Modifying MED for Model Selection

Kristyn Pantoja

1/23/2020

MED Overview

Minimum Energy Design

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

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

Theorem: If $q = \frac{1}{f^{1/2p}}$, the **limiting distribution**[1] of the design points is target distribution, f.

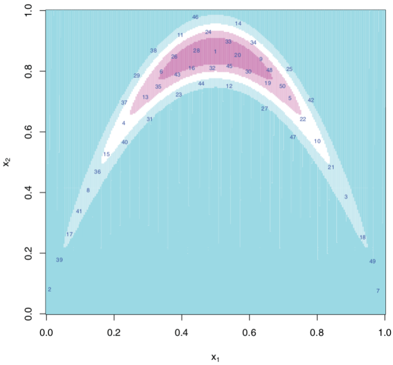


Figure 1: Sampling the "Banana" function

[1] "Sequential Exploration of Complex Surfaces Using Minimum Energy Designs," Joseph et. al. 2015, Result 1

MED for Model Selection

Goals

A 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

Define q in terms of $f_D(x)$, a normalized Wasserstein distance between $y|H_0, X$ and $y|H_1, X$, assuming a bounded design space.

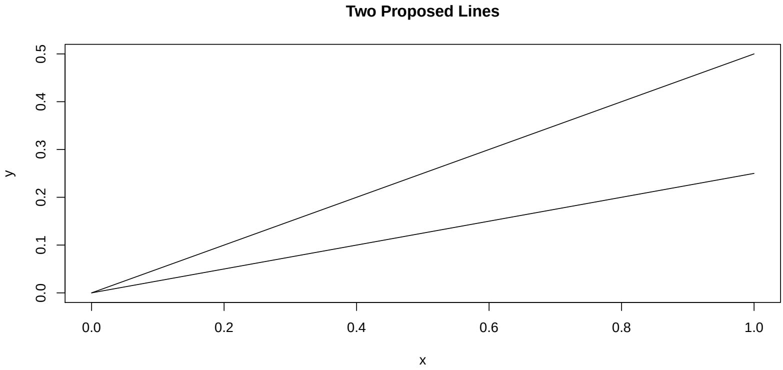
Modified Objective

$$q = \frac{1}{f_D^{1/2p}}$$

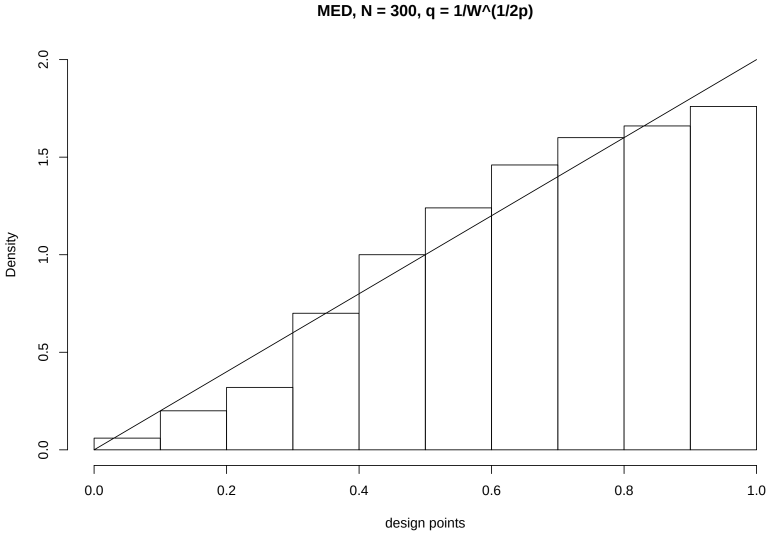
where $f_D(\mathbf{x}) = \text{Wasserstein}(\phi_{0,\mathbf{x}}, \phi_{1,\mathbf{x}})$,

- Here, the regions that are important for distinguishing the two models have high density.
- A tuning parameter α adjusts the space-filling aspect: $q_\alpha = 1/f_D^{\alpha/2p}$

