Modifying MED for Model Selection

Kristyn Pantoja

6 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.

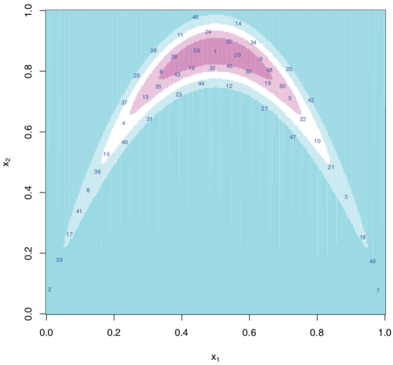


Figure 1: Sampling the "Banana" function

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

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.

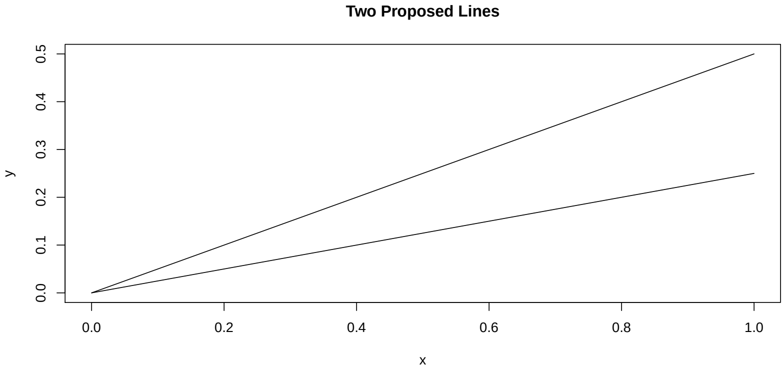
Modified Objective

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

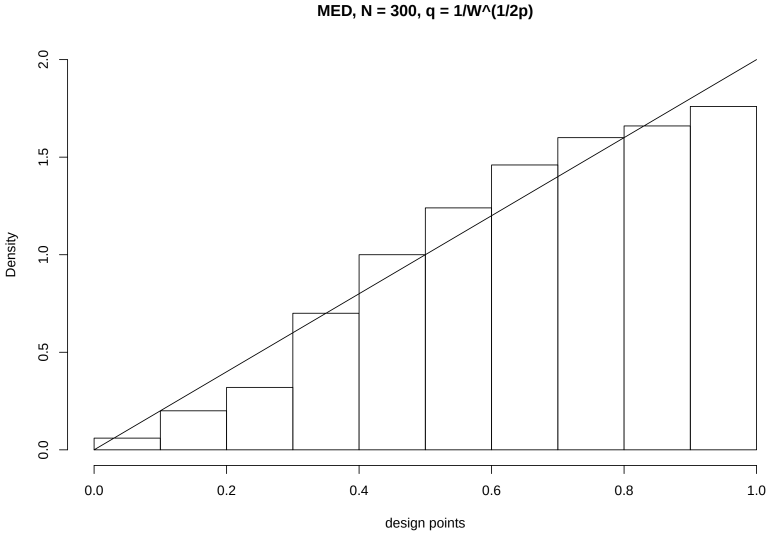
where $f_D(\mathbf{x}) = \text{Wasserstein}(\phi_{0,\mathbf{x}}, \phi_{1,\mathbf{x}})$,

