6-2022

Bootstrapped Fractional Designs Applied to Models with Both Mixture and Process Variables

Laura Vinton
Union College - Schenectady, NY

Bootstrapped Fractional Designs Applied to Models with Both Mixture and Process Variables

Laura Vinton
Professor Roger Hoerl
03/17/2022

Table of Contents:

Executive Summary:

Mixture variables are unique in that all of the components must sum to 1. This causes problems when trying to create a model when there is interaction between mixture and process variables. The generally considered best model is the fully linearized model, but this can get quite large very quickly. Thus we began by comparing models on the full data sets of the Cornell Fish Patty Data, both the full and reduced Prescott data sets, and Wang's Melting data set. These models include the fully linearized model, the additive linear model, the KCV model, the SHB nonlinear model, and Zhong's nonlinear model. After seeing that both nonlinear models appear to be the most viable alternatives, we used the systematically selected fractions of each data set in order to obtain both an in and out of sample RMSE. This allows us to see if there is evidence of overfitting, how well the model predicts out of sample, and how well the model fits the training data. Upon looking at these results it became clear that Zhong's nonlinear model has serious overfitting issues, and the SHB nonlinear model and KCV model now appear to be the best potential alternatives to the fully linearized model. At this point, there are still very limited data points to look at in terms of how well the model fits. To increase the number of data points, we used bootstrapping to create a random sample that is proportional to the size of the full data set. This ensured that each model would run with limited warning messages. The resulting RMSEs indicated that Zhong's nonlinear model and the fully linearized model had extreme evidence of overfitting as the out of sample RMSE were extremely large. Thus, we considered the other 3 models as better options. These were not great for every data set. There was evidence of overfitting for these models with Cornell and Wang's data sets. The models seemed to do better for the Prescott data, which is interesting as it is the largest data set and the only data set where the mixture variables have constraints and the process variables have 3 levels. Within the Prescott data sets, the SHB nonlinear appears to perform the best with the least evidence of overfitting. As it also did well for the systematic fractions and full data sets, we conclude that this is in general the best alternative to the fully linearized model.

Introduction:

A mixture, which is also known as a formulation, blend, or composition is different from other types of variables when used in experimentation. Mixtures or formulations are seen frequently in everyday life and often have many responses of interest. A few real world examples include mixing a cake batter or the contents of paints. In both examples there are proportions of ingredients that give each cake or paint certain properties. A formulation variable has properties that result from proportions of ingredients rather than total amount of ingredients. When working with a formulation the sum of the proportions adds to 1. This fact makes formulations unique. As all the components must add to 1, the design space is changed. This change is caused by the fact that if you know all but one level of a component then you can compute the last by subtracting the sum of the known components from 1. (Snee and Hoerl 2016)

This property affects the geometry of the design space. In general, the design space involving formulations results in a space with a dimensionality one fewer than the factorial space. A few basic examples that can easily be visualized include a design space with 2 independent variables that results in a square while the design space with formulations results in a line, or a design space with 3 independent variables results in a cube while the design space with formulations results in a plane that is shaped like a triangle.This triangle is known as a simplex. Both of these examples are shown in Figure 1. One way to display formulation compositions is using trilinear coordinates. As an example, consider a region for 3 components. The resulting region will be a simple triangle with 3 vertices and 3 edges with each component running from a unique vertex to the center of the opposite edge. The constraints for each component run parallel to the edge that the component runs to. The triangle that results from this example has 2 independent dimensions where the intersection of any 2 lines is a point. (Snee and Hoerl 2016)

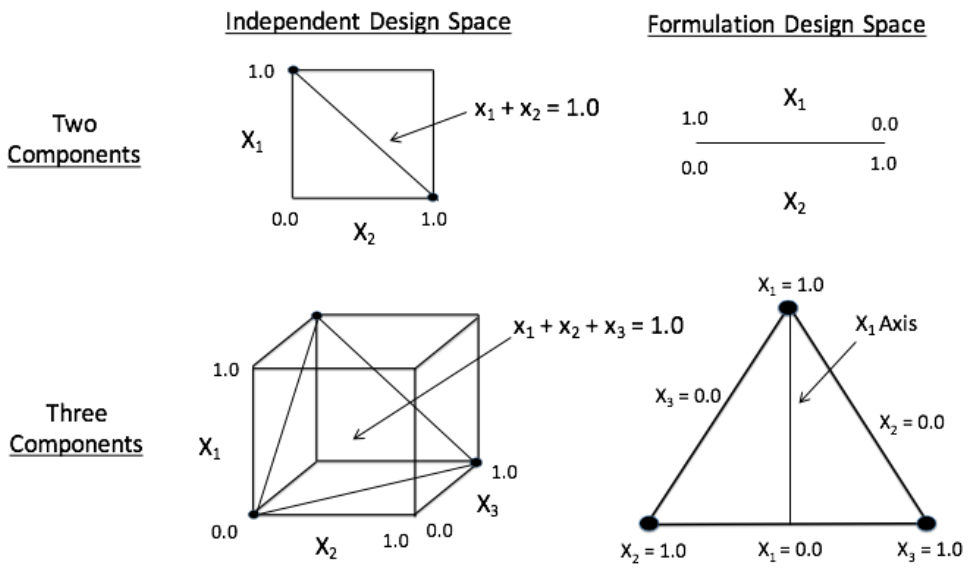

Figure 1: Figure 1.1 from Strategies for Formulations Development (Snee and Hoerl 2016)

An experiment involving formulation is different from one that only involves independent variables. One big difference is that the formulations have constraints, that is, the variable must be between 0 and 1. Another difference is that the components must sum to 1. These facts lead to the consequence that the dimensionality of an experiment involving formulations is 1 less than a factorial model. Regardless, experimental design and statistical models are related as the design enables the estimation of specific models. However once a design is completed the options for a model are limited. Scheffe (1963) was one of the first to start to lay the groundwork for mixture designs and modeling. He introduced the idea of a simple lattice design, which in general includes q pure blends ($^{q}_{2}$) 50-50 blends, ($^{q}_{3}$) ⅓-⅓-⅓ blends, and so on. Figure 2 will show an example of a simplex centroid with 3 components. It also includes a single centroid. This leads to the full simple lattice having 2^{q}-1 runs that can become an impractical amount of runs. Thus, often in practice a reduced simplex centroid is used. This includes the pure blend, 50-50 blends, and a centroid. Going forward this reduced design will be referred to as the simplex centroid. Even though replication is a very important part of the design, replicating all the points can become unfeasible. Thus, replication typically occurs at the

centroid or at the pure blends. Another popular option for replication is using a “checkpoint” that checks the halfway points between the vertices and the centroid.

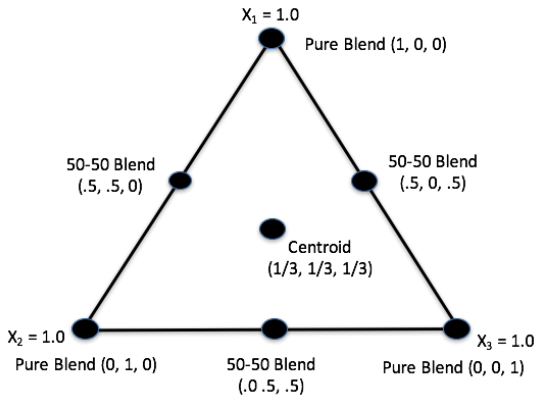

Figure 2: Figure 3.2 from Strategies for Formulations Development (Snee and Hoerl 2016)

Process variables are variables that are not subject to the constraint that they must sum to 1 like the mixture components are. If an experiment has both process and mixture variables then designing and modeling becomes more complex. If the two types of variables do not interact then it’s possible to use separate designs and additive models that combine the mixture and process models. That is, use a linear additive model, which simply adds the formulation model to the mixture model. However if the mixture and process variables interact then integrated designs and models are needed to ensure that the interaction is accounted for. (Snee and Hoerl 2016)

Standard designs for combining both process and mixture variables will often cross a factorial with a mixture by either running the factorial design at every point in the simplex or running the simplex at every point in the factorial design. An example of this can be seen in Figure 3. The designs involving both mixture and process variables can get quite large very quickly, so considering fractional designs is a way to try to be more efficient with time and resources. This is done by running a fraction of the process design at each point in the simplex or running a fraction of the simplex at each point in the process design. When doing the former it is common to run the full process design at the centroid to try to get a better understanding of

what is going on in the interior of the design. Regardless of which is used it is important to alternate which fraction is run at each point, so that all interactions can be estimated. It is also important to note that one can arbitrarily reverse the fractional design. (Snee and Hoerl 2016)

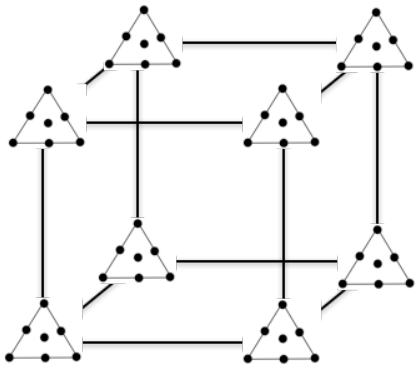

Figure 3: Figure 9.1 from Strategies for Formulations Development (Snee and Hoerl 2016)

