Response Surface Models: To Reduce or Not to Reduce?

Byran J. Smucker[1], David J. Edwards[2], and Maria L. Weese[3]

[1]Department of Statistics, Miami University, 311 Upham Hall, 100 Bishop Circle, Oxford, OH 45056

[2]Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, 4124 Grace E. Harris Hall, 1015 Floyd Avenue, Richmond, VA 23284

[3]Department of Information Systems and Analytics, Miami University, 800 E High St. RM 3095, Oxford, OH 45056

Abstract

In classical response surface methodology, the optimization step uses a small number of important factors. However, in practice, experimenters sometimes fit a second-order model without previous experimentation. In this case, the true model is uncertain and the full model may overfit. Here, we use an extensive simulation to evaluate several analysis strategies in terms of their optimum locating ability, and use both simulation and published experiments to evaluate their general prediction facility. We consider traditional (reducing via p-values; forward selection), regularization (LASSO; Gauss-LASSO), and Bayesian analysis methods.

Keywords Confirmation runs; Model selection; Optimization; Prediction; Regularization methods; Second-order model

1 Introduction

In traditional Response Surface Methodology (RSM) (Box and Wilson 1951; Box and Draper 1987; Myers et al. 2016), the optimization step is ideally and typically undertaken with just a few factors. The hope is that if there were a large number of factors at the beginning of the RSM process, they have been screened down to a few during preliminary experiments. However, Ockuly et al. (2017)

found that out of 129 response surface studies sampled from the applied RSM literature, 83 of them included no mention of screening. Although sequential RSM is ideal, the "one-shot" approach to response surface screening and optimization is sometimes necessary; for instance, Lawson (2003) cites challenges such as work schedules, operator training, fixed deadlines, prototype development, and pilot plants as examples when traditional sequential RSM may be impractical or overly time consuming.

If there has been no *a priori* screening experiments and there are as many as four or more factors being studied, many of the terms in the full quadratic model are likely to be unimportant (see Ockuly et al. 2017). This leads to an interesting and important question: for a second-order RSM experiment that includes apparently inactive terms, should the analyst retain the full second-order fitted model when using the model for optimization and/or prediction, or should the model be reduced? And if you should select a model, what analysis method should be used?

The RSM literature is ambiguous and mixed in its advice. Several authors (Peixoto 1987, 1990; Nelder 2000) argue strongly about how empirical models might be reduced, in principle. They advocate for "well-formed polynomials", for which any second-order term includes all of its marginal terms. Thus, these authors implicitly allow for model reduction, but only of a particular type. We do note that Peixoto (1987) does admit that if the goal of the study is strictly prediction and transforming the factors is not of interest, then other reductions may be allowable. Other authors, such as Chipman (1996), allow for the possibility of fitting models that do not exhibit strong or even weak heredity. Montgomery et al. (2005) suggests that in some cases a model will possess inferior predictive performance if strong effect heredity is enforced. None of these articles provide an empirical study of RSM models or investigate the quality of various model selection methods that include an assessment of prediction quality on out-of-sample data. Furthermore, a survey of several experimental design and RSM textbooks, including Khuri and Cornell (1996), Dean et al. (2017), Wu and Hamada (2011), and Montgomery (2017), don't address the issue directly but do provide examples in which they reduce the model. In RSM textbooks, the development of classical canonical analysis of the second-order model requires that the full model be retained. However, even so, Myers et al. (2016) provides examples and/or exercises in which models are reduced. On the other hand, texts like Box and Draper (2007) and Mead et al. (2012) have a simple-to-complex

model-building perspective, rendering moot the idea of starting with a full model and reducing it. In our review, there does not appear to be explicit recommendations.

In this article, we focus on second-order response surface models for which the goal is prediction generally and optimization specifically. Given an optimization objective, the right metric to test the quality of competing models is the accuracy with which they identify the optimum point. One may also be interested in the quality of the methods to predict more generally across the design space, though if prior experimentation has not been performed, lack-of-fit is a concern. Specifically, in "one-shot RSM" experiments, the "region of experimentation" and the "region of operability" often coincide. In such cases, fitting a second-order polynomial model over a large region of operability will potentially lead to a poor fit, particularly in higher dimensions. Thus, in the cases in which screening and line searches have not been performed, we expect models to optimize and predict more poorly than they would otherwise.

If certain model-selection strategies are better than others, they can demonstrate it with more accurate identification of the optimum point, as well as superior out-of-sample prediction performance. Roecker (1991) performed a simulation study on out-of-sample predictive performance for various model selection strategies in general regression settings. The study concluded that the full model will typically be better for prediction unless more than half of the terms in the model are unnecessary. Hawkins (2004) also discusses the problem of overfitting, but does not perform any sort of large-scale empirical or simulation analysis. In the context of designed experiments, Jensen (2016) provides a recent overview of the confirmation run literature. In accordance with Jensen's usage, we use "confirmation run" to refer to any out-of-sample run used to confirm a model (in this case a response surface model).

Here, we consider a sample of response surface studies from the applied RSM literature that include both second-order experimental data and at least one confirmation run. We test several model-selection methods in terms of their out-of-sample prediction properties. In addition to the empirical analysis of existing RSM experiments, we also simulate from a variety of response surface models and compare the same analysis methods both in terms of their ability to locate the true optimum as well as the quality of their predictions.

In Section 2 we provide some background on RSM, the analysis methods we use, and how we

measure model quality. In Section 3, we discuss the sample of RSM papers and empirically evaluate the various analysis methods using these data. In Section 4, we make a similar comparison, except with simulated data. We discuss the results and make several conclusions in Section 5.

2 Background and Setting

Response Surface Methodology (RSM) (Box and Wilson 1951; Myers et al. 2016) refers to a set of statistical tools used when an experimenter would like to optimize a process for which the inputs can be controlled. One way to describe the process is as follows:

0. Screening. If necessary, reduce the number of factors from many to hopefully just a few, via experimentation.

1. Initial improvement. If far from a process optimum, rapid improvement is likely possible using a simple design and first-order model, which will point in the direction of steepest ascent/descent. This initial improvement process is iterative, and may include multiple designs and line searches. These searches will be terminated when curvature is detected, indicating that a more complicated model should be used to capture the complexity of the response surface near a local optimum.

2. Optimization. Once the experimenter nears a local optimum, an approximate optimum can be ascertained using a second-order model. Throughout this article, and without loss of generality, we assume that the goal of optimization is to maximize the response variable within the experimental region, R.

Note that the above description is not exhaustive, but simply a high level view of a typical, classical RSM process. For Step 2, we represent the standard second-order model for observation Y as

$$Y = \mathbf{f}'(X_1, \ldots, X_m)\boldsymbol{\beta} + \epsilon, \tag{1}$$

where

$$\mathbf{f}'(X_1, \ldots, X_m) = (1, X_1, \ldots, X_m, X_1^2, \ldots, X_m^2, X_1X_2, \ldots, X_{m-1}X_m),$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_m, \beta_1^2, \ldots, \beta_m^2, \beta_{12}, \ldots, \beta_{m-1,m})',$$

and we include standard independence and normality assumptions on ϵ. Note that the β_i^2 notation does not indicate a nonlinear regression parameter; it only specifies linear parameters associated with squared predictors. We use this notation because in Section 4 we need to accommodate simulated models from higher-order polynomials. In this classical setting, after fitting model (1) to the data, the estimated optimum operating conditions are obtained as $[X_1^*, \ldots, X_m^*] = \underset{[X_1,\ldots,X_m]\in R}{\arg\max}\ \mathbf{f}'(X_1, \ldots, X_m)\hat{\boldsymbol{\beta}}$ where $\hat{\boldsymbol{\beta}}$ is an estimate of $\boldsymbol{\beta}$.

As mentioned in Section 1, it is likely that in many realizations of RSM, experimenters jump directly to Step 2. If so, then although the full second-order model is given in (1), it could be that some subset of the full model provides a better empirical approximation to the response surface in the sense of its ability to estimate the optimum design point, or in terms of its predictive accuracy. If optimization is the primary objective of the model, then the goal of the experimenter should be to discover the method that best recovers the optimum.

2.1 Response Surface Analysis Methods

In this article we will evaluate both response surface studies from the applied RSM literature and simulated response surfaces using a variety of analysis methods. We will use the analysis methods to construct a predictive model from the RSM experimental data and optimize as well as make predictions on out-of-sample runs. We consider the following analysis methods:

1. The full second-order model. That is, fitting the model from (1).

2. A reduced second-order model, based upon p-values and $\alpha = 0.05$. That is, any term with a p-value less than α is retained, and the model is refit to include only these terms.

3. A reduced second-order model using p-values adjusted using the False Discovery Rate (FDR) (Benjamini and Hochberg 1995), with $\alpha = 0.05$.

4. Forward selection using AICc as the selection criterion.

5. LASSO-regularized regression (Tibshirani 1996), described below.

6. The Gauss-LASSO (Rigollet and Tsybakov 2011), described below.

7. Bayesian optimization using the posterior predictive distribution (Peterson 2004). Note that for the noninformative priors we use for this method, predictions and standard errors are equivalent to those made by the full second-order model. Thus, any results throughout the article that evaluate prediction quality will not explicitly include the Bayesian method. However, the estimated optimum is typically different and so the Bayesian approach is highlighted when evaluating the optimization of response surfaces. This procedure is described in more detail in Section 2.2.

We note that for the p-value-based methods, using $\alpha = 0.05$ to choose "significant" terms has been criticized extensively on various grounds (see Wasserstein et al. 2016, 2019, and accompanying articles). Here, we will simply consider the p-value a crude measure of the evidence that an effect is important and study its performance as a model-selection method in the RSM context.

Regularized regression methods, such as the Dantzig selector (Candes and Tao 2007) and LASSO (Tibshirani 1996), have been successfully used to analyze supersaturated designs (Draguljić et al. 2014; Weese et al. 2015) as well as Definitive Screening Designs (Errore et al. 2017; Weese et al. 2018). Because Central Composite designs and Box-Behenken designs have adequate degrees of freedom for estimation of the full second order model we use the LASSO for variable selection here. LASSO estimates for the parameters in Eq. (1) are the values that minimize:

$$\sum_{i=1}^{n}(Y_i - \mathbf{f}'(X_{1i},\ldots,X_{mi})\boldsymbol{\beta})^2 + \lambda\sum_{i'=1}^{p}|\beta_{i'}|, \tag{2}$$

where $\mathbf{f}$ and $\boldsymbol{\beta}$ are as defined in Eq. (1), $p = 1 + 2m + \binom{m}{2}$, and $\beta_{i'}$ is the i'^{th} element of $\boldsymbol{\beta}$. The choice of λ in Eq. (2) is often made by cross-validation; however, this appears to be ineffective in the analysis of small, structured designs (Yuan et al. 2007; Draguljić et al. 2014). Instead, we use the LASSO coefficient estimates at the value of λ that minimizes $AICc = AIC + \frac{2p(p+1)}{n-p-1}$ where

$$AIC = n\log\left(\frac{(Y - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda})'(Y - \mathbf{X}\hat{\boldsymbol{\beta}}_L^{\lambda})}{n}\right) + 2p, \tag{3}$$

$\mathbf{X}$ is the expanded design matrix with the same form as $\mathbf{f}'(X_1,\ldots,X_m)$ and $\hat{\boldsymbol{\beta}}_L^{\lambda}$ are the LASSO

estimates for a particular value of λ.

