Combining Adaptive and Designed Statistical Experimentation: Process Improvement, Data Classification, Experimental Optimization, and Model Building

by

Chad Ryan Foster

B.S., Colorado School of Mines (1998)
M.S., University of Texas at Austin (2000)

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

Author . . . . . .
Department of Mechanical Engineering
March 11, 2009

Certified by . . . . . .
Daniel D. Frey
Associate Professor
Thesis Supervisor

Accepted by . . . . . .
David E. Hardt
Chairman, Department Committee on Graduate Theses

Combining Adaptive and Designed Statistical Experimentation: Process Improvement, Data Classification, Experimental Optimization, and Model Building

by

Chad Ryan Foster

Submitted to the Department of Mechanical Engineering
on March 11, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Research interest in the use of adaptive experimentation has returned recently. This historic technique adapts and learns from each experimental run but requires quick runs and large effects. The basis of this renewed interest is to improve experimental response and it is supported by fast, deterministic computer experiments and better post-experiment data analysis. The unifying concept of this thesis is to present and evaluate new ways of using adaptive experimentation combined with the traditional statistical experiment. The first application uses an adaptive experiment as a preliminary step to a more traditional experimental design. This provides experimental redundancy as well as greater model robustness. The number of extra runs is minimal because some are common and yet both methods provide estimates of the best setting. The second use of adaptive experimentation is in evolutionary operation. During regular system operation small, nearly unnoticeable, variable changes can be used to improve production dynamically. If these small changes follow an adaptive procedure there is high likelihood of improvement and integrating into the larger process development. Outside of the experimentation framework the adaptive procedure is shown to combine with other procedures and yield benefit. Two examples used here are an unconstrained numerical optimization procedure as well as classification parameter selection.

The final area of new application is to create models that are a combination of an adaptive experiment with a traditional statistical experiment. Two distinct areas are examined, first, the use of the adaptive experiment to determine the covariance structure, and second, the direct incorporation of both data sets in an augmented model. Both of these applications are Bayesian with a heavy reliance on numerical computation and simulation to determine

the combined model. The two experiments investigated could be performed on the same physical or analytical model but are also extended to situations with different fidelity models. The potential for including non-analytical, even human, models is also discussed.

The evaluative portion of this thesis begins with an analytic foundation that outlines the usefulness as well as the limitations of the procedure. This is followed by a demonstration using a simulated model and finally specific examples are drawn from the literature and reworked using the method.

The utility of the final result is to provide a foundation to integrate adaptive experimentation with traditional designed experiments. Giving industrial practitioners a solid background and demonstrated foundation should help to codify this integration. The final procedures represent a minimal departure from current practice but represent significant modeling and analysis improvement.

Thesis Supervisor: Daniel D. Frey
Title: Associate Professor

Acknowledgements

This work was possible only with the guidance of my many mentors in and around MIT including Dan Frey, Maria Yang, Warren Seering, Chris Magee, Alex Slocum, David Rumpf, and Tony Patera, plus the lively discussion and interaction of my fellow researchers Nandan Sudarsanam, Xiang Li, Jagmeet Singh, Sidharth Rupani, Yiben Lin, Rajesh Jugulum, Hungjen Wang, Ben Powers, Konstantinos Katsikopoulos, Lawrence Neeley, Helen Tsai, and Jonathan Evans, as well as international correspondence with Monica Altamirano, Leslie Zachariah, Catherine Chiong-Meza, Michiel Houwing, Paulien Herder, and Pepijn Dejong and finally the financial support of MIT, NSF, MIT Pappalardo Fellowship, Cummins, GE, Landis, and, of course, the support and encouragement of Elise and the rest of my family.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Many experimental situations require a setup, tuning, or a variable importance decision before running a designed experiment. If this other procedure is run as an adaptive experiment there can be additional benefit to the subsequent designed experiment. The adaptive experiment of focus is the adaptive-One-Factor-at-a-Time (aOFAT) experiment described in Frey et al. (2003), to be combined with a number of different statistically designed experiments including fractional factorial, Box-Behnken, Plackett-Burman, and D-Optimal as well as other procedures including evolutionary operation, article classification, and unconstrained optimization. The hypothesis is that there is an appropriate and beneficial place within designed experimentation to combine an adaptive experiment with a traditional statistical experiment.

Design-of-experiments (DOE) is a frequently used tool to understand and improve a system. The experimental technique began as support for long-term agricultural projects that allowed the development of methods such as blocking, randomization, replication, and fractional factorial analysis (Box et al., 2005). Many of these practices are considered fundamental to good experimentation, and are widely used today. The next advancement

to experimentation were achieved by industrial practitioners. In the chemical and manufacturing industries experiments ran more quickly, but were still expensive. Sequential experimentation specifically designed for regression analysis became the standard. The experiment was tied to a particular underlying physical model and could accurately estimate the required model parameters with minumum excessive runs. In current experimentation research design parameters are separated from noise parameters to allow robustness tuning, with the most popular technique being crossed arrays (Wu and Hamada, 2000). These methods rely on a single design paradigm, the statistical experiment. The previous method of changing a one factor at a time (OFAT) (Daniel, 1973) has been discounted as lacking the statistical power and requiring too many runs (Wu and Hamada, 2000). The advantages of learning from each run and approaching a maximum quickly are under appreciated and over criticized. This adaptive approach is also easy to explain and implement and does not require an extensive statistical background.

The literature on experimentation (Wu and Hamada, 2000; Montgomery, 1996; Box et al., 2005) is primarily from a statistical viewpoint and differing in paradigm from the previous one-factor approach, as Kuhn (1996) would say, the discussions between the two options may be incommensurable. The arguments for the statistical approach are based on a language and perspective that does not exist with the one-factor methodology. Even with a preponderance of evidence in support of the one-factor approach in certain situations, yielding slightly is tantamount to questioning the foundation for a statistical approach. The suggestion forwarded in this work is partially that an opportunity exists to bridge the paradigms of one-factor and statistical experiments. It is not to belittle the advancement of statistical experiments but to expand the framework to consider the system of application. A parallel can be drawn to Newtonian and relativistic physics. While it is accepted that for high speed and short time applications the Einstein view is more correct, for the majority of

earth bound physics the Newtonian approach is more useful. Einstein (1919) also suggests that his theory does not supersede Newtonian physics and finding accessible situations to measure any difference is difficult. From a practical standpoint, accepting the validity of Einstein does not reduce the ubiquitous utility of Newtonian physics in daily engineering activities. The same approach could be taken in experimentation. While acknowledging the validity of statistical experimentation there are situations where one-factor methodologies are more practical. Taking this openness even further there are opportunities to benefit from both a one-factor design as well as a statistical experiment. The analogy would be initial predictions using Newtonian physics to be later refined with relativistic calculations. For many instruments and situations the initial method would be sufficient but the confirmation and refinement using a relativistic approach would support the results.

Although the statistical and adaptive approaches are traditionally used in different situations this work will present opportunities to combine the results from both types of experiments into a complete testing framework. This combination is challenging to accept by both the academic as well as the industrial community. The academics question the pragmatic utility while most practitioners are unwilling to challenge the foundation of their six-sigma training. Although it may be impossible to bridge the incommensurate points of view, this work is an attempt to present some specific examples that demonstrate the utility of using both methodologies.

