Predictive Response Surface Models: To Reduce or Not to Reduce?

Byran Smucker
@ByranSmucker

Department of Statistics
Miami University, Oxford, OH

October 4, 2018

Fall Technical Conference 2018, West Palm Beach, FL

Joint work with colleagues

Maria Weese
@MFWeese
Miami University

David Edwards
Virginia Commonwealth
University

Overview

Outline

Setup

Imagine you are an experimenter attempting to optimize a process; **making good predictions** is your objective. You are ready to do a response surface experiment on your **four factors**. You do the experiment and analyze the data. But since you were unable to do a screening experiment prior to this one, you find that the trusty 'ol p-values indicate that **6 of the 14 terms** from your full second-order model are "not significant".

Taking a poll

Given this scenario, what would you say?

Taking a poll

Given this scenario, what would you say?

1. Retain the full second-order model
2. Reduce the model in a reasonable way
3. It doesn't matter; either way will predict with about the same quality

Motivation

Most of the RSM experiments we examined from the literature (83 out of 129) failed to mention a prior screening experiment.

Motivation

Most of the RSM experiments we examined from the literature (83 out of 129) failed to mention a prior screening experiment.

In such a setting, reasonable that not all of the factors are important, and even more likely that certain interactions/quadratic terms will be unimportant.

Motivation

Most of the RSM experiments we examined from the literature (83 out of 129) failed to mention a prior screening experiment.

In such a setting, reasonable that not all of the factors are important, and even more likely that certain interactions/quadratic terms will be unimportant.

But if the full quadratic model includes specious terms, could overfitting reduce the quality of prediction for the full model?

The literature is not much help

Ambiguous recommendations regarding RSM

1. Sampling of RSM/design textbooks give little explicit advice
2. Some (e.g. Peixoto, 1987, 1990, Nelder, 2000) argue philosophically in favor of retaining any terms marginal to retained second-order terms, though Peixoto (1987) makes concession if prediction is only purpose
3. Montgomery et al. (2005) suggests that retaining full model can result in inferior predictive performance

Any other literature that studies this question in RSM context and gives specific recommendations?

The questions before us

If your goal is to predict new observations using a second-order model:

1. should you use the full second-order model for response surface analysis or should you reduce it?
2. if you should reduce, what method should you use?

Outline

RSM Encapsulated

One way to describe RSM is as follows:

0. Screening. Reduce the number of factors from potentially many to hopefully just a few.
1. Initial improvement. If far from a process optimum, rapid improvement is likely possible using a first-order model to point in the direction of steepest ascent/descent.
2. Optimization. Once the experimenter nears a local optimum, an approximate optimum can be ascertained using a second-order model.

In this work, we focus on Step 2 when the goal is prediction based upon a quadratic model.

How we will investigate overfitting in RSM

Measure *out-of-sample* predictive performance.

Two data sources:

1. Sample of papers from the applied RSM literature.
2. An extensive simulation of a wide variety of response surfaces.

Empirical dataset: Sample of RSM studies from the literature

12 papers with published RSM experiments plus at least one validation run.

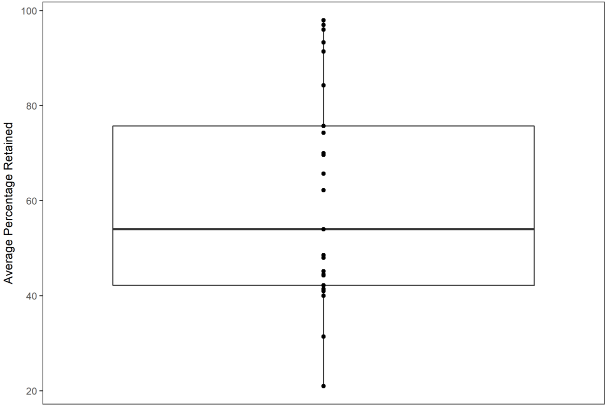
From a larger sample of RSM studies by Ockuly et al. (2017).

25 responses from these 12 papers.

54 validation runs from these 25 responses.

Compared quality of prediction for the 54 validation runs for each analysis method.

Empirical dataset: % terms retained



Simulation dataset (1/4)

Used an RSM simulation testbed from McDaniel and Ankenman (2000).