- 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

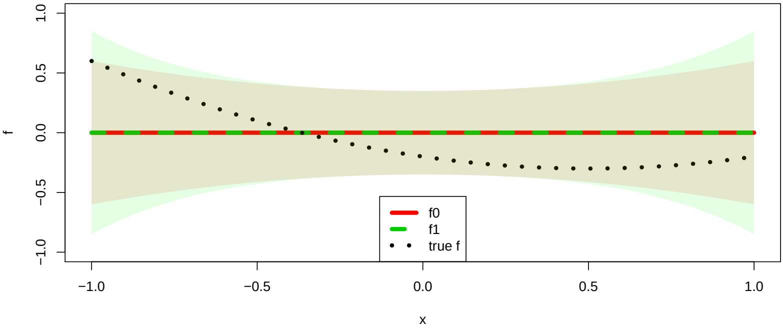


Cautionary Example

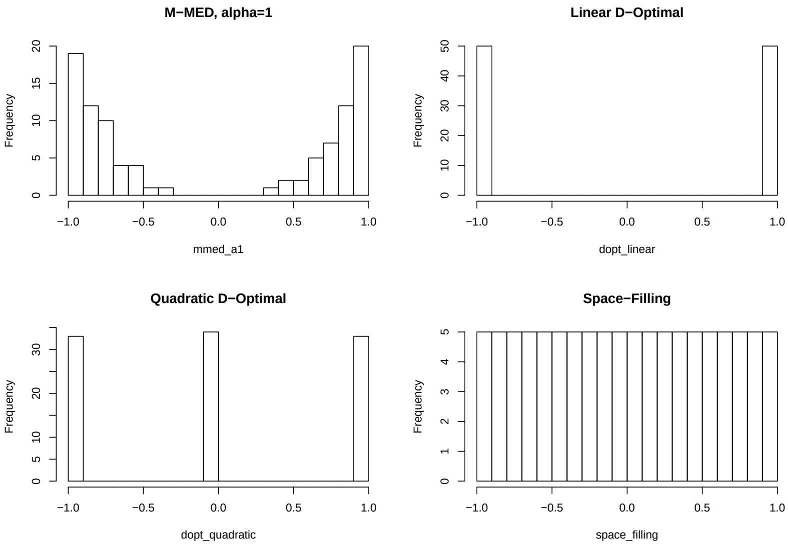
Suppose we want to consider a linear model and quadratic model:

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

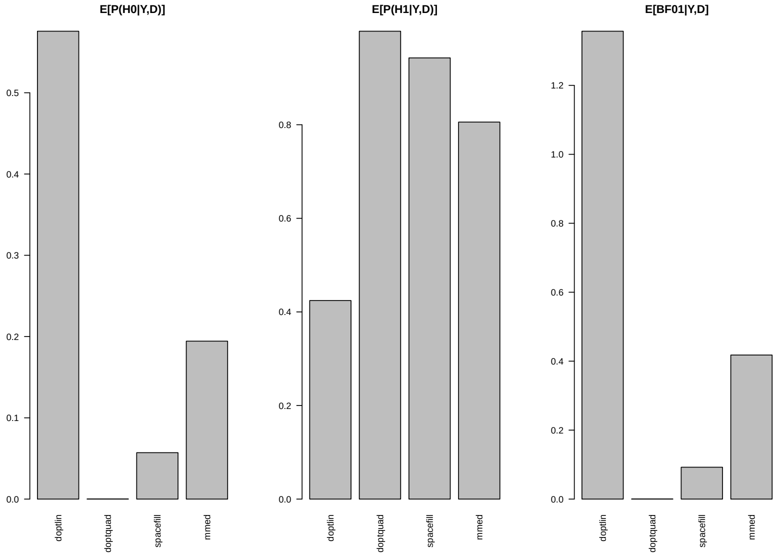
Consider the case where the true model is quadratic:
$\beta_T = (-0.2, -0.4, 0.4)$