The first situation of interest is reusing runs from a prior adaptive experiment. By reusing runs the intent is to increase the number of common runs between the two experiments. The adaptive experiment cannot be preplanned and so the potential reuse in the subsequent experiment is stochastic. The procedure investigated begins with an aOFAT experiment. The first follow-up experiment is a traditional fractional factorial design. The number of runs reused is dependent on the fraction used, the number of variables, and size

of fraction. This number asymptotes to approximately twenty percent of the total adaptive runs. This run reuse is demonstrated on a number of actual experiments as well as surrogate experiments. If the follow-up experiment is more flexible in design, one option investigated was the non-balanced D-optimal design. As suggested in Wu and Hamada (2000), a fully orthogonal non-balanced D-optimal design is a good alternative to a fractional factorial. This change dramatically improves run reuse to all but one run, although it requires design planning after the initial aOFAT is complete. In addition to simulating the results of this improvement the independence of the two resultant maximum settings is demonstrated. Running an adaptive experiment before a statistical experiment creates an opportunity for run reuse while providing an independent maximum setting estimates.

This adaptive approach could also be used on the manufacturing floor. The method of evolutionary operation (EVOP) is revisited with a focus on utilizing adaptive experimentation. The alignment of this continuous improvement technique with the sequential maximization nature of an aOFAT provides a positive pairing. The use of these adaptive procedures was discussed by Box and Draper (1969) to the conclusion that the methodology was na ive. This conclusion is challenged here by investigating actual system responses, and showing a place for sequential adaptive experiments. Instead of using small fractional factorial experiments, repeated single steps in an adaptive procedure is shown to be more robust to initial and subsequent variable selection. Because of the stochastic nature of the repeated procedure a modified Gibbs sampler is introduced to minimize the additional runs while converging to a better variable setting. An offshoot of this procedure is the use of an adaptive experiment in computational function maximization.

The modified sequential simplex procedure was originally developed for evolutionary operation (Spendley et al., 1962). This rank-based geometric procedure was used frequently in the 1970's and 1980's although it languished in the 1990's for more complex

derivative-based methods. More recently it has returned to popularity with the increased use of computer simulations. As a robust method it is able to handle discontinuities and noise at the cost of more function evaluations. There are implementations of the simplex in most numerical programs for unconstrained optimization. The typical initial setup is based on changing one variable at a time (Press et al., 2007). This is improved by adding an adaptive element and performing an aOFAT for the initialization. The aOFAT procedure is modified to align the geometric center of the starting points to that of the non-adaptive method to permit equivalent comparisons. The adaptive procedure improves the overall convergence and reduces the number of function evaluations. Combining the adaptive procedure with the simplex starts the geometric procedure towards the maximum gradient for improved convergence. The benefit of this change is demonstrated on a test suite for numerical optimization (Moré et al., 1981).

Outside of the optimization another issue addressed here is variable selection. Using the Mahalanobis-Taguchi Strategy (MTS) from Taguchi and Jugulum (2002), data classification is based on a statistical distance. One hurdle to using this system is in selecting the best variables for classification. Traditionally orthogonal arrays are used to select a subset of variables. This method can be improved by using an aOFAT experiment combined with the Mahalanobis distance. This procedure is specifically applied to an image classification system where the variables of interest are the coefficients of a wavelet transform. In this case the addition of variables adds to the computational load of the classification system reducing its performance. It is important to add the minimum number of variables while maximizing their usefulness. The superior peroformance of the aOFAT combined approach is demonstrated and has been published in Foster et al. (2009).

In addition to dual results and as a starting procedure, aOFAT can be used as one experiment that combines the results into a single model. Combining two different types of

data was approached in a Bayesian framework. The use of a correlated gaussian random variable to make a posterior prediction has been used successfully by Joseph (2006). Part of this methodology is to use a correlation matrix for the input variables. Instead of using a larger experiment the information was divided between an early aOFAT experiment to create the correlation matrix followed by a highly aliased Plackett-Burman design (Plackett and Burman, 1946). The goal of this aspect of the work is to combine the relative strengths of both the aOFAT and traditional experimental procedures. The aOFAT can be used to create a variable ranking while the aliased design is able to efficiently define the model. A procedure to define the correlation matrix is created that benefits from published data regularities (Wu and Hamada, 2000) and variable distribution (Li and Frey, 2005). This methods performance is equivalent to using an uninformed correlation matrix and a larger experimental design with equal total runs. The procedure is demonstrated on a number of published examples as well as surrogate functions.

The last aspect of combined model building is to use experiments of different accuracy such as Qian and Wu (2008). Combining computational and physical experiments is one example of these different accuracies. The use of adaptive experiments uses a minimum number of runs while increasing the likelihood of having points near the maximum. A new method of calculating convergence is presented as well as a procedure to maximize each simulated markov chain. The result is a procedure that provides a good model using both data types that is more accurate at the maximum values.

The ultimate goal of this work is to create a foundation for the integration of adaptive experimentation into statistical experiments. Simple techniques are presented for using setup runs and getting benefit from those runs. This continues to manufacturing where evolutionary operation (EVOP) can be improved and simplified with adaptive experiments. A numerical maximization procedure is improved through a better starting approach, and

a classification procedure is shown to benefit from an adaptive parameter selection technique. The final area focused on using data from an adaptive experiment and a traditional experiment to build a single model. First, the covariance estimation was improved to yield more accurate and smaller models with the same number of runs. Second, incorporating data from two different accuracy sources is shown to benefit from making one of the experiments adaptive. The overriding goal for all of these procedures is to extend the framework for combining adaptive techniques with traditional experiments to reach a greater audience and provide examples and tools necessary for their application.

Bibliography

Box, G. E. P. and Draper, N. R. (1969). *Evolutionary Operation: A Statistical Method for Process Improvement*. John Wiley & Sons, Inc.

Box, G. E. P., Hunter, S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. John Wiley & Sons.

Daniel, C. (1973). One-at-a-time plans (the fisher memorial lecture, 1971). *Journal of the American Statistical Association*, 68:353368.

Einstein, A. (November 28, 1919). What is the theory of relativity? *The London Times*.

Foster, C., Frey, D., and Jugulum, R. (2009). Evaluating an adaptive one-factor-at-a-time search procedure within the mahalanobis taguchi system. *International Journal of Industrial and Systems Engineering*.

Frey, D. D., Englehardt, F., and Greitzer, E. M. (2003). A role for "one-factor-at-a-time" experimentation in parameter design. *Research in Engineering Design*, 14:65–74.

Joseph, V. R. (2006). A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48:219–229.

Kuhn, T. (1996). *The Structure of Scientific Revolutions*. University of Chicago Press, 3 edition edition.

Li, X. and Frey, D. D. (2005). A study of factor effects in data from factorial experiments. In *Proceedings of IDETC/CIE*.

Montgomery, D. C. (1996). *Design and Analysis of Experiments*. John Wiley & Sons.

Moré, J. J., Garbow, B. S., and Hillstrom, K. E. (1981). Testing unconstrained optimization software. *ACM Transactions on Mathematical Software*, 7(1):17–41.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33:305–325.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing (3rd Edition)*. Cambridge University Press.

