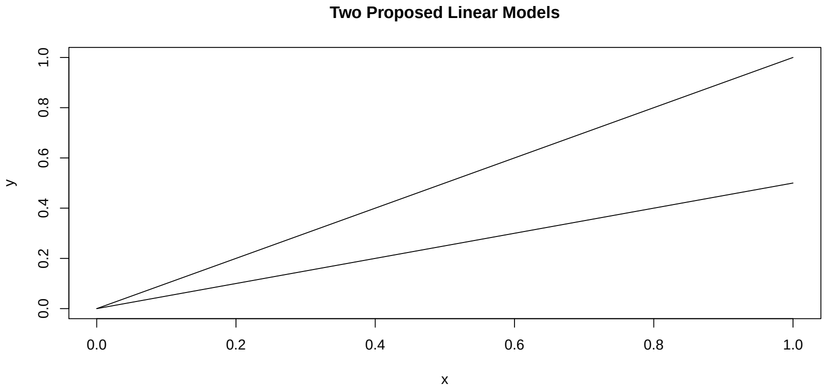
MED for Model Selection

Kristyn Pantoja

6/25/2019

Simple Linear Regression: Unknown Slope

Design an Experiment that Estimates Slope



- Goal: Choose design $\mathbf{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ to gather data that will
  1. help distinguish these two slopes
  2. allow adequate estimation of β.
- Idea: Minimum Energy Design!

Minimum Energy Design

Minimum energy design (MED) is a deterministic sampling method which makes use of evaluations of the target distribution f to obtain a weighted space-filling design.

Definition:

Design $\mathbf{D} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ is a MED if it minimizes the total potential energy, given by:

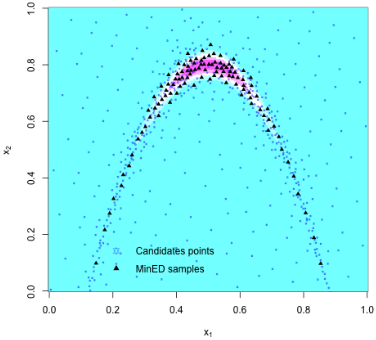
$$\sum_{i \neq j} \frac{q(\mathbf{x}_i)q(\mathbf{x}_j)}{d(\mathbf{x}_i, \mathbf{x}_j)}$$

Choose the charge function, $q = \frac{1}{f^{1/2p}}$ so that the limiting distribution of the design points is target distribution, f.

Advantages of MED

Sampling the "Banana" Function

- $N = 109$
- $K = 6$
- $NK = 654$ evaluations of f



Compared to other sampling methods, MED

- has fewer points and hence (unlike MCMC)
- requires fewer evaluations of f (unlike MCMC)
- is not prone to missing high-density regions (unlike QMC)

Simple Linear Regression without Intercept

- Assume $y_i = x_i\beta + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$ and $\beta \sim N(\mu, \nu^2)$.
- $y|\beta, X \sim N(X\beta, \sigma^2 I)$
- $y|X \sim N(X\mu, \sigma^2 I + \nu^2 XX^T)$ after marginalizing out β

Hypotheses

Suppose we suspect $\beta = \mu_0$ or $\beta = \mu_1$, i.e.

$$H_0 : \beta \sim N(\mu_0, \nu_0^2)$$
$$H_1 : \beta \sim N(\mu_1, \nu_1^2)$$

MED design may distinguish these two hypotheses and allow for adequate estimation of β.

Evaluating the Designs

Evaluating Methods

- Posterior Variance, i.e. $Var[\beta|y, X]$
- Expected Posterior Probabilities of Hypotheses & Bayes Factor
- Design Criteria:
  - Total Potential Energy
  - Criterion for One-at-a-Time Algorithm
  - Criterion for Fast Algorithm

Interpretations

- A design that is better for estimating β may have smaller posterior variance.
- A design that is better for hypothesis testing may give a larger expected posterior probability to the true model from simulated responses.