One approach to modeling a design with both process and mixture variables and their interactions is using a multiplicative model, which in general is not linear in its parameters. Thus there is no simple way to get a least squares solution, we would need to use sophisticated software. One way to solve this is to linearize, or multiply out the individual terms in the two models, creating a linear model, for example, when there are 3 mixture and 2 process variables:

$$c(x,z) = f(x) * g(z) = b_1x_1 * a_0 + b_1x_1 * a_1z_1 + \ldots + b_{123}x_1x_2x_3 * a_{12}z_1z_2$$

where

$$f(x) = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$$

$$g(z) = a_0 + a_1z_1 + a_2z_2 + a_{12}z_1z_2$$

There are a few problems with this approach that result from the large number of terms to be estimated. These include; a complicated model that likely violates parsimony, the idea that models should be as small and simple as possible, and a difficulty with interpretation. (Snee and Hoerl 2016)

Another way to try to solve this problem is to just fit the equation as non-linear. This is still an overspecificed equation with no unique least squares solution, which can be seen by

multiplying f(x) by a non-zero constant and dividing g(z) by that same constant. We fix a_0=1 in order to solve this problem. Note that any non-zero constant can be used here, but 1 is helpful with interpretation of the coefficients. It is important to note that the resulting coefficients are not equal to what they would be in a linearized solution. There are a few problems with this approach as well, which include that this approach is not as flexible as the linearized ones and as this is a newer concept. It also has less underlying theory developed. (Snee, Hoerl and Bucci 2016)

Bootstrapping is a way to create a random fractional sample of data. This method selects a data point at random and puts it into the sample, then that data point is put back in, so it can potentially be picked again. This means that bootstrapping is a method of sampling with replacement. Bootstrapping is usually done multiple times, that is, multiple samples are created in this same fashion, creating many random fractional samples of the data. This allows for the creation of a possible alternative to the original systematic fractional designs that have been used previously for testing models in sample and out of sample. These samples can then be fed into multiple different models, and since there is replacement there will be some amount of the data remaining to predict with the model. This is important as both the in-sample root mean square error, or RMSE, which indicates how well the model fits the data within the sample, and the out of sample RMSE, which indicates how well the model can fit the data out of the sample, can be evaluated. These metrics indicate if the model fits the data while still showing if this fit is too good, that is, there is overfitting. Comparing these results with the results of systematic fractions will help indicate if this is an alternative to the original fractional method.

Literature Review:

Scheffe (1963) was one of the first to start to lay the groundwork for mixture designs and modeling. He introduces the idea of a simplex centroid design that includes the overall centroid of the simplex and the centroids of the lower dimensional simplexes that are contained in the overall simplex. When modeling the data gathered from this design he explains how the

constant term from a typical regression model is eliminated and thus that the resulting coefficients can't be interpreted in the same way as a model involving just process variables. Scheffe also explored designing experiments with both process and mixture variables. He recommended running a complete design of the process variables at each point in the simplex centroid. Scheffe even went further to explain the different cases that exist based on the levels of the process variables. The designs that combine both mixture and process variables get very large, so Scheffe introduced a way to reduce the size of these designs through fractionation. The fractional design that he explained can only be used when the process variables have two levels. He explains a rule for creating these fractions; if a point is included in a fraction then any other point that has the same value of the process variable and any point that includes a level of the component variable must also be included. An example of this rule for the component being at the centroid and the process variables being at a low level would result in all components and all process variables at low levels. Another way that he mentions for making the size of the model smaller is to delete higher-order terms that are less likely to be of great importance in advance, resulting in a smaller augmented simplex lattice, with which a smaller model can be run. Many have used these base ideas as the basis of their own research.

Some mixture problems complicate Scheffe's original design, that is , if a component has a constraint. This changes the design. For a design with only formulations, it is no longer a simplex; rather it becomes an extreme vertices design. The type of constraint each component has will guide the type of design. For example if there are only lower bounds then a simplex can still be utilized. However this simplex will look like a subset of a simplex with the full range of its components as seen in Figure 4. If the components have both lower and upper bounds the design space will not look like a simplex. This can be seen in Figure 5 that contains an example with 3 components with both lower and upper bounds. It is also important to notice that the centroid is not in the same place as it was with a simplex. The centroid is found by averaging all the vertices. Thus, it is different for non-simplex designs. These design spaces can have edges

that are very different in length due to some components having wide ranges and others have short ranges. This can negatively impact the quality of the predictions of the models. To try to avoid this, one can include the midpoints of the long edges in the design. On the other extreme, if there are many vertices near each other creating a cluster, it can be more efficient to replace these points with a singular centroid. The points that are clustered in this way are called pseudo- replicates and they typically all lie on the same plane. In general when an extreme vertices design is chosen to be the best approach, then we typically start with a quadratic model, which is

$$E(y) = \sum_{i=1}^{q} b_i x_i + \sum \sum_{1 \le i < j}^{q} b_{ij} x_i x_j$$

It is important to remember that this is not a guaranteed appropriate model, but it is typically a good place to start. (Snee and Hoerl 2016)

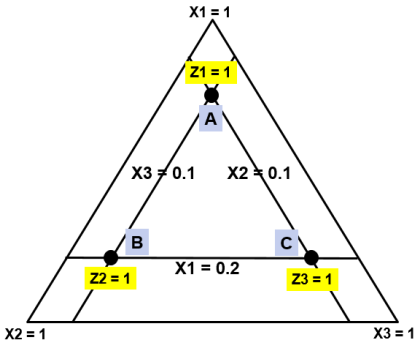

Figure 4: Figure 6.1 from Strategies for Formulations Development (Snee and Hoerl)

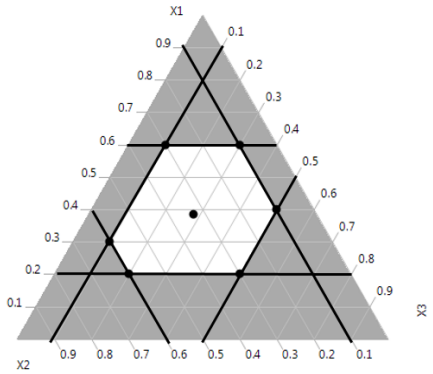

Figure 5: Figure 6.2 from Strategies for Formulation Development (Snee and Hoerl)

In order to build any successful model there is a process that should be followed to try to ensure success. Model building doesn't start with the data, rather it starts with defining purposes and objectives. This is an important first step as without clear goals it is impossible to evaluate the model and determine its adequacy. Two examples of purposes include creating a better understanding of the relationship between the formulations and the responses and predicting future values of responses. After the goals are clarified it is time to get a better understanding of the data. To do this look at the pedigree, simple plots, and summary statistics. Then it is time to produce the first model by taking into account experimental design, trends in the data, and subject matter knowledge. Once the model is created it must be evaluated using model metrics and residuals (actual values- predicted values). A very important model metric is the RMSE, which is an estimate of the standard deviation of the errors. This metric is very important in determining how well the model fits the data. A lower RMSE indicates a better fit. Both in and out of sample RMSEs can be calculated and compared in order to evaluate fit and how well the model predicts out of sample values. One should not be concerned if their first model does not adequately represent the data and meet the goals of the model as the process of modeling requires many loops through model building and evaluating the model. (Snee and Hoerl 2016)

The most common type of model for formulations is the Scheffe models. Thus, these could make a reasonable starting place for models with formulations. These are different due to the fact that the components sum to one and so a linear regression model can't be estimated. One difference between the Scheffe linear model and the linear regression model is that Scheffe's model drops the constant term. Therefore its formula for the expected value, or long term average value that would be expected to observe these component levels is

$$E(y) = \sum_{i=1}^{q} b_i x_i$$

The model can also be written as follows that includes the error term

$$y_i = \sum_{i=1}^{q} b_i x_i + e_i$$

This is for a single specific observation of y. As the components sum to 1 the absolute value of the coefficients is not as meaningful as the relationships between the coefficients. In order to calculate the effect of a certain component use the following equation

$$E_i = (b_i - \overline{b_i})$$

Where $\overline{b_i}$ is the average of all other coefficients other than b_i. This equation only works for linear blending models that cover the full range of each component. If a model has curvature use the quadratic Scheffe model that incorporates cross product terms as follows

$$E(y) = \sum_{i=1}^{q} b_i x_i + \sum \sum_{1 \le i \le j}^{q} b_{ij} x_i x_j$$

As the components sum to 1 the cross product terms no longer represent interaction, instead they model non-linear blending. If there is severe non-linear blending, more terms can be added using the special cubic Scheffe model

$$E(y) = \sum_{i=1}^{q} b_i x_i + \sum \sum_{1 \le i \le j}^{q} b_{ij} x_i x_j + \sum \sum \sum_{1 \le i \le j \le k}^{q} b_{ijk} x_i x_j x_k$$

This does not include all the third order terms but is instead the quadratic model plus all the 3 term cross products. Again, the cubic terms are considered nonlinear blending rather than 3-factor interaction. The basic models for process variables are as follows:

The linear process model is:

$$E(y) = a_0 + \sum_{i=1}^{r} a_i z_i$$

While the quadratic process model is:

$$E(y) = a_0 + \sum_{i=1}^{r} a_i z_i + \sum \sum_{1 \le i \le j}^{r} a_{ij} z_i z_j$$

