Adaptive sequential design for regression on Schauder Basis

Serge Cohen, Sébastien Gadat

Adaptive sequential design for regression on Schauder Basis

Serge Cohen[a], Sébastien Gadat[a]

[a] *Institut de Mathématiques de Toulouse*
Laboratoire de Statistique et Probabilités

Abstract

We present a new sequential algorithm to build both optimal design and model selection in a multi-resolution family of functions. This algorithm relies on a localization property of discrete sequential D and A-optimal designs for Schauder Basis. We use these property with a simulated annealing strategy to obtain our stochastic algorithm. We illustrate its efficiency on several numerical experiments.

Key words: Sequential Optimal Design, Model Selection, Stochastic Algorithm
2000 MSC: 62K05, 62L05, 93E35

1. Introduction

This paper presents a new algorithm for building optimal design to recover an unknown signal f. The main interest of the work is included in the adaptive nature of this algorithm which will be sequential. Following classical ideas of multi-resolution analysis, we want to find both a design $(\xi_1, \dots \xi_n)$ and a family of linearly independent functions in an optimal way to expand f.

A large amount of recent works deal with some model selection approaches from a theoretical point of view often using L^1 penalized strategy to obtain sparse decomposition, and yielding for instance LASSO [8] or LARS [9] algorithms and at last the Dantzig selector [10]. But to the best of our knowledge, there does not exist some optimization method of the design dedicated to these sparse model selection methods.

Regarding now the community of optimal design research, lot of works are concerned with finding design represented as continuous measure and there exist scarce explicit results to find a good design. Some explicit discrete designs can be however found (see [5] and the reference therein) but most of time, good designs are located using some numerical algorithms. Moreover, a large amount of these numerical methods yields some continuous designs although discrete designs are easier to handle from a practical point of view. At last, from an optimal design point of view, there exists some advances in sequential methods (see [16]). Although it seems natural to fit both the model and the design with

Email addresses: Serge.Cohen@math.univ-toulouse.fr (Serge Cohen), Sebastien.Gadat@math.univ-toulouse.fr (Sébastien Gadat)

sequential measures of an unknown signal, no approach with model selection has been considered yet.

Our contribution is twofold, we first provide a new theoretical localization of optimal designs for a multi-resolution family of functions (the Schauder Basis), and these localization results can be tensorized to dimensions higher than one. Then, we infer from this multi-resolution family an adaptive strategy to build a model selection coupled with a sequential optimal design method.

This optimality remains to be properly defined and to obtain a precise mathematical criterion, we will quite naturally use some classical ideas of optimal design theory (see e.g. [1],[2] for the main tool and we will use [4, 5] for general ideas on optimal designs), adaptive regression [3] and multi-resolution analysis [6, 7].

To clearly define our objective and settings, denote E the space where the variable t is living , we assume the signal f to be expanded in a basis $(\Lambda_{j,k})_{j,k}$. We want to successively select some measure points of the design ξ_i and some element of $(\Lambda_{j,k})_{j,k}$ to reach a correct approximation $\hat{f}$ of f.

Our objective is twofold: first we want to recursively find an appropriate design $\mathbf{x} = \{\xi_1, \dots, \xi_n\}$ which will be adaptive to the sequential measurements done on the unknown signal f. Secondly, our goal is to select an appropriate subset of functions Λ to keep the variability of the reconstructed (approximated) signal $\hat{f}$ low and build thus a sparse representation of $\hat{f}$ of the true unknown f. To measure both well-suited designs and set of functions Λ, $\hat{f}$ will naturally be deduced from $(f(\xi_i))_{i\in\{1,\dots n\}}$ and Λ using a classical linear model. We will not adopt a penalization approach as [8, 9, 10] since in this framework, the effect of the chosen design $\mathbf{x}$ on the variance of the reconstructed $\hat{f}$ is not explicit which makes the first step of building an optimal $\mathbf{x}$ very hard, and moreover these methods are not exclusively dedicated to recover $\hat{f}$ using as less observations as possible.

Our work is organized as follow: next section presents some definitions and and classical considerations of optimal design theory and then describes the general behavior of our adaptive algorithm (model selection and sequential design). Section 3 gives some theoretical results on the localization of the sequential optimal design and proves consistency of our method without model selection. Section 4 precisely describes the stochastic algorithm which builds the model selection. At last, Section 5 provides some experimental comparisons, especially with L^1-penalized approaches which are widely used now.

2. Model

2.1. Basis expansion

We will use the one dimensional framework but all our results can be extended to the multi-dimensional case. Denote $E = [0; 1]$ the one-dimensional space and $x \in E$, f is supposed to be expanded in the "triangle" Schauder Basis defined by:

$$\Lambda_{0,0}(t) = \frac{1}{2} - \left| x - \frac{1}{2} \right| \quad \text{and} \quad \Lambda_{j,k}(t) = 2^{j/2}\Lambda_{0,0}(2^j t - k). \tag{1}$$

Some examples of $\Lambda_{j,k}$ are plotted in figure 7.

Note here that we have chosen to normalize functions $\Lambda_{j,k}$ so that $\|\Lambda_{j,k}\|_2 = 1/4$ but this choice will not have some important consequence. In this basis, the unknown f is given by

$$f(t) = \underbrace{\sum_{j,k} \alpha_{j,k}\Lambda_{j,k}(t)}_{:=\eta(t)=\mathbb{E}[f(t)]} + \epsilon(t), \tag{2}$$

where $\epsilon(t)$ is a Gaussian white noise $\mathcal{N}(0, \sigma^2)$.

Remark 1. *In this work, we have chosen to use the Triangle Schauder Basis instead of a true multi-resolution wavelet basis for one main reason. Indeed, finding optimal design using such basis will be almost explicit since we will determine for each subset of functions $(\Lambda_i)_{i\in I}$ a finite set of points to build optimal designs.*

2.2. General description of the sequential algorithm

The objective is to build an "optimal" pair $(\mathbf{x}, I)$ where $\mathbf{x}$ denotes the design of the linear model built whereas I is the index of functions used to build a linear model with the design $\mathbf{x}$. Indeed, as our framework is the classical statement of optimal design, one must understand that the measurement of some $f(x_i)$ is considered as a costly task and the searched algorithm will have to select a few points among E to approximate well the signal f over E. Consequently, it will be impossible to explore both all possible indexes I and n-sets $\mathbf{x}$ (and thus to compute the associated $f(\xi_i)$ for $\xi_i \in \mathbf{x}$) and to choose among them the best fitted linear model. Thus, we will follow a sub-optimal strategy where we will successively build the design $\mathbf{x_n}$ and the set I_n recursively. Initialization of $(\mathbf{x_n}, I_n)$ will be detailed in the sequel. To build $\mathbf{x_{n+1}}$ from $\mathbf{x_n}$, obviously we will not erase some points from $\mathbf{x_n}$ in the design at step $n+1$, because it has been costly to evaluate f on the design $\mathbf{x_n}$. Hence it will be imperative to keep the former points of $\mathbf{x_n}$ in $\mathbf{x_{n+1}}$ so that $\mathbf{x}_{n+1} \subset \mathbf{x}_n$.

To infer a criterion and an "optimality" for $(\mathbf{x_n}, I_n)$, we will detail first some classical element of optimal designs theory before we adapt them to our initial motivation of finding both good adaptive designs and set of functions.

The number of observations will be fixed in the beginning of the algorithm.

2.3. Integrated mean square error (IMSE)

Following notations of [1], we call J the IMSE inferred from any design $\mathbf{x}$ using a set of functions Λ indexed by a I:

$$J(\mathbf{x}, I) = \frac{\Omega}{\sigma^2} \int_E \left[\mathbb{E}\hat{f}_{\mathbf{x},I}(t) - \eta(t)\right]^2 dt, \tag{3}$$

where $\hat{f}_{\mathbf{x},I}$ is the estimator of f based on a standard linear model computed on $\mathbf{x}$ and $f(\mathbf{x})$ using the linear combinations of $\left(\Lambda_{(j,k)}\right)_{(j,k)\in I}$. In the last formula, Ω is the volume of the domain E. More precisely, denote $f(\mathbf{x})$ the column vector given by the signal observed on the points of the design $\mathbf{x}$ of length l:

$$f(\mathbf{x}) = \begin{pmatrix} f(\xi_1) \\ \vdots \\ f(\xi_l) \end{pmatrix},$$

and use the notation $\Lambda_I(\mathbf{x})$ for the rectangular $(p \times l)$ matrix:

$$\Lambda_I(\mathbf{x}) = \begin{pmatrix} \Lambda_{(j_1,k_1)}(\xi_1) & \dots & \Lambda_{(j_1,k_1)}(\xi_l) \\ \vdots & \dots & \vdots \\ \Lambda_{(j_p,k_p)}(\xi_1) & \dots & \Lambda_{(j_p,k_p)}(\xi_l) \end{pmatrix}.$$

Then, the linear estimator $\hat{f}_{\mathbf{x},I}$ is defined as

$$\hat{f}_{\mathbf{x},I} = \sum_{(j,k)\in I} \hat{\alpha}_{(j,k)} \Lambda_{(j,k)},$$

where the vector $\hat{\alpha} = (\hat{\alpha}_{(j,k)})_{(j,k)\in I}$ is given by:

$$\hat{\alpha} = \left({}^t\Lambda_I(\mathbf{x})\Lambda_I(\mathbf{x})\right)^{-1} \Lambda_I(\mathbf{x}) f(\mathbf{x}). \tag{4}$$

Finally, expanding the IMSE definition (3) yields the classical bias/variance trade off:

$$J(\mathbf{x}, I) = \frac{\Omega}{\sigma^2} \underbrace{\int_E Var[\hat{f}_{\mathbf{x},I}(t)]dt}_{:=V_{\mathbf{x},I}} + \frac{\Omega}{\sigma^2} \underbrace{\int_E \left(\mathbb{E}[\hat{f}_{\mathbf{x},I}(t)] - \eta(t)\right)^2 dt}_{:=B_{\mathbf{x},I}}. \tag{5}$$

Recall now that our goal is to find both design $\mathbf{x}$ and decomposition subset Λ_I to minimize (5). In the last equation, obviously, the bias term $B_{\mathbf{x},I}$ is untractable since it depends principally on η which is unknown for all t over E and which is approximated by $f(\mathbf{x})$ at the design points $\mathbf{x}$. Thus, equation (5) is not good enough to recover good pairs $(\mathbf{x}, I)$. Consequently, it will be necessary to slightly modify and bound the "energy term" $J(\mathbf{x}, I)$ to get something one can expect to minimize.

2.4. Energy term and the adaptive strategy

2.4.1. The bias term

We start pointing a first method to handle the bias term even if we will not use this model to run our algorithm for computational reason.

