PRIM related correlations

Axel Mathéi

Montefiore Institute - University of Liège

19th Belgian Workshop on Mathematical Optimization

7th March 2013

Table of contents

Problem statement

An industrial decision maker owns a production line having **dozens** of parameters. Currently, he's tuning them by **trial and error**, but wants a program finding the right "zone" to be in.

- Input: Data (parameters + quality)
- Output: Parameter intervals
  - Resulting in the best quality box possible.
  - Limited to a given number of parameters.
  - Containing a minimum number of measures (points).

Interpretability

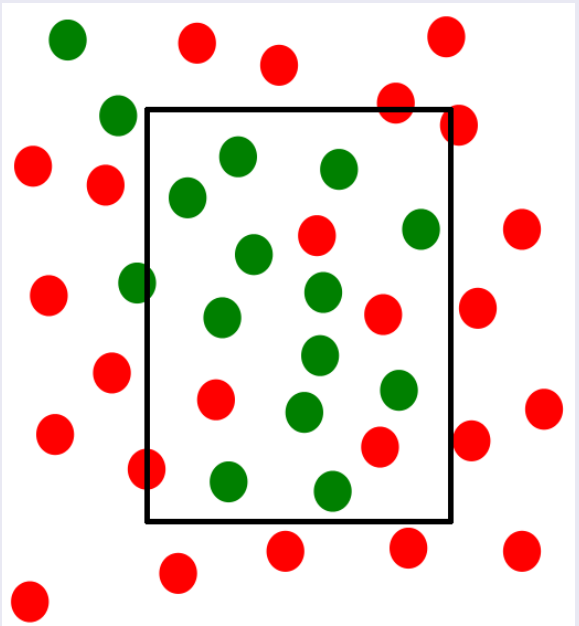
Why boxes?

$\rightarrow$ **Interpretability**

The decision maker prefers to know directly which parameters to controls and in which **ranges** to put them.

That would be impossible with ellipsoid for example.

Box



First idea: Brute force, look at all the possible boxes $\rightarrow \mathcal{O}(d.N^3)$