Posterior Variance

In the Bayesian linear regression framework,

$$y|\beta, X \sim N(X\beta + \sigma^2 I)$$
$$\beta \sim N(\mu, V)$$

with $X \in \mathbb{R}^{N \times p}, \beta \in \mathbb{R}^p, V \in \mathbb{R}^{p \times p}$,

- $\hat{\beta} = \frac{1}{\sigma^2}\Sigma_B(X^T y + \sigma^2 V^{-1}\mu)$ with posterior distribution

  $$\beta|y, X \sim N\left(m_B, \Sigma_B\right)$$

  where

  $$\Sigma_B = \sigma^2(X^T X + \sigma^2 V^{-1} I)^{-1}$$
  $$m_B = \frac{1}{\sigma^2}\Sigma_B(X^T y + \sigma^2 V^{-1}\mu)$$

- Σ_B can be used to evaluate a design **D**'s ability to estimate β

Posterior Probabilities of Hypotheses

- Posterior Probability of model $H_\ell, \ell \in 1, ..., M$:

$$P(H_\ell|y, X) = \frac{\pi_\ell L(y|H_\ell, X)}{\sum_{m=1}^{M} \pi_m L(y|H_m, X)}$$

  where π_m is the prior on H_m (typically $\pi_m = \frac{1}{M}$), and $L(y|H_m, X)$ is the model evidence.
- $P(H_\ell|y, X)$ tells which hypothesis is more likely to give the correct model.
- $E[P(H_\ell|y, X)|H_r]$ may be estimated using MC approximation from simulated responses y under a chosen hypothesis H_r.
- $E[P(H_\ell|y, \mathbf{D})|H_r]$ can be used to evaluate a design $\mathbf{D}$'s ability to distinguish hypotheses

Estimate Expected Posterior Probability of a Hypothesis

Estimate the expected posterior probability of hypothesis H_ℓ for J simulations of Y under H_r, given design $\mathbf{D} = \{x_1, ..., x_N\}$:

1. For $j = 1, \ldots, J$:
   1.1 Draw $\beta \sim N(\mu_r, \nu_r^2)$
   1.2 Draw $y_i^{(j)}|\mathbf{D} \sim N(\mathbf{x}_i\beta, \sigma_r^2)$, $\forall \mathbf{x}_i \in \mathbf{D}$
   1.3 $\forall m \in \{1, ..., M\}$, calculate model evidences $L(y^{(j)}|H_m, \mathbf{D})$
      - model evidence $L(y|H_m, \mathbf{D})$ is the marginal likelihood $N(\mathbf{D}\mu_m, \sigma^2 I + \nu^2 \mathbf{D}\mathbf{D}^T)$ evaluated at y and $\mathbf{D}$.
   1.4 Calculate the posterior probability of H_ℓ, $P(H_\ell|y^{(j)}, \mathbf{D})$, from simulation j

$$P(H_\ell|y^{(j)}, \mathbf{D}) = \frac{\pi_\ell P(y^{(j)}|H_\ell, \mathbf{D})}{\sum_{m=1}^{M} \pi_m P(y^{(j)}|H_m, \mathbf{D})}$$

2. Average the estimated posterior probabilities of H_ℓ over $\forall j$ to obtain MC estimate of $E[P(H_\ell|y, \mathbf{D})|H_r]$

MED Criteria

1. The Total Potential Energy, which both algorithms aim to minimize:
$$\sum_{i \neq j} \frac{q(\mathbf{x}_i)q(\mathbf{x}_j)}{d(\mathbf{x}_i, \mathbf{x}_j)}$$

2. One-at-a-Time Algorithm: minimize
$$\left\{ \sum_{i \neq j} \left(\frac{q(\mathbf{x}_i)q(\mathbf{x}_j)}{d(\mathbf{x}_i, \mathbf{x}_j)} \right)^k \right\}^{1/k}$$
which gives the Total Potential Energy criterion when $k = 1$.

3. Fast Algorithm: minimize
$$\max_{i \neq j} \frac{q(\mathbf{x}_i)q(\mathbf{x}_j)}{d(\mathbf{x}_i, \mathbf{x}_j)}$$

MED-generating Algorithms

One-at-a-Time Algorithm (2015)

Steps to obtain MED using One-at-a-Time algorithm:

1. Obtain *numCandidates* candidate points, $\mathbf{x}$, in $[0, 1]$.
2. Initialize D_N by choosing $\mathbf{x}_j$ to be the candidate $\mathbf{x}$ which optimizes f, where $f(\mathbf{x}) = \text{Wasserstein}(\phi_{0,\mathbf{x}}, \phi_{1,\mathbf{x}})$ and

$$\phi_{0,\mathbf{x}} = N(\mu_0 \mathbf{x}, \sigma_0^2 + \mathbf{x}^2 \nu_0^2),$$
$$\phi_{1,\mathbf{x}} = N(\mu_1 \mathbf{x}, \sigma_1^2 + \mathbf{x}^2 \nu_1^2)$$

3. Choose the next point $\mathbf{x}_{j+1}$ by:

$$\mathbf{x}_{j+1} = \arg\min_{\mathbf{x}} \sum_{i=1}^{j} \left(\frac{q(\mathbf{x}_i) q(\mathbf{x})}{d(\mathbf{x}_i, \mathbf{x})} \right)^k$$

where $q = 1/f^{(1/2p)}$, $d(x, y)$ is Euclidean distance and $k = 4p$.