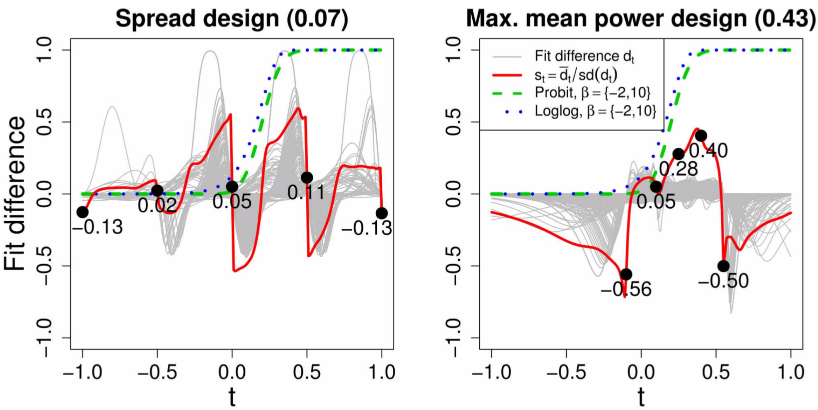
D-Optimal and Space-filling Designs



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

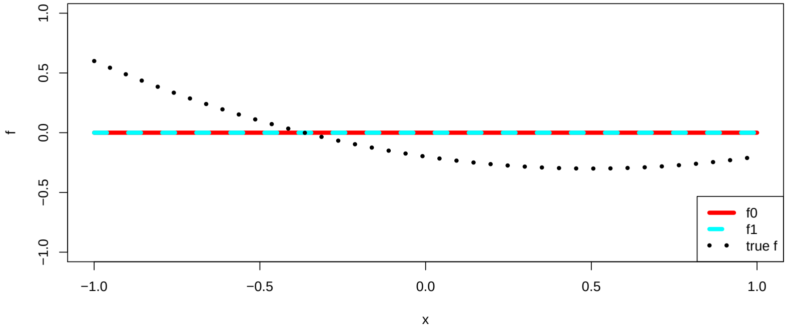
Case 1: Quadratic true model

Hypothesized and True Models

Consider the cautionary example again.

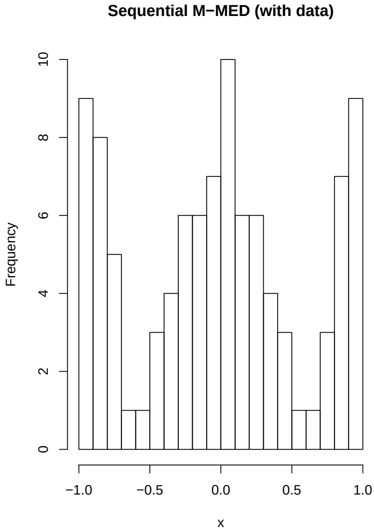
$$H_0 : \beta \sim N((0,0)^T, \nu^2 I_2)$$
$$H_1 : \beta \sim N((0,0,0)^T, \nu^2 I_3)$$

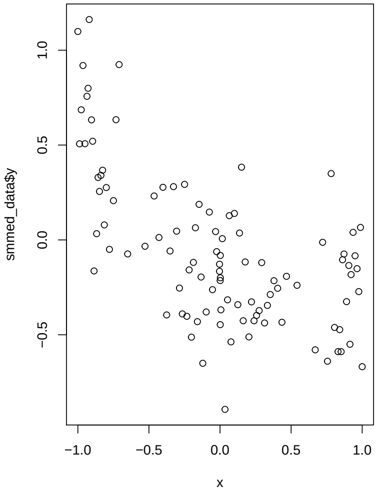
Consider the case where the true model is quadratic:
$\beta_T = (-0.2, -0.4, 0.4)$



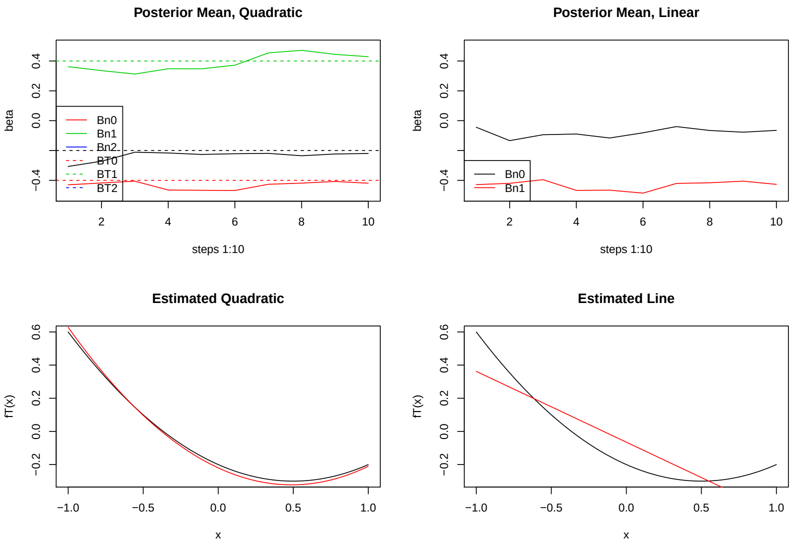
Sequential M-MED (using data)

