 x)\,\xi(\mathrm{d}x)$
➠ $\xi_D^* = \xi_A^* = \ldots = \delta_{1/\theta^0}$

Objective: remove the dependence in nominal value θ^0

3 main classes of methods (related)

6 Nonlocal optimum design

Ex: $\eta(x,\theta) = \exp(-\theta x)$, $y_i = \eta(x_i, \bar{\theta}) + \varepsilon_i$, $\theta > 0$, $x \in \mathscr{X} = [0, \infty)$

➔ $M(\xi, \theta^0) = \int_{\mathscr{X}} x^2 \exp(-2\theta^0 x)\, \xi(\mathrm{d}x)$

➠ $\xi_D^* = \xi_A^* = \ldots = \delta_{1/\theta^0}$

Objective: remove the dependence in nominal value θ^0

3 main classes of methods (related)

➊ Average optimum design: maximize $\mathsf{E}_\theta\{\phi(X,\theta)\}$ (or $\mathsf{E}_\theta\{\phi(\xi,\theta)\}$)

6 Nonlocal optimum design

Ex: $\eta(x,\theta) = \exp(-\theta x)$, $y_i = \eta(x_i,\bar{\theta}) + \varepsilon_i$, $\theta > 0$, $x \in \mathscr{X} = [0,\infty)$

➔ $M(\xi,\theta^0) = \int_{\mathscr{X}} x^2 \exp(-2\theta^0 x)\, \xi(\mathrm{d}x)$

➠ $\xi_D^* = \xi_A^* = \ldots = \delta_{1/\theta^0}$

Objective: remove the dependence in nominal value θ^0

3 main classes of methods (related)

❶ Average optimum design: maximize $\mathrm{E}_\theta\{\phi(X,\theta)\}$ (or $\mathrm{E}_\theta\{\phi(\xi,\theta)\}$)

❷ Maximin optimum design: maximize $\min_\theta\{\phi(X,\theta)\}$ (or $\min_\theta\{\phi(\xi,\theta)\}$)

➠ Between ❶ and ❷: regularized maximin criteria, quantiles and probability level criteria

6 Nonlocal optimum design

Ex: $\eta(x,\theta) = \exp(-\theta x)$, $y_i = \eta(x_i,\bar{\theta}) + \varepsilon_i$, $\theta > 0$, $x \in \mathscr{X} = [0,\infty)$

➔ $M(\xi,\theta^0) = \int_{\mathscr{X}} x^2 \exp(-2\theta^0 x)\,\xi(\mathrm{d}x)$

➠ $\xi_D^* = \xi_A^* = \ldots = \delta_{1/\theta^0}$

Objective: remove the dependence in nominal value θ^0

3 main classes of methods (related)

❶ Average optimum design: maximize $\mathrm{E}_\theta\{\phi(X,\theta)\}$ (or $\mathrm{E}_\theta\{\phi(\xi,\theta)\}$)

❷ Maximin optimum design: maximize $\min_\theta\{\phi(X,\theta)\}$ (or $\min_\theta\{\phi(\xi,\theta)\}$)

➠ Between ❶ and ❷: regularized maximin criteria, quantiles and probability level criteria

❸ Sequential design

A/ Average Optimum design

Nothing special: probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$

$$\boxed{\phi(\cdot,\theta^0) \rightarrow \phi_{AO}(\cdot) = \int_\Theta \phi(\cdot,\theta)\,\mu(\mathrm{d}\theta)}$$

No difficulty if $\Theta = \{\theta^{(1)},\ldots\theta^{(M)}\}$ finite and $\mu = \sum_{i=1}^{M} \alpha_i \delta_\theta^{(i)}$ (integral $\rightarrow$ finite sum)

A/ Average Optimum design

Nothing special: probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$

$$\boxed{\phi(\cdot,\theta^0) \to \phi_{AO}(\cdot) = \int_\Theta \phi(\cdot,\theta)\,\mu(\mathrm{d}\theta)}$$

No difficulty if $\Theta = \{\theta^{(1)},\ldots\theta^{(M)}\}$ finite and $\mu = \sum_{i=1}^M \alpha_i \delta_\theta^{(i)}$ (integral $\to$ finite sum)

☛ Approximate design theory:

$\phi_{AO}(\cdot)$ is concave when each $\phi(\cdot,\theta)$ is concave

➠ same properties and same algorithms as in Section 3 for design measures

A/ Average Optimum design

Nothing special: probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$

$$\boxed{\phi(\cdot,\theta^0) \to \phi_{AO}(\cdot) = \int_\Theta \phi(\cdot,\theta)\,\mu(\mathrm{d}\theta)}$$

No difficulty if $\Theta = \{\theta^{(1)},\ldots\theta^{(M)}\}$ finite and $\mu = \sum_{i=1}^M \alpha_i \delta_\theta^{(i)}$ (integral $\to$ finite sum)

☛ Approximate design theory:

$\phi_{AO}(\cdot)$ is concave when each $\phi(\cdot,\theta)$ is concave

➠ same properties and same algorithms as in Section 3 for design measures

☛ Exact design: same algorithms as in Section 3

(for continuous distributions μ use stochastic approximation to avoid evaluations of integrals [P & Walter 1985])

A Bayesian interpretation:

Suppose μ=prior distribution has a density $\pi(\theta)$

$\rightarrow$ entropy $-\int_\Theta \pi(\theta)\,\log[\pi(\theta)]\,\mathrm{d}\theta$

Posterior distribution of θ: $\pi(\theta|X,\mathbf{y}) = \frac{\varphi(\mathbf{y}|X,\theta)\pi(\theta)}{\varphi(\mathbf{y}|X)}$

Gain in information = decrease of entropy

Entropy may increase, but expected gain in information $\mathcal{I}(X)$ is always positive [Lindley 1956]

➠ $\mathcal{I}(X) = \mathsf{E}_{\mathbf{y}}\{\int_\Theta(\pi(\theta|X,\mathbf{y})\,\log[\pi(\theta|X,\mathbf{y})] - \pi(\theta)\,\log[\pi(\theta)])\,\mathrm{d}\theta\}$

where the expectation $\mathsf{E}_{\mathbf{y}}\{\cdot\}$ is for the marginal $\varphi(\mathbf{y}|X)$

A Bayesian interpretation:

Suppose μ=prior distribution has a density $\pi(\theta)$

$\to$ entropy $-\int_\Theta \pi(\theta)\, \log[\pi(\theta)]\, d\theta$

Posterior distribution of θ: $\pi(\theta|X, \mathbf{y}) = \frac{\varphi(\mathbf{y}|X,\theta)\pi(\theta)}{\varphi(\mathbf{y}|X)}$

Gain in information = decrease of entropy

Entropy may increase, but expected gain in information $\mathcal{I}(X)$ is always positive [Lindley 1956]

➠ $\mathcal{I}(X) = \mathsf{E}_{\mathbf{y}}\{\int_\Theta(\pi(\theta|X, \mathbf{y})\, \log[\pi(\theta|X, \mathbf{y})] - \pi(\theta)\, \log[\pi(\theta)])\, d\theta\}$

where the expectation $\mathsf{E}_{\mathbf{y}}\{\cdot\}$ is for the marginal $\varphi(\mathbf{y}|X)$

If the experiment informative enough (σ^2 small, n large enough):

maximizing $\mathcal{I}(X) \overset{\approx}{\Leftrightarrow}$ maximizing $\int \log\det \mathbf{M}(X, \theta)\, \pi(\theta)\, d\theta$

Which prior $\pi(\theta)$? Expected gain in information maximum when $\pi(\cdot) =$ noninformative prior (Jeffrey) — which depends on ξ

$\pi^*(\theta) = \frac{\det^{1/2} \mathbf{M}(\xi,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\xi,\theta)\, d\theta}$ ➠ maximize $\int \det^{1/2} \mathbf{M}(\xi,\theta)\, d\theta$

Which prior $\pi(\theta)$? Expected gain in information maximum when $\pi(\cdot) =$ noninformative prior (Jeffrey) — which depends on ξ

$\pi^*(\theta) = \frac{\det^{1/2} \mathbf{M}(\xi,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\xi,\theta)\, d\theta}$ ➠ maximize $\int \det^{1/2} \mathbf{M}(\xi,\theta)\, d\theta$

$\pi_\nu(\theta) = \frac{\det^{1/2} \mathbf{M}(\nu,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\nu,\theta)\, d\theta}$ $\rightarrow$ uniform distribution of responses $\eta(\cdot,\theta)$ (for the metric defined by ν) [Bornkamp 2011]

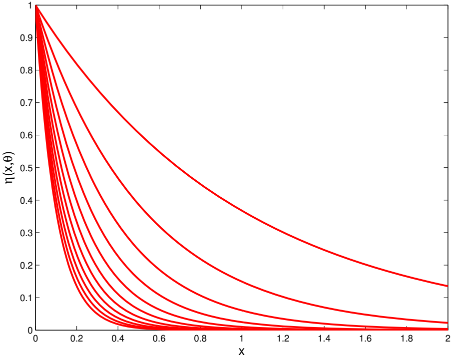
Which prior $\pi(\theta)$? Expected gain in information maximum when $\pi(\cdot) =$ noninformative prior (Jeffrey) — which depends on ξ

$\pi^*(\theta) = \frac{\det^{1/2} \mathbf{M}(\xi,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\xi,\theta)\, d\theta}$ ➠ maximize $\int \det^{1/2} \mathbf{M}(\xi,\theta)\, d\theta$

$\pi_\nu(\theta) = \frac{\det^{1/2} \mathbf{M}(\nu,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\nu,\theta)\, d\theta}$ $\rightarrow$ uniform distribution of responses $\eta(\cdot,\theta)$ (for the metric defined by ν) [Bornkamp 2011]

Ex: $\eta(x,\theta) = \exp(-\theta x)$, α-quantiles of $\eta(x,\theta)$ for different π

