30

Sequential Minimum Energy Designs for Model Selection

*Kristyn Pantoja[1], **David Jones**[1], Rui Tuo[2], Huiyan Sang[1], Texas A&M University [1]Department of Statistics, [2]Department of Industrial Engineering*

T3: TEXAS A&M TRIADS FOR TRANSFORMATION
A President's Excellence Fund Initiative

Introduction

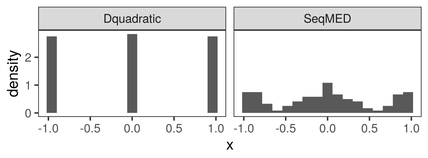
Statistical experimental designs specify what data should be collected to best achieve the experimental goals at hand.

Experiments typically prioritize one or more of the following four goals:

1. *Parameter estimation*
2. *Prediction*
3. *Model checking (is the model adequate?)*
4. *Model comparison / model selection / hypothesis testing*

Good designs for Goal 4 often perform poorly at the other three Goals, e.g., by only sampling very localized regions where the models to be compared differ most.

Contribution: We introduce a new design approach which prioritizes Goal 4 while also performing well at the other three Goals. The key is ensuring a space-filling property of the designs (right panel below).



Left: Example D-optimal design. **Right:** Corresponding SeqMED design we propose.

Applicable to wide ranging areas such as 3D printing, spatial modeling (e.g., soil quality), and dose-response analysis.

SeqMED

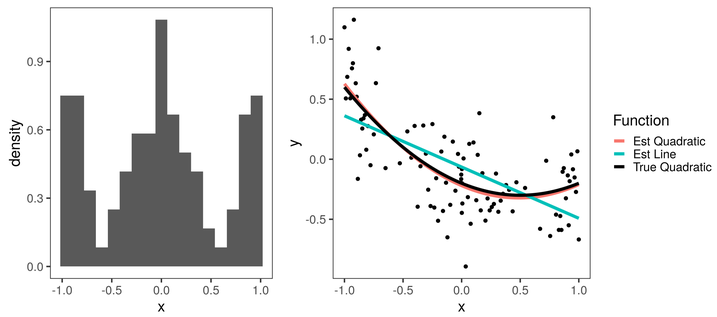
A space-filling design that emphasizes areas of the design space that can distinguish two competing models. Criterion to be optimized:

$$\sum_{i \neq j} \frac{q(x_i)q(x_j)}{d(x_i, x_j)}$$

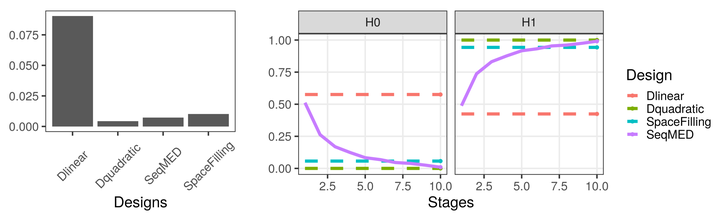
Where $q(\cdot)$ quantifies the separation of the two models at a given x and $d(\cdot,\cdot)$ is the distance between two design points. Analogy to potential energy of charged particles.

Model Discrimination

Take away: when one of hypothesized models is correct, SeqMED performs similarly to optimal design (the D-optimal design in this case).



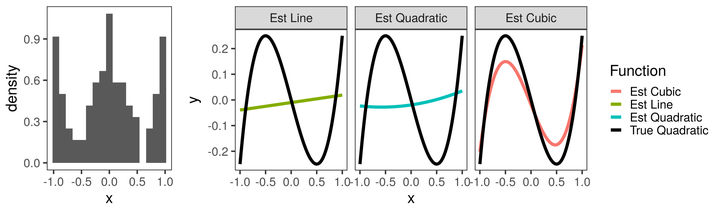
Left: SeqMED for comparison of a linear and quadratic regression model. **Right:** Fitted models and data generated from the true, quadratic function using SeqMED design inputs.



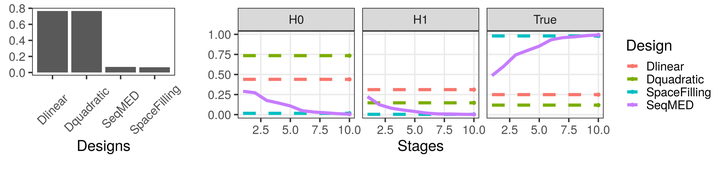
Left: Expected mean-squared error of estimates for the quadratic coefficient. **Right:** Expected posterior probabilities of the two hypotheses, where H_0 = linear model and H_1 = quadratic model.

Model Checking

Take away: when neither hypothesized model is correct, SeqMED can figure this out and fit the correct model, whereas the D-optimal design now fails.



Left: SeqMED when the true function is cubic. **Right:** Fit of linear, quadratic, and cubic models to the data generated from the true cubic function.