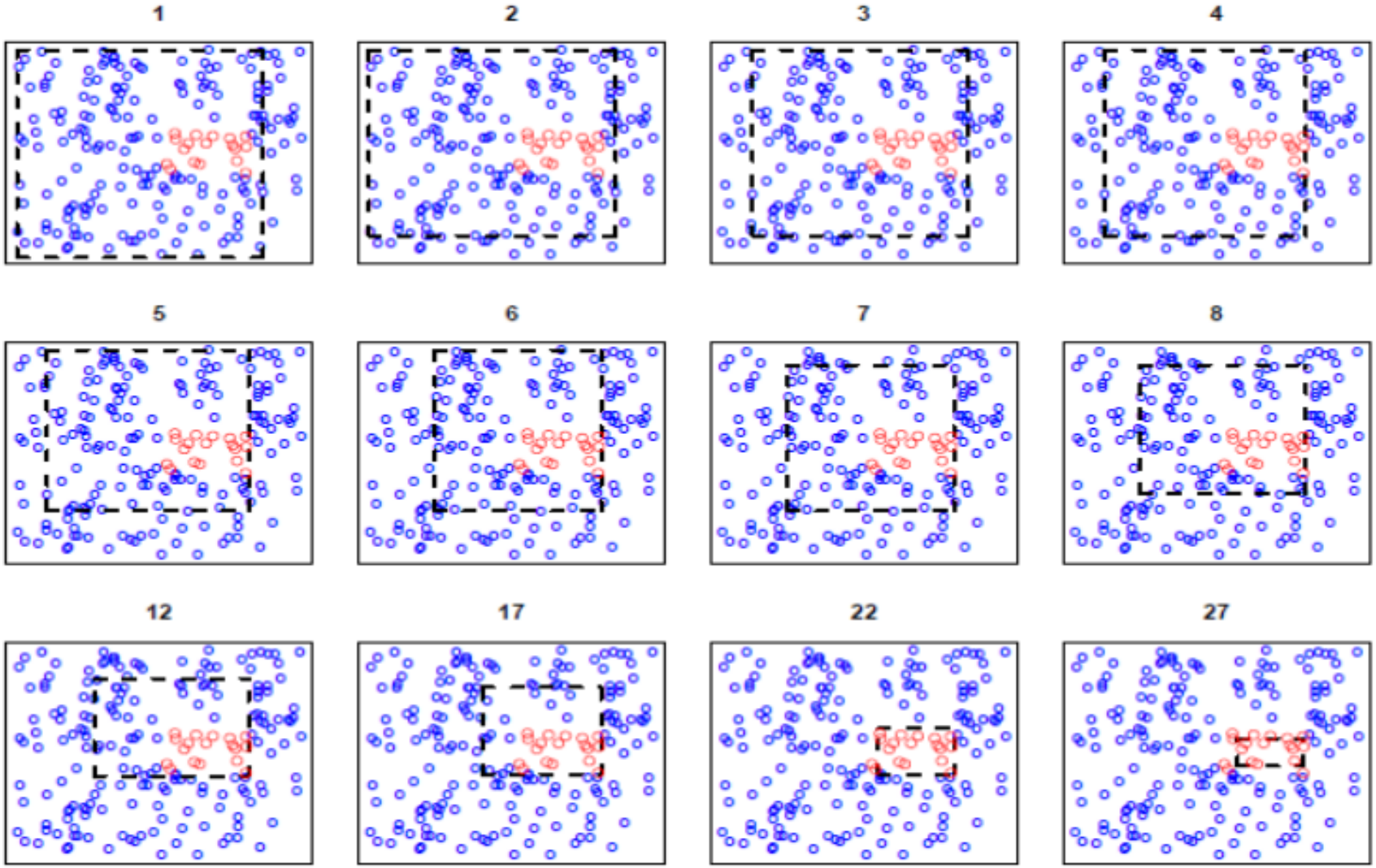
PRIM

Patient **R**ule **I**nduction **M**ethod, *J. Friedman & N. Fisher* (1998)
Idea:

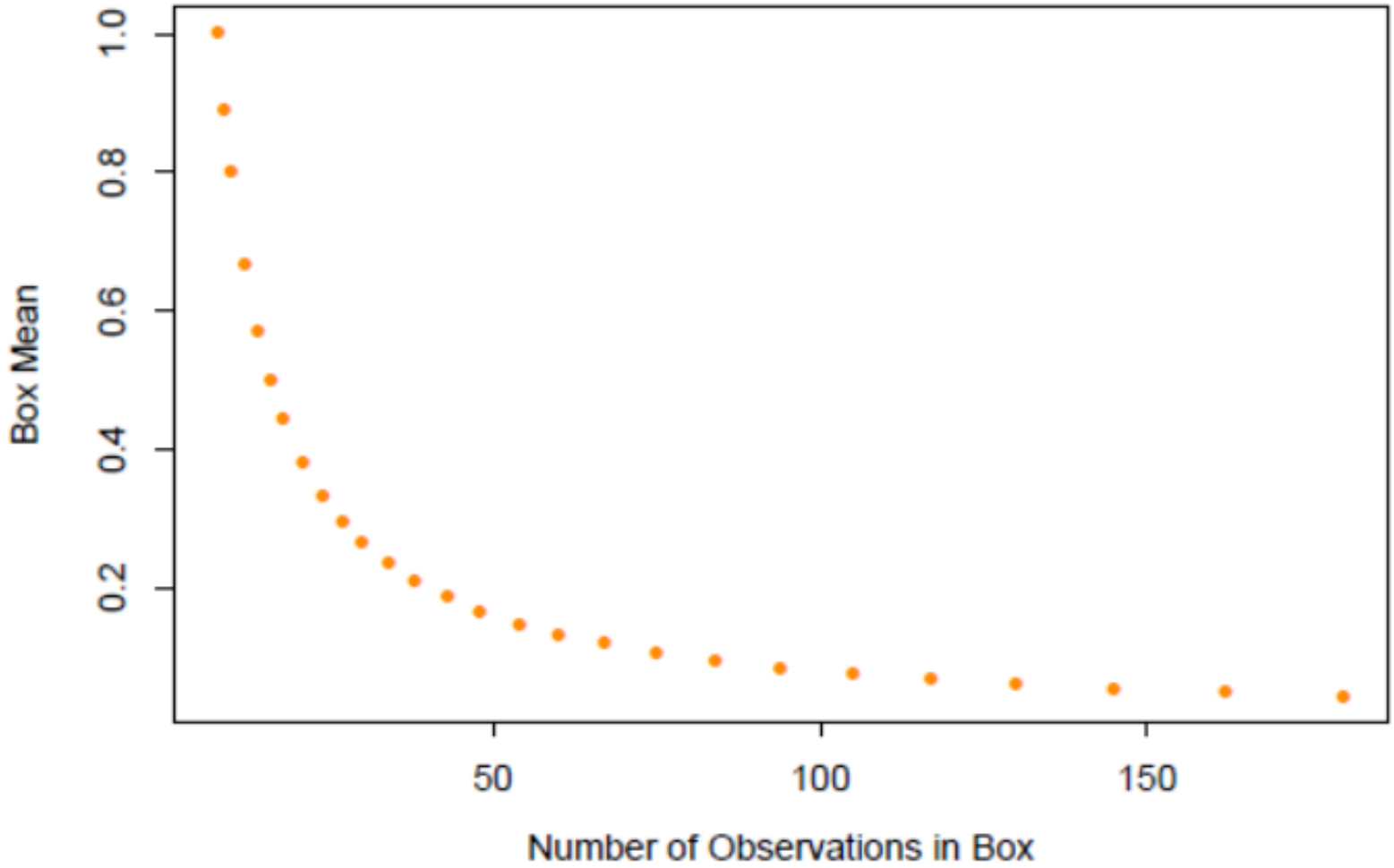
1. Starts with all points.
2. For every dimension, removes a fraction α from the points at an extreme then at the other.
3. Chooses the best among the $2d$ sets generated and restart at step 2 until it remains a fraction β from the starting points.