Qian, P. Z. G. and Wu, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50:192–204.

Spendley, W., Hext, G. R., and Himsworth, F. R. (1962). Sequential application of simplex design in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461.

Taguchi, G. and Jugulum, R. (2002). *The Mahalanobis Taguchi Strategy: A Pattern Technology System*. John Wiley and Sons.

Wu, C.-F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, Inc.

Chapter 2

Experimental Background

2.1 Early Experimental Developments

The science and art of designed experimentation began as agriculture experimentation by Ronald A. Fisher (Figure 2-1) at the Rothamsted Experimental Station in England where he studied crop variation. The techniques that he developed were the basis to test different seed/soil/and rotation parameters in a noisy field environment (Fisher, 1921). This early work cumulated in two important books on the use of statistical methods in scientific investigation (Fisher, 1925, 1935). A parallel development was being made by William S. Gosset (Figure 2-2), also in agriculture but this time related to small samples of barley for beer production. These two early pioneers developed some of the foundations of statistics and experimentation including blocking, randomization, replication, and orthogonality. Another contribution that was made was progress on small sample distributions, thus for smaller experiments the estimates of significance and error could be calculated (Student, 1908).

The fundamentals of these early experiments were foundational to further experimental

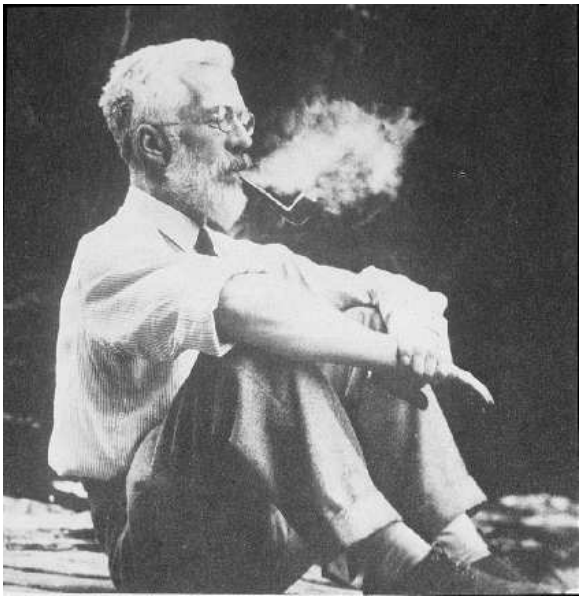

Figure 2-1: Ronald A. Fisher

Figure 2-2: William S. Gosset

development and continue to be utilized today. Replication utilizes repeated experiments at identical settings, although not run sequentially but at random. The principle of replication allows for an overall experimental error estimate. If this error is low compared with with the experimental response, the confidence is high that the experiment is representative of the population in general. The reverse is also true that given a desired error margin (or risk), it is possible to estimate the required number of replicates. Randomization suggests that the order of changes should vary randomly. By making adjustments in random order, any significance in the results is more likely due to the experimental variables and not some other latent, or hidden, variable. A latent variable is something that changes throughout the experiment but is not directly changed by the experimenter. These variables could be obvious like the temperature of the room, to something more hidden like the predilection of boys to use their right foot. If the experimental changes are applied in a random fashion then it is unlikely that these latent variables will affect the result. The next aspect introduced is if there are some uncontrolled variables that are too difficult or expensive randomize. One method to deal with these variables is through blocking. Identical sets of experiments can be run in blocks, and the different blocks can be run at different settings of these uncontrolled variables. An example of blocking would be two different manufacturing plants that would each run an identical experiment. Although the differences between plants are large, the changes within a plant should be similar. The goal for blocked experiments is for the within block variation to be low compared with the between block variation. The last aspect of early experimentation was input variable orthogonality. If the variables in an experiment are arranged such that there is zero correlation between them they are considered orthogonal. Most designed experiments are arranged to guarantee this property, which simplifies analysis.

The experimental designs that were developed began with full-factorial designs at two

levels. These designs are complete enumerations of all variable combinations. The first variable switches from the low to high setting every run, the second variable every two runs, the third every four, etc. This led to 2^n number of runs for each replication where n is the number of factors or variables. The runs should be randomized, blocked if possible, and replicated. These large designs had sufficient runs to estimate the main effects, and all interactions, the main drawback was they were too large for all but the simplest experiments. To reduce the number of runs fractions of these experiments were developed. The fractional designs begin with a smaller full-factorial design and to add additional factors that are combinations of the existing factors are used. Each factor run is orthogonal to the others so multiplying two or more factor runs together yields a new run that is orthogonal to those. The design of these is complicated in finding good variable combinations that yield orthogonal results to the greatest number of other factors. The factors that are not separable are called aliased. For example, given a three factor, full-factorial design, multiplying the first, second, and third factors (ABC) gives you a fifth factor (D). This design is a 2^{4-1} design with resolution IV, called so because the number of factors multiplied together to get the identity is four ($ABCD = I$). In general, a resolution IV design has no n-way interaction with any other $(5 - n)$-way interaction. This design is obviously aliased in any effects of ABC would not be distinguishable from main effect D. There is a tremendous research history on the fractional factorial concept and Yates (1935); Fisher (1935); Box and Hunter (1961b,a) are some good starting points. Fractional factorial designs are the workhouse of designed experimentation. Today research focuses on incorporating noise variables, identifying concomitant or lurking variables, and exploiting covariats, through such things as highly fractioned, non-replicated, or non-randomized designs (Sitter, 2002). There are other techniques for designing an experiment, but most industrial experiments rely on the fractional factorial.

One of the other techniques is called optimal design, it was first described by Smith (1918) but the lack of computational power prevented its popularity until later. The primary motivation of optimal design was to focus on the inferential power of the design versus the algebraic properties of its construction (such as rotatability) (Kotz and Johnson, 1993). This work will be limited to linear models and so a complete definition of optimal designs is unwarranted. The basics are the comparison of different potential designs against a criterion or calculation of merit. Numerical methods search through potential designs before selecting one with the best criterion. Given a linear model:

$$Y = X * \beta \tag{2.1}$$

The best linear estimate of β is $(X^T * X)^{-1} X^T * Y$ and a measure of the variance on this estimate (given uncorrelated, homoscedastic noise with variance σ) is:

$$\sigma^2 * (X * X^T)^{-1} \tag{2.2}$$

One measure of good design is the size of this matrix. There is no complete metric for the size of this matrix and so a number of alternatives have been proposed. One popular one is the D-optimality condition that seeks to minimize the determinant of this matrix. Others are the A-optimality for the trace of the matrix, or E-optimality minimizes the largest eigenvalue of the matrix. There are a number of other potential optimality conditions, here the focus is on D-optimality because it offers a clear interpretation, and is invariant to scale transforms. It is not the only choice for optimal designs but has been suggested as good starting location by Kiefer and Wolfwitz (1959). The main utility of optimal designs as stated in more recent texts Wu and Hamada (2000) is to augment previous runs. The draw-

back of this approach is the dependency on the underlying model before creating a design. By limiting the cases to those where the linear-model determinant is a global minimum it forces orthogonal models.

