

2006 Joint Research Conference on Statistics in  
Quality, Industry, and Technology  
June 7-9, Knoxville, Tennessee

# Bayesian Optimal Design of Choice Experiments

Bradley Jones



**Roselinde Kessels**

Martina Vandebroek



Peter Goos



# Choice experiment in marketing

Which of the two race bicycles would you prefer if the options only differ with respect to the attributes shown?



Carbon frame

Aluminum frame

Classic frame

Sloping frame

Straight fork

Bent fork

Bontrager Race Lite wheels

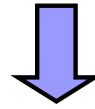
Shimano WH-7701 wheels

Shimano Ultegra groupset

Shimano Dura-Ace groupset



**Choice set**



Which of the two race bicycles would you prefer if the options only differ with respect to the attributes shown?



Aluminum frame

Carbon frame

Sloping frame

Classic frame

Straight fork

Bent fork

Shimano WH-7701 wheels

Mavic Ksyrium SL wheels

Shimano Tiagra groupset

Campagnolo Record groupset



# Setup

- Respondents evaluate several sets of (hypothetical) products or services
- More specifically, respondents indicate the alternative they like most in each choice set  
⇒ 0/1 outcomes
- These alternatives are called profiles
- The profiles are described in terms of (categorical) attributes
- Choice sets typically consist of two, three or four profiles

# Design problem for example

- 15 choice sets of 2 profiles
- 5 attributes:
  - material of the frame (carbon, aluminum)
  - type of frame (classic, sloping)
  - type of fork (straight, bent)
  - type of wheels (3 levels)
  - type of groupset (5 levels)
- $2 \times 2 \times 2 \times 3 \times 5 = 120$  possible profiles
- which  $15 \times 2 = 30$  profiles will be used?
- how will we partition them in sets of two?
- goals:
  - estimate the value respondents attach to the levels of each attribute
  - predict respondents' choices

## Statistical model

- multinomial logit model
- based on the random utilities model

$$U_{js} = \mathbf{x}'_{js} \boldsymbol{\beta} + \varepsilon_{js}$$

where  $\mathbf{x}_{js}$  represents the attribute levels and  $\boldsymbol{\beta}$  is the set of parameter values


- probability of choosing alternative  $j$  in choice set  $s$

$$p_{js} \left( \begin{array}{l} \text{option } j \text{ chosen} \\ \text{in choice set } s \end{array} \right) = \frac{e^{\mathbf{x}'_{js} \boldsymbol{\beta}}}{\sum_{t=1}^J e^{\mathbf{x}'_{ts} \boldsymbol{\beta}}}$$

## *Estimation-based design criteria*

- seek to minimize variances of estimators
- minimize a function of the variance-covariance matrix of the estimators:

$$\begin{aligned}\text{var}(\mathbf{X}, \boldsymbol{\beta}) &= \left( \sum_{s=1}^S \mathbf{X}'_s (\mathbf{P}_s - \mathbf{p}_s \mathbf{p}'_s) \mathbf{X}_s \right)^{-1} \\ &= \mathbf{M}^{-1}(\mathbf{X}, \boldsymbol{\beta})\end{aligned}$$

- minimize the trace: A-optimality criterion
- minimize the determinant: D-optimality criterion  
     equivalent to maximizing the determinant of the information matrix  $\mathbf{M}$

## Prediction-based design criteria

- seek to minimize variances of predicted probabilities
- minimize the maximum prediction variance: G-optimality criterion
- minimize the average prediction variance: V-optimality criterion
- talking about choice probabilities requires choice sets to be specified  
➡ we list all possible choice sets of size  $J$

e.g.:  $\binom{120}{2} = 7,140$  choice sets or 14,280 profiles



## Prediction-based design criteria

- mathematically:

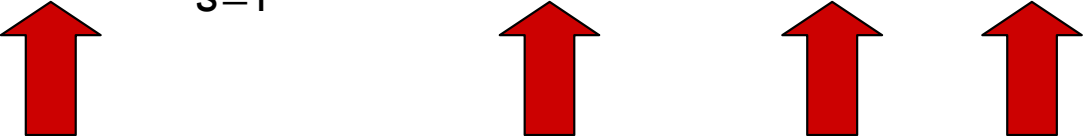
$$G = \max_{\mathbf{x}_{jq} \in \mathcal{X}} \mathbf{c}'(\mathbf{x}_{jq}) \mathbf{M}^{-1}(\mathbf{X}, \boldsymbol{\beta}) \mathbf{c}(\mathbf{x}_{jq})$$

$$V = \int_{\mathcal{X}} \mathbf{c}'(\mathbf{x}_{jq}) \mathbf{M}^{-1}(\mathbf{X}, \boldsymbol{\beta}) \mathbf{c}(\mathbf{x}_{jq}) d\mathbf{x}_{jq}$$

$$\text{with } \mathbf{c}(\mathbf{x}_{jq}) = \frac{\partial p_{jq}(\mathbf{x}_{jq}, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = p_{jq} \left( \mathbf{x}_{jq} - \sum_{t=1}^J p_{tq} \mathbf{x}_{tq} \right) \text{ and}$$

