Comparing Sampling Strategies for Constructing D-Optimal Designs

Pedro Bruel

March 21, 2019

Optimizing problems with a large number of configurable parameters, or *factors*, is one of the main objectives of our transparent Design of Experiments approach to autotuning. Despite the high dimensionality of these problems, it is usually the case that only a small number of factors impacts the measured performance metrics significantly. Our approach aims to identify the most significant factors using an initial performance model, D-Optimal designs, and Analysis of Variance (ANOVA).

Our experiments with SPAPT kernels obtained results similar to a uniform sampling strategy, while using significantly less evaluations. Since it was unfeasible to evaluate all possible configurations of each SPAPT kernel due to their large search spaces, the experimental designs used in each iteration of our iterative approach were selected from a limited sample of configurations. The size of these samples were always small relative to the size of each search space, due to the high dimension and number of values for each factor.

A practical concern derived from the limitation on the size of samples to choose a design from is the large amount of time it takes to evaluate a relatively large number of candidate configurations. Each search space also had several constraints on factor levels, reducing the valid search space and slowing the process of finding candidates to test. Another concern is the intuition that the quality of the designs constructed from uniformly sampled candidates should always be poorer than in the case where designs are selected from complete search spaces. This property should be more pronounced when certain types of performance models are selected, such as quadratic models, due to the fact that practically none of the points sampled uniformly from a high-dimensional search space would be in the central region of the space.

Table 1: Design construction techniques and target performance models

Sampling & Construction Strategy	Performance Models
Uniform	Linear & quadratic
Biased	Linear
Biased + Normal	Quadratic
Uniform + *Fedorov*	Linear & quadratic
Biased + *Fedorov*	Linear
Biased + Normal + *Fedorov*	Quadratic

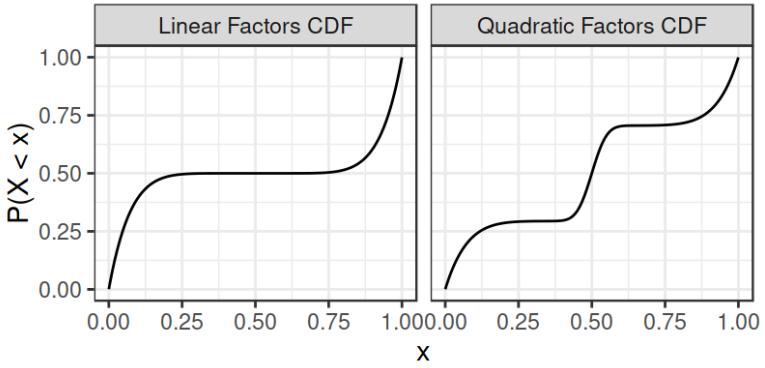

Figure 1: Cumulative Distribution Functions used for Biased Samples

These concerns motivated the development of a methodology to evaluate the quality of the designs produced by *Fedorov's algorithm*, the chosen D-Optimal construction technique, in the cases where designs were selected from samples with different properties. We studied design efficiency, measured by the *D-criterion*, for designs

constructed using the performance models and construction techniques listed in Table 1. Sampling strategies consisted of obtaining samples using the *cumulative distribution functions* in Figure 1, and adding samples from a normal distribution to the quadratic model biased samples.

This report is organized as follows. Section 1 presents the three metrics we used to compare each of the sampling strategies in Table 1, Section 2 describe the experiments we performed to compare sampling strategies, and Section 3 discusses the results obtained.

1 Comparing Sampling Strategies

In addition to the D-Criterion value, we also measured the quality of the designs generated by each technique using the *euclidean distance* between the vector of responses generated by different models fitted using the results of a design's experiments. We compared the results of a model using all factors and of a re-fitted model using only significant factors identified with ANOVA. We performed this comparison for all sampling strategies in Table 1.

We decided to sample new functions for each repetition of our experiments with D-Criteria and model fits. This was initially done by sampling values for the coefficients for each model term, generating a new function that was used to compute the results corresponding to the factor values of each of a design's experiments. Coefficient sampling is described in more detail in the following section.