Bias bound using discrepancy

As pointed in the last paragraph, we need to bound $J(\mathbf{x}, I)$ to yield numerically tractable equation. It is possible to use the Koksma-Hlawka inequality [11, 12] inferred from the discrepancy of $\mathbf{x}$:

$$\int_E \left(\mathbb{E}[\hat{f}_{\mathbf{x},I}(t)] - \eta(t)\right)^2 dt \leq \frac{1}{l}\sum_{i=1}^{l} \left[\mathbb{E}\hat{f}_{\mathbf{x},I}(\xi_i)] - \eta(\xi_i)\right]^2 + D_l^*(\mathbf{x}) V\left((\mathbb{E}\hat{f}_{\mathbf{x},I} - \eta)^2\right), \tag{6}$$

where $D_l^*(\mathbf{x})$ is the so called star-discrepancy of $\mathbf{x}$ up to l, and V is a variation of the function $t \mapsto (\mathbb{E}(\hat{f}_{\mathbf{x},I}(t)) - \eta(t))^2$. However, equation (6) may not be satisfactory again since we cannot really compute the variation! This is why it may be natural to replace the last term involving η by a penalized term

$$\int_E \left(\mathbb{E}[\hat{f}_{\mathbf{x},I}(t)] - \eta(t)\right)^2 dt \leq \frac{1}{l}\sum_{i=1}^{l} \left[\mathbb{E}\hat{f}_{\mathbf{x},I}(x_i)] - \eta(x_i)\right]^2 + \lambda_I D_l^*(\mathbf{x}) := B_1(\mathbf{x}, I). \tag{7}$$

In this last equation, λ_I is a penalization term replacing the total variation of $(\mathbb{E}\hat{f}_{\mathbf{x},I} - \eta)^2$. This last term may increase with the highest resolution of maps composing Λ_I.

Remark 2. *We provide this bound for theoretical sake of completeness. We propose to simply consider the case $\lambda_I = 0$ instead for numerical reasons since the solution of our optimization step will be almost explicit.*

Fast approximation of the bias term

Recall that the bias term expression is

$$B(\mathbf{x}, I) = \int_E \left(\mathbb{E}[\hat{f}_{\mathbf{x},I}(t)] - \eta(t)\right)^2 dt.$$

In the sequel, we will need to control the power of bias reduction of each function Λ_I. To do so, we will simply approximate this term and his derivatives by an empirical mean. The section 3 will detail the use of such approximations.

2.4.2. The Variance term

Following standard argument of optimal design theory, $V(\mathbf{x}, I)$ can be computed from the definitions of $\hat{\alpha}$ given in equation (4) and $f = \eta + \epsilon$. Immediate computation (see [13] for instance) yields

$$V_{\mathbf{x},I} = \int_E \sigma^2 \, Var\left[{}^t\Lambda_I(t)\left[{}^t\Lambda_I(\mathbf{x})\Lambda_I(\mathbf{x})\right]^{-1} {}^t\Lambda_I(\mathbf{x})\right] dt = \sigma^2 Tr\left(\mu_{1,1}(I) M_{\mathbf{x},I}^{-1}\right).$$

Let's recall that in the last formula, $M_{\mathbf{x},I}$ is the information matrix of the design $\mathbf{x}$ with the basis function Λ_I stated as

$$M_{\mathbf{x},I} = \Lambda_I(\mathbf{x})^t \Lambda_I(\mathbf{x}),$$

where ${}^tA$ is the transposed of A, and $\mu_{1,1}(I)$ is the first moment matrix given by

$$\mu_{1,1}(I) = \int_E {}^t\Lambda_I(t)\Lambda_I(t)dt.$$

To sum up the two last paragraphs, we obtain naturally the energy term

$$\mathcal{E}(\mathbf{x}, I) = Tr\left(\mu_{1,1}(I) M_{\mathbf{x},I}^{-1}\right) + \frac{1}{\sigma^2}\frac{1}{l}\sum_{i=1}^{l}\left[\mathbb{E}\hat{f}_{\mathbf{x},I}(\xi_i)] - \eta(\xi_i)\right]^2. \tag{8}$$

Some further investigations will be necessary to handle a more general setting with $\lambda_I > 0$ and we will give some statistical idea of many open questions concerning some future developments.

2.4.3. The adaptive strategy

In our adaptive framework, we need to choose successively some new points in the design $\mathbf{x}$ while we can decide or not to update the set Λ_I. As pointed in the introductory paragraph, we do not delete points of the design $\mathbf{x}$. Consequently, the algorithm is necessarily of the following form:

Step 0

- Fix any initial set of functions Λ_{I_0}. For instance in a one dimensional setting with $E = [0; 1]$, we can choose naturally $I_0 = \{(0, 0); (1, 0); (1, 1)\}$ since we do not have any prior on the unknown function f.
- Compute the optimal design $\mathbf{x_0}$ which minimizes the Variance term:

$$\mathbf{x_0} = \arg\min_{\mathbf{x}} Tr\left(\mu_{1,1}(I) M_{\mathbf{x},I_0}^{-1}\right).$$

Note that this choice implies immediately that $\mathbf{x_0}$ minimize $\mathcal{E}(., I_0)$.

Step n

- Update the set of functions Λ_{I_n} to minimize $\mathcal{E}(\mathbf{x_n}, .)$. We will describe a suboptimal strategy below. Note that this suboptimal strategy will build $\Lambda_{I_{n+1}}$ from Λ_{I_n} with and addition of one son of one of the maps in Λ_{I_n} or deleting one map of Λ_{I_n}.
- Choose the optimal design $\mathbf{x_{n+1}}$ deduced from $\mathbf{x_n}$ with an addition of one point ξ_{n+1}

$$\mathbf{x_{n+1}} = \mathbf{x_n} \cup \{\xi_{n+1}\},$$

using the former set of functions $\Lambda_{I_{n+1}}$ previously computed.

The next section will describe how one can fix a fast algorithm to run both steps of each iteration of the algorithm.

3. Optimization steps

In the last algorithm, two steps are needed to be detailed, the first is the update of the optimal design and the second is how can we deduce to modify or not the set of functions Λ_{I_n}.

3.1. Scheme of the variance minimization

This paragraph is dedicated to the iteration $\mathbf{x_n} \mapsto \mathbf{x_{n+1}}$ and will use standard argument of optimal design theory. Recall that one has to determine the optimal ξ such that $\mathbf{x_n} \cup \xi$ will generate a minimum variance term in $\mathcal{E}$ while the set I_n is fixed:

$$\xi_{n+1} = \arg\min_{\xi} Tr(\mu_{1,1}(I_n) M^{-1}_{\mathbf{x}\cup\xi, I_n}) \tag{9}$$

Note that this last optimization procedure does not depend on any computation of f on any new point. The simplest natural way to find ξ_{n+1} is using a simulated annealing algorithm, but in the very particular case of triangle Schauder basis functions, this minimization step always yields very special solution as shown in the one dimensional figure 7.

Indeed, the minimization step always yields solutions that are dyadic points $\lambda = \frac{k}{2^j}$, whose resolutions j are bounded by the maximal resolution of the maps in Λ_I. This important fact clearly improved the numerical resolution of the equation (9). We will show some theoretical supporting proof in the next paragraph.

3.2. Theoretical study of the variance minimization for the Schauder Triangle Basis

In optimal design theory, there classically exists three optimality criteria of experimental designs. All these optimality are based on the information matrix $M_{\mathbf{x}}$, the D-optimal design is based on the minimization of

$$\Phi_0(M_{\mathbf{x}}) = \det M^{-1}_{\mathbf{x}},$$

while A-optimal design or E-optimal design are based on the maximization of

$$\Phi_{1,C}(M_{\mathbf{x}}) = Tr\left(C M^{-1}_{\mathbf{x}}\right) \qquad \text{and} \qquad \Phi_{\infty}(M_{\mathbf{x}}) = \sup_{\lambda \in Sp(M_{\mathbf{x}})} |\lambda|.$$

Obviously, in most cases, these three criteria does not yield equivalent designs. In our work, we have mainly focused on the two first criteria (D and A optimal designs). Note here that we do not handle some continuous measure on the design because our goal is to identify optimal "points" to measure f. Thus it is not possible to easily recover classical results on D and A optimal design applying some classical equivalence theorems [2, 13, 14] since our parameterization is not a convex function of the points of the design. In the next two paragraphs, we are studying the localization problem for optimal designs dedicated to the Schauder Triangle Basis defined by (1). Of course these properties may be false in a more general multi-resolution basis even if the adaptive algorithm principle remains unchanged.

3.2.1. First optimal design criterium

We provide first a study concerning the D-optimal design criterion. In our approach, we need to fix the first elements of the design $\mathbf{x}$ and find a point the location of $\xi \in E$ such that $\mathbf{x} \cup \xi$ is D-optimal. Next theorem shows that in fact, ξ is necessarily a dyadic point which is the maximum of one of the map of Λ_I. The proof of this theorem is deferred to the appendix.

Theorem 1. *Let* $\mathbf{x}$ *be any fixed design and* Λ_I *be any finite subset of functions extracted from the Schauder triangle basis, then*

$$\arg\min_{\xi \in E} \det M^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I) \subset \underbrace{\bigcup_{i\in I} \arg\max_{t\in E} \Lambda_i(t) \bigcup_{i\in I} \partial Supp(\Lambda_i)}_{:=\mathcal{E}},$$

where $Supp(f)$ *is the support of* f.

This theorem is very useful following our adaptive strategy to build $\mathbf{x}_{n+1}$ at step $n+1$ while adding a new point to the design $\mathbf{x}_n$ at step n. Indeed, it is sufficient to explore the small finite number of dyadic points $\mathcal{E}$, described above and select the point which maximizes the D-criterion. We also provide a result which generalizes the last theorem regarding the D-optimal design criterion for LASSO regression.

Theorem 2. *Let* $\mathbf{x}$ *be any fixed design and* Λ_I *be any finite subset of functions extracted from the Schauder triangle basis, then the*

$$\arg\min_{\xi \in E} \det \left(M_{\mathbf{x}\cup\xi}(\Lambda_I) + \alpha I_d\right)^{-1} \subset \mathcal{E}.$$

3.2.2. The trace optimal design criterion

We give here some element of the study of A-optimal design criterion, but this study is not complete yet since one of the key point remains open. We search ξ which maximizes the criterion given the last observations $\mathbf{x}$. Remark first that one can re-write the A-optimal design criterion using the equation (16):

$$\begin{aligned} Tr\left(\mu_{1,1} M^{-1}_{\mathbf{x}\cup\xi}\right) &= Tr\left(\mu_{1,1} M^{-1}_{\mathbf{x}} - \frac{\mu_{1,1} M^{-1}_{\mathbf{x}} \Lambda_I^t \Lambda_I M^{-1}_{\mathbf{x}}}{1 +^t \Lambda_I M^{-1}_{\mathbf{x}} \Lambda_I}\right) \\ &= Tr\left(\mu_{1,1} M^{-1}_{\mathbf{x}}\right) - \frac{^t\Lambda_I M^{-1}_{\mathbf{x}} \mu_{1,1} M^{-1}_{\mathbf{x}} \Lambda_I}{1 +^t \Lambda_I M^{-1}_{\mathbf{x}} \Lambda_I}. \end{aligned}$$

Thus, the location of the optimal point ξ at step $n+1$ is deduced from step n minimizing the second term of the last equation and from a numerical point of view, optimization of this last term is thus performed easily.