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

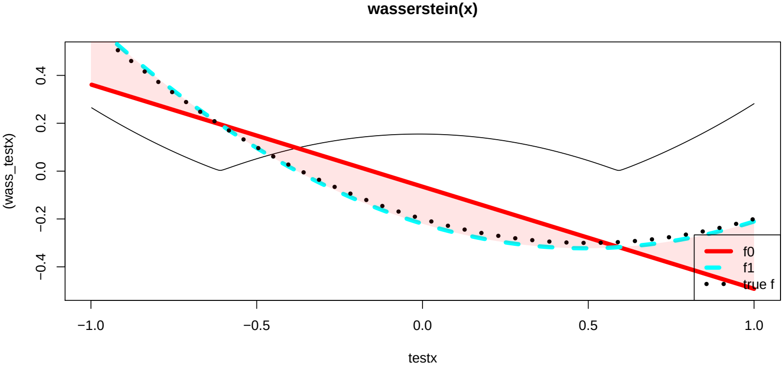




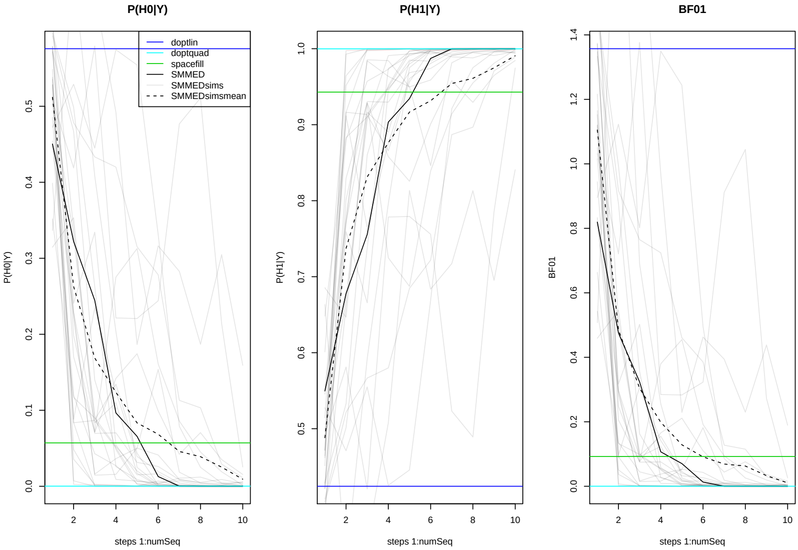
Linear and Quadratic Fits



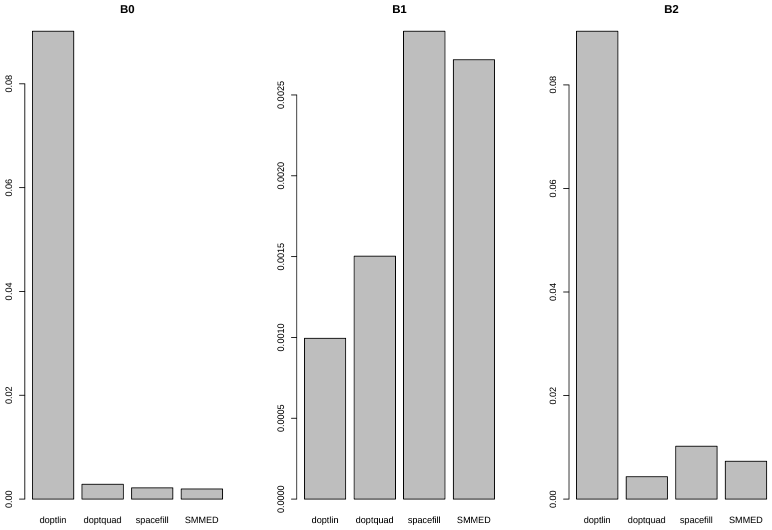
High Density Areas



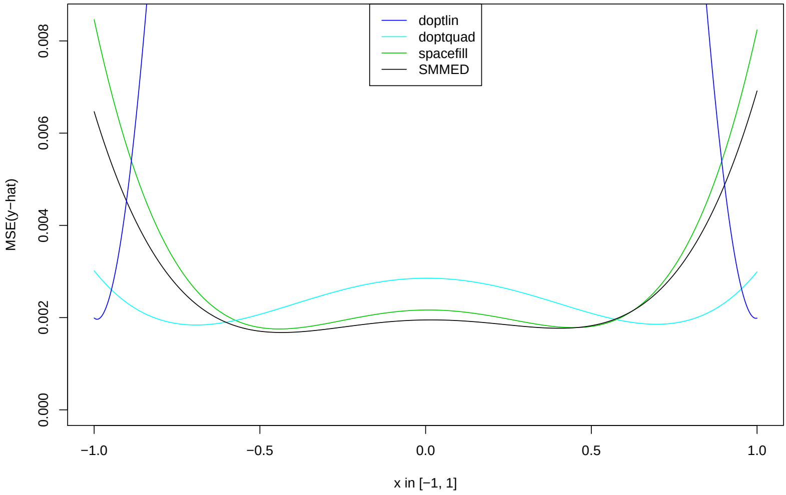
Hypothesis Testing



Parameter Estimation: MSE(Bn)