Provides some control over effect heredity, effect sparsity, and the bumpiness of the response surface.

We tested 20 designs (half CCD; half BBD) ranging from 3 to 7 factors.

For each design, 27 settings of the testbed, so 540 total scenarios, 1000 simulations each. Error variance $\sigma = 0.5$.

Simulation dataset (2/4)

Simulation assumed no previous screening: possible that all m factors important; possible that only a subset $m_r < m$ of factors are active.

Once m_r established, wide variety of true response surfaces tested, governed by testbed inputs $\mathbf{S}$, $\mathbf{T}$, and r:

- Some assumed just a proportion of terms of each type active (ME's, 2fi's, quadratic terms)
- Some assume all 2fi's and/or all quadratic terms
- Some assume a true response surface much more complex than quadratic

Simulation dataset (3/4)

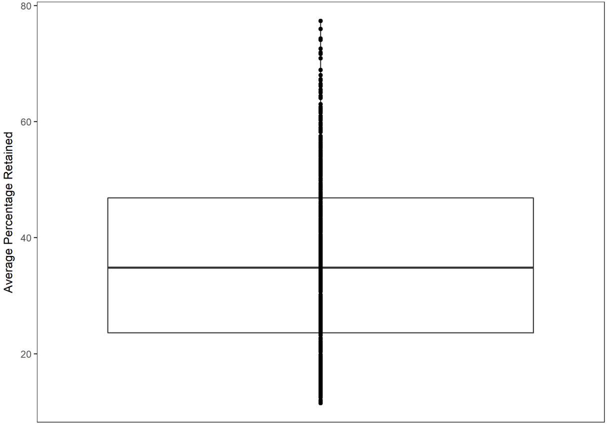
Testbed gives us response values from the design + response values from 9 validation points that we chose.

Validation points:

- 1 center run;
- 4 randomly selected non-center run design points;
- 4 randomly selected design points from the design space.

We compare the different analysis methods based on their predictive performance on the 9 validation points.

Simulation dataset (4/4)



Analysis methods to compare

- Full second-order model
- Reduced model based on p-values < 0.05
- Reduced model based on False Discovery Rate-adjusted p-values < 0.05
- Forward selection using AICc as the criterion
- Lasso
- Gauss-Lasso

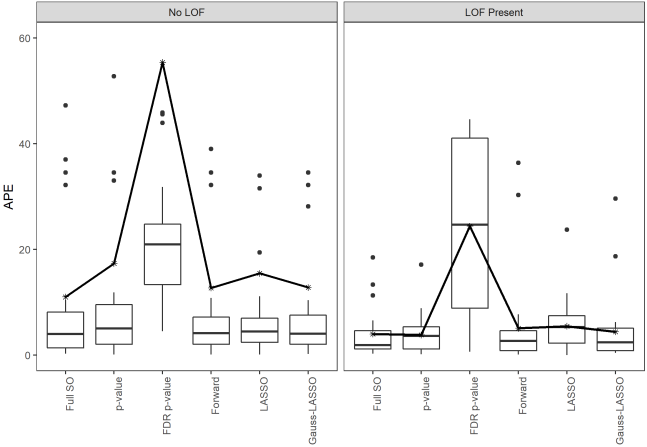
Outline

Overall results

Method	APE	Ranking	Terms Retained (%)
Full	7.8 (3.6)	2.5 (3)	100 (100)
p-value	11.1 (4.1)	3.4 (3.5)	53.8 (44.4)
FDR p-value	41.0 (22.9)	5.8 (6)	47.9 (37.1)
Forward selection	9.2 (3.6)	3.0 (3)	67.4 (66.7)
LASSO	10.8 (5.0)	3.4 (5)	72.6 (75.0)
Gauss-LASSO	8.9 (3.6)	2.8 (3)	62.5 (55.6)

Results from RSM studies

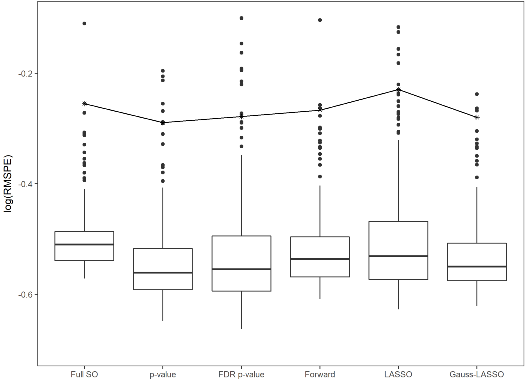
Extreme outliers have been omitted.