Remark 3. *From a theoretical point of view, we provide to the best of our knowledge, two unsolved conjectures which has always numerically been checked in our experiments. Note that the second one is stronger than the first one.*

Conjecture 1. *For any non negative t, we have the following localization property*

$$\arg\min_{\xi\in E} \det\left(tId + M^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I)\right) \subset \bigcup_{i\in I} \arg\max_{t\in E} \Lambda_i(t) \bigcup_{i\in I} \partial Supp(\Lambda_i).$$

Conjecture 2. *For any symmetric positive matrix C and non negative t, we have the following localization property*

$$\arg\min_{\xi\in E} \det\left(tC + M^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I)\right) \subset \bigcup_{i\in I} \arg\max_{t\in E} \Lambda_i(t) \bigcup_{i\in I} \partial Supp(\Lambda_i).$$

These two conjectures allow us to assert the next property which locate the A optimal designs for the Triangle Schauder Basis.

Theorem 3. *If the conjecture (1) is true, then*

$$\arg\max_{\xi\in E} Tr\left(M^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I)\right) \subset \bigcup_{i\in I} \arg\max_{t\in E} \Lambda_i(t) \bigcup_{i\in I} \partial Supp(\Lambda_i).$$

If the conjecture (2) is true, then

$$\arg\max_{\xi\in E} Tr\left(CM^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I)\right) \subset \bigcup_{i\in I} \arg\max_{t\in E} \Lambda_i(t) \bigcup_{i\in I} \partial Supp(\Lambda_i).$$

3.2.3. Convergence in the case of fixed basis I

We detail here the convergence of the parameter estimate $\hat{\alpha}$ while following the strategy of sequential optimal design detailed in the last paragraphs when the basis I remains fixed. As both of the two previous criterion yield same optimal design, we are only concerned by the study of sequential strategy:

$$\mathbf{x_{n+1}} = \mathbf{x_n} \cup \xi_{n+1} \qquad \text{and} \qquad \xi_{n+1} = \arg\max_{\xi} \det(M_{\mathbf{x_n}\cup\xi}),$$

while $\hat{\alpha}$ is classically given by

$$\hat{\alpha}_n = M^{-1}_{\mathbf{x_n}} \Lambda_I(\mathbf{x_n}) f(\mathbf{x_n}).$$

This asymptotic behavior is detailed in the next theorem whose proof is defered to the appendix

Theorem 4. *Let f and η be given as in (2) with a fixed basis I, then the sequential optimal design is consistent: $\hat{\alpha}_n \to \alpha$ a.s. Moreover, there exists a positive constant C such that*

$$\|\alpha_n - \alpha\|_\infty \leq C\sqrt{\frac{\log n}{n}}.$$

Remark 4. *The last theorem ensures the consistency of $\hat{\alpha}_n$ provided that the signal η is decomposed in the good basis function Λ_I. Note that when $\eta \notin Span(\Lambda_I)$, the convergence to the natural projection of η into $Span(\Lambda_I)$ also holds.*

4. Stochastic Model Selection

This section presents a stochastic algorithm to update I_n to build a coupled model selection with the sequential design strategy. For this, we need first some tool to estimate the efficiency of each function in I_n. This is done looking at the Bias term of (5).

4.1. Bias Optimization

This paragraph is dedicated to the optimization of the bias term defined through the bound (6). As pointed above, this term $B(\mathbf{x}, I)$ is replaced by the sum of the empirical L2 loss. We provide here some heuristics to update the basis functions $(\Lambda_i)_{i\in I_n}$. The first problem is to measure the efficiency of each Λ_i, $i \in I_n$ and we detail this measurement in the next paragraph. Then, we are facing the difficult problem to decide either to add, delete some functions in I_{n+1} or leave I_n unchanged. The paragraph will then provide some hints coming from classical acceptation-reject procedures of Metropolis stochastic algorithms.

4.1.1. Ranking criterion

We propose in the sequel to use one of three ranking criterion for functions of I_n into a stochastic simulated annealing like algorithm.

The ANOVA ranking. We detail here how we can measure the efficiency of each Λ_i where $i \in I_n$. The first natural idea is to use the ANOVA (Analysis of Variance) strategy. For each $i \in I_n$, compute the ratio

$$q_{anova}(i) = \frac{\sum_{x\in\mathbf{x_n}} \left[\hat{f}_{\mathbf{x_n},I_n\setminus\{i\}}(x) - f(x)\right]^2}{\sum_{x\in\mathbf{x_n}} \left[\hat{f}_{\mathbf{x_n},I_n}(x) - f(x)\right]^2}.$$

q_{anova} is classically related to the efficiency of each Λ_i to predict the unknown η, $q_{anova}(i)$ is weak when Λ_i is not relevant, and high when Λ_i is important for the linear model. Thus, the several ratio in q_{anova} provide a natural hint to rank the functions of Λ_i.

The LASSO ranking. We detail very briefly the LASSO procedure to rank variables in linear regression. The model introduced in [8] is to find $(a_i)_{i\in I_n}$ solution of the penalized $l1$ least square problem:

$$a^t = \arg\min_{\|a\|_1\le t} \sum_{x\in\mathbf{x_n}} \left[\sum_{i\in I} a_i\Lambda_i(x) - f(x)\right]^2, \tag{10}$$

where t is a non negative control parameter. Such optimization problem is well known to produce some sparse solutions of a^t (see [8, 9] for instance), the amount of sparsity in a^t is highly dependent on the value of t, sparse representations occurring for small values of t. Moreover, solutions of (10) satisfy the nice property:

$$\forall(h,t) > 0 \qquad a_i^t \neq 0 \Longrightarrow a_i^{t+h} \neq 0.$$

Since we recover the standard linear model estimate when t goes to infinity, we can thus rank variables by decreasing order of importance by increasing t yielding the classical Forward Stagewise linear Regression selection.

The empirical gradient ranking. We propose here to use a direct approximation of the bias gradient by an empirical approach, B is given by

$$B(\mathbf{x_n}, I_n) = \int_E \left(\mathbb{E}[\hat{f}_{\mathbf{x}_n, I_n}(t)] - \eta(t)\right)^2 dt,$$

and we can decompose $\mathbb{E}\hat{f}_{\mathbf{x}_n, I_n}(t)$ in our basis

$$\mathbb{E}\hat{f}_{\mathbf{x}_n, I_n}(t) = \sum_{i \in I} \hat{\alpha}_i \Lambda_i.$$

Now, compute each partial derivative to measure the power of each Λ_i

$$\frac{\partial B(\mathbf{x}, I)}{\partial \hat{a}_i} = \left| 2 \int_E \Lambda_i(t) \left[\mathbb{E}\hat{f}_{\mathbf{x}_n, I_n}(t) - \eta(t)\right] dt \right|.$$

Obviously, the exact computation of this last term is intractable, we approximate this term using naturally the former points in the design $\mathbf{x}$:

$$q_{bias}(i) = \left| \sum_{\xi \in \mathbf{x}_n} |Supp(\Lambda_i)| \Lambda_i(\xi) \left[\hat{f}_{\mathbf{x}_n, I_n}(\xi) - \eta(\xi)\right] \right|.$$

We will present in our experiments results based on this last empirical Bias criterion. We have also used the ANOVA or the LASSO ranking but we do not have found some significant differences with the approach based on q_{bias}.

4.1.2. Stochastic Learning of I_n with a Simulated Annealing dynamic

Following the last paragraph, it is possible to measure the efficiency of each map $\Lambda_i, i \in I_n$ since each element of I is described by an efficiency criterion (q_{bias}, q_{anova} and a_i^t). We now propose a method to modify I_n. This algorithm is largely inspired of classical stochastic methods of Metropolis-Hastings. Remind first that

$$\mathcal{E}_{emp}(\mathbf{x}, I) = Tr\left(\mu_{1,1}(I) M_{\mathbf{x},I}^{-1}\right) + \frac{1}{\sigma^2} \frac{1}{l} \sum_{i=1}^{l} \left[\hat{f}_{\mathbf{x},I}(x_i)] - f(x_i)\right]^2.$$

To update I_n, we will use a Simulated Annealing strategy which is classically decomposed in a proposition step and an acceptation rule adapted to a stationary measure criterion. We first recall some classical elements of Simulated Annealing theory before. Then we will describe precisely our proposition algorithm and the acceptation ratio.

Generality on Simulated Annealing algorithm. The Simulated Annealing procedure produces an algorithm to optimize a non negative functional cost C. This method involves simulating a non-homogeneous Markov chain whose invariant distribution at iteration n is $\mu_n \propto \mu^{1/T_n}$ where $(T_n)_{n \geq 0}$ is a temperature decreasing cooling scheme such that $T_n \to 0$. Under classical conditions (see [18] for instance), μ_∞ concentrates itself on the set of minima of C.

The situation is as follows: let Ω a measurable set with a measure m and let μ be a measure on Ω with density (also denoted μ) w.r.t m. The S.A. method with stationary distribution μ and proposal distribution $q(I, I')$ works as follow:

- from state $I \in \Omega$, first propose a state I' with probability $q(I, I')$

- then, accept the transition with a probability which is adjusted so that μ is invariant.

We assume the following property: for all $I \in \Omega$,

$$q(I, I') > 0 \Longleftrightarrow q(I', I) > 0.$$

The probability to accept the transition I to I' at iteration n is then defined as:

$$\forall I' \neq I \qquad Q_n(I, I') \quad = \quad \min\left\{\frac{\mu_n(I')q(I', I)}{\mu_n(I)q(I, I')}, 1\right\}. \tag{11}$$

When μ corresponds to a Gibbs field associated to a cost function C ($\mathcal{E}_{emp}$ in our case), this ratio is in fact given by

$$\forall I' \neq I \qquad Q_n(I, I') = \min\left\{e^{\frac{C(I)-C(I')}{T_n}}\frac{q(I', I)}{q(I, I')}, 1\right\}. \tag{12}$$

Reversible Jump proposal. We propose to use as transition kernel q a reversible MCMC [19]. The main difficulty is to ensure the weak reversibility condition given in the former paragraph:

$$q(I, I') > 0 \Longleftrightarrow q(I', I) > 0.$$

In our framework, we start with $I_0 = \{(0, 0\}$ and we decide to use the following dynamic for the set $I_n \mapsto I_{n+1}$ based on

$\mathcal{B}$: Birth of any element $i \notin I_n$, which is associated to a function Λ_{j_i,k_i}, provided that there exists an element $i' \in I_n$ such that Λ_{j_i,k_i} is a son or the father of $\Lambda_{j_{i'},k_{i'}}$. (The meaning of son and father is to be understood with respect to the complete dyadic tree considered as a family tree.)

$\mathcal{D}$: Death of any element $i \in I_n$ provided that one of its son or its father is still in I_n.

Please remark that the set of vertices in I_n are not connected in general in the dyadic tree, it is a consequence of the reversibility condition. These moves are defined by heuristic considerations, the only condition to be fulfilled is to maintain the correct invariant distribution described in equation (12).