2.1.1 Higher Order Models

The previous models limited the analysis to linear and interaction terms. If it is desirable to estimate quadratic effects then one obvious extension would be to run a 3^n full-factorial experiment. The drawback of this large experiment is that most of the runs are used to estimate high order, improbable, interactions. Given the principle of hierarchy from Hamada and Wu (1992) which states that lower order effects are more important than higher order effects and effects of the same order are equal, most of these terms are insignificant, and so these runs are wasted. Utilizing fractional factorial designs has greater run economy while normally yielding the same models. There are also situations where the number of levels is a mixture of two and three level factors. This leads to a large number of potential experimental designs with different resolution and confounding structure. A small, but significant, change in approach is to view the experiment as an opportunity to efficiently fit a proposed model. If this alternative view is used then designs could be more efficient and much smaller. In an early advance, Box and Wilson (1951) showed how to overcome the problem where the usual two-level factorial designs were unable to find a ridge. These central composite designs (CCD) were efficient and rotatable (Box and Hunter, 1957), meaning that the variance estimate was comparable in any direction. The CCD consists of three ports first the corner or cube points (2^n) second the axial or star points ($2*n$) and the center points ($\approx 3-5$ Montgomery (1996)). With a defined goal of building a quadratic model these designs are highly efficient and are normally employed to search for more optimal operating

conditions. One selection that needs to be made by the experimenter is the distance of the star points. These points are located α times further than the corner points. The selection of $\alpha = 1$ is called the face centered cubic and has only three levels for each variable. Another popular selection is to make the design rotatable, or have a constant distance to the center point, so $\alpha = \sqrt{n}$. The last selection of α makes the cube points and the star points orthogonal blocks. This property is useful if they are going to be run sequentially in this case $\alpha = \sqrt{k(1 + n_{a0}/n_a)/(1 + n_{c0}/n_c)}$, where n_a is the number of axial points, and n_{a0} is the axial center points and n_c and n_{c0} is the same for the corner points of k variables. One drawback of the CCD design is that the corner points are run at all the variable extremes, and it is also not as efficient as some other deigns. If the experiment is going to be run at only three levels an improvement is the Box-Behnken design (Box and Behnken, 1960). This design is slightly more compact than the traditional CCD, and does not have any of the corner points. It was created by combining a number of incomplete block designs, and so also has potential for orthogonal blocking. For four variables the Box-Behnken design and CCD ($\alpha = \sqrt{n}$) are rotations of each other, one having points at the corners and the other not. This feature is not the case for more variables.

The Plackett-Burman designs are very efficient experimental designs. The metric of redundancy factor (Box and Behnken, 1960) is going to be used to describe these designs. If a designed experiment of k factors is going to be used to fit a polynomial model of order d then it has to be able to separably estimate $(k + d)!/k!d!$ model factors. For example, a full-factorial design of p-levels (normally 2 or 3) can at most estimate a model of order $p-1$. To estimate a quadratic model at least three points are necessary given a full-factorial design has p^k runs. The redundancy factor is the ratio of the number of runs to the number of parameters that can be separately estimated. For the full factorial design it is $p^k(p - 1)!k!/(k + p - 1)!$, which for a 2^5 design is 5.3 and for a 3^5 design is 11.6. The ratios for the

N	Vector
12	++-+++---+-
20	++--++++-+-+----++-
24	+++++-+-++--++--+-+----
36	-+-+++---+++++-+++--+----+-+-++--+-
44	++--+-+--+++-++++++--+-+++-----+---++-+-++-

Table 2.1: Plackett-Burman Generating Row

full factorial designs are very large. For the Plackett-Burman designs with the number of variables $k = 3, 7, 11, \ldots,$ or $4i-1$, the two-level ($p = 2$) require only $r = 4, 8, 12, 16, \ldots, 4i$ runs. Thus their redundancy factor is unity. This minimal redundancy is normally not used in practice as they have no residual data that can be used to check the validity of the model. The primary area of utility of this design is in screening experiments. If it is known in advance that a number of the variables will probably be unimportant then those extra runs can be used for model validity checks.

The construction of a Plackett-Burman design is completed in a cyclic fashion. A generating row is used initially as in Table 2.1. This generating row is then shifted one entry to the right, and the last entry is placed first. This procedure is repeated until the entire generating row has been have cycled through. The final row of all -1's is added to complete the design.

All of these designs and the general process of making design decisions are described in the original classic text on experimentation of Box et al. (1978) which has been updated in Box et al. (2005).

2.2 Adaptive Designs

During the second world war a number of statisticians and scientists were gathered by the United States government to from the Scientific Research Group (SRG). This group

worked on pertinent war studies such as the most effective anti-aircraft ordinance size and the settings for proximity fuses. One area of research that came from this group was the idea of sequential analysis. Instead of running an entire experiment before analyzing the results they considered the power of analyzing during the experiment (Friedman and Friedman, 1999). Out of the early work of Wald (1947) further researchers have proposed ways to not just analyze but to modify the experiment sequentially such as yan Lin and xin Zhang (2003). These methods are prominent in clinical trials such as Tsiatis and Mehta (2003) and Chow and Chang (2006). One of the ideas is now termed response-adaptive randomization (RAR) Hu and Rosenberger (2006) which was introduced as a rule called 'play-the-winner' by (Zelen, 1969). The idea is to bias the randomization of sequential trials by the preceding results. This fundamental idea will be used in this thesis in the chapter on evolutionary operation (Chapter 4) and again in the chapter on aOFAT integrated improvement (Chapter 7).

An additional area of research that began with the SRG was using repeated experiments to find a maximum by Friedman and Savage (1947). This was one of the foundations for Frey et al. (2003) and Frey and Jugulum (2003) work on the subject. In the work here repeated experiments are run with each subsequent experiment reducing the variable range. In the end the variable range spans the function maximum for linear convex variables.

The statistical design approach has been used as a starting point to optimization processes. One example is the question posed by Box (1957), could the evolutionary operation statistical experimentation procedure be made automatic enough to be run on a digital computer. This original question drove Spendley et al. (1962) to develop a geometric optimization procedure called the sequential simplex. This procedure will be investigated here because it has properties of interest. First the objective is to maximize a few runs, an adaptive procedure will have the biggest effect. As the number of runs grow the ability of the

statistical experiment to measure variable importance grows. The second reason that this application is appropriate is the goal is to search for a maximum.

Those two areas will play an important role in this thesis and are the motivation for much the work. A simple definition of these two main system aspects are those that first use very few experimental runs and second desire function maximization. There are many practical areas where these properties are desirable especially within the context of applied industrial experimentation. Taken to an extreme the logical goal is to maximize the value of each run and limit the total number of runs. As Daniel (1973) and Frey and Geitzer (2004) point out, there are numerous experimental situations where adaptation is desirable and stopping the experiment early is a frequent occurrence.