(Snee and Hoerl 2016)

When there are both mixture and process variables in an experiment, the design and resulting model becomes more complex. The way to go about determining which designs and models to use is based on the amount of interaction between the process and mixture variables. If there is no interaction, an additive model would be helpful, but if there is interaction then the model should have terms that account for this. The additive models do not do this. Thus, a different type of model should be used, and these tend to get quite large very quickly. The best way to attack this problem is by using a statistical engineering approach that focuses on how to utilize statistical concepts and tools and integrate them with other sciences to generate improved results. The strategy to solve this problem is sequential in nature and involves a lot of smaller concepts within statistics. The strategy also leaves room for improvement as if the resulting model is not strong enough, that is, there is a next step that can reasonably improve the model. (Snee, Hoerl, and Bucci 2016)

Now when there are both process and mixture variables and no expected interaction between the different types, the two types of model will be fit separately in a linear additive model:

$$c(x, z) = f(x) + g(z)$$

where f(x) is the mixture model and g(z) is the process model. If there is expected interaction a multiplicative model should be used:

$$c(x, z) = f(x) * g(z)$$

which is not linear in its parameters. There are multiple methods to address this problem, including linearizing (multiplying out the equation completely) the model and using a nonlinear model. Both these approaches come with downsides. The fully linearized solution is often quite large, which can become costly and time consuming while the nonlinear solution is not as flexible as the linearized solution and has less underlying theory. (Snee, Hoerl, and Bucci 2016)

Another approach from Kowalski, Cornell and Vining (2000) suggests a combined model based on Taylor series approximation. The idea suggested borrows the idea of running a subset

of the combinations and the idea of selecting a fraction using D-efficiency. Suppose the true model for the n process variables is :

$$\eta_{pv} = \alpha_0 + \sum_{k=1}^{n} \alpha_k z_k + \sum_{k=1}^{n} \alpha_{kk} z_k^2 + \sum \sum_{k<l}^{n} \alpha_{kl} z_k z_l$$

which includes n pure quadratic terms. Combining this with Scheffe's quadratic model for formulations we get:

$$\eta_{pv} = \sum_{i=1}^{q} \beta_i x_i + \sum \sum_{i<j}^{q} \beta_{ij} x_i x_j + \sum_{k=1}^{n} \alpha_{kk} z_k^2 + \sum \sum_{k<l}^{n} \alpha_{kl} z_k z_l + \sum_{i=1}^{q} \sum_{k=1}^{n} \gamma_{ik} x_i z_k$$

This model can be used even if the pure quadratic terms are not needed by omitting those n terms.

In most investigations the model is specified first and then a design is picked that will support the model and the design must include at least as many terms as there are parameters. To determine the design one suggestion is the central composite design (ccd) where there are 2^n factorial points, 2n axial points with ±α for 1 factor, and zero for the rest. It is also suggested to use at least one center point replicate. The design discussed in this article starts with the ccd for the process variables, but this design doesn't have to be used for every data set. Then for the combined design the subset of the simplex centroid is placed at each point in the ccd. There are 2 designs considered for the equation that combine the model of the process variables and Scheffe's model. Both include vertices of the simplex at ½ of the 2^n factorial point with midedge points at the other half. This spreads the blends evenly among the process variables. This is important because if a process variable is unimportant then there is still information on the blends with the other process variables. The only difference between the two designs occurs at the center of the process variables, which can be seen in Figure 6 below. The first design includes a full simplex at the center (F) while the other includes only the centroid blend at the center (C). This article looks at 5 combinations of mixture and process variables (MC=mixture component and PV= process variable): 3MC, 2PV; 3MC, 3PV; 4MC, 2PV; 4MC, 3PV; 3MC, 2PV

(with constraints on MC). The authors also utilize a central composite process design that means that there are 5 levels of each process variable. (Kowalski, Cornell and Vining 2000)

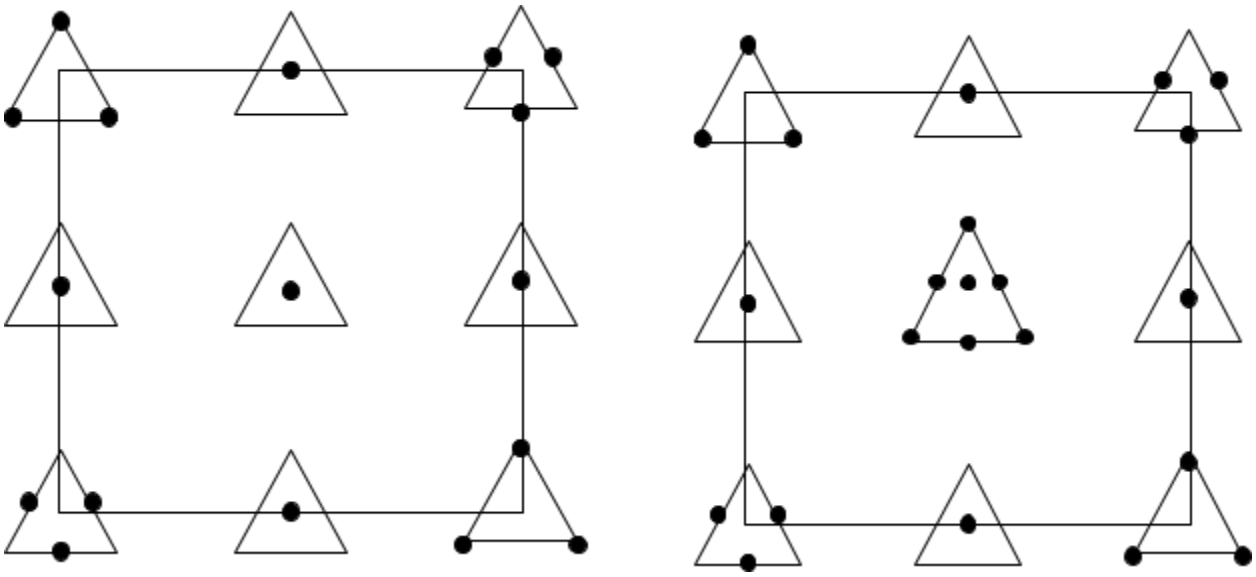

Figure 6: created based off FIG 1 from Kowalski, Cornell and Vining 2000

If the number of design points for either of these suggestions is less than the number when crossing the ccd with the full design in mixture components, then use a computer program to find the design using optimality criterion, such as D-optimality. One can also choose to use something slightly different if another design has other properties like symmetry or orthogonality. The suggested designs are compared to the designs chosen by PROC OPTEX in SAS, which requires a point list and a model to be fit. The point list for the suggested model is the simplex design at each point in ccd. The relative efficiencies of the 2 proposed designs are the D-criterion for the proposed design divided by the D-criterion for PROC OPTEX design with the same number points. After examining the combinations above we see both designs are at least 73% as efficient as the D-optimal design except for the combination of 3MC, 2PV, which is higher. Generally the design with just the centroid blend at the center is more efficient than the design with the full simplex in the center. This indicates that the additional points from the full simplex doesn't seem beneficial in terms of the D-criterion, but it has benefits in analysis.

Generally as the design size increases the D-criterion increases to a limit and then plateaus. (Kowalski, Cornell and Vining 2000)

Each design starts with N degrees of freedom (df) and to estimate the terms in the model p-1 degrees of freedom are needed. This leaves N-p df for an error source that can test the significance of the terms. Most cases have enough df to estimate the model, but when wanting to do significance testing it is preferred to have a full simplex in the center. Another way to ensure there are enough dfs is to run replicates at the center. When testing the effects ideally begin with the interaction between mixture and process variables. This involves asking the main question: Is the effect of each process variable constant for all blends of the mixture components? Followed by either: If so, is the effect significantly different from 0? or: If not, among which blends is the effect different? Hypothesis tests can then be run based on these questions to help determine if there is interaction between the mixture and process variables and if the process variables affect the mixture components equally. Regardless of the results, it can still be informative to look at each part of the model separately. It is possible that the results can suggest terms not to include or ones to add. Before eliminating a term it is important to test the model for lack of fit. (Kowalski, Cornell and Vining 2000)

Mushan Zhong (2019) looked at other nonlinear models that addressed a problem within the SHB nonlinear model. The SHB model looks as follows (for 2 mixture variables and 1 process variable):

$$c(x,z) = (b_1x_1 + b_2x_2 + b_{12}x_1x_2) * (a_0 + a_1z_1)$$

(Snee, Hoerl, and Bucci 2016) There are problems with interpreting the coefficient a_1 as it represents both the interaction of z and the mixture variables as well as the pure impact of z (i.e., the main effect of z). To try to solve this four models were tested. The first model adds cross product terms between process and mixture to g(z):

$$c(x,z) = (b_1x_1 + b_2x_2 + b_{12}x_{12}) * (a_0 + a_1z_1 + a_2z_2 + a_{12}z_1z_2 + (ab)_{11}z_1x_1 + (ab)_{12}z_1x_2 +$$

$$(ab)_{21}z_2x_1 + (ab)_{22}z_2x_2)$$

The second model adds x terms that are crossed with higher order z terms:

$$c(x,z) = (b_1x_1 + b_2x_2 + b_{12}x_{12}) * (a_0 + a_1z_1 + a_2z_2 + a_{12}z_1z_2 + (ab)_{11}z_1x_1 + (ab)_{12}z_1x_2 +$$

$$(ab)_{21}z_2x_1 + (ab)_{22}z_2x_2 + (ab)_{121}z_1z_2x_1 + (ab)_{122}z_1z_2x_2)$$