Remark 5. *These moves are not so classical since basically one could make evolving the set I_n using birth or deletion steps following the natural structure of dyadic trees. This evolution would generate connected trees (from root to leaves) but such trees are not consistent with a sparse representation of the signal. At last, the necessary reversible jump condition is fulfilled provided the definition of $\mathcal{B}$ and $\mathcal{D}$.*

Given any iteration n, an "optimal design" $\mathbf{x}_n$ and a basis I_n, we use first one of the three ranking criterion defined above to propose a new state. We first fix a real $p_n \in]0; 1[$ which will be the probability to propose an addition of one function to I_n. Conversely, the real q_n will be the probability to delete one element of I_n. At last, r_n will be the probability of the birth of the initial

element $\Lambda_{0,0}$. If this initial element belongs to I_n, we set $r_n = 0$. Otherwise, (p_n, q_n, r_n) are chosen such that

$$p_n + q_n + r_n = 1 \quad p_n > 0 \quad q_n > 0 \quad r_n > 0.$$

In the birth case, denote I_n^{birth} the set of elements in I_n such that one of their sons is not present in I_n. Then, propose the birth of a descendant of some element $\Lambda_i, i \in I_n^{birth}$ where we sample i with a discrete probability p_{birth} which is an increasing function of q_{bias} or q_{anova}. For instance, one can simply choose

$$\forall i \in I_n \qquad p_{birth}(i) = \frac{q_{bias}(i)}{\sum_{j \in I_n} q_{bias}(j)}.$$

In the death case, denote I_n^{death} the set of elements in I_n such that one of their descendant or ascendant is in I_n and propose the death of one of the poorest predictor using a decreasing function of q_{bias} or q_{anova}.

The resulting transition kernel of the simulated Markov chain is then a mixture of the different transition kernel associated with the moves described above. We choose now classically $T_n = \frac{C_1}{C_2 + \log(n)}$ and this yields the transition kernel q and the acceptation ratio Q_n.

5. Experimental results

This section present two examples, each time an unknown signal η must be recovered from as few observations as possible. The first example deals with the approximation of some unknown functions that cannot be developed in the triangle Schauder basis. The second example illustrates the database of Motorcycle impact experiment ([20]). We will compare our method with some other approximations obtained with regular designs, or model selection strategy such as the LASSO. The numerical criterion to draw this comparison will be the Integrated Mean Square Error Rate. As pointed in the section 3, if the conjectures 1 and 2 are satisfied, designs obtained for standard linear model or for LASSO models are equivalent since the several minimum of the variance criterion are the same. Note that for all of our experiments, we normalize the observations to get $\Omega = [0; 1]$.

5.1. Description of the data

We investigate first the approximation obtained when the function η is unknown. We set first η_1 to be a sinus cardinal type function, more precisely, we get

$$\forall x \in [0; 1] \qquad \eta_1(x) = a \times \frac{\sin\left[k(x - 1/2)\right]}{k(x - 1/2)}.$$

In addition, we define f_1 as

$$\forall x \in [0; 1] \qquad f_1(x) = \eta_1(x) + \sigma w_1(x), \tag{13}$$

where $(dw_1(x))_{x \in [0;1]}$ is a normal centered independent white noise model. The parameters a and σ permit to modify the Signal to Noise Ratio.

We want to compare our regression method to recover η_1 with a small number of experiments. We finally initialize the triangle basis functions I_0 to $\Lambda_{I_0} =$

$\{\Lambda_{0,0}; \Lambda_{1,0}; \Lambda_{1;1}\}$. This initialization Λ_{I_0} is shown on figure 7 besides some realizations of equation (13) are shown on figure 7.4.

The next synthetic function to be approximated is a mixture of localized Gaussian kernel. This example will enable us to see whether our method is good adaptive to the successive noisy measurements of f. For this, we set η_2 to be localized around some values of Ω, $0.25, 0.5$ and 0.75, with different amplitudes and frequencies.

$$\forall x \in [0;1] \qquad \eta_2(x) = 5e^{-1000(x-0.25)^2} + 5e^{-100(x-0.75)^2} + 20e^{-100(x-0.5)^2}$$

In the last case, f_2 is defined as

$$\forall x \in [0;1] \qquad f_2(x) = \eta_2(x) + \sigma w_2(x), \tag{14}$$

and some realizations of equation (14) are shown on figure 7.5.

5.2. Methods

We run our algorithm setting $C_1 = 10, C_2 = 1$ and $p_n = 0.8, q_n = 0.2$ or $p_n = 0.75, q_n = 0.15$ and $r_n = 0.1$ (depending on $\Lambda_{0,0}$ belongs to I_n or not). Moreover, we assume that $\sigma = 1$.

To obtain a reliable integrated mean square estimation of the several methods, we repeat our experiments 1000 times and compute the IMSE between the true signal η_1 or η_2 and our estimates f_1 or f_2. The next figures will show the performance of the several methods used listed above:

- Method 1: linear model on "optimal" design $\mathbf{x}_n$ with learned I_n.
- Method 2: linear model on regular design $(i/n)_{i=1..n}$ with learned I_n.
- Method 3: LASSO model on "optimal" design with learned I_n.
- Method 4: LASSO model on regular design with learned I_n.
- Method 5: LASSO model on "optimal" design with full I_{max}.
- Method 6: LASSO model on regular design with full I_{max}.

Each time, we plot the evolution of the IMSE with the number of experiments n for the six methods, we also show the density of designs. Remark that all methods except Method 6 are dependent on our algorithm (I_n or $\mathbf{x}_n$) and our result will be compared to the standard Method 6. Moreover, we present Method 5 for sake of completeness even if our main contributions are Methods 1 to 4.

Note at last that the LASSO procedure has been run with a cross validation procedure to compute the best sparsity parameter t.[1].

[1]We use the implementation of the LASSO described in [8] downloadable at http://www.applied-mathematics.net/download.php?id=45

5.3. Results

Function η_1. In figure 7.6 we put the histogram of the selected design points when we run 1000 Monte Carlo simulations for 20 iterations. We choose to restrict us to the first 20 iterations since we want to exhibit the most important experiments performed to approach η_1.

We can remark that our algorithm choose to measure the signal in the neighborhood of the changing point $1/2 + m\pi/2k$ of the Sinus Cardinal function η_1 ($m \in \mathbb{Z}$). The algorithm appears to localize the important "changing point" of the signal.

Moreover, the next figures 7.7 and 7.8 illustrate the good behavior of our method following the evolution of the Integrated Mean Square Error Rate. We remark that in both cases of low or high variance ($\sigma = 0.5$ or $\sigma = 2$), the quadratic loss is always decreasing with the number of experiments for Methods 1/3, this result was not so obvious since I_n can be modified each time n is increased (for instance the IMSE of Method 6 can increase for some iteration). We also remark that using an optimized LASSO (with k-fold cross validation) based on the design $\mathbf{x}_n$ and a basis defined *via* I_n has an equivalent IMSE.

This confirms the usefulness of the selection of functions in I_n to obtain good interpolation results. Method 2 is completely outperformed by the Method 1 and 3. This is not surprising since a regular design may not be adapted to an irregular structure of I_n. The linear model can be really bad-conditioned when the resolution of some functions in I_n is high whereas the design is not adapted to these high resolution functions.

The LASSO algorithm runs on our basis I_n with a regular design (Method 4) is generally better than Method 2 since it solves the problem of bad-conditioned linear models with an automatic deletion of the high resolution functions which yield bad conditioned linear systems. But this point is false when we use a LASSO algorithm on the full basis function (maximum resolution) on a non-regular design. Indeed, Method 5 was the worse of the interpolation algorithms we used. At last, the LASSO on regular design and full basis functions (Method 6) performs generally well when the number of experiment is not too small (at last 60 experiments), but is completely outperformed by Method 1 or 3 for small number of experiments.

Note also that on the example of the Sinus Cardinal signal, the variance term σ does not seem to have a real influence on the ranking of the methods. Of course, the IMSE is better when σ is small, but methods 1 and 3 seem to be the best among all the proposed algorithms.

Function η_2. The same conclusions can be drawn following the results described in figure 7.9.

Considering now the evolution of the IMSE (figures 7.10 and 7.11) with the number of experiments, we remark here that in the low variance case, our algorithm (method 1) may not be very relevant compared to LASSO interpolation on regular design with a good basis function I_n. But in the case of larger variance term, methods 1 and 3 appear to be the more reliable (see figures 7.12 and 7.13). This is also illustrated considering the interpolation obtained in figures 7.12 and

7.13. At last, the LASSO method computed on the full basis of functions and a regular design remains good for sufficient number of experiment as pointed in figure 7.13 (number of experiments greater than 50). One explanation of the efficiency of methods 1 and 3 in the high variance case is that these methods use a control on the real value of the variance although the LASSO method 4 on regular designs, as all penalized methods, use a penalization heuristic to control the variance term.

5.4. Motorcycle impact experiment

We end the simulation paragraph by using a real dataset of Motorcycle Impact Experiment (see [20] for a brief description of the data). This experiment is designed to measure the efficiency of crash helmets and especially the minimum and the maximum values of the signal. Silverman [20] uses a spline smoothing approach to estimate the underlying curve. One may ask whether the experimenters really need the 133 observations to interpolate the curve response well. We decide to scale the 133 observations between 0 and 1 and we first compute the kernel smoothing interpolation described in [20]. Moreover, we decide to use this interpolation as the "true" response to compare our methods. Indeed, this does not yield the true response but we use it as an indicator of the quality of the design strategies. At last, we randomize the kernel interpolation by an addition of a white noise. We find again a good performance of the methods 1 and 3 compared to other methods. We only show some examples of interpolation obtained with 50 experiments. These results are plotted in figures 7.14 and 7.15. We remark that the results are satisfactory with methods 1 or 3 particularly on the first slope of the signal. Other methods are visually outperformed, when the IMSE is compared to the kernel smoothing approach: the estimated IMSE appears to be around 54 for methods 1 and 3, around 65 for the whole LASSO methods 4 and 6 and greater than 1000 for methods 2 and 5 in the high variance case described in figure 7.15.

6. Conclusion

The adaptive method developed in this paper is working well on numerical and real data examples compared to previous method in the literature. But, on the theoretical side, many questions remain open.

First, it would be very fruitful to generalize the localization result to multi-resolution set of smooth functions.

The problem to know how to handle the discrepancy term λ_{I_n} seems interesting since one can imagine to decrease this penalization term should slowly decrease with the number of experiments but increase when the resolution of one map in I_n is increased as pointed in equation (6).

At last, some future work will address the difficult question of the nature and rate of convergence of the stochastic coupled algorithm (I_n and $\mathbf{x}_n$ evolving). To do so, it is necessary to fix a precise cooling strategy to use the consistency result of theorem 4.