1.1 Sampling Function Coefficients

Consider the functions $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}\colon \mathbf{X} \in \mathbb{R}^n \to \mathbb{R}$ of the form

$$f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}) = \alpha_0 + \sum_{i=1}^{n} \left(\alpha_i x_i + \beta_i x_i^2\right) + \varepsilon,$$

where $x \in [-1, 1]\ \forall x \in \mathbf{X}$, with *coefficients of linear terms* $\boldsymbol{\alpha} \in \mathbb{R}^{n+1}$, *coefficients of quadratic terms* $\boldsymbol{\beta} \in \mathbb{R}^n$, and normally distributed error ε. The procedure we used to sample new coeficients at each experiment starts by choosing which of the $\alpha_i \in \boldsymbol{\alpha}$, $\beta_i \in \boldsymbol{\beta}$ will be significant. With our target scenario in mind, each coefficient had a 15% chance of being significant. A uniformly sampled value in the interval $[-x_{\text{sup}}, -x_{\text{inf}}] \cup [x_{\text{inf}}, x_{\text{sup}}]$ is chosen for each significant coefficient, where $x_{\text{sup}}, x_{\text{inf}} \in \mathbb{R}$, $x_{\text{sup}} > x_{\text{inf}} > 0$. Small normally distributed values with mean zero and standard deviation x_{sd} are assigned to the other factors. Figure 2 shows one coefficient sample obtained by this process, where $|\boldsymbol{\alpha}| + |\boldsymbol{\beta}| = 2|\mathbf{X}| + 1 = 121$, $x_{\text{inf}} = 1$, $x_{\text{sup}} = 7$, and $x_{\text{sd}} = 0.04$.

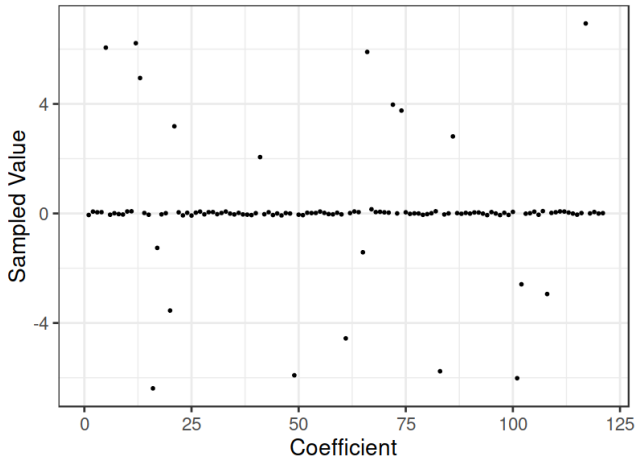

Figure 2: Sampled coefficients

1.2 Comparing Coefficient Vectors

Consider a design $\xi_{n,k}$ with factors $x_{k1}, \ldots, x_{kn}$, experiments $\mathbf{X}_1, \ldots, \mathbf{X}_k$, and design matrix given by

$$\xi_{n,k} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \ldots & x_{kn} \end{bmatrix},$$

where the k experiments were chosen by Fedorov's algorithm from a relatively large sample of size K.

Using a function $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ sampled as described in the previous section, we can compute the value or *response vector* $\mathbf{y} = (y_1, \ldots, y_k)$ for each experiment vector $\mathbf{X}_i = (x_{i1}, \ldots, x_{in})$, obtaining the design $\xi'_{n,k}$ with design matrix given by

$$\xi'_{n,k} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} & f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_1) \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} & f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \ldots & x_{kn} & f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_k) \end{bmatrix}.$$

We then use $\xi'_{n,k}$ to perform a linear regression with model

$$f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_i) = y_i = \alpha'_0 + \mathbf{X}_i^\top (\alpha'_1, \ldots, \alpha'_n) + \left(\mathbf{X}_i^2\right)^\top (\beta'_1, \ldots, \beta'_n) + \varepsilon,$$

obtaining the coefficient vector $\mathbf{C}' = \boldsymbol{\alpha}' \cup \boldsymbol{\beta}'$ which approximates $\mathbf{C} = \boldsymbol{\alpha} \cup \boldsymbol{\beta}$. The *coefficient vector distance* $\text{cvd}(\mathbf{C}, \mathbf{C}') = \|\mathbf{C} - \mathbf{C}'\|^2$ provides a measure of how well the fitted coefficients approximate the real function coefficients, and enables comparing two models.