Prediction: MSE(y-hat)



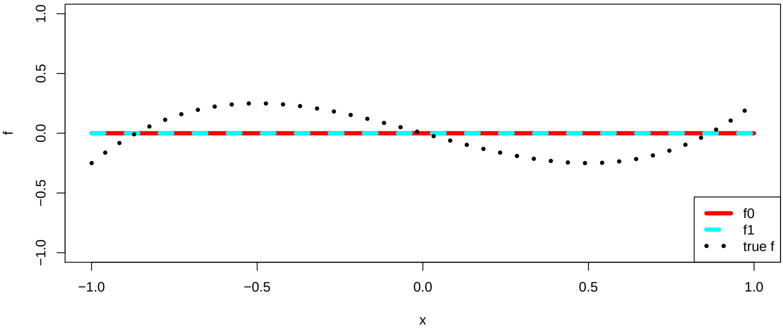
Case 2: Cubic

f0, f1, true f

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

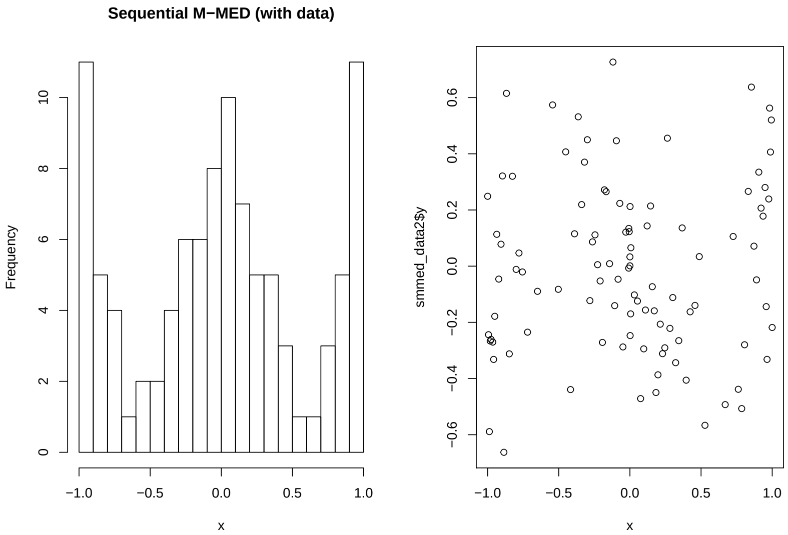
$$H_0 : \beta \sim N((0,0)^T, V_0)$$
$$H_1 : \beta \sim N((0,0,0)^T, V_0)$$