7. Appendix

We will denote l as the number of points in a fixed design $\mathbf{x}$ and p the cardinal of I. We suppose for this the trivial assumption $l+1 \geq p$ and denote F the map given by $F(\xi) = \det(M_{\mathbf{x}\cup\xi})$. We will show that F is a convex map on every interval where she is differentiable. Assuming ξ to be suitably chosen among differentiable points of Λ_I, we will note the $\Lambda_I'(\xi)$ the vector composed of the differentiable maps of Λ_I computed at point ξ and the squared matrix

$$M_\xi' = \left((\Lambda_{i_1}\Lambda_{i_2}' + \Lambda_{i_1}'\Lambda_{i_2})(\xi)\right)_{i_1,i_2\in I} = \frac{\mathrm{d}}{\mathrm{d}\xi}\left(M_{\mathbf{x}\cup\xi}\right).$$

Using the standard euclidean scalar product on $\mathbb{R}^p$, one can check immediately that

$$\forall U \in \mathbb{R}^p \qquad M_\xi' U = \langle \Lambda_I(\xi); U\rangle \Lambda_I'(\xi) + \langle \Lambda_I'(\xi); U\rangle \Lambda_I(\xi).$$

First, we state some classical results on matrices whose proofs are based on standard argument on matrices of rank 1. Some details can be found in [15] and in chapter one of [13].

Proposition 1. *Provided $M_{\mathbf{x}\cup\xi}^{-1}$ and $M_{\mathbf{x}}^{-1}$ are non-singular, they obey the relations*

$$M_{\mathbf{x}}^{-1} = M_{\mathbf{x}\cup\xi}^{-1} + \frac{M_{\mathbf{x}\cup\xi}^{-1}\Lambda_I(\xi)^t\Lambda_I(\xi)M_{\mathbf{x}\cup\xi}^{-1}}{1 -^t \Lambda_I(\xi)M_{\mathbf{x}\cup\xi}^{-1}\Lambda_I(\xi)} \tag{15}$$

$$M_{\mathbf{x}\cup\xi}^{-1} = M_{\mathbf{x}}^{-1} - \frac{M_{\mathbf{x}}^{-1}\Lambda_I(\xi)^t\Lambda_I(\xi)M_{\mathbf{x}}^{-1}}{1 +^t \Lambda_I(\xi)M_{\mathbf{x}}^{-1}\Lambda_I(\xi)}. \tag{16}$$

Moreover, $|M_{\mathbf{x}\cup\xi}|$ and $|M_{\mathbf{x}}|$ satisfies

$$\frac{\det M_{\mathbf{x}\cup\xi}}{\det M_{\mathbf{x}}} = \frac{1}{1 -^t \Lambda_I M_{\mathbf{x}\cup\xi}\Lambda_I} \tag{17}$$

$$\frac{\det M_{\mathbf{x}}}{\det M_{\mathbf{x}\cup\xi}} = \frac{1}{1 +^t \Lambda_I M_{\mathbf{x}}\Lambda_I} \tag{18}$$

We now establish two technical lemmas useful to establish our localization theorem.

Lemma 1. *For any symmetric matrix S, we have the relation*

$$Tr(M_\xi' S) = Tr(SM_\xi') = 2\langle S\Lambda_I(\xi); \Lambda_I'(\xi)\rangle. \tag{19}$$

Proof: Consider first the case where $\{\Lambda_I(\xi), \Lambda_I'(\xi)\}$ are linearly independent in $\mathbb{R}^p$. A short calculus show that

$$M_\xi' S\Lambda_I(\xi) = \langle S\Lambda_I(\xi); \Lambda_I(\xi)\rangle\Lambda_I'(\xi) + \langle S\Lambda_I'(\xi); \Lambda_I(\xi)\rangle\Lambda_I(\xi),$$

and

$$M_\xi' S\Lambda_I'(\xi) = \langle S\Lambda_I(\xi); \Lambda_I'(\xi)\rangle\Lambda_I'(\xi) + \langle S\Lambda_I'(\xi); \Lambda_I'(\xi)\rangle\Lambda_I(\xi).$$

Since the rank of M_ξ' is 2, we can find a basis adapted to the family $(\Lambda_I(\xi); \Lambda_I'(\xi))$ such that the endomorphism described by $M_\xi' S$ in the basis is

$$\begin{pmatrix} \langle S\Lambda_I'(\xi); \Lambda_I(\xi)\rangle & \langle S\Lambda_I(\xi); \Lambda_I(\xi)\rangle & 0 & \dots & 0 \\ \langle S\Lambda_I'(\xi); \Lambda_I'(\xi)\rangle & \langle S\Lambda_I(\xi); \Lambda_I'(\xi)\rangle & 0 & \vdots & 0 \\ 0 & \dots & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & 0 \\ 0 & \dots & 0 & \dots & 0 \end{pmatrix}.$$

Thus in this case

$$Tr(M'_\xi S) = 2\langle S\Lambda_I(\xi); \Lambda'_I(\xi)\rangle.$$

Suppose now that $\{\Lambda_I(\xi), \Lambda'_I(\xi)\}$ are linearly dependent, from this assumption we get

$$\langle S\Lambda_I(\xi); \Lambda_I(\xi)\rangle\Lambda'_I(\xi) = \langle S\Lambda'_I(\xi); \Lambda_I(\xi)\rangle\Lambda_I(\xi),$$

and applying the same argument as above with the endomorphism $M'_\xi S$ whose rank is one in this case, we also obtain

$$Tr(M'_\xi S) = 2\langle S\Lambda_I(\xi); \Lambda'_I(\xi)\rangle. \qquad \square$$

For sake of simplicity, we will omit the parameter Λ_I of the information matrices M written in the next two proofs. If we denote by $Com(M)$ the matrix ${}^t cof(M)$, where $cof(M)$ is the matrix of cofactors of A, we have the following result.

Lemma 2. *Assume ξ to be a regular point for the map Λ_I, and that $M_{\mathbf{x}}$, $M_{\mathbf{x}\cup\xi}$ are non-singular, then*

$$Tr\left({}^t Com(M_{\mathbf{x}\cup\xi})M'_\xi\right) = Tr\left({}^t Com(M_{\mathbf{x}})M'_\xi\right).$$

Proof: Using lemma 1 applied first to $S = M^{-1}_{\mathbf{x}\cup\xi}$, we get

$$Tr(M^{-1}_{\mathbf{x}\cup\xi}M'_\xi) = 2\langle M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I; \Lambda'_I\rangle. \tag{20}$$

Moreover, lemma 1 applied now to $S = M^{-1}_{\mathbf{x}\cup\xi}\Lambda^t_I\Lambda_I M^{-1}_{\mathbf{x}\cup\xi}$ yields

$$Tr\left(M^{-1}_{\mathbf{x}\cup\xi}\Lambda^t_I\Lambda_I M^{-1}_{\mathbf{x}\cup\xi}M'_\xi\right) = 2\langle M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I \underbrace{{}^t\Lambda_I M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I}_{=\langle\Lambda_I; M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I\rangle}; \Lambda'_I\rangle.$$

Thus

$$Tr\left(M^{-1}_{\mathbf{x}\cup\xi}\Lambda^t_I\Lambda_I M^{-1}_{\mathbf{x}\cup\xi}M'_\xi\right) = 2\langle\Lambda_I; M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I\rangle\langle\Lambda'_I; M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I\rangle. \tag{21}$$

From (15),(20) and (21), we get

$$\begin{aligned}
Tr(M^{-1}_{\mathbf{x}}M'_\xi) &= 2\langle M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I; \Lambda'_I\rangle + \frac{2\langle\Lambda_I; M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I\rangle\langle\Lambda'_I; M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I\rangle}{1 -^t \Lambda_I M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I} \\
&= 2\langle M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I; \Lambda'_I\rangle\left(1 + \frac{\langle\Lambda_I; M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I\rangle}{1 -^t \Lambda_I M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I}\right) \\
Tr(M^{-1}_{\mathbf{x}}M'_\xi) &= \frac{Tr(M^{-1}_{\mathbf{x}\cup\xi}M'_\xi)}{1 -^t \Lambda_I M^{-1}_{\mathbf{x}\cup\xi}\Lambda_I}.
\end{aligned}$$

Now, use (16) and the relation $A^{-1} = \frac{{}^t(Com(A))}{\det(A)}$ to reach the conclusion of the lemma:

$$Tr\left({}^t Com(M_{\mathbf{x}\cup\xi})M'_\xi\right) = Tr\left({}^t Com(M_{\mathbf{x}})M'_\xi\right). \quad \square$$

Proof of theorem 1: We will note $F(\xi) = \det\left(M_{\mathbf{x}\cup\xi}\right)$. Suppose first that $M_{\mathbf{x}}$ is non-singular and ξ is not a dyadic point described by the set $\mathcal{E}$. In this case, classical differentiation used with lemma 2 yields

$$F'(\xi) = Tr\left({}^t Com(M_{\mathbf{x}\cup\xi})M'_\xi\right) = Tr\left({}^t Com(M_{\mathbf{x}})M'_\xi\right).$$

Finally, since Tr is a linear map, we immediately get

$$F"(\xi) = Tr\left({}^tCom(M_{\mathbf{x}})M"_{\xi}\right) = Tr\left({}^tCom(M_{\mathbf{x}})\Lambda'_I\,{}^t\Lambda'_I\right) = {}^t\Lambda'_I\,{}^tCom(M_{\mathbf{x}})\Lambda'_I \geq 0.$$

Thus F is a convex function on each interval outside of $\mathcal{E}$ Consequently, its maximum are located on some dyadic points of $\mathcal{E}$. This is equivalent to the assertion of the proposition.

