Modifying MED for Model Selection

Kristyn Pantoja

12 February 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.

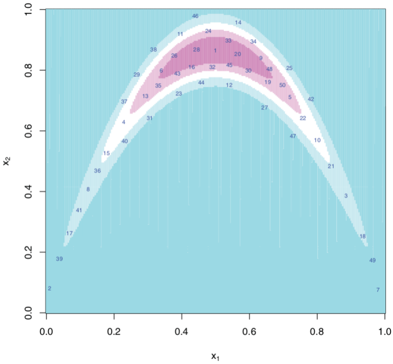


Figure 1: Sampling the "Banana" function

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

Modifying 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 distance between $y|H_0, x$ and $y|H_1, x$:

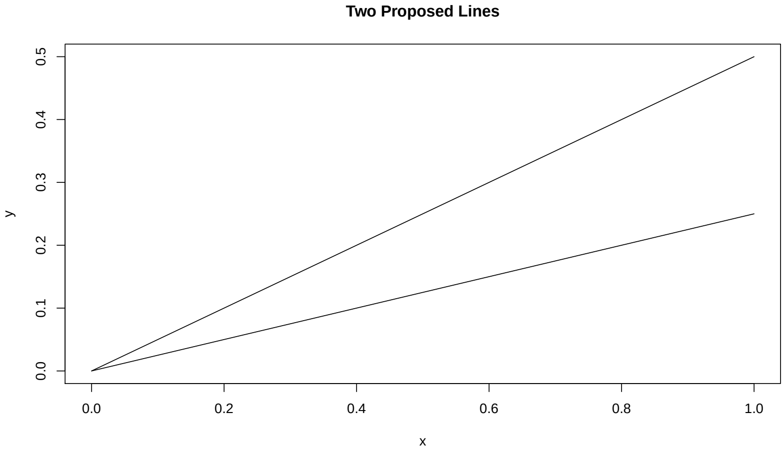
Modified Objective

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

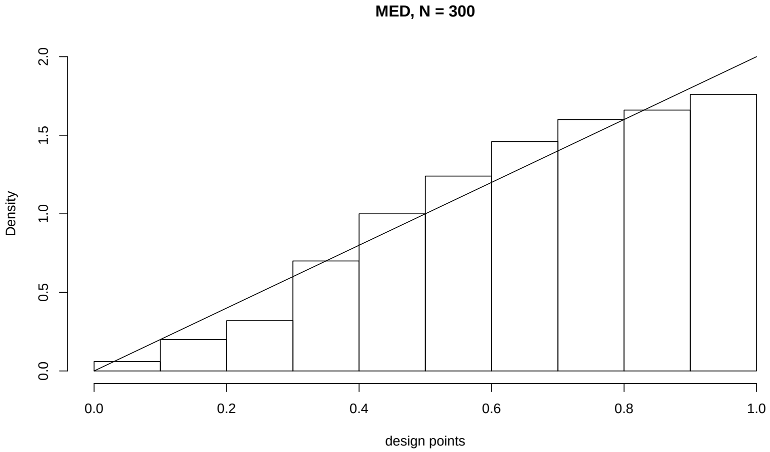
where $f_D(x) = c\text{Wasserstein}(\phi_{0,\mathbf{x}}, \phi_{1,\mathbf{x}})$

- f_D can be normalized to be a probability density, assuming a bounded design space.
- Here, the regions that are important for distinguishing the two models have high density.