1.3 Comparing Response Vectors

Consider designs $\xi_{n,k}$ and $\xi'_{n,k}$, constructed as in the last section. We can use $\xi'_{n,k}$ to perform a linear regression with model

$$f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_i) = y_i = \alpha'_0 + \mathbf{X}_i^\top (\alpha'_1, \ldots, \alpha'_n) + \left(\mathbf{X}_i^2\right)^\top (\beta'_1, \ldots, \beta'_n) + \varepsilon,$$

obtaining coefficients that define the function $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$ which approximates $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}$. Evaluating the fitted model in the original large sample of size K gives the response vector

$$\mathbf{Y}' = (f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}(\mathbf{X}_1), \ldots, f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}(\mathbf{X}_K)),$$

and evaluating the true sampled function gives the response vector

$$\mathbf{Y} = (f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_1), \ldots, f_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_K)).$$

The *response vector distance* $\text{rvd}(\mathbf{Y}, \mathbf{Y}') = \|\mathbf{Y} - \mathbf{Y}'\|^2$ provides a measure of how well the fitted model approximates the real function in the initial sample of size K, and enables the comparison of two models for the same data.

1.4 Comparing D-Criteria

Consider a D-Optimal design $\xi_{n,k}$ constructed using Fedorov's algorithm, with n factors, experiments $\mathbf{X}_1, \ldots, \mathbf{X}_k$ chosen from a large candidate set of K experiments, and design matrix given by

$$\xi_{n,k} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \ldots & x_{1n} \\ x_{21} & x_{22} & x_{23} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & x_{k3} & \ldots & x_{kn} \end{bmatrix}.$$

The quadratic performance model used to construct this design was

$$\mathbf{y} = \boldsymbol{\beta}\mathbf{M} + \varepsilon,$$

where $\boldsymbol{\beta}$ is the vector of $2n+1$ coefficients for linear, quadratic, and intercept terms, ε is the normally distributed error term, and $\mathbf{M}$ is the *model matrix*, a $(2n+1) \times k$ matrix containing each factor level as it would appear in the performance model, given by

$$\mathbf{M} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1n} & x_{11}^2 & \ldots & x_{1n}^2 \\ 1 & x_{21} & \ldots & x_{2n} & x_{21}^2 & \ldots & x_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{k1} & \ldots & x_{kn} & x_{k1}^2 & \ldots & x_{kn}^2 \end{bmatrix}.$$

The model matrix $\mathbf{M}$ is used to compute the *D-Criterion* $D(\mathbf{M}_{2n+1,k}) \in [0, 1]$. The D-Criterion of a design depends on the performance model used to construct it, and is defined as

$$\text{D}(\mathbf{M}_{2n+1,k}) = \left| \frac{\mathbf{M}^\top \mathbf{M}}{k} \right| \left(\frac{1}{2n+1} \right).$$

2 Experiments

The experiments described in this section use our Design of Experiments approach to optimize sampled linear and quadratic functions by identifying their most significant factors. First, we used the function coefficient sampling strategy described in Section 1.1 to generate target objective functions with large numbers of factors, but with few significant ones. Then, we used the biased sampling strategies described in Section 1 to generate sets of candidate points, which were used to construct D-Optimal designs using Fedorov's algorithm. The response vectors of each design were evaluated using the generated functions, which enabled running ANOVA tests to identify the most significant factors. Finally, new models were fit to the design's data using only the identified factors. The relevant parameters of each experiment are described in Table 2.