Original Motivating Example



Limiting Distribution

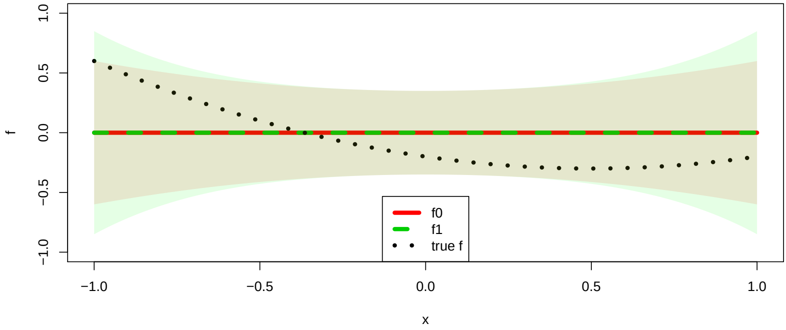


Cautionary Example

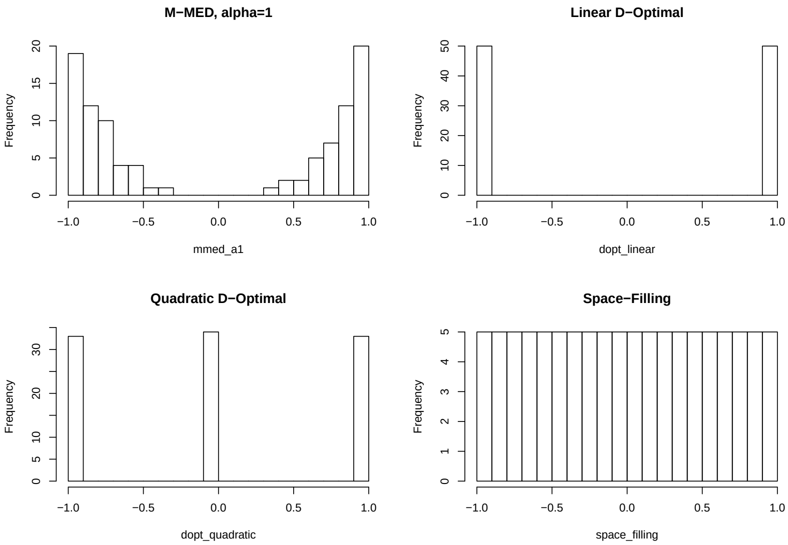
Suppose we want to consider a linear model and quadratic model:

$$H_0 : \beta \sim N((0,0)^T, \nu^2 I_2)$$
$$H_1 : \beta \sim N((0,0,0)^T, \nu^2 I_3)$$

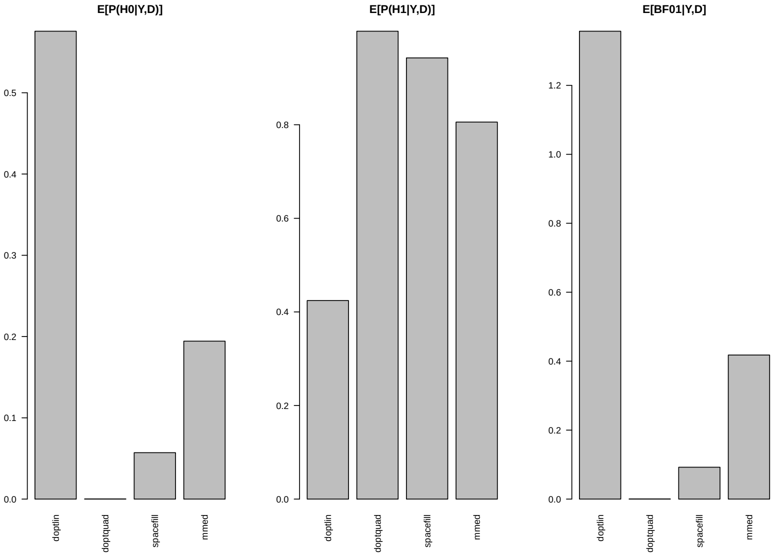
Consider the case where the true model is quadratic:
$\beta_T = (-0.2, -0.4, 0.4)$



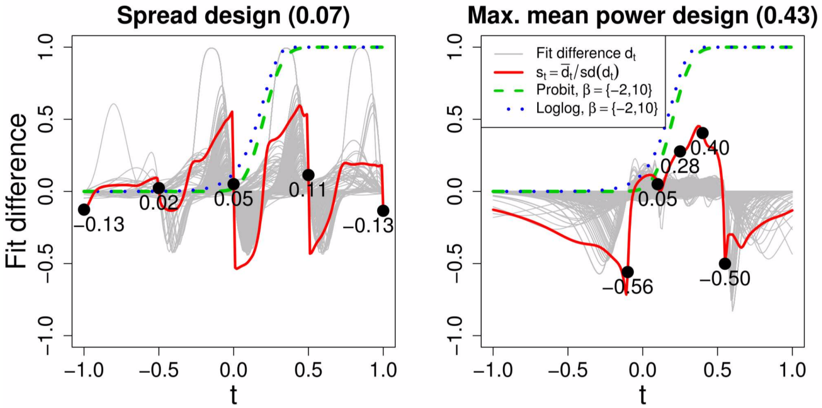
D-Optimal and Space-filling Designs



Posterior Probabilities



Points for Estimation



Points in the middle do not show large difference between the two models, but are importaint for constraining the models to be distinguished[2]

[2] "Designing Test Information and Test Information in Design", Jones & Meng

Sequential Modified MED

Sequential Design

If an experiment setting allows for data to be gathered sequentially, the modified MED (M-MED) can be adjusted to take into account data from previous experiments.

Currently, we have $q_\alpha = 1/f_D^{\alpha/2p}$, where $f_D(\mathbf{x}) = \text{Wasserstein}(\phi_{0,\mathbf{x}}, \phi_{1,\mathbf{x}})$

- M-MED: $\phi_{\ell,\mathbf{x}}$ is the marginal distribution of $y|H_\ell, X$

Taking data into account

- Sequential M-MED: $\phi_{\ell,\mathbf{x}}$ is the posterior predictive distribution[3] of $y|H_\ell, X$.

[3] See appendix

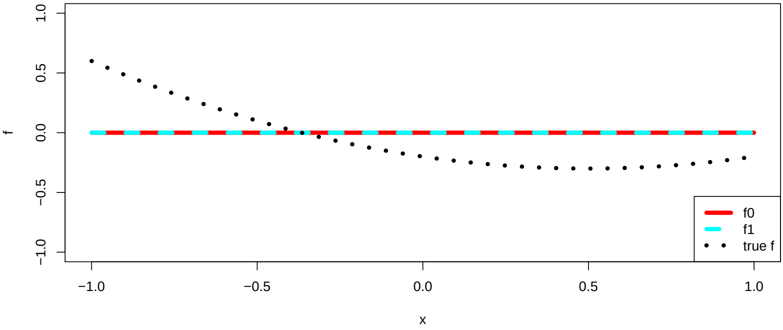
Case 1: Quadratic true model

Hypothesized and True Models

Consider the cautionary example again.