π uniform on $\Theta = [1, 10]$



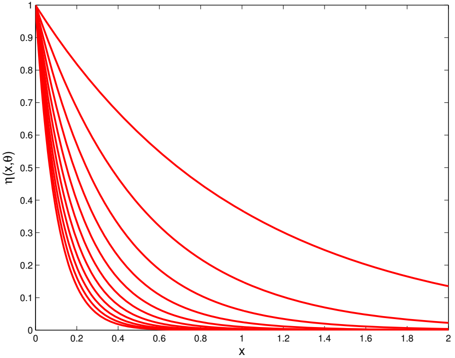
Which prior $\pi(\theta)$? Expected gain in information maximum when $\pi(\cdot)$ = noninformative prior (Jeffrey) — which depends on ξ

$$\pi^*(\theta) = \frac{\det^{1/2} \mathbf{M}(\xi,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\xi,\theta)\,\mathrm{d}\theta}$$ ➠ maximize $\int \det^{1/2} \mathbf{M}(\xi,\theta)\,\mathrm{d}\theta$

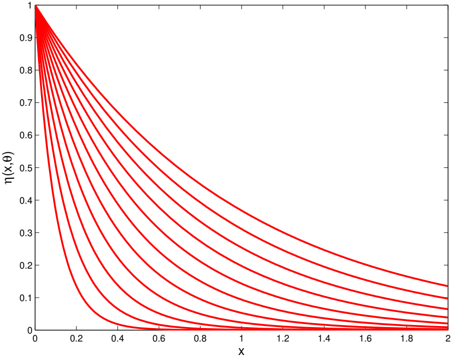
$$\pi_\nu(\theta) = \frac{\det^{1/2} \mathbf{M}(\nu,\theta)}{\int_\Theta \det^{1/2} \mathbf{M}(\nu,\theta)\,\mathrm{d}\theta}$$ → uniform distribution of responses $\eta(\cdot,\theta)$ (for the metric defined by ν) [Bornkamp 2011]

Ex: $\eta(x,\theta) = \exp(-\theta x)$, α-quantiles of $\eta(x,\theta)$ for different π

π uniform on $\Theta = [1, 10]$



π_ν, ν uniform on $\mathscr{X} = [0, 2]$



B/ Maximin Optimum design

$$\boxed{\phi(\cdot,\theta^0) \rightarrow \phi_{MmO}(\cdot) = \min_{\theta\in\Theta} \phi(\cdot,\theta)}$$

☛ Exact design:

Θ finite $\rightarrow$ same algorithms as in Section 3

Θ compact subset of $\mathbb{R}^p$ $\rightarrow$ relaxation method to solve a sequence of maximin problems with finite (and growing) sets $\Theta^{(k)}$ [P & Walter 1988]

B/ Maximin Optimum design

$$\phi(\cdot, \theta^0) \to \phi_{MmO}(\cdot) = \min_{\theta \in \Theta} \phi(\cdot, \theta)$$

- Exact design:
  - Θ finite $\to$ same algorithms as in Section 3
  - Θ compact subset of $\mathbb{R}^p$ $\to$ relaxation method to solve a sequence of maximin problems with finite (and growing) sets $\Theta^{(k)}$ [P & Walter 1988]
- Approximate design:
  - $\phi_{MmO}(\cdot)$ concave when each $\phi(\cdot, \theta)$ is concave
  - ▲ but $\phi_{MmO}(\cdot)$ is non-differentiable!
  - ➠ maximize $\phi_{MmO}(\xi)$ using a specific algorithm for concave non-differentiable maximization (cutting plane, level method... see Section 3)

How to check optimality of ξ^*?

$\phi(\cdot,\theta^0)$ differentiable: $\max_{x\in\mathscr{X}} F_\phi(\xi^*;\delta_x,\theta^0) \leq 0$?
$\rightarrow$ plot $F_\phi(\xi^*;\delta_x,\theta^0)$ as a function of x

$\phi_{MmO}(\cdot)$ not differentiable: $\max_{\nu\in\Xi} F_{\phi_{MmO}}(\xi^*;\nu) \leq 0$ cannot be exploited directly

How to check optimality of ξ^*?

$\phi(\cdot,\theta^0)$ differentiable: $\max_{x\in\mathscr{X}} F_\phi(\xi^*;\delta_x,\theta^0) \leq 0$?
$\rightarrow$ plot $F_\phi(\xi^*;\delta_x,\theta^0)$ as a function of x

$\phi_{MmO}(\cdot)$ not differentiable: $\max_{\nu\in\Xi} F_{\phi_{MmO}}(\xi^*;\nu) \leq 0$ cannot be exploited directly

Equivalence Theorem:

ξ^* maximizes $\phi_{MmO}(\xi) \Leftrightarrow \max_{\nu\in\Xi} F_{\phi_{MmO}}(\xi^*;\nu) \leq 0$
$\Leftrightarrow \max_{\nu\in\Xi} \min_{\theta\in\Theta(\xi^*)} F_\phi(\xi^*;\nu,\theta) \leq 0$
with $\Theta(\xi) = \{\theta : \phi(\xi,\theta) = \phi_{MmO}(\xi)\}$
$\Leftrightarrow \max_{x\in\mathscr{X}} \int_{\Theta(\xi^*)} F_\phi(\xi^*;\delta_x,\theta)\mu^*(\mathrm{d}\theta) \leq 0$
for some probability measure μ^* on $\Theta(\xi^*)$

How to check optimality of ξ^*?