In the second version of LASSO, referred to as "*Gauss*-LASSO" (Rigollet and Tsybakov 2011), the LASSO is first used to identify the important predictors (i.e. those with $\hat{\boldsymbol{\beta}}_L^{\lambda} \neq 0$) for each λ and least squares estimates are obtained for the reduced set of variables. The selection of λ is again made via $AICc$ where the LASSO estimates in (3) are replaced by least squares estimates. This two stage procedure reduces the bias of the LASSO coefficients while retaining this regularization method's benefit of reduced variance.

Note that there was no attempt to maintain effect heredity in any of these modeling strategies. Instead, we focus purely on predictive models.

2.2 Response Surface Optimization

The analysis of second-order response surfaces typically includes the determination of the stationary point and its associated canonical analysis (e.g. Myers et al. 2016). However, it is not uncommon for the experimenter to find either the location or nature of the stationary point to be of little value. For instance, the stationary point may be a saddle point when a maximum is desired or, alternatively, the stationary point may lie outside the experimental region. In such cases, numerical optimization approaches are useful for estimating $[X_1^*, \ldots, X_m^*]$, the optimum point in the factor space. Optimization of the full second-order fitted response surface model is overwhelmingly emphasized in the RSM literature. However, for a given frequentist method considered in this article, we first choose a model and then find the point which optimizes the chosen model.

As a Bayesian alternative, Peterson (2004) and Rajagopal and Del Castillo (2005) recommend a posterior predictive approach for response surface optimization. As pointed out by Del Castillo (2007), a clear advantage of the Bayesian approach over optimization based on fitted models is that uncertainty in model parameters are directly incorporated into the analysis. Suppose we wish to predict a new observation, $\tilde{Y}$, at a new combination of factor levels, $\tilde{X}_1, \ldots, \tilde{X}_m$. Then, the posterior predictive distribution is

$$p(\tilde{Y}|Y, \tilde{X}_1, \ldots, \tilde{X}_m) = \int_{\boldsymbol{\theta}} p(\tilde{Y}|\boldsymbol{\theta}, \tilde{X}_1, \ldots, \tilde{X}_m) p(\boldsymbol{\theta}|Y) d\boldsymbol{\theta},$$

where, for the usual linear model, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma)$. Peterson (2004) assigns a noninformative prior distribution for the model parameters; that is, $p(\boldsymbol{\beta}, \sigma) \propto 1/\sigma$. Under this prior distribution, it is shown that the posterior predictive density follows a non-central t-distribution with $n - p$ degrees of freedom. Specifically,

$$p(\tilde{Y}|Y, \tilde{X}_1, \ldots, \tilde{X}_m) \sim t_{n-p}\left(\mathbf{f}'(\tilde{X}_1, \ldots, \tilde{X}_m)\hat{\boldsymbol{\beta}}, s^2\left(1 + \mathbf{f}'(\tilde{X}_1, \ldots, \tilde{X}_m)\mathbf{X}'\mathbf{X}\mathbf{f}(\tilde{X}_1, \ldots, \tilde{X}_m)\right)\right).$$

Del Castillo (2007) notes that while Bayesian estimates and standard errors coincide with frequentist results under noninformative priors, there is "considerable value for predictive inference". Although other choices of prior distribution are common (e.g. conjugate prior), we only consider the noninformative case.

Given that the form of the posterior predictive distribution is known, it is straightforward to compute the probability that a future response, at a new treatment combination, will conform to some desirable quality level. Suppose that the experimenter desires a future response to be in some subset, R, of the response space. Then, the probability of conformance, $P(\tilde{Y} \in R|Y, \tilde{X}_1, \ldots, \tilde{X}_m)$, is computed as

$$P(\tilde{Y} \in R|Y, \tilde{X}_1, \ldots, \tilde{X}_m) = \int_R p(\tilde{Y}|Y, \tilde{X}_1, \ldots, \tilde{X}_m)d\tilde{Y}.$$

Peterson (2004) refers to a conformance probability as a Bayesian reliability. Response surface optimization is conducted by maximizing Bayesian reliabilities over the experimental region; this can be accomplished via a grid search or general optimization techniques.

In this article, we compare the frequentist and Bayesian approaches for estimating the true optimum within the experimental region.

2.3 Assessing Prediction Quality

We measured the quality of the analysis methods in terms of their predictive performance on out-of-sample points. Specifically:

1. Root mean squared prediction error (RMSPE), to evaluate the simulation results (Section 4). If a method provides m out-of-sample predictions $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_m$ for true out-of-sample values

$Y_1, Y_2, \ldots, Y_m$, we measure a rough "average" error as:

$$RMSPE = \sqrt{\frac{\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2}{m}}$$

2. Absolute percentage error (APE), to evaluate the empirical results (Section 3). For an out-of-sample prediction $\hat{Y}_i$ and a true out-of-sample value Y_i, a scale-free measure of a prediction's quality is:

$$APE = \frac{|Y_i - \hat{Y}_i|}{Y_i} \times 100\%$$

3. Ranks, to evaluate the empirical results (Section 3). If we have t different predictions, $\hat{Y}_{1i}, \hat{Y}_{2i}, \ldots, \hat{Y}_{ti}$, of an out-of-sample value Y_i, we can assign 1 to the best prediction, 2 to the next-best prediction, ..., and t to the worst prediction.

3 Empirical Evaluation from the RSM Literature

We first consider a sample of RSM studies from the literature which report both the original second-order experiment as well as out-of-sample confirmation runs.

3.1 Description of Dataset

Ockuly et al. (2017) used a stratified random sample of papers from the applied RSM literature to estimate various characteristics of RSM experiments and analysis. From this sample, we determined which papers reported validation run results. Since each paper reported the original experiment and results, we reanalyzed these data using the methods described in Section 2.1 and evaluated the quality of predictions using the measures in Section 2.3. In all, we used response surface experiment results from 12 papers, which included a total of 25 responses and 54 validation runs. Of the 25 total responses, 8 of them involved full second-order models that exhibited lack-of-fit at the 0.05 level. Out of the 12 papers, 11 used RSM to ultimately find an optimum experimental value. In 10 of those cases, the confirmation run was performed at suggested optimum values (possibly based upon simultaneous optimization, in cases of multiple response) and one performed confirmation

runs for "randomly chosen" conditions in the experimental region. The only paper that does not mention finding an optimum reports a single confirmation run with no discussion regarding how it was chosen.

This is a modest sample, but provides an "in the wild" reality check that may be lost in the complications of the simulation in Section 4.

3.2 Comparison of Analysis Methods using RSM Studies

Since the response variables used in the sampled RSM studies may be on drastically different scales, we used absolute percent error (APE) for each prediction as our primary metric of comparison. Additionally, for each validation run we rank the analysis methods with respect to RMSPE and compared the distribution of their rankings.

Roecker (1991) identifies the size of the true model as a key factor in determining whether the model should be reduced, in the context of general regression. To construct a proxy for the size of the true model, for each of the 25 responses, we calculated the average percent terms retained over each of the methods (excluding the full model, which always retains 100% of the terms), and call this "% Retained" throughout. In Table 1, we provide this percentage broken out by each method, and in the left panel of Figure 1 we show the distribution of "% Retained" across the 25 responses in our sample. Most of the responses are at or greater than 50%, and there is only one response below 30.

Table 1: Mean (median) absolute percent error (APE), ranking in terms of RMSPE, and percent terms retained for each method compared over all 25 responses.

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)

Table 1 demonstrates that in this set of real RSM studies, the full second-order model fares well in comparison with the other methods. It has a smaller mean APE than its competitors, and it ranks better, on average, as well. The models based upon an FDR p-value adjustment clearly fare the worst, and there is also the suggestion that the p-value method and the LASSO are also worse

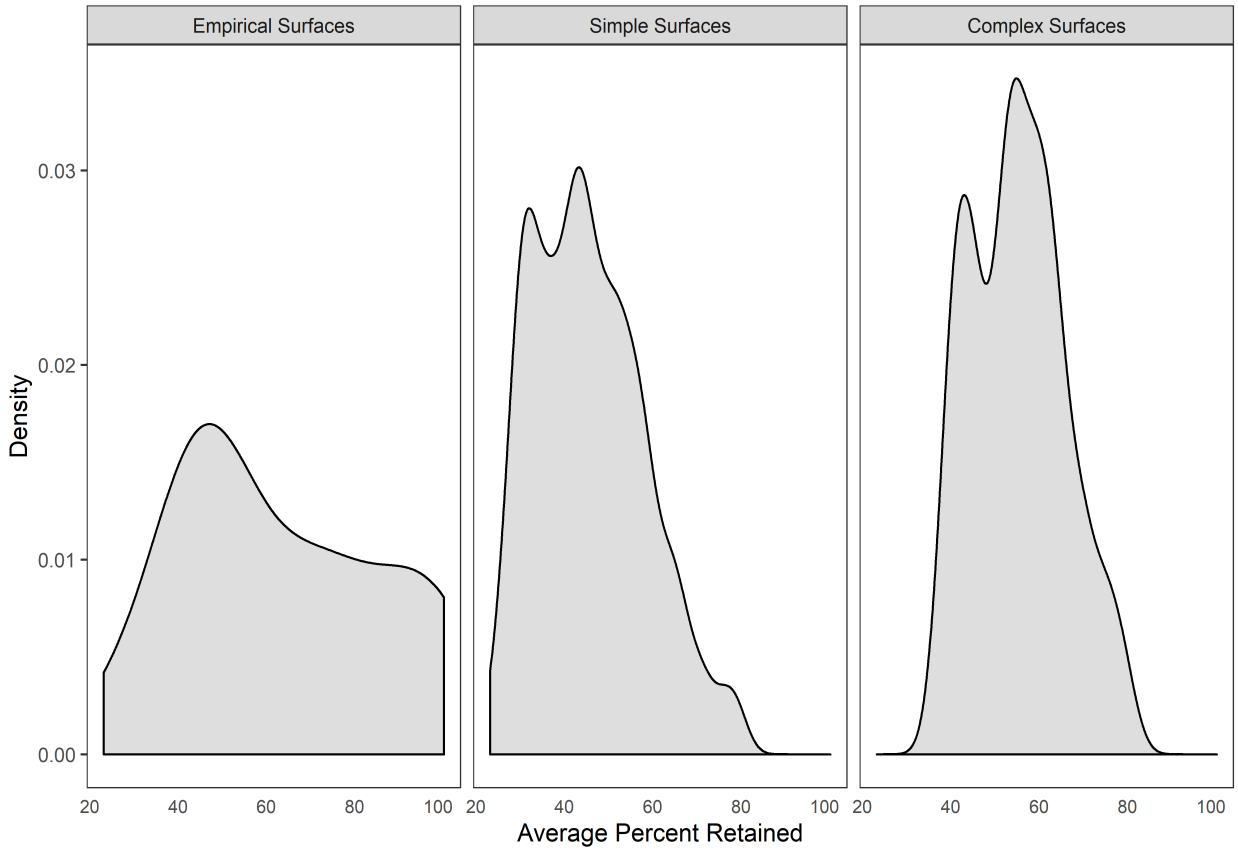

Figure 1: The left most panel contains the average model size for the 25 responses from the literature in terms of percent of full model, averaged over all analysis methods (except the full second-order model method). The next two panels are from simulation results; see Section 4 for details

than the other methods.

