MED for Model Selection

Kristyn Pantoja

6/4/2019

Simple Linear Regression: Unknown Slope

MED-generating Algorithms

Other Designs

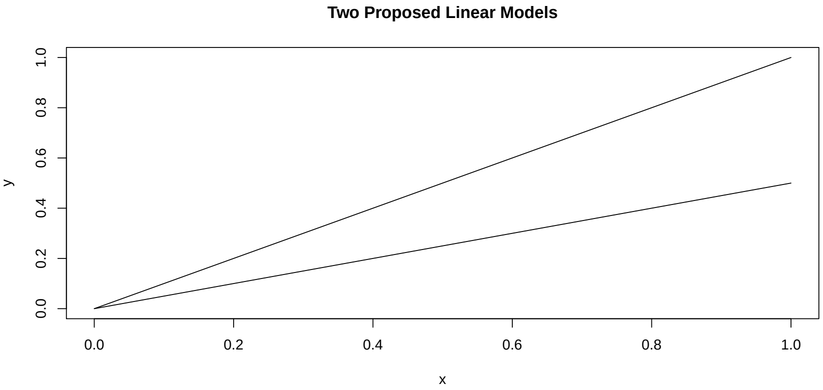
The Table

Simple Linear Regression: Unknown Slope and Intercept

Linear vs Quadratic

Simple Linear Regression: Unknown Slope

Design an Experiment that Estimates Slope



- Want to 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 minimum energy design 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.

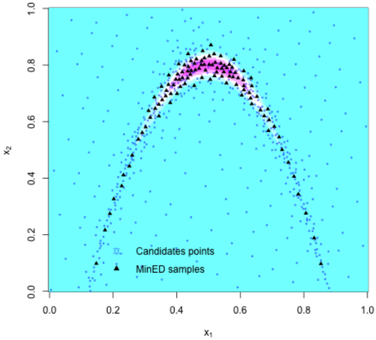
Objective:

$$\max_{i \neq j} \frac{1}{f^{1/2p}(\mathbf{x}_i) f^{1/2p}(\mathbf{x}_j) d(\mathbf{x}_i, \mathbf{x}_j)}$$

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 $\beta \sim N(\mu, \nu^2)$.
- $Y|\beta \sim N(X\beta, \sigma^2 I)$
- $Y \sim N(X\mu, \sigma_m^2 I + \nu^2 XX^T)$ 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 will help 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
- 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 might have smaller posterior variance.
- A design that is better for hypothesis testing will give a larger expected posterior probability to the true model from simulated responses.

Posterior Variance

In the linear regression model $Y \sim N(X\beta + \sigma^2 I)$ with $\beta \sim N(\mu, V)$ where $X \in \mathbb{R}^{N \times p}, \beta \in \mathbb{R}^p, V \in \mathbb{R}^{p \times p}$,

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

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

  where

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

The posterior variance Σ_B does not depend on the response y.

Posterior Probabilities of Hypotheses

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

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

  where π_m is the prior on H_m (typically $\pi_m = \frac{1}{M}$), and $P(Y|H_m)$ is the model evidence.
- The posterior probability of hypotheses tells which hypothesis is more likely to give the correct model.
- The expected posterior probability of the hypotheses $E[P(H_\ell|Y)|H_r]$ may be estimated using MC approximation from simulated responses $Y = \{y_1, \ldots, y_N\}$ under a chosen hypothesis H_r.

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 $y_i^{(j)} \sim N(\mathbf{x}_i \beta, \sigma_r^2)$, $\forall \mathbf{x}_i \in \mathbf{D}$
    1.3 $\forall m \in \{1, ..., M\}$, calculate model evidences $P(Y^{(j)} | H_m, \mathbf{D})$
    - model evidence $P(Y | H_m, \mathbf{D})$ is the marginal likelihood $N(\mathbf{D}\mu_m, \sigma_m^2 I + \nu^2 \mathbf{D}\mathbf{D}^T)$ evaluated at Y and $\mathbf{D}$.

    1.4 Calculate the posterior probability of H_ℓ, $P(H_\ell | Y^{(j)})$, from simulation j

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

2. Average the estimated posterior probabilities of H_ℓ over $\forall j$ to obtain a Monte Carlo estimate of the expected posterior probability of H_ℓ, $E[P(H_\ell | Y^{(j)}) | H_r]$

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 criterion tries to 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 becomes the Total Potential Energy Criterion when $k = 1$.