$\phi(\cdot,\theta^0)$ differentiable: $\max_{x\in\mathscr{X}} F_\phi(\xi^*;\delta_x,\theta^0) \leq 0$?
$\rightarrow$ plot $F_\phi(\xi^*;\delta_x,\theta^0)$ as a function of x

$\phi_{MmO}(\cdot)$ not differentiable: $\max_{\nu\in\Xi} F_{\phi_{MmO}}(\xi^*;\nu) \leq 0$ cannot be exploited directly

Equivalence Theorem:

ξ^* maximizes $\phi_{MmO}(\xi) \Leftrightarrow \max_{\nu\in\Xi} F_{\phi_{MmO}}(\xi^*;\nu) \leq 0$
$\Leftrightarrow \max_{\nu\in\Xi} \min_{\theta\in\Theta(\xi^*)} F_\phi(\xi^*;\nu,\theta) \leq 0$
with $\Theta(\xi) = \{\theta : \phi(\xi,\theta) = \phi_{MmO}(\xi)\}$
$\Leftrightarrow \max_{x\in\mathscr{X}} \int_{\Theta(\xi^*)} F_\phi(\xi^*;\delta_x,\theta)\mu^*(\mathrm{d}\theta) \leq 0$
for some probability measure μ^* on $\Theta(\xi^*)$

Once ξ^* is determined, solve a LP problem:
μ^* on $\Theta(\xi^*)$ minimizes $\max_{x\in\mathscr{X}} \int_{\Theta(\xi^*)} F_\phi(\xi^*;\delta_x,\theta)\mu(\mathrm{d}\theta)$
➠ plot $\int_{\Theta(\xi^*)} F_\phi(\xi^*;\delta_x,\theta)\mu^*(\mathrm{d}\theta)$ (should be ≤ 0)

Ex: $\eta(x,\theta) = \theta_1 \exp(-\theta_2 x)$, $p = 2$, $\mathscr{X} = [0,2]$, $\theta_2 \in [0, \theta_{2_{\max}}]$

$$\phi(\xi,\theta) = \frac{\det^{1/p} \mathbf{M}(\xi,\theta)}{\det^{1/p} \mathbf{M}(\xi_D^*,\theta)} \quad (D \text{ efficiency}, \in [0,1])$$

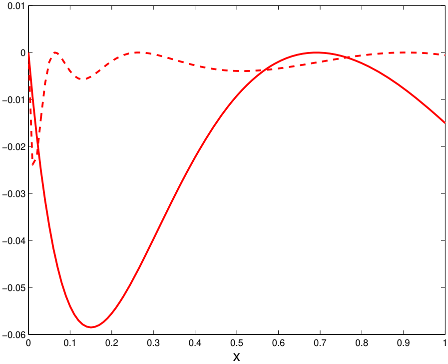
Ex: $\eta(x,\theta) = \theta_1 \exp(-\theta_2 x)$, $p = 2$, $\mathscr{X} = [0,2]$, $\theta_2 \in [0, \theta_{2_{\max}}]$

$$\phi(\xi,\theta) = \frac{\det^{1/p} \mathbf{M}(\xi,\theta)}{\det^{1/p} \mathbf{M}(\xi_D^*,\theta)}$$ (D efficiency, $\in [0,1]$)

$\int_{\Theta(\xi^*)} F_\phi(\xi^*; \delta_x, \theta)\mu^*(\mathrm{d}\theta)$ for

$\theta_{2_{\max}} = 2$ (solid line, 2 support points) and

$\theta_{2_{\max}} = 20$ (dashed line, 4 support points)



C/ Regularized Maximin Optimum design

Suppose $\phi(\cdot,\theta) > 0$ for all θ; μ a probability measure on $\Theta \subset \mathbb{R}^p$ compact

$\phi_{MmO}(\xi) = \min_{\theta\in\Theta} \phi(\cdot,\theta) \leq \overline{\phi}_q(\xi) = \left[\int_\Theta \phi^{-q}(\xi,\theta)\mu(\mathrm{d}\theta)\right]^{-\frac{1}{q}}$ (differentiable)

with $\overline{\phi}_{-1}(\xi) = \phi_{AO}(\xi)$, $\overline{\phi}_0(\xi) = \exp\left\{\int_\Theta \log[\phi(\xi,\theta)]\mu(\mathrm{d}\theta)\right\}$ and

$\boxed{\overline{\phi}_q(\xi) \to \phi_{MmO}(\xi) \text{ as } q \to \infty}$ (and $\overline{\phi}_q(\cdot)$ concave for $q \geq -1$)

Moreover, $\Theta = \{\theta^{(1)},\ldots,\theta^{(M)}\}$, $\mu = \frac{\sum_i \delta_{\theta^{(i)}}}{M} \implies \frac{\phi_{MmO}(\xi_q^*)}{\phi_{MmO}^*} \geq M^{-1/q}$

C/ Regularized Maximin Optimum design

Suppose $\phi(\cdot,\theta) > 0$ for all θ; μ a probability measure on $\Theta \subset \mathbb{R}^p$ compact

$\phi_{MmO}(\xi) = \min_{\theta\in\Theta} \phi(\cdot,\theta) \leq \overline{\phi}_q(\xi) = \left[\int_\Theta \phi^{-q}(\xi,\theta)\mu(\mathrm{d}\theta)\right]^{-\frac{1}{q}}$ (differentiable)

with $\overline{\phi}_{-1}(\xi) = \phi_{AO}(\xi)$, $\overline{\phi}_0(\xi) = \exp\left\{\int_\Theta \log[\phi(\xi,\theta)]\mu(\mathrm{d}\theta)\right\}$ and