Outline

Overall results

Extreme outliers have been trimmed from this plot, and scenarios are omitted for which more than 10% of the simulations results exhibit lack-of-fit.



Clue from the literature

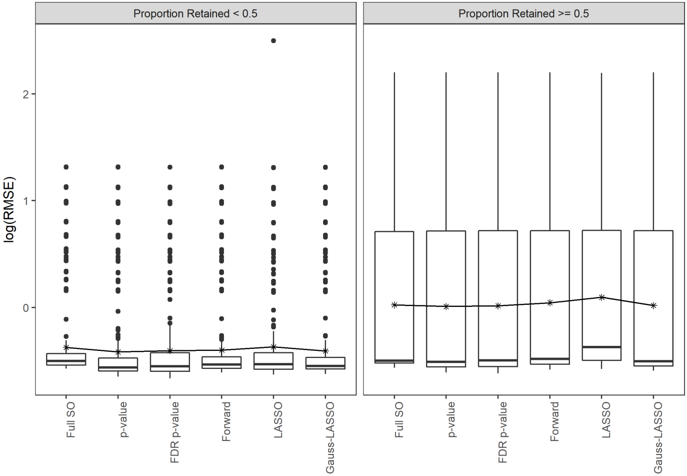
Roecker (1991) says how much you reduce has an impact.

In a general regression setting, she suggested using a reduced model for prediction if less than 50% of terms are retained.

So let's look at results broken up by the proportion of terms retained.

Also, as before we only look at scenarios for which less than 10% of the simulations showed lack-of-fit.

Results categorized by sparsity



Statistical Modeling of Simulation Results

To strengthen conclusions, we used our simulation results to perform formal modeling of log(RMSPE).

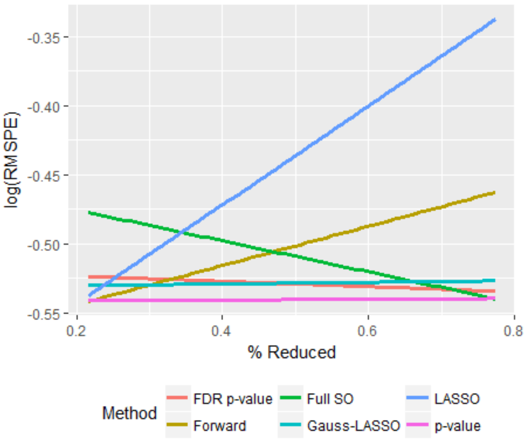
Included as factors: **S**, **T**, r, number of runs, number of factors, design type (CCD/BBD), analysis method, and % terms retained.

Used forward selection with AICc as the stopping criterion.

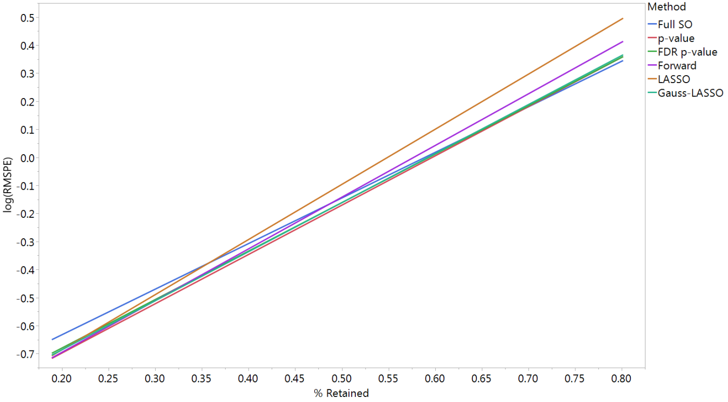
All main effects and two-factor interactions included as possible terms.

We validated these results by running another set of simulations with a larger error variance $\sigma = 1$.