Table 2: Parameters of experiments with different regression models

Parameter	Linear Model	Quadratic Model
Factors	60	60
Design Size	70	130
Fedorov Samples	1000	1000
Testing Samples	1000	1000
Normal Samples Ratio	–	8%
Normal Samples Standard Deviation	–	0.04
Sampling Strategies	Uniform, Biased, Fedorov	Uniform, Biased, Normal, Fedorov
Sampled Objective Function	As in Section 1.1, $\boldsymbol{\beta} = 0$	As in Section 1.1
Noise Standard Deviation	$[2^0, \ldots, 2^6]$	$[2^0, \ldots, 2^6]$
Regression Model	Linear	Quadratic
Pr(> F) Threshold	$[0.001, 0.01, 0.1]$	$[0.001, 0.01, 0.1]$
Coefficients	61	121
Coefficient Significance Probability	15%	15%
Coefficient Noise Standard Deviation	0.01	0.01
Coefficient Range	($x_{inf} = 1$, $x_{sup} = 7$)	($x_{inf} = 1$, $x_{sup} = 7$)
Repetitions	100	100

Figure 3 lists the objects that can be observed and compared after applying our approach to build a performance model, represented by the function $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$, from observations of a function $f_{\boldsymbol{\alpha},\boldsymbol{\beta}} + \varepsilon$. In this experiment using function coefficient sampling, we are capable of observing $f_{\boldsymbol{\alpha},\boldsymbol{\beta}} + \varepsilon$, but we also have access to $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ and to the actual model coefficients $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. This means we can compare the accuracy of not only the responses predicted by $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$, but also of the coefficients $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$.

In real autotuning experiments we would not be capable of observing either the actual responses without added noise or the real model coefficients, and the only comparison possible is between the observed responses with noise $f_{\boldsymbol{\alpha},\boldsymbol{\beta}} + \varepsilon$ and the predicted responses $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$.

The experiment $\mathbf{X}$ in Figure 3 is one of the observed experiments $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$ that compose the D-Optimal designs used for ANOVA steps and the additional experiments in the test set. In the experiment using function coefficient sampling, the experiments $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$ consist of experiments selected for the D-Optimal designs, plus an additional uniform sample used later for testing.

The comparison between $f_{\boldsymbol{\alpha},\boldsymbol{\beta}} + \varepsilon$ and $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$ is meaningful only with regards to sampled regions of the search space. What we would actually like to know is the accuracy of $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$ for the regions of the search space where local and global minima of $f_{\boldsymbol{\alpha},\boldsymbol{\beta}}$ are found.

Comparing the coefficient vectors $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ is a better option for our experiment, but it is not feasible in real autotuning experiments. Despite that, we will show in Section 3.2 that comparing $f_{\boldsymbol{\alpha},\boldsymbol{\beta}} + \varepsilon$ with $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$, and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $(\boldsymbol{\alpha}', \boldsymbol{\beta}')$ leads to similar conclusions in this experiment. This motivates using the comparison between $f_{\boldsymbol{\alpha},\boldsymbol{\beta}} + \varepsilon$ and $f_{\boldsymbol{\alpha}',\boldsymbol{\beta}'}$ in experimental settings where $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is not available for comparison.

3 Results

3.1 D-Criterion

Since the evaluation of the D-Criterion, as described in Section 1.4, does not depend on the actual response values, we can group the experiments using the same regression model but different values of Pr(> F), as well as the experiments with and without added normal samples.

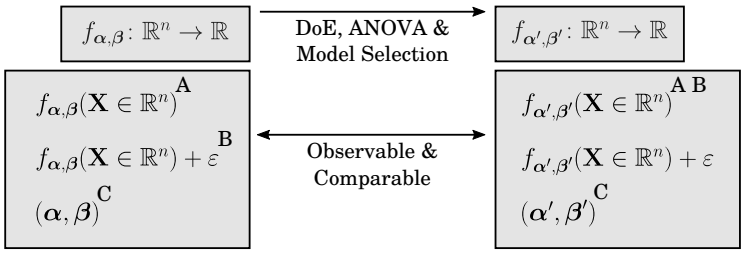

Figure 3: Comparing coefficients and output of real and predicted models

We see in Figure 4 that the D-Criteria was much higher for experiments using the linear regression model, and that the biased samples given to Fedorov's algorithm helped finding higher values of D. The magnitude of the improvements produced by using a biased sample were not expected for the linear model, given the high number of factors. These improvements might be explained by the fact that most uniformly sampled points in such high dimension would be located in the faces of the search space hypercube, while few or no points at all would be located in the hypercube corners or center regions. The improvements could be due to our biased sampling strategy forcing most points to the corners of the hypercube.