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

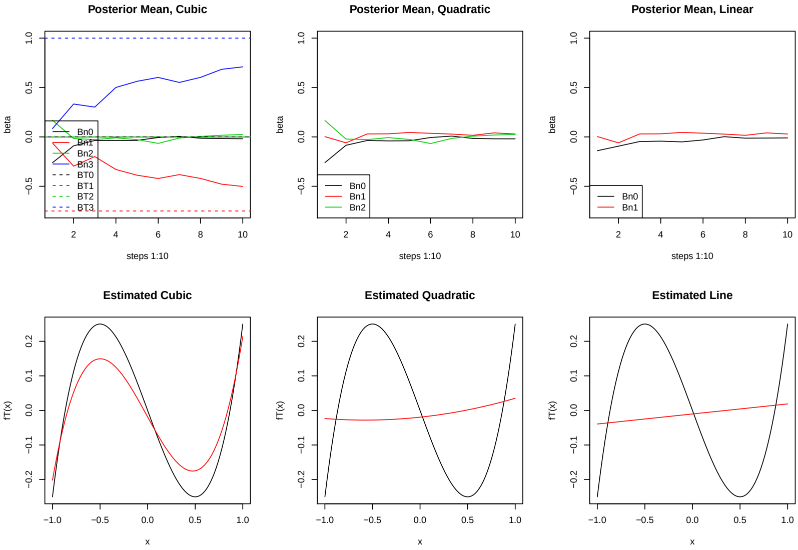


Sequential M-MED With Data

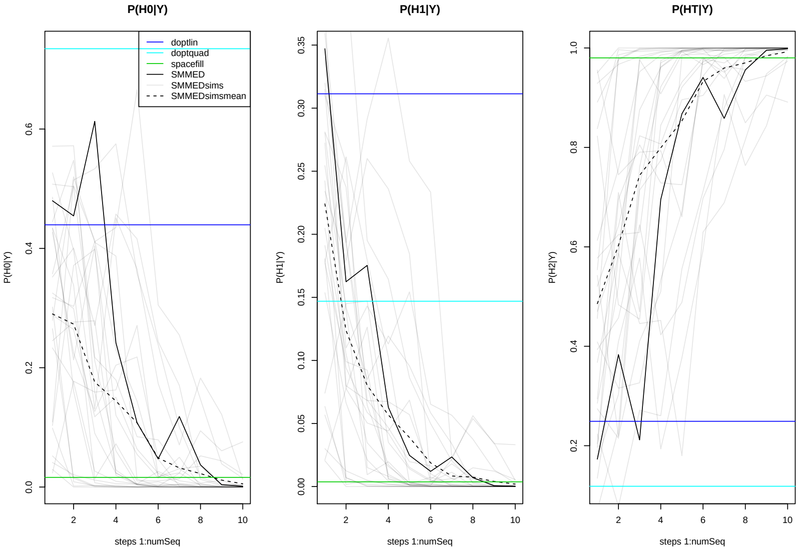
A sequence of 10 steps, generating 10 points in each step, resulting in 100 points:



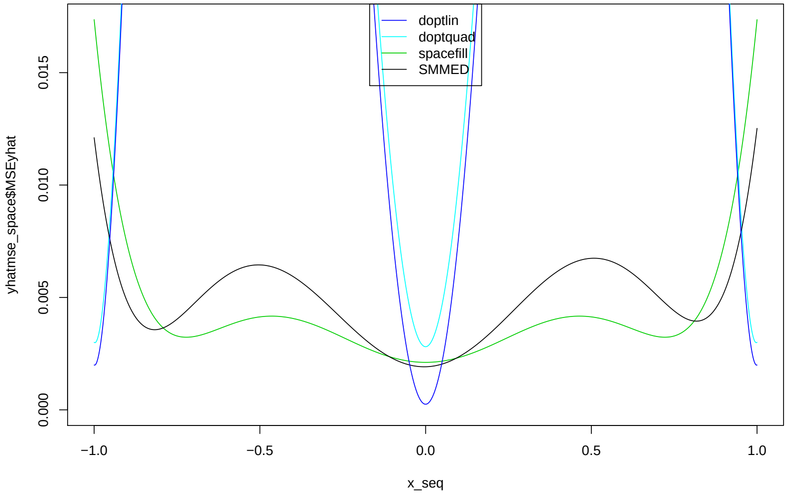
Linear, Quadratic, Cubic Fits



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

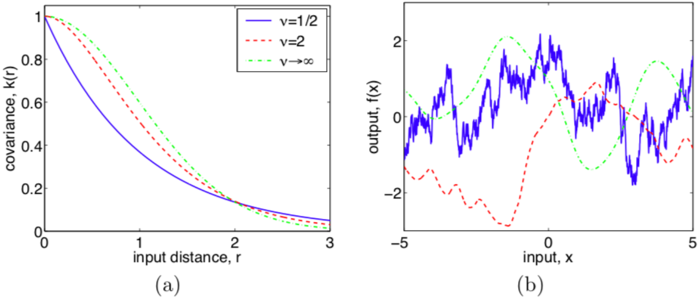


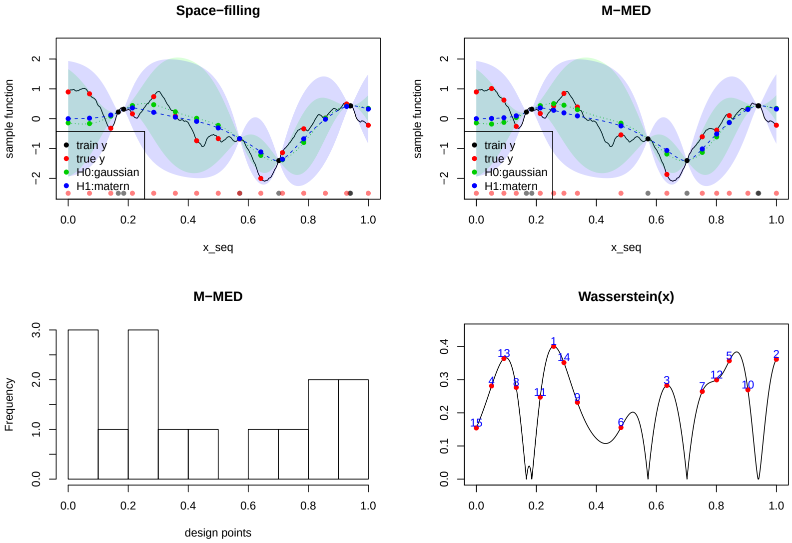
Figure 4.1: Panel (a): covariance functions, and (b): random functions drawn from Gaussian processes with Matérn covariance functions, eq. (4.14), for different values of ν, with $\ell = 1$. The sample functions on the right were obtained using a discretization of the x-axis of 2000 equally-spaced points.

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

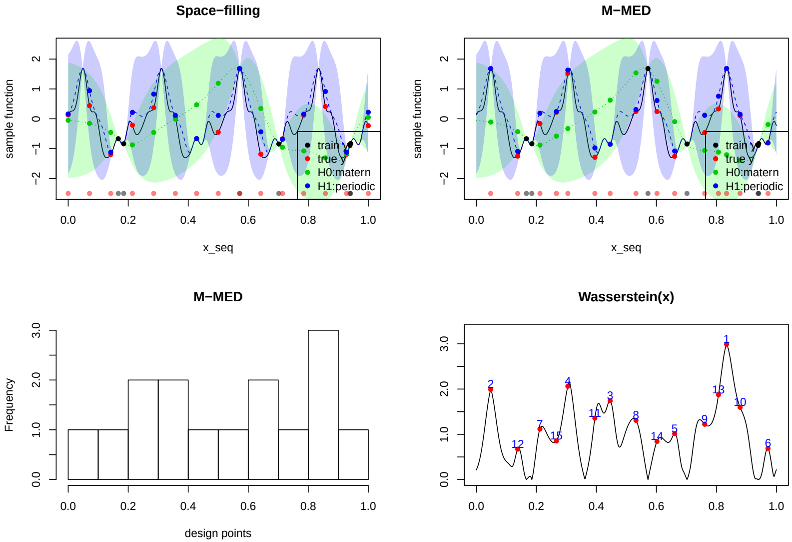
Applying M-MED to Gaussian Process Model Selection

- 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.

Sq Exponential vs. Matern (f generated from Matern)



Matern vs. Periodic (f generated from Periodic)



Sequential M-MED for GP

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)

- 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)

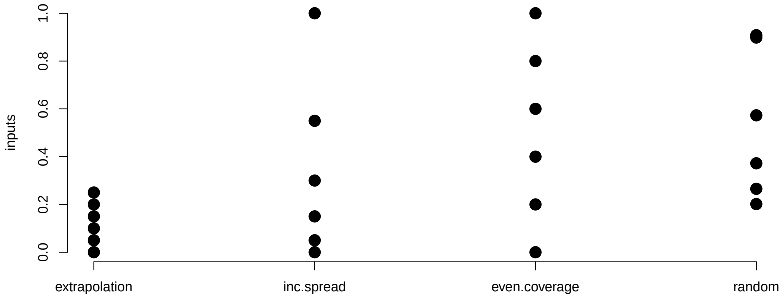
$$\frac{P(Y|H_1, X)\pi_1}{P(Y|H_0, X)\pi_0 + P(Y|H_1, X)\pi_1}$$

Metrics are used to compare MMED to space-filling and random designs.