- This is a greedy algorithm for choosing points one at a time

Fast Algorithm (2018)

In each of S stages, create a new design to iteratively minimize

$$\max_{i \neq j} \frac{q(\mathbf{x}_i)q(\mathbf{x}_j)}{d(\mathbf{x}_i, \mathbf{x}_j)}$$

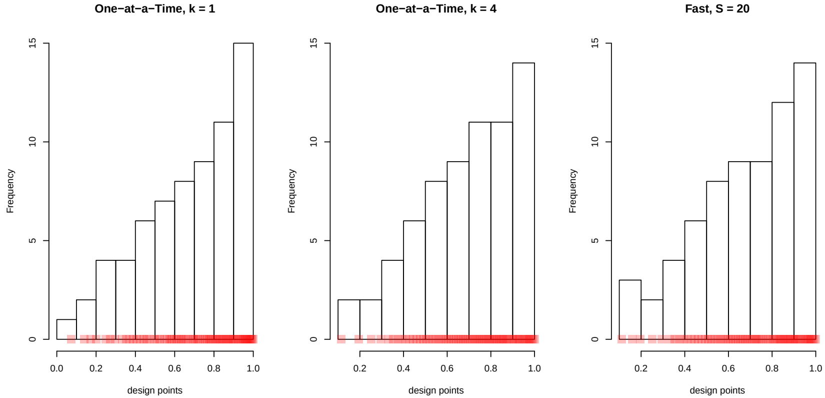
1. Initialize space-filling design $\mathbf{D}_1 = \{\mathbf{x}_1^{(1)} \ldots \mathbf{x}_N^{(1)}\}$
2. For $s = 1, \ldots, S-1$ steps, obtain each design point $\mathbf{x}_j^{(s+1)} \in \mathbf{D}_{s+1}$ by:

$$\begin{aligned}
\mathbf{x}_j^{s+1} &= \arg\min_{\mathbf{x} \in \mathbf{C}_j^{s+1}} \max_{i=1:(j-1)} \frac{1}{f^{\gamma_s}(\mathbf{x}_i) f^{\gamma_s}(\mathbf{x}) d^{(2p)}(\mathbf{x}_i, \mathbf{x})} \\
&= \arg\min_{\mathbf{x} \in \mathbf{C}_j^{s+1}} \max_{i=1:(j-1)} \frac{q^{\gamma_s}(\mathbf{x}_i) q^{\gamma_s}(\mathbf{x})}{d(\mathbf{x}_i, \mathbf{x})}
\end{aligned}$$

where $\gamma_s = s/(S-1)$ and $\mathbf{C}_j^{s+1}$ is the candidate set for $\mathbf{x}_j^{(s+1)}$

- Points migrate to more optimal locations in each stage

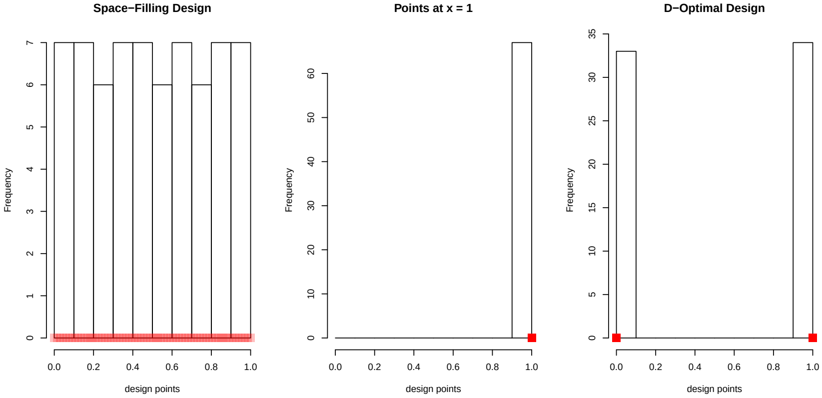
Designs from MED-Generating Algorithms



Other Designs

Other Designs

- Random designs: 50 random designs ($\mathbf{x} \sim U([0,1]^p)$, $\forall \mathbf{x} \in \mathbf{D}$).
- Space-Filling Design: evenly spaced points over $[0,1]$
- $\mathbf{D} = \mathbf{1}$: $\forall \mathbf{x} \in \mathbf{D}, \mathbf{x} = 1$.
- D-optimal Design: seeks to minimize the variance of the estimated regression coefficients.
  - generated by `AlgDesign` (using Federov's exchange algorithm).