The values of the D-Criterion do not seem to change much in the experiments with the quadratic model, but we will show in the next sections that the models fitted to the data provided by different designs had different results. We attempted to add extra samples to the center region of the search space with a normally distributed sample with small standard deviation. Despite that, the observed D-Criteria remained much lower than the ones observed in linear model experiments. The smaller efficiency of designs could also be due to having double the number of factors and experiments than the linear model experiments.

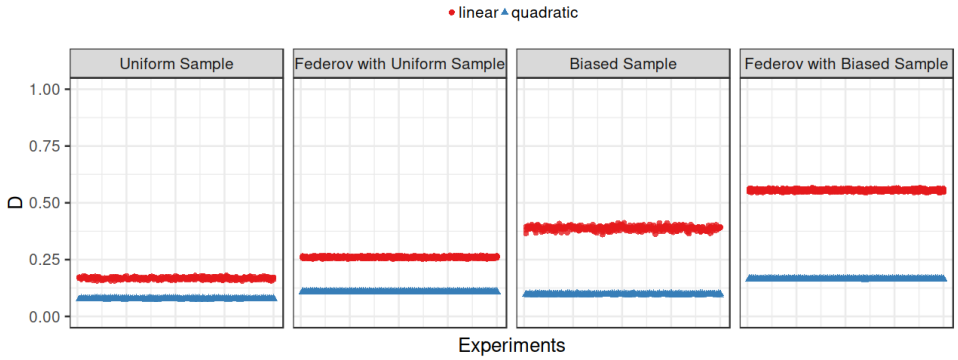

Figure 4: Comparison of D-Criterion values for linear and quadratic models

3.2 Coefficient & Response Vector Distances

We compared the sampling strategies in Table 2 regarding the distances between the coefficient vectors of fitted models and the real sampled model, as is described in Section 1.1 and shown in Figure 3. The results are shown in Figures 5 and 6 for the linear and quadratic regression models, respectively.

We see that the predicted vectors of coefficients get farther from the sampled coefficient vector as the standard deviation of the noise increases, for both regression models. The distances between models produced with uniform and biased samples also increases with the noise for the linear model, but it has a more complex relationship with the noise for the quadratic model. The $\Pr(>\text{F})$ threshold for factor selection impacts only the distances of the models selected after performing the ANOVA test, shown in the second column of each figure.

The horizontal lines in each plot represent a least squares regression using a constant model for the distances of each experiment group. The lines tagged with labels on the left represent experiments using a uniform sample, while the lines tagged with labels on the right represent experiments using a biased sample. Additionally, the dashed lines represent the results obtained with a model without filtering by ANOVA's $\Pr(>\text{F})$ thresholds. Full lines represent experiments using $\Pr(>\text{F})$ thresholds of 0.001, or "three stars", of 0.01, or "two stars", and of 0.1, or "dot" in R summary tables.

We can see the impact of different Pr(> F) thresholds after the ANOVA test by comparing labels of each line in Figures 5 and 6.

In the linear model experiments, shown in Figure 5, all threshold values produced improvements relative to the initial naive models. Higher thresholds achieved larger improvements than the smaller thresholds in general, and this difference increases with the standard deviation of the noise. Lower thresholds produced almost no improvements for higher noise values. This is also observed in the quadratic model experiments, shown in Figure 6, where it is more pronounced.

The chosen threshold values can be more important than the sample given to the D-Optimal design construction algorithm. This is seen in the quadratic model experiments where, even for lower noise values, the improvements achieved by high thresholds using uniform samples surpass the improvements achieved by low thresholds using biased samples.

In the linear and quadratic model experiments, we also observe that for higher noise standard deviations there is a split between closer and farther groups of coefficient vectors, with an empty region in the middle. This happens only after a model is selected using the results of ANOVA tests, and represents the impact of incorrect identification of significant coefficients.