$\mathcal{X} = \left\{ \left\{ \mathbf{x}_{1q}, \dots, \mathbf{x}_{Jq} \right\} \mid q = 1, \dots, Q \right\}$ , all  $Q$  possible  
choice sets of size  $J$

# Dependency on the unknown model parameter $\beta$

$$\mathbf{M}(\mathbf{X}, \beta) = \sum_{s=1}^S \mathbf{X}'_s (\mathbf{P}_s(\beta) - \mathbf{p}_s(\beta) \mathbf{p}'_s(\beta)) \mathbf{X}_s$$


## ***Bayesian optimal designs:***

- construct a prior distribution for the parameters
  - find design that performs best on average
    - Sándor & Wedel (2001, 2002, 2005)

# Computational results

(Kessels, Goos and Vandebroek 2006)

- design problem involves
  - 2 attributes with 3 levels + 1 attribute with 2 levels
  - $3 \times 3 \times 2 = 18$  possible profiles
- design sizes
  - 12 choice sets of 2 profiles  
 $Q = \binom{18}{2} = 153$  choice sets or 306 profiles
  - 8 choice sets of 3 profiles  
 $Q = \binom{18}{3} = 816$  choice sets or 2,448 profiles

# Computational results

(Kessels, Goos and Vandebroek 2006)

- Monte Carlo sampling

$$\pi(\boldsymbol{\beta}) = N(\boldsymbol{\beta} \mid \boldsymbol{\beta}_0, \mathbf{I}_5) \text{ with } \boldsymbol{\beta}_0 = [-1, 0, -1, 0, -1]'$$

 1000 draws

- modified Fedorov algorithm

- exchange of profiles from the candidate set
- 200 tries or random starting designs

Monte Carlo modified Fedorov algorithm (MCMF)

# Optimal designs

(Kessels, Goos and Vandebroek 2006)

Choice sets with  
two alternatives

CS	Alt	D			A			G			V		
		Attributes			Attributes			Attributes			Attributes		
		1	2	3	1	2	3	1	2	3	1	2	3
1	I	2	3	1	3	2	1	3	1	2	2	2	2
	II	1	2	1	3	1	1	2	2	2	1	1	1
2	I	2	2	2	2	3	1	3	2	1	2	1	2
	II	1	1	1	1	2	1	2	3	1	1	2	1
3	I	1	2	2	2	1	2	1	2	1	1	2	2
	II	3	1	2	1	2	2	2	1	1	3	1	1
1	I	3	2	1	1	3	2	2	2	1	1	2	2
	II	2	1	1	2	3	1	3	1	2	3	1	1
	III	1	2	2	1	2	1	1	3	1	2	2	1
2	I	1	1	1	1	1	1	3	2	2	3	2	1
	II	2	2	1	2	1	1	1	1	2	1	3	1
	III	1	3	2	1	2	1	3	1	2	2	1	1

Choice sets with  
three alternatives

# Computing times

(Kessels, Goos and Vandebroek 2006)

- for 1 try of MCMF
- performed in the SAS procedure IML

Design criterion	# Alternatives	
	2	3
D	00h:05m	00h:05m
A	00h:05m	00h:05m
G	02h:30m	11h:00m
V	02h:30m	11h:00m

*Computation of Bayesian G- and V-optimal designs is practically infeasible using MCMF*

# Improved approach

(Kessels, Jones, Goos and Vandebroek 2006)

- Huge reduction in computing times and better designs
- As a result of using
  1. a small designed sample of prior parameters
  2. a coordinate-exchange algorithm
  3. updates of the Cholesky decomposition of the information matrix
  4. a handy formula for the computation of the Bayesian V-optimality criterion

Adaptive algorithm

# Improved results

(Kessels, Jones, Goos and Vandebroek 2006)

- Huge reduction in computing times
  - for 1 try of the **adaptive algorithm**
  - performed in MATLAB 7
  - ➡ comparison with computing times of **MCMF** in MATLAB 7

Design criterion	# Alternatives		
	2	3	4
D	00:00:03	00:00:04	00:00:05
A	00:00:03	00:00:04	00:00:05
G	00:00:07	00:00:32	00:04:23
V	00:00:03	00:00:05	00:00:08

Design criterion	# Alternatives		
	2	3	4
D	00:08:00	00:08:00	—
A	00:08:00	00:08:00	—
G	03:00:00	12:00:00	—
V	03:00:00	12:00:00	—