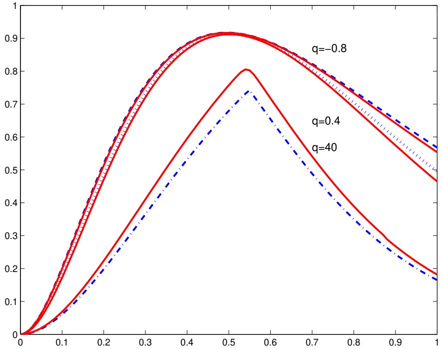
$\overline{\phi}_q(\xi) \to \phi_{MmO}(\xi)$ as $q \to \infty$ (and $\overline{\phi}_q(\cdot)$ concave for $q \geq -1$)

Moreover, $\Theta = \{\theta^{(1)},\ldots,\theta^{(M)}\}$, $\mu = \frac{\sum_i \delta_{\theta^{(i)}}}{M} \implies \frac{\phi_{MmO}(\xi_q^*)}{\phi_{MmO}^*} \geq M^{-1/q}$

Ex: $\eta = \exp(-\theta x)$

$\phi(\xi,\theta) = \frac{M(\xi,\theta)}{M(\xi^*,\theta)}$ (= efficiency)

Plot of $\overline{\phi}_q(\delta_x)$ function of x



D/ Quantiles and probability level criteria

A/ ξ^*_{AO} good for $\mu(\mathrm{d}\theta)$ on Θ, may be bad for some θ

▲ for $\psi(\cdot) \nearrow$, maximizing $\psi\left[\int_\Theta \phi(\cdot,\theta)\,\mu(\mathrm{d}\theta)\right]$ (AO-opt.)
is different from maximizing $\int_\Theta \psi[\phi(\cdot,\theta)]\,\mu(\mathrm{d}\theta)$

B/ ξ^*_{MmO} often depends on the boundary of Θ
(➔ we often simply replace the dependence on θ^0 by a dependence on θ_{max})

D/ Quantiles and probability level criteria

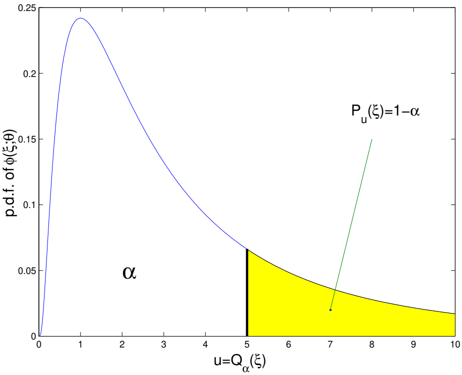
A/ ξ^*_{AO} good for $\mu(\mathrm{d}\theta)$ on Θ, may be bad for some θ

▲ for $\psi(\cdot) \nearrow$, maximizing $\psi\left[\int_\Theta \phi(\cdot,\theta)\,\mu(\mathrm{d}\theta)\right]$ (AO-opt.)
is different from maximizing $\int_\Theta \psi[\phi(\cdot,\theta)]\,\mu(\mathrm{d}\theta)$

B/ ξ^*_{MmO} often depends on the boundary of Θ
(➔ we often simply replace the dependence on θ^0 by a dependence on $\theta_{\max}$)

u given $\rightarrow$
$P_u(\xi) = \mu\{\phi(\xi,\theta) \geq u\}$

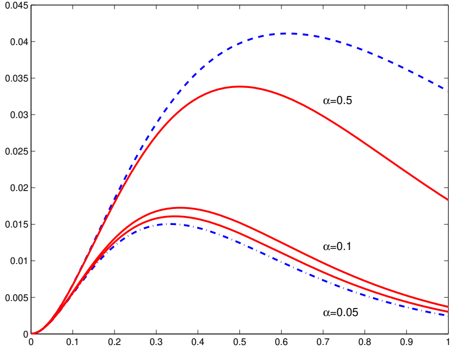
$\alpha \in (0,1)$ given $\rightarrow$
$Q_\alpha(\xi) = \max\{u : P_u(\xi) \geq 1-\alpha\}$



- ➤ maximizing $P_u(\xi) = \mu\{\phi(\xi, \theta) \geq u\}$ is well adapted to
  $\phi(\xi, \theta) =$ efficiency $\in (0, 1)$
- ➤ $Q_\alpha(\xi) \to \phi_{MMO}$ as $\alpha \to 0$
- ➤ for $\psi(\cdot) \nearrow$, using $\psi[\phi(\xi, \theta)]$ does not change $P_u(\xi)$ and $Q_\alpha(\xi)$
- ▲ $P_u(\xi)$ and $Q_\alpha(\xi)$ generally not concave!
  (but we can compute directional derivatives and maximize)