Original Motivating Example



Limiting Distribution

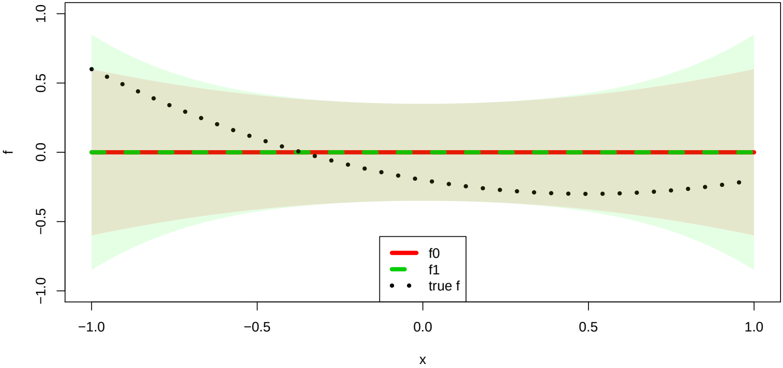


Cautionary Example

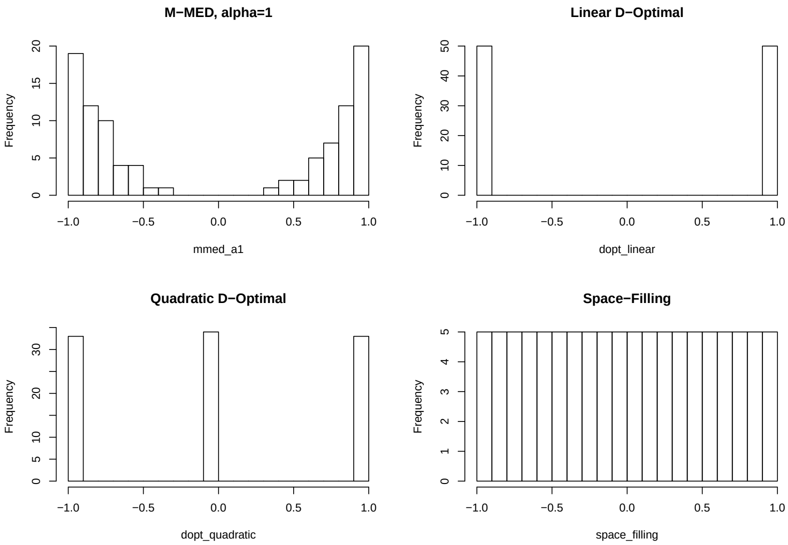
Suppose we want to consider a linear model and quadratic model:

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

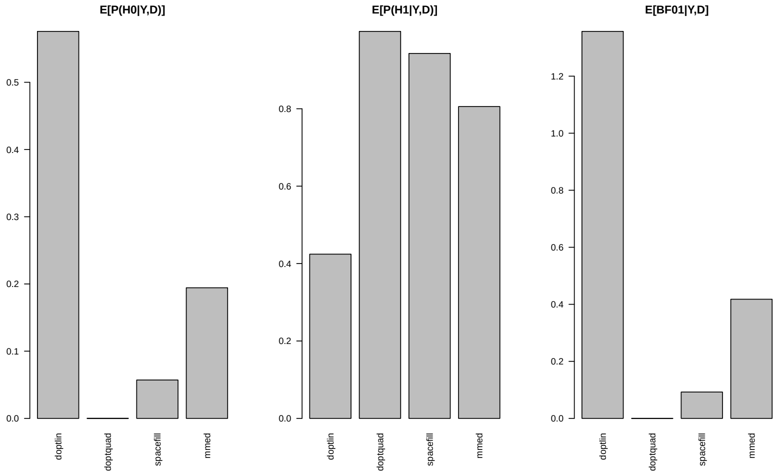
and the true model is quadratic, with $\beta_T = (-0.2, -0.4, 0.4)^T$



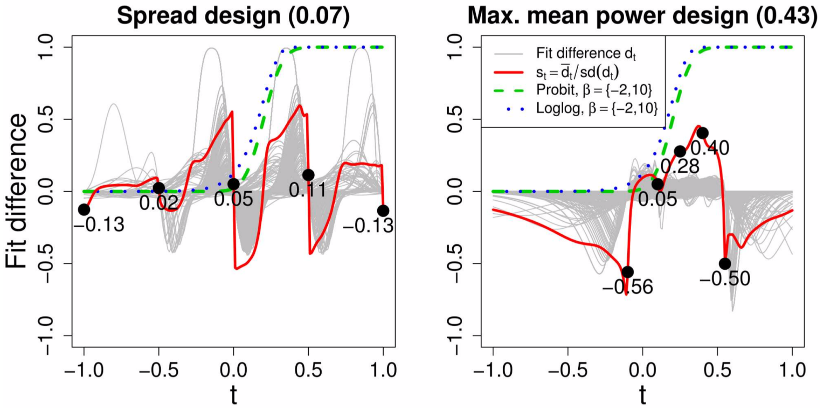
D-Optimal and Space-filling Designs



Expected Posterior Probabilities of Hypotheses



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 M-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 have $q = 1/f_D^{1/2p}$, 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 of $y|H_\ell, X$.