Since 8 of the 25 fitted full second-order models exhibited lack-of-fit, in Figure 2 we present graphical comparisons of the six methods, for those responses that showed 0.05-level evidence of lack-of-fit and those that didn't. To highlight the comparisons between the methods, we've omitted the extreme outliers from the figures, but the mean line on Figure 2 shows how skewed the distributions are, particularly for those responses without a lack-of-fit. Clearly, the FDR p-value method is the worst, and in terms of median APE there is little difference between the responses with and without lack-of-fit in terms of prediction quality. Figure 3 suggests that, with the exception of FDR p-value, there is not a clear difference between any of the methods when the cases are categorized "less than 50% retained" and "greater than (or equal to) 50% retained".

Interpreting plots and statistics is perilous with so little data. Ultimately, the results from the

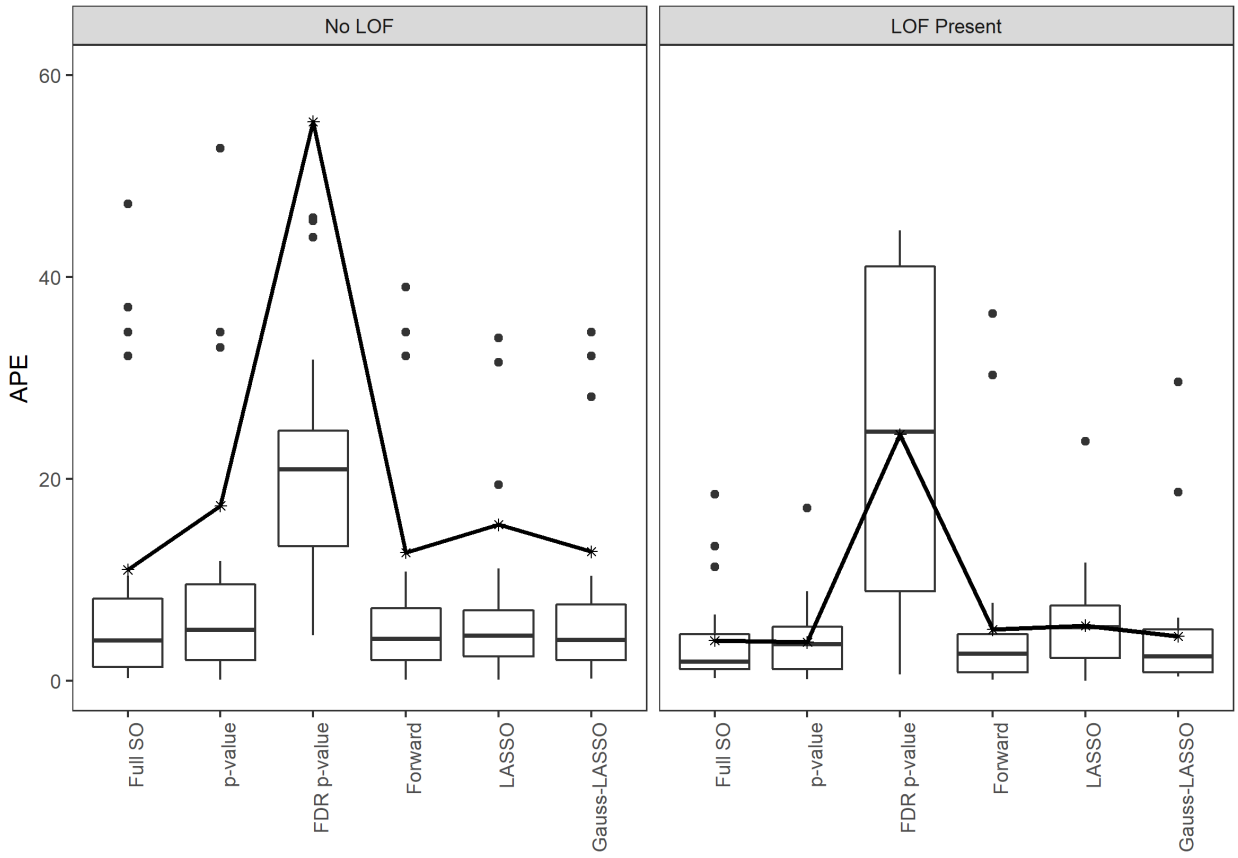

Figure 2: Absolute percentage errors for the six methods, over the 54 validation runs from the 25 responses in the literature, for models that exhibited lack-of-fit and for those that did not. The line represents the mean of each group. Extreme outliers have been omitted; a full plot can be found in Appendix A.

empirical studies provide some initial evidence that the full second-order model is a reasonable default for the analysis of RSM studies, and that the FDR p-value method should be avoided. However, the lack of data—particularly for responses which are reduced to less than 50% of the full model—and our inability to check these methods for quality of optimization requires us to use simulation to investigate these analysis methods more thoroughly.

4 Simulation to Evaluate Analysis Methods

In this section, we simulate response surfaces and analyze them via the methods described in Section 2.1, to provide additional insight regarding the analysis of RSM studies under a range of realistic conditions.

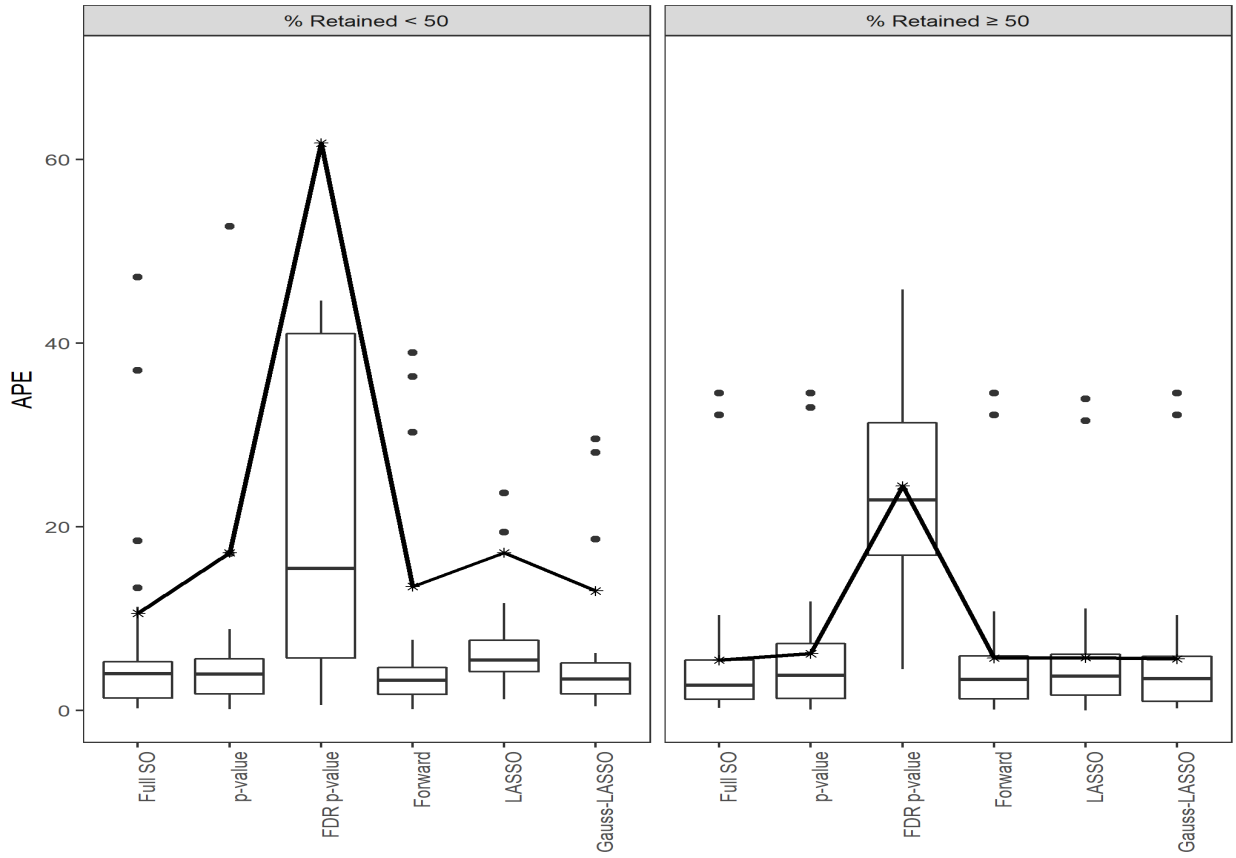

Figure 3: Absolute percentage errors for the six methods, over the 54 confirmation runs from the 25 responses in the literature, for responses that had less than 50% of the model terms retained and for those that had more than 50%. The line represents the mean of each group. Extreme outliers have been omitted; a full plot can be found in Appendix A.

4.1 Description of Simulation

Though simulation of general regression data is common and relatively straightforward, simulating from response surface models in a way that induces reasonable topologies is challenging. Our goal is to produce a wide variety of response surfaces that bear some similarity to the response surfaces observed in our real examples in Section 3 and in the larger sample from Ockuly et al. (2017).

4.1.1 Designs from which we Simulated

We used both central composite designs (CCDs) and Box-Behnken designs (BBDs) in our simulations, varying the number of factors from $m = 3$ to $m = 7$. For each design type with m factors, we considered designs with 4 center points. For the CCDs, we used axial distances of 1

(i.e. face-centered) and $\sqrt{m}$. Table 3 in Appendix B provides more details about the designs.

4.1.2 Generation of Response Surfaces

To simulate response surfaces, we used the testbed approach described in McDaniel and Ankenman (2000), which allows a measure of control over aspects of the response surface such as effect sparsity, effect heredity, bumpiness, and flatness. The testbed works by specifying three main sets of parameters:

1. $\mathbf{S}$, a 6×2 matrix (with elements $s_{\ell 1}$ and $s_{\ell 2}$, such that $\ell = 1, 2, \ldots, 6$) which specifies the probability of pure polynomial main effects of order 1-6 appearing in the model, conditional on the appearance or non-appearance of the main effect of one order lower. More specifically, $s_{11} = P(\beta_i \neq 0)$, $s_{12} = 0$, $s_{\ell 1} = P(\beta_i^{\ell} \neq 0 \,|\, \beta_i^{\ell-1} \neq 0)$, $2 \leq \ell \leq 6$, and $s_{\ell 2} = P(\beta_i^{\ell} \neq 0 \,|\, \beta_i^{\ell-1} = 0)$, $2 \leq \ell \leq 6$, where $i = 1, \ldots, m$ indexes the m experimental factors and β_i^{ℓ} represents a linear parameter associated with pure polynomial predictor X_i^{ℓ}.

2. $\mathbf{T}$, a 2×2 matrix which specifies the probability of interactions of order 2 and 3 appearing in the model, conditional on the appearance or non-appearance of an interaction term one order lower. The elements of $\mathbf{T}$ are given as $t_{11} = P(\beta_{ij} \neq 0 \,|\, \beta_i \neq 0 \cup \beta_j \neq 0)$, $t_{12} = P(\beta_{ij} \neq 0 \,|\, \beta_i = 0 \cap \beta_j = 0)$, $t_{21} = P(\beta_{ijk} \neq 0 \,|\, \beta_{ij} \neq 0 \cup \beta_{ik} \neq 0 \cup \beta_{jk} \neq 0)$, and $t_{22} = P(\beta_{ijk} \neq 0 \,|\, \beta_{ij} = 0 \cap \beta_{ik} = 0 \cap \beta_{jk} = 0)$, where i, j, and k refer to particular experimental factors.

3. r, a flatness index. Smaller values of r (e.g. $r = 0.5$) are associated with steeper response surfaces while larger values (e.g. $r = 2$) indicate a flatter surface.