The third model removes terms that purely deal with z from the second equation:

$$c(x,z) = (b_1x_1 + b_2x_2 + b_{12}x_{12}) * (a_0 + (ab)_{11}z_1x_1 + (ab)_{12}z_1x_2 + (ab)_{21}z_2x_1 + (ab)_{22}z_2x_2$$

$$+ (ab)_{121}z_1z_2x_1 + (ab)_{122}z_1z_2x_2)$$

The fourth model deletes the main factor z terms from the first equation in g(z):

$$c(x,z) = (b_1x_1 + b_2x_2 + b_{12}x_{12}) * (a_0 + a_{12}z_1z_2 + (ab)_{11}z_1x_1 + (ab)_{12}z_1x_2 + (ab)_{21}z_2x_1 + (ab)_{22}z_2x_2)$$

The goal in looking at these equations is to find greater flexibility to fit the data. Multiple data sets were used to compare these 4 models, and the RMSE was used to compare the models. (Zhong and Hoerl 2019)

After comparing these models with each other with 6 data sets, Zhong saw that models 2 and 3 are identical and models 1 and 4 are identical. Model 2 is the overparameterized version of 3 and 4 is the overparameterized version of 1. Models 2 and 3 result in the lower RMSE, and since 2 is overparameterized, Zhong considers model 3 to be the best of the 4 proposed models. Next the third model was compared to the existing nonlinear model (SHB nonlinear), the fully linearized model, and the additive linear model. The same 6 data sets- Cornell's Fish Patty Data, Prescott's Bread data, Wang's Ice Cube Data, Cornell's Brazil Trees Data, Atenolol Data, and Twist's Disintegration Data- were used to analyze how well the models fit the data. After running these models on all the data sets, the RMSE was lowest for the fully linearized model for 3 of the 6 sets. This would be expected given that this model is the "gold standard". The other 3 data sets had the smallest RMSE with the newly added third model. It is also important to note that the linear additive model does very poorly for many of these data sets,

indicating interaction between the process and mixture variables in the data set. (Zhong and Hoerl 2019)

In all these situations the designs can get quite large. Thus it is important to have a strategy to make the designs smaller. One such strategy is fractionation. One approach to fractionation belongs to Cornell and this entails running a fractional process design at each point in the formulation. This is a useful idea as the process variable designs tend to cleanly fractionate. Another option is to run a full process design at certain points in the formulation design and a portion of the process design at other points of the design. An example of this would be running the full process design at the centroid, running one half at the pure blends, and the other half at the 50-50 blends, which can be seen in Figure 7. In this figure it can be seen that the process variable design fractions switch between pure blends and the 50-50 blends, this is to ensure that all terms will be estimated. One concern with this approach is what happens if there is interaction and not enough terms. One way to avoid this is to use a sequential strategy that allows for smaller designs at first while still keeping the possibility of larger designs. (Snee and Hoerl 2016)

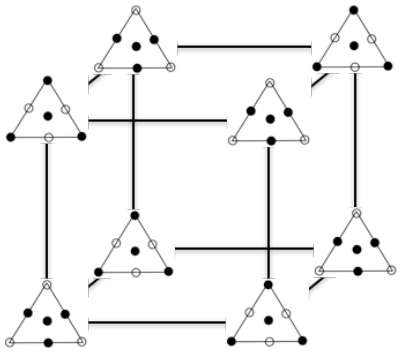

Figure 7: Figure 9.3 from Strategies for Formulations Development (Snee and Hoerl 2016)

Snee, Hoerl, and Bucci (2016) pieced together many of these concepts to come up with a strategy for designing and modeling when process and mixture variables are present. They focused on the additive linear, linearized, and their nonlinear mode, which we will call the SHB nonlinear model. First run a fractional design and fit both the additive linear model and the SHB

nonlinear model. Since only part of the full design is run, there won't be enough degrees of freedom to estimate all the terms in the linearized model. Note that if the SHB nonlinear performs better than the additive linear model there is noteworthy interaction between mixture and process variables. Evaluate these models using statistics like the RMSE and the residual plots. If these models are not strong enough then run the other part of the design that has not been run. This will result in a complete design. Again try the different types of model until reaching once that is up to standards. However now it is important to mention that these models from the full design will need to include a degree of freedom to block for time, i.e. the first run versus the second run.

A key question that exists through this research is how to best approach a problem when there is likely interaction between mixture and process variables, but we are not confident in this. This problem becomes a statistical engineering problem whose key is the integration of multiple tools. Figure 8 shows a flow chart of this strategy. We see that the first step is to run a fractional factorial design with the goal of reducing the amount of experimentation needed to fit the model. After running this design fit both the linear additive and non-linear models to this data. If the non-linear model fits better than it is likely that there is important interaction between the mixture and process variables. Then, intensely evaluate both model adequacy using the typical strategies like plotting the residuals. If one is not satisfied with the first fraction then follow up by running the other part of the full design. This allows for the full linearized model to be estimated if needed with the only downfall of losing a degree of freedom to blocking time (the two fractions). When evaluating the model's adequacy one metric often used is the RMSE, but in some cases for the full linearized model there are no available degrees of freedom for error. Thus, something called Lenth's method must be used. This identifies a subset of the important terms while dropping the rest. This frees up degrees of freedom from the dropped terms to estimate a pseudo RMSE. Just remember when comparing this to other RMSEs that this is only an estimate. (Snee and Hoerl 2016)

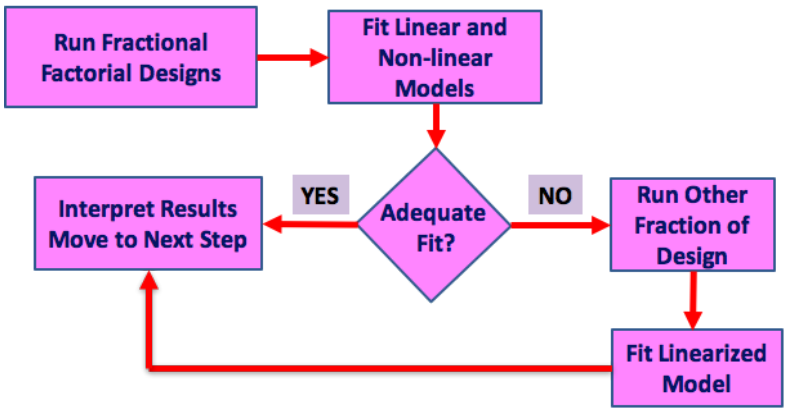

Figure 8: Figure 9.4 from Strategies for Formulations Development (Snee and Hoerl 2016)

One important conclusion from this process comes from the Prescott bread data. This data looked at bread volume with 3 different types of flour as well as poofing time and mixture time. It is important to note that each type of flour was constrained, thus pseudo components were used to utilize the simplex design. When Prescott did this experiment, he used a 90 run design to estimate 42 terms. All three types of models were looked at by Snee, Hoerl, and Bucci, and the resulting RMSE were compared. The linearized model was the best with the smallest RMSE, but the SHB nonlinear model resulted in only a slightly smaller RMSE (21.0 for linearized compared to 23.1 for SHB nonlinear). Then a fractional design was created in order to help look at how well the model predicted out of sample and to see if a fractional design can be a potential alternative to the full design. The fractional design used for this data set had an entire 3^2 design run at the centroid, points where $z_1z_2=0$ run at pure blends, and points where $z_1=z_2=0$ or $z_1z_2=+/-1$ run at the edge points; (⅔, ⅓, 0), (0, ⅔,⅓), (⅓,0,⅔). The other fractional design would involve flipping the points run at the edge points and pure blends. Either design results in a 39 run design, reducing the size of the original design by more than half. (Snee, Hoerl, and Bucci 2016)

When looking at the resulting RMSE, again the linearized model did the best, and it even had a lower RMSE than when a linearized model was used for the full design. However this can be explained by the fact that Prescott determined the parameters needed in the

linearized model from the full data set. Regardless, the linearized model RMSE for the hold out data was very high indicating that the model would be poor at predicting values outside of the sample. After examining the RMSEs for both the additive linear and SHB nonlinear models with the fractional design, it became clear that these models can serve as valid alternatives to the linearized model. The RMSEs were reasonably small and reasonably close to the RMSE of the linearized model, and the RMSEs for the hold out sample were smaller than the corresponding RMSE for the linearized model. Thus, indicating that these models may not fit the data as well, but will likely be better at predicting out of sample values. Overall, the fractional design can be used as a cost-reducing alternative in this experiment to running the entire design. (Snee, Hoerl, and Bucci 2016)

Zhong also continued her research by evaluating her new model with fractionated designs. Three of the data sets were also fractionated to test how well the models predicted out of sample. This revealed that the new nonlinear model did very poorly when predicting out of sample. Combining this fact with the fact that it did well in fitting the original (training) data indicates that there is likely overfitting in the new model. The severity of this increases as the number of terms in the model increases. One potential cause of the overfitting is there are not enough degrees of freedom for RMSE since there are a large number of terms. One way to solve this problem is to delete terms from the third equation that are statistically insignificant, that is, they have small t ratios. Another possible cause of the overfitting is multicollinearity. To solve this problem one could delete terms with high correlations while focusing on those terms with smaller t ratios. It is important to note that as these solutions decrease the hold out RMSE, the within RMSE will increase. Thus it is important to find a balance. Overall, this new model can be considered among the existing models as a potential solution to modeling data with interaction between process and mixture variables. (Zhong and Hoerl 2016)

This research and these ideas for designing and modeling when there are both process and mixture variables serve as a point of comparison when trying to find the best ways to model

this type of data. The fractions discussed in this research are systematically chosen with the purpose of being able to model part of the data and predict the rest with as little evidence of overfitting as possible. This leads to another interesting comparison between how these fractions perform when compared to models created from data samples that are bootstrapped rather than systematically chosen. These samples will be random and have the potential to include the same data point more than once. Looking at these different models with the different types of data sets and subsets of them results in a better understanding of when to use each type of model when trying to model process and mixture variables as well as their interaction. This is due to the fact that bootstrapping will result in multiple comparisons rather than just two fractions.