Suppose now $M_{\mathbf{x}}$ is singular, we can find a sequence $M_{\mathbf{x},\epsilon_n} = M_{\mathbf{x}} + \epsilon_n Id$ which is non-singular such that

$$\lim_{n\mapsto+\infty} M_{\mathbf{x}} + \epsilon_n Id = M_{\mathbf{x}}.$$

Consider now the function $F_{\epsilon_n}(\xi)$ defined as

$$F_{\epsilon_n}(\xi) = \det\left(M_{\mathbf{x},\epsilon_n\cup\xi}\right)$$

We can use the same arguments as before to conclude that $\arg\max F_{\epsilon_n} \subset \mathcal{E}$ since these arguments only rely on a slight modification of lemma 2 which becomes:

$$F'_{\epsilon_n}(\xi) = Tr\left({}^tCom(M_{\mathbf{x},\epsilon_n\cup\xi})M'_{\xi}\right) = Tr\left({}^tCom(M_{\mathbf{x},\epsilon_n})M'_{\xi}\right).$$

Now, remark that $\mathcal{E}$ is a finite set which is not varying with ϵ_n and

$$\forall\xi \qquad F_{\epsilon_n}(\xi) \leq \max_{x\in\mathcal{E}} F_{\epsilon_n}(x).$$

Taking the limit in the relation above yields the conclusion of the proof. □

Proof of theorem 3: Remark first that $t \mapsto \det\left(tId + M^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I)\right)$ is a polynomial function of t whose degree p is the size of Λ_I. This polynomial function is developed in

$$\det\left(tId + M^{-1}_{\mathbf{x}\cup\xi}(\Lambda_I)\right) = t^p - Tr\left(M^{-1}_{\mathbf{x}\cup\xi_1}(\Lambda_I)\right)t^{p-1} + Q_{\xi}(t)$$

where $deg(Q_{\xi}) \leq p-2$. Now for $\xi_1, \xi_2 \in E$ satisfying

$$Tr\left(M^{-1}_{\mathbf{x}\cup\xi_1}(\Lambda_I)\right) \geq Tr\left(M^{-1}_{\mathbf{x}\cup\xi_2}(\Lambda_I)\right),$$

we can immediately check that for sufficiently large t, we have

$$\det\left(tId + M^{-1}_{\mathbf{x}\cup\xi_1}(\Lambda_I)\right) \leq \det\left(tId + M^{-1}_{\mathbf{x}\cup\xi_2}(\Lambda_I)\right).$$

Consequently, the solutions of the trace maximization problem are the same as the one deduced from the determinant minimization problem and this remark ends the first point. To get the more general second conclusion, we just have to apply in a similar way conjecture 2. □

Remark 6. *To extend now the proof to higher dimensions with some tensorized family of Schauder functions, one just have to remark that both lemma 1 and 2 are still valid. Then a similar argument to the one used in the proof of theorem 1 shows the convexity of F except in the neighborhood of dyadic points.*

Proof of theorem 4: This proof is largely inspired from [16], himself directly related to theorem 1 of [17] which states the almost sure convergence of $\hat{\alpha}_n$ to α provided the two conditions

C1 $\lambda_{\min}\left[M_{\mathbf{x_n}}\right] \to \infty \qquad a.s.$

C2 $\log\left(\lambda_{\max}\left[M_{\mathbf{x_n}}\right]\right) = o\left(\lambda_{\min}\left[M_{\mathbf{x_n}}\right]\right) a.s.$

where $\lambda_{\min}(M)$ denotes the minimum eigenvalue of M and $\lambda_{\max}(M)$ the maximum eigenvalue of M.

We establish first the condition C1. Remark first that as the map $(\Lambda_i)_{i\in I}$ are linearly independent, we can find $\rho > 0$ such that

$$B(0,\rho) \subset \underbrace{\overline{Conv(\Lambda_I(t), t \in E)} \cup \overline{-Conv(\Lambda_I(t), t \in E)}}_{:=\mathcal{G}},$$

where $Conv$ denotes the convex hull of a set. Now, we have for any symmetric positive definite M

$$\max_{y\in B(0,\rho)} {}^t y M^{-1} y = \lambda_{min}(M)^{-1}\rho^2,$$

and since $y \mapsto {}^t y M^{-1} y$ is convex, we can state that

$$\max_{x\in E} {}^t\Lambda_I(x) M^{-1} \Lambda_I(x) \geq \frac{\rho^2}{\lambda_{min}(M)}. \tag{22}$$

Remark that all maps in Λ_I are continuous and E is compact, thus

$$\exists\, L > 0 \quad \forall t \in E \qquad \left\| {}^t\Lambda_I(t)\Lambda_I(t) \right\|_2 \leq L,$$

where $\|A\|_2 := \sup_{x\in B(0,1)} \|Ax\|$, where we take the Euclidean norm in the last definition. Now, the spectral radius satisfies the triangular inequality and

$$\lambda_{\max}\left(\frac{M_{\mathbf{x_k}}}{k}\right) \leq \frac{\sum_{i=1}^{k} \lambda_{\max}(\Lambda_I(\xi_i)^t\Lambda_I(\xi_i))}{k} \leq L.$$

Defining $I_k = M_{\mathbf{x_k}}/k$, the last inequality yields

$$\lambda_{\max}(I_k) \leq L. \tag{23}$$

Next define $\rho_k = \det(I_k)$ and $d_k(t) = {}^t\Lambda_I(t) I_k^{-1} \Lambda_I(t)$, from proposition 1 equation (18), we have

$$\rho_{k+1} = \left(\frac{k}{k+1}\right)^p \left(1 + \frac{d_k(\xi_{k+1})}{k}\right) \rho_k \geq \rho_k \left(\frac{k}{k+1}\right)^p.$$

Thus, for any $\epsilon > 0$, we can find $K_1 \geq 1$ such that

$$\forall k \geq K_1 \qquad \rho_{k+1} \geq (1-\epsilon)\rho_k, \tag{24}$$

and a simple induction shows that $\rho_k \geq (1-\epsilon)^{k-K_1}\rho_{K_1}$. Let $A_k = (1-\epsilon)^{k-K_1}\rho_{K_1}$, since $A_k \to 0$ as $k \to \infty$, we can find $K_2 \geq K_1$ such that

$$\forall k \geq K_2 \qquad \frac{\rho^2}{A_k^{1/p}} > 2p \qquad \text{and} \qquad \left(\frac{k+1}{k}\right)^p \leq 1 + \frac{2p}{k}. \tag{25}$$

We show now by induction that ρ_k is bounded from below by $(1-\epsilon)A_{K_2}$ for sufficiently big k. This is obviously true for $k = K_2 + 1$.

Suppose now that $\rho_k \geq (1-\epsilon)A_{K_2}$. If $\rho_k \geq A_{K_2}$, in view of (24) we immediately obtain $\rho_{k+1} \geq (1-\epsilon)A_{K_2}$. We must thus study the case $A_{K_2} > \rho_k \geq (1-\epsilon)A_{K_2}$. From the definition of d_k and (22), we have

$$\max_{x\in E} d_k(x) \geq k\frac{\rho^2}{\lambda_{\min}(M_{\mathbf{x_k}})} \geq k\frac{\rho^2}{\det(M_{\mathbf{x_k}})^{1/p}} \geq \frac{\rho^2}{\rho_k^{1/p}}.$$

From equation (25) and our assumption on ρ_k, we obtain

$$\max_{x\in E} d_k(x) \geq \frac{\rho^2}{A_{K_2}^{1/p}} > 2p$$

Finally, the definition of ξ_{k+1} yields

$$\rho_{k+1} = \rho_k \left(\frac{k}{k+1}\right)^p \left(1 + \frac{d_k(\xi_{k+1})}{k}\right) = \rho_k \left(\frac{k}{k+1}\right)^p \left(1 + \frac{\max_{x\in E} d_k(x)}{k}\right) \geq \rho_k$$

This last inequality ends the induction and ρ_k is bounded from below by a constant Γ. Now, remark that

$$\lambda_{\min}(I_k)\lambda_{\max}(I_k)^{p-1} \geq \det(I_k) \geq \Gamma,$$

and we obtain from equation (23)

$$\lambda_{\min}(M_{\mathbf{x_k}}) \geq k\frac{\Gamma}{L^{p-1}} \to +\infty \qquad \text{as} \qquad k \to +\infty,$$

this last equation proves condition (C1).

Regarding condition (C2), simple algebra yields

$$\frac{\lambda_{\min}(M_{\mathbf{x_k}})}{\log\left(\lambda_{\max}(M_{\mathbf{x_k}})\right)} \geq \frac{k\Gamma}{L^{p-1}\log\left(kL\right)} \to \infty \qquad \text{as} \qquad k \to +\infty,$$

and this last equation proves condition (C2).

With notation of theorem 1 of [17], take $\delta = 0$ and apply now this theorem to conclude that

$$\|\hat{\alpha}_n - \alpha\|_\infty = O\left(\left[\frac{\log\left(\lambda_{\max}(M_{\mathbf{x_k}})\right)}{\lambda_{\min}(M_{\mathbf{x_k}})}\right]^{1/2}\right) = O\left(\sqrt{\frac{\log n}{n}}\right) \qquad \square$$

References

[1] Box G. E. P. and Draper N. R. (1959). A Basis for the Selection of a Response Surface Design. *Journal of the American Statistical Association*, 54(287) 622–654.

[2] Kiefer J. and Wolfowitz J. (1959). Optimum designs in regression problem. *Annals of Mathematical Statistics*, 30, 271–294.

[3] Friedman J. H (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19,1–141.

[4] Khuri A.I. and Cornell J.A. (1987), Response Surfaces: Designs and Analyses . *New York : Marcel Dekker, Inc.*

[5] Dette H. and Studden W. J. (1997). The theory of canonical moments with applications in statistics, probability, and analysis. *Wiley Series in Probability and Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1997.*

[6] Meyer Y. (1990). Ondelettes et OpÈrateurs I. *Hermann*.

[7] Donoho D. L. and Johnstone I. M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*.

[8] Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (B)*, 58 (1), 267–288.

[9] Efron B., Johnstone I., Hastie T. and Tibshirani R. (2003). Least angle regression. *The Annals of Statistics*, 32(2), 407-499.

[10] Candes E. and Tao T. (2007). The Dantzig selector: Statistical estimation when p is much smaller than n. *The Annals of Statistics*, 35(6), 2313–2351.

[11] Hlawka E. (1961). Funktionen von beschr‰nkter Variatiou in der Theorie der Gleichverteilung. *Annali di Matematica Pura ed Applicata*, 54(1), 325–333.

[12] Koksma J.F. (1942).Een algemeene stelling uit de theorie des gelijkmatige verdeeling modulo1. *Mathematica B (Zutphen)*, 11, 7–11.

[13] Fedorov V.V. (1969), Theory of Optimal Experiments. *Academic Press, New York*.

[14] Karlin S. and Studden W. J. (1966). Optimal experimental designs. *Annals of Mathematical Statistics*, 37, 783–815.

[15] Meyer R. K. and Nachtsheim C. J. (1995). The Coordinate-Exchange Algorithm for Constructing Exact Optimal Experimental Designs. *Technometrics*, 37(1), 60–69.

[16] Pronzato L. (2000). Adaptive optimization and D-optimum experimental design. *The Annals of Statistics*, 28 (6), 1743–1761.

[17] Lai T.L. and Wei C.Z.. (1982). Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10 (1), 154–166.

[18] Geman S. and Geman D. (1984). Stochastic Relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6), 721–741.

[19] Green P.J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82, 711–732.

[20] Silverman B.W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *Journal of the Royal Statistical Society, Series B*, 47, 1–52.

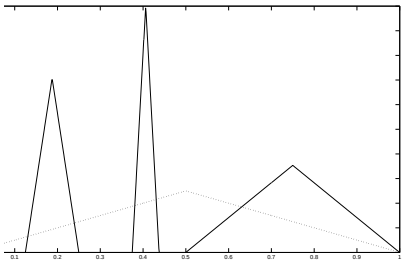

Figure 7.1: Several functions $\Lambda_{j,k}$, here (j,k) equals to $\{(0,0);(1,1)(3,1)(4,6)\}$.