3. Fast Algorithm tries to 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

$$\phi_{0,\mathbf{x}} = N(\tilde{\beta}_0\mathbf{x}, \sigma^2_{\epsilon_0} + \mathbf{x}^2\sigma^2_{\beta_0}),$$
$$\phi_{1,\mathbf{x}} = N(\tilde{\beta}_1\mathbf{x}, \sigma^2_{\epsilon_1} + \mathbf{x}^2\sigma^2_{\beta_1})$$

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 (suggested from experiments) $k = 4p$.

- this is a greedy algorithm which picks points one at a time
- when $k = 1$, the criterion becomes the Total Potential Energy

Fast Algorithm (2018)

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

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

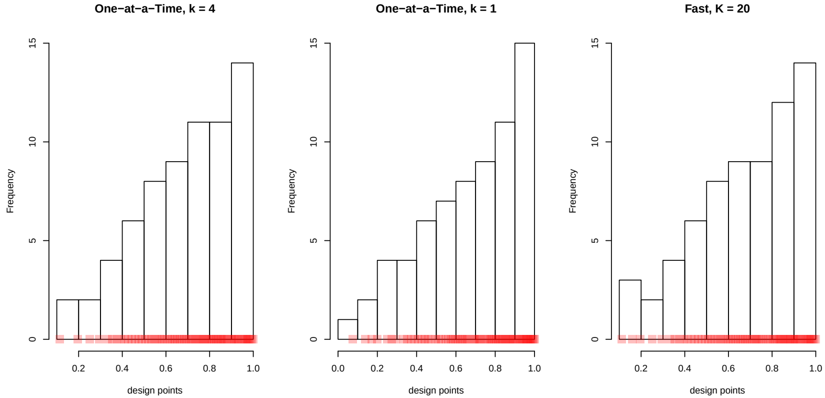
1. Initialize space-filling design $\mathbf{D}_1 = \{\mathbf{x}_1^{(1)} \dots \mathbf{x}_N^{(1)}\}$
2. For $k = 1, \dots, K-1$ steps, obtain each design point $\mathbf{x}_j^{(k+1)}$ of the next stage $\mathbf{D}_{k+1}$ by:

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

where $\gamma_k = k/(K-1)$ and $\mathbf{C}_j^{k+1}$ is the candidate set for design point $\mathbf{x}_j$ at stage $k+1$.

- points migrate to optimal location at each stage
- candidates are different for each design point

Designs from MED-Generating Algorithms



One−at−a−Time, k = 4

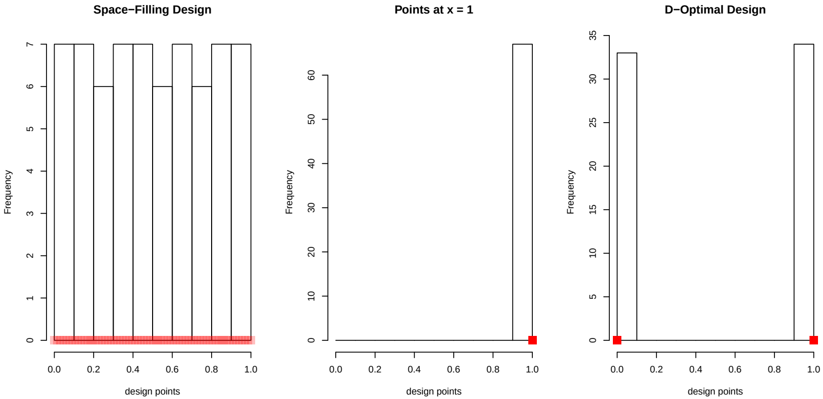
One−at−a−Time, k = 1

Fast, K = 20

Other Designs

Other Designs

- Random designs: 10 simulated random designs ($\mathbf{x} \sim U([0,1]^p)$, $\forall \mathbf{x} \in \mathbf{D}_{\text{random}}$).
  - Note: there is large variability for the criteria in designs with randomly chosen design points.
- Space-Filling Design: evenly spaced points.
- $X = 1$: $\forall \mathbf{x} \in \mathbf{D}, \mathbf{x} = 1$.
- D-optimal Design: seeks to minimize the variance of the estimated regression coefficients.
  - Alphabet-optimal desings are generated using `AlgDesign` package (using Federov's exchange algorithm).


Space−Filling Design
Frequency
0
1
2
3
4
5
6
7
0.0
0.2
0.4
0.6
0.8
1.0
design points
Points at x = 1
Frequency
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
design points
D−Optimal Design
Frequency
0
5
10
15
20
25
30
35
0.0
0.2
0.4
0.6
0.8
1.0
design points

The Table

Results!