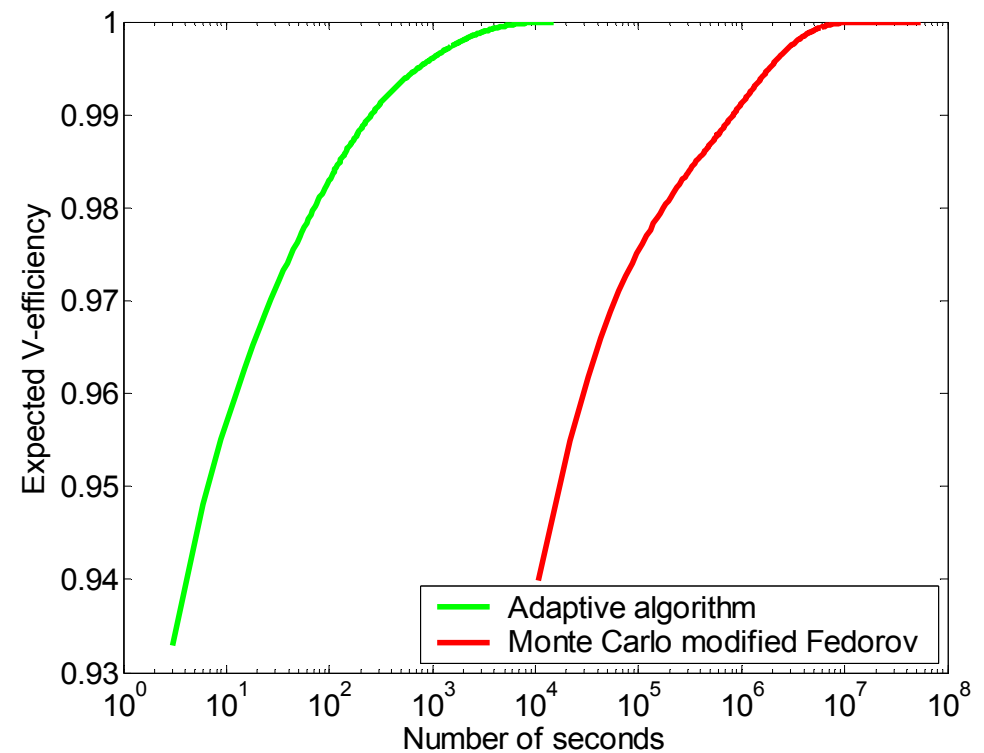
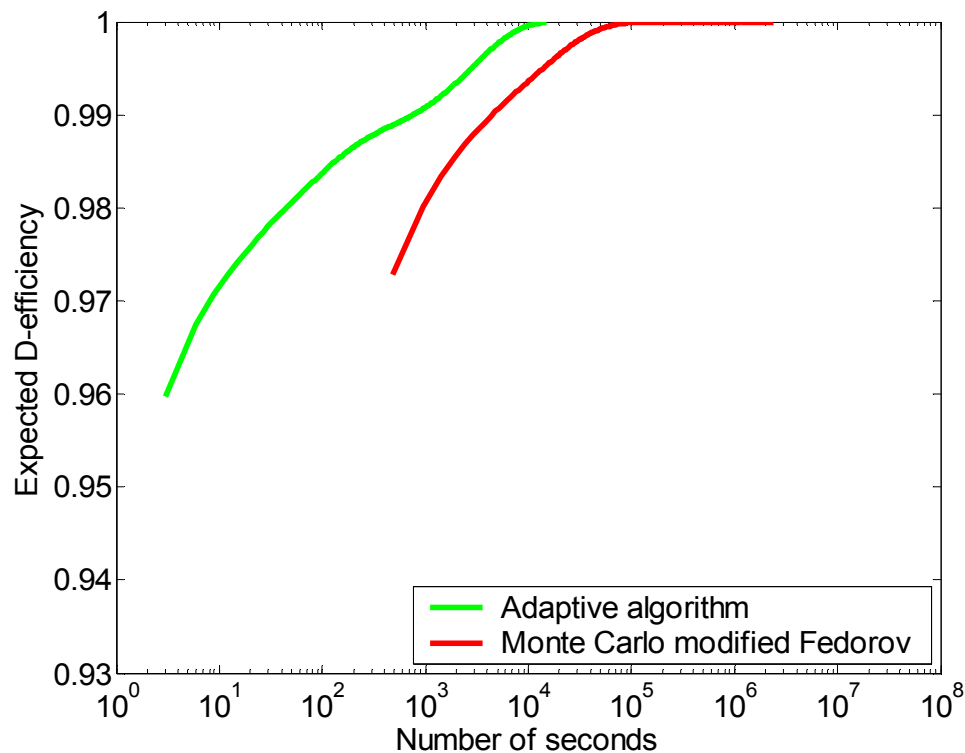


# Improved results

(Kessels, Jones, Goos and Vandebroek 2006)

## CONSEQUENCE:

The adaptive algorithm is computationally more effective in time than MCMF



# Improved results

(Kessels, Jones, Goos and Vandebroek 2006)

- Better or tied D-, A-, G- and V-optimal designs
  - from 1000 tries

Optimal design	2 alternatives		3 alternatives		4 alternatives	
	Adaptive	MCMF	Adaptive	MCMF	Adaptive	MCMF
D	0.73024	0.73024	0.75362	0.76617	0.86782	—
A	6.55212	6.60563	5.97903	6.02261	6.57135	—
G	0.49887	0.51997	0.51051	0.51843	0.60494	—
V	0.07184	0.07219	0.06267	0.06285	0.05728	—

# Adaptive algorithm

(Kessels, Jones, Goos and Vandebroek 2006)