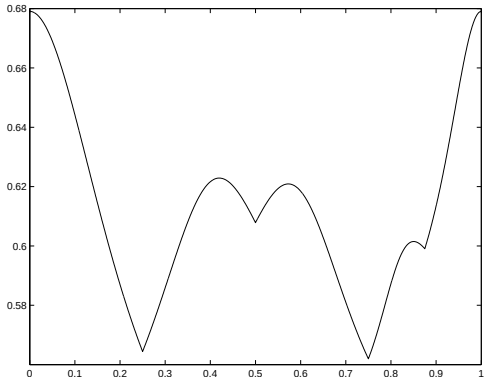

Figure 7.2: Evolution with respect to ξ of the variance term while $I = \{(0,0);(1,0);(1,1);(2,3)\}$.

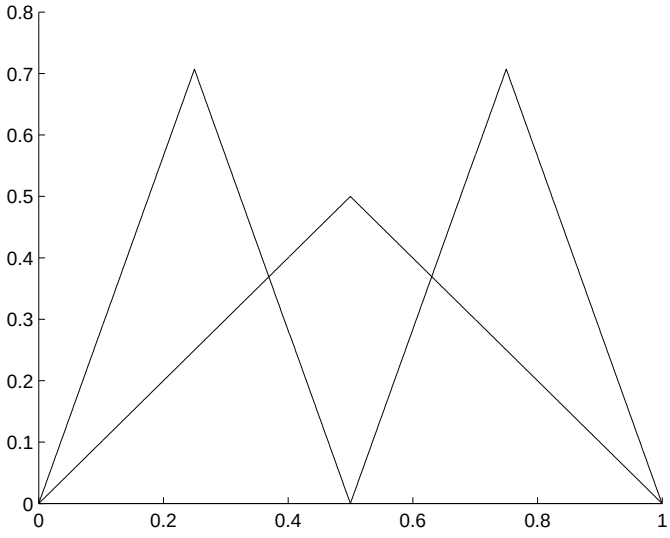

Figure 7.3: Several functions $\Lambda_{j,k}$, here (j,k) equals to $\{(0,0);(1,1)(3,1)\}$.

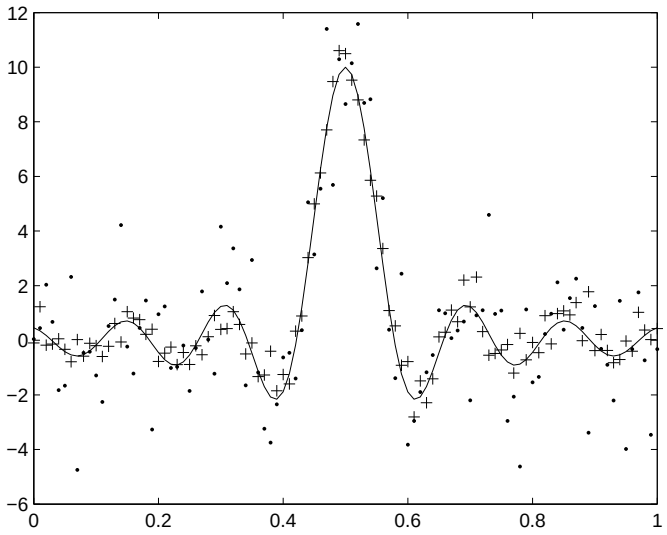

Figure 7.4: Function η_1 with $a = 10, k = 40$ and some realizations of $f_1(x)$ with $\sigma = 0.5$ (crossed curve) or $\sigma = 2$ (dashed curve).

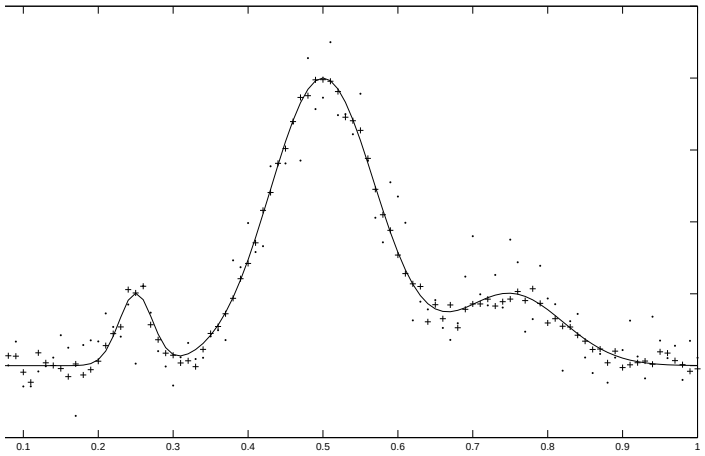

Figure 7.5: Function η_2 with and some realizations of $f_2(x)$ with $\sigma = 0.5$ (crossed curve) or $\sigma = 2$ (dashed curve).

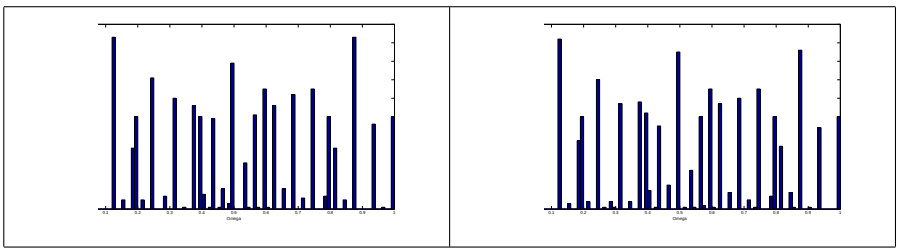

Figure 7.6: Mean Design selected by our algorithm for the estimation of η_1 $(a = 10, k = 40)$ among the first 20 iterations for $\sigma = 0.5$ (left) and $\sigma = 2$ (right).

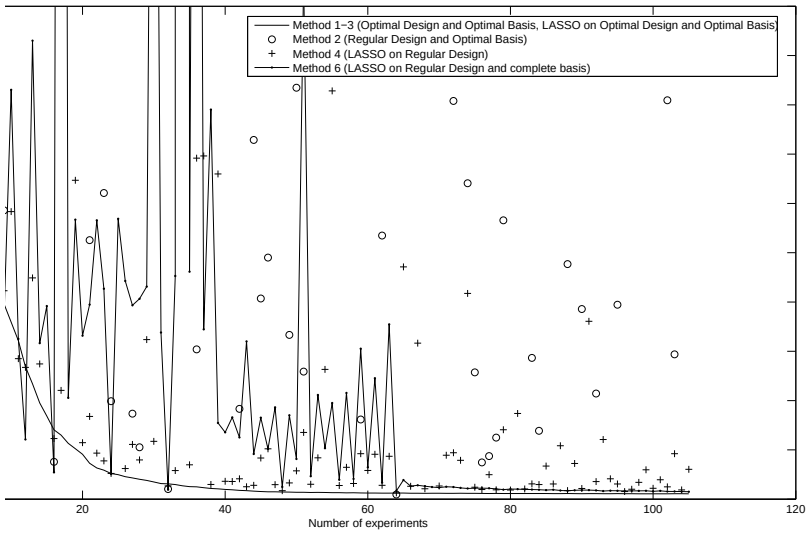

Figure 7.7: Evolution of the IMSE for the estimation of η_1 ($a = 10, k = 40$, $\sigma = 0.5$) with the number of experiments for 5 of the 6 methods listed above method 5 is omitted).

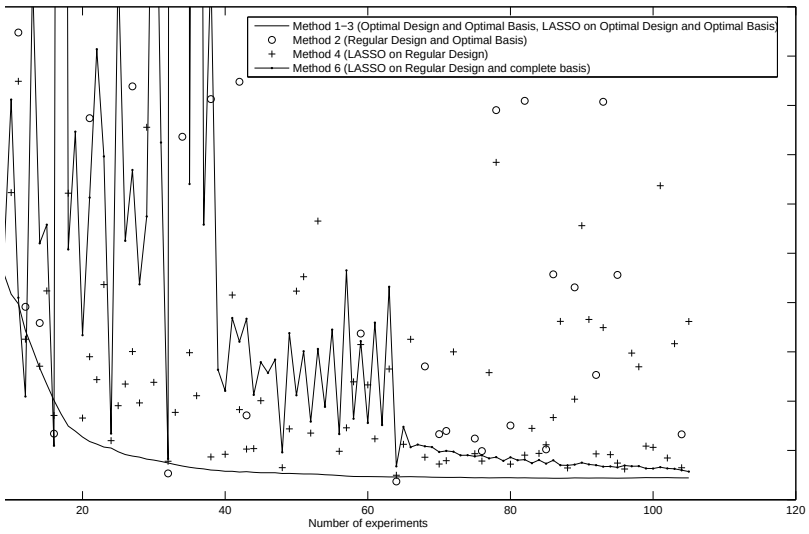

Figure 7.8: Evolution of the IMSE for the estimation of η_1 ($a = 10, k = 40$, $\sigma = 2$) with the number of experiments for 5 of the 6 methods listed above method 5 is omitted).

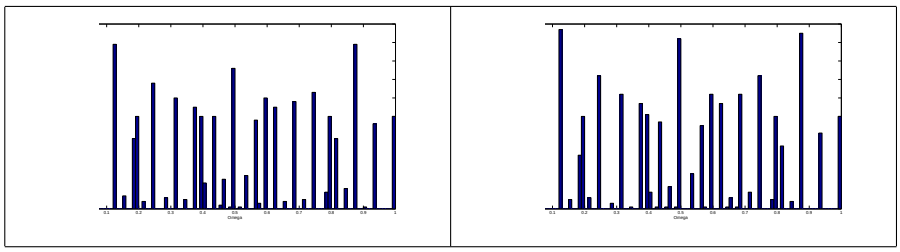

Figure 7.9: Mean Design selected by our algorithm for the estimation of η_2 among the first 20 iterations for $\sigma = 0.5$ (left) and $\sigma = 2$ (right).

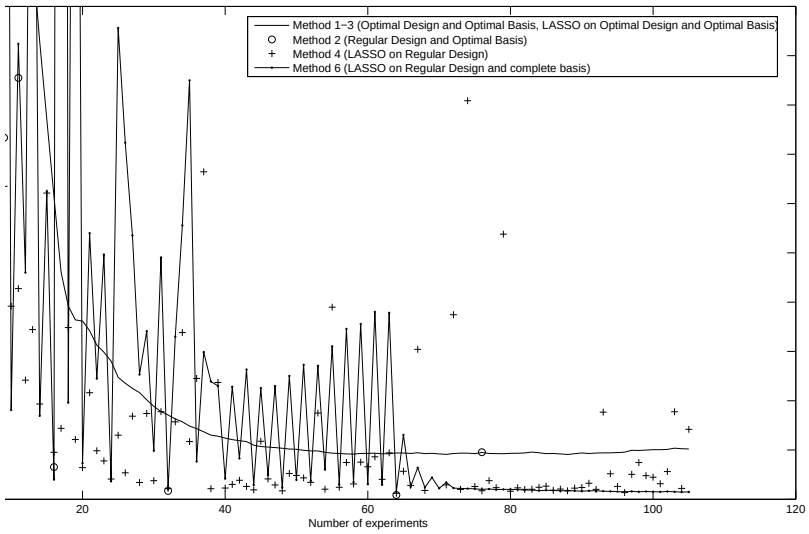

Figure 7.10: Evolution of the IMSE for the estimation of η_2 ($\sigma = 0.5$) with the number of experiments for 5 of the 6 methods listed above method 5 is omitted).