- ➤ maximizing $P_u(\xi) = \mu\{\phi(\xi,\theta) \geq u\}$ is well adapted to
  $\phi(\xi,\theta) =$ efficiency $\in (0,1)$
- ➤ $Q_\alpha(\xi) \to \phi_{MMO}$ as $\alpha \to 0$
- ➤ for $\psi(\cdot) \nearrow$, using $\psi[\phi(\xi,\theta)]$ does not change $P_u(\xi)$ and $Q_\alpha(\xi)$
- ▲ $P_u(\xi)$ and $Q_\alpha(\xi)$ generally not concave!
  (but we can compute directional derivatives and maximize)

Ex: $\eta = \exp(-\theta x)$
$\phi(\xi,\theta) = M(\xi,\theta)$
Plot of $Q_\alpha(\delta_x)$ function of x
(with $\phi_{AO}(\delta_x)$ and $\phi_{MmO}(\delta_x)$)



Ongoing work: conditional value at risk (also called superquantile)

$$\phi_\alpha(\xi) = \frac{1}{\alpha} \int_{\{\theta:\phi(\xi,\theta)\leq Q_\alpha(\xi)\}} \phi(\xi,\theta)\, \mu(\mathrm{d}\theta)$$

which is concave in ξ when $\phi(\cdot,\theta)$ is concave for all θ, see (Valenzuela et al., 2015; Guerra, 2016)

E/ Sequential design

$\theta^0 \rightarrow$ design: $X^1 = \arg\max_X \phi(X, \theta^0)$

$\rightarrow$ observe: $\mathbf{y}^1 = \mathbf{y}^1(X^1)$

$\rightarrow$ estimate: $\hat{\theta}^1 = \arg\min_\theta J(\theta; \mathbf{y}^1, X^1)$

$\rightarrow$ design: $X^2 = \arg\max_X \phi(\{X^1, X\}, \hat{\theta}^1)$

$\rightarrow$ observe: $\mathbf{y}^2 = \mathbf{y}^2(X^2)$

$\rightarrow$ estimate: $\hat{\theta}^2 = \arg\min_\theta J(\theta; \{\underbrace{\mathbf{y}^1, \mathbf{y}^2}_{\text{growing}}\}, \{\underbrace{X^1, X^2}_{\text{growing}}\})$

$\rightarrow$ design: $X^3 = \arg\max_X \phi(\{X^1, X^2, X\}, \hat{\theta}^2)$

...etc.

E/ Sequential design

$\theta^0 \to$ design: $X^1 = \arg\max_X \phi(X, \theta^0)$
$\to$ observe: $\mathbf{y}^1 = \mathbf{y}^1(X^1)$
$\to$ estimate: $\hat{\theta}^1 = \arg\min_\theta J(\theta; \mathbf{y}^1, X^1)$
$\to$ design: $X^2 = \arg\max_X \phi(\{X^1, X\}, \hat{\theta}^1)$
$\to$ observe: $\mathbf{y}^2 = \mathbf{y}^2(X^2)$
$\to$ estimate: $\hat{\theta}^2 = \arg\min_\theta J(\theta; \{\underbrace{\mathbf{y}^1, \mathbf{y}^2}_{\text{growing}}\}, \{\underbrace{X^1, X^2}_{\text{growing}}\})$
$\to$ design: $X^3 = \arg\max_X \phi(\{X^1, X^2, X\}, \hat{\theta}^2)$
...etc.

☞ Replace unknown θ by best current guess $\hat{\theta}^k$
(there exist variants with Bayesian estimation and average optimality)

E/ Sequential design

$\theta^0 \to$ design: $X^1 = \arg\max_X \phi(X, \theta^0)$
$\to$ observe: $\mathbf{y}^1 = \mathbf{y}^1(X^1)$
$\to$ estimate: $\hat{\theta}^1 = \arg\min_\theta J(\theta; \mathbf{y}^1, X^1)$
$\to$ design: $X^2 = \arg\max_X \phi(\{X^1, X\}, \hat{\theta}^1)$
$\to$ observe: $\mathbf{y}^2 = \mathbf{y}^2(X^2)$
$\to$ estimate: $\hat{\theta}^2 = \arg\min_\theta J(\theta; \{\underbrace{\mathbf{y}^1, \mathbf{y}^2}_{\text{growing}}\}, \{\underbrace{X^1, X^2}_{\text{growing}}\})$
$\to$ design: $X^3 = \arg\max_X \phi(\{X^1, X^2, X\}, \hat{\theta}^2)$
...etc.

☞ Replace unknown θ by best current guess $\hat{\theta}^k$
(there exist variants with Bayesian estimation and average optimality)

▲ Consistency of $\hat{\theta}^n$?
Asymptotic normality (for design based on **M**)?
(X^k depends on $\mathbf{y}^1, \ldots, \mathbf{y}^{k-1} \Longrightarrow$ independence is lost)

➟ No problem if each X^i has size $\geq p = \dim(\theta)$ (batch sequential design)

➠ No problem if each X^i has size $\geq p = \mathrm{dim}(\theta)$ (batch sequential design)
If n observation in total, two stages only: size of first batch?
$\rightarrow$ should be proportional to $\sqrt{n}$ (but it does not say much . . .)

➠ No problem if each X^i has size $\geq p = \dim(\theta)$ (batch sequential design)
If n observation in total, two stages only: size of first batch?
$\rightarrow$ should be proportional to $\sqrt{n}$ (but it does not say much ...)