As an example, consider the information in Table 2, taken from Ockuly et al. (2017) who investigated notions of effect sparsity and heredity in response surface studies. Note that the last entry in the table does not appear in Ockuly et al. (2017) but was computed using the dataset from that paper.

Table 2: Summary of relevant sparsity and heredity information from data described in Ockuly et al. (2017).

Type	Proportion	Example of Active Terms
{Main}	0.55	A, B
{Quadratic|Main active}	0.58	A, A^2
{Quadratic|Main not active}	0.33	A^2
{2fi|Both parents active}	0.33	A, B, AB
{2fi|Exactly one parent active}	0.16	A, AB
{2fi|No parents active}	0.07	AB
{2fi|At least one parent active}	0.242	A, AB or A, B, AB

We can use much of the information in Table 2 to build $\mathbf{S}$ and $\mathbf{T}$ matrices that produce response surfaces that reflect these quantities. Specifically, we can specify

$$\mathbf{S}^T = \begin{bmatrix} 0.55 & 0.58 & 0 & 0 & 0 & 0 \\ 0 & 0.33 & 0 & 0 & 0 & 0 \end{bmatrix}$$

and

$$\mathbf{T} = \begin{bmatrix} 0.242 & 0.07 \\ 0 & 0 \end{bmatrix}.$$

The $\mathbf{S}$ matrix entries provide a reasonable quantification of the chance of linear main effects being active, along with quadratic main effects, depending upon whether the linear term has entered or not. Similarly, the $\mathbf{T}$ matrix provides similar estimates regarding two-factor interactions, depending upon whether at least one factor from the interaction is active or not. Besides the $\mathbf{S}$ and $\mathbf{T}$ matrices above, we used two other $\mathbf{S}$ matrices: 1) all linear and quadratic main effects active; 2) all pure polynomial effects up to 6^{th}-order active. Similarly, we used two other $\mathbf{T}$ matrices: 1) all two-factor interactions active; 2) all second and third-order interactions active. We also used three r values, $r = \{0.5, 1, 2\}$. Altogether, we simulated over the 27 combinations of $\mathbf{S}$ matrices, $\mathbf{T}$ matrices, and r values. The particular $\mathbf{S}$ and $\mathbf{T}$ matrices used are given in Appendix B in Tables 4 and 5. Note that

the **S** matrix that included all high-order polynomial effects (the third matrix in Table 4, which we call $\mathbf{S} = 3$) produced anomalous simulation results. Thus, in the main simulation results analysis in Section 4.2, we omit the scenarios with $\mathbf{S} = 3$ and consider them separately in Section 4.3.

The RSM simulation testbed of McDaniel and Ankenman (2000) uses the regression parameter values to shape the topology of the response surface, so the user cannot specify the true parameter values directly. Instead, a range of response values must be specified. We chose the response range to be similar to that produced in prior simulation studies where coefficients were selected to be between -3 and 3. In addition, based on coefficients generated using the testbed, we specify an error standard deviation of $\sigma = 0.5$.

Since this paper focuses primarily on RSM scenarios in which careful *a priori* screening was not done, we allow the number of active factors to be chosen at random from 2 up to m, before the **S** and **T** matrices are applied to further define the true underlying models. If screening had been performed, we would expect a less sparse set of true models. In the middle panel in Figure 1 we see that most of the scenarios retain less than 50% of the full model. Overall, compared to the empirical responses in the left panel of Figure 1, the simulated models are more sparse. This works for our purposes, because we wish to investigate the performance of RSM analysis methods in relatively sparse settings where screening has not previously been performed. With the 15 designs and 18 response surface types in our main simulation scenarios (excluding those with higher-order polynomial true models), we had a total of 270 simulation scenarios, each simulated 1000 times. We examined the $15 \times 9 = 135$ scenarios that include high-order polynomials separately, in Section 4.3.

4.1.3 Optimization and Prediction of Simulated Response Surfaces

In order to compare the analysis methods, we wish to evaluate their performance in identifying the true optimum point in the design space as well as measuring their predictive performance on out-of-sample points.

Response surface optimization (whether it be optimization of the true model, a fitted model, or a Bayesian reliability) is performed via the L-BFGS-B algorithm (Byrd et al. 1995) as implemented in the R function `optimx`. Starting values are selected using a random Latin hypercube. For the

Bayesian approach, the response subspace, R, is specified to be $R = [Y^{(0.95)}, \infty)$ where $Y^{(0.95)}$ is the 95^{th} percentile of the simulated response.

For confirmation runs, we used nine design points: one center point, four randomly selected design points, and four randomly selected points from the design region. The points from the design region were selected at random from a set of 500 points from a minimum potential space-filling design. The minimum potential designs were generated using JMP® (2018).

4.2 Simulation Results

Simulating as described in Section 4.1, we compared the analysis methods in terms of how accurately they identified the true optimum point in the design space, as well as their RMSPE on the confirmation runs described in Section 4.1.3. We analyzed the simulation results using both plots and formal linear models. We will discuss the plots primarily, but also provide some details on the modeling we performed. Since we performed many different simulations leading up to those reported in this paper, we used an explore/validate approach in which we ran a set of simulations with $\sigma = 0.5$, analyzed the results, and then validated them by running another set of simulations in the same way except with $\sigma = 1$. In both the graphs and the models, the $\sigma = 0.5$ results were validated by the second set of simulations.

We simulated over a fairly wide-ranging set of response surfaces, including those with three-factor interactions, so for some scenarios the full second-order model may fit poorly. Therefore, for analysis, we omitted any scenario for which more than 10% of the 1000 simulations resulted in a lack-of-fit (using $\alpha = 0.05$) when fitting the full second-order model. Unless otherwise specified, all results reported and discussed in this section omit these ill-fitting scenarios. That is, we report results based on 193 of the 270 scenarios.

4.2.1 Graphical Simulation Results ($\sigma = 0.5$)

To compare the quality of the various methods to optimize response surfaces, we measured the distance between estimated optima and true optima. In Figure 4, we observe that the methods based upon p-values, along with the Bayesian approach, appear to be more accurate than the other methods, in terms of identifying response surface optima. Figure 5 plots the distance between the

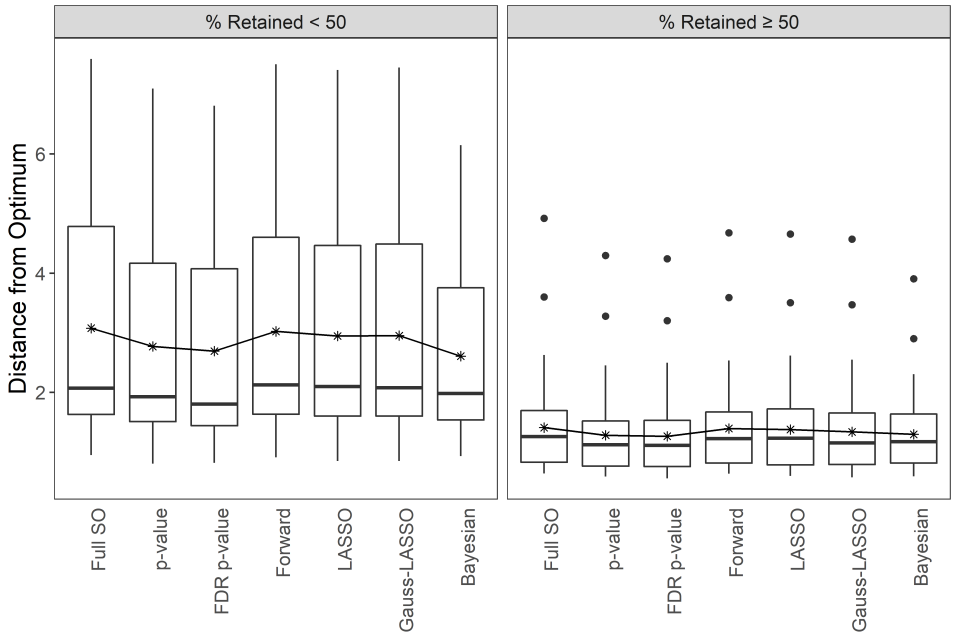

Figure 4: Euclidean distance between the point estimated as optimum and the true optimum, by % retained category. The stars, connected by lines, represent the mean of each distribution.

estimated and true optima by the number of factors. Although a similar pattern to Figure 4 is evident, the Bayesian and FDR p-value methods clearly differentiate themselves for larger k. For smaller k (most notably for $k = 3$), the methods exhibit fewer differences.

Figure 6 is an overall comparison of the analysis methods described in Section 2.1, in terms of predicting confirmation runs; log(RMSPE) is used to compare prediction performance. Note, again, that the Bayesian method is not included because with noninformative priors, it uses the full second-order model to make point predictions. We have omitted extreme outliers so it is easier to see differences between the methods. In Appendix C, the full version of the plot is displayed. From Figure 6, there is some evidence that the full second-order model and the LASSO perform worse than the other methods. Similarly, it appears that both the p-value methods as well as the Gauss-LASSO may perform slightly better than the others for prediction.

We note in passing that the same general patterns in Figures 4 and 6 hold up across the different design types (BBD, face-centered CCD, CCD with $\sqrt{m}$ axial distance), though for the CCD's with larger axial distances, predictions are slightly less precise and the ability to locate the optimum is worse than for the other designs. Also, the pattern in Figure 6 is similar to what we see across different confirmation run types (center runs, design points, random interior points). Likewise, the

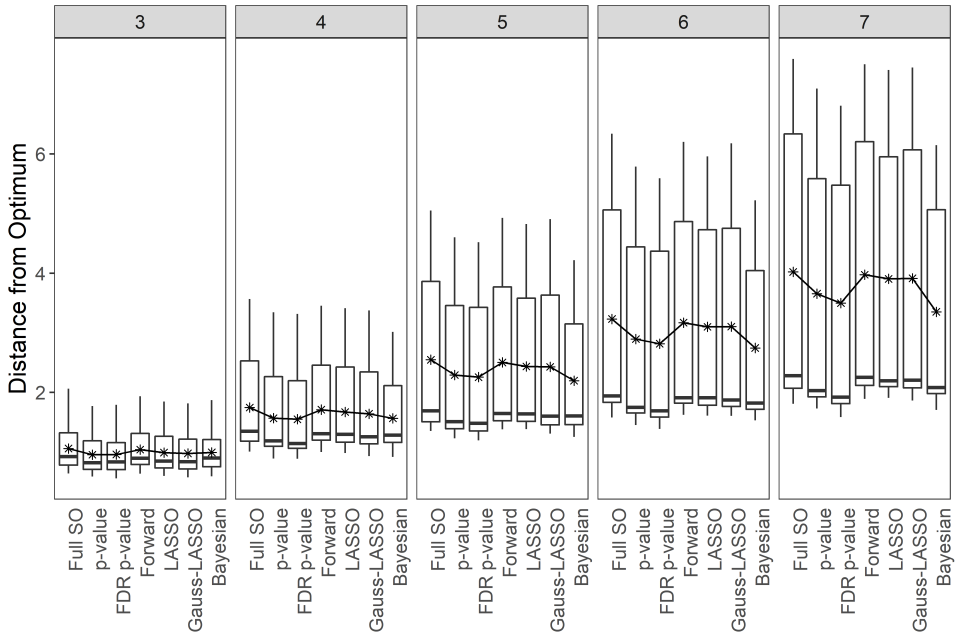

Figure 5: Euclidean distance between the point estimated as optimum and the true optimum, by k. The stars, connected by lines, represent the mean of each distribution.

same patterns emerge when comparing methods across the number of factors. LASSO is particularly poor for $k = 3$ but improves as k increases. These plots are omitted for brevity.