Space−Filling Design
Points at x = 1
D−Optimal Design
Frequency
design points

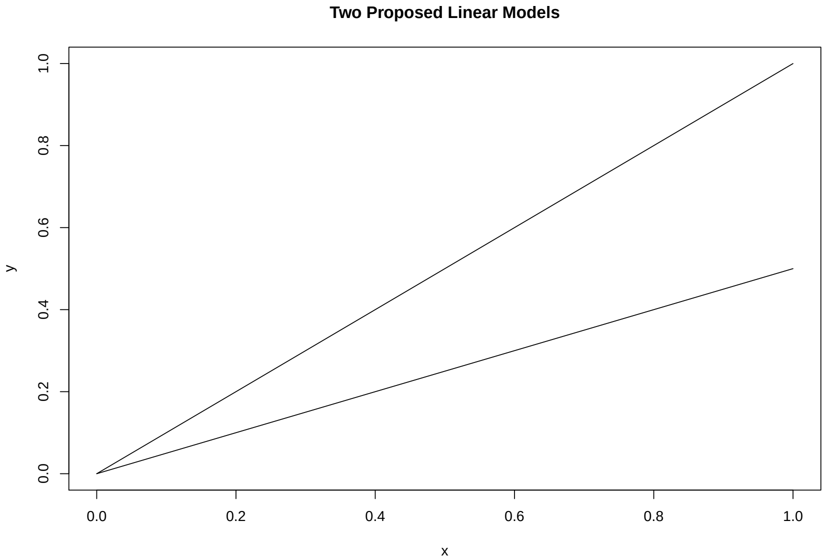
Results

Results!

	1atT,k=1	1atT,k=4	Fast	Random	Space	D = 1	D-opt
E[P(H0|Y,D)|H0]	0.999	0.999	0.999	0.999	0.999	0.999	0.999
E[BF01 | H0]	3.67e+15	3.55e+16	5.21e+16	6.53e+15	5.3e+14	6.63e+16	6.3e+15
E[P(H1|Y,D)|H1]	0.999	0.999	0.999	0.999	0.999	0.999	0.999
E[BF01|H1]	0.0108	0.00142	0.000913	0.0522	0.00318	0.00104	0.000738
PostVar b x10e-4	6.33	6.25	6.28	9.15	9.09	3.47	6.41
TPE x10e3	2810	2870	2820	8050000	Inf	Inf	Inf
Fast x10e3	97.5	43.7	44.5	7810000	Inf	Inf	Inf
1atT(k=4) x10e3	120	92.5	94.1	7810000	Inf	Inf	Inf
Mean(D)	0.674	0.689	0.684	NA	0.5	1	0.507
sd(D)	0.247	0.219	0.23	NA	0.295	0	0.504

Simple Linear Regression: Unknown Slope and Intercept

Design an Experiment that Estimates Slope and Intercept



SetUp

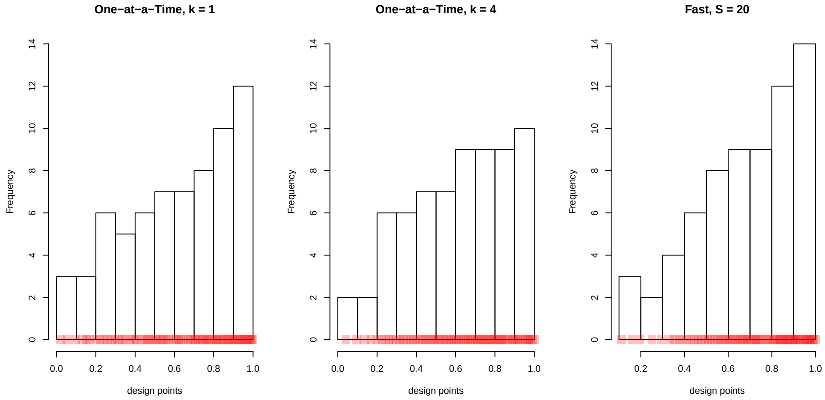
Similar to the unknown slope case,

- Assume $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ and $\beta \sim N(\mu, V), \mu = (\mu_0, \mu_1)^T, V = \text{diag}(\nu_0^2, \nu_1^2)$.
- $y|\beta, X \sim N(X\beta, \sigma^2 I)$
- $y|X \sim N(X\mu, \sigma^2 I + XVX^T)$