One can detect many hyperrectangles (or "boxes") by removing the points from the box found and restarting the algorithm on the remaining points.

PRIM - 2D Example



PRIM - Top-down peeling



PRIM - Variables selection

It is difficult to control a lot of parameters at the same time, so we want to limit the number of output intervals.

To select a reduced number of dimensions, we iterate:

1. Release the interval for each dimension, one by one.
2. Remove the dimension which, when released, gives the best resulting box.
3. If we still have too many dimensions, go to step 1.

PRIM - First results

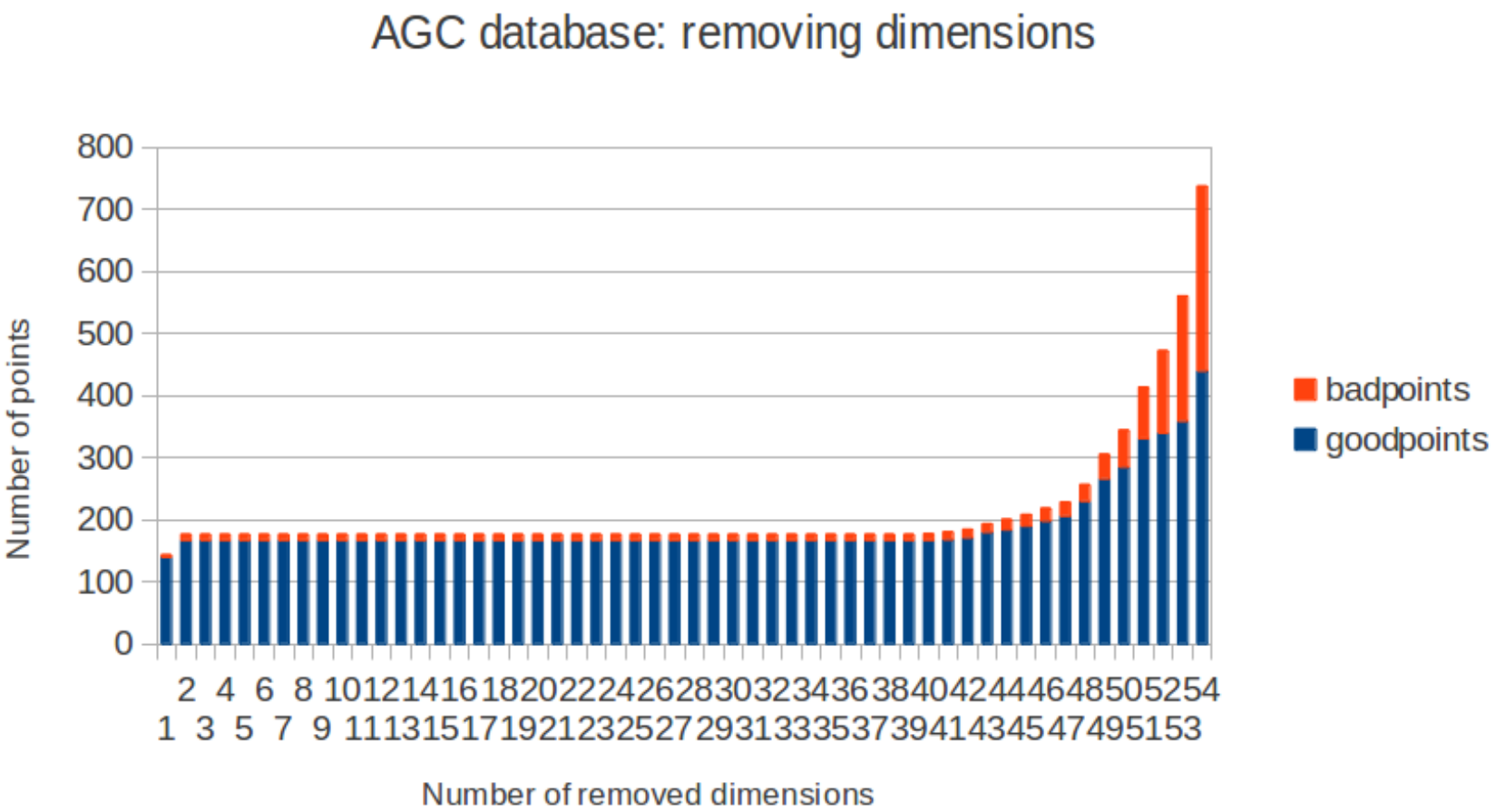
- Database : **AGC**
- Total: 748 points, 440 good ones.
- Dimensions: 53

- Parameters: $\alpha = 5\%$, $\beta = 20\%$

$\Rightarrow$ **5s**, 144 points, 140 good ones $\rightarrow$ **97**%

$\Rightarrow$ 5D limitation : 306 points (266 good ones, 40 bad ones) $\rightarrow$ **87**%

PRIM - First results



Ongoing research - Correlations

When a box is defined by a few dimensions, we could want to exchange one of them keeping a good box, without restarting the algorithm.