Roecker (1991) suggests that, at least in a general regression setting, the benefit of reducing a regression model depends upon the size of the true model. In particular, she pointed out that when more than 50% of the full model terms were discarded, predictions were improved when the model was reduced. Taking a cue from this work, we develop additional insight by observing Figure 7. For our simulations, it is clear that the full second-order model is inferior to several other methods, particularly when less than 50% of the terms in the full second-order model have been retained. (Recall that we estimated "% Retained" by averaging the number of terms retained across the five reducing methods, and dividing this average by the number of terms in the full second-order model.) When more than 50% of the model terms are retained, the full second-order model improves relative to the other methods, though it still appears worse than both p-value methods and the Gauss-LASSO.

Surprisingly, the p-value based methods perform most consistently, with the unadjusted p-value method providing good performance across the "% Retained" categories. The LASSO performs poorly overall, but the Gauss-LASSO is a solid performer across the two "% Retained" categories,

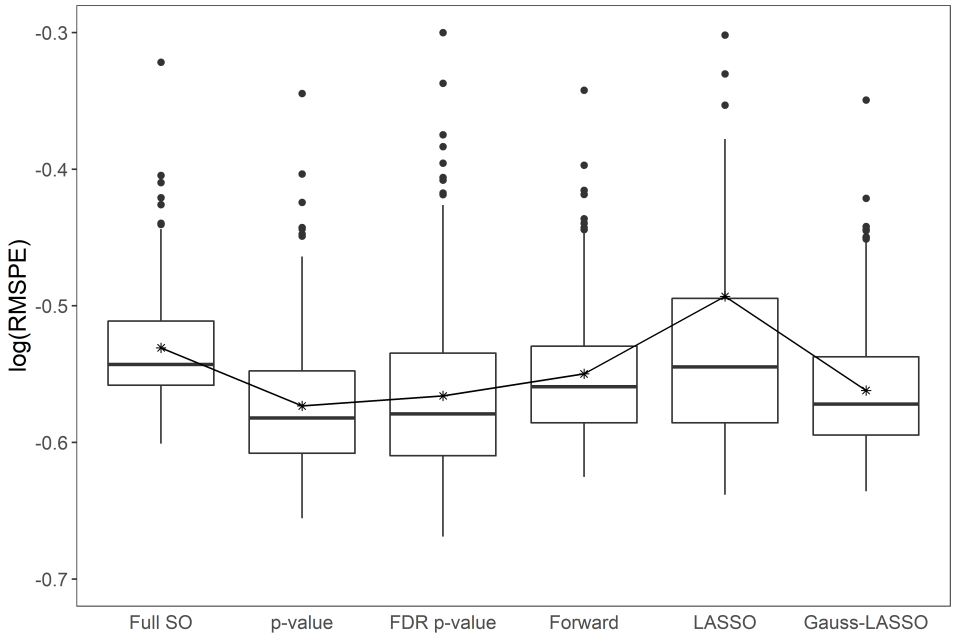

Figure 6: Comparison of confirmation run predictive accuracy of the six methods, based on simulated data. The stars, connected by lines, represent the mean of each distribution. Extreme outliers have been trimmed from this plot; for a full plot, see Appendix C.

comparable to the p-value method.

Figures 8 and 9 show 95% prediction interval (PI) widths and coverages for the least-squares-based methods. Bayesian 95% credible intervals coincide with the frequentist PIs for the full second-order model. As expected, the full second-order model produces the widest PI's, but maintains full coverage. The practice of reducing a statistical model and performing inference using the degrees of freedom pooled into the error by the inactive terms is statistically dubious, and there is an active literature discussing solutions to this problem (e.g. Berk et al. 2013; Barber et al. 2015; Lee et al. 2016; Tibshirani 1996). In this setting we see that the negative affect of unadjusted post-selection inference is clear but not overly drastic.

4.2.2 Formal Analysis of Simulation Results

To formalize the conclusions from the simulation, we develop a linear model for log(RMSPE) by treating **S** (1, 2), **T** (1, 2, 3), r (0.5, 1, 2), design type (CCD, Box-Behnken), and analysis method (Method) as categorical predictors, and run size (n), number of factors (m), and percent of terms retained (% Retained) as continuous predictors. Forward selection with AICc as the stopping

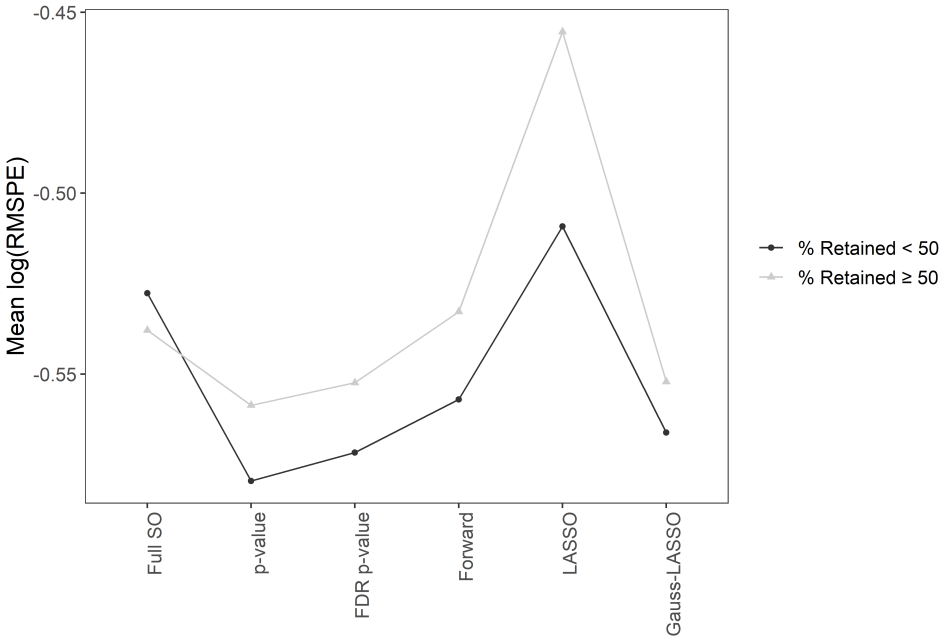

Figure 7: Mean log(RMSPE) for confirmation runs, by analysis methods and % retained category.

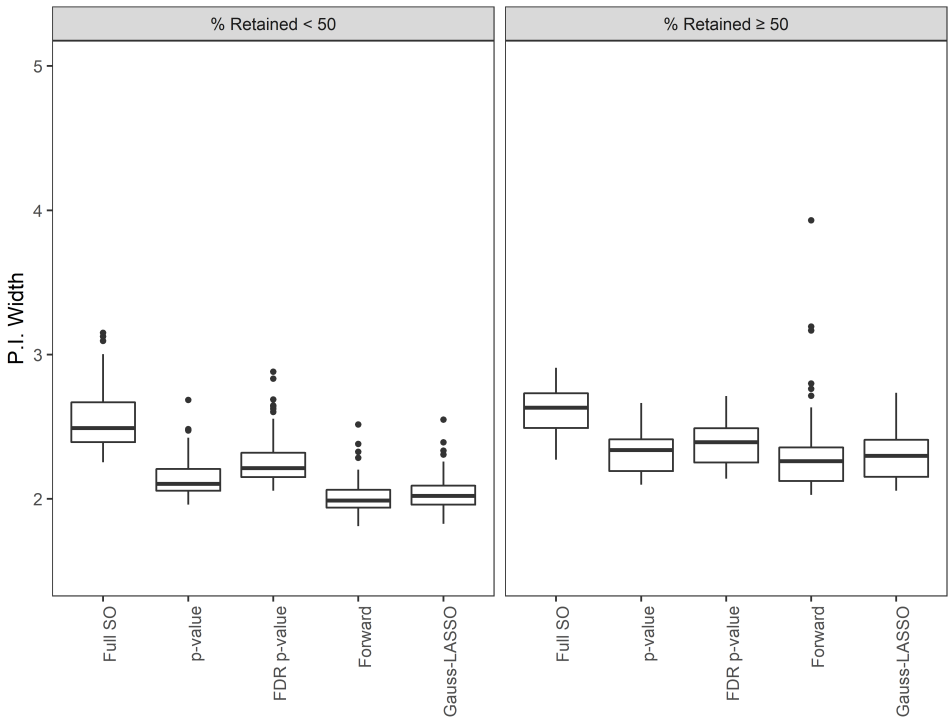

Figure 8: Width of prediction intervals using RSM analysis methods fit using least squares, for simulated response surfaces from Section 4.2.1.

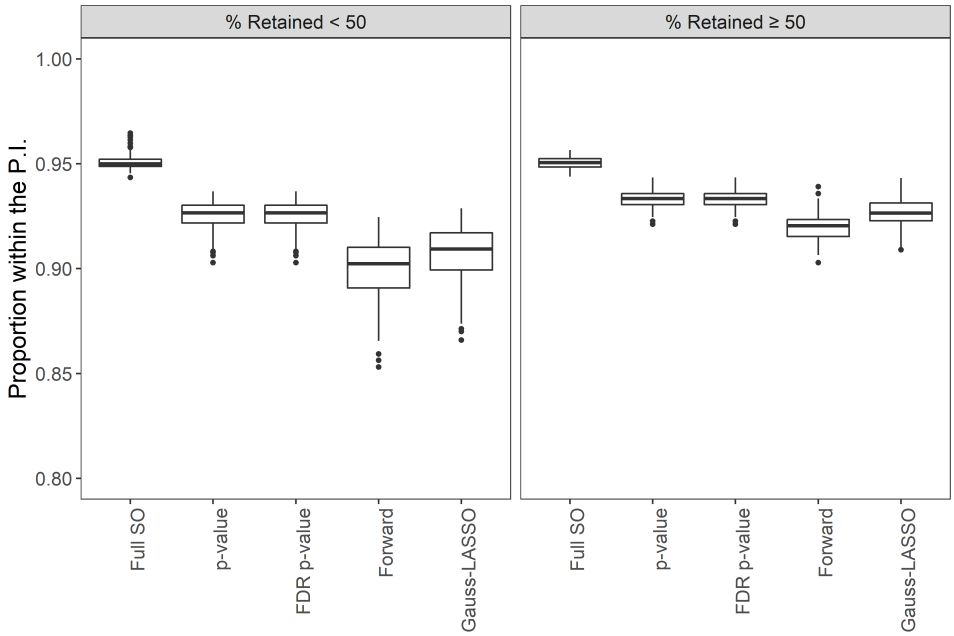

Figure 9: Coverage for prediction intervals using RSM analysis methods fit using least squares, for simulated response surfaces from Section 4.2.1.

criterion was used for model building. All main effects and two-factor interactions were included as candidate effects.

Using the fitted model, the main effect for Method (p-value < 0.0001) is displayed in Figure 10. Using a Tukey p-value adjustment to compare means, the LASSO exhibits a significantly larger log(RMSPE) than all other methods. No other pairs showed evidence of being different. Although a number of interaction terms involving Method are deemed active, one that is particularly noteworthy is Method*% Retained (p-value < 0.0001); see Figure 11. In particular, when the average percent retained is less than roughly 35%, all five reducing methods outperform the full second-order model with respect to prediction error. As expected, the competitive advantage of the reducing methods relative to the full model is diminished as the average percent retained increases; at about 75% terms retained, the full second-order model becomes roughly as good as the best of the reducing methods. The LASSO followed by forward selection exhibit the largest values of log(RMSPE) when percent retained is larger than roughly 55%.