Methodology:

In order to compare a variety of models both in full and with fractions, a few data sets with different qualities is an important starting point. Three data sets were used for comparison. These are Cornell's fish patty data, Prescott's bread data, and Wang's melting data. Cornell' fish patty data has three mixture variables, the types of fish in the patty, which are mullet, sheepshead, and croaker. This data set also has 3 process variables that have to do with how each patty is cooked, which are baking times, baking temperatures, and frying times. The response variable is the texture of the patties. The design used for the mixture variables in this data set is a seven-run simplex that includes pure blends, 50-50 blends, and the centroid while the design used for the process variables is an eight run 2^3 design. This results in a base mixture model of

$$f(x) = b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + b_{123}x_1x_2x_3$$

and a base model for the process variables of

$$g(z) = a_0 + a_1z_1 + a_2z_2 + a_3z_3 + a_{12}z_1z_2 + a_{13}z_1z_3 + a_{23}z_2z_3 + a_{123}z_1z_2z_3$$

(Snee, Hoerl, and Bucci 2016)

Prescott's bread data has 3 mixture variables that represent the types of flours used in the bread, which are Tjalve, Folke, and Norwegian. This data set also has 2 process variables, which are proofing time and mixing time. The response variable is bread loaf volume. One interesting feature of this data set is that there were constraints on the types of flour, that is, $0.25 \leq x_1 \leq 1$, $0 \leq x_2 \leq 0.75$, and $0 \leq x_3 \leq 0.75$, where Tjalve is x_1, Folke is x_2, and Norwegian is x_3. Thus, a pseudo component design was used for the mixture variables, which was a 10 run simplex lattice design. A 9-run 3^2 design was used for the process variables. This leads to the same base mixture model as Cornell's fish patty data, but the resulting base model of the process variables is as followed

$$g(z) = a_0 + a_1 z_1 + a_2 z_2 + a_{12} z_1 z_2 + a_{11} z_1^2 + a_{22} z_2^2$$

After considering several models, Prescott came to the conclusion that a reduced model would be the ideal final model. That is,

$$y = 522.8x_1 + 448.1x_2 + 599.3x_3 + 13.0x_1 z_1 + 1.7x_2 z_1 + 54.3x_3 z_1 + 56.3x_1 z_2 +$$

$$37.2x_2 z_2 + 73.8x_3 z_2 - 39.4x_1 z_1^2 + 3.7x_2 z_1^2 - 46x_3 z_1^2 - 10.2x_1 z_2^2 + 28.4x_2 z_2^2 + 1x_3 z_2^2$$

When looking at this data set both the reduced model and the full model were looked at separately. For each model run with the Prescott data, one model was run with every term included while another model was run only including terms that stem from the reduced model above. (Snee, Hoerl, and Bucci 2016)

Wang's melting data has 3 mixture variables as well. These are water, milk, and juice. This data set includes 2 process variables that are the amount of sugar and type of milk used. This data has a response variable of melting time of the ice produced. The mixture design for Wang's data includes pure blends, 50-50 blends, checkpoint, and the centroid, and the process design is a 2^2 design. These result in the same base mixture model as both the Cornell and Prescott data sets and the base model for process variables of

$$g(z) = a_0 + a_1 z_1 + a_2 z_2 + a_{12} z_1 z_2$$

(Snee, Hoerl, and Bucci 2016)

For each data set each model of interest was run on the full data to see how well each model would fit the data by looking at the RMSE. The additive linear model, which does not account for interaction between the mixture variables and the process variables is

$$c(x, z) = f(x) + g(z)$$

where the two base models are simply added together. The fully linearized model, which based on previous research is the best at modeling data is

$$c(x, z) = f(x) * g(z)$$

where each term in the base mixture model is multiplied by each term in the base model for process variables. For example with Wang's melting data the fully linearized model would be

$$c(x, z) = b_1 x_1 * a_0 + b_1 x_1 * a_1 z_1 + b_1 x_1 * a_2 z_2 + b_1 x_1 * a_{12} z_1 z_2 + b_2 x_2 * a_0 + \ldots + b_{123} x_1 x_2 x_3 * a_{12} z_1 z_2$$

This model can get quite large and thus for the Cornell data the number of terms in the model equals the number of runs in the design. Therefore there are no degrees of freedom left to calculate RMSE. In order to get a calculation here Lenth's pseudo standard error was used for this model's standard error estimation. These 2 models are the earliest ideas of how to deal with modeling mixture variables, process variables, and their interactions. (Snee, Hoerl, and Bucci 2016)

Another approach to modeling these variables and their interactions is the model suggested by Kolwalski, Cornell, and Vining (2000). They suggest a method of a multiplicative model that is similar to the fully linearized model. However, they suggest eliminating higher factor interaction terms, that is, terms with more than two factors should be eliminated. This results in a model with fewer terms, with the hope of still capturing the data as in general the terms with higher factor interactions are less significant. In general this model is

$$c(x,z) = \sum_{i=1}^{q} b_i x_i + \sum\sum_{i<j}^{q} b_{ij} x_i x_j + \sum_{k=1}^{n} a_{kk} z_k^2 + \sum\sum_{k<l}^{n} a_{kl} z_k z_l + \sum_{i=1}^{q}\sum_{k=1}^{n} ab_{ik} x_i z_k$$

This model for the the Wang data looks as follows

$$c(x,z) = b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + a_{12} z_1 z_2 + ab_{11} x_1 z_1 + ab_{12} x_1 z_2 + ab_{21} x_2 z_1 + ab_{22} x_2 z_2 + ab_{31} x_3 z_1 + ab_{32} x_3 z_2$$

The other approach to modeling when there is expected interaction between the mixture and process variables is using some type of nonlinear model. Two main nonlinear models were examined. First was the SHB nonlinear model (Snee, Hoerl, and Bucci 2016). This nonlinear model finds $c(x,z) = f(x) * g(x)$ directly, rather than multiplying terms out. For example, this would look as follows for the Wang data

$$c(x,z) = (b_1 x_1 + b_2 x_2 + b_3 x_3 + b_{12} x_1 x_2 + b_{13} x_1 x_3 + b_{23} x_2 x_3 + b_{123} x_1 x_2 x_3) * (a_0 + a_1 z_1 + a_2 z_2 + a_{12} z_1 z_2)$$

In order to try to obtain a unique least squares solution, a_0 must be set to a non-zero constant. For ease and consistency, a_0 is set to 1 for each model. Zhong (and Hoerl 2019) also did work with nonlinear models. Her model 3 is used here as another nonlinear model for comparison. This nonlinear model for Wang's data is

$$c(x,z) = (b_1 x_1 + b_2 x_2 + b_{12} x_{12}) * (a_0 + (ab)_{11} z_1 x_1 + (ab)_{12} z_1 x_2 + (ab)_{21} z_2 x_1 + (ab)_{22} z_2 x_2 + (ab)_{121} z_1 z_2 x_1 + (ab)_{122} z_1 z_2 x_2)$$

A third nonlinear model was considered as well. This model took Zhong's third model (above) and eliminated all the terms that had to do with one of the mixture variables. As the mixture variables sum to one, the one variable can be determined by the other. This model was only run on the Cornell data to observe how it did in comparison to the other models.