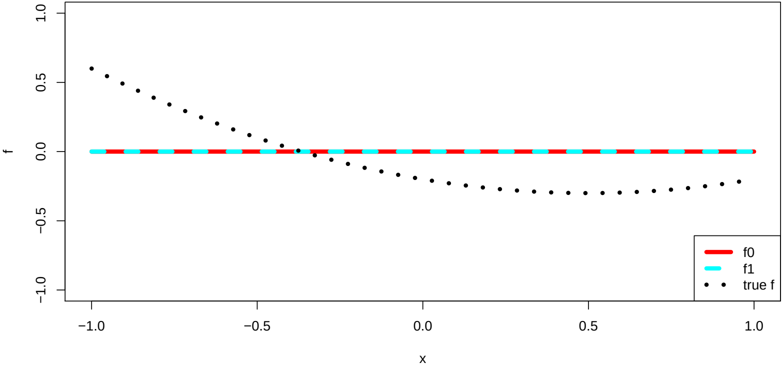
Case 1: Quadratic true model

True Model and Hypothesized Models

Consider the cautionary example again.

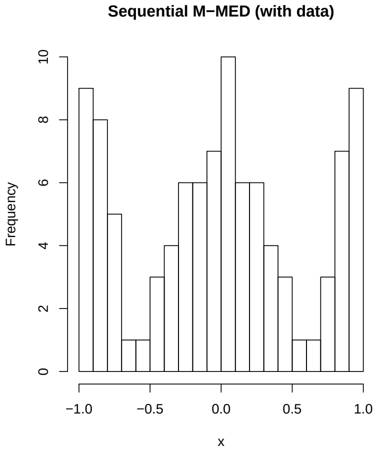
$$H_0 : \beta \sim N((0,0)^T, \nu^2 I_2)$$
$$H_1 : \beta \sim N((0,0,0)^T, \nu^2 I_3)$$

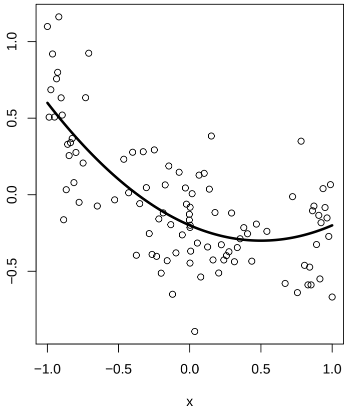
where the true model is quadratic: $\beta_T = (-0.2, -0.4, 0.4)$



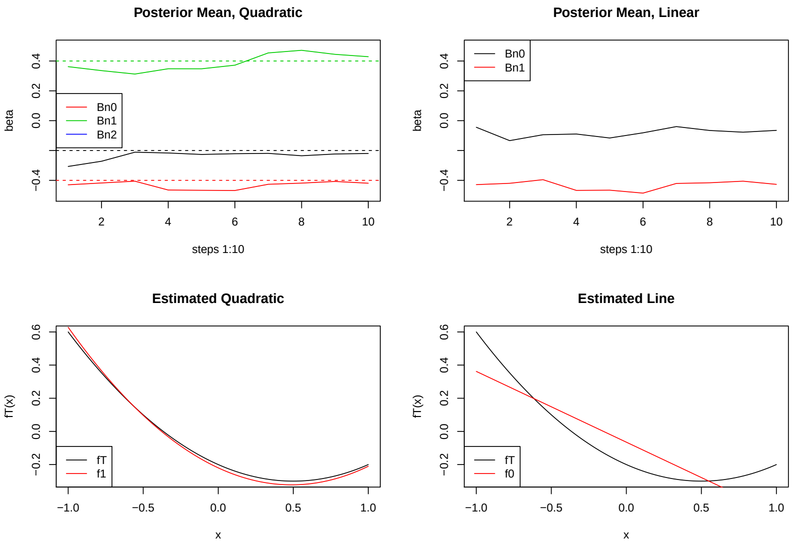
Sequential M-MED (and simulated data)