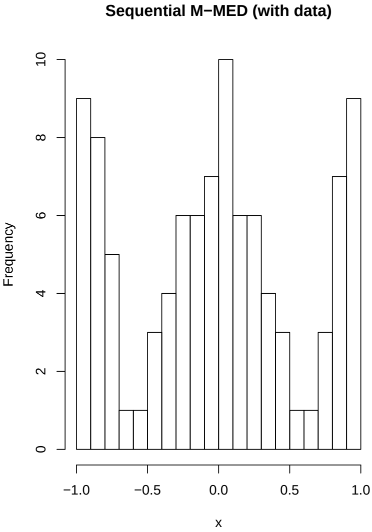
$$H_0 : \beta \sim N((0,0)^T, \nu^2 I_2)$$
$$H_1 : \beta \sim N((0,0,0)^T, \nu^2 I_3)$$

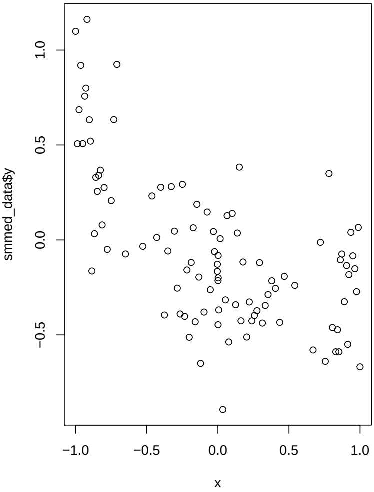
Consider the case where the true model is quadratic:
$\beta_T = (-0.2, -0.4, 0.4)$



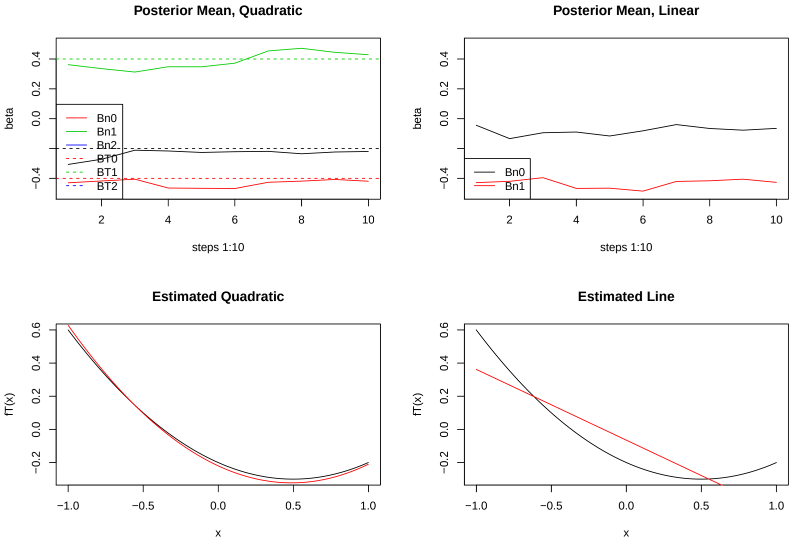
Sequential M-MED (using data)

A sequence of 10 steps, generating 10 points in each step, resulting in 100 points:

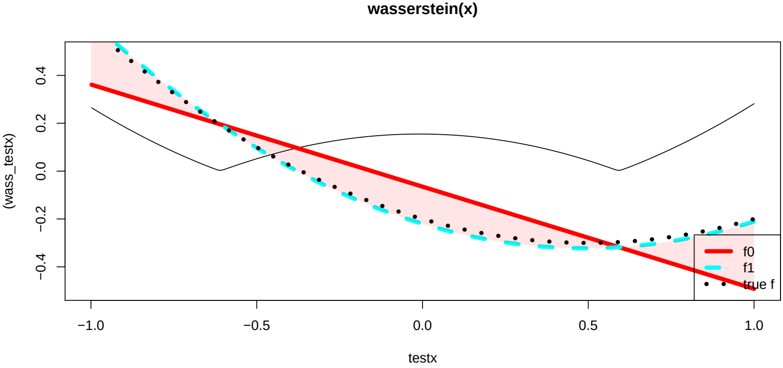




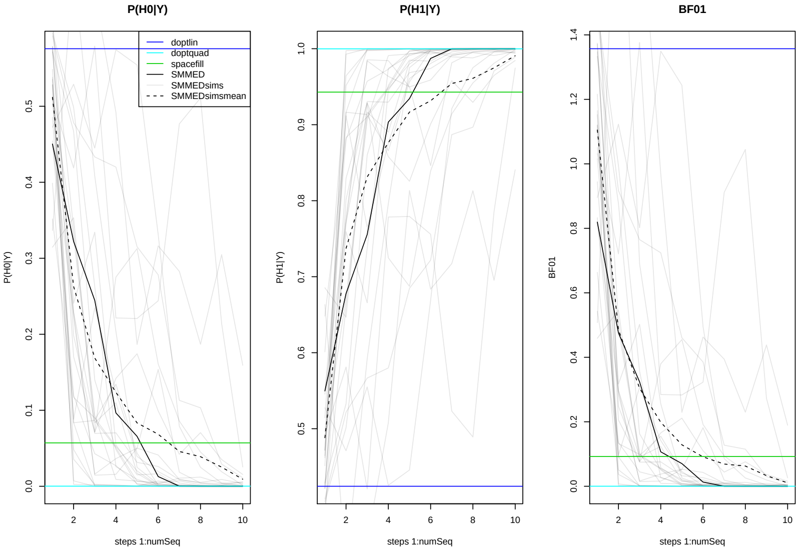
Linear and Quadratic Fits



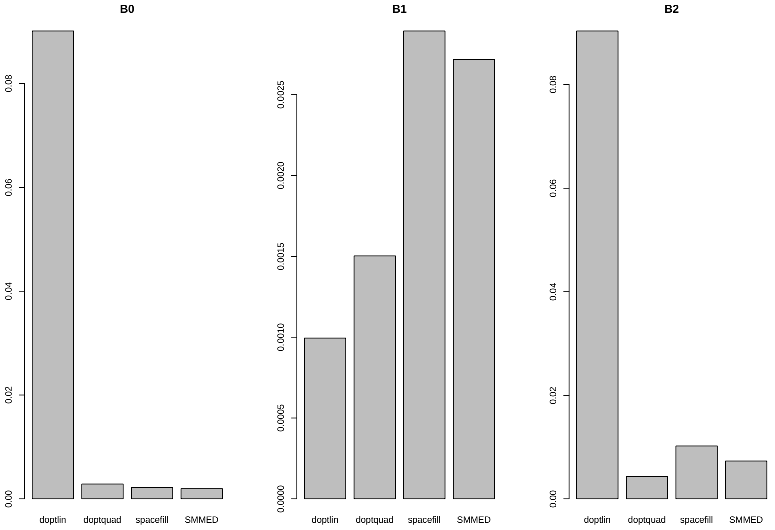
High Density Areas



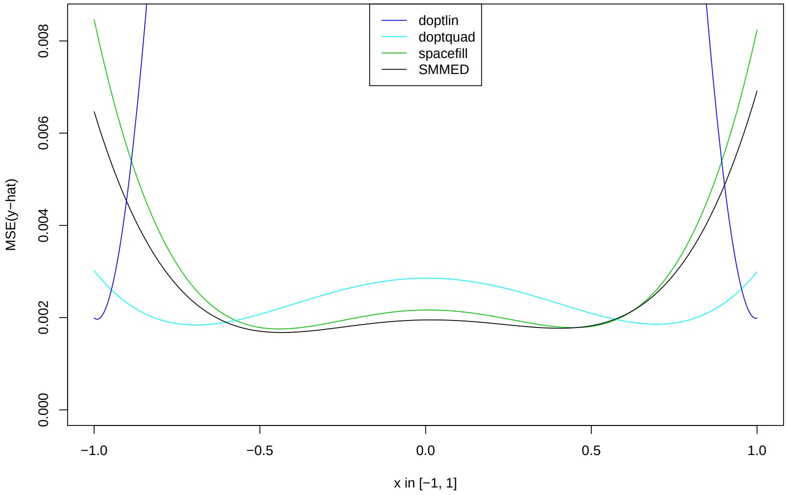
Hypothesis Testing



Parameter Estimation: MSE(Bn)



Prediction: MSE(y-hat)



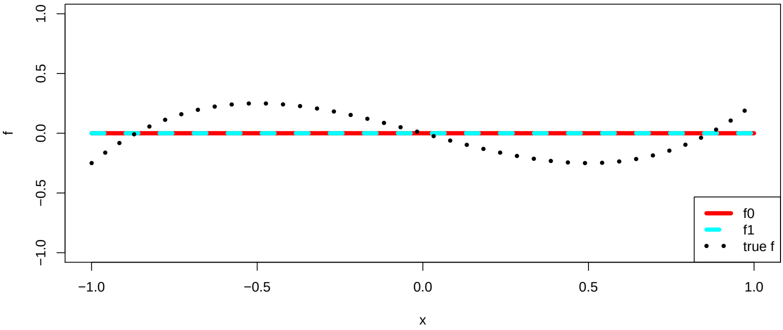
Case 2: Cubic

f0, f1, true f

Suppose we want to consider a linear model and quadratic model:

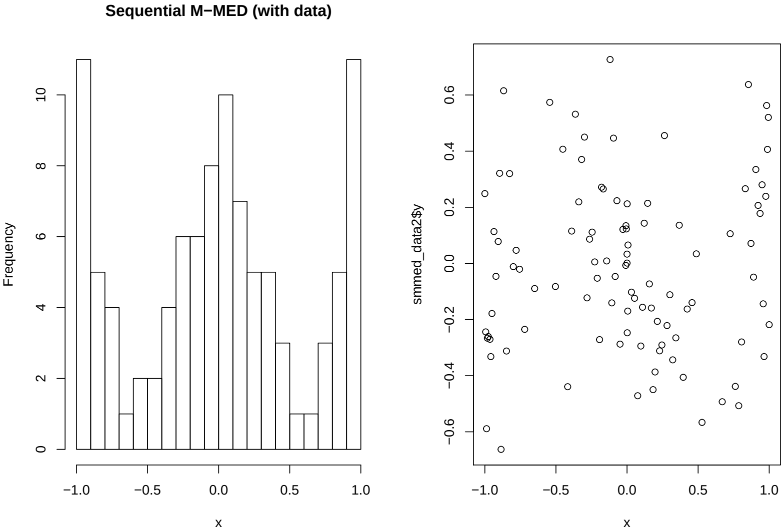
$$H_0 : \beta \sim N((0,0)^T, V_0)$$
$$H_1 : \beta \sim N((0,0,0)^T, V_0)$$