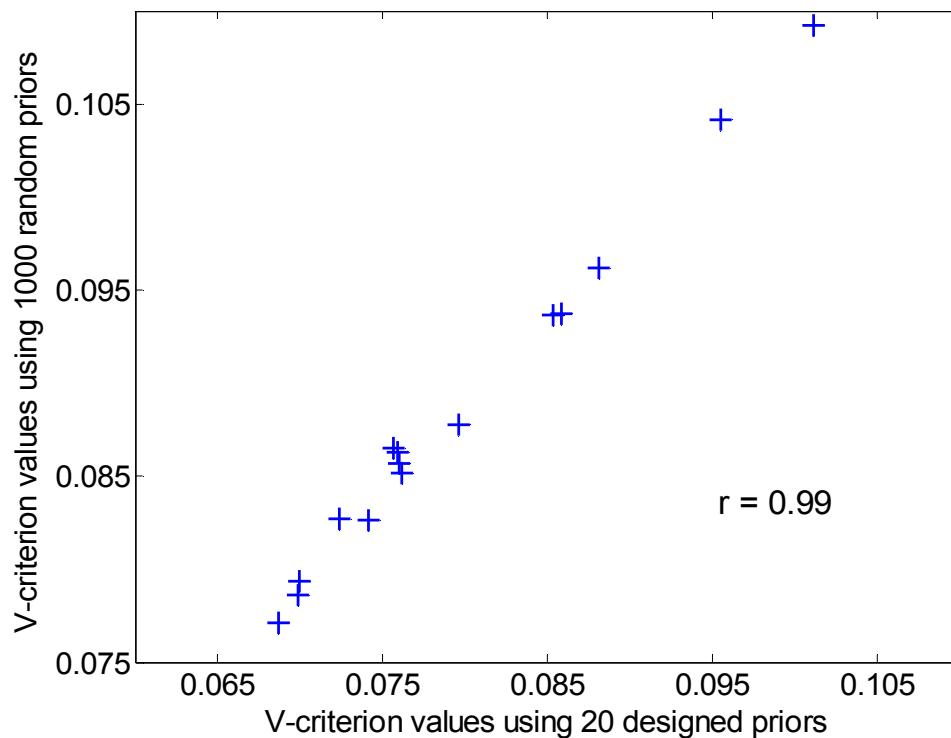
- Computation of the Bayesian designs produced by 1000 tries
  - based on 20 designed prior parameters instead of 1000 random ones
- Re-evaluation of the Bayesian designs from 1000 tries and selection of the optimal design
  - based on 1000 random prior parameters

Computational savings of up to **98%** within each try of the algorithm!

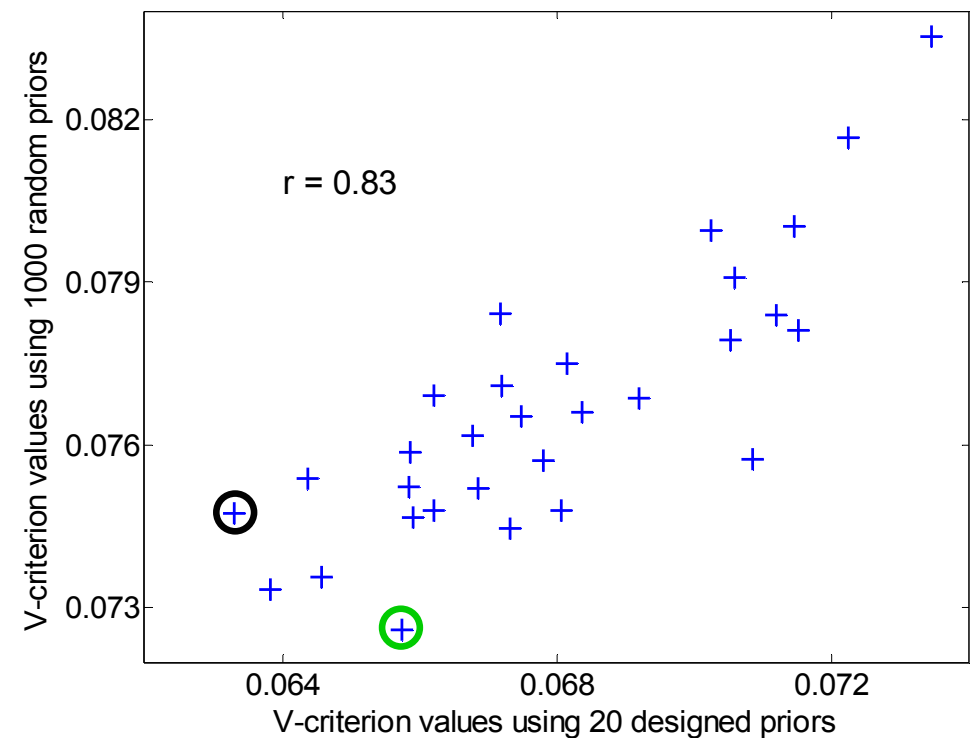
# Why an adaptive algorithm?