We also fit a similar model for "distance of estimated optimum from true optimum" as the response. Figure 12 shows the Method main effect plot ($p < 0.0001$). The Bayesian approach produces the smallest distance and is significantly different from all other methods (all Tukey

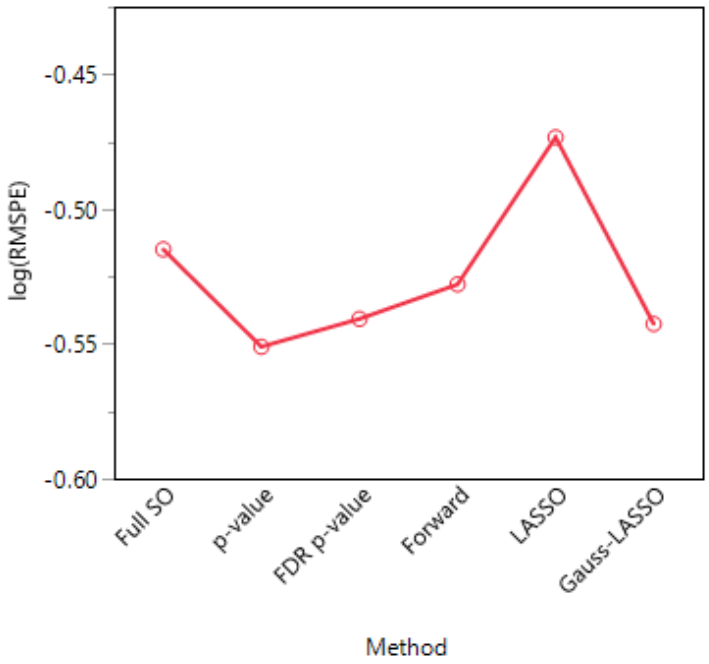

Figure 10: Method main effect plot for model with log(RMSPE) as the response.

adjusted p-values < 0.0001). Finally, there is strong evidence (Tukey p-values range from $p < 0.0001$ to $p = 0.0370$) that the full second-order model is worse than all other methods.

As a means of validating the forward selection model for our analysis of the simulations, we ran a second complete set of simulations; this time, the error variance was set to $\sigma = 1$. Again excluding scenarios with $S = 3$ as well as those with more than 10% LOF, the forward selection model for log(RMSPE) was fit to these validation simulation results. The Method main effect is again statistically significant ($p < 0.0001$); the validation model continues to indicate the inferiority of the LASSO for prediction. In contrast, we note that a Tukey test does not indicate significant log(RMSPE) differences among LASSO, the full second-order model, and the FDR p-value approach. The Method*% Retained interaction remained statistically significant (p-value < 0.0001) and a plot similar to Figure 11 was produced that demonstrated the advantage of the reducing methods over the full second-order model for smaller values of percent retained.

Likewise, the forward selection model for "distance of estimated optimum from true optimum" was fit to the second set of simulation results. The model results are identical with the exception that the Bayesian and FDR p-value approaches are now superior to all other methods; they are

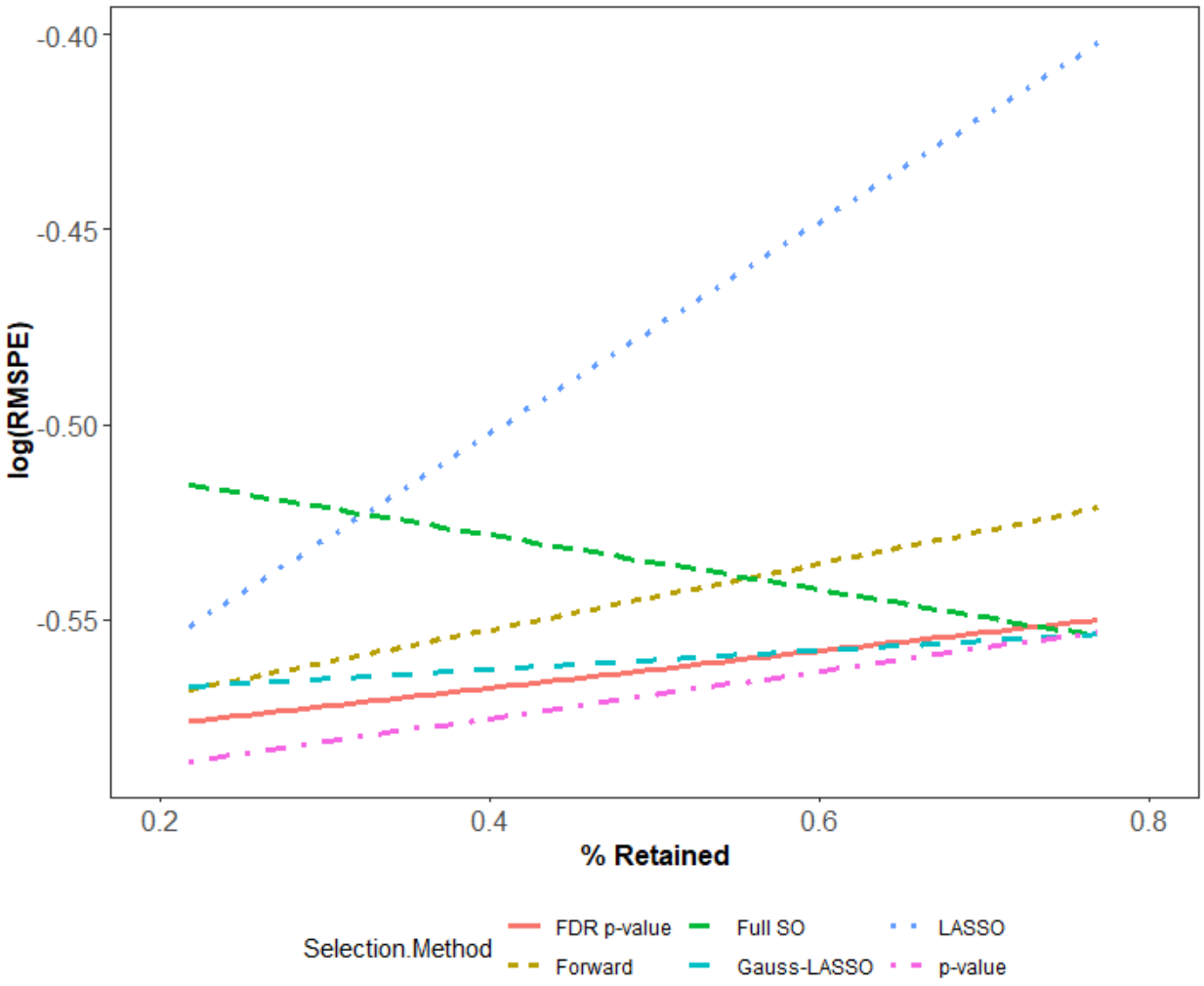

Figure 11: Method × % Retained Interaction interaction plot for model with log(RMSPE) as the response.

not significantly different from each other ($p = 0.6268$). The full second-order model and forward selection are inferior to all others.

4.3 Simulation Results for Complex Response Surfaces

We also explored simulation scenarios in which the true model included polynomial terms up to sixth-order. This corresponds to the third $\mathbf{S}$ matrix in Table 4 ($\mathbf{S} = 3$) of Appendix B. Clearly, true response surfaces with this level of complexity will typically prevent even the full quadratic model from adequately representing the actual response surface. As in the previous simulations, we used the same three levels of $\mathbf{T}$ and r, along with 15 designs and $\mathbf{S} = 3$, for a total of 135 simulation

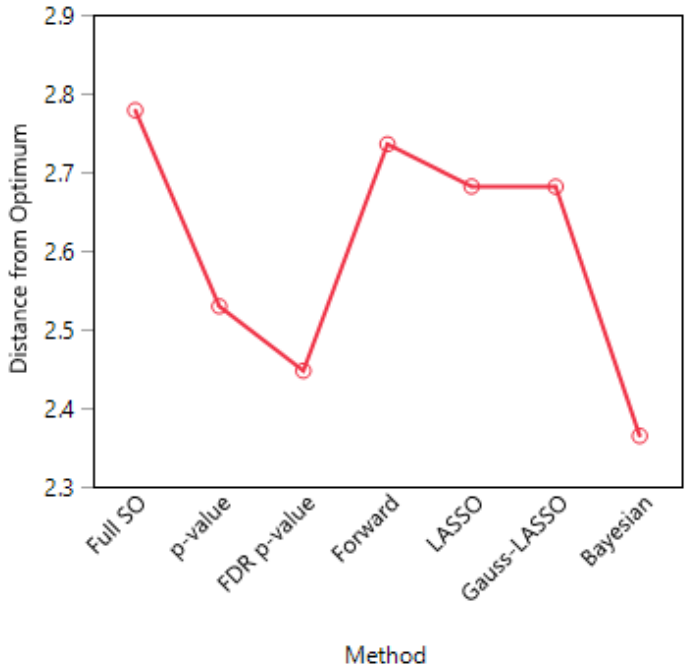

Figure 12: Method main effect plot for model with "Distance from Optimum" as the response.

scenarios. Recall that before choosing the model form (which in this case includes higher-order polynomial terms), we choose the number of important factors. This explains why the true model is complex but the proportion of terms retained when reducing the full second-order model is often much less than one, as shown in the right-most panel in Figure 1. Under this condition, the analysis methods produced larger models than in the simulation with simpler true models (see the middle and right panel of Figure 1).

Given the high-order polynomial nature of $\mathbf{S} = 3$, models with no more than second-order terms should objectively detect lack-of-fit. However, in about 44% of the scenarios they did not (that is, we detected lack-of-fit in less than 10% of the simulations for 44% of the scenarios). Interestingly, every scenario including a CCD with an axial distance of $\sqrt{m}$ was eliminated, probably because these designs include more extreme points which will be more likely to detect discrepancies between the true and fitted models.

Figure 13 shows that for these complex surfaces there is virtually no difference between the analysis methods for large or small models, in terms of prediction accuracy for confirmation runs. We conjecture the surfaces generated under the $\mathbf{S} = 3$ condition are too complex to distinguish

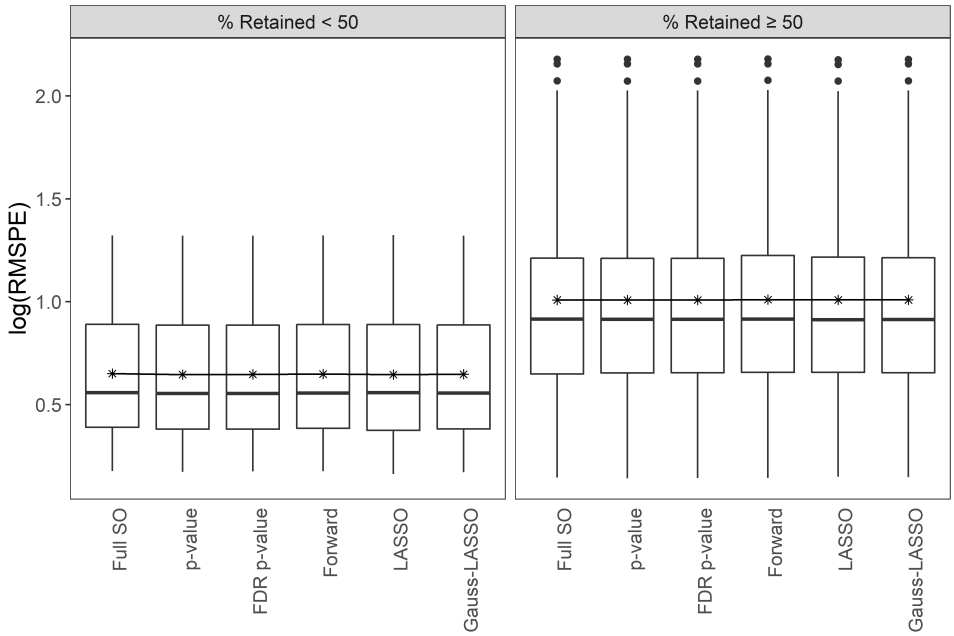

Figure 13: For complex response surfaces and the simulations with $\sigma = 0.5$, comparison of six methods by model size, in terms of log(RMSPE). The line represents the mean of each method.

between methods.