and suppose $\beta_T = (0, -0.75, 0, 1)$

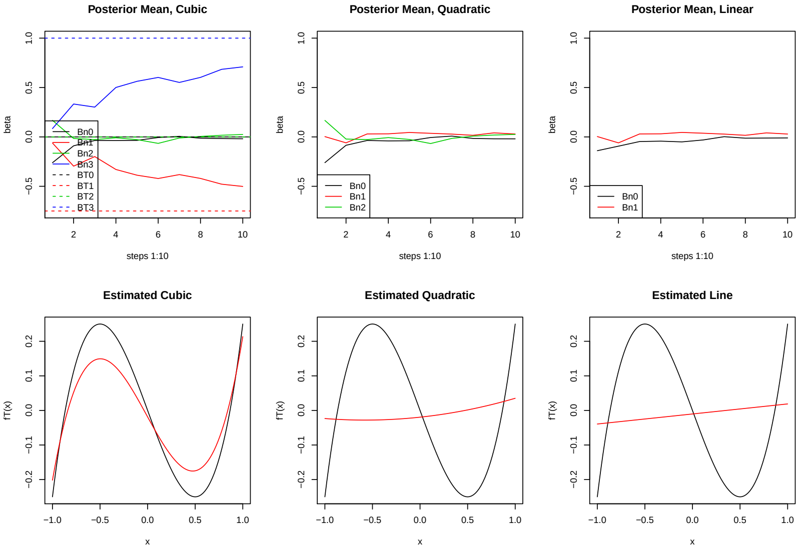


Sequential M-MED With Data

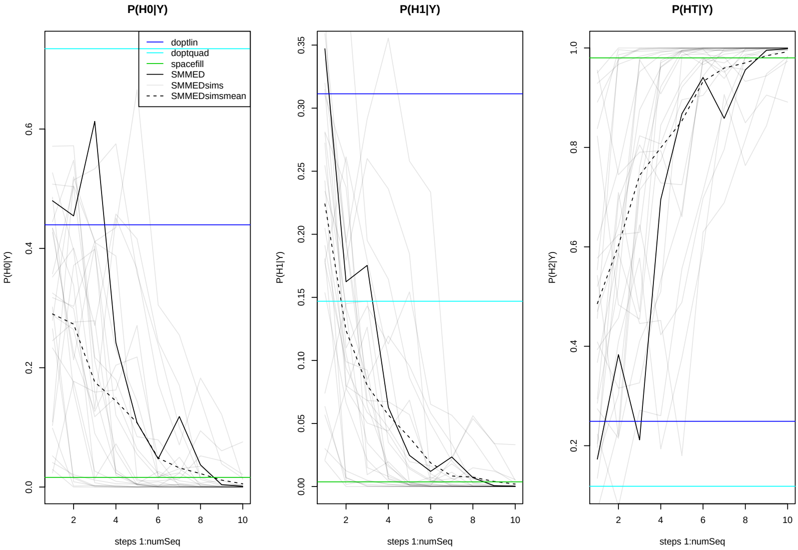
A sequence of 10 steps, generating 10 points in each step, resulting in 100 points:



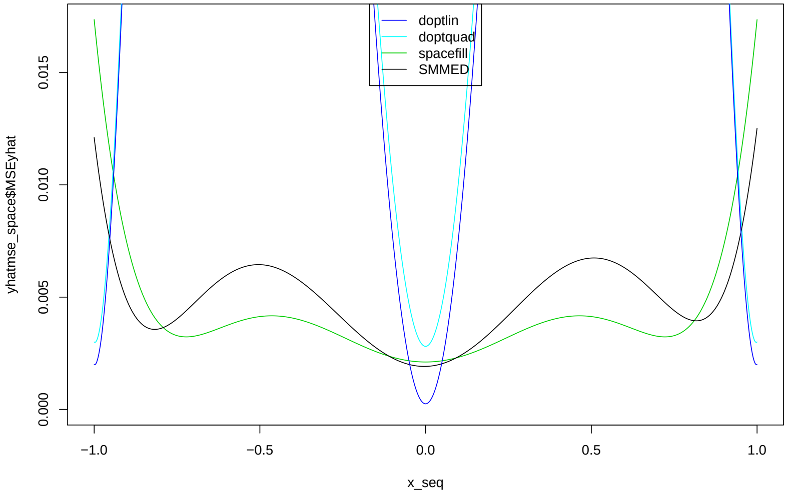
Linear, Quadratic, Cubic Fits



Hypothesis Testing



Prediction: MSE(y-hat)



Gaussian Process Application

Applying MED to Gaussian Process Model Selection

- Several covariance function options for Gaussian Process[4]. How to choose between two good options?
  - Squared Exponential: infinitely differentiable, standard choice
  - Matern: more reasonable smoothness assumptions
  - non-stationary options to capture structure in data

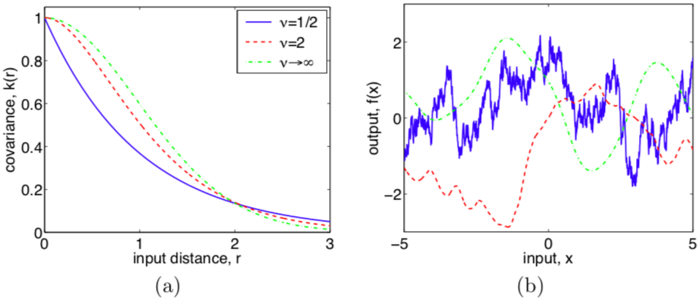


Figure 4.1: Panel (a): covariance functions, and (b): random functions drawn from Gaussian processes with Matérn covariance functions, eq. (4.14), for different values of ν, with $\ell = 1$. The sample functions on the right were obtained using a discretization of the x-axis of 2000 equally-spaced points.