(Kessels, Jones, Goos and Vandebroek 2006)

## Improvements in 1 try



## Designs from 30 tries



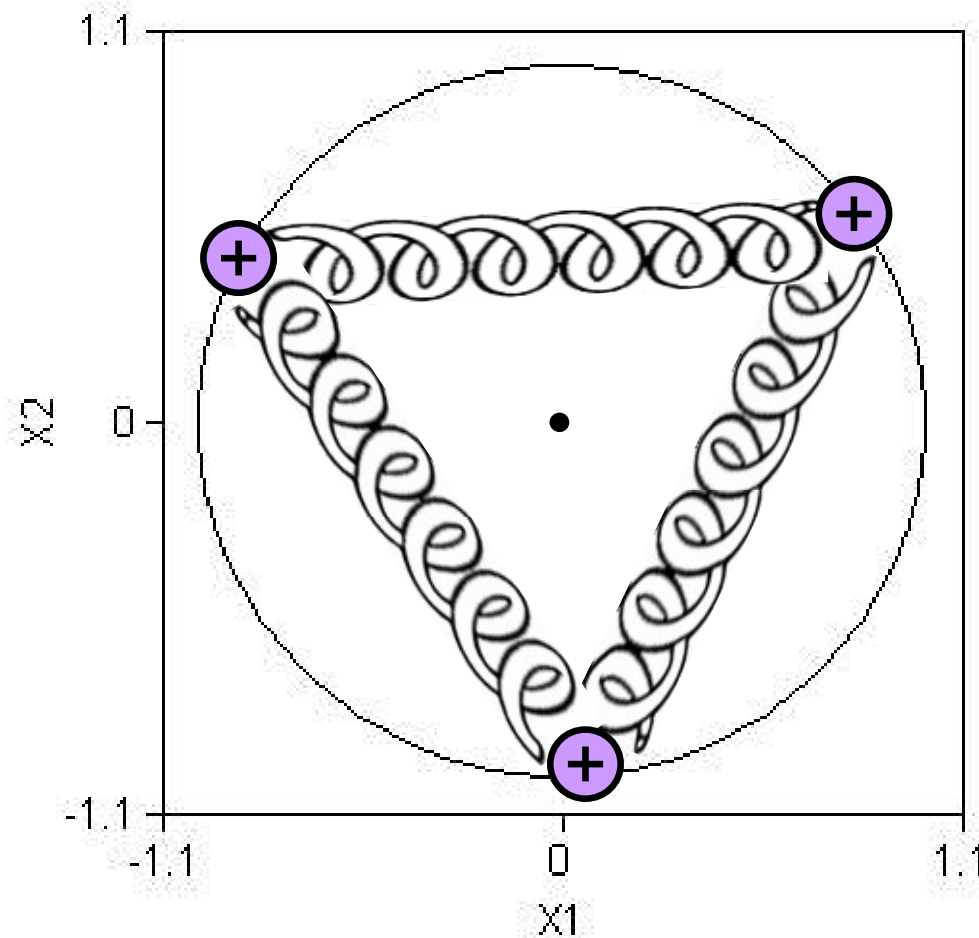
# Designed prior parameters

(Kessels, Jones, Goos and Vandebroek 2006)

- spherically symmetric and with (near) constant separation from each other around the prior mean at a distance of 2 units (radius =  $2\sigma$ )
- = minimum potential designs or space filling designs, created in JMP 6
  - let  $d_{ef}$  be the distance between points  $e$  and  $f$
  - minimize  $E_{pot} = \sum_{ef, e \neq f} (d_{ef}^2 + 1/d_{ef})$   
with  $d_{ef}^2$  the energy in a spring when you pull it and  $1/d_{ef}$  the energy between two like charged particles

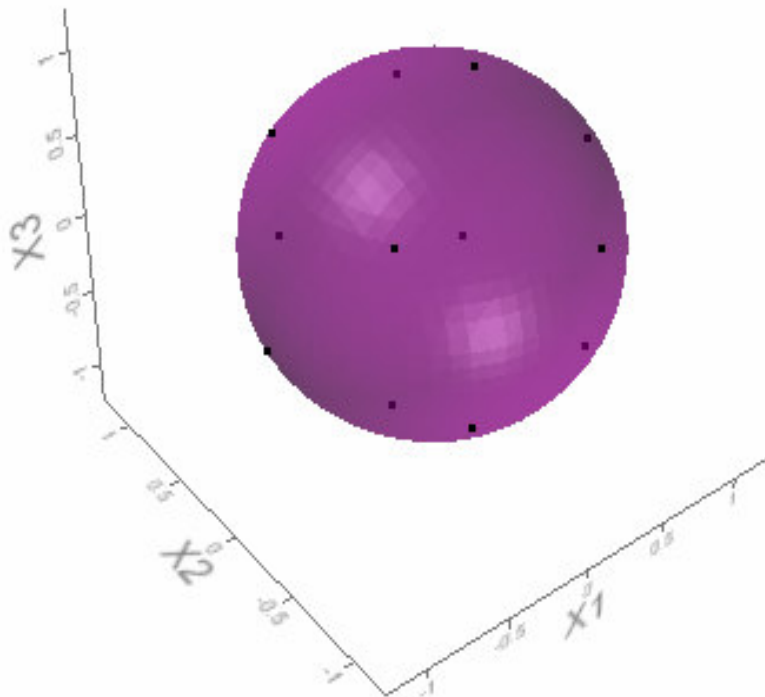
# Three equally spaced points on the circumference of a circle

(Kessels, Jones, Goos and Vandebroek 2006)



# 12 equally spaced points on the surface of a sphere

(Kessels, Jones, Goos and Vandebroek 2006)