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

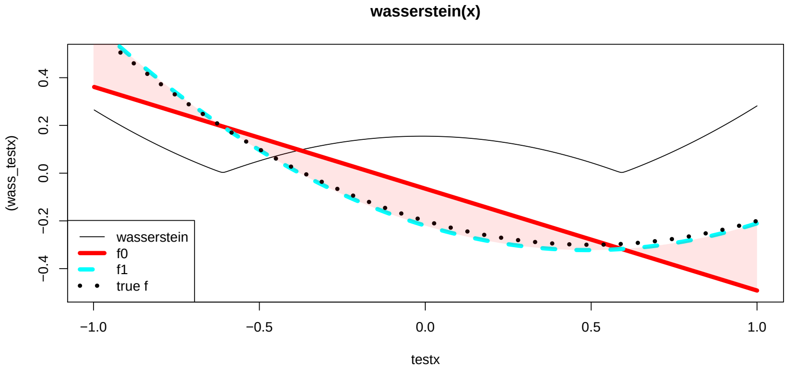




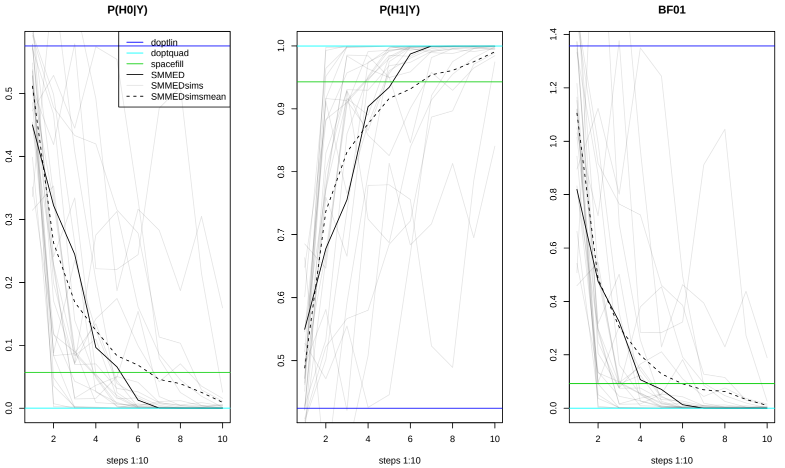
Linear and Quadratic Fits



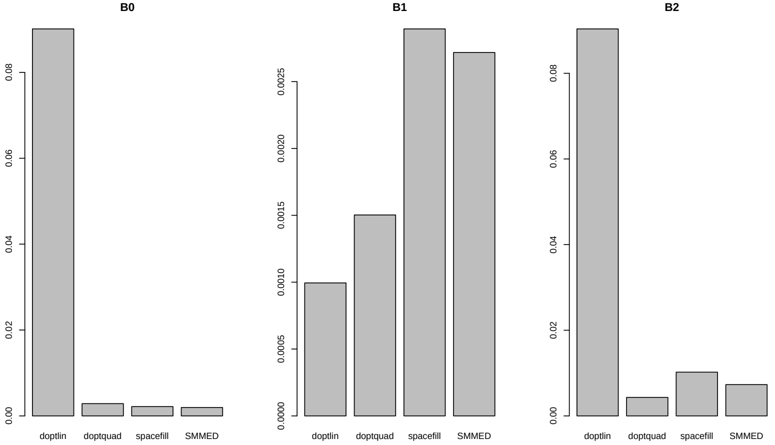
High Density Areas



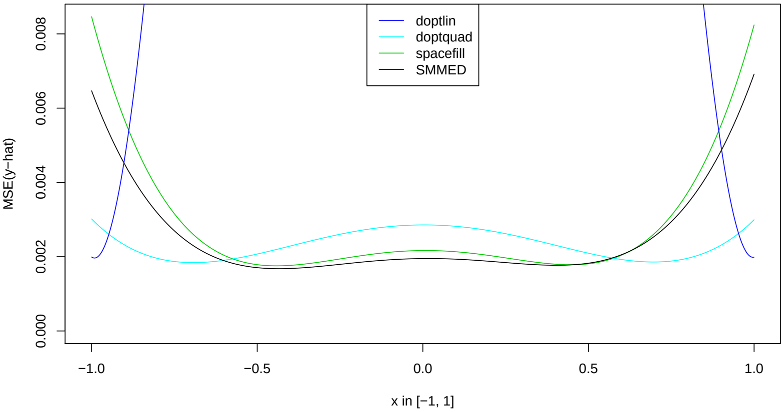
Model Selection: E[P($H_\ell | Y, D$)]