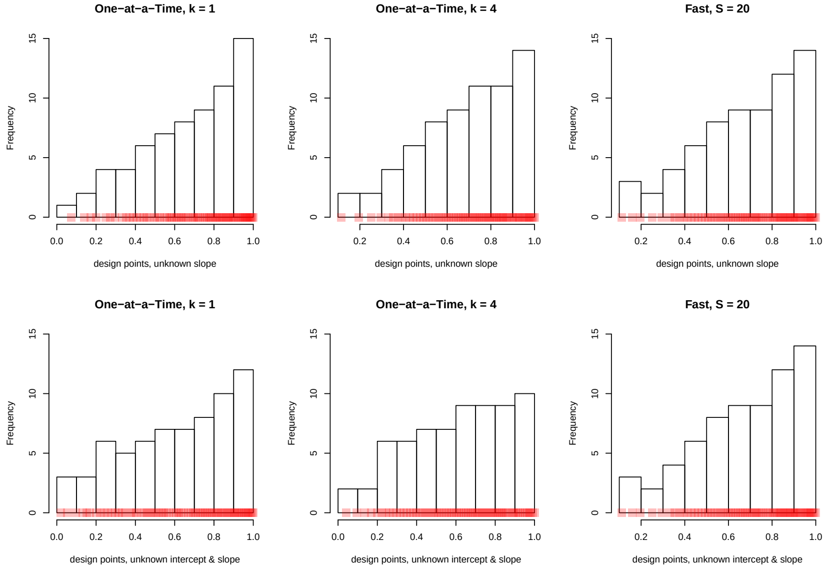
Hypotheses

Suppose we suspect $\beta = \mu_0$ or $\beta = \mu_1$, i.e.

$$
\begin{aligned}
H_0 : \beta &\sim N(\mu_0, V_0), \\
\mu_0 &= (\mu_{00}, \mu_{01})^T, \\
V_0 &= \text{diag}(\nu_{00}^2, \nu_{01}^2) \\
H_1 : \beta &\sim N(\mu_1, V_1), \\
\mu_1 &= (\mu_{10}, \mu_{11})^T, \\
V_1 &= \text{diag}(\nu_{10}^2, \nu_{11}^2)
\end{aligned}
$$


One−at−a−Time, k = 1
One−at−a−Time, k = 4
Fast, S = 20
Frequency
design points
0 2 4 6 8 10 12 14
0.0 0.2 0.4 0.6 0.8 1.0