2.3 Background for One-Factor-at-a-Time (OFAT)

While it is almost impossible to investigate the history of the intuitive OFAT (one-factor-at-a-time) experiment more recent investigations into comparative one-factor options is available. Daniel (1973) was an early proponent of the technique within the statistical community. He discussed the opportunity and the required effect size to make it worthwhile. His main concern was with changing each variable in order and the comparison to a regular fractional factorial experiment. While the motivation for each of these different types of experiments is disparate the runs and analysis is similar. Because of the risk of time-trends and the inability to estimate interactions it was determined that the ratio of effect to noise had to be around four. This high resolution gave sufficient power to this historic method. There were five different types of one-factor experiments presented by Daniel (1973). These five types are strict, standard, paired, free, and curved. Strict varies each subsequent variable beginning with the previous setting. If the experimenter was test-

ing *a*(where only *a* is at the high setting) then *ab*(with both *a* and *b*) then *abc* this is an example of a strict OFAT. The advantages to this arrangement is that it transverses the design space and can be easily augmented by starting at the beginning and removing factors, the experiment above could be extended by adding *bc* and *c*. The standard OFAT runs each variable in order *a*, *b*, *c*, and *d*. This order focuses the runs on one corner of the experiment, which increases knowledge around that area but does not improve estimates of interactions. The paired order is designed for runs that are typically run on parallel experimental setups. Each setup completes a pair of runs that can estimate the main effects and separate the interactions. The first two runs for the first setup could be *a* and *(1)*(all values low) while the second would run *abcd* and *bcd*. These two standard OFAT experiments are combined to yield variable information after two runs of each setup, thus decisions can be made about future experiments. The free OFAT is only touched on briefly but brings a level of adaptiveness. After a part of a traditional experiment is complete, some response assumptions are made to reduce the additional runs. If the initial highly fractioned experiment shows A+BC is important then choose additional runs to separate out A from BC assuming the rest of the effects are negligible. The final OFAT experiment is a curved design. This separates out easy to change from difficult to change variables. The easy to change variables are swept through their range of values while the others remain constant. A subsequent set would change all of the variables and run the sweep again. These five represent the basic set of publicized OFAT experiments. The practitioners of this experimentation technique often wanted an easy way to gain factor importance in situations where the experimental error was low and results were quickly obtained.

2.4 Adaptive One-Factor-at-a-Time (aOFAT)

The one-factor-at-a-time (OFAT) experiment was once regarded as the correct way to do experiments, and is probably the default in many non-statistical frameworks. Inside the statistical framework it is possible to view full-factorial designs as a series of OFAT experiments. Given a 2^3 experiment in standard order runs $(1, 2, 3, 5)$, $(8, 7, 6, 4)$ are two OFAT experiments that yield the same runs as a full-factorial experiment.

Daniel (1973) discusses this option and the utility benefits of OFAT to experimenters. It is possible to learn something after each experimental run, and not require the entire set of runs to be complete. The power of this analysis requires the effect to be three or four times as great as the noise, and in many situations these are the only effects of interest.

The four basic issues brought up against OFAT experiments, and repeated in different contexts are (Wu and Hamada, 2000):

- Requires more runs for same effect estimation precision
- Cannot estimate some interactions
- Conclusions are not general
- Can miss optimum settings

These are legitimate issues with the methodology but the effect in practice depends significantly on the experimental purpose and scope. Taking each of these points out of the experimental context to blindly support a statistical based approach ignores some situations where this methodology has clear advantages.

These same negative arguments are repeated in (Czitrom, 1999) where the author give specific examples where the choice of a OFAT experiment is inferior to a regular statistical

experiment. First, the discussion does not address realistic experimentation nor does it discuss additional information sources. Both of these possibilities are discussed in this work (Chapter 3 and 7). To support the statistical experiment the author gives an example of two variables where the experimenter wants to run an OFAT of three points, temperature and pressure. The number of replicas was decided in advance as well as the variable range. The first concern is around how that data was collected and how it could be combined with the experimental results. Second, the entirety of all the experiments are planned in advance, if the outcome is to search for a maximum, there are better options (as discussed in (Friedman and Savage, 1947)). There is no argument against the majority of the examples presented in (Czitrom, 1999) (examples two and three), and the statistical experimental framework is superior to a traditional OFAT approach. The reality that OFAT is inferior in certain situations does not eliminate the possibility that OFAT has a useful place in the experimental toolbox. This work explores a handful of those opportunities.

The uses forwarded in this work augment, instead of replace the statistical experimentation. There are many situations that benefit from an adaptive framework, important example situations include:

- Insufficient planning resources
- Immediate improvement needed
- Variable ranges and effect magnitude unknown

Although there may other specific situational examples, these are the situations described in Frey and Geitzer (2004) and Daniel (1973).

If the resources to plan the experiment and layout and perform the runs are not available is no experimentation possible? Some situations are limited by time and resource pressure

and only overhead-free experimentation, such as OFAT, is possible. There are other situations that demand some immediate improvement to the running condition. Additional, and more complete, experiments can be run afterwards to tune the system but an initial change needs to be made that has a high likelihood of succeeding (such as adaptive-OFAT (aOFAT)). Many experiments are run on processes and factors where little is known. It may not be possible to determine the variable ranges for the experiment with a reasonable degree of confidence. The only way to determine the possible ranges is to experiment on the system, and a OFAT framework can determine the maximum and minimum settings. These general situations have specific examples that have shown to benefit from the OFAT approach. There are potentially many other situations where this technique may be beneficial, but there has not yet been a serious inquiry. For example, one area may be to reduce the number of variable changes. The OFAT and aOFAT experiment could be compared to options such as Gray codes (Gray, 1953). It is infeasible to predict all the opportunities but as the technique gains greater publication its use should expand.

As the statistical approach is accepted, many authors (Wu, 1988; Box et al., 2005; Myers and Montgomery, 2002) suggest an adaptive framework where a sequence of experiments is performed. These experiments could be changing because of newly discovered interactions or to change the variable ranges to search for a better operating condition. The minimum experimental process suggested is a two or three factor experiment (in Box et al. (2005), for example), but if this is reduced to the extreme then their procedure also reduces to an aOFAT sequential experimentation procedure. The procedure outlined in Myers and Montgomery (2002) uses this sequential procedure and as the value nears a maxima, the experiment is expanded to study more of the interactions or quadratic effects. This adaptive sequential procedure is revisited in this work with the initial experiment being the minimal aOFAT followed by a statistically based procedure.

There have been some recent comparisons between the aOFAT methodology and more traditional orthogonal arrays in Frey et al. (2003). They found that for the same number of runs, the aOFAT was able to discover the maximum setting with high probability. The successful resultant of the procedure should be limited to those situations where the maximum number of runs is small (limited to the number of variables plus one). Thus the comparison is normally between aOFAT and Resolution III Fractional Factorials (later in this work Plackett-Burman Designs will also be included). If there are additional resources there is limited information about what would be the next steps. If the goal is to match a standard factorial experiment, Daniel (1973) suggests running a series of OFAT experiments. These experiments cover the runs for a reduced factorial design and so an adaptive addition is unnecessary. Friedman and Savage (1947) suggest that a series of adaptive experiments can be used to search for a maximum. More recently, Sudarsanam (2008) proposes running a number of aOFAT experiments and ensemble the results. Most authors are silent on the subject of additional runs and instead offer direct comparisons to specific experimental designs. One could conclude that the current methodology for sequential experimentation could be utilized just replacing the fractional factorial design with an adaptive design. This extension has yet to be demonstrated in practice and does not prevent methodologies that combine aOFAT experiments and other experiments.