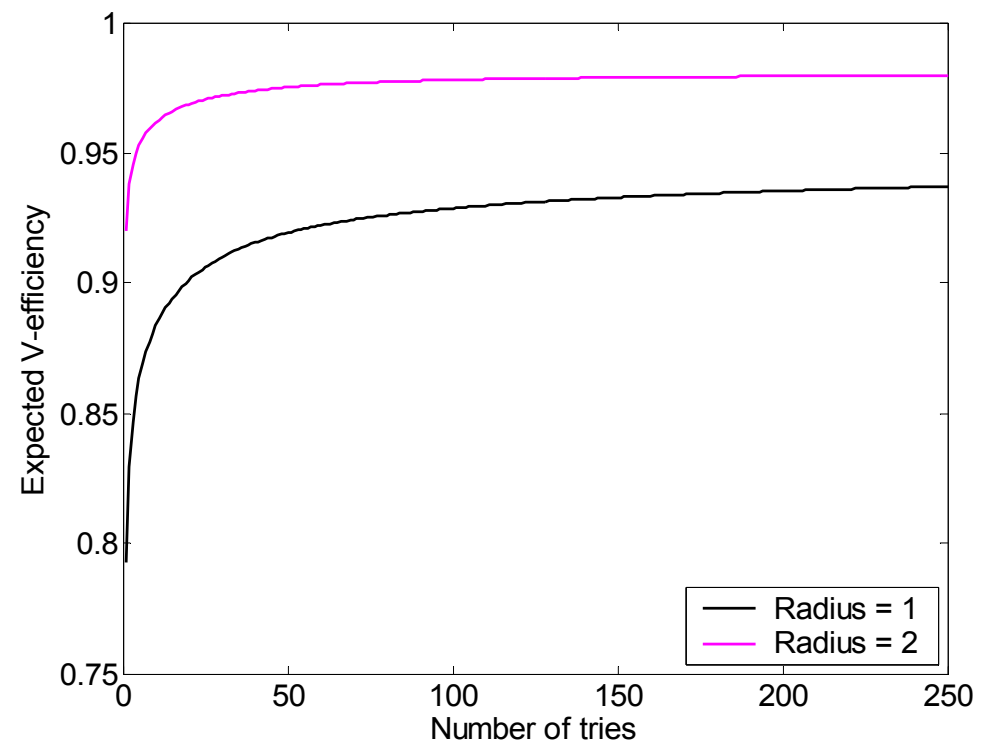
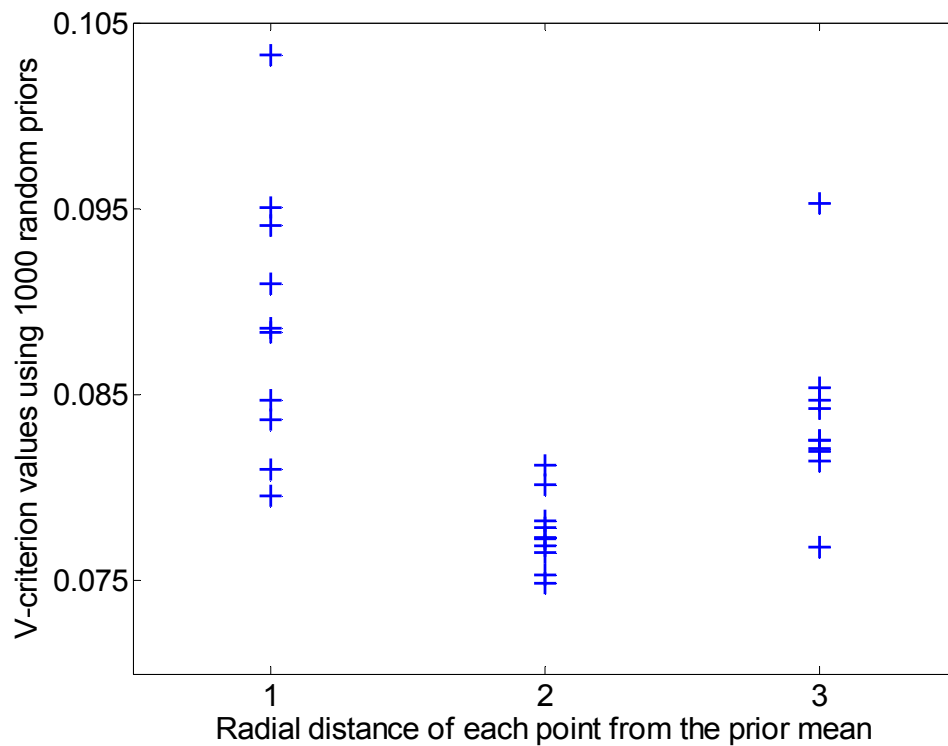