Figure 14 shows the distance between the optimum found by each analysis method and the true optimum, by % retained. We see a similar pattern for the complex surfaces as we did for the non-complex surfaces, with the p-value/FDR p-value and Bayesian approaches increasing their lead on the rest of the methods; this is particularly evident when % retained is less than 50%.

5 Discussion and Conclusion

In this article, we have studied the analysis of second-order response surface models. Since many practitioners of RSM appear to be performing and analyzing second-order experiments without previous experimentation, we examined whether analyzing these experiments with the full second-order model will overfit. Specifically, we compared using the full second-order model with several alternative analysis strategies, in terms of their ability to optimize the response as well as their prediction quality on out-of-sample runs. We used two datasets to compare the analysis methods. First, we used a sample of published RSM studies that reported at least one confirmation run. Secondly, we undertook an extensive simulation which we studied for both optimization and

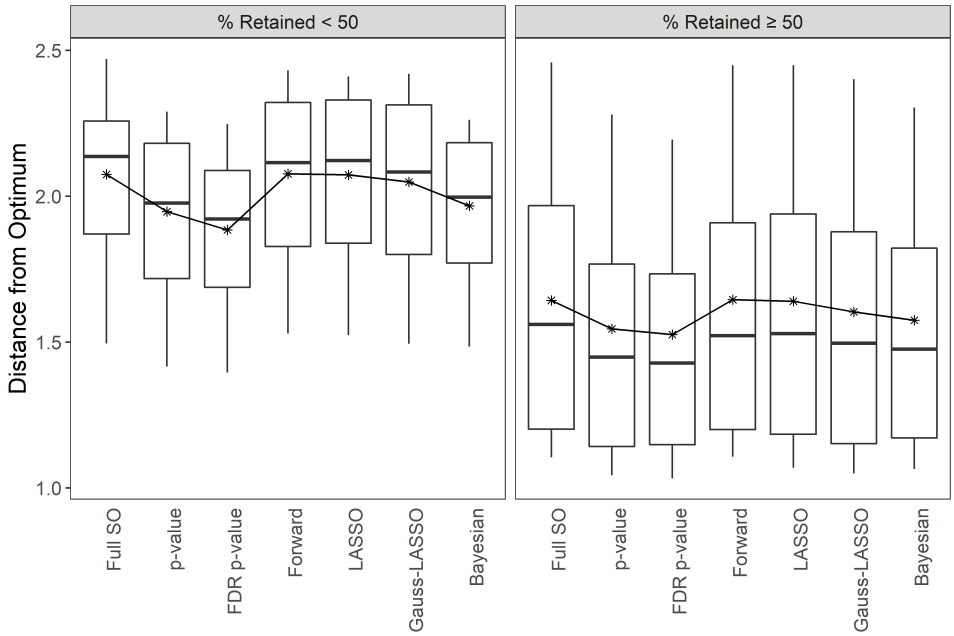

Figure 14: For complex response surfaces and the simulations with $\sigma = 0.5$, comparison of six methods by distance from the true optimum value. The line represents the mean of each method.

prediction.

The results of our study suggest several conclusions. First, optimization is best done using the Bayesian method, or by first reducing using the FDR-adjusted p-value method. Optimizing using the model reduced using unadjusted p-values is competitive as well. For complex surfaces, these also appear to be the preferred methods. Using the full second-order model is not recommended for optimization, if there are inactive terms.

Secondly, for more general prediction of out-of-sample points, using the unadjusted p-value method will be effective. The FDR-adjusted p-value method performs well too, though in the small sample from the literature it did quite poorly. The Gauss-LASSO also predicts reasonably well.

Thirdly, the full second-order model is not recommended when many terms are inactive. However, based on the results in Section 4.2.2, as the underlying true models get larger, the full second-order model eventually predicts as well as the best of the other methods, while the LASSO and forward selection get worse. The unadjusted p-value method does well across model sizes, further illustrating its effectiveness as a simple model-selection scheme for designs with little multicollinearity.

Fourth, for the simulations we explored (excluding the complex surfaces of Section 4.3), we found that coverage for prediction intervals only slightly declines when reducing the model using p-values or by the Gauss-LASSO. Though we don't recommend performing inference after pooling, we simply report the empirical evidence that we've found: in this RSM setting, nominally 95% PI's should be considered between 90% and 95% when based on models reduced by unadjusted or FDR-adjusted p-values at $\alpha = 0.05$.

To summarize our main conclusions: the Bayesian method is preferred for optimization, reducing the model via unadjusted p-values will perform well for both optimization and prediction, and the full second-order model is not recommended unless the number of unimportant terms is small.

It is worth noting that the idea of performing the final step of response surface methodology without the earlier steps (screening, first-order modeling) deviates from Box's classical perspective. In Box and Draper (2007), it is clear that the model-building strategy he promulgated was sequential, starting with simple models and then—if those models failed to capture the complexity of the experimental data—moving on to more complicated models. In contrast, one-shot experiments evident in the literature don't assume that initial experiments have explored simpler, first-order relationships before the second-order model is fit. As a reviewer pointed out, a one-shot experiment not only means screening hasn't been done, but also first-order line searches have been neglected and the quality of the model and the optimum is likely degraded because the design region is necessarily larger. Thus, the setting described in this paper is not ideal; we prefer that every response surface study would be sequential. However, when this path cannot be reasonably followed, we hope that this article provides some insight regarding how to optimize and predict.

As we have seen, the Bayesian approach has much to recommend in this setting: it is effective at finding the optimal point even when all regression terms are retained. On the other hand, with noninformative priors its out-of-sample predictions are the same as those for the poorly performing full second-order model. There is a well-known Bayesian interpretation of the LASSO (Park and Casella 2008; Hans 2009), which suggests a connection between the Bayesian and regularization methods considered in this article. Future work could explore how to retain or further improve the optimization results of the Bayesian approach, while strengthening its ability to make predictions across the design space.

References

Barber, R. F., Candès, E. J., et al. (2015), "Controlling the false discovery rate via knockoffs," *The Annals of Statistics*, 43, 2055–2085.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, 289–300.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013), "Valid post-selection inference," *The Annals of Statistics*, 41, 802–837.

Box, G. and Wilson, K. (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society. Series B (Methodological)*, 13, 1–45.

Box, G. E. and Draper, N. R. (1987), *Empirical model-building and response surfaces.*, John Wiley & Sons.

— (2007), *Response surfaces, mixtures, and ridge analyses*, vol. 649, John Wiley & Sons.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995), "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, 16, 1190–1208.

Candes, E. and Tao, T. (2007), "The Dantzig Selector: Statistical Estimations when p is Much Larger than n," *The Annals of Statistics*, 35(6), 2313–2351.

Chipman, H. (1996), "Bayesian variable selection with related predictors," *Canadian Journal of Statistics*, 24, 17–36.

Dean, A., Voss, D., and Draguljić, D. (2017), *Design and analysis of experiments*, Springer.

Del Castillo, E. (2007), *Process optimization: a statistical approach*, vol. 105, Springer Science & Business Media.

Draguljić, D., Woods, D. C., Dean, A. M., Lewis, S. M., and Vine, A.-J. E. (2014), "Screening Strategies in the Presence of Interactions," *Technometrics*, 56(1), 1–16.

Errore, A., Jones, B., Li, W., and Nachtsheim, C. J. (2017), "Using definitive screening designs to identify active first-and second-order factor effects," *Journal of Quality Technology*, 49, 244–264.

Hans, C. (2009), "Bayesian lasso regression," *Biometrika*, 96, 835–845.

Hawkins, D. M. (2004), "The problem of overfitting," *Journal of chemical information and computer sciences*, 44, 1–12.

Jensen, W. A. (2016), "Confirmation runs in design of experiments," *Journal of Quality Technology*, 48, 162–177.

JMP® (2018), "JMP Software, Version 13," SAS Institute Inc., Cary, NC.

Khuri, A. I. and Cornell, J. A. (1996), *Response surfaces: designs and analyses*, vol. 152, CRC press.

Lawson, J. (2003), "One-step screening and process optimization experiments," *The American Statistician*, 57, 15–20.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016), "Exact post-selection inference, with application to the lasso," *The Annals of Statistics*, 44, 907–927.

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

Mead, R., Gilmour, S. G., and Mead, A. (2012), *Statistical principles for the design of experiments: applications to real experiments*, vol. 36, Cambridge University Press.

Montgomery, D. C. (2017), *Design and analysis of experiments*, John wiley & sons.

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, 197–201.

Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. (2016), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, John Wiley & Sons.

Nelder, J. (2000), "Functional marginality and response-surface fitting," *Journal of Applied Statistics*, 27, 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.

Park, T. and Casella, G. (2008), "The bayesian lasso," *Journal of the American Statistical Association*, 103, 681–686.

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

— (1990), "A property of well-formulated polynomial regression models," *The American Statistician*, 44, 26–30.

Peterson, J. J. (2004), "A posterior predictive approach to multiple response surface optimization," *Journal of Quality Technology*, 36, 139–153.

Rajagopal, R. and Del Castillo, E. (2005), "Model-robust process optimization using Bayesian model averaging," *Technometrics*, 47, 152–163.

Rigollet, P. and Tsybakov, A. (2011), "Exponential screening and optimal rates of sparse estimation," *The Annals of Statistics*, 731–771.

Roecker, E. B. (1991), "Prediction error and its estimation for subset-selected models," *Technometrics*, 33, 459–468.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Wasserstein, R. L., Lazar, N. A., et al. (2016), "The ASA's statement on p-values: context, process, and purpose," *The American Statistician*, 70, 129–133.

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019), "Moving to a World Beyond "p¡ 0.05"," .

Weese, M., Smucker, B., and Edwards, D. (2015), "Searching for Powerful Supersaturated Designs," *Journal of Quality Technology*, 47(1).

Weese, M. L., Ramsey, P. J., and Montgomery, D. C. (2018), "Analysis of definitive screening designs: Screening vs prediction," *Applied Stochastic Models in Business and Industry*, 34, 244–255.

Wu, C. J. and Hamada, M. S. (2011), *Experiments: planning, analysis, and optimization*, vol. 552, John Wiley & Sons.

Yuan, M., Joseph, V. R., and Lin, Y. (2007), "An efficient variable selection approach for analyzing designed experiments," *Technometrics*, 49, 430–439.