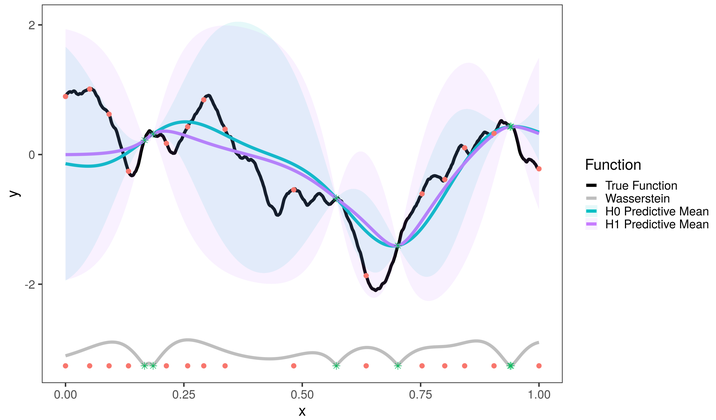


Left: Expected mean-squared error of cubic coefficient estimates. **Right:** Expected posterior probabilities for two hypotheses and the true cubic model.

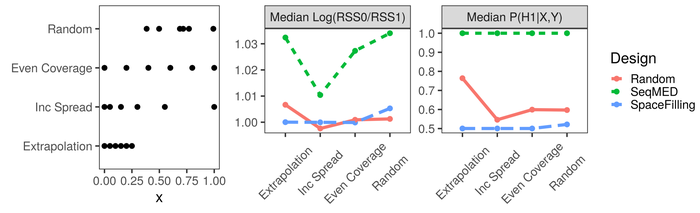
Application to Kernel Discrimination

A Gaussian process is a highly flexible way to model a curve or function, especially when the general shape is unknown and we only have information about the smoothness or about any periodic behavior present.

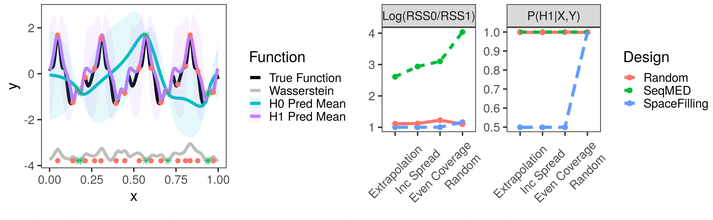
Smoothness and other structure is captured through the kernel function.



Above: A Gaussian Process using the squared exponential kernel (blue) and a Gaussian process using the Matern kernel (purple) fit to initial data (black stars) from the true function (black line), which is generated under the Matern kernel. SeqMED (red points) makes use of a measure of the pointwise separation between the two competing models (gray line).



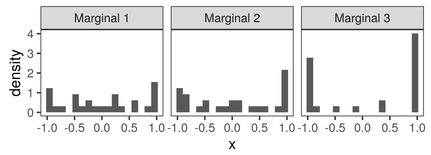
Left: For four different initial data sets, SeqMED outperforms two competing designs in both prediction and model selection. **Middle:** Prediction is evaluated based on the true model's prediction error (RSS) relative to the alternative model. **Right:** Model selection ability is measured by the expected posterior probability of the true model.



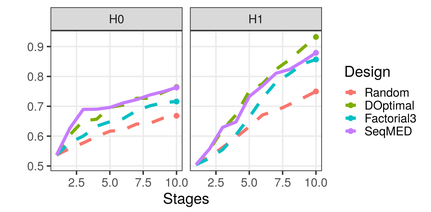
Left: Gaussian Process using Matern kernel (blue) and periodic kernel (purple) fit to initial, which is generated under the periodic kernel. **Middle and Right:** analogous to above. SeqMED again outperforms the two competing designs.

Application to Variable Selection

We often want to select the most important variables or predictors to include in a model.



Above: SeqMED for choosing whether a linear model should have 2 or 3 factors. **Take away:** designs for variables 1 and 2 are evenly spread, design for variable 3 prioritizes determining if variable is needed.



The expected posterior probabilities of the true hypothesis, where H_0 = two factors and H_1 = three factors. **Left:** H_0 true. **Right:** H_1 true.

Similar results were obtained when selecting variables for Gaussian Process kernels.

Conclusion

SeqMED is a space-filling design that automatically selects data to collect for distinguishing two models. Its space-filling aspect ensures information is available for the additional goals of model checking, estimation, and prediction. Competing designs typically do well at either model selection, or the additional goals, but not both.

Key Reference

V. Roshan Joseph, Tirthankar Dasgupta, Rui Tuo, and C. F. Jeff Wu. Sequential Exploration of Complex Surfaces Using Minimum Energy Designs.Technometrics, 57(1):64–74, January 2015. ISSN 0040-1706, 1537-2723. doi:10.1080/00401706.2014.881749.
URL:https://www.tandfonline.com/doi/full/10.1080/00401706.2014.881749.