	Fast	1atT,k=4	1atT,k=1	Random	Space	X = 1	D-opt
E[P(H0|Y)|H0]	0.699	0.699	0.699	0.703	0.717	0.717	0.717
E[P(H1|Y)|H0]	0.301	0.301	0.301	0.297	0.283	0.283	0.283
E[BF01 | H0]	12.6	12.6	12.6	11.7	13.7	13.8	13.8
E[P(H0|Y)|H1]	0.296	0.296	0.296	0.298	0.297	0.296	0.297
E[P(H1|Y)|H1]	0.704	0.704	0.704	0.702	0.703	0.704	0.703
E[BF01 | H1]	0.936	0.938	0.94	0.968	0.89	0.889	0.89
Post Var Slope	0.000287	0.000285	0.000289	0.000466	0.000442	0.000149	0.000293
TPE	2820000	2870000	2810000	7.28e+09	Inf	Inf	Inf
Fast Crit	44500	43700	97500	7.05e+09	Inf	Inf	Inf
1atT Crit (k=4)	94100	92500	120000	7.05e+09	Inf	Inf	Inf
Mean(D)	0.684	0.689	0.674	NA	0.5	1	0.507
sd(D)	0.23	0.219	0.247	NA	0.295	0	0.504

More Evaluations (D, De, A, I, Ge)

- D-efficiency, $De = \frac{\det(X^T X)^{(1/p)}}{N}$, is the relative number of runs (expressed as a percent) required by a hypothetical orthogonal design to achieve the same determinant value. It provides a way of comparing designs across different sample sizes.
    - When a design is orthogonal (all parameters can be estimated independently of each other), $De = 1$
    - De is proportional to the criterion of D-Optimal design, which seeks to maximize $\det(X^T X)$ (or minimize $\det((X^T X)^{-1})$). Hence, we want De close to 1.
- The A-Optimal design minimizes the average variance of the estimates of the regression coefficients: $\mathrm{tr}((X^T X)^{-1})/p$.
- The I-Optimal design seeks to minimize the average prediction variance over the design space.
- Ge, or G-efficiency, is available as a standard of design quality. It is good for minimizing the maximum variance of the predicted values.
    - Ge provides a lower bound on De for approximate theory: $De \geq exp(1 - \frac{1}{Ge})$. Hence, the closer to 1, the better.

When the model is given by $f(x) = (x)^T$,

	Fast	1atT,k=4	1atT,k=1	Space	X = 1	D-opt	I-opt
D	34.8	34.9	34.4	22.5	67	34	32.9
De	0.519	0.521	0.514	0.336	1	0.507	0.491
A	1.93	1.92	1.95	2.98	1	1.97	2.04
I	0.642	0.639	0.649	0.992	0.333	0.657	0.679
Ge	0.519	0.521	0.514	0.336	1	0.507	0.491

- Best performances:
  - Highest De: $X = 1$ design (MEDs were next best, but D-optimal design was close)
  - Lowest A: $X = 1$ design (MEDs were next best)
  - Lowest I: $X = 1$ design (MEDs were next best)
  - Highest Ge: $X = 1$ design (MEDs were next best)
- Note: Optimal designs were not optimized to this form.

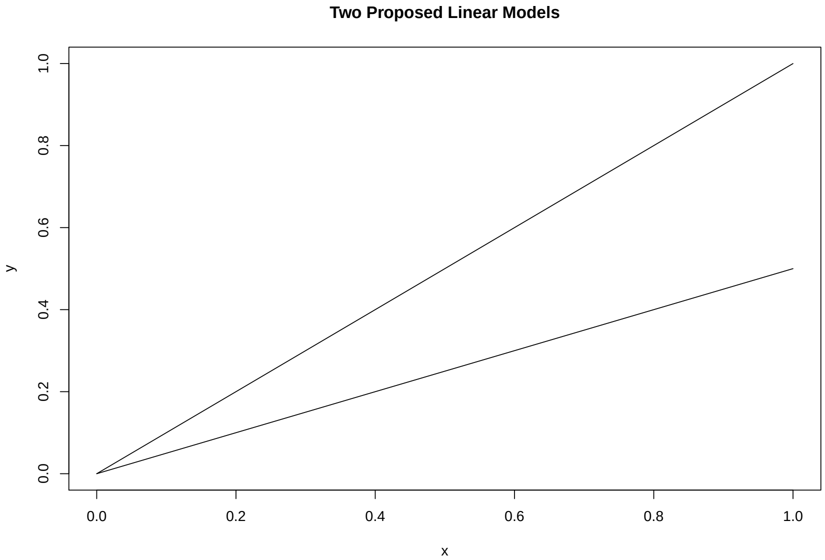
When $f(x) = (1, x)^T$,

	Fast	1atT,k=4	1atT,k=1	Space	X = 1	D-opt	I-opt
D	233	212	269	385	0	1120	1110
De	0.228	0.217	0.245	0.293	0	0.5	0.498
A	14.6	16.1	12.6	7.78	NA	3.02	3.01
I	3.25	3.52	2.89	1.97	NA	1.33	1.34
Ge	0.2	0.181	0.233	0.511	NA	0.985	0.982