Regardless of which model is used for which data set, using the full design to fit the model only results in a single RMSE of how well the model fits the data. One way to get another

RMSE is to fractionate the data set, that is, run the model on a proportion of the data set. By fractionating the data, there are points remaining that can be used to evaluate the model based on how well the model can predict this hold-out sample. Each data set has a systematic way of separating the data. For the Cornell data set one fractional design includes every pure blend where $z_1z_2z_3$=1, every 50-50 blend where $z_1z_2z_3$=-1, and every centroid is included. This results in 32 points. This is enough to run all the models except the fully linearized model. It is also important to run both fractions, that is, also run the fractional design where $z_1z_2z_3$=-1 at all the pure blends and $z_1z_2z_3$=1 at the 50-50 blends. One of Prescott's fractional designs includes all the full 3^2 design at the centroid, the pure blends where z_1z_2=0, and the points where z_1=z_2=0 as well as at the edge points (⅔,⅓,0), (0,⅔, ⅓), and (⅓, 0, ⅔) where z_1z_2=±1. Again be sure to reverse the pure blends and edge points to obtain the second fraction. The fractional design was also too small to run the fully linearized model on the full Prescott data, but it was large enough to run all the other models for the full Prescott models as well as all the reduced models. Wang's fractional design is obtained by running the full 2^2 design at the centroid and at pure blends as well as including all the 50-50 edge points where z_1z_2=1 and all the checkpoints where z_1z_2=-1, ensuring to run the other fraction reserving the 50-50 edge points and the checkpoints. (Snee, Hoerl, and Bucci 2016)

These fractions allowed for the evaluation of the models, but they were all systematically chosen, raising the question of what would happen if these fractions were chosen more at random. In order to answer this question bootstrapping is introduced to create samples. By bootstrapping, not only is the resulting sample random, but it can include replicated points, something that is not included in the fractional designs. Bootstrapping also allows samples to be generated quickly. Thus, give the ability to run multiple bootstrapped samples for each model and find the average RMSE for both in and out of sample. First the bootstrapped samples for each model had the same proportion of points that were in the fractional designs for each data set, that is, 32 for Cornell, 39 for Prescott, and 28 for Wang. The number of bootstraps was kept

consistent at 500 replications and the resulting in and out of sample RMSEs were each averaged. However it is important to know that each bootstrapped sample can result in a different size sample for the data points not included in the original sample due to sampling with replacement. This fact also allowed for a fractional design that was proportional to the size of the full design, that is, for Cornell 56 points, for Prescott 90 points, and for Wang 40 points. All the models were run on both the types of bootstrapped samples that resulted in both an average in sample and out of sample RMSE for each model.

Results:

Regardless of which data set that is being modeled, it is expected in general that the linearized model, which is the standard, will be one of the best options. However, using this model is not always feasible due to its size. Thus, trying to find alternative models is important. In general, as expected, the linearized model was one of the best models for the full version of each data set, as it had one of the lowest RMSE. However, the SHB nonlinear model had a lower RMSE for the fish patty data than the fully linearized model. The fully linearized model of the fish patty data had exactly as many terms in the model as degrees of freedom. Thus, there were no degrees of freedom left to estimate RMSE, so Lenth's method is needed in order to calculate the RMSE. This fact indicates that the linearized model may not be the best if it means there are no degrees of freedom left to calculate RMSE. Thus, for data sets that result in this situation, I would recommend using a different model with fewer terms. This situation is also seen in the Wang data set when using a fractional design. Again for the fully linearized model, Lenth's method was needed to calculate both the in-sample RMSE. For comparison to the full design fully linearized model of the fish patty data, we will focus on the in-sample RMSEs. For both fractional designs, the RMSE (in sample only) is lowest for the SHB nonlinear model of that data. Combining this with the fish patty results indicates that when the data set requires Lenth's method for the linearized model it would be valuable to run a SHB nonlinear model of the data for a better fit.

If one is looking for a general alternative to the linearized model for a full data set, it is also indicated by the five that were evaluated here to start with a nonlinear model, particularly Zhong's suggested nonlinear model. Regardless of which data set is being studied this model results in the second smallest (with the exception of the fish patty, discussed above) RMSE indicating the second best fit. The additive linear model does not appear to be a great alternative to the fully linearized model, which makes sense as this model does not account for the interaction between the mixture components and process variables. However since most of the time the additive linear model is the smallest (has the least amount of terms), if one is unsure if their data set has interaction between the mixture and process variables, it could be valuable to run the additive linear model. If one is sure based on subject matter knowledge that there is interaction between these two types of variables, then I would recommend not running the additive linear model unless one would like to compare the linear additive fit with the nonlinear fit.

The SHB nonlinear model also appears to be a valuable option for a smaller model. The RMSEs for each data set is not significantly larger than linearized or Zhong's nonlinear model. The SHB nonlinear model can also be much smaller than Zhong's nonlinear model depending on the number of variables in the data sets. The KCV models seem to result in a fit somewhere between the nonlinear models and the additive linear models in regards to modeling a full data set. With this in mind, I would not use this model as my first alternative to the fully linearized model, but I would consider it if I was not satisfied with the nonlinear models. This is because in the four data sets that have been studied, there is one where the KCV does better than the SHB nonlinear model and very similarly to Zhong's nonlinear model. This is for the full Prescott data, which is interesting as Prescott not only didn't look at this model, but he did not use as many terms in his original models. I question how and why Prescott came up with the reduction he did as across the board each model run with the reduced number of terms resulted in a worse fitting

model. The majority of the models' RMSEs were larger for the reduced model indicating that the model is not as strong as the models for the full models.

	Linearized	Additive Linear	SHB Nonlinear	KCV	Zhong's Nonlinear
Fish Patty	0.1735124	0.243215	0.1571212	0.176387	0.1376118
Prescott Bread- Reduced	21.04146	27.3367	23.134817	26.74669	20.991646
Prescott Bread-Full	20.49767	24.36848	23.084599	21.60752	21.403988
Wang- Melting	1.35763	2.377313	2.0512868	2.324935	1.7745609

Table 1: Resulting RMSEs of the full data set

In order to reduce the size of the design, using a fractional design is helpful. This raises a question of which types of model are the best to use to both fit the data and predict out of sample. Since the entire design is no longer being used, there is a new ability to predict the points that are not being used in the model. When looking at the in sample RMSEs the smallest are mainly from the models that gave the lowest RMSE of the whole data set. That is, when possible the fully linearized model tends to fit the data the best. Both nonlinear models typically have smaller in sample RMSEs as well. However for Zhong's nonlinear model the fish patty data set resulted in RMSEs, both in and out of sample, that didn't converge. This leads to some hesitation with this model. The KCV model appears to do better in the fractional data sets than in the full data set. This means that the KCV is an even more viable model for data fitting when using the fractional designs.

Fish Patty				
	Full	Fraction 1	Fraction 2	Bootstrapped(average of 50 replications)
Linearized (in)	0.1735124	NA	NA	NA
Linearized (out)	NA	NA	NA	NA
Additive Linear (in)	0.243215	0.272191	0.241152	0.2102847
Additive Linear (out)	NA	0.25275	0.2787393	0.3087357

SHB Nonlinear (in)	0.1571212	0.1571804	0.1643394	0.1286442
SHB Nonlinear(out)	NA	0.2215759	0.2174062	0.2273087
KCV(in)	0.176387	0.166781	0.175383	0.1347747
KCV(out)	NA	0.235781	0.235915	0.2825416
Zhong Nonlinear (in)	0.1376118	NA	NA	0.2129316*
Zhong Nonlinear (out)	NA	NA	NA	3900.37*

Table 2: Resulting RMSEs for Cornell's Fish Patty data

Prescott Bread (full)				
	Full	Fraction 1	Fraction 2	Bootstrapped(average of 50 replications)
Linearized (in)	20.49767	NA	NA	14.10342*
Linearized (out)	NA	NA	NA	47.7697*
Additive Linear (in)	24.36848	24.67436	23.18139	22.6761
Additive Linear (out)	NA	20.5303	31.2386096	27.63029
SHB Nonlinear (in)	23.084599	23.104138	21.680231	21.8397
SHB Nonlinear(out)	NA	25.280685	30.016727	25.05579
KCV(in)	21.60752	21.6561	19.48697	19.64363
KCV(out)	NA	25.34705	28.85494	25.75777
Zhong Nonlinear(in)	21.403988	18.13025	15.811674	17.96013*
Zhong Nonlinear(out)	NA	29.55676	33.481127	28.6293*

Table 3: Resulting RMSEs for Prescott's bread data- full

Prescott Bread (reduced)				
	Full	Fraction 1	Fraction 2	Bootstrapped(average of 50 replications)
Linearized (in)	21.04146	20.32851	17.2164	19.03954
Linearized (out)	NA	27.18814668	30.2583	26.31582
Additive Linear (in)	27.3367	24.89664	23.93275	23.33573
Additive Linear (out)	NA	25.310381	21.6788	26.20833
SHB Nonlinear (in)	23.134817	23.632421	22.572241	22.18256
SHB Nonlinear(out)	NA	24.37375	25.901494	25.01325
KCV(in)	26.74669	27.82497	30.26884	25.37099
KCV(out)	NA	28.0344606	27.622992	29.17136

Zhong Nonlinear (in)	20.991646	20.369677	17.011075	19.23087
Zhong Nonlinear(out)	NA	26.795825	30.779495	24.72038

Table 4: Resulting RMSEs for Prescott's bread data -reduced

Wang Melting				
	Full	Fraction 1	Fraction 2	Bootstrapped(average of 50 replications)
Linearized (in)	1.357363	2.188401	3.964085	0.3264438*
Linearized (out)	NA	3.460649	4.0191621	568385.9*
Additive Linear (in)	2.377313	2.468037	2.664512	2.052332
Additive Linear (out)	NA	2.93376265	2.30349897	3.018968
SHB Nonlinear (in)	2.0512868	2.0185971	2.232527	1.690962
SHB Nonlinear(out)	NA	3.0144979	2.4363005	2.797079
KCV(in)	2.324935	2.364035	2.878017	1.701689
KCV(out)	NA	2.6738147	1.5460833	3.829843
Zhong Nonlinear (in)	1.7745609	1.4859205	1.9637178	1.72237*
Zhong Nonlinear (out)	NA	2.8020316	2.287202	4.52699*