Run	X1	X2	X3	Minimum Distance	Nearest Point	Radius
1	0.231845	-0.57289	0.786146	0.52572	8	0.999984
2	-0.69986	-0.09182	0.708348	0.52572	11	0.999995
3	-0.23184	0.572899	-0.78614	0.52572	12	0.999995
4	0.557661	-0.80912	-0.18524	0.52572	5	0.999991
5	-0.46192	-0.88486	0.060249	0.52572	4	0.999996
6	-0.55766	0.809128	0.185242	0.52572	9	0.999989
7	0.949872	0.030723	0.311119	0.52572	4	0.999996
8	0.172680	0.474046	0.863401	0.52572	1	1
9	0.461922	0.884865	-0.06024	0.52572	6	0.999988
10	-0.17267	-0.47404	-0.86339	0.52572	12	0.999994
11	-0.94986	-0.03072	-0.31111	0.52572	2	0.999991
12	0.699855	0.091823	-0.70835	0.52572	3	0.999989

# Radius of $2\sigma$

(Kessels, Jones, Goos and Vandebroek 2006)

## Comparison to radii of $1\sigma$ and $3\sigma$





# Coordinate-exchange algorithm

(Kessels, Jones, Goos and Vandebroek 2006)

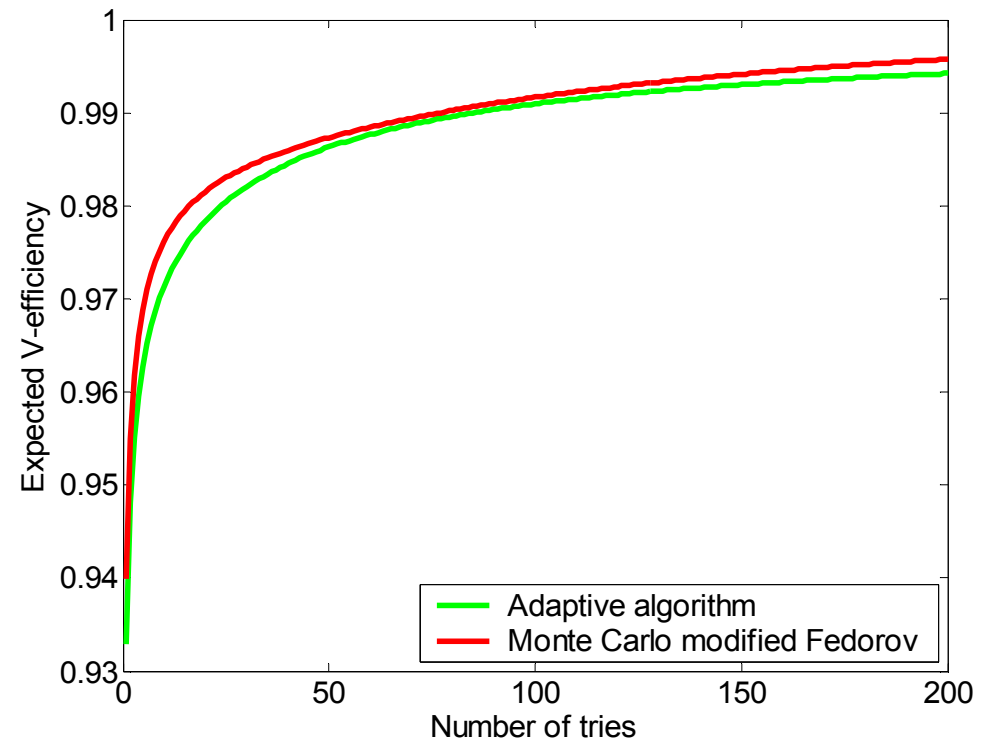
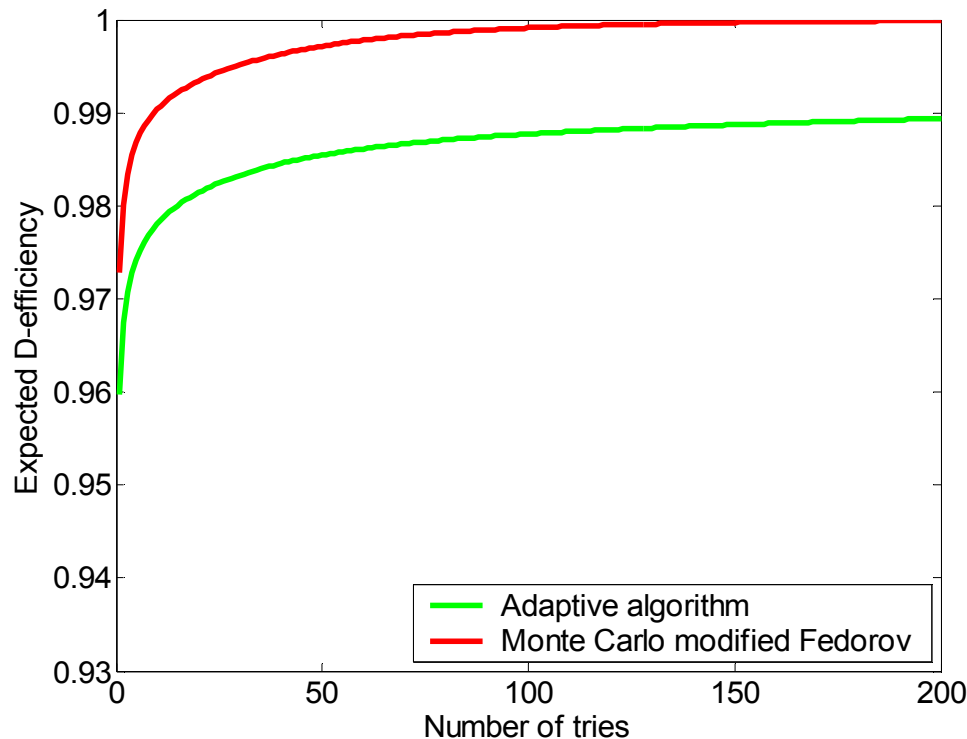
- greedy profile exchange algorithm: instead of possibly changing every coordinate/attribute level of a profile, one only changes one
- candidate-set-free algorithm
- very fast: reductions in computing time increase with the dimensions of the design problem (attributes, attribute levels, design profiles)
- less effective per try than MF (*but not in time!*): it takes more tries to find the global optimum
- originally developed by Meyer and Nachtsheim (1995)