Parameter Estimation: MSE(B_n)



Prediction: MSE($\hat{y}$)



Case 2: Cubic true model

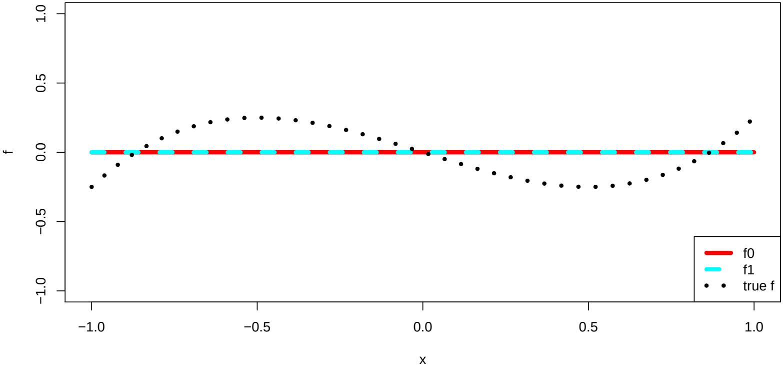
True Model and Hypothesized Models

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

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

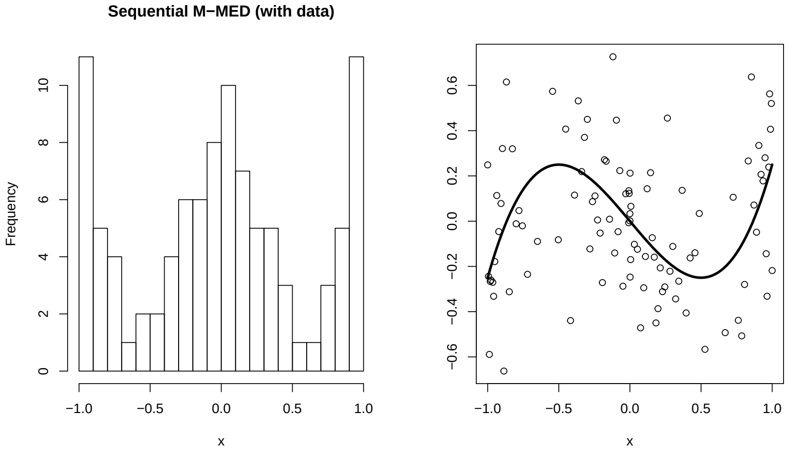
$$H_1 : \beta \sim N((0,0,0)^T, V_0)$$

and the true model is cubic with coefficients $\beta_T = (0, -0.75, 0, 1)^T$

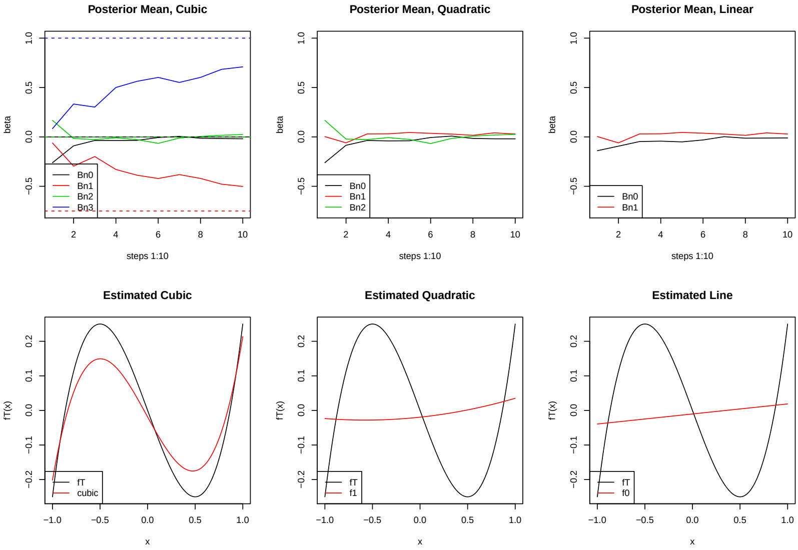


Sequential M-MED With Data