Table 5: Resulting RMSEs for Wang's melting data

For the fully linearized models and Zhong's nonlinear models, the RMSE of the out of sample values is much larger than the in sample RMSE as seen in Tables 2-5. This is an indication of overfitting. It appears that certain data sets are more susceptible to overfitting as for the fish patty data set both the SHB nonlinear and KCV models demonstrate RMSEs that indicate possible overfitting. This trend also follows for the Prescott data when using all possible terms of the model. Interestingly these two data sets are the two data sets where there was not enough data to run a fully linearized model. Regardless of the overfitting, the only other model that was looked at was the additive linear model, and it does not appear to fit the data as well for reasons that have been previously mentioned. That said, I think trying other types of models for these data sets to see if they also result in overfitting would be very useful. In general, I would recommend a similar strategy for modeling these fractional designs as the full designs. If one knows by subject matter knowledge that there are interactions between mixture and process variables, the additive model can be surpassed. In terms of picking an alternative for

the fully linearized model, the SHB nonlinear model seems to generally be the next best option when you are considering both fitting the data and prediction. However, the KCV model appears to fit the data better than the SHB nonlinear model, but there is more evidence of overfitting. Thus, if prediction is of less importance I would consider the KCV a viable alternative to the linearized model. It is still important to be careful as this model tends to have some overfitting problems.

After observing the severity of overfitting present in Zhong's suggested nonlinear model another nonlinear model was attempted. This model was created based on Zhong's original model. It is a reduction of her model, given the properties that mixture variables have, that is, that they sum to 1, the reduction was eliminating every term with one of the mixture variables in it. Since each mixture variable can be determined by calculating 1- the sum of the other mixture variables, we wanted to see if reducing Zhong's model would eliminate the overfitting and non convergence issues we have. To check this the fish patty data was used first. A similar model was tried three times, each one eliminating a different term, but in all three cases the fractional models would not converge. This led to the decision to not go further with this model as it is unlikely that this model would solve these problems.

There were a few general problems that emerged when looking at the RMSEs that resulted from the original bootstrapped samples that were proportional to the fractional designs. First is for all the linear models there is significant evidence of overfitting. The linearized models have some ridiculously large out of sample RMSEs such as 253.52 for the full Prescott data, which is an indicator of extreme overfitting. This indicates to me that these models are not valuable due to the overfitting. When looking at both the KCV and linear additive models for all the data sets at first glance it appears that the bootstrapped models are better than the original fractional ones as the lowest in sample RMSEs belong to the bootstrapped model in each data set. However when looking closer a small problem emerges, that is, that there is more evidence of overfitting in each of the bootstrapped models. This is indicated by larger out of sample

RMSEs. In almost every data set for both the linear additive and KCV models, the bootstrapped model has the largest out of sample RMSE while having the lowest in sample RMSE. This means that the bootstrapped models are fitting the initial data better, but not doing as well at predicting out of sample. This indicates overfitting. One thing that could be causing the overfitting is that for most of these models there is a warning message from R. This message indicates that some of these predictions may be misleading due to a rank-deficiency. Regardless of this, it appears that overall the original fractions are better suited for modeling these data sets as there is more overfitting for the bootstrapped samples.

Another general problem that is seen with this size bootstrapped sample and number of replicates is that the vast majority of nonlinear models do not converge. The only nonlinear model that ran properly with these parameters is the SHB nonlinear reduced Prescott model. This model resulted in very similar results to one of the two original fractional models for this data set. However, this bootstrapped model did show more evidence of overfitting than both of the original fractional models for this data set. The rest of the nonlinear models, that is, all the other data sets for the SHB nonlinear models and all the data sets with new model 1, resulted in an error where R could not run the model due to lack of convergence. This could be caused by an abnormal bootstrapped sample that does not support the model being estimated, so moving forward changing the size of the sample and number of samples run could result in the nonlinear models running.

In order to try to solve this problem the sample size was changed. This time it was changed to have the size of the sample match the size of the original data set. After trying this some of the nonlinear models were still struggling to converge so the starting values were set to the values of the coefficients in the original full model. This eliminated most of the problems within the nonlinear models. These changes also eliminated the majority of the rank-deficiency warnings from the nonlinear models. In order to ensure that each model in each data set produced a reasonable in and out of sample RMSE the number of replications was reduced to

50. This was the largest amount of replications where each model would run and produce a reasonable in sample RMSE. A few models still had a few warnings, particularly for the linearized model and Zhong's nonlinear model, and these are indicated in Tables 2-5 by a star. The lack of errors indicate that these parameters are likely more successful than the previous parameters for bootstrapping.

Bootstrapping with the same proportion to the full data set resulted in in-sample RMSEs that were among the lowest for each model and data set when compared to both the systematically chosen fractions and the full data set. However, when you compare these in-sample RMSEs to the out of sample RMSEs there appears to be potential problems with overfitting. Two models appear to have worse overfitting issues than the others. These are the two models with the largest number of terms in general, the linearized model and Zhong's nonlinear model. When looking closely at the out of sample RMSEs of these models, it is clear that some of these models when run on the bootstrapped samples have such extreme overfitting that I would likely not want to consider the model as an accurate representation of trends about the data set. A few of these examples are the linearized model for the melting data and Zhong's nonlinear model for the fish patty data. Both of the out of sample RMSEs are in the thousands, which is very extreme overfitting. The other three models, the additive linear, SHB nonlinear model, and the KCV model, have varying evidence of overfitting throughout the different data sets as seen in Tables 2-5.

Most of these other models have evidence of overfitting for each data set, but for two of the data sets this evidence is not very strong. These are both of the Prescott bread data sets seen in Tables 3 and 4. In these data sets the in sample RMSEs of the three models in question are the smallest when compared to the systematically chosen fractions and the full data set. This fact alone would indicate that the models run on the bootstrapped samples fit the data the best. However it is also important to see how the models that are run on portions of the data set predict the remaining data points. This is where an interesting pattern emerges. The out of

sample RMSEs are not always the largest for the bootstrapped samples. This means that the evidence of overfitting for the bootstrapped samples is weaker than for the fractions. This is not consistent for every model and every fraction, but there are multiple situations where this is the case including at least one fraction for the SHB nonlinear model for both data sets and at least one fraction for both the additive linear and KCV models for the full Prescott data set. This would indicate that the choice of the fraction for this data set may not be as important. However it is also important to note that this data set is the largest. Thus, the size of the bootstrapped sample is also the largest.

The other two data sets, which have smaller bootstrapped samples, reveal that there are more overfitting problems for the models resulting from the bootstrapped samples than the systematically chosen fractions. In both data sets, the fish patty and the melting, the difference between the in sample and out of sample RMSEs are the largest for the bootstrapped samples as seen in Tables 2 and 4. This is true for each of the three models that were being compared in the above paragraph. This large difference indicates that there is the strongest evidence for overfitting in these models run on the bootstrapped samples. This result contrasts the result from the Prescott data sets, that is, that systematic choice of the fraction is important. Since these two groups of data sets show contrasting results, there must be differences between the data sets that account for these differences. A few differences in terms of properties of the data sets that could cause some of these differences include the fact that the bread data has further constraints on the mixture variables and the bread data also has more levels of its process variables than the other data sets as well as the large sample size. The first two of these facts change the shapes of the designs of the data set. Thus there is some evidence that the bootstrapping samples can be an adequate alternative to the full model and the systematic fractions for data sets that have similar properties to the Prescott bread data sets.

Regardless of the data set and the type of sample that is being modeled it appears that the SHB nonlinear model is a generally solid alternative to the fully linearized model. This can

be seen as not only does this model present the best alternative when modeling the full data set in general, but it also seems to perform well with the fractions. Even though the KCV model performs better on some of the fractions in terms of fitting a model, I still think that overall the SHB nonlinear model is the best alternative. This is because of how each model does with the out of sample RMSEs. The KCV models tend to have higher out of sample RMSEs, which is a strong indicator that the models are overfitting the data. There is less evidence of this in the SHB nonlinear models. The out of sample RMSEs tend to be closer to the in sample RMSEs, which shows that there is less evidence of overfitting. Overall, the SHB seems to be the strongest alternative to the fully linearized model.

Key Conclusions:

When comparing the KCV model in all the data sets with the full data set there is some initial thought that it could be considered a viable alternative to the linearized model even though the RMSEs are not as small as for the linearized and nonlinear models. This thought is heightened after evaluating the KCV model with systematically chosen fractions of the data sets. This model appears to show strong evidence of overfitting for two of the four data sets. These are the fish patty data and the full bread data set. However, there is very minimal evidence of overfitting with this model for the other two data sets, and the in sample RMSEs for these data sets are only slightly larger than the RMSE for the full data set. These factors indicate that the fractional design with a KCV model may be a viable alternative to the full design with a KCV model. It is important to note that the concern for overfitting occurs in the data sets that have more terms. Thus, I would only recommend trying the KCV on smaller data sets.

Bootstrapping is random in its nature and allows for replication due to the sampling with replacement. Thus, a sample size of proportion size to the original sample is possible. Looking at a larger sample size as well as providing the correct coefficients as starting points for the nonlinear models allowed for the majority of the models to run without error or warning.