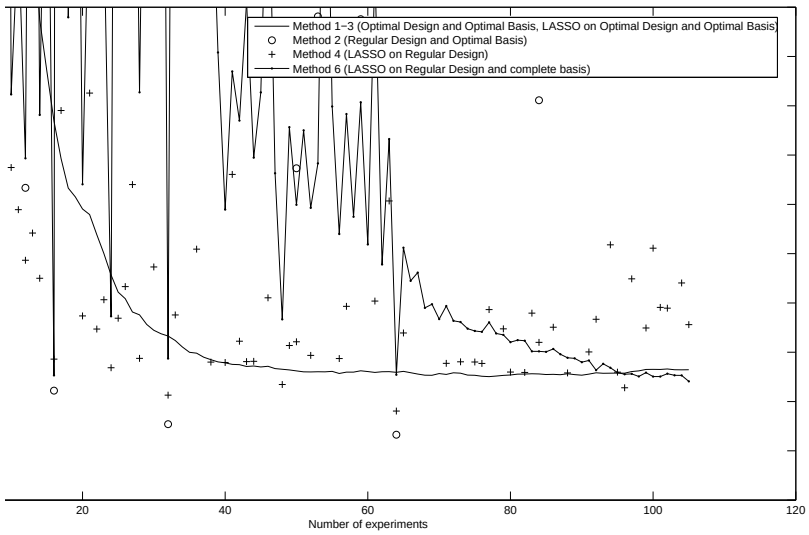

Figure 7.11: Evolution of the IMSE for the estimation of η_2 ($\sigma = 2$) with the number of experiments for 5 of the 6 methods listed above method 5 is omitted).

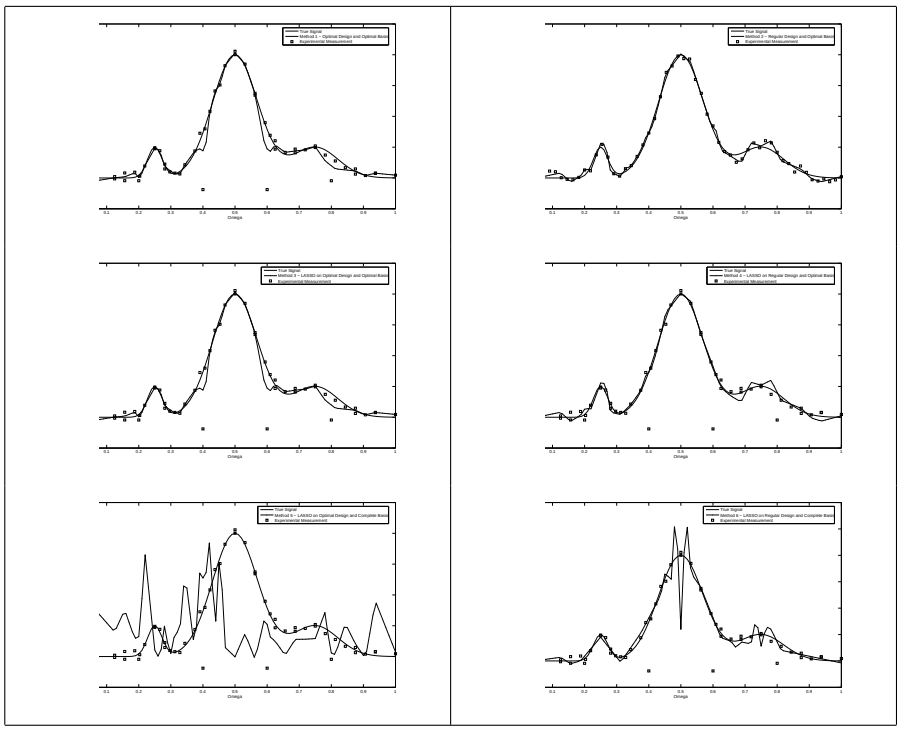

Figure 7.12: Interpolation of the Mixture Signal η_2 using the 6 methods and a low variance term $\sigma = 0.5$.

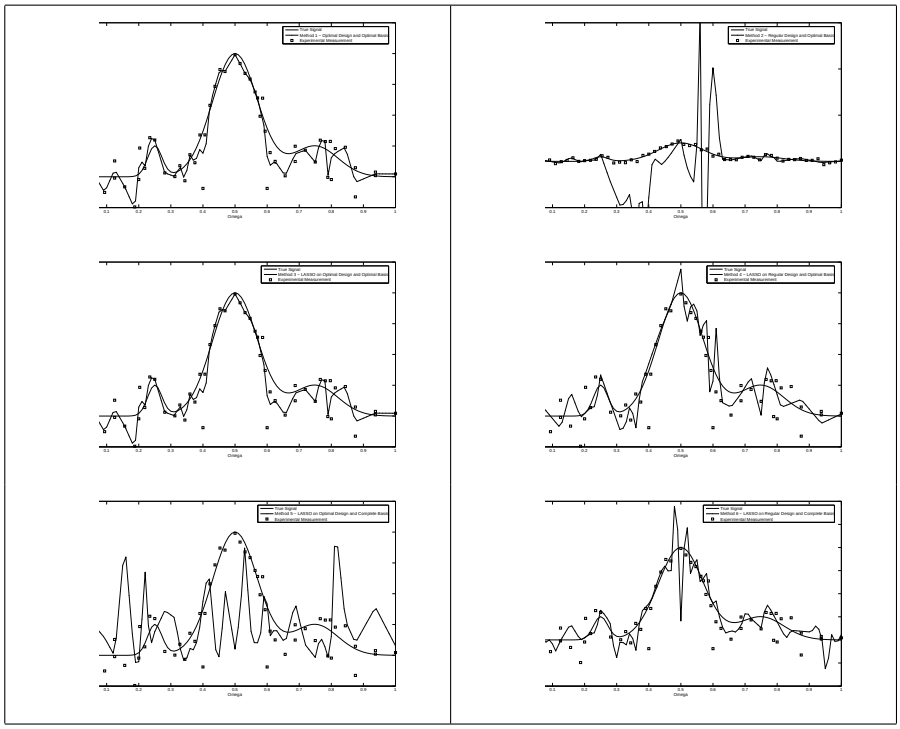

Figure 7.13: Interpolation of the Mixture Signal η_2 using the 6 methods and a high variance term $\sigma = 2$.

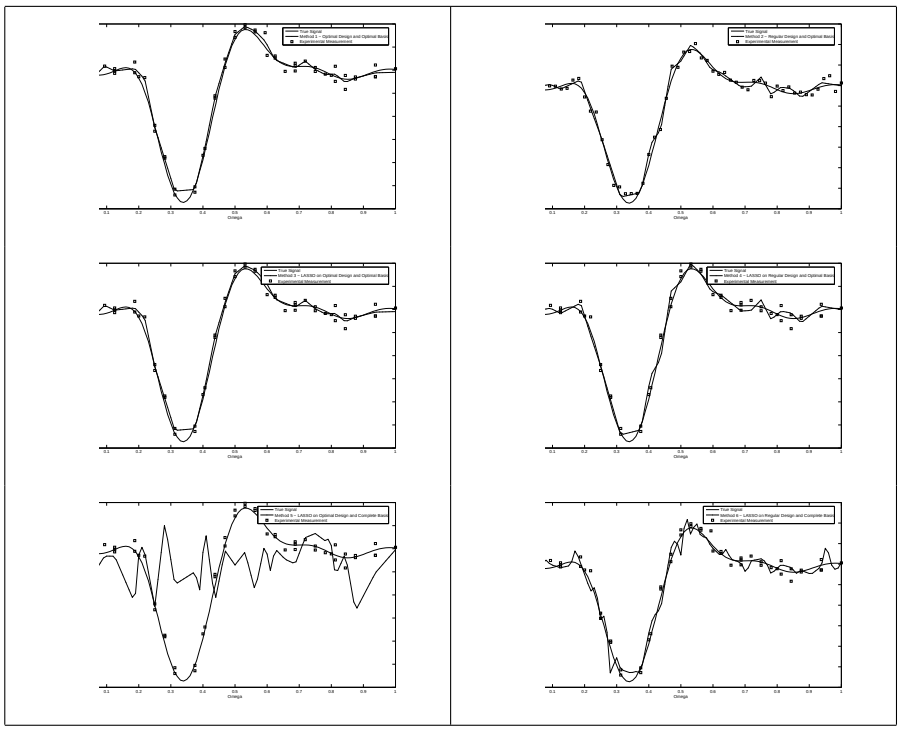

Figure 7.14: Interpolation of the Motorcycle Signal using the 6 methods and a low variance term $\sigma = 5$.

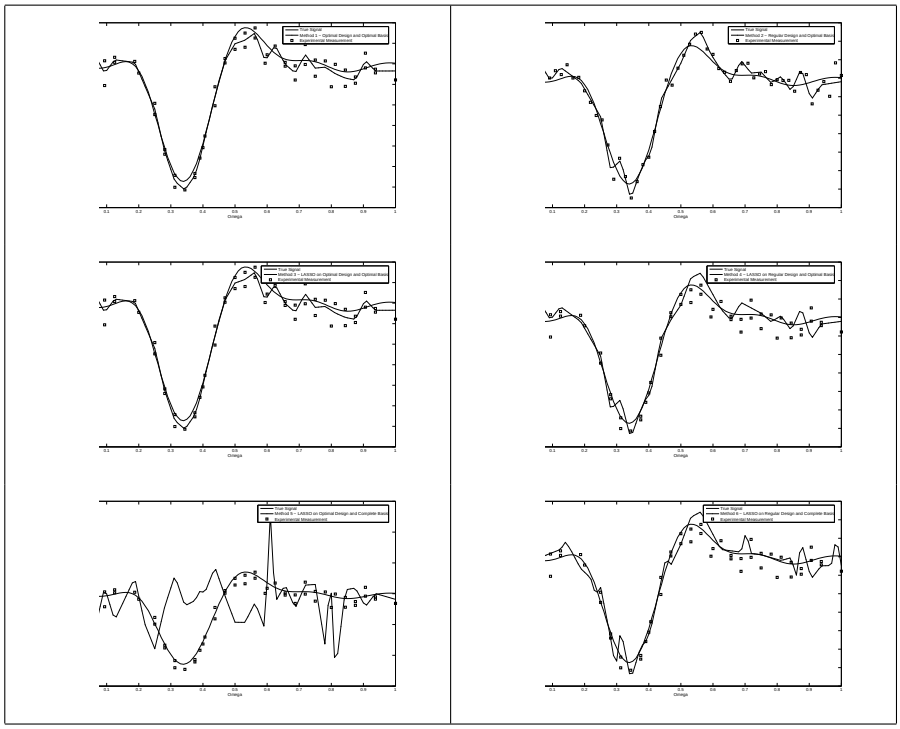

Figure 7.15: Interpolation of the Motorcycle Signal using the 6 methods and a high variance term $\sigma = 10$.