Appendix A: Full plots (with outliers) for empirical results from Section 3

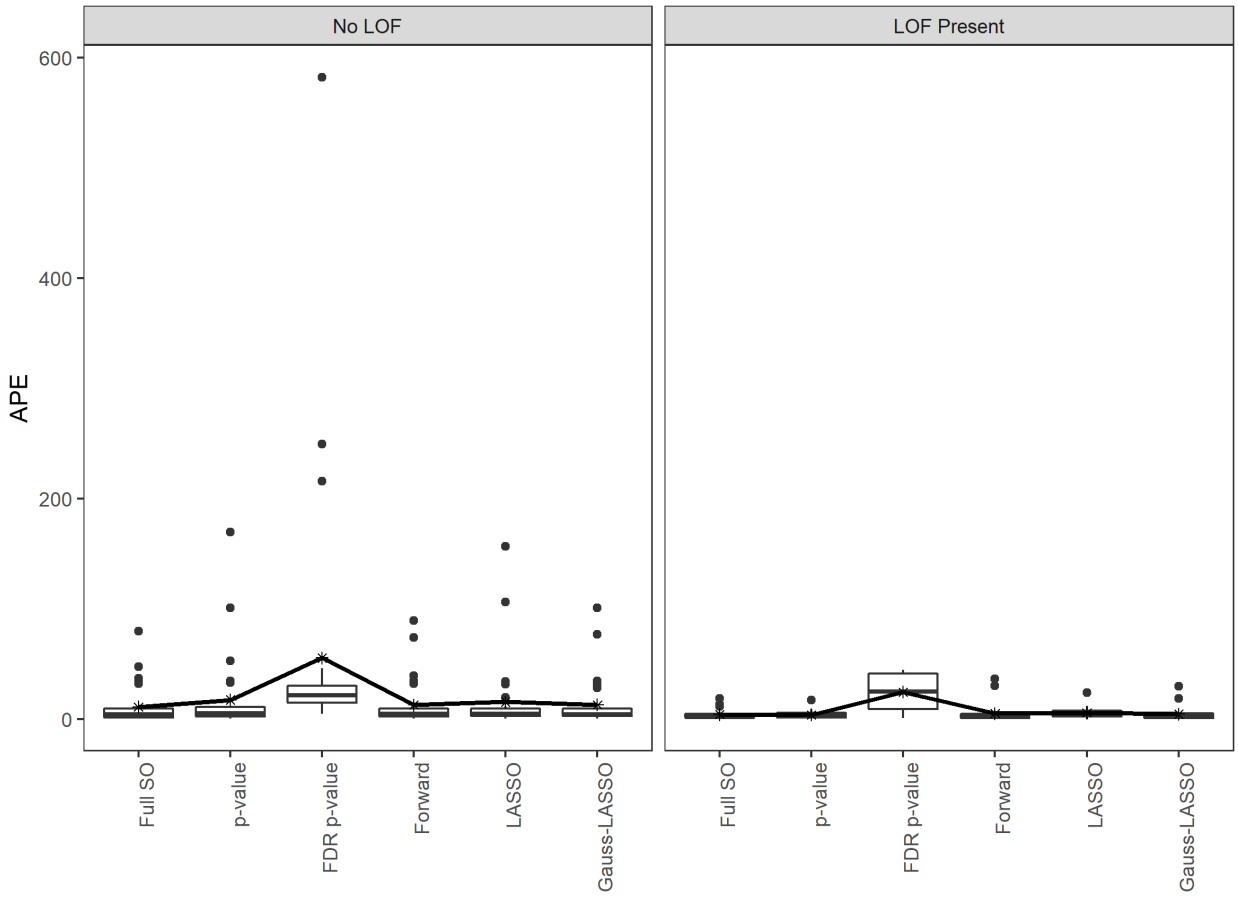

Figure 15: Absolute percentage errors for the six methods, over the 54 confirmation runs from the 25 responses in the literature, for models that exhibited lack-of-fit and for those that did not. The line represents the mean of each group. This is the full version of Figure 2.

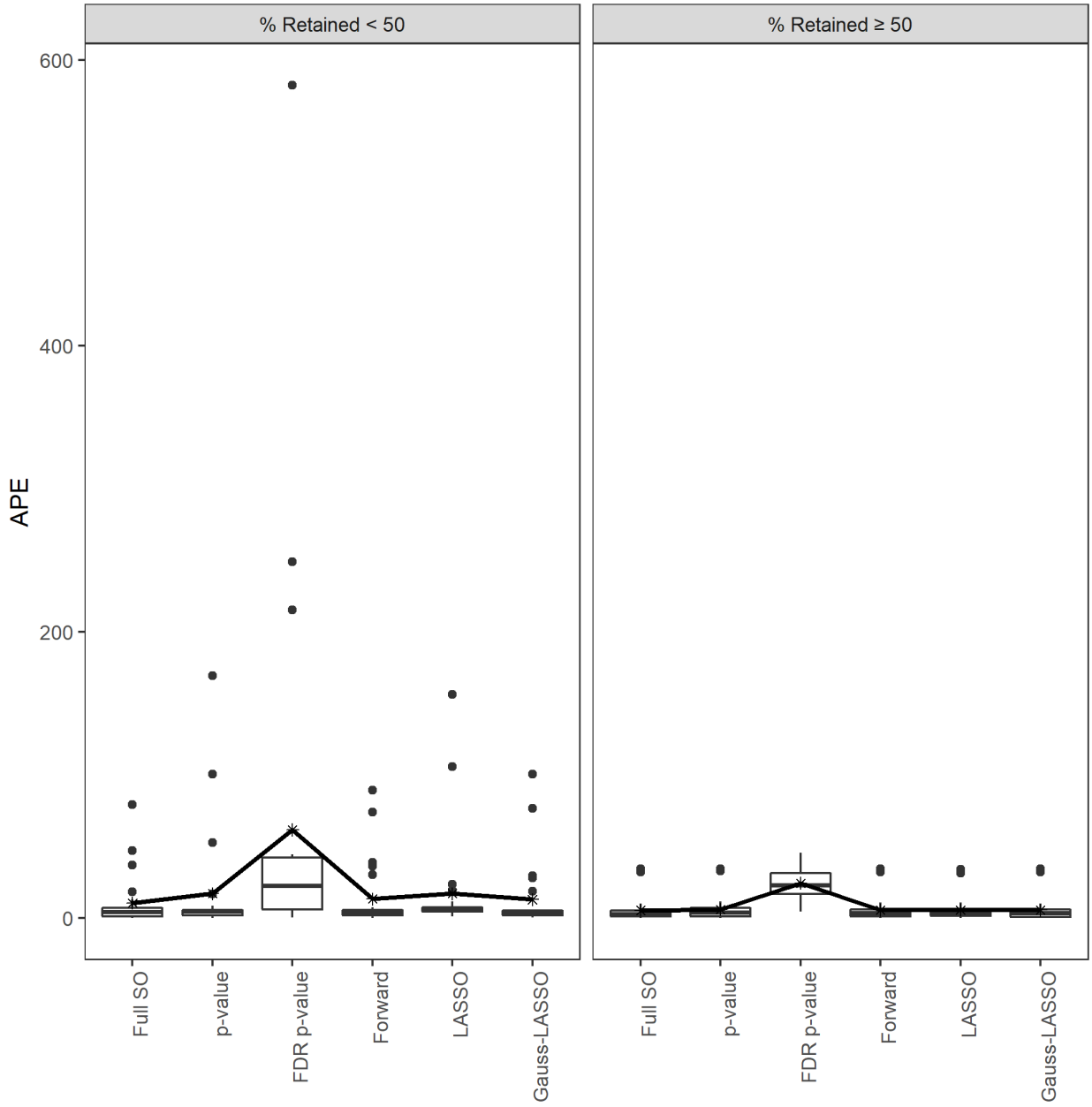

Figure 16: Absolute percentage errors for the six methods, over the 54 confirmation runs from the 25 responses in the literature, for models that were reduced more than 50% and those that were not. The line represents the mean of each group. This is the full version of Figure 3.

Table 4: Values of the **S** matrix used to generate various response surfaces.

Surface Properties	**S matrix**
Some linear and some quadratic	$\mathbf{S}^T = \begin{bmatrix} 0.55 & 0.58 & 0 & 0 & 0 & 0 \\ 0 & 0.33 & 0 & 0 & 0 & 0 \end{bmatrix}$
All linear and quadratic	$\mathbf{S}^T = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$
All up to 6^{th} order effects	$\mathbf{S}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

Appendix B: Collection of designs, S, T, and r quantities used in simulation

Table 3: Designs used in simulations. Central Composite Designs used an axial distance of $\sqrt{m}$ or 1 All designs used 4 center points.

Design	m	n
CCD	3	18
	4	28
	5	30
	6	48
	7	82
BBD	3	16
	4	28
	5	44
	6	52
	7	60

Table 5: Values of the **T** matrix used to generate various response surfaces.

Surface Properties	**T matrix**
Some 2-factor interactions	$\mathbf{T} = \begin{bmatrix} 0.242 & 0.07 \\ 0 & 0 \end{bmatrix}$
All 2-factor interactions	$\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$
All 2-factor and 3-factor interactions	$\mathbf{T} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

Appendix C: Full version of plot from simulation results

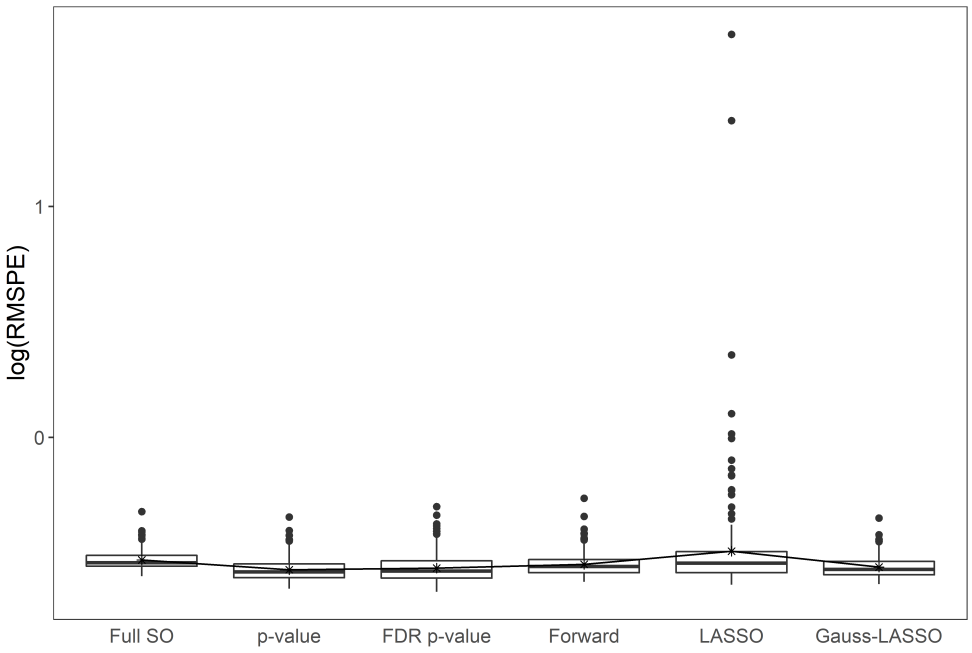

Figure 17: Comparison of six methods, in terms of log(RMSPE), used to analyze simulated RSM data. The line represents the mean of each method. This is the full version of Figure 6.