% Retained × Method without scenarios with high-order polynomial terms



% Retained × Method Interaction



Outline

Recap

- We examined whether to reduce a full-second order RSM model or just use the full model
- We considered a variety of analysis methods to reduce the model
- We compared the predictive quality of the methods for a set of real RSM experiments from the literature
- We also made a comparison based upon an extensive simulation of response surfaces and RSM designs

Conclusions

1. *Full model works well unless there is an extreme reduction of the model*
2. For models with more than about 35%-50% terms retained, don't use the LASSO (and maybe not forward selection either), but little difference between the other methods.
3. Based on the small sample from the literature, I'd also be nervous using an FDR-adjustment to p-values to reduce.

Answer to Poll Question

Imagine you are an experimenter attempting to optimize a process; **making good predictions** is your objective. You are ready to do a response surface experiment on your **four factors**. You do the experiment and analyze the data. But since you were unable to do a screening experiment prior to this one, you find that the trusty 'ol p-values indicate that **6 of the 14 terms** from your full second-order model are "not significant". Given this scenario, which would you choose?

1. Retain the full second-order model
2. Reduce the model in a reasonable way
3. It doesn't matter; either way will predict with about the same quality

Answer to Poll Question

Imagine you are an experimenter attempting to optimize a process; **making good predictions** is your objective. You are ready to do a response surface experiment on your **four factors**. You do the experiment and analyze the data. But since you were unable to do a screening experiment prior to this one, you find that the trusty 'ol p-values indicate that **6 of the 14 terms** from your full second-order model are "not significant". Given this scenario, which would you choose?

1. Retain the full second-order model
2. Reduce the model in a reasonable way
3. It doesn't matter; either way will predict with about the same quality

Surprise answer: **all answers are correct**!

References

McDaniel, W. R. and Ankenman, B. E. (2000). A response surface test bed. *Quality and Reliability Engineering International*, 16(5):363–372.

Montgomery, D. C., Myers, R. H., Carter, W. H., and Vining, G. G. (2005). The hierarchy principle in designed industrial experiments. *Quality and reliability engineering international*, 21(2):197–201.

Nelder, J. (2000). Functional marginality and response-surface fitting. *Journal of Applied Statistics*, 27(1):109–112.

Ockuly, R. A., Weese, M. L., Smucker, B. J., Edwards, D. J., and Chang, L. (2017). Response surface experiments: A meta-analysis. *Chemometrics and Intelligent Laboratory Systems*, 164:64–75.

Peixoto, J. L. (1987). Hierarchical variable selection in polynomial regression models. *The American Statistician*, 41(4):311–313.

Peixoto, J. L. (1990). A property of well-formulated polynomial regression models. *The American Statistician*, 44(1):26–30.

Advertisement: Open Challenges in Industrial Statistics

Quality Engineering

Submit your problems within the industrial statistics community

Print Email Tweet Share

Are you a practitioner of industrial statistics?

Can you think of a problem within the industrial statistics community?

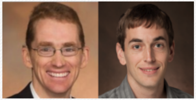
As part of our new dedicated column within the *Quality Engineering* journal we will be looking at the issues you face within industrial statistics. Solutions will then be published as regular articles with the wider aim of improving collaboration between the industry and academia.

Would you like the industry to help solve your problems in industrial statistics? Then send us your problems, and they may be included in our dedicated open access column!

Motivation:

Industrial statistics research motivated by real-world problems and based on real data is extremely valuable, but relatively rare.

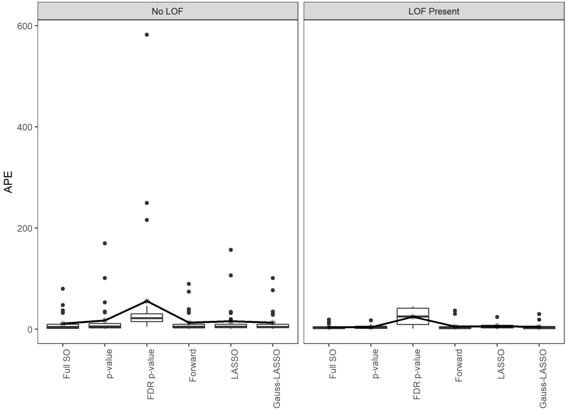
Column Editors



Column Editor:
Willis Jensen
W.L. Gore & Associates
wjensen@wlgore.com

Column Co-editor:
Byran Smucker
Miami University (OH)
smuckerb@miamioh.edu

Extra slide: Empirical results with extreme outliers



Extra slide: Overall simulation results with extreme outliers