- Here, MEDs did not do so well, but they weren't optimized for this model.
  - Highest *De*: D-optimal design (MEDs $>$ space-filling design)
  - Lowest *A*: I-optimal design (space-filling $>$ MEDs)
  - Lowest *I*: D-optimal design (space-filling $>$ MEDs)
  - Highest *Ge*: I-optimal design (space-filling $>$ MEDs)
  - This may suggest that MED might not be very robust...
- Note: `NA`s for $X = 1$ design are due to invertibility of $X^T X$ for design matrix $X = (\mathbf{1}_N \; \mathbf{1}_N)$

Simple Linear Regression: Unknown Slope and Intercept

Design an Experiment that Estimates Slope and Intercept



SetUp

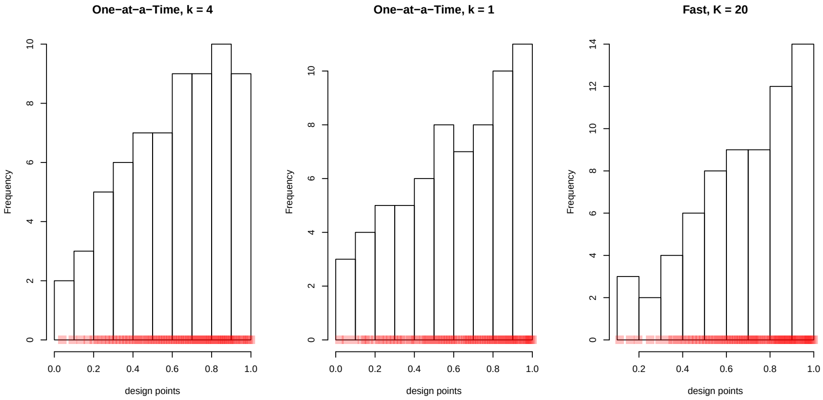
The set-up is similar, but there are some slight differences:

- Assume $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ and $\beta \sim N(\mu, V), V = \text{diag}(\nu_1^2, \nu_2^2)$.
- $Y|\beta \sim N(X\beta, \sigma^2 I)$
- $Y \sim N(X\mu, \sigma_m^2 I + XVX^T)$ after marginalizing out β

Hypotheses

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

$$H_0 : \beta \sim N\left(\mu_0, V_0\right)$$
$$H_1 : \beta \sim N\left(\mu_1, V_1\right)$$