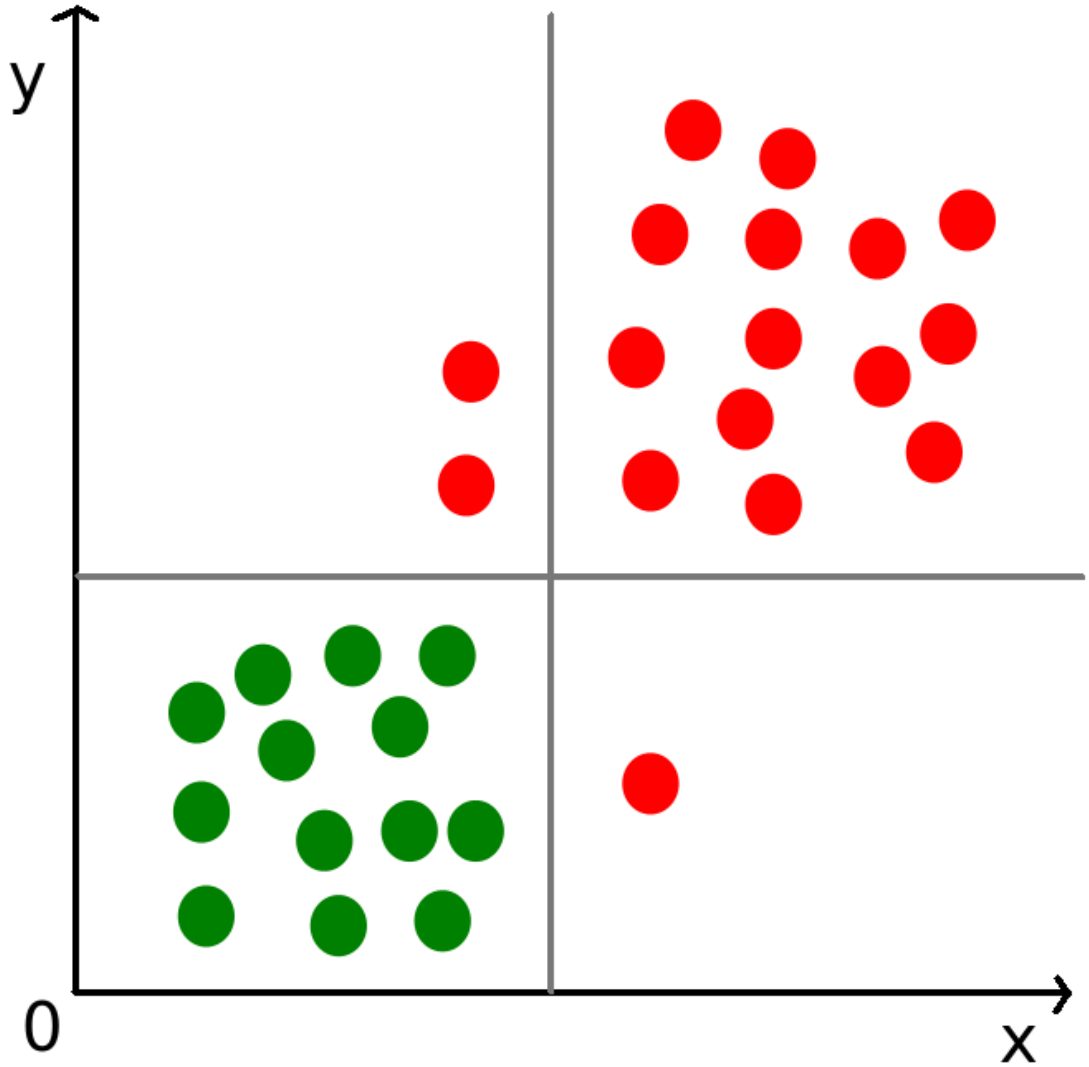
It can be useful if a parameter is more difficult or dangerous to control, and the others are already tuned.

2 kind of scores we can maximize:

- Mean: best possible box
  $\rightarrow$ maximum-density segment problem ($\mathcal{O}(n)$)
- Similarity: X_1 is the constrained box; X_2 is X_1 without the removed constraint; searching a X_2 subarray maximizing:

$$\frac{|X_1 \cap X_2|}{|X_1 \cup X_2|} \tag{1}$$

Ongoing research - Difference between scores



Ongoing research - First results

- Database : marketing (Friedman & Fisher)
- Total: 9409 points on 14 dimensions
- Goal: isolate largest income and swap the most important variable

- Parameters: $\alpha = 5\%$, $\beta = 20\%$, 3 important variables
- Most important variable: dual income

$\Rightarrow$ Mean: Swaping it with "householder status" even increased the score by 3%

$\Rightarrow$ Similarity: Swaping it with "marital status" gave a similarity score of 57%, decreasing the mean score by only 2%

Complexity

- PRIM:
  - 1D projections: $\mathcal{O}(d.N.\log(N))$
  - Top-down peeling: $-\log(N)/\log(1-\alpha)$ steps
- Swap:
  - Mean: $\mathcal{O}(d.N)$
  - Similarity: $\mathcal{O}(d.N^2)$
    For each removed dimension:
    - Intersection: $\mathcal{O}(N)$ (with hash set)
    - Max intersection score subarray: $\mathcal{O}(N^2)$

References

1. Jerome H. Friedman and Nicholas I. Fisher. Bump hunting in high-dimensional data. *Statistics and Computing*, 9(2):123–143, April 1999.

2. Michael H. Goldwasser, Ming-Yang Kao, and Hsueh-I Lu. Linear-time algorithms for computing maximum-density sequence segments with bioinformatics applications. *CoRR*, cs.DS/0207026, 2002.

Thank you for your attention!