[5]"Gaussian Processes for Machine Learning" Rasmussen et. al. 2005

Applying M-MED to Gaussian Process Model Selection

- Goal: Choose a design that will distinguish the two gaussian process models.
- Distinguishing functions vs. distributions over functions:
  - For regression models, we use $f_D(\mathbf{x}) = \text{Wasserstein}(\phi_{0,\mathbf{x}}, \phi_{1,\mathbf{x}})$. What is the distance function now? What are $\phi_{0,\mathbf{x}}, \phi_{0,\mathbf{x}}$?
  - Key Question: Do we need to consider the predictive distribution for each GP model?
    - Doing so would give us an option for $\phi_{0,\mathbf{x}}, \phi_{0,\mathbf{x}}$.
    - We would need to have at least some data in order to model each Gaussian Process (training set) and use M-MED to select points for comparing them.

Simulations Set-Up

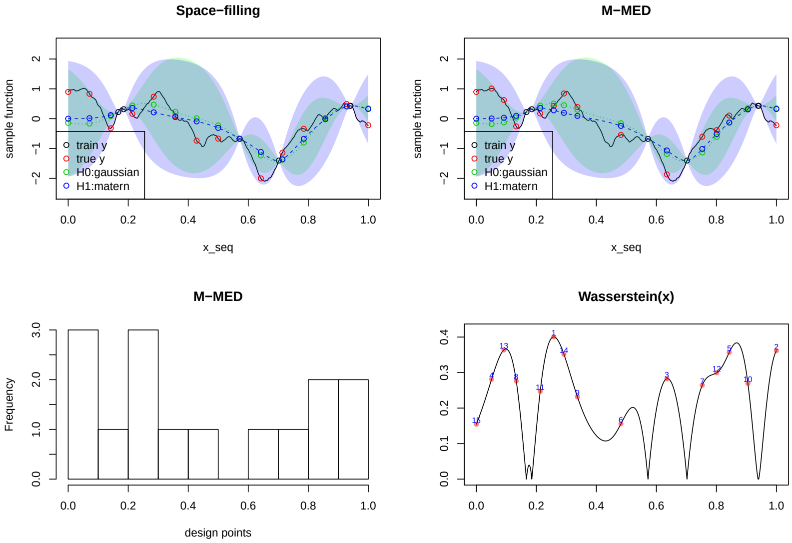
- I consider two cases:
  - Gaussian vs. Matern kernels, where the true function is generated from the Matern kernel
  - Matern vs. Periodic kernels, where the true function is generated from the Periodic kernel
- To evaluate MED for each case, I draw uniformly selected input points for my training set, and then apply MED to the data.
- I consider two measures for comparing MED to a space-filling design:
  - ratio of RSS for each hypothesized kernel:

$$\frac{\sum_{i \in \mathbf{D}} (y_i^{\text{pred}_0} - y_i^{\text{new}})^2}{\sum_{i \in \mathbf{D}} (y_i^{\text{pred}_1} - y_i^{\text{new}})^2}$$

  - likelihood ratio:

$$\frac{L(y^{\text{new}} | \xi, y^{\text{obs}}, \mathbf{X}^{\text{obs}}, \Theta = 0)}{L(y^{\text{new}} | \xi, y^{\text{obs}}, \mathbf{X}^{\text{obs}}, \Theta = 1)}$$

Gaussian vs. Matern (simulation)



Gaussian vs. Matern: log(RSS0/RSS1)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.62	0.053	0.25	0.54	0.62	5.5

Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.42	0.042	0.14	0.42	0.39	4.8

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 0.68
```

Gaussian vs. Matern: log ratio of predictive densities

(after removing `NA`s caused from non-invertible matrix)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-Inf	-2.9e+08	-4.8e+07	-Inf	-9.9e+05	-1.3e+05

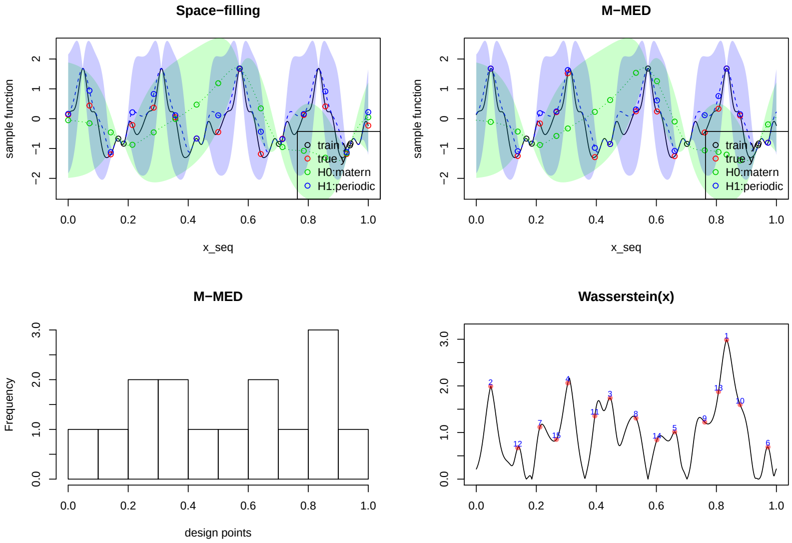
Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.7e+10	-1.2e+06	-3.7e+05	-1.1e+09	-9.6e+04	-1.8e+04

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 0.875
```

Matern vs. Periodic (simulation)



Matern vs. Periodic: log(RSS0/RSS1)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.56	1.3	1.8	2	2.2	3.6

Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.094	0.72	1	1.1	1.4	3

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 1
```

Matern vs. Periodic: log ratio of predictive densities

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-280	-160	-130	-140	-110	-65

Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-Inf	-99	-85	-Inf	-60	-38

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

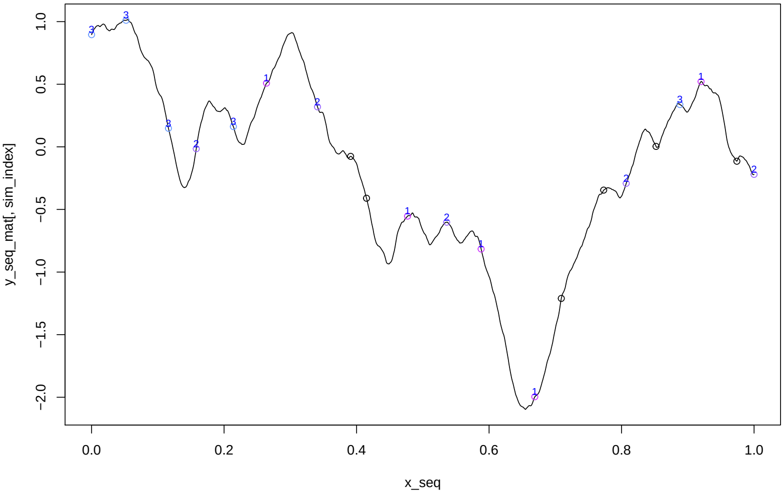
```
## [1] 0.92
```

Will a sequential design improve results?

For the sequential designs, I:

1. Start with 6 input data
2. Use SMMED to sequentially gather 15 new data points in 3 steps, with 5 new points
3. To compare SMMED to a space-filling design, I use the previous evaluations on the 15 new points (pretending that data was not gathered for them yet)

Gaussian vs. Matern (sequentially)



Gaussian vs. Matern: log(RSS0/RSS1)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.49	-0.082	0.27	0.48	0.64	2.8

Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-0.53	-0.21	0.24	0.45	0.57	2.5

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 0.64
```

Gaussian vs. Matern: log likelihood ratio (predictive)

(after removing `NA`s caused from non-invertible matrix)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-8.7e+05	-1.2e+05	-5.5e+04	-1.3e+05	-3.1e+04	-5.8e+03

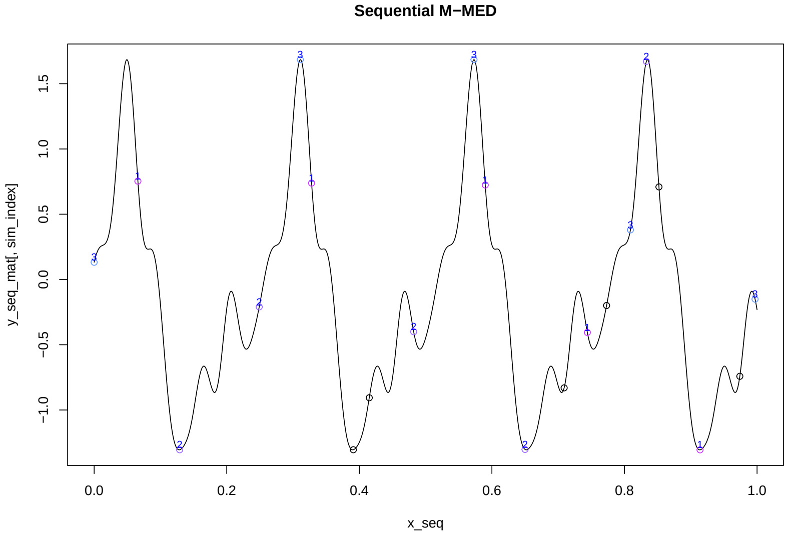
Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-9.6e+08	-1.0e+07	-2.8e+06	-6.6e+07	-3.6e+05	-2.9e+04

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 0
```

Matern vs. Periodic (sequentially)



Matern vs. Periodic: log(RSS0/RSS1)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.26	1.2	1.5	1.6	2	4.7

Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.35	0.9	1.3	1.3	1.7	2.5

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 0.56
```

Matern vs. Periodic: log likelihood ratio (predictive)

(after removing `NA`s caused from non-invertible matrix)

M-MED

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-4.9e+02	-2.5e+02	-1.6e+02	-2.1e+02	-1.3e+02	-8.9e+01

Space-filling

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-1.5e+02	-9.6e+01	-7.5e+01	-7.7e+01	-5.5e+01	-3.2e+01

Percentage of simulations that resulted in M-MED evaluations that were larger than Space-filling evaluations

```
## [1] 1
```

Appendix

Posterior Predictive Distribution of y

$[\tilde{y}|\tilde{x}, X, y, \sigma_\varepsilon^2, H_i, V_i]$ for brevity, call it $\tilde{y}|y$

$$f(\tilde{y}|y) = \int f(\tilde{y}|\beta; \tilde{x}, \sigma_\varepsilon^2) f(\beta|y, X, V_i, \sigma_\varepsilon^2) d\beta$$

where $f(\tilde{y}|\beta; \tilde{x}, \sigma_\varepsilon^2)$ is the pdf of $N(\tilde{x}^T\beta, \sigma_\varepsilon^2)$ and $f(\beta|y, X, V_i, \sigma_\varepsilon^2)$ is the posterior distribution of β; we denote the posterior mean and variance β_n and Σ_n, respectively.