Frey et al. (2006); Frey and Sudarsanam (2008); Frey and Wang (2005) have looked into the mechanism behind aOFAT that leads to improvement. This research is empirically based and shows that for low levels of experimental error or relatively high amounts of interaction aOFAT is superior to Resolution III Fractional Factorial designs (Frey et al., 2003). The comparative advantage with high interaction suggests that there might be a complementary relationship between aOFAT and Fractional Factorial designs. Given this relationship are there other options for additional resources? Some possibilities are inves-

tigated in this work including, run reuse in another experiment and searching for a maxima through a sequential simplex. The other area of investigation was utilizing the relationships in Frey and Wang (2005) to apply a Bayesian framework to maximize the utility of the aOFAT experiment as a prior predictor.

The underlying system structure requires low noise for good system estimates. Daniel (1973) suggests that the effect magnitude should be 4σ while Frey et al. (2003) suggests that 1.5σ is sufficient. These estimates are based on different data sets and may be different for a particular experiment. The other requirement was the speed to collect data samples, both Daniel (1973); Frey et al. (2003) suggest that sampling should be quick. This requirement limits the effect of drift or time series effects. It is possible to account for some of these effects by running multiple experiments, but the lack of randomization limits the extent of this improvement.

There are many experimental techniques the two presented here are adaptive-one-factor-at-a-time (aOFAT) and statistical experiments. Both have situations where they are superior but due to an adversarial relationship there is limited research on the combination of the two methodologies. This research begins to bridge the OFAT and specifically aOFAT experiments with statistical experimental techniques. The areas of application are run-reuse, maxima seeking, variable selection, and applications in a Bayesian Framework including prior prediction and dual data integration.

2.5 aOFAT Opportunities

The combination of statistical and adaptive experiments is seen as a starting point that can leverage the strengths of each technique. Instead of choosing between the two techniques the goal is to combine the two to improve the outcome. As mentioned previously the areas

under investigation are for system maximization where there is little risk of time trends affecting the results. The initial approach is to improve the traditional industrial experiment. These experiments are normally part of a six-sigma process such as Breyfogle (2003). Given some process variables, noise, and an output variable junior-level engineers design an experiment to improve their process. This has been instituted in companies such as GE with the green-belt and black-belt certification (GE, 2009). Within these experiments the application areas are broad but the experiments of interest require some physical setup and should have relatively low expected levels of time dependent noise. Many of these systems could be replaced completely with adaptive experimental techniques although there are added benefits to look at experimental integration. Adaptive experiments can augment these traditional experiments to provide additional benefit with little experimental risk. This integration is initially presented in Chapter 3 to run an adaptive experiment during setup or to initially test the system. This is then followed by a traditional statistical experiment. The integration of these two methods is presented as the ability to reuse some of the runs from the adaptive experiment in the subsequent statistical experiment. This combination does not integrate the analysis but provides two experiments with fewer runs than both separately. This technique is general enough to be applied to most experimental situations without affecting the results of the designed experiment. It is also possible to integrate the results from both experiments into a single prediction. There are two ares explored here and both are Bayesian. The use of classical statistics was poorly equipped because the problem integrates two sources of data to estimate the model. If the system knowledge is sufficient to choose a system of models then a traditional approach may be used, although the experimental setups would differ. Many others have also investigated this data integration including Qian et al. (2006); Kennedy and O'Hagan (2001); Goldstein and Rougier (2004) who have looked at mostly empirical Bayesian approaches. This technique will be

employed here in using the initial prediction for the covariance matrix as in Chapter 8 well as for the use of two different experimental costs in Chapter 9. The empirical approach is one method, some of these models could also use a closed form posterior distribution. For academic implementation the empirical approach is flexible and interpretable, further industrial use could gain speed and computational flexibility by calculating the posterior distributions. There are many other areas of application to combine two sources of data. The goal in this work was to investigate the breadth of looking at additional runs in an experiment and combining multiple different experiments. One could investigate additional models options outside of the linear models explored here. One option is the kringing models such as Joseph et al. (2008), or other patch models such as radial basis functions in Yang (2005). The general models used here should provide a background to drive greater complexity and application specific model options. Outside of model building the opportunities extend to replacing the use of orthogonal arras or other extremely fractionated designs. In Chapter 6 an investigation was made into a classification system that historically used orthogonal arrays. Replacing the aOFAT in these situations improves the resolution at minimal cost. The application of tuning a classification system fits with the previous requirements, there are few available runs compared with the number of variables, and the goal is to maximize the ability of the classifier. This example emphasizes the strengths of the aOFAT technique within a classification context. In addition to traditional response model the classification model can also be helped with the adaptive experiments. There are other classification techniques, such as Yang et al. (2008), that could be investigated to use an adaptive data collection approach. Outside of modeling, a promising area of application is in simple optimization.

The opportunity within the optimization field is around techniques that are relatively simple and do not use need to calculate derivatives. Originally the investigation focused on

optimization techniques that started as statistical experiments. Box (1957)'s evolutionary operation (EVOP) procedure is a particularly good starting point. There are many opportunities within the optimization literature and some identified as statical optimization techniques in Nocedal and Wright (1999). To demonstrate the adaptive application a historically related unconstrained optimization procedure known as sequential simplex was selected. This technique was originally developed from the EVOP procedure but is now popular with computer simulations. This fundamental technique is well publicized and aligns well with an adaptive opportunity. Other opportunities have not been investigated although there may be a handful of possibilities outside of the intersection of statistical experimentation and numerical optimization.

2.6 Prediction Sum of Squares (PRESS)

When comparing different experimental model-building methods it is difficult to assess 'better'. One model may be larger and more accurate, but the other uses fewer variables. The predicted sum of squares (PRESS) from Allen (1971b), also known as the predicted residual sum of squares (Liu et al., 1999), is a metric for model variable selection. This metric originated when Allen (1971a) improved upon the traditional residual sum of squares with a metric that would not always suggest additional regression variables improve accuracy. The accuracy of a prediction point that was not in the regression would decrease as the model was over-fit. This metric would increase as the fit improved at that point and then decrease after it was over fit. This new approach to model building focused on prediction accuracy. The model was now sensitive to the point choice for this calculation. His procedure was to take each point individually in the data set, fit the model without that point, and check the error at that point. In the statistical learning community this is known

as leave-one-out cross-validation. Tibshirani (1996, pg. 215) recommends the low-bias and high variance properties for this method but warns that the calculation burden could be significant. The major motivation in using this method is that a time-saving shortcut exists for linear models.