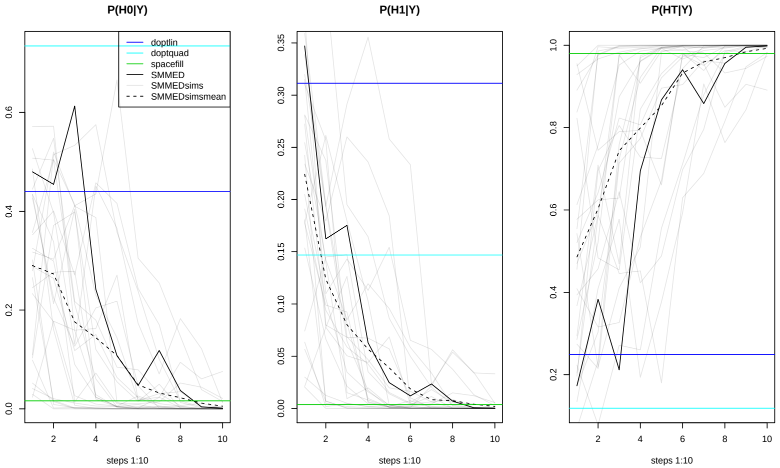
A sequence of 10 steps, generating 10 points in each step, resulting in 100 points:



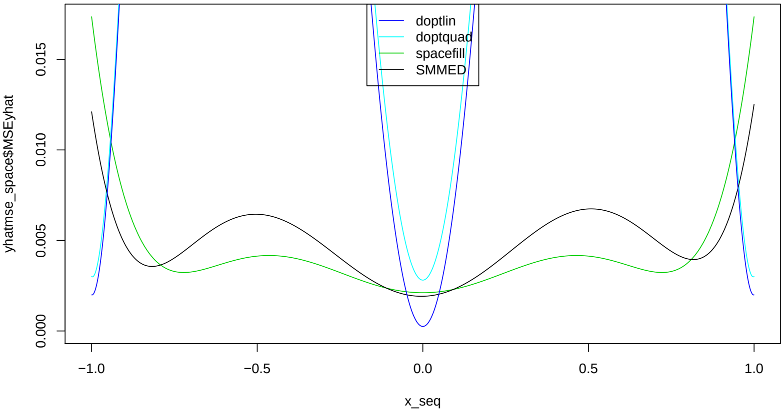
Linear, Quadratic, Cubic Fits



Model Selection: $\mathrm{E}[\mathrm{P}(H_\ell | Y, D)]$



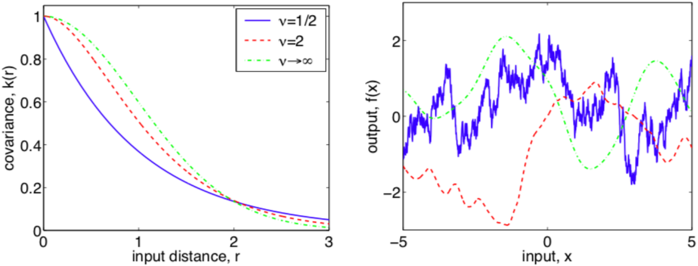
Prediction: MSE($\hat{y}$)



Application to Gaussian Process Model Selection

Gaussian Process Model Selection

- Several covariance function options for Gaussian Process[3]. 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



(a) (b)