One−at−a−Time, k = 4

One−at−a−Time, k = 1

Fast, K = 20

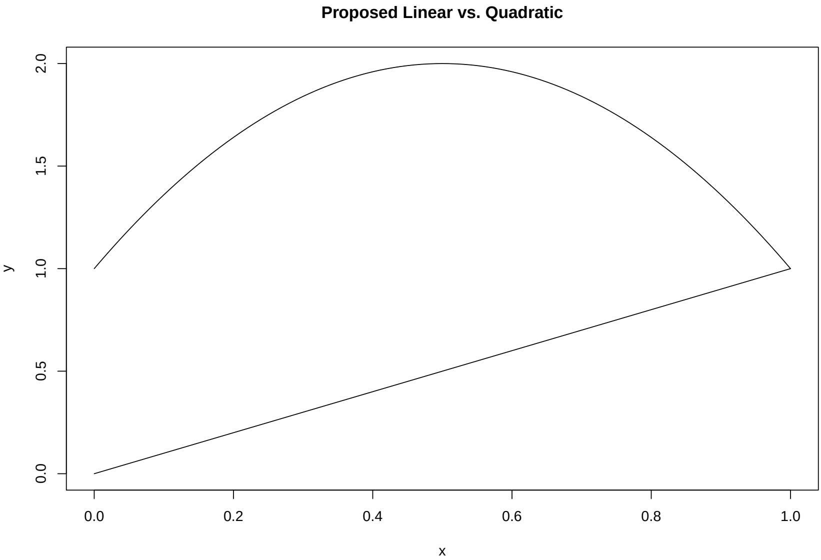
Table

	Fast	1atT,k=4	1atT,k=1	Space	X = 1	D-opt
E[P(H0|Y)|H0]	6.69e-63	4.58e-63	1.15e-65	4.89e-64	1.76e-67	1.44e-99
E[P(H1|Y)|H0]	1	1	1	1	1	1
E[BF01 | H0]	6.69e-63	4.58e-63	1.15e-65	4.89e-64	1.76e-67	1.44e-99
E[P(H0|Y)|H1]	6.58e-63	1.17e-63	2.29e-63	3.05e-64	4.2e-69	6.57e-98
E[P(H1|Y)|H1]	1	1	1	1	1	1
E[BF01|H1]	6.58e-63	1.17e-63	2.29e-63	3.05e-64	4.2e-69	6.57e-98
PostVar Int	0.000332	0.000291	0.000283	0.000235	0.000535	0.000197
PostVar Slope	0.00057	0.000572	0.000551	0.000558	0.000535	0.000345
TPE	2820000	2200000	2250000	Inf	Inf	Inf
Fast Crit	44500	23500	56000	Inf	Inf	Inf
1atT Crit (k=4)	94100	62100	77400	Inf	Inf	Inf
Mean(D)	0.684	0.61	0.606	0.5	1	0.507
sd(D)	0.23	0.254	0.274	0.295	0	0.504

- Compared to the alphabet-optimal designs, the MED methods allow the experimenter to determine how similar the intercepts and slopes are and determines the design points accordingly.

Linear vs Quadratic

Linear Model vs. Quadratic Model



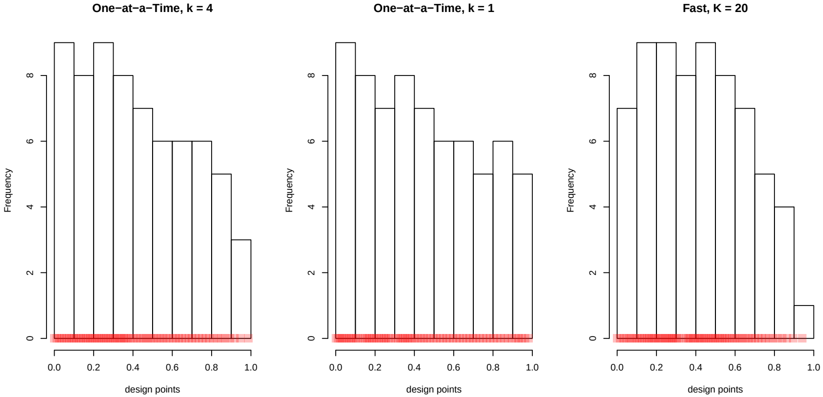
SetUp

As before,

- $Y|\beta \sim N(X\beta, \sigma^2 I)$
- $Y \sim N(X\mu, \sigma_m^2 I + XVX^T)$ after marginalizing out β

Hypotheses

$$\begin{aligned} H_0 : \beta &\sim N\left(\mu_0, V_0\right), \\ \mu_0 &= \left(\beta_{00}, \beta_{01}\right)^T, \\ V &= \text{diag}(\nu_{01}^2, \nu_{02}^2) \\ H_1 : \beta &\sim N\left(\mu_1, V_1\right), \\ \mu_0 &= \left(\beta_{10}, \beta_{11}, \beta_{12}\right)^T, \\ V &= \text{diag}(\nu_{11}^2, \nu_{12}^2, \nu_{13}^2) \end{aligned}$$


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

Table

	Fast	1atT,k=4	1atT,k=1	Space	X = 1	D-opt
E[P(H0|Y)|H0]	1	1	1	1	3.44e-28	0.987
E[P(H1|Y)|H0]	0	0	0	0	1	0.0129
E[BF01 | H0]	Inf	Inf	Inf	Inf	3.44e-28	1.8e+69
E[P(H0|Y)|H1]	0.94	0.934	0.95	0.996	0.527	4.02e-214
E[P(H1|Y)|H1]	0.06	0.0662	0.0496	0.00434	0.473	1
E[BF01|H1]	7.21e+85	2.01e+91	1.48e+93	1.92e+113	1.12	4.02e-214
PostVar b0	0.000222	0.00022	0.000225	0.000245	0.000682	0.000211
PostVar b1	0.000705	0.00069	0.000677	0.000672	0.000682	0.000604
PostVar b2	0.00079	0.000751	0.000726	0.000703	0.000682	0.000604
TPE	469000	439000	502000	Inf	Inf	Inf
Fast Crit	8240	5990	24600	Inf	Inf	Inf
1atT Crit (k=4)	14200	13100	27600	Inf	Inf	Inf
Mean(D)	0.417	0.426	0.449	0.5	1	0.507
sd(D)	0.249	0.274	0.291	0.295	0	0.504