Given a model

$$Y = X \cdot \beta + \varepsilon \tag{2.3}$$

with data X of dimension nxp and Y of dimension nx1, the least squares predictor of β would be

$$\hat{\beta} = (XX^T)^{-1}X^T Y \tag{2.4}$$

so $\hat{y}_i = x_i^T * \hat{\beta}$ and let $\hat{\beta_{(i)}}$ be the estimate of β with the ith observation removed. The PRESS is defined as

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - x_i^T \hat{\beta_{(i)}})^2 \tag{2.5}$$

This would be computationally challenging without this simplification.

$$\text{PRESS} = \sum_{i=1}^{n} \frac{y_i - \hat{y}_i}{1 - H_{ii}}^2 \tag{2.6}$$

Where H_{ii}'s are the diagonals of the H, hat matrix (because it puts a 'hat' on y).

$$H = X(XX^T)^{-1}X^T \tag{2.7}$$

The diagonals are equal to the leverage of the observation i. This simplification requires only a single calculation of H and then using the diagonals and $\hat{y} = HY$, the PRESS statistic is a summation. To compare with other measurements of error such as Root-Mean-Square-Error (RMSE) and Standardized-RSME (SRSME) this work will frequently report

the $\sqrt{PRESS}$

2.7 Empirical Bayesian Statistics

Given data x a goal is to determine the most probable underlying error distribution that would yield that data. In practice we assume that the form of the distribution is known but, based on some unknown parameter (λ). This distribution parameter is assumed to be a random variable from a known distribution G.

The unconditional probability distribution on x is given as:

$$p(x) = \int p(x|\lambda) dG(\lambda) \tag{2.8}$$

Our goal is to determine *postereri* distribution on λ given the data x. This is accomplished by looking at the error to any given estimator function $\psi(x)$.

$$\begin{aligned} E(\psi(x) - \lambda)^2 &= E[E[(\psi(x) - \lambda)^2|\lambda]] \\ &= \int \sum_x p(x|\lambda)[\psi(x) - \lambda]^2 dG(\lambda) \\ &= \sum_x \int p(x|\lambda)[\psi(x) - \lambda]^2 dG(\lambda) \end{aligned} \tag{2.9}$$

for a fixed x we can solve for the minimum value if the expected value by solving for the interior equation $I(x)$-

$$I(x) = \int p(x|\lambda)(\psi(x) - \lambda)^2 dG(\lambda) \tag{2.10}$$

fixing x so $\psi(x) = \psi$ this equation can be expanded given a constant function $\psi(x) = \psi$

-

$$
\begin{aligned}
I &= y^2 \int pdG - 2\psi \int p\lambda dG + \int p\lambda^2 dG \\
&= \int pdG(\psi - \frac{\int p\lambda dG}{\int pdG})^2 + \left[\int p\lambda^2 dG - \frac{(\int p\lambda dG)^2}{\int pdG} \right]
\end{aligned} \tag{2.11}
$$

and is at a minimum when

$$
\psi(x) = \frac{\int (p|\lambda)\lambda dG(\lambda)}{\int p(x|\lambda) dG(\lambda)} \tag{2.12}
$$

This is the posterior estimate of λ. This is the empirical Bayesian approach to estimate the distribution parameter given the data x. The biggest challenge to this approach is to determine a valid initial distribution G to yield a good estimate of the distribution parameter. Gelman et al. (2003) discourages the term empirical Bayes for this method because it implies that the full Bayesian approach is somehow not empirical although they both are experimental.

2.8 Gaussian Process (GP)

The Gaussian Stochastic Processes, or Gaussian Process (GP), is also known as a Gaussian Random Function Model. Given a fixed input space that is greater than a single variable, an output Y is a GP if for any vector x in the input space the output Y has a multivariate normal distribution. In practice the GP correlation function is selected to be non-singular. Thus for any given input vector the covariance matrix as well as the output distribution is also non-singular. The GP can be specified by a mean function and a covariance function. The mean is typically constant and normally zero although for one process in this work it is one

instead. The covariance function determines the relationship between the input variables. This is a stationary process and so only the difference in the input values is needed. There are two main choices for the correlation function, first choice is the Gaussian or power exponential:

$$R(x_1 - x_2) = \exp(-\theta \cdot (x_1 - x_2)^2) \tag{2.13}$$

The second correlation function changes the square to an absolute value and the resultant GP is called a Ornstein-Uhlembeck process (Santner et al., 2003). Both of these correlation functions will be used in this work. The Gaussian is infinitely differentiable at the origin and is useful to represent smooth processes. The Orstein-Uhlembeck process has more random fluctuations and is more representative of observed data with random error.

2.9 Hierarchical Probability Model

A realistic and representative model generator will be used to test the different methodologies presented in this thesis specifically in Chapter 3 for reusing aOFAT runs as well as Chapter 7 where the aOFAT in incorporated into a correlation matrix. This model, and the coefficients used here, come from Frey and Wang (2005). The basic idea is taken from Chipman et al. (1997) with the intent of generating a population of models that exhibit data regularities from Wu and Hamada (2000) such as effect sparsity, hierarchy, and inheritance. Using Equations 2.14 to 2.23 a large population of functions can be generated that mimic actual experimental systems. The coefficients (p, p_{ij}, p_{ijk}, β_i, β_{ij}, β_{ijk}, c, σ_N, σ_ε, s_1, s_2) come from an analysis of 113 full-factorial experiments (of sizes 2^3, 2^4, 2^5,and 2^6) that come from published journals.

$$y(x_1, x_2, \ldots, x_n) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j>i}}^{n} \beta_{ij} x_i x_j + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j>i}}^{n} \sum_{\substack{k=1 \\ k>j}}^{n} \beta_{ijk} x_i x_j x_k + \varepsilon \tag{2.14}$$

$$x_i \sim NID(0, \sigma_N^2) \quad i \in 1 \ldots m \tag{2.15}$$

$$x_i \in \{+1, -1\} \quad i \in m+1 \ldots n \tag{2.16}$$

$$\varepsilon \sim \mathrm{NID}(0, \sigma_\varepsilon^2) \tag{2.17}$$

$$\Pr(\delta_i = 1) = p \tag{2.18}$$

$$\Pr(\delta_{ij} = 1 | \delta_i, \delta_j) = \begin{cases} p_{00} & \text{if } \delta_i + \delta_j = 0 \\ p_{01} & \text{if } \delta_i + \delta_j = 1 \\ p_{11} & \text{if } \delta_i + \delta_j = 2 \end{cases} \tag{2.19}$$

$$\Pr(\delta_{ijk} = 1 | \delta_i, \delta_j, \delta_k) = \begin{cases} p_{000} & \text{if } \delta_i + \delta_j + \delta_k = 0 \\ p_{001} & \text{if } \delta_i + \delta_j + \delta_k = 1 \\ p_{011} & \text{if } \delta_i + \delta_j + \delta_k = 2 \\ p_{111} & \text{if } \delta_i + \delta_j + \delta_k = 3 \end{cases} \tag{2.20}$$

$$f(\beta_i | \delta_i) = \begin{cases} \mathrm{N}(0, 1) & \text{if } \delta_i = 0 \\ \mathrm{N}(0, c^2) & \text{if } \delta_i = 1 \end{cases} \tag{2.21}$$

$$f(\beta_{ij} | \delta_{ij}) = \frac{1}{s_1} \begin{cases} \mathrm{N}(0, 1) & \text{if } \delta_{ij} = 0 \\ \mathrm{N}(0, c^2) & \text{if } \delta_{ij} = 1 \end{cases} \tag{2.22}$$

$$f(\beta_{ijk}|\delta_{ijk}) = \frac{1}{s_2}\begin{cases} \mathrm{N}(0,1) & \text{if } \delta_{ijk} = 0 \\ \mathrm{N}(0,c^2) & \text{if } \delta_{ijk} = 1 \end{cases} \tag{2.23}$$

There are important attributes of this model that should be noted. The model encapsulates the three data regularities published in Wu and Hamada (2000); sparsity, or the fact that only a few effects will be significant; hierarchy, or that the biggest effects are main effects followed by two-way and then three-way interactions; and finally inheritance, or if a variable has a significant main effect it is likely to be significant in a two and three-way interactions. Next, the effects follow a normal distribution and so have an equal probability of being positive or negative. This model includes only main effects and interactions, higher order effects and other model non-linearities are not present. The use of a multi-variate linear model is appropriate in this case because the experimental design under study is very low order. The resulting experimental model is of lesser complexity than the model used to create the HPM.