Compare Unknown Slope to Unknown Intercept & Slope

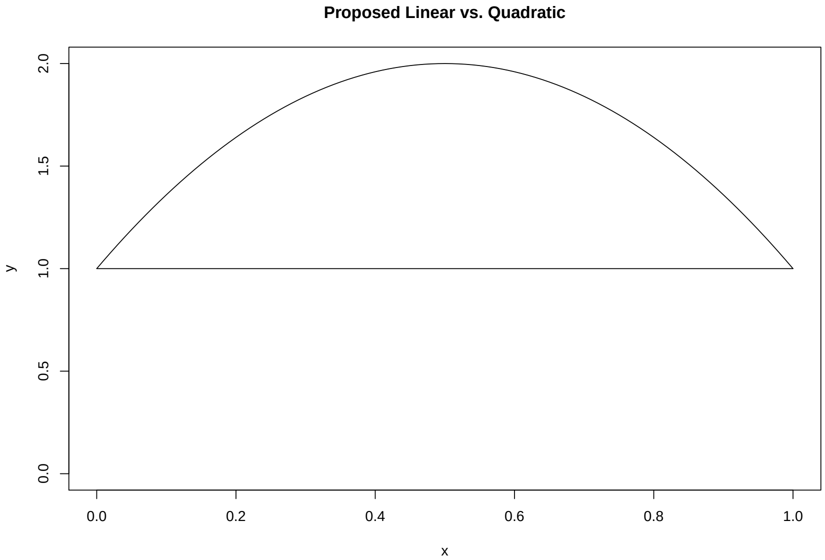


Table

	1atT,k=1	1atT,k=4	Fast	Space	D = 1	D-opt
E[P(H0|Y,D)|H0]	0.997	0.995	0.994	0.994	0.991	0.998
E[BF01 | H0]	3.99e+10	3.66e+09	1.87e+10	4.32e+10	5.72e+09	1.84e+14
E[P(H1|Y,D)|H1]	0.992	0.993	0.994	0.994	0.986	0.997
E[BF01|H1]	0.0925	0.0427	0.0919	1.03	0.144	0.0417
PostVar b0 x10e-4	10.3	10.8	13	8.01	25.9	5.9
PostVar b1 x10e-4	21.2	22.6	23.5	21	25.9	10.9
TPE x10e3	2270	2190	2820	Inf	Inf	Inf
Fast x10e3	56.2	24.1	44.5	Inf	Inf	Inf
1atT(k=4) x10e3	80.1	62.1	94.1	Inf	Inf	Inf
Mean(D)	6.66e+09	0.611	0.684	0.5	1	0.507
sd(D)	1.63e+10	0.255	0.23	0.295	0	0.504

Linear vs Quadratic

Linear Model vs. Quadratic Model



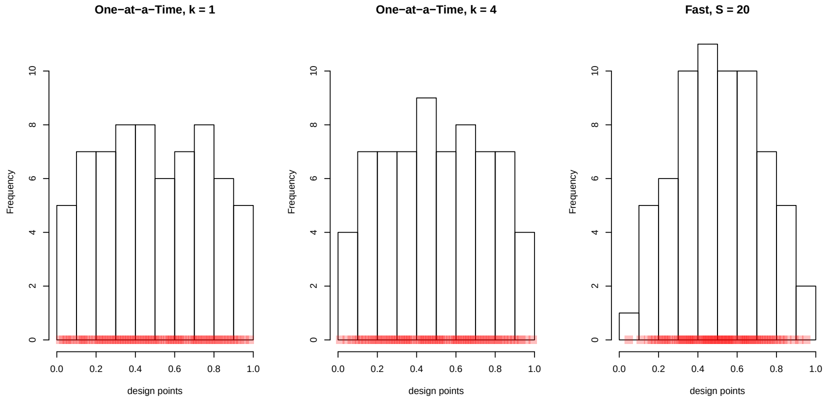
SetUp

We compare the linear model $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$ with the quadratic model $y_i = \beta_0 + x_i\beta_1 + +x_i^2\beta_2 + \varepsilon_i$

- $y|\beta, X \sim N(X\beta, \sigma^2 I)$
- $y|X \sim N(X\mu, \sigma_m^2 I + XVX^T)$

Hypotheses

$$
\begin{aligned}
H_0 : \beta &\sim N\left(\mu_0, V_0\right), \\
\mu_0 &= \left(\mu_{00}, \mu_{01}\right)^T, \\
V_0 &= \mathrm{diag}(\nu_{00}^2, \nu_{01}^2) \\
H_1 : \beta &\sim N\left(\mu_1, V_1\right), \\
\mu_0 &= \left(\mu_{10}, \mu_{11}, \mu_{12}\right)^T, \\
V_1 &= \mathrm{diag}(\nu_{10}^2, \nu_{11}^2, \nu_{12}^2)
\end{aligned}
$$



One−at−a−Time, k = 1
One−at−a−Time, k = 4
Fast, S = 20
Frequency
design points
0
2
4
6
8
10
0.0
0.2
0.4
0.6
0.8
1.0

Table

	1atT,k=1	1atT,k=4	Fast	Space	D = 0.5	D = 1	D-opt
E[P(H0|Y,D)|H0]	1	1	1	1	1	0.511	0.52
E[BF01 | H0]	5.66e+77	3e+75	5.69e+68	7.33e+88	1.83e+47	1.06	1.12
E[P(H1|Y,D)|H1]	1	1	1	1	1	0.51	0.516
E[BF01|H1]	1.18e-38	1.73e-40	1.09e-30	2.17e-50	1e-19	0.997	1.01
PostVar b0 x10e-4	8.39	8.67	9.77	8.12	14	33.7	6.2
PostVar b1 x10e-4	30.9	31.2	32.9	30.3	41	33.7	28.1
PostVar b2 x10e-4	32.2	32.4	34.7	30.9	47.7	33.7	28.1
TPE x10e3	872	786	973	Inf	Inf	Inf	Inf
Fast x10e3	40.2	11.7	12.9	Inf	Inf	Inf	Inf
1atT(k=4) x10e3	43.6	22.3	30	Inf	Inf	Inf	Inf
Mean(D)	0.494	0.501	0.51	0.5	0.5	1	0.507
sd(D)	0.272	0.263	0.217	0.295	0	0	0.504