Integrating out β leads to a normal distribution with mean

$$E[\tilde{y}|y] = E[E[\tilde{y}|\beta, y]] = E[\tilde{x}^T\beta|y] = \tilde{x}^T\beta_n$$

and with variance

$$\begin{aligned}\mathrm{Var}[\tilde{y}|y] &= E[\mathrm{Var}[\tilde{y}|\beta, y]] + \mathrm{Var}[E[\tilde{y}|\beta, y]] \\ &= \sigma_\varepsilon^2 + \mathrm{Var}[\tilde{x}^T\beta|y] = \sigma_\varepsilon^2 + \tilde{x}^T\Sigma_n\tilde{x}\end{aligned}$$

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 $\mathbf{D}_N$ by choosing $\mathbf{x}_1$ 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. For $j = 1, \ldots, N$, 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 $s = 1, \ldots, S-1$ stages, obtain each design point $\mathbf{x}_j^{(s+1)} \in \mathbf{D}_{s+1}$ by:

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

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

Posterior Probabilities of Hypotheses

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

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

  where π_m is the prior on H_m (typically $\pi_m = \frac{1}{M}$), and $f(y|H_m, X)$ is the model evidence, i.e. density of $N_N(X\mu_\ell, \sigma_\varepsilon^2 I + XV_\ell X^T)$ evaluated at a given y and design $\mathbf{D}$ with N design points.
- $P(H_\ell|y, X)$ tells which hypothesis is more likely to give the correct model.
- $E[P(H_\ell|y, X)|H_r, X]$ may be estimated using MC approximation from simulated responses y.
- $E[P(H_\ell|y, \mathbf{D})|H_r, \mathbf{D}]$ 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 $y_i^{(j)}|\mathbf{x}_i \sim N(\mathbf{x}_i^T \beta_T, \sigma_\varepsilon^2)$, $\forall \mathbf{x}_i \in \mathbf{D}$, so $y^{(j)} \in R^N$.
   1.2 $\forall m = \{0, 1\}$, calculate model evidences $f(y|H_m, \mathbf{D})$
   1.3 Calculate the posterior probability of H_ℓ, $P(H_\ell|y^{(j)}, \mathbf{D})$, from simulation j

$$P(H_\ell|y^{(j)}, \mathbf{D}) = \frac{f(y^{(j)}|H_\ell, X)}{f(y^{(j)}|H_0, X) + f(y^{(j)}|H_1, X)}$$

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

Note that $y^{(j)}$ are generated from $N_N(X\beta_T, \sigma_\varepsilon^2 I)$ and are independent, while the model evidence for H_m marginalizes out β and evaluates $y^{(j)}$ using $f(y|H_m, \mathbf{D})$, the density of $N_N(X\mu_m, \sigma_\varepsilon^2 I + XV_mX^T)$, in which they are no longer assumed to be independent.

Closed Form MSE of Posterior Mean

For notation, call $E[\beta|Y] = \beta_n$.

$$\begin{aligned} MSE(\beta_n) &= Var[\beta_n] + (E[\beta_n] - \beta_T)^2 \\ &= Var[\beta_n] + (E[\beta_n])^2 - 2\beta_T E[\beta_n] + \beta_T^2 \end{aligned}$$

where

$$\begin{aligned} Var[\beta_n] &= Var[\frac{1}{\sigma^2}\Sigma_B(X^T y + \sigma^2 V^{-1}\mu)] = Var[\frac{1}{\sigma^2}\Sigma_B X^T y] \\ &= (\frac{1}{\sigma^2})^2 \Sigma_B X^T Var[y] X \Sigma_B = (\frac{1}{\sigma^2})^2 \Sigma_B X^T (\sigma^2 I) X \Sigma_B \\ &= \frac{1}{\sigma^2}\Sigma_B X^T X \Sigma_B \\ E[\beta_n] &= E[\frac{1}{\sigma^2}\Sigma_B(X^T y + \sigma^2 V^{-1}\mu)] = \frac{1}{\sigma^2}\Sigma_B(X^T E[y] + \sigma^2 V^{-1}\mu) \\ &= \frac{1}{\sigma^2}\Sigma_B(X^T X\beta_T + \sigma^2 V^{-1}\mu) = \frac{1}{\sigma^2}\Sigma_B X^T X \beta_T + \Sigma_B V^{-1}\mu \end{aligned}$$

where $\Sigma_B = Var[\beta|y] = \sigma^2(X^T X + \sigma^2 V^{-1})^{-1}$ and $y \sim N(X\beta_T, \sigma^2 I)$

Closed Form MSE of y-hat

For an unseen point $\mathbf{x}_*$, its predicted response $\hat{y} = \mathbf{x}_*^T \beta_n$, where β_n is the posterior mean of β.

$$\begin{aligned} MSE(\hat{y}) &= Var[\hat{y}] + Bias^2(\hat{y}) \\ &= Var[\mathbf{x}_*^T \beta_n] + E[\hat{y} - y_T]^2 \\ &= \mathbf{x}_*^T Var[\beta_n] \mathbf{x}_* + E[\mathbf{x}_*^T \beta_n] - \mathbf{x}_*^T \beta_T \\ &= \mathbf{x}_*^T Var[\beta_n] \mathbf{x}_* + \mathbf{x}_*^T E[\beta_n] - \mathbf{x}_*^T \beta_T \end{aligned}$$

where $E[\beta_n]$ and $Var[\beta_n]$ were calculated in the previous slide.

T-Optimal Designs

Comparing linear model with fixed parameters against the quadratic model parameters allowed to vary

points	weights
-1	0.25
0	0.50
1	0.25

E[P(Hi|Y,D)] with T-Optimal Designs