The HPM is going to be used in a number of studies in this thesis to test the effectiveness of different experimental routines. Along with the HPM analysis of a proposed method, actual examples are pulled from the literature to demonstrate the method. The use of the HPM is designed to test a variety of models and determine the robustness of the different methods, while the example is used to ground model in one specific example.

2.10 Opportunities

The use of adaptive experimentation has a long past, and historically it was the only way to experiment. After the current statistical movement eliminated nearly all adaptive experiments, a new found place has been emerging for these experiments such as in (Frey et al.,

2003) and (Frey et al., 2006). This work focuses on the more pragmatic experimentalist that finds a good place for the intuitive adaptive experiment along with the statistical fractional factorial, CCD, or Box-Behnken design. As the computational processing techniques advance, the potential to use the Gaussian process in an Empirical Bayes framework extends the utility of these adaptive experiments combined with traditional statistical experiments. When comparing multiple experimental techniques the cross-validated PRESS statistic will be employed to help differentiate the models with different numbers of factors.

Bibliography

Allen, D. M. (1971a). Mean square error of prediction as a criterion for selection variables. *Technometrics*, 13(3):469–473.

Allen, D. M. (1971b). The prediction sum of squares as a criterion for selecting predictor variables. Technical Report Tech. Report 23, University of Kentucky Departement of Statistics.

Box, G. E. P. (1957). Evolutionary operation: A method for increasing industrial productivity. *Applied Statistics*, 6:81–101.

Box, G. E. P. and Behnken, D. W. (1960). Some three level designs for the study of quantiative variables. *Technometrics*, 2:455–476.

Box, G. E. P. and Hunter, J. S. (1957). Multifactor experimental designs for exploring response surfaces. *Annals of Machematical Statistics*, 28:195–241.

Box, G. E. P. and Hunter, J. S. (1961a). The 2k-p fractional factorial designs part ii. *Technometrics*, 3:449–458.

Box, G. E. P., Hunter, S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. John Wiley & Sons.

Box, G. E. P. and Hunter, T. S. (1961b). The 2k-p fractional factorial designs part i. *Technometrics*, 3:311–351.

Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Expimenters: An Introduction to Design, Data Analysis, and Model Building*. John Wiley & Sons.

Box, G. E. P. and Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 13:1–45.

Breyfogle, F. (2003). *Implementing Six Sigma, Second Endition*. John Wiley & Sons.

Chipman, H. M., Hamada, M., and Wu, C. F. J. (1997). Bayesian variable-selection apporach for analyzing designed experiments with complex aliasing. *Technometrics*, 39:372381.

Chow, S.-C. and Chang, M. (2006). *Adaptive Design Methods in Clinical Trials*. Chapman & Hall/CRC Press.

Czitrom, V. (1999). One-factor-at-a-time versus designed experiments. *The American Statistician*, 53(2):126–131.

Daniel, C. (1973). One-at-a-time plans (the fisher memorial lecture, 1971). *Journal of the American Statistical Association*, 68:353368.

Fisher, R. A. (1921). Studies on crop variation. i. an examination of the yield of dressed grain from broadbalk. *Journal of Agricultural Science*, 11:107–135.

Fisher, R. A. (1925). *Statistical Methods for Resarch Workers*. Eidenburg and London, Oliver and Boyd.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburg and London, Oliver and Boyd.

Frey, D. D., Englehardt, F., and Greitzer, E. M. (2003). A role for "one-factor-at-a-time" experimentation in parameter design. *Research in Engineering Design*, 14:65–74.

Frey, D. D. and Geitzer, E. M. (2004). one step at a time. *Mechanical Engineering*, 126(7):36.

Frey, D. D. and Jugulum, R. (2003). How one-factor-at-a-time experimentation can lead to greater improvements than orthogonal arrays. In *Proceedings of the ASME Design Engineering Technical Conference*.

Frey, D. D. and Sudarsanam, N. (2008). An adaptive one-factor-at-a-time method for robust parameter design: Comparison with crossed arrays via case studies. *ASME Journal of Mechanical Design*.

Frey, D. D., Sudarsanam, N., and Persons, J. B. (2006). An adaptive one-factor-at-a-time method for robust parameter design: Comparison with crossed arrays via case studies. In *Proceedings of 2006 ASME International Design Engineering Technical Conference and Computers and Information in Engineering Conference, DETC2006*.

Frey, D. D. and Wang, H. (2005). Towards a theory of experimentation for expected improvement. In *Proceedings of IDETC/CIE*.

Friedman, M. and Friedman, R. D. (1999). *Two Lucky People: Memoirs*. University of Chicago Press.

Friedman, M. and Savage, L. J. (1947). *Planning Experiments Seeking Maxima*, chapter 13, pages 365–372. Techniques of Statistical Analysis - eds. Eisenhart C, Hastay MW, Wallis WA. New York: McGraw-Hill.

GE (2009). www.ge.com/sixsigma/. Website.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Goldstein, M. and Rougier, J. C. (2004). Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, 26:448–466.

Gray, F. (1953). Pulse code communication, u.s. patent number 2632058, filed in 1947.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137.

Hu, F. and Rosenberger, W. F. (2006). *The Theory of Response Adaptive Randomization in Clinical Trials*. John Wiley.

Joseph, R. V., Hung, Y., and Sudjianto, A. (2008). Blind kriging: A new method for developing metamodels. *ASME Journal of Mechanical Design*, 130:1–8.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 63:425–464.

Kiefer, J. and Wolfwitz, J. (1959). Optimum designs in regression problems. *Annuals of Mathematical Statistics*, 30:271–294.

Kotz, S. and Johnson, N. L., editors (1993). *Breakthroughs in Statictics: Volume1: Foundations and Basic Theory*. Springer.

Liu, H., Weiss, R. E., Jennrich, R. I., and Wenger, N. S. (1999). Press model selection in repeated measures data. *Computational Statistics and Data Analysis*, 30:169–184.

Montgomery, D. C. (1996). *Design and Analysis of Experiments*. John Wiley & Sons.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methoology*. Wiley.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, C. F. J. (2006). Building surrogate models based on detalined and approximate simulations. *ASME Journal of Mechanical Design*, 128:668–677.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer.

Sitter, J. L. L. R. R. (2002). Analyzing unrepliced blocked or split-plot fractional factorial designs. *Jouurnal of Quality Technology*, 34:229–243.

Smith, K. (1918). On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations. *Biometrika*, 12:1–85.

Spendley, W., Hext, G. R., and Himsworth, F. R. (1962). Sequential application of simplex design in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461.

Student (