It is interesting to note that, as the standard deviation of the noise increases, it is increasingly equivalent to fit a model without filtering significant parameters and to fit a model with low Pr(> F) thresholds, for both linear and quadratic model experiments. With high noise, it produces better results when we are less restrictive regarding parameter significance by using higher thresholds of Pr(> F).

Figures 7 and 8 show the results of comparing response vectors, as is described in Section 1.3 and shown in Figure 3. The data in Figures 7 and 8 was obtained during the same experiments show in Figures 5 and 6.

We observe essentially the same behaviours in the comparisons between coefficient and response vectors, despite the limited size of the sets used for comparing the models built using data from D-Optimal designs with the real models at each step. This motivates using the response vector distance as a comparison metric for models in experimental settings where the real coefficients are not available for comparison, such as real autotuning problems or experiments using more complex function sampling techniques.

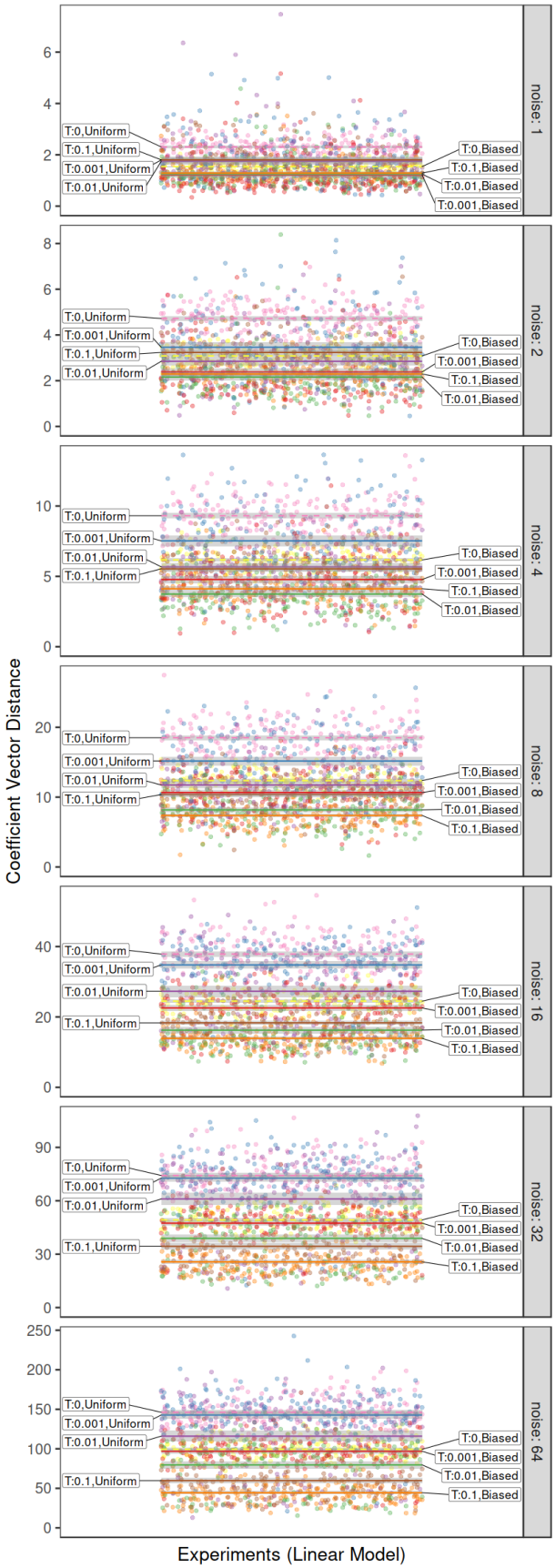

Figure 5: Coefficient distances for the linear model

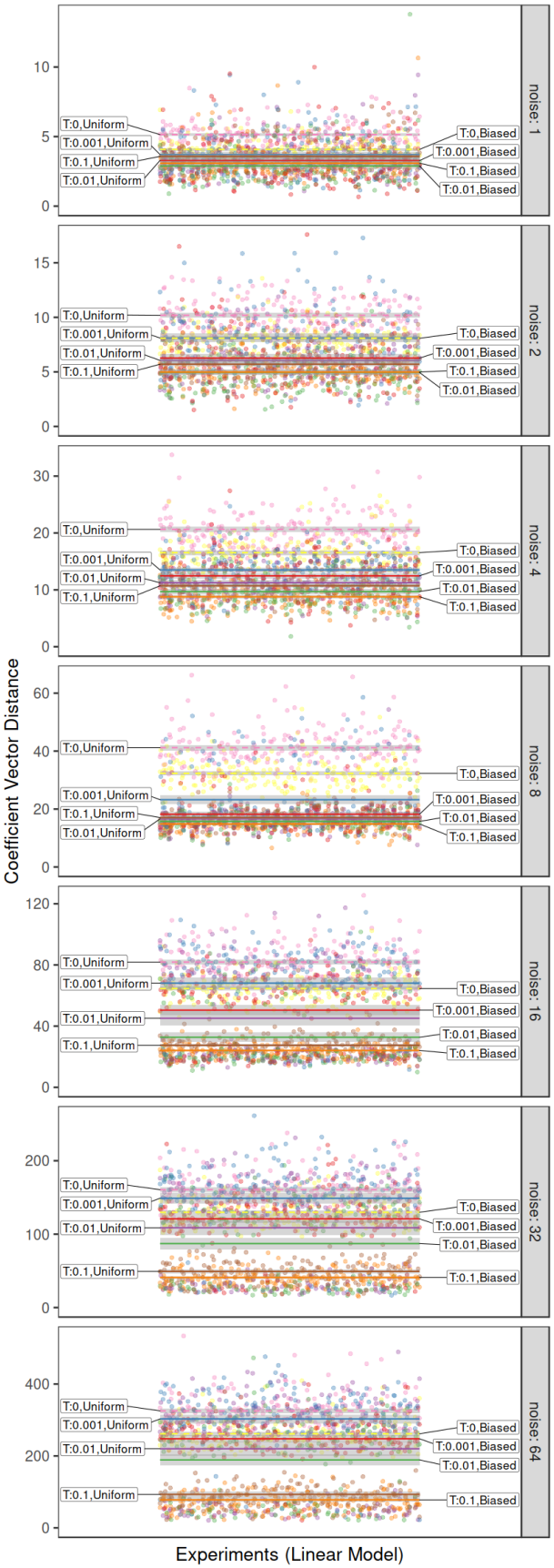

Figure 6: Coefficient distances for the quadratic model

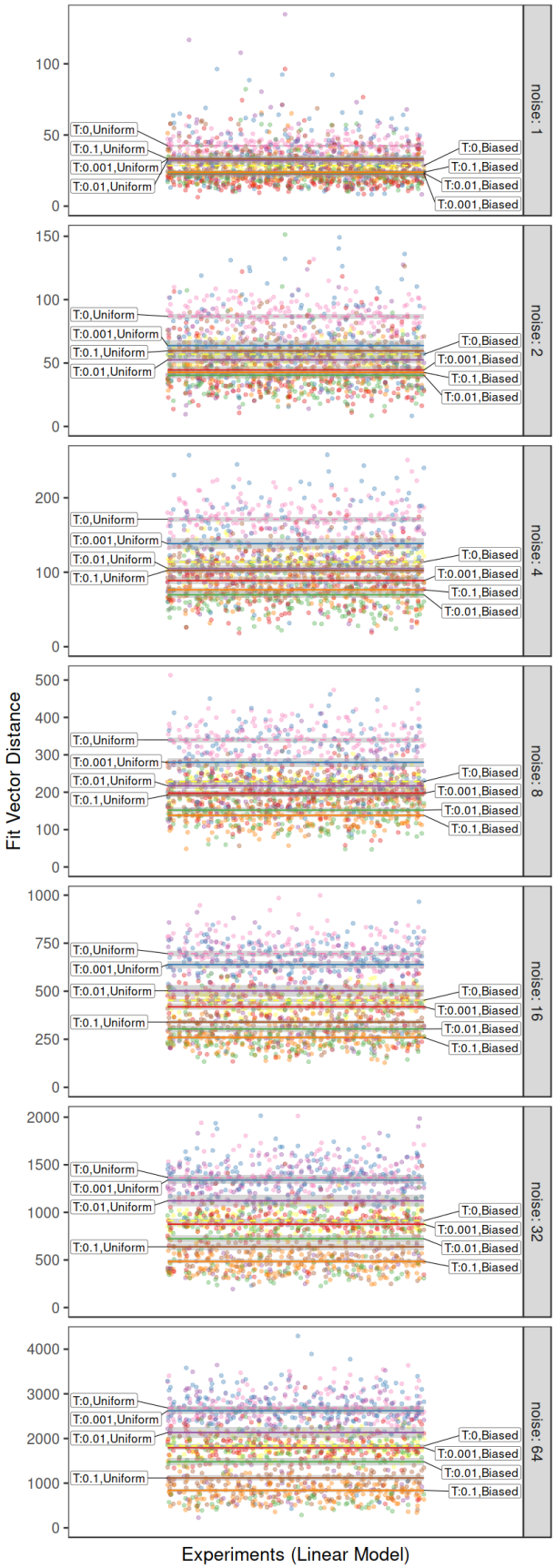

Figure 7: Fit distance for the linear model

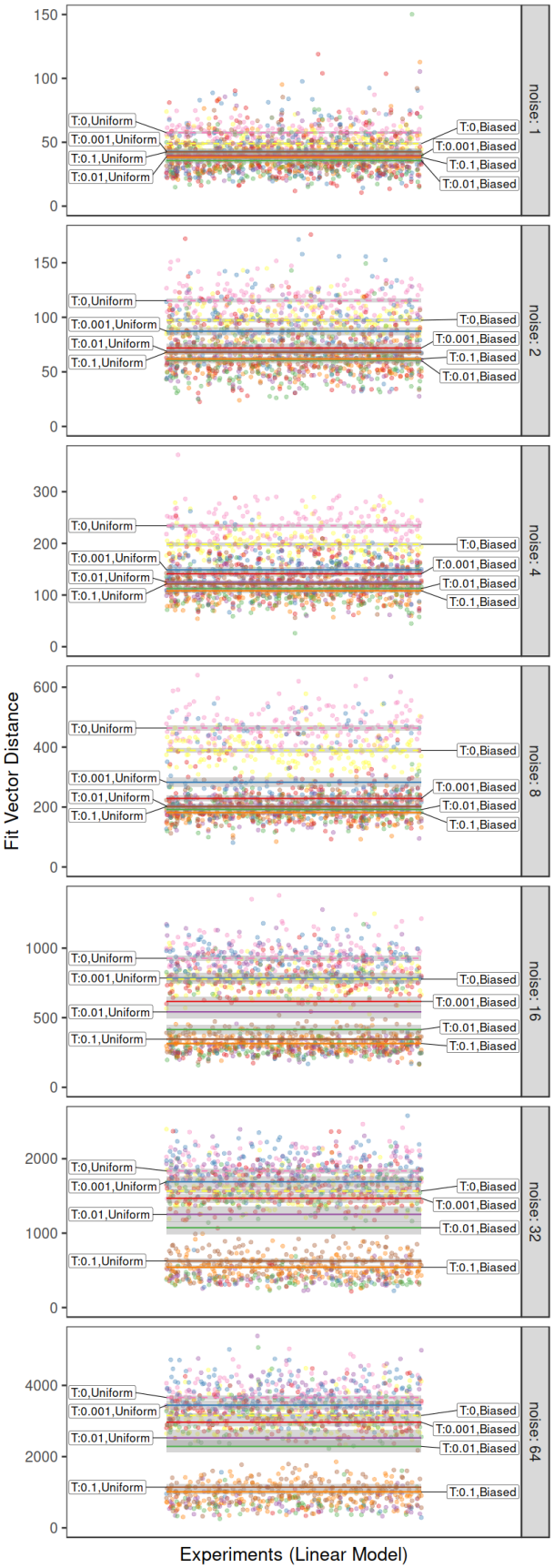

Figure 8: Fit distance for the quadratic model