➠ Full sequential design: $X^k = \{x_k\}$ (batches of size 1)
$\rightarrow$ convergence properties difficult to investigate

$$\mathbf{M}(X_{k+1},\hat{\theta}^k) = \frac{k}{k+1}\mathbf{M}(X_k,\hat{\theta}^k) + \frac{1}{k+1}\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta}\Big|_{\hat{\theta}^k}\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta^\top}\Big|_{\hat{\theta}^k}$$

with $x_{k+1} = \arg\max_{\mathscr{X}} F_\phi(\xi^k;\delta_x|\hat{\theta}^k) \Leftrightarrow$ Wynn's algorithm [1970] with $\alpha_k = \frac{1}{k+1}$

➠ No problem if each X^i has size $\geq p = \dim(\theta)$ (batch sequential design)
If n observation in total, two stages only: size of first batch?
$\rightarrow$ should be proportional to $\sqrt{n}$ (but it does not say much ...)

➠ Full sequential design: $X^k = \{x_k\}$ (batches of size 1)
$\rightarrow$ convergence properties difficult to investigate

$$\mathbf{M}(X_{k+1}, \hat{\theta}^k) = \frac{k}{k+1}\mathbf{M}(X_k, \hat{\theta}^k) + \frac{1}{k+1}\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta}\Big|_{\hat{\theta}^k}\frac{\partial\eta(x_{k+1},\theta)}{\partial\theta^\top}\Big|_{\hat{\theta}^k}$$

with $x_{k+1} = \arg\max_{\mathscr{X}} F_\phi(\xi^k; \delta_x|\hat{\theta}^k)$ $\Leftrightarrow$ Wynn's algorithm [1970] with $\alpha_k = \frac{1}{k+1}$
➤ some CV results for Bayesian estimation [Hu 1998]
➤ no general CV results for LS and ML estimation
some results when $\mathscr{X}$ is finite ($\mathscr{X} = \{x^{(1)}, \ldots, x^{(\ell)}\}$) [P 2009, 2010]

References I

Atkinson, A., Cox, D., 1974. Planning experiments for discriminating between models (with discussion). Journal of Royal Statistical Society B36, 321–348.

Atkinson, A., Fedorov, V., 1975. The design of experiments for discriminating between two rival models. Biometrika 62 (1), 57–70.

Atwood, C., 1973. Sequences converging to *D*-optimal designs of experiments. Annals of Statistics 1 (2), 342–352.

Bates, D., Watts, D., 1980. Relative curvature measures of nonlinearity. Journal of Royal Statistical Society B42, 1–25.

Böhning, D., 1985. Numerical estimation of a probability measure. Journal of Statistical Planning and Inference 11, 57–69.

Böhning, D., 1986. A vertex-exchange-method in *D*-optimal design theory. Metrika 33, 337–347.

Box, G., Hill, W., 1967. Discrimination among mechanistic models. Technometrics 9 (1), 57–71.

Chernoff, H., 1953. Locally optimal designs for estimating parameters. Annals of Math. Stat. 24, 586–602.

Clarke, G., 1980. Moments of the least-squares estimators in a non-linear regression model. Journal of Royal Statistical Society B42, 227–237.

D'Argenio, D., 1981. Optimal sampling times for pharmacokinetic experiments. Journal of Pharmacokinetics and Biopharmaceutics 9 (6), 739–756.

Dette, H., Melas, V., 2011. A note on de la Garza phenomenon for locally optimal designs. Annals of Statistics 39 (2), 1266–1281.

Fedorov, V., 1972. Theory of Optimal Experiments. Academic Press, New York.

Fedorov, V., Leonov, S., 2014. Optimal Design for Nonlinear Response Models. CRC Press, Boca Raton.

References II

Fisher, R., 1925. Statistical Methods for Research Workers. Oliver & Boyd, Edimbourgh.

Gauchi, J.-P., Pázman, A., 2006. Designs in nonlinear regression by stochastic minimization of functionnals of the mean square error matrix. Journal of Statistical Planning and Inference 136, 1135–1152.

Goodwin, G., Payne, R., 1977. Dynamic System Identification: Experiment Design and Data Analysis. Academic Press, New York.

Guerra, J., 2016. Optimisation multi-objectif sous incertitude de phénomènes de thermique transitoire. Ph.D. Thesis, Université de Toulouse.

Hamilton, D., Watts, D., 1985. A quadratic design criterion for precise estimation in nonlinear regression models. Technometrics 27, 241–250.

Hill, P., 1978. A review of experimental design procedures for regression model discrimination. Technometrics 20, 15–21.

Hougaard, P., 1985. Saddlepoint approximations for curved exponential families. Statistics & Probability Letters 3, 161–166.

Hu, I., 1998. On sequential designs in nonlinear problems. Biometrika 85 (2), 496–503.

Kelley, J., 1960. The cutting plane method for solving convex programs. SIAM Journal 8, 703–712.

Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canadian Journal of Mathematics 12, 363–366.

Ljung, L., 1987. System Identification, Theory for the User. Prentice-Hall, Englewood Cliffs.

Mitchell, T., 1974. An algorithm for the construction of “D-optimal” experimental designs. Technometrics 16, 203–210.

Nesterov, Y., 2004. Introductory Lectures to Convex Optimization: A Basic Course. Kluwer, Dordrecht.

References III

Niederreiter, H., 1992. Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia.

Pázman, A., 1986. Foundations of Optimum Experimental Design. Reidel (Kluwer group), Dordrecht (co-pub. VEDA, Bratislava).

Pázman, A., 1993. Nonlinear Statistical Models. Kluwer, Dordrecht.

Pázman, A., Pronzato, L., 1992. Nonlinear experimental design based on the distribution of estimators. Journal of Statistical Planning and Inference 33, 385–402.

Pázman, A., Pronzato, L., 1996. A Dirac function method for densities of nonlinear statistics and for marginal densities in nonlinear regression. Statistics & Probability Letters 26, 159–167.

Pázman, A., Pronzato, L., 2014. Optimum design accounting for the global nonlinear behavior of the model. Annals of Statistics 42 (4), 1426–1451.

Pronzato, L., 2009. Asymptotic properties of nonlinear estimates in stochastic models with finite design space. Statistics & Probability Letters 79, 2307–2313.

Pronzato, L., 2010. One-step ahead adaptive D-optimal design on a finite design space is asymptotically optimal. Metrika 71 (2), 219–238, (DOI: 10.1007/s00184-008-0227-y).

Pronzato, L., Pázman, A., 1994. Second-order approximation of the entropy in nonlinear least-squares estimation. Kybernetika 30 (2), 187–198, *Erratum* 32(1):104, 1996.

Pronzato, L., Pázman, A., 2013. Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties. Springer, LNS 212, New York.

Pronzato, L., Walter, E., 1985. Robust experiment design via stochastic approximation. Mathematical Biosciences 75, 103–120.

References IV

Pronzato, L., Walter, E., 1988. Robust experiment design via maximin optimization. Mathematical Biosciences 89, 161–176.

Pronzato, L., Zhigljavsky, A., 2014. Algorithmic construction of optimal designs on compact sets for concave and differentiable criteria. Journal of Statistical Planning and Inference 154, 141–155.

Pukelsheim, F., 1993. Optimal Experimental Design. Wiley, New York.

Pukelsheim, F., Reider, S., 1992. Efficient rounding of approximate designs. Biometrika 79 (4), 763–770.

Schwabe, R., 1995. Designing experiments for additive nonlinear models. In: Kitsos, C., Müller, W. (Eds.), MODA4 – Advances in Model-Oriented Data Analysis, Spetses (Greece), june 1995. Physica Verlag, Heidelberg, pp. 77–85.

Silvey, S., 1980. Optimal Design. Chapman & Hall, London.

Titterington, D., 1976. Algorithms for computing D-optimal designs on a finite design space. In: Proc. of the 1976 Conference on Information Science and Systems. Dept. of Electronic Engineering, John Hopkins University, Baltimore, pp. 213–216.

Torsney, B., 1983. A moment inequality and monotonicity of an algorithm. In: Kortanek, K., Fiacco, A. (Eds.), Proc. Int. Symp. on Semi-infinite Programming and Applications. Springer, Heidelberg, pp. 249–260.

Torsney, B., 2009. W-iterations and ripples therefrom. In: Pronzato, L., Zhigljavsky, A. (Eds.), Optimal Design and Related Areas in Optimization and Statistics. Springer, Ch. 1, pp. 1–12.

Valenzuela, P., ROjas, C., Hjalmarsson, H., 2015. Uncertainty in system identification: learning from the theory of risk. IFAC-PapersOnLine 48 (28), 1053–1058.

Vila, J.-P., 1990. Exact experimental designs via stochastic optimization for nonlinear regression models. In: Proc. Compstat, Int. Assoc. for Statistical Computing. Physica Verlag, Heidelberg, pp. 291–296.

References V

Vila, J.-P., Gauchi, J.-P., 2007. Optimal designs based on exact confidence regions for parameter estimation of a nonlinear regression model. Journal of Statistical Planning and Inference 137, 2935–2953.

Walter, E., Pronzato, L., 1994. Identification de Modèles Paramétriques à Partir de Données Expérimentales. Masson, Paris, 371 pages.

Walter, E., Pronzato, L., 1997. Identification of Parametric Models from Experimental Data. Springer, Heidelberg.

Welch, W., 1982. Branch-and-bound search for experimental designs based on D-optimality and other criteria. Technometrics 24 (1), 41–28.

Wu, C., 1978. Some algorithmic aspects of the theory of optimal designs. Annals of Statistics 6 (6), 1286–1301.

Wynn, H., 1970. The sequential generation of D-optimum experimental designs. Annals of Math. Stat. 41, 1655–1664.

Yang, M., 2010. On de la Garza phenomenon. Annals of Statistics 38 (4), 2499–2524.

Yang, M., Biedermann, S., Tang, E., 2013. On optimal designs for nonlinear models: a general and efficient algorithm. Journal of the American Statistical Association 108 (504), 1411–1420.

Yu, Y., 2010. Strict monotonicity and convergence rate of Titterington's algorithm for computing D-optimal designs. Comput. Statist. Data Anal. 54, 1419–1425.

Yu, Y., 2011. D-optimal designs via a cocktail algorithm. Stat. Comput. 21, 475–481.

Zarrop, M., 1979. Optimal Experiment Design for Dynamic System Identification. Springer, Heidelberg.