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

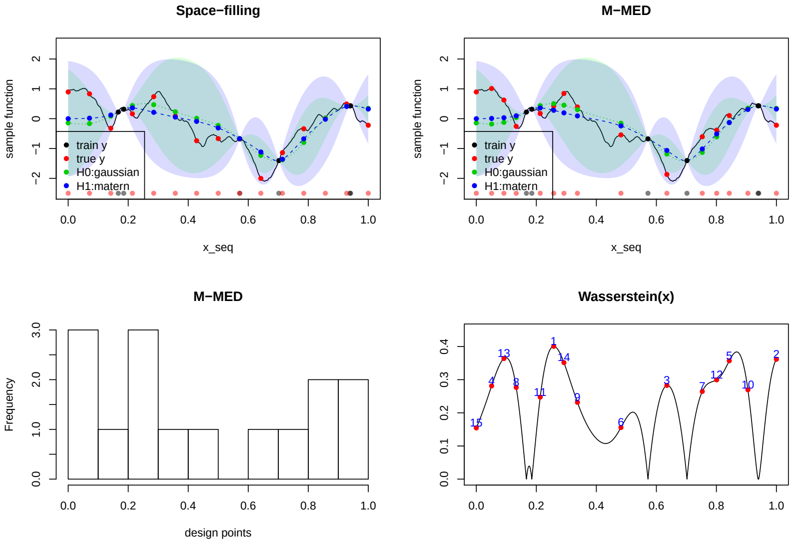
Applying SM-MED to Gaussian Process

- 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 $\phi_{0,\mathbf{x}}, \phi_{0,\mathbf{x}}$?
  - Why consider the predictive distribution for each GP model:
    - We can use posterior predictive distributions for $\phi_{0,\mathbf{x}}, \phi_{0,\mathbf{x}}$ in f_D.
    - We would need to have at least some initial data in order to model each Gaussian Process and use M-MED to select points for comparing them.

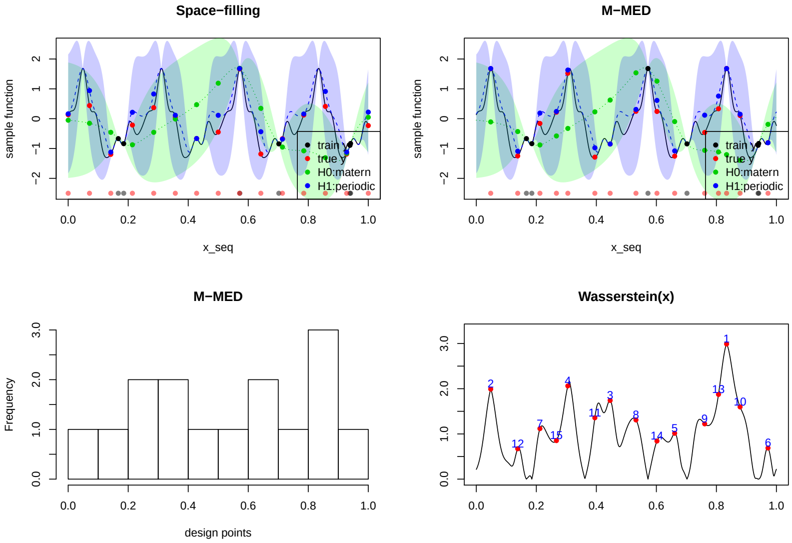
Hypothesis tests

1. Squared Exponential vs. Matern, f is generated from Matern
2. Matern vs. Periodic, f is generated using from Periodic

Sq Exponential vs. Matern (f generated from Matern)



Matern vs. Periodic (f generated from Periodic)



Different Input Cases

Simulations Set-Up

To generate points sequentially with MMED:

1. Start with 6 input data
2. Use SMMED to sequentially gather 15 new data points in 3 steps (5 new points at each step)

Comparing SM-MED to other designs

I compare SM-MED to space-filling and uniformly distributed design points using these metrics:

- RSS0/RSS1 (prediction)

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

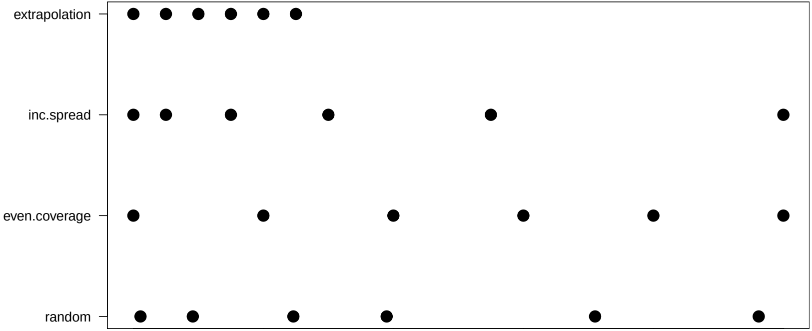
- $P(H_1|X, Y)$ (model selection)

$$\frac{P(Y|H_1, X)\pi_1}{P(Y|H_0, X)\pi_0 + P(Y|H_1, X)\pi_1}$$

4 Input Cases

We consider 4 cases which use different input points for each of the 2 hypothesis tests and 3 designs:

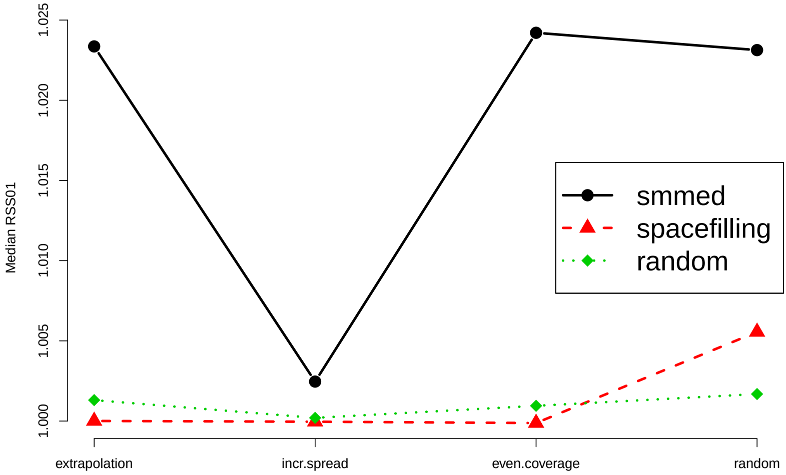
- Extrapolation
- Increasing spread
- Even coverage
- Random



Squared Exponential vs. Matern

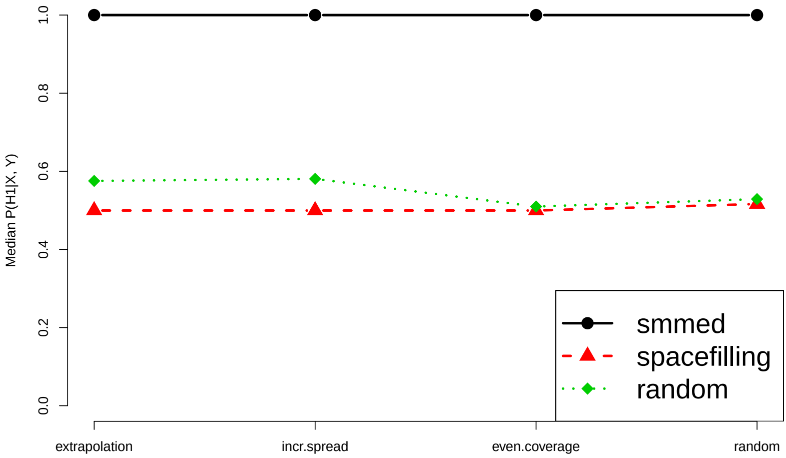
Sq Exponential vs. Matern: RSS Ratio (01)

Compare the median $\frac{RSS0}{RSS1}$ for each case and each design



Sq Exponential vs. Matern: P($H_1|Y, X$)

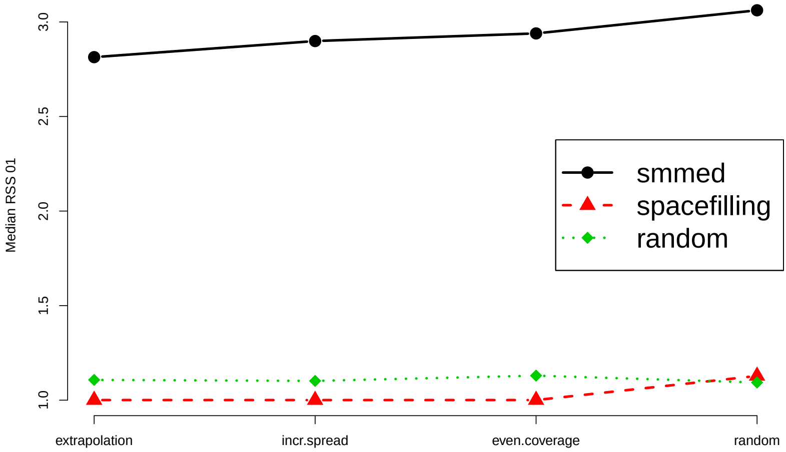
Compare the median $P(H_1|Y, X)$ for each case and each design



Matern vs. Periodic

Matern vs. Periodic: RSS Ratio (01)

Compare the median $\frac{RSS0}{RSS1}$ for each case and each design



Matern vs. Periodic: $P(H_1|Y,X)$

Compare the median $P(H_1|Y,X)$ for each case and each design