However, a few models still resulted in RMSEs that indicated that these models are not ideal for modeling the data due to the issues with predicting out of sample. Particularly the linearized models and Zhong's nonlinear models have this problem. When observing the in sample RMSEs for each data set of these models, it is almost always smaller than the RMSE for the same model when run on the corresponding full data set. This alone would indicate a strong fit, but the out of sample RMSEs reveal the major problem. Each of the out of sample RMSEs are significantly larger than the in sample RMSE. Some are even in the thousands. This is very strong evidence of overfitting. Thus, I would not recommend running these larger models on observational data.

The other three models, additive linear, KCV and the SHB nonlinear model, have less of this problem. However, for two of the data sets, the fish patty and the melting data sets, I still would not recommend running these models on a bootstrapped sample of this size due to strong evidence of overfitting. Both the Prescott data sets have limited evidence of overfitting for these three models. Even though the out of sample RMSE is larger than the in sample RMSE for each model with bootstrapping, there are multiple situations where the difference between the RMSEs is larger in one of the systematically chosen fractions. This is an indication that there is less evidence of overfitting in the model run on the bootstrapped sample than the model run on the systematically chosen fraction. The SHB nonlinear model is the only of the three models that does this for both data sets. Thus, using this nonlinear model on observational data of this size could be considered an adequate alternative to the systematically chosen sample for these data sets.

Since there are contrasts in results between the different data sets, it is possible that the reason is based on the differences between the data sets. This would mean that there are certain qualities that would make a data set better to be run with a bootstrapped sample than others. When looking at the Prescott data compared to the other two data sets the first thing that jumps out is that it has the most data. That is, the Prescott data has 90 points compared to 56

for Cornell's data and 40 for Wang's data. This means that the bootstrapped sample for the Prescott data was the largest. Thus, it is possible that the larger size of the data set results in better results. The Prescott data set also has a very different design space than the other two data sets. This is for two reasons. First the mixture variables have further constraints. This means that the basic design for the mixture variables is no longer a full simplex. The other reason is that the process variables have 3 levels, whereas the other data sets have process variables with 2 levels. This changes the basic design of the process variables. These unique properties of the Prescott data could indicate that the bootstrapped sample is an adequate alternative if using a data set with similar qualities. However to confirm this I would want to study more data sets with properties similar to the Prescott data to see if these trends hold up.

Overall, if I were to recommend a model as an alternative to the fully linearized model I would recommend the SHB nonlinear model. This model performs decently when compared to the other models when modeling the full data set. However what makes it more convincing as a solid alternative is the fact that this model seems to perform better than most regardless of the fraction or size of the data set. This means that this model is very robust, that is, it can handle data with many different qualities, and still produce a reasonable fit while still having confidence in the model's ability to predict. This is seen as in general the SHB model produces one of the smaller in sample RMSEs for both the systematically chosen fractions and the bootstrapped samples while still producing a smaller out of sample RMSEs simultaneously. This is one of the qualities that makes the SHB model better overall than say the KCV model as even though the KCV model has some smaller in sample RMSEs, the out of sample RMSEs indicate strong evidence of overfitting. Thus, in general I would recommend the SHB nonlinear model as the best alternative to the fully linearized model.

Opportunities for Further Research:

After seeing that the Prescott data, both the reduced and full models, performed better with the bootstrapped samples than the Cornell and Wang data sets, I would like to further

investigate the potential why for this. There are three properties that make the Prescott data sets different from the other two; the larger size of the data set, the constraint on the mixture variable, and the fact that the process variables have 3 levels. Since there are multiple things that are unique about this data set, if I were to continue this research further I would like to try to isolate each quality and try to determine if the trend continues. In order to do this I would need to find data sets of similar size as well as data sets of a smaller size. Then run the 3 models that seem to get somewhat reasonable in and out of sample RMSEs on bootstrapped samples, that is, the additive linear, the SHB nonlinear, and the KCV models. I would compare the two chunks of data sets to see if the RMSEs indicated a pattern based on the size of the data set. I would like to complete a similar process for data sets with higher levels of process variables (most likely compare 3 levels to 2 levels to keep the size of the models reasonable) as well as for data sets with further constraints on the mixture variables. By running these models with bootstrapped samples on multiple data sets with similar qualities I hope to determine if the pattern seen with the Prescott data holds for other data sets with similar properties. By trying to isolate each property I also hope to see if one of these properties leads to a better model with a bootstrapped sample than others.

Summary:

In order to try to find reasonable alternatives for a fully linearized model, we compared many existing models using 4 data sets- the Cornell Fish Patty Data, the full Prescott Bread Data, the reduced Prescott Bread Data, and Wang's Melting Data. These models include the additive linear model, the SHB nonlinear model, the KCV model, and Zhong's nonlinear model. We started by running all of these models as well as the fully linearized model on each full data set. This revealed that the two nonlinear models appear to be the best alternatives to the fully linearized model as both nonlinear models have the smallest RMSEs for a majority of the data sets. The KCV model did not appear to be as valuable as an alternative, but it in general did not

have as large a RMSE as the additive linear model. Thus it should still be in consideration as a viable alternative.

In order to get both an in sample and out of sample RMSE we then decided to run these models for fractional designs of these data sets. This would allow us to see how well the models do at prediction and would help ensure that the model does not overfit the data. We used a systematic method to choose which points were in each fraction and ensured to run the reverse of each fraction as well. Upon running the models on the fractional designs a few things become clear. First there appears to be significant evidence of overfitting with Zhong's nonlinear models. This is seen as the out of sample RMSEs are much larger than the in sample RMSEs in general. Thus, this model may not be as strong an alternative to the fully linearized model. The KCV model appears to be a stronger alternative when looking at the models run on fractional designs. However, there are some problems with overfitting with this model as well. For the larger data sets, the Cornell and full Prescott data sets, the out of sample RMSEs are large enough when compared to the in sample RMSEs to have concern for overfitting. With this in mind, we would be more confident considering the KCV model for smaller models and data sets.

The systematically chosen fractions only allowed for two data points to be observed. In order to look at more data points, we decided to use bootstrapping to create multiple samples to run the different models on. Bootstrapping is also random in nature, which means if these samples result in low RMSEs, then the choice of what is in the fraction is not important. After determining that there were a multitude of problems in running all the models on bootstrapped samples that are proportional to the systematic fractions, we decided to run the models on samples proportional to the size of the full data set. When looking at the resulting in and out of sample RMSEs there are a few issues that pop up immediately. These issues are very strong evidence of overfitting for both the fully linearized models and for Zhong's nonlinear models. Some of the out of sample RMSEs for these models are even in the thousands, which is

evidence of extreme overfitting. These models generally are the models with the most terms. Therefore, we would not recommend running these models on observational data.

The 3 remaining models (SHB nonlinear, KCV, and additive linear) demonstrate less evidence of overfitting as the difference between the in and out of sample RMSEs is smaller. However, only two of the data sets show minimal evidence of overfitting for these models. The resulting RMSEs from the Cornell and Wang data sets have strong evidence of overfitting. Both Prescott data sets show limited evidence of overfitting for these models. Even though the out of sample RMSE is always larger than the in sample RMSE for these models with the Prescott data sets, the difference between the RMSEs is sometimes smaller for the models run with bootstrapped samples than the systematically chosen fractions. This is an indication that there is less overfitting for the models run on the bootstrapped samples than the systematically chosen fractions. The model that seems to be the best alternative when using observational data, as this is the only model that has the small difference between in and out of sample RMSEs for both the full and reduced Prescott data sets.

It is also very interesting that there is significantly stronger evidence of overfitting in two of the data sets. There are some distinct properties that differ between the groups of data sets that could be the cause of the difference. One property is the size of the data set as the Prescott data set is the largest, so the size of the bootstrapped sample is also the largest. The other two properties impact the shape of the design space, and they are the fact that the mixture variables have added constraints and the fact that the process variables have 3 levels when the other data sets have process variables with 2 levels. It is unclear which of these reasons would result in the trend seen here and this is something that we would need to conduct further research to uncover.

References:

Kowalski, Scott, John Cornell, and Geoffrey Vining. “A new model and class of designs for mixture experiments with process variables.” *Communications in Statistics- Theory and Methods,* vol.29, no.9-10, 2000, pp.2255-2280. DOI: 10.1080/03610920008832606

Scheffe, Henry. “The Simplex-Centroid Design for Experiments with Mixtures.” *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 25, no.2, 1963, pp.235-263. https://www.jstor.org/stable/2984294

Snee, Ronald, Roger Hoerl, and Gabriella Bucci. “A statistical engineering strategy for mixture problems with process variables.” *Quality Engineering,* vol.28, no.3, 2016, pp.263-279.DOI: 10.1080/08982112.2015.1100733

Snee, Ronald, and Roger Hoerl. 2016. *Strategies for Formulations Development: A Step-by-Step Guide Using JMP.* Cary, NC:SAS Institute Inc.

Zhong, Mushan and Roger Hoerl. “Use of Nonlinear Models in Analyzing Experiments with Both Mixture and Process Variables.” 2019

Appendix A- Link to the R code for bootstrapping

https://www.dropbox.com/s/ilxqx28fuqu1de7/Bootstrapping%20models.Rmd?dl=0