# Coordinate-exchange algorithm

(Kessels, Jones, Goos and Vandebroek 2006)

## CONSEQUENCE:

The adaptive algorithm is computationally less effective per number of tries than MCMF



# Updating the Cholesky decomposition of the information matrix $\mathbf{M}$

(Kessels, Jones, Goos and Vandebroek 2006)

- feasible because  $\mathbf{M}$  is symmetric as a result of symmetric information matrices  $\mathbf{M}_s = \mathbf{X}'_s (\mathbf{P}_s - \mathbf{p}_s \mathbf{p}'_s) \mathbf{X}_s = \mathbf{X}'_s \mathbf{C}_s \mathbf{X}_s$ , with  $\mathbf{C}_s$  symmetric
- by definition:  $\mathbf{M} = \mathbf{L}'\mathbf{L}$  with  $\mathbf{L}$  the upper triangular matrix named the Cholesky factor
- $\mathbf{M}^{-1} = (\mathbf{L}'\mathbf{L})^{-1} \Rightarrow \mathbf{M}^{-1} = \mathbf{L}_{inv} \mathbf{L}'_{inv}$
- updating  $\mathbf{L}$ : changing a profile in  $\mathbf{X}$ 
  - deletion of the current profile:  $\mathbf{L1}$
  - insertion of a new profile:  $\mathbf{L2}$

# D- and A-optimality criteria in terms of the Cholesky factor

(Kessels, Jones, Goos and Vandebroek 2006)

$$\begin{aligned} D &= \left( \det(\mathbf{M}^{-1}) \right)^{1/k} \\ &= 1 / \left( \det(\mathbf{M}) \right)^{1/k} \\ &= 1 / \left( \det(\mathbf{L}') \det(\mathbf{L}) \right)^{1/k} \\ &= 1 / \left( \text{prod}(\text{diag}(\mathbf{L})) \right)^{2/k} \end{aligned}$$

$$\begin{aligned} A &= \text{tr}(\mathbf{M}^{-1}) \\ &= \text{tr}(\mathbf{L}_{inv} \mathbf{L}'_{inv}) \\ &= \text{SS}(\text{elements in } \mathbf{L}_{inv}) \end{aligned}$$

# V- and G-optimality criteria in terms of the Cholesky factor

(Kessels, Jones, Goos and Vandebroek 2006)

## ■ V: computational short cut

→ prediction variance  $\mathbf{c}'(\mathbf{x}_{jq})\mathbf{M}^{-1}\mathbf{c}(\mathbf{x}_{jq}) = \text{scalar}$

→  $\mathbf{c}'(\mathbf{x}_{jq})\mathbf{M}^{-1}\mathbf{c}(\mathbf{x}_{jq}) = \text{tr}\left(\mathbf{c}'(\mathbf{x}_{jq})\mathbf{M}^{-1}\mathbf{c}(\mathbf{x}_{jq})\right)$

→  $\text{tr}\left(\mathbf{c}'(\mathbf{x}_{jq})\mathbf{M}^{-1}\mathbf{c}(\mathbf{x}_{jq})\right) = \text{tr}\left(\mathbf{c}(\mathbf{x}_{jq})\mathbf{c}'(\mathbf{x}_{jq})\mathbf{M}^{-1}\right)$

→ let  $\mathbf{W}_{jq} = \mathbf{c}(\mathbf{x}_{jq})\mathbf{c}'(\mathbf{x}_{jq}) \Rightarrow \mathbf{W} = \frac{1}{JQ} \sum_{j=1}^J \sum_{q=1}^Q \mathbf{W}_{jq}$

→  $V = \int_{\mathcal{X}} \mathbf{c}'(\mathbf{x}_{jq})\mathbf{M}^{-1}\mathbf{c}(\mathbf{x}_{jq}) d\mathbf{x}_{jq} = \text{tr}\left(\mathbf{W}\mathbf{L}_{inv}\mathbf{L}'_{inv}\right)$


## ■ G: trick does not work since all the individual variances have to be computed to find the worst variance

# Large choice designs

(Kessels, Jones, Goos and Vandebroek 2006)

- Easy to compute using the adaptive algorithm
- **Race bicycle example:** 15 choice sets of size 2  
5 attributes:  $2^3 \times 3 \times 5$   
extension: 10 choice sets of size 3

Computing times per try



Design criterion	# Alternatives	
	2	3
D	00:00:08	00:00:14
V	00:00:15	00:04:05