Inputs

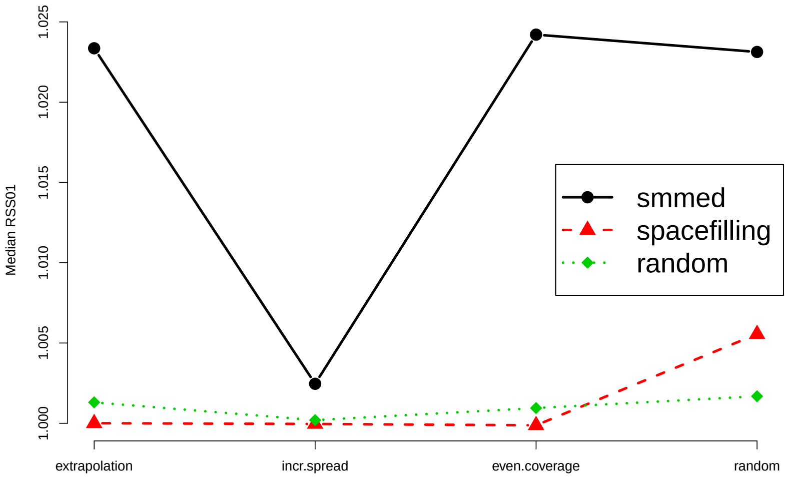
We consider 4 cases which use different input points for each of the 2 hypothesis tests and 3 designs:

- Extrapolation
- Increasing spread
- Even coverage
- Random



Gaussian vs. Matern: How do the different input cases compare?

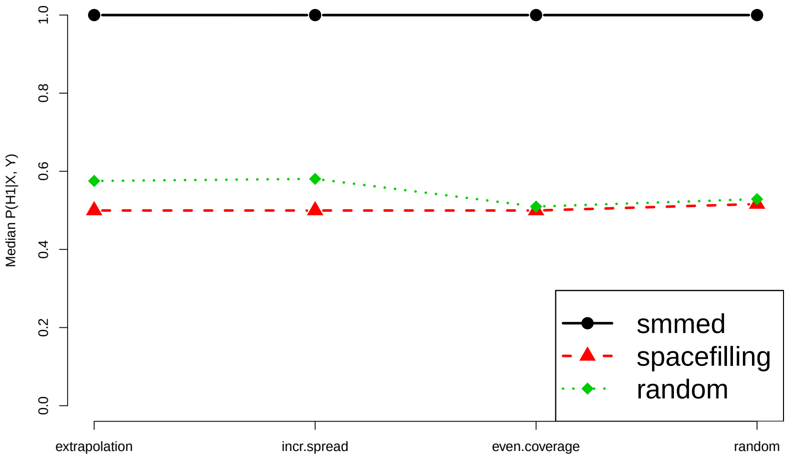
RSS Ratio (01)



1.025
1.020
1.015
1.010
1.005
1.000
Median RSS01
extrapolation
incr.spread
even.coverage
random
smmed
spacefilling
random

Posterior Probability of H1

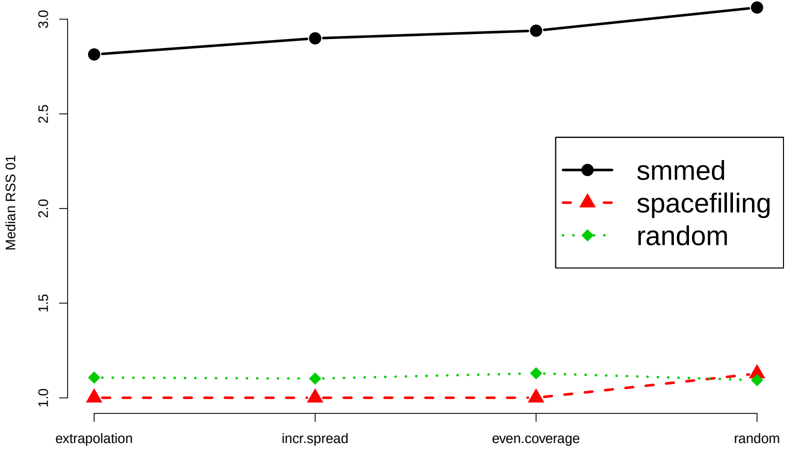
Compare the median $P(H_1|X, Y)$ for each case and design



Matern vs. Periodic: How do the different cases compare?

RSS Ratio (01)

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



Posterior Probability of H1

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