Experimental design in nonlinear models: small-sample properties

LUC PRONZATO

CNRS, Université Côte d'Azur

1 Introduction

Regression model

$$\underbrace{y_i = y(x_i)}_{\text{observation at } x_i} \quad = \quad \overbrace{\eta(x_i, \bar{\theta})}^{\text{model response at } x_i} \quad + \underbrace{\varepsilon_i}_{\text{error}}$$

where the ε_i are i.i.d., with $\mathrm{E}\{\varepsilon_i\} = 0$ and $\mathrm{E}\{\varepsilon_i^2\} = \sigma^2$

1 Introduction

Regression model

$$\underbrace{y_i = y(x_i)}_{\text{observation at } x_i} = \overbrace{\eta(x_i, \bar{\theta})}^{\text{model response at } x_i} + \underbrace{\varepsilon_i}_{\text{error}}$$

where the ε_i are i.i.d., with $\mathsf{E}\{\varepsilon_i\} = 0$ and $\mathsf{E}\{\varepsilon_i^2\} = \sigma^2$

$\mathbf{X}_n = (x_1, \ldots, x_n)$ the design

$\mathbf{y} = (y_1, \ldots, y_n)^\top$ the vector of observations

$\boldsymbol{\eta}(\theta) = (\eta(x_1, \theta), \ldots, \eta(x_n, \theta))^\top$ the vector of model responses

$\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ the errors ($\rightarrow \mathsf{E}\{\boldsymbol{\varepsilon}\} = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n$)

$\bar{\theta}$ = true value of the model parameters $\theta \in \mathbb{R}^p$

Least Squares (LS) estimator: $\boxed{\hat{\theta}^n = \arg\min_\theta \|\mathbf{y} - \boldsymbol{\eta}(\theta)\|^2}$

Information matrix (at θ^0, normalised — per observation)

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

(a $p \times p$ matrix, with $p = \dim(\theta)$)

A. Linear regression

$\eta(x,\theta) = \mathbf{f}^\top(x)\theta \rightarrow \frac{\partial \eta^\top(\theta)}{\partial\theta} = \mathbf{F}^\top = (\mathbf{f}(x_1),\ldots,\mathbf{f}(x_n))$ and

$$\boxed{\hat{\theta}^n = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{y}}$$

normalised information matrix: $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\mathbf{F}^\top\mathbf{F}$

$\rightarrow$ choose the x_i such that $\mathbf{M}_n$ has full rank

A. Linear regression

$\eta(x,\theta) = \mathbf{f}^\top(x)\theta \rightarrow \frac{\partial \eta^\top(\theta)}{\partial\theta} = \mathbf{F}^\top = (\mathbf{f}(x_1),\ldots,\mathbf{f}(x_n))$ and

$$\boxed{\hat{\theta}^n = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{y}}$$

normalised information matrix: $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\,\mathbf{F}^\top\mathbf{F}$

➞ choose the x_i such that $\mathbf{M}_n$ has full rank

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

$$\Rightarrow \mathrm{E}\{\hat{\theta}^n\} = \bar{\theta}$$

$$\Rightarrow \mathrm{Var}(\hat{\theta}^n) = \mathrm{E}\{(\hat{\theta}^n - \bar{\theta})(\hat{\theta}^n - \bar{\theta})^\top\} = \frac{\sigma^2}{n}\,\mathbf{M}_n^{-1}$$

➞ choose the x_i to minimise a scalar function of $\mathbf{M}_n^{-1}$
or maximise a function $\Phi(\mathbf{M}_n)$ (information function (Pukelsheim, 1993))

A. Linear regression

$\eta(x,\theta) = \mathbf{f}^\top(x)\theta \to \frac{\partial \eta^\top(\theta)}{\partial\theta} = \mathbf{F}^\top = (\mathbf{f}(x_1), \ldots, \mathbf{f}(x_n))$ and

$$\boxed{\hat{\theta}^n = (\mathbf{F}^\top\mathbf{F})^{-1}\mathbf{F}^\top\mathbf{y}}$$

normalised information matrix: $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\mathbf{F}^\top\mathbf{F}$

→ choose the x_i such that $\mathbf{M}_n$ has full rank

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

$$\Rightarrow \mathrm{E}\{\hat{\theta}^n\} = \bar{\theta}$$

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

→ choose the x_i to minimise a scalar function of $\mathbf{M}_n^{-1}$
or maximise a function $\Phi(\mathbf{M}_n)$ (information function (Pukelsheim, 1993))

Normal errors $\epsilon \sim \mathscr{N}(\mathbf{0}, \sigma^2\,\mathbf{I}_n)$ ➠ $\boxed{\hat{\theta}^n \sim \mathscr{N}(\bar{\theta}, \frac{\sigma^2}{n}\mathbf{M}^{-1}(\mathbf{X}_n))}$

→ no particular problem with *small data*

B. Nonlinear regression

$\eta(x, \theta)$ nonlinear in θ

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

$\boxed{\hat{\theta}^n \stackrel{\text{a.s.}}{\to} \bar{\theta} \text{ as } n \to \infty}$ (strong consistency) [but $\mathsf{E}\{\hat{\theta}^n\} \neq \bar{\theta}$ ($\hat{\theta}^n$ is biased)]

B. Nonlinear regression

$\eta(x,\theta)$ nonlinear in θ

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

$\boxed{\hat{\theta}^n \stackrel{\text{a.s.}}{\rightarrow} \bar{\theta} \text{ as } n \rightarrow \infty}$ (strong consistency) [but $\mathsf{E}\{\hat{\theta}^n\} \neq \bar{\theta}$ ($\hat{\theta}^n$ is biased)]

normalised information matrix <u>at θ</u>: $\mathbf{M}_n(\theta) = \mathbf{M}(\mathbf{X}_n,\theta) = \frac{1}{n}\sum_{i=1}^n \frac{\partial\eta(x_i,\theta)}{\partial\theta}\frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}$

Under «standard» regularity assumptions ($\eta(x,\theta)$ twice continuously differentiable w.r.t. θ for any $x\ldots$), for a suitable sequence (x_i),

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

with $\mathbf{M}(\theta) = \lim_{n\rightarrow\infty} \mathbf{M}_n(\theta)$

B. Nonlinear regression

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

$\boxed{\hat{\theta}^n \stackrel{\text{a.s.}}{\to} \bar{\theta} \text{ as } n \to \infty}$ (strong consistency) [but $\mathsf{E}\{\hat{\theta}^n\} \neq \bar{\theta}$ ($\hat{\theta}^n$ is biased)]

normalised information matrix at θ: $\mathbf{M}_n(\theta) = \mathbf{M}(\mathbf{X}_n, \theta) = \frac{1}{n} \sum_{i=1}^n \frac{\partial\eta(x_i,\theta)}{\partial\theta} \frac{\partial\eta(x_i,\theta)}{\partial\theta^\top}$

Under «standard» regularity assumptions ($\eta(x,\theta)$ twice continuously differentiable w.r.t. θ for any $x\ldots$), for a suitable sequence (x_i),

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

→ choose the x_i to minimise a scalar function of $\mathbf{M}_n^{-1}(\theta^0)$,
or maximise a function $\Phi(\mathbf{M}_n(\theta^0))$, for a prior guess θ^0 (local design)

= **classical approach for DoE in nonlinear models**
(based on asymptotic normality)

- 1) DoE for linear models (local design for nonlinear models, for a given θ^0)
  - Which information function Φ?
  - How to construct an optimal design for Φ?

- 1) DoE for linear models (local design for nonlinear models, for a given θ^0)
  - Which information function Φ?
  - How to construct an optimal design for Φ?
- 3,4,5,6) Small-sample issues

- 1) DoE for linear models (local design for nonlinear models, for a given θ^0)
  - Which information function Φ?
  - How to construct an optimal design for Φ?
- 3,4,5,6) Small-sample issues
- 7) nonlocal DoE for nonlinear models (based on asymptotic normality)

2 DoE for linear models

Design criterion Φ

- *A*-optimality: minimise $\mathrm{trace}[\mathbf{M}^{-1}] \Leftrightarrow$ maximise $\Phi(\mathbf{M}) = 1/\mathrm{trace}[\mathbf{M}^{-1}]$
  $\Leftrightarrow$ minimise sum of lengthes^2 of axes of (asymptotic) confidence ellipsoids

2 DoE for linear models

Design criterion Φ

- *A*-optimality: minimise $\mathrm{trace}[\mathbf{M}^{-1}] \Leftrightarrow$ maximise $\Phi(\mathbf{M}) = 1/\mathrm{trace}[\mathbf{M}^{-1}]$
  $\Leftrightarrow$ minimise sum of lengthes^2 of axes of (asymptotic) confidence ellipsoids
- *E*-optimality: maximise $\Phi(\mathbf{M}) = \lambda_{\mathrm{min}}(\mathbf{M})$
  $\Leftrightarrow$ minimise longest axis of (asymptotic) confidence ellipsoids

2 DoE for linear models

Design criterion Φ

- A-optimality: minimise $\mathrm{trace}[\mathbf{M}^{-1}] \Leftrightarrow$ maximise $\Phi(\mathbf{M}) = 1/\mathrm{trace}[\mathbf{M}^{-1}]$
  $\Leftrightarrow$ minimise sum of lengthes^2 of axes of (asymptotic) confidence ellipsoids
- *E*-optimality: maximise $\Phi(\mathbf{M}) = \lambda_{\min}(\mathbf{M})$
  $\Leftrightarrow$ minimise longest axis of (asymptotic) confidence ellipsoids
- *D*-optimality: maximise $\Phi(\mathbf{M}) = \det^{1/p}(\mathbf{M})$ $[p = \dim(\theta)]$
  $\Leftrightarrow$ minimise volume of (asymptotic) confidence ellipsoids
  (proportional to $1/\sqrt{\det(\mathbf{M})}$)
  Very much used:
  - a D-optimum design is invariant by reparameterisation

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

  - often leads to repeat the same experimental conditions (replications)

Construction of an optimal design

A/ Exact design

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

Maximise $\Phi(\mathbf{M}_n)$ w.r.t. $\mathbf{X}_n$ with $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{f}(x_i)\mathbf{f}^\top(x_i)$

Construction of an optimal design

A/ Exact design

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

Maximise $\Phi(\mathbf{M}_n)$ w.r.t. $\mathbf{X}_n$ with $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{f}(x_i)\mathbf{f}^\top(x_i)$

- If $n \times d$ not too large $\rightarrow$ standard algorithm
  [but there exist constraints ($x_i \in \mathscr{X}$ for all i), local optimas. . .]

Construction of an optimal design

A/ Exact design

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

Maximise $\Phi(\mathbf{M}_n)$ w.r.t. $\mathbf{X}_n$ with $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{f}(x_i)\mathbf{f}^\top(x_i)$

- If $n \times d$ not too large ➞ standard algorithm

  [but there exist constraints ($x_i \in \mathscr{X}$ for all i), local optimas…]
- Otherwise ➞ take the particular form of the problem into account

Exchange methods: (Fedorov, 1972; Mitchell, 1974)
At iteration k, exchange **one** *support point* x_j by a better one x^* in $\mathscr{X}$ in the sense of $\Phi(\cdot)$

$$\mathbf{X}_n^k = (x_1, \dots, \boxed{\begin{array}{c} x_j \\ \updownarrow \\ x^* \end{array}}, \dots, x_n)$$

Construction of an optimal design

A/ Exact design

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

Maximise $\Phi(\mathbf{M}_n)$ w.r.t. $\mathbf{X}_n$ with $\mathbf{M}_n = \mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}(x_i)\mathbf{f}^\top(x_i)$

- If $n \times d$ not too large ➞ standard algorithm
  [but there exist constraints ($x_i \in \mathscr{X}$ for all i), local optimas...]
- Otherwise ➞ take the particular form of the problem into account

Exchange methods: (Fedorov, 1972; Mitchell, 1974)
At iteration k, exchange **one** *support point* x_j by a better one x^* in $\mathscr{X}$ in the sense of $\Phi(\cdot)$

$$\mathbf{X}_n^k = (x_1, \ldots, \boxed{\begin{array}{c} x_j \\ \updownarrow \\ x^* \end{array}}, \ldots, x_n)$$

- Branch and bound (Welch, 1982), rounding an optimal design measure (Pukelsheim and Reider, 1992)

B/ Design measures: approximate design theory

(Chernoff, 1953; Kiefer and Wolfowitz, 1960; Fedorov, 1972; Silvey, 1980; Pázman, 1986; Pukelsheim, 1993; Fedorov and Leonov, 2014)

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^n \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

[with $\mathbf{M}(\mathbf{X}_n) = \mathbf{M}(\mathbf{X}_n, \theta^0)$ and $\mathbf{f}(x_i) = \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0}$ in a nonlinear model]
The additive form is essential (comes from the independence of observations)

B/ Design measures: approximate design theory

(Chernoff, 1953; Kiefer and Wolfowitz, 1960; Fedorov, 1972; Silvey, 1980; Pázman, 1986; Pukelsheim, 1993; Fedorov and Leonov, 2014)

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

[with $\mathbf{M}(\mathbf{X}_n) = \mathbf{M}(\mathbf{X}_n, \theta^0)$ and $\mathbf{f}(x_i) = \frac{\partial\eta(x_i,\theta)}{\partial\theta}\Big|_{\theta^0}$ in a nonlinear model]
The additive form is essential (comes from the independence of observations)

Repeat r_i observations at the same $x_i \in \mathscr{X}$ (r_i replications):
→ only $m \leq n$ different x_i

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \sum_{i=1}^{m} \frac{r_i}{n}\, \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

$$\begin{aligned}\frac{r_i}{n} &= \text{proportion of observations collected at } x_i\\ &= \text{«percentage of experimental effort» at } x_i\\ &= \text{weight } w_i \text{ of support point } x_i\end{aligned}$$

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \sum_{i=1}^{m} w_i \, \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

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

➔ normalised discrete distribution on $\mathscr{X}$, with $\boxed{\text{constraints } r_i/n = w_i}$

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \sum_{i=1}^{m} w_i \, \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

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

➔ normalised discrete distribution on $\mathscr{X}$, with $\boxed{\text{constraints } r_i/n = w_i}$

Release the constraints ➞ $w_i \geq 0$ with $\sum_{i=1}^{m} w_i = 1$

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

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \sum_{i=1}^m w_i\, \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

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

➔ normalised discrete distribution on $\mathscr{X}$, with $\boxed{\text{constraints } r_i/n = w_i}$

Release the constraints ➞ $w_i \geq 0$ with $\sum_{i=1}^m w_i = 1$

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

More general expression: ξ = any probability measure on $\mathscr{X}$ $(\int_{\mathscr{X}} \xi(\mathrm{d}x) = 1)$

$$\boxed{\mathbf{M}(\xi) = \int_{\mathscr{X}} \mathbf{f}(x)\mathbf{f}^\top(x)\, \xi(\mathrm{d}x)}$$

$$\boxed{\mathbf{M}(\mathbf{X}_n) = \sum_{i=1}^m w_i\, \mathbf{f}(x_i)\mathbf{f}^\top(x_i)}$$

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

➔ normalised discrete distribution on $\mathscr{X}$, with $\boxed{\text{constraints } r_i/n = w_i}$

Release the constraints ➞ $w_i \geq 0$ with $\sum_{i=1}^m w_i = 1$

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

More general expression: ξ = any probability measure on $\mathscr{X}$ ($\int_{\mathscr{X}} \xi(\mathrm{d}x) = 1$)

$$\boxed{\mathbf{M}(\xi) = \int_{\mathscr{X}} \mathbf{f}(x)\mathbf{f}^\top(x)\, \xi(\mathrm{d}x)}$$

$\mathbf{M}(\xi) \in$ convex closure of the set of rank 1 matrices $\mathbf{f}(x)\mathbf{f}^\top(x)$
$\mathbf{M}(\xi)$ is symmetric $p \times p$, belongs to a $\frac{p(p+1)}{2}$-dimensional space
Caratheodory Theorem ➞ for any ξ, there exists a discrete probability measure ξ_d with $\frac{p(p+1)}{2} + 1$ support points at most, such that $\mathbf{M}(\xi_d) = \mathbf{M}(\xi)$
(true in particular for the optimum design)

Maximise $\Phi[\mathbf{M}(\xi)]$, $\Phi(\cdot)$ concave (e.g., A, E, D-optimality) and $\mathbf{M}(\xi)$ linear in ξ

→ convex programming

Usually, $\mathscr{X}$ is first discretised

→ optimise a vector of weights
(possibly high dimensional, but the solution is sparse)

Typical algorithm when Φ is differentiable (A, D-optimality):
Frank-Wolfe conditional gradient (called vertex-direction algorithm in DoE), with predefined (Wynn, 1970) or optimal (Fedorov, 1972) step-size

[but there exist more efficient methods]

More difficult if Φ not differentiable (E-optimality), but feasible

Application to models with complete product-type interactions

Single factor models: $\eta_k(x,\theta^{(k)}) \triangleq [\mathbf{f}^{(k)}(x)]^\top \theta^{(k)}$

global model for d factors $\mathbf{x} = (\{\mathbf{x}\}_1, \{\mathbf{x}\}_2, \ldots, \{\mathbf{x}\}_d)^\top$:

$$\eta(\mathbf{x},\gamma) = [\mathbf{f}_1(\{\mathbf{x}\}_1) \otimes \cdots \otimes \mathbf{f}_d(\{\mathbf{x}\}_d)]^\top \gamma$$

Application to models with complete product-type interactions

Single factor models: $\eta_k(x,\theta^{(k)}) \triangleq [\mathbf{f}^{(k)}(x)]^\top \theta^{(k)}$

global model for d factors $\mathbf{x} = (\{\mathbf{x}\}_1, \{\mathbf{x}\}_2, \ldots, \{\mathbf{x}\}_d)^\top$:

$$\eta(\mathbf{x},\gamma) = [\mathbf{f}_1(\{\mathbf{x}\}_1) \otimes \cdots \otimes \mathbf{f}_d(\{\mathbf{x}\}_d)]^\top \gamma$$

In particular, if η_k = polynomial of degree d_k $(\dim(\theta^{(k)}) = p_k = 1 + d_k)$,

$$\eta = \text{polynomial with total degree } \sum_{k=1}^d d_k \; (\dim(\gamma) = \prod_{k=1}^d p_k)$$

Example:

$$\begin{aligned}
\mathbf{f}^\top(\mathbf{x})\gamma &= (\theta_0^{(1)} + \theta_1^{(1)}\{\mathbf{x}\}_1 + \theta_2^{(1)}\{\mathbf{x}\}_1^2) \times (\theta_0^{(2)} + \theta_1^{(2)}\{\mathbf{x}\}_2 + \theta_2^{(2)}\{\mathbf{x}\}_2^2) \\
&= \gamma_0 + \gamma_1\{\mathbf{x}\}_1 + \gamma_2\{\mathbf{x}\}_2 + \gamma_{12}\{\mathbf{x}\}_1\{\mathbf{x}\}_2 + \gamma_{11}\{\mathbf{x}\}_1{}^2 + \gamma_{22}\{\mathbf{x}\}_2{}^2 \\
&\quad + \gamma_{112}\{\mathbf{x}\}_1{}^2\{\mathbf{x}\}_2 + \gamma_{122}\{\mathbf{x}\}_1\{\mathbf{x}\}_2{}^2 + \gamma_{1122}\{\mathbf{x}\}_1{}^2\{\mathbf{x}\}_2{}^2
\end{aligned}$$

Application to models with complete product-type interactions

Single factor models: $\eta_k(x, \theta^{(k)}) \triangleq [\mathbf{f}^{(k)}(x)]^\top \theta^{(k)}$

global model for d factors $\mathbf{x} = (\{\mathbf{x}\}_1, \{\mathbf{x}\}_2, \ldots, \{\mathbf{x}\}_d)^\top$:

$$\eta(\mathbf{x}, \gamma) = [\mathbf{f}_1(\{\mathbf{x}\}_1) \otimes \cdots \otimes \mathbf{f}_d(\{\mathbf{x}\}_d)]^\top \gamma$$

In particular, if η_k = polynomial of degree d_k $(\dim(\theta^{(k)}) = p_k = 1 + d_k)$,

$$\eta = \text{polynomial with total degree } \sum_{k=1}^d d_k \ (\dim(\gamma) = \prod_{k=1}^d p_k)$$

Example:

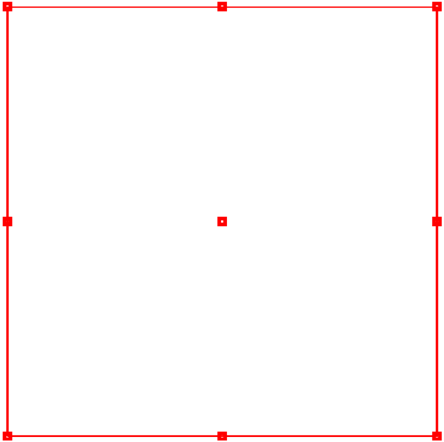
$$\begin{aligned}
\mathbf{f}^\top(\mathbf{x})\gamma &= (\theta_0^{(1)} + \theta_1^{(1)}\{\mathbf{x}\}_1 + \theta_2^{(1)}\{\mathbf{x}\}_1^2) \times (\theta_0^{(2)} + \theta_1^{(2)}\{\mathbf{x}\}_2 + \theta_2^{(2)}\{\mathbf{x}\}_2^2) \\
&= \gamma_0 + \gamma_1\{\mathbf{x}\}_1 + \gamma_2\{\mathbf{x}\}_2 + \gamma_{12}\{\mathbf{x}\}_1\{\mathbf{x}\}_2 + \gamma_{11}{\{\mathbf{x}\}_1}^2 + \gamma_{22}{\{\mathbf{x}\}_2}^2 \\
&\quad + \gamma_{112}{\{\mathbf{x}\}_1}^2\{\mathbf{x}\}_2 + \gamma_{122}\{\mathbf{x}\}_1{\{\mathbf{x}\}_2}^2 + \gamma_{1122}{\{\mathbf{x}\}_1}^2{\{\mathbf{x}\}_2}^2
\end{aligned}$$

D, *A* and *E*-optimal design measure = tensor product of the d optimal design measures (Schwabe, 1996)

(true for any complete product-type interaction model — not only for polynomials)

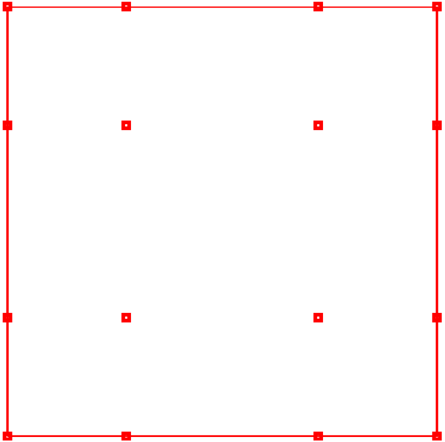
Polynomial with degree k: D-optimal design supported on $k+1$ points
(on $[-1,1]$: roots of $(1-t^2)P_k'(t)$, with $P_k(t) \triangleq k$-th de Legendre polynomial),
all with the same weight $1/(k+1)$

dimension 2, $d_1 = d_2 = 2$
ξ^* has 9 support points, weights $=1/9$



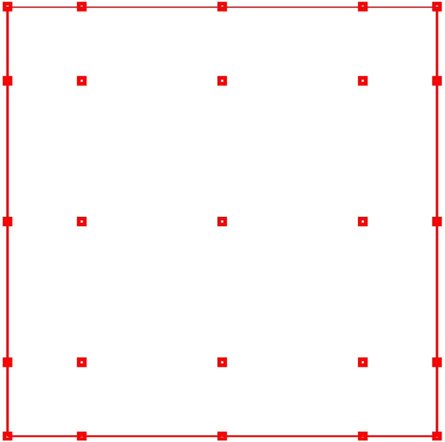
Polynomial with degree k: D-optimal design supported on $k+1$ points
(on $[-1,1]$: roots of $(1-t^2)P_k'(t)$, with $P_k(t) \triangleq$ k-th de Legendre polynomial),
all with the same weight $1/(k+1)$

dimension 2, $d_1 = d_2 = 3$
ξ^* has 16 support points, weights $=1/16$



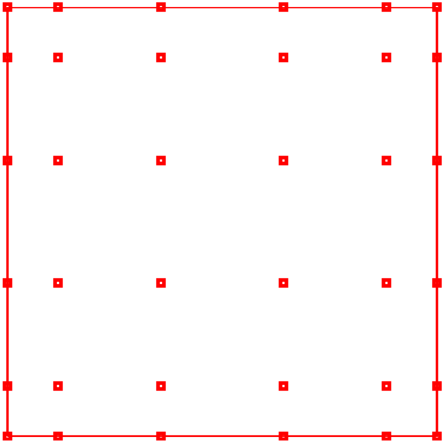
Polynomial with degree k: D-optimal design supported on $k+1$ points
(on $[-1,1]$: roots of $(1-t^2)P_k'(t)$, with $P_k(t) \triangleq k$-th de Legendre polynomial),
all with the same weight $1/(k+1)$

dimension 2, $d_1 = d_2 = 4$
ξ^* has 25 support points, weights $=1/25$



Polynomial with degree k: D-optimal design supported on $k+1$ points
(on $[-1,1]$: roots of $(1-t^2)P_k'(t)$, with $P_k(t) \triangleq$ k-th de Legendre polynomial),
all with the same weight $1/(k+1)$

dimension 2, $d_1 = d_2 = 5$
ξ^* has 36 support points, weights $=1/36$



Application to models with intercept, no interaction

Single factor models: $\eta_k(x,\theta^{(k)}) \triangleq \theta_0^{(k)} + \sum_{i=1}^{d_k} \theta_i^{(k)} f_i^{(k)}(x)$

global model for d factors: $\eta(\mathbf{x},\gamma) = \theta_0 + \sum_{k=1}^{d} \sum_{i=1}^{d_k} \theta_i^{(k)} f_i^{(k)}(\{\mathbf{x}\}_k)$

Application to models with intercept, no interaction

Single factor models: $\eta_k(x,\theta^{(k)}) \triangleq \theta_0^{(k)} + \sum_{i=1}^{d_k} \theta_i^{(k)} f_i^{(k)}(x)$

global model for d factors: $\eta(\mathbf{x},\gamma) = \theta_0 + \sum_{k=1}^{d} \sum_{i=1}^{d_k} \theta_i^{(k)} f_i^{(k)}(\{\mathbf{x}\}_k)$

In particular, if η_k = polynomial of degree d_k $(\dim(\theta^{(k)}) = p_k = 1 + d_k)$,

η = polynomial with total degree $\max_k^d d_k$ $(\dim(\gamma) = 1 + \sum_{k=1}^{d} d_k)$

Example:

$$
\begin{aligned}
\mathbf{f}^\top(\mathbf{x})\gamma &= (\theta_0^{(1)} + \theta_1^{(1)}\{\mathbf{x}\}_1 + \theta_2^{(1)}\{\mathbf{x}\}_1^2) + (\theta_0^{(2)} + \theta_1^{(2)}\{\mathbf{x}\}_2 + \theta_2^{(2)}\{\mathbf{x}\}_2^2) \\
&= \gamma_0 + \gamma_1\{\mathbf{x}\}_1 + \gamma_2\{\mathbf{x}\}_2 + \gamma_{11}\{\mathbf{x}\}_1{}^2 + \gamma_{22}\{\mathbf{x}\}_2{}^2
\end{aligned}
$$

D-optimal design measure = tensor product of d D-optimal measures (Schwabe, 1996)

Application to models with intercept, no interaction

Single factor models: $\eta_k(x,\theta^{(k)}) \triangleq \theta_0^{(k)} + \sum_{i=1}^{d_k} \theta_i^{(k)} f_i^{(k)}(x)$

global model for d factors: $\eta(\mathbf{x},\gamma) = \theta_0 + \sum_{k=1}^{d} \sum_{i=1}^{d_k} \theta_i^{(k)} f_i^{(k)}(\{\mathbf{x}\}_k)$

In particular, if η_k = polynomial of degree d_k $(\dim(\theta^{(k)}) = p_k = 1 + d_k)$,

η = polynomial with total degree $\max_k^d d_k$ $(\dim(\gamma) = 1 + \sum_{k=1}^{d} d_k)$

Example:

$$\begin{aligned} \mathbf{f}^\top(\mathbf{x})\gamma &= (\theta_0^{(1)} + \theta_1^{(1)}\{\mathbf{x}\}_1 + \theta_2^{(1)}\{\mathbf{x}\}_1^2) + (\theta_0^{(2)} + \theta_1^{(2)}\{\mathbf{x}\}_2 + \theta_2^{(2)}\{\mathbf{x}\}_2^2) \\ &= \gamma_0 + \gamma_1\{\mathbf{x}\}_1 + \gamma_2\{\mathbf{x}\}_2 + \gamma_{11}\{\mathbf{x}\}_1{}^2 + \gamma_{22}\{\mathbf{x}\}_2{}^2 \end{aligned}$$

D-optimal design measure = tensor product of d D-optimal measures (Schwabe, 1996)

Hardly manageable in high dimension

(d polynomials of degree k ➠ $(k+1)^d$ support points),

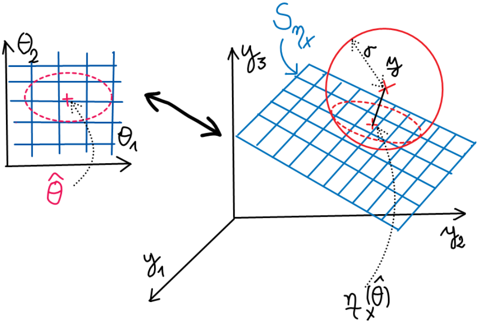
but maybe a useful message for Gaussian Process models and kriging:

➔ put more points along the boundaries than deeply inside

(Dette and Pepelyshev, 2010)

3 Linear and nonlinear models

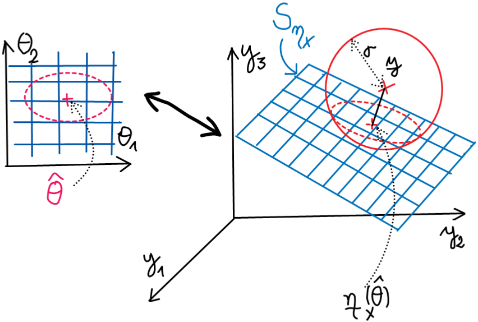
Linear regression: easy



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

3 Linear and nonlinear models

Linear regression: easy

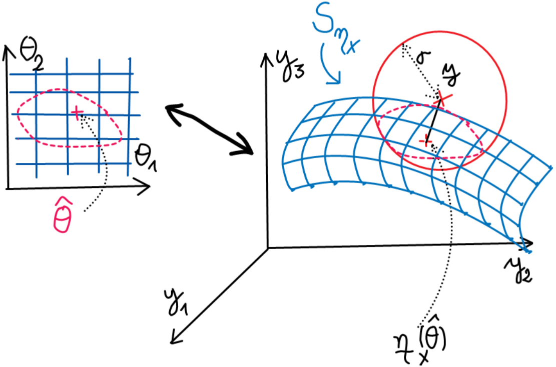


The expectation surface $\mathbb{S}_\eta = \{\boldsymbol{\eta}(\theta) = (\eta(x_1, \theta), \ldots, \eta(x_n, \theta))^\top : \theta \in \mathbb{R}^p\}$ is flat and linearly parameterised

$\mathbf{M}(\mathbf{X}_n, \theta)$ does not depend on θ

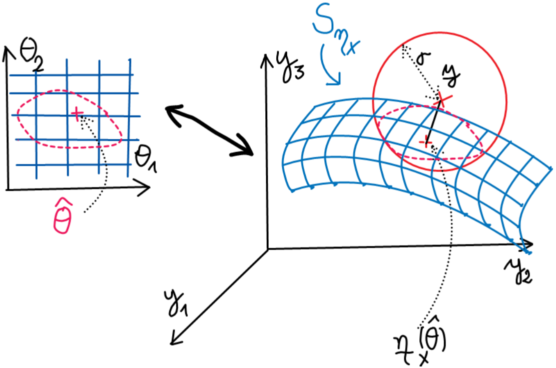
Normal errors $\epsilon \sim \mathscr{N}(\mathbf{0}, \sigma^2\, \mathbf{I}_n)$ ➠ $\boxed{\hat{\theta}^n \sim \mathscr{N}(\bar{\theta}, \frac{\sigma^2}{n}\mathbf{M}^{-1}(\mathbf{X}_n))}$

Nonlinear regression: maybe a bit tricky...



$\mathbb{S}_\eta$ is curved (intrinsic curvature) and nonlinearly parameterised (parametric curvature) (Bates and Watts, 1980)

Nonlinear regression: maybe a bit tricky...

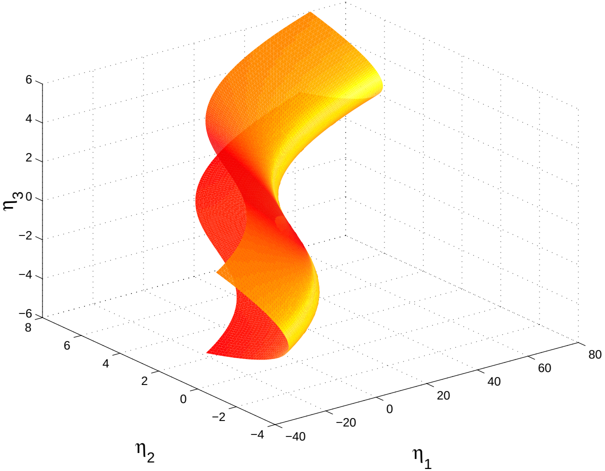


$\mathbb{S}_\eta$ is curved (intrinsic curvature) and nonlinearly parameterised (parametric curvature) (Bates and Watts, 1980)

$\mathbf{M}(\mathbf{X}_n, \theta)$ does depend on θ

Normal errors $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ ➠ $\boxed{\hat\theta^n \sim \mathbf{?}}$

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



Two major difficulties with nonlinear models:

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

use linear approximation

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

➠ small-sample properties

Two major difficulties with nonlinear models:

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

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 linearise, where do we linearise?
(choice of a nominal value θ^0)
➠ nonlocal optimum design

4 Small-sample properties

Asymptotically ($n \to \infty$) ➠ $\boxed{\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{M}^{-1}(\mathbf{X}_n, \bar{\theta}))}$

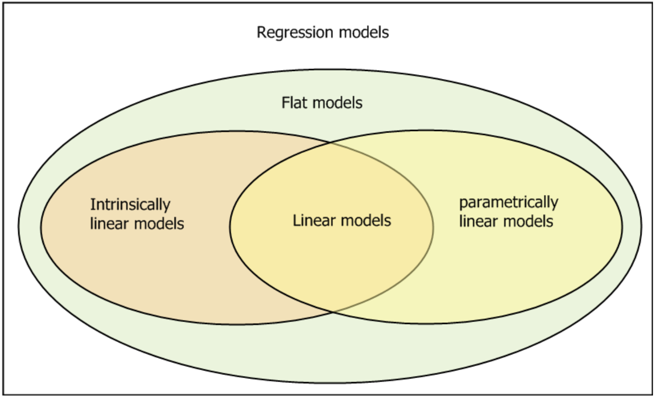
but what is the small sample precision?

4 Small-sample properties

Asymptotically ($n \to \infty$) ➠ $\boxed{\sqrt{n}(\hat{\theta}^n - \bar{\theta}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{M}^{-1}(\mathbf{X}_n, \bar{\theta}))}$

but what is the small sample precision?

A classification of regression models (Pázman, 1993)



➔ Consider projection on the expectation surface $\mathbb{S}_\eta$:

➤ $\mathbf{P}_{\theta^0}$ = orthogonal projector onto the tangent space to $\mathbb{S}_\eta$ at $\boldsymbol{\eta}(\theta^0)$:

$$\mathbf{P}_{\theta^0} = \frac{1}{n} \frac{\partial \boldsymbol{\eta}(\theta)}{\partial \theta^\top}\Big|_{\theta^0} \mathbf{M}^{-1}(\mathbf{X}_n, \theta^0) \frac{\partial \boldsymbol{\eta}^\top(\theta)}{\partial \theta}\Big|_{\theta^0}$$

(an $n \times n$ matrix, depends on $\mathbf{X}_n$)
Bates and Watts (1980) intrinsic and parametric-effect measures of nonlinearity:

$$\begin{aligned} C_{int}(\mathbf{X}_n, \theta; \mathbf{u}) &= \frac{\|[\mathbf{I}_n - \mathbf{P}_\theta] \sum_{i,j=1}^p u_i \mathbf{H}_{\cdot ij}(\theta) u_j\|}{n\, \mathbf{u}^\top \mathbf{M}(\mathbf{X}_n, \theta)\mathbf{u}} \\ C_{par}(\mathbf{X}_n, \theta; \mathbf{u}) &= \frac{\|\mathbf{P}_\theta \sum_{i,j=1}^p u_i \mathbf{H}_{\cdot ij}(\theta) u_j\|}{n\, \mathbf{u}^\top \mathbf{M}(\mathbf{X}_n, \theta)\mathbf{u}} \end{aligned}$$

with $\mathbf{u} \in \mathbb{R}^p$ and $\mathbf{H}_{\cdot ij}(\theta) = \frac{\partial^2 \boldsymbol{\eta}(\theta)}{\partial \theta_i \partial \theta_j}$

Intrinsic curvature: $C_{int}(\mathbf{X}_n, \theta) = \sup_{\mathbf{u} \in \mathbb{R}^p \setminus \{\mathbf{0}\}} C_{int}(\mathbf{X}_n, \theta; \mathbf{u})$
Parametric curvature: $C_{par}(\mathbf{X}_n, \theta) = \sup_{\mathbf{u} \in \mathbb{R}^p \setminus \{\mathbf{0}\}} C_{par}(\mathbf{X}_n, \theta; \mathbf{u})$

Intrinsically linear models

- The expectation surface $\mathbb{S}_\eta = \{\boldsymbol{\eta}(\theta) : \theta \in \mathbb{R}^p\}$ is flat (plane) — intrinsic curvature $\equiv 0$
- There exists a reparameterisation (continuously differentiable) that makes the model linear
- $\mathbf{P}_\theta \mathbf{H}_{\cdot ij}(\theta) = \mathbf{H}_{\cdot ij}(\theta)$, where $\mathbf{H}_{\cdot ij}(\theta) = \frac{\partial^2 \boldsymbol{\eta}(\theta)}{\partial\theta_i \partial\theta_j}$

Intrinsically linear models

- The expectation surface $\mathbb{S}_\eta = \{\boldsymbol{\eta}(\theta) : \theta \in \mathbb{R}^p\}$ is flat (plane)
  — intrinsic curvature $\equiv 0$
- There exists a reparameterisation (continuously differentiable) that makes the model linear
- $\mathbf{P}_\theta \mathbf{H}_{ij}^{\cdot\cdot}(\theta) = \mathbf{H}_{ij}^{\cdot\cdot}(\theta)$, where $\mathbf{H}_{ij}^{\cdot\cdot}(\theta) = \frac{\partial^2 \boldsymbol{\eta}(\theta)}{\partial\theta_i \partial\theta_j}$

Observing at p different x_i only (replications) makes the model intrinsically linear
$[p = \dim(\theta)]$

Parametrically linear models

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

Parametrically linear models

- $\mathbf{M}(\mathbf{X}_n, \theta) =$ 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

Parametrically linear models

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

Linear models

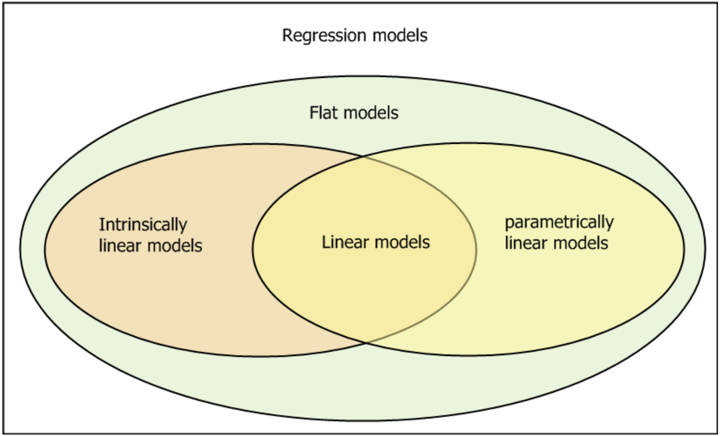
- $\eta(x, \theta) = \mathbf{f}^\top(x)\theta + c(x)$
- the model is intrinsically and parametrically linear

Flat models

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

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

A classification of regression models (Pázman, 1993)



Density of the LS estimator (we suppose $\varepsilon \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$)

<u>**Intrinsically linear models**</u> (in particular, repetitions at p points):

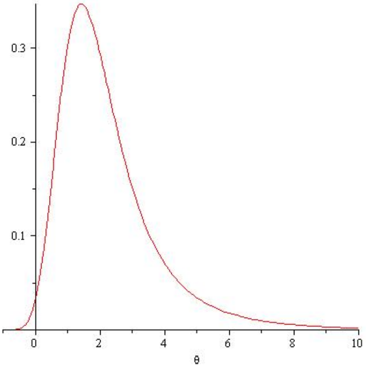
$\rightarrow$ exact distribution $\boxed{\hat{\theta}^n \sim q(\theta|\bar{\theta}) = \frac{n^{p/2} \det^{1/2} \mathbf{M}(\mathbf{X}_n, \theta)}{(2\pi)^{p/2}\, \sigma^p} \exp\left\{-\frac{1}{2\sigma^2} \|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\bar{\theta})\|^2\right\}}$

Density of the LS estimator (we suppose $\varepsilon \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$)

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

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

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

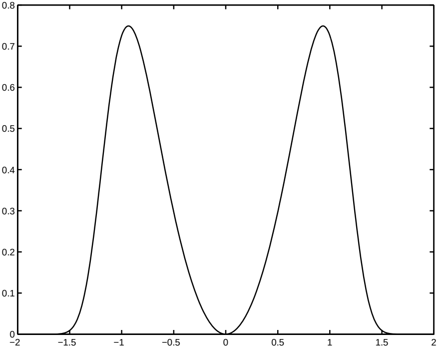


Density of the LS estimator (we suppose $\varepsilon \sim \mathscr{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$)

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

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

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



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

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

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

There exists other approximations (more complicated) for models with $R_{hijk}(\theta) \not\equiv 0$ (non-flat)

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

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

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

There exists other approximations (more complicated) for models with $R_{hijk}(\theta) \not\equiv 0$ (non-flat)

➡ Design of experiments? (since $q(\theta|\bar{\theta})$ depends on $\mathbf{X}_n$)

5 DOE based on small sample precision

(P & Pázman, 2013, Chap. 6)

1) Minimise the MSE $\mathrm{E}\{\|\hat{\theta}^n(\mathbf{y}) - \bar{\theta}\|^2\}$

The approximation of Clarke (1980) requires the 4th-order derivatives of $\boldsymbol{\eta}(\theta)$

5 DOE based on small sample precision

(P & Pázman, 2013, Chap. 6)

1) Minimise the MSE $\mathrm{E}\{\|\hat{\theta}^n(\mathbf{y}) - \bar{\theta}\|^2\}$

The approximation of Clarke (1980) requires the 4th-order derivatives of $\boldsymbol{\eta}(\theta)$

→ Use the (approximate) density $q(\theta|\bar{\theta})$

→ $\boxed{\text{Minimise } \int_\Theta \|\theta - \bar{\theta}\|^2 q(\theta|\bar{\theta})\, \mathrm{d}\theta \text{ w.r.t. } \mathbf{X}_n}$ using stochastic approximation

5 DOE based on small sample precision

(P & Pázman, 2013, Chap. 6)

1) Minimise the MSE $\mathrm{E}\{\|\hat{\theta}^n(\mathbf{y}) - \bar{\theta}\|^2\}$

The approximation of Clarke (1980) requires the 4th-order derivatives of $\boldsymbol{\eta}(\theta)$

→ Use the (approximate) density $q(\theta|\bar{\theta})$

→ $\boxed{\text{Minimise } \int_\Theta \|\theta - \bar{\theta}\|^2 q(\theta|\bar{\theta})\, \mathrm{d}\theta \text{ w.r.t. } \mathbf{X}_n}$ using stochastic approximation

Problem: we need to force θ to remain in Θ

→ the integral can be made equal to 0

5 DOE based on small sample precision

(P & Pázman, 2013, Chap. 6)

1) Minimise the MSE $\mathrm{E}\{\|\hat{\theta}^n(\mathbf{y}) - \bar{\theta}\|^2\}$

The approximation of Clarke (1980) requires the 4th-order derivatives of $\boldsymbol{\eta}(\theta)$

→ Use the (approximate) density $q(\theta|\bar{\theta})$

→ $\boxed{\text{Minimise } \int_\Theta \|\theta - \bar{\theta}\|^2 q(\theta|\bar{\theta})\, \mathrm{d}\theta \text{ w.r.t. } \mathbf{X}_n}$ using stochastic approximation

Problem: we need to force θ to remain in Θ

→ the integral can be made equal to 0

Solution: approximate the density $\tilde{q}_w(\theta|\bar{\theta})$ of a penalised LS estimator $\tilde{\theta}^n$

$\tilde{\theta}^n = \arg\min_\theta \left\{\|\mathbf{y} - \boldsymbol{\eta}(\theta)\|^2 + 2w(\theta)\right\}$

where $w(\theta)$ forces θ to remain in Θ [$w(\theta) = +\infty$ outside Θ]

→ $\boxed{\text{Minimise } \int_\Theta \|\theta - \bar{\theta}\|^2 \tilde{q}_w(\theta|\bar{\theta})\, \mathrm{d}\theta \text{ w.r.t. } \mathbf{X}_n}$

[also covers the case of max. *a posteriori* estimation (relate $w(\theta)$ to the prior on θ)]
(P & Pázman, 1992; Pázman and Gauchi, 2006)

2) Use a small-sample variant of D-optimal design

A D-optimal design minimises

(*i*) the volume of asymptotic (ellipsoidal) confidence regions

(*ii*) the (Shannon) entropy of the asymptotic distribution of $\hat{\theta}^n$

2) Use a small-sample variant of *D*-optimal design

A *D*-optimal design minimises

(*i*) the volume of asymptotic (ellipsoidal) confidence regions

(*ii*) the (Shannon) entropy of the asymptotic distribution of $\hat{\theta}^n$

Hamilton and Watts (1985) minimize the (approximate) volume $V(\mathbf{X}_n, \theta^0)$ of (approximate) confidence regions ($V(\mathbf{X}_n, \theta^0)$ has an explicit form and a geometrical interpretation)

Vila (1990); Vila and Gauchi (2007) minimize the expected volume of exact confidence regions (not ellipsoidal, not necessarily of minimum volume), using stochastic approximation

→ Choose $\mathbf{X}_n$ that minimises the approximate entropy of the approximate distribution of $\hat{\theta}^n$ (P & Pázman, 1994b)

$$\boxed{\text{Minimise } \mathrm{Ent}[q(\cdot|\bar{\theta})] = -\int_{\mathbb{R}^n} \log[q(\hat{\theta}^n(\mathbf{y})|\bar{\theta})]\varphi(\mathbf{y}|\mathbf{X}_n,\bar{\theta})\,\mathrm{d}\mathbf{y} \text{ w.r.t. } \mathbf{X}_n}$$

where $\varphi(\mathbf{y}|\mathbf{X}_n,\bar{\theta})$ corresponds to $\mathbf{y} \sim \mathscr{N}(\boldsymbol{\eta}(\bar{\theta}),\sigma^2\mathbf{I}_n)$

→ Choose $\mathbf{X}_n$ that minimises the approximate entropy of the approximate distribution of $\hat{\theta}^n$ (P & Pázman, 1994b)

$$\boxed{\text{Minimise } \mathrm{Ent}[q(\cdot|\bar{\theta})] = -\int_{\mathbb{R}^n} \log[q(\hat{\theta}^n(\mathbf{y})|\bar{\theta})]\varphi(\mathbf{y}|\mathbf{X}_n,\bar{\theta})\,\mathrm{d}\mathbf{y} \text{ w.r.t. } \mathbf{X}_n}$$

where $\varphi(\mathbf{y}|\mathbf{X}_n,\bar{\theta})$ corresponds to $\mathbf{y} \sim \mathscr{N}(\boldsymbol{\eta}(\bar{\theta}),\sigma^2\mathbf{I}_n)$

Use a 2nd order Taylor development of $\log[q(\hat{\theta}^n(\mathbf{y})|\bar{\theta})]$ around $\mathbf{y} = \boldsymbol{\eta}(\bar{\theta})$:

$$\mathrm{Ent}[q(\cdot|\bar{\theta})] = -\log q(\bar{\theta}|\bar{\theta}) - \frac{\sigma^2}{2}\sum_{i=1}^{N} \frac{\partial^2 \log q[\hat{\theta}(\mathbf{y})|\bar{\theta}]}{\partial y_i^2}\bigg|_{\boldsymbol{\eta}(\bar{\theta})} + \mathcal{O}(\sigma^4)$$

After some (lengthy) calculations...

$$\begin{aligned}
\mathrm{Ent}[q(\cdot|\bar{\theta})] &= \overbrace{\frac{p}{2}[1+\log(2\pi\sigma^2)] - \frac{1}{2}\log\det[n\mathbf{M}(\mathbf{X}_n,\bar{\theta})]}^{\text{entropy of asymptotic normal distribution}} \\
&- \frac{\sigma^2}{2n}\sum_{h,i,j,k=1}^{p}\Bigg(\{\mathbf{M}^{-1}(\mathbf{X}_n,\bar{\theta})\}_{ij}\Bigg[\frac{1}{n}\{\mathbf{M}^{-1}(\mathbf{X}_n,\bar{\theta})\}_{kh}[R_{kjhi}(\bar{\theta})+U^h_{kij}(\bar{\theta})] \\
&- G^h_{ki}(\bar{\theta})\,G^k_{hj}(\bar{\theta}) - G^k_{kh}(\bar{\theta})\,G^h_{ij}(\bar{\theta})\Bigg]\Bigg) + \mathcal{O}(\sigma^4)
\end{aligned}$$

$$\text{where } U^h_{kij}(\theta) = \frac{\partial^3\boldsymbol{\eta}^\top(\theta)}{\partial\theta_k\partial\theta_i\partial\theta_j}\,\frac{\partial\boldsymbol{\eta}(\theta)}{\partial\theta_h}$$

$$G^k_{ij}(\theta) = \frac{1}{n}\sum_{h=1}^{p}\frac{\partial\boldsymbol{\eta}^\top(\theta)}{\partial\theta_h}\,\mathbf{H}_{\cdot ij}\,\{\mathbf{M}^{-1}(\mathbf{X}_n,\bar{\theta})\}_{hk}$$

with $R_{hijk}(\theta) = T_{hjik}(\theta) - T_{hkij}(\theta)$, $T_{hjik}(\theta) = [\mathbf{H}_{\cdot hj}(\theta)]^\top[\mathbf{I}_n - \mathbf{P}_\theta]\mathbf{H}_{\cdot ik}(\theta)$ and $\mathbf{H}_{\cdot ij}(\theta) = \frac{\partial^2\boldsymbol{\eta}(\theta)}{\partial\theta_i\partial\theta_j}$

3) Related work using the approximate density $q(\theta|\bar{\theta})$

3a) (approximate) marginal densities of $\hat{\theta}^n$ (Pázman & P, 1996)

Denote $\gamma = h(\theta)$ [with $\gamma = \theta_i$ for some $i \in \{1, \ldots, p = \dim(\theta)\}$ as particular case]

$$\boxed{q(\gamma|\bar{\theta}) = \frac{1}{\sqrt{2\pi}\sigma\|\mathbf{b}_\gamma\|}\ \exp\left\{-\frac{1}{2\sigma^2}\|\mathbf{P}_\gamma[\boldsymbol{\eta}(\theta_\gamma) - \boldsymbol{\eta}(\bar{\theta})]\|^2\right\}}$$

where

$$\begin{aligned}
\theta_\gamma &= \arg\min_{\theta:h(\theta)=\gamma} \|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\bar{\theta})\|^2 \\
\mathbf{b}_\gamma &= \frac{1}{n}\,\frac{\partial\boldsymbol{\eta}(\theta)}{\partial\theta^\top}\Big|_{\theta_\gamma} \mathbf{M}^{-1}(\mathbf{X}_n, \theta_\gamma)\, \frac{\partial h(\theta)}{\partial\theta}\Big|_{\theta_\gamma} \\
\mathbf{P}_\gamma &= \frac{\mathbf{b}_\gamma \mathbf{b}_\gamma^\top}{\|\mathbf{b}_\gamma\|^2}
\end{aligned}$$

[There also exist more precise approximations, more complicated;
the difficulty compared to (Tierney et al., 1989) is that $\hat{\theta}^n(\mathbf{y})$ is not known explicitly]

→ Can be used to compare experiments

Ex: a two-compartment model in pharmacokinetics (P & Pázman, 2001)
Observe $y(t) = x_C(t)/V + \varepsilon(t)$ where $x_C(t)$ evolves according to

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

errors $\epsilon(t_i)$ i.i.d. $\mathscr{N}(0, \sigma^2)$

$\rightarrow$ Can be used to compare experiments

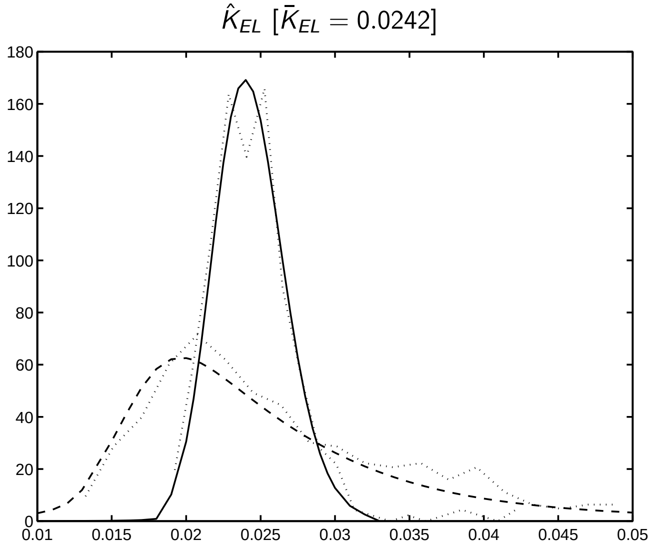
Ex: a two-compartment model in pharmacokinetics (P & Pázman, 2001)
Observe $y(t) = x_C(t)/V + \varepsilon(t)$ where $x_C(t)$ evolves according to

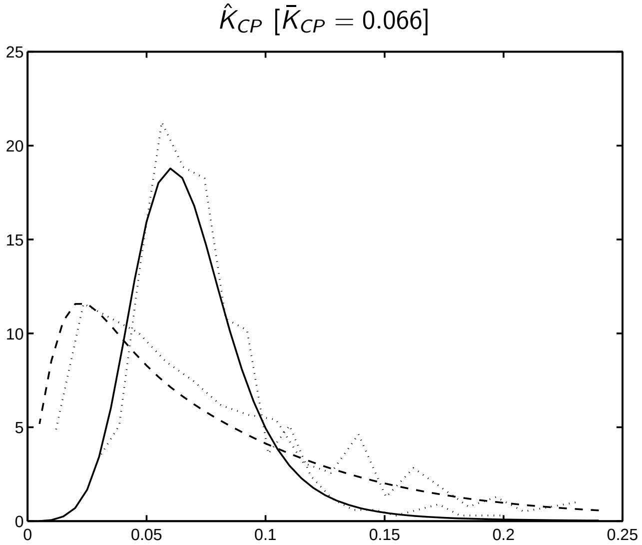
$$\left\{ \begin{array}{rcllcl} \frac{dx_C(t)}{dt} & = & (-K_{EL}- & K_{CP})x_C(t) & + & K_{PC}x_P(t) + u(t) \\ \frac{dx_P(t)}{dt} & = & & K_{CP}x_C(t) & - & K_{PC}x_P(t) \end{array} \right.$$

errors $\epsilon(t_i)$ i.i.d. $\mathscr{N}(0, \sigma^2)$

$\rightarrow$ 4 unknown parameters $\theta = (K_{CP}, K_{PC}, K_{EL}, V)^\top$

Compare 2 designs (8 observation times each) using simulated experiments with a given true $\bar{\theta}$


$\hat{K}_{EL}$ $[\bar{K}_{EL} = 0.0242]$
180
160
140
120
100
80
60
40
20
0
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05


$\hat{K}_{CP}$ [$\bar{K}_{CP} = 0.066$]
25
20
15
10
5
0
0
0.05
0.1
0.15
0.2
0.25

3b) Bias correction for LS estimation in nonlinear regression

$$\mathbf{b}(\bar{\theta}) = \text{bias of } \hat{\theta}^n = \mathsf{E}_{\mathbf{X}_n,\bar{\theta}}\{\hat{\theta}^n(\mathbf{y})\} - \bar{\theta}$$
$$= \underbrace{-\frac{\sigma^2}{2n^2}\,\mathbf{M}^{-1}(\mathbf{X}_n,\bar{\theta})\,\frac{\partial \boldsymbol{\eta}^\top(\theta)}{\partial\theta}\Big|_{\bar{\theta}}\sum_{i,j=1}^{p}\mathbf{H}_{ij}(\bar{\theta})\,\{\mathbf{M}^{-1}(\mathbf{X}_n,\bar{\theta})\}_{ij}}_{=\tilde{\mathbf{b}}(\bar{\theta})\ \text{(Box, 1971)}} + \mathcal{O}(\sigma^4)$$

We can write $\hat{\theta}^n = \mathbf{b}(\bar{\theta}) + \bar{\theta} + \omega$, with $\mathsf{E}_{\mathbf{X}_n,\bar{\theta}}\{\omega\} = \mathbf{0}$

Two-stage LS: solve $\boxed{\hat{\theta}^n = \mathbf{b}(\theta) + \theta}$ for $\theta \rightarrow \hat{\theta}^{n,*}$

[$\hat{\theta}^{n,*}$ unbiased when $\mathbf{b}(\theta) = \mathbf{A}\theta + \mathbf{c}$ for all θ with $\mathbf{I}_p + \mathbf{A}$ nonsingular]

1st method: $\hat{\theta}^{n,0} = \hat{\theta}^n$ given, then

$$\hat{\theta}^{n,1} = \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,0})$$

[…sometimes more biased than $\hat{\theta}^n$ (Picard and Prum, 1992)]

1st method: $\hat{\theta}^{n,0} = \hat{\theta}^n$ given, then

$$\begin{aligned}
\hat{\theta}^{n,1} &= \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,0}) \\
&\quad [\ldots \text{sometimes more biased than } \hat{\theta}^n \text{ (Picard and Prum, 1992)}] \\
\hat{\theta}^{n,2} &= \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,1}) \\
\vdots &= \vdots \\
\hat{\theta}^{n,*} = \hat{\theta}^{n,\infty} &= \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,\infty})
\end{aligned}$$

that is, $\hat{\theta}^{n,*} + \mathbf{b}(\hat{\theta}^{n,*}) = \hat{\theta}^n$, or

$$\mathsf{E}_{\mathbf{X}_n,\hat{\theta}^{n,*}}\{\hat{\theta}^n(\mathbf{y})\} = \boxed{\int_{\mathbb{R}^n} \hat{\theta}^n(\mathbf{y})\,\varphi(\mathbf{y}|\mathbf{X}_n,\hat{\theta}^{n,*})\,\mathrm{d}\mathbf{y} = \hat{\theta}^n}$$

<u>1st method</u>: $\hat{\theta}^{n,0} = \hat{\theta}^n$ given, then

$$\begin{aligned}
\hat{\theta}^{n,1} &= \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,0}) \\
&\quad [\ldots \text{sometimes more biased than } \hat{\theta}^n \text{ (Picard and Prum, 1992)}] \\
\hat{\theta}^{n,2} &= \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,1}) \\
\vdots &= \vdots \\
\hat{\theta}^{n,*} = \hat{\theta}^{n,\infty} &= \hat{\theta}^n - \mathbf{b}(\hat{\theta}^{n,\infty})
\end{aligned}$$

that is, $\hat{\theta}^{n,*} + \mathbf{b}(\hat{\theta}^{n,*}) = \hat{\theta}^n$, or

$$\mathsf{E}_{\mathbf{X}_n,\hat{\theta}^{n,*}}\{\hat{\theta}^n(\mathbf{y})\} = \boxed{\int_{\mathbb{R}^n} \hat{\theta}^n(\mathbf{y})\, \varphi(\mathbf{y}|\mathbf{X}_n, \hat{\theta}^{n,*})\, \mathrm{d}\mathbf{y} = \hat{\theta}^n}$$

Solve for $\hat{\theta}^{n,*}$ using stochastic approximation (P & Pázman, 1994a)

2nd method (approximate): use $\tilde{\mathbf{b}}$ instead of $\mathbf{b}$

2nd method (approximate): use $\tilde{\mathbf{b}}$ instead of $\mathbf{b}$

Solve for $\tilde{\theta}^{n,*}$: $\boxed{\tilde{\theta}^{n,*} + \tilde{\mathbf{b}}(\tilde{\theta}^{n,*}) = \hat{\theta}^n}$

that is

$$\tilde{\theta}^{n,*} - \frac{\sigma^2}{2n^2}\,\mathbf{M}^{-1}(\mathbf{X}_n, \tilde{\theta}^{n,*})\,\frac{\partial \boldsymbol{\eta}^\top(\theta)}{\partial \theta}\Big|_{\tilde{\theta}^{n,*}}\,\sum_{i,j=1}^{p} \mathbf{H}_{:ij}(\tilde{\theta}^{n,*})\,\{\mathbf{M}^{-1}(\mathbf{X}_n, \tilde{\theta}^{n,*})\}_{ij} = \hat{\theta}^n$$

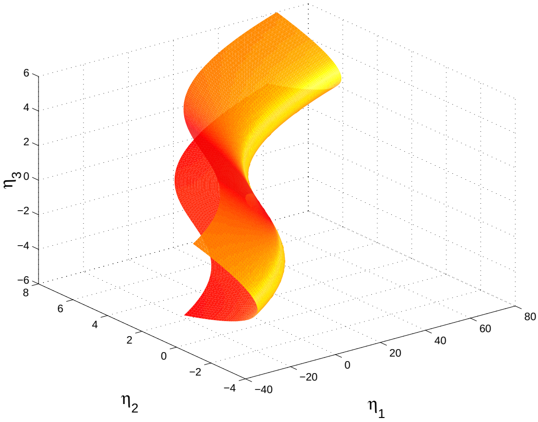
[Different from the score-corrected estimator $\hat{\theta}^n_{sc}$ of (Firth, 1993):

$\rightarrow$ solve $\boxed{\frac{\partial \boldsymbol{\eta}^\top(\theta)}{\partial \theta}[\mathbf{y} - \boldsymbol{\eta}(\theta)] - \mathbf{M}(\mathbf{X}_n, \theta)\tilde{\mathbf{b}}(\theta) = \mathbf{0}}$ for θ]

(Pázman & P, 1998) gives the (approximate) joint and marginal densities of $\tilde{\theta}^{n,*}$ and $\hat{\theta}^n_{sc}$

6 Extended optimality criteria

(P & Pázman, 2013, Chap. 7)



$\mathbb{S}_\eta$ may overlap, there may be local minimisers for the LS problem...

Important and difficult problem, often neglected

What can we do at the design stage?

➠ extensions of usual optimality criteria

➔ Avoid situations where $\|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\bar{\theta})\|$ can be small when $\|\theta - \bar{\theta}\|$ is large:

$$\text{maximise } \phi_{eE}(\mathbf{X}_n, \theta^0) = \min_{\theta} \frac{\|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta^0)\|^2}{\|\theta - \theta^0\|^2}$$

corresponds to E-optimal design ($\Leftrightarrow$ maximise $\lambda_{\min}[\mathbf{M}(\mathbf{X}_n)]$) when η is linear

What can we do at the design stage?

➠ extensions of usual optimality criteria

➔ Avoid situations where $\|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\bar{\theta})\|$ can be small when $\|\theta - \bar{\theta}\|$ is large:

$$\text{maximise } \phi_{eE}(\mathbf{X}_n, \theta^0) = \min_{\theta} \frac{\|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta^0)\|^2}{\|\theta - \theta^0\|^2}$$

corresponds to E-optimal design ($\Leftrightarrow$ maximise $\lambda_{\min}[\mathbf{M}(\mathbf{X}_n)]$) when η is linear

Extensions of E-, G- and c-optimal design in (Pázman & P, 2014)

What can we do at the design stage?

➟ extensions of usual optimality criteria

➔ Avoid situations where $\|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\bar{\theta})\|$ can be small when $\|\theta - \bar{\theta}\|$ is large:

$$\text{maximise } \phi_{eE}(\mathbf{X}_n, \theta^0) = \min_{\theta} \frac{\|\boldsymbol{\eta}(\theta) - \boldsymbol{\eta}(\theta^0)\|^2}{\|\theta - \theta^0\|^2}$$

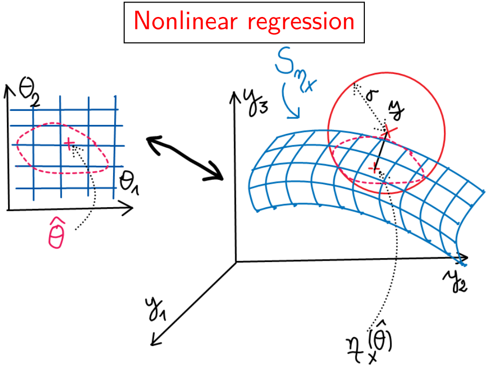
corresponds to E-optimal design ($\Leftrightarrow$ maximise $\lambda_{\min}[\mathbf{M}(\mathbf{X}_n)]$) when η is linear

Extensions of E-, G- and c-optimal design in (Pázman & P, 2014)

Extensions to generalised regression models and other design criteria in the Ph.D. thesis (Sternmüllerová, 2019)

7 Nonlocal DoE for nonlinear models

(P & Pázman, 2013, Chap. 8)

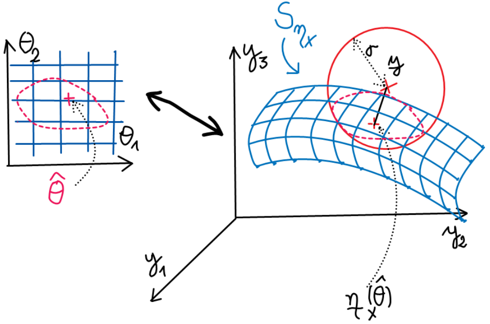


Nonlinear model ➠ **everything is local**

7 Nonlocal DoE for nonlinear models

(P & Pázman, 2013, Chap. 8)

Nonlinear regression



Nonlinear model ➠ **everything is local**

$\phi(\cdot)$ an information criterion, to be maximised with respect to the design $\mathbf{X}_n$:

$$\phi(\mathbf{X}_n) = \phi(\mathbf{X}_n, \theta), \text{ but which } \theta?$$

Local optimum design: based on a nominal value θ^0 → maximize $\phi(\mathbf{X}_n, \theta^0)$
[concerns all methods considered so far,
based on asymptotic normality (AN) or small-sample properties]

Local optimum design: based on a nominal value θ^0 → maximize $\phi(\mathbf{X}_n, \theta^0)$
[concerns all methods considered so far,
based on asymptotic normality (AN) or small-sample properties]

Objective of nonlocal DoE: remove the dependence in θ^0
3 main classes, essentially for $\phi(\xi, \theta) = \Phi[\mathbf{M}(\mathbf{X}_n, \theta)]$ (based on AN)

Local optimum design: based on a nominal value $\theta^0 \rightarrow$ maximize $\phi(\mathbf{X}_n, \theta^0)$
[concerns all methods considered so far,
based on asymptotic normality (AN) or small-sample properties]

Objective of nonlocal DoE: remove the dependence in θ^0
3 main classes, essentially for $\phi(\xi, \theta) = \Phi[\mathbf{M}(\mathbf{X}_n, \theta)]$ (based on AN)

❶ Average optimum design: maximise $\mathsf{E}_\theta\{\phi(\mathbf{X}_n, \theta)\}$ (or $\mathsf{E}_\theta\{\phi(\xi, \theta)\}$)

Local optimum design: based on a nominal value θ^0 ➞ maximize $\phi(\mathbf{X}_n, \theta^0)$
[concerns all methods considered so far,
based on asymptotic normality (AN) or small-sample properties]

Objective of nonlocal DoE: remove the dependence in θ^0
3 main classes, essentially for $\phi(\xi, \theta) = \Phi[\mathbf{M}(\mathbf{X}_n, \theta)]$ (based on AN)

➊ Average optimum design: maximise $\mathsf{E}_\theta\{\phi(\mathbf{X}_n, \theta)\}$ (or $\mathsf{E}_\theta\{\phi(\xi, \theta)\}$)
➋ Maximin optimum design: maximise $\min_\theta\{\phi(\mathbf{X}_n, \theta)\}$ (or $\min_\theta\{\phi(\xi, \theta)\}$)
➠ Between ➊ and ➋: regularised maximin criteria, quantiles and probability level criteria

Local optimum design: based on a nominal value θ^0 ➞ maximize $\phi(\mathbf{X}_n, \theta^0)$
[concerns all methods considered so far,
based on asymptotic normality (AN) or small-sample properties]

Objective of nonlocal DoE: remove the dependence in θ^0
3 main classes, essentially for $\phi(\xi, \theta) = \Phi[\mathbf{M}(\mathbf{X}_n, \theta)]$ (based on AN)

❶ Average optimum design: maximise $\mathsf{E}_\theta\{\phi(\mathbf{X}_n, \theta)\}$ (or $\mathsf{E}_\theta\{\phi(\xi, \theta)\}$)
❷ Maximin optimum design: maximise $\min_\theta\{\phi(\mathbf{X}_n, \theta)\}$ (or $\min_\theta\{\phi(\xi, \theta)\}$)
➠ Between ❶ and ❷: regularised maximin criteria, quantiles and probability level criteria
❸ Sequential design

➊ Average Optimum design

Probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$ ($\neq$ Bayesian estimation)

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

[No difficulty if Θ is finite and $\mu = \sum_{i=1}^{M} \alpha_i \delta_\theta^{(i)}$ (integral $\to$ finite sum); otherwise, use stochastic approximation to avoid evaluations of integrals]

❶ Average Optimum design

Probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$ ($\neq$ Bayesian estimation)

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

[No difficulty if Θ is finite and $\mu = \sum_{i=1}^M \alpha_i \delta_\theta^{(i)}$ (integral $\rightarrow$ finite sum); otherwise, use stochastic approximation to avoid evaluations of integrals]

$\phi_A(\xi)$ is concave in ξ when each $\phi(\xi,\theta)$ is concave

➠ same properties and same algorithms as for local design

❶ Average Optimum design

Probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$ ($\neq$ Bayesian estimation)

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

[No difficulty if Θ is finite and $\mu = \sum_{i=1}^{M} \alpha_i \delta_\theta^{(i)}$ (integral $\rightarrow$ finite sum); otherwise, use stochastic approximation to avoid evaluations of integrals]

$\phi_A(\xi)$ is concave in ξ when each $\phi(\xi, \theta)$ is concave

➠ same properties and same algorithms as for local design

❷ Maximin Optimum design

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

$\phi_M(\xi)$ is concave in ξ when each $\phi(\xi, \theta)$ is concave, but $\phi_M(\cdot)$ is non-differentiable

❶ Average Optimum design

Probability measure $\mu(\mathrm{d}\theta)$ on $\Theta \subseteq \mathbb{R}^p$ ($\neq$ Bayesian estimation)

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

[No difficulty if Θ is finite and $\mu = \sum_{i=1}^M \alpha_i \delta_\theta^{(i)}$ (integral $\to$ finite sum); otherwise, use stochastic approximation to avoid evaluations of integrals]

$\phi_A(\xi)$ is concave in ξ when each $\phi(\xi,\theta)$ is concave

➠ same properties and same algorithms as for local design

❷ Maximin Optimum design

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

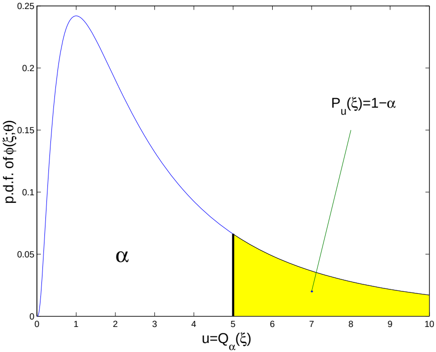
$\phi_M(\xi)$ is concave in ξ when each $\phi(\xi,\theta)$ is concave, but $\phi_M(\cdot)$ is non-differentiable

Problems

❶ Optimal design for $\phi_A(\cdot)$ not invariant by a monotone transformation of $\phi(\cdot,\theta)$

❷ Optimal design for $\phi_M(\cdot)$ very sensitive to the choice of the boundary of Θ

Between ❶ and ❷: quantiles and probability level criteria



➞ maximise P_u for a given u, or maximise Q_α for a given α

(when $\alpha \to 0$, tends to maximin optimality)

Directional derivatives, algorithms . . . but the criteria are not concave:

$\rightarrow$ no guarantee of successful maximisation

Directional derivatives, algorithms ... but the criteria are not concave:
→ no guarantee of successful maximisation

A related very promising approach: maximise the conditional value at risk (or superquantile) as proposed by Valenzuela et al. (2015)

$$\phi_\alpha(\mathbf{X}_n) = \max_{t\in\mathbb{R}} \left\{ t + \frac{1}{\alpha} \int_\Theta \min\left[0, \phi(\mathbf{X}_n;\theta) - t\right] \mu(\mathrm{d}\theta) \right\}$$

When μ has a density (w.r.t. Lebesgue measure on Θ) then

$$\phi_\alpha(\mathbf{X}_n) = \frac{1}{\alpha} \int_{\{\theta:\phi(\mathbf{X}_n;\theta)<Q_\alpha(\mathbf{X}_n)\}} \phi(\mathbf{X}_n;\theta)\,\mu(\mathrm{d}\theta)$$

$\phi(\xi,\theta)$ concave in $\xi \Rightarrow \phi_\alpha(\xi)$ concave
$\phi_1(\mathbf{X}_n) = \phi_A(\mathbf{X}_n)$ and $\phi_\alpha(\mathbf{X}_n) \to \phi_M(\mathbf{X}_n)$ as $\alpha \to 0$
[part of the Ph.D. thesis (Sternmüllerová, 2019)]

❸ Sequential design

$\theta^0 \to$ design: $\mathbf{X}^1 = \arg\max_{\mathbf{X}} \phi(\mathbf{X}, \theta^0)$

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

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

$\to$ design: $\mathbf{X}^2 = \arg\max_{\mathbf{X}} \phi(\{\mathbf{X}^1, \mathbf{X}\}, \hat{\theta}^1)$

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

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

$\to$ design: $\mathbf{X}^3 = \arg\max_{\mathbf{X}} \phi(\{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}\}, \hat{\theta}^2)$

...etc.

❸ Sequential design

$\theta^0 \to$ design: $\mathbf{X}^1 = \arg\max_{\mathbf{X}} \phi(\mathbf{X}, \theta^0)$

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

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

$\to$ design: $\mathbf{X}^2 = \arg\max_{\mathbf{X}} \phi(\{\mathbf{X}^1, \mathbf{X}\}, \hat{\theta}^1)$

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

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

$\to$ design: $\mathbf{X}^3 = \arg\max_{\mathbf{X}} \phi(\{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}\}, \hat{\theta}^2)$

. . . etc.

➞ Replace unknown θ by best current guess $\hat{\theta}^k$

(there exist variants with Bayesian estimation and average optimality)

❸ Sequential design

$\theta^0 \to$ design: $\mathbf{X}^1 = \arg\max_{\mathbf{X}} \phi(\mathbf{X}, \theta^0)$

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

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

$\to$ design: $\mathbf{X}^2 = \arg\max_{\mathbf{X}} \phi(\{\mathbf{X}^1, \mathbf{X}\}, \hat{\theta}^1)$

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

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

$\to$ design: $\mathbf{X}^3 = \arg\max_{\mathbf{X}} \phi(\{\mathbf{X}^1, \mathbf{X}^2, \mathbf{X}\}, \hat{\theta}^2)$

...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 designs based on $\mathbf{M}$)?

(difficulty: $\mathbf{X}^k$ depends on $\mathbf{y}^1, \ldots, \mathbf{y}^{k-1} \Longrightarrow$ independence is lost)

Each $\mathbf{X}^i$ has size q

➠ No big difficulty if $q \geq p = \dim(\theta)$ (batch sequential design)

Each $\mathbf{X}^i$ has size q

➠ No big difficulty if $q \geq p = \dim(\theta)$ (batch sequential design)

If n observation in total, two stages only: size of first batch?

$\rightarrow$ should be proportional to $\sqrt{n}$ (*which does not say much . . .*)

Each $\mathbf{X}^i$ has size q

➠ No big difficulty if $q \geq p = \dim(\theta)$ (batch sequential design)

If n observation in total, two stages only: size of first batch?

$\rightarrow$ should be proportional to $\sqrt{n}$ (*which does not say much ...*)

➠ Full sequential design: $\mathbf{X}^k = \{x_k\}$ ($q = 1$)

$\rightarrow$ <u>convergence properties are difficult to investigate</u>...

Each $\mathbf{X}^i$ has size q

➠ No big difficulty if $q \geq p = \dim(\theta)$ (batch sequential design)
If n observation in total, two stages only: size of first batch?
$\rightarrow$ should be proportional to $\sqrt{n}$ (*which does not say much* ...)

➠ Full sequential design: $\mathbf{X}^k = \{x_k\}$ $(q = 1)$
$\rightarrow$ convergence properties are difficult to investigate...
When

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

with $x_{k+1} = \arg\max_{\mathscr{X}} \underbrace{F_\phi(\xi^k; \delta_x | \hat{\theta}^k)}_{\text{directional derivative}}$ $\Leftrightarrow$ conditional gradient algorithm with step-size $\frac{1}{k+1}$ (Wynn, 1970)

➤ some CV results for Bayesian estimation (Hu, 1998)
➤ no general CV results for LS and ML estimation,
[unless $\mathscr{X} = \{x^{(1)}, \ldots, x^{(\ell)}\}$ finite (P 2009, 2010)]

8 Conclusions

DoE for nonlinear models with small data:

- ✓ Using the small-sample properties of the estimator can be a bit tricky

8 Conclusions

DoE for nonlinear models with small data:

✓ Using the small-sample properties of the estimator can be a bit tricky

✓ Numerical simulations are useful: for instance,
construct a locally optimum design (at θ^0),
simulate data (for another θ^1),
estimate θ (correct the bias),
check closeness to θ^1 (plot marginals),
repeat (for other θ^0 and θ^1), etc.

8 Conclusions

DoE for nonlinear models with small data:

- ✓ Using the small-sample properties of the estimator can be a bit tricky
- ✓ Numerical simulations are useful: for instance,
  construct a locally optimum design (at θ^0),
  simulate data (for another θ^1),
  estimate θ (correct the bias),
  check closeness to θ^1 (plot marginals),
  repeat (for other θ^0 and θ^1), etc.
- ✓ When applicable, sequential (adaptive) design is a good remedy to the dependence of the optimal design in θ

8 Conclusions

DoE for nonlinear models with small data:

- ✓ Using the small-sample properties of the estimator can be a bit tricky
- ✓ Numerical simulations are useful: for instance,
  construct a locally optimum design (at θ^0),
  simulate data (for another θ^1),
  estimate θ (correct the bias),
  check closeness to θ^1 (plot marginals),
  repeat (for other θ^0 and θ^1), etc.
- ✓ When applicable, sequential (adaptive) design is a good remedy to the dependence of the optimal design in θ

Thank you for your attention !

References I

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

Box, M., 1971. Bias in nonlinear estimation. Journal of Royal Statistical Society B33, 171–201.

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.

Dette, H., Pepelyshev, A., 2010. Generalized latin hypercube design for computer experiments. Technometrics 52 (4), 421–429.

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.

Firth, D., 1993. Bias reduction of maximum likelihood estimates. Biometrika 80 (1), 27–38.

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.

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

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

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

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

References II

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., 1998. Approximate densities of two bias–corrected nonlinear LS estimators. In: Atkinson, A., Pronzato, L., Wynn, H. (Eds.), MODA'5 – Advances in Model–Oriented Data Analysis and Experimental Design, Proceedings of the 5th Int. Workshop, Marseille, 22–26 juin 1998. Physica Verlag, Heidelberg, pp. 145–152.

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

Picard, D., Prum, B., 1992. The bias of the MLE, an example of the behaviour of different corrections in genetic models. Statistics 23, 159–169.

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., July 1994a. Bias correction in nonlinear regression via two-stages least-squares estimation. In: Blanke, M., Söderström, T. (Eds.), Prep. 10th IFAC/IFORS Symposium on Identification and System Parameter Estimation. Vol. 1. Danish Automation Society, Copenhagen, pp. 137–142.

References III

Pronzato, L., Pázman, A., 1994b. 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., 2001. Using densities of estimators to compare pharmacokinetic experiments. Computers in Biology and Medicine 31 (3), 179–195.

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

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., 1996. Optimum Designs for Multi-Factor Models. Springer, New York.

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

Sternmüllerová, K., 2019. Optimum design in nonlinear models. Ph.D. Thesis, Comenius University, Bratislava, Slovakia.

Tierney, L., Kass, R., Kadane, J., 1989. Approximate marginal densities of nonlinear functions. Biometrika 76 (3), 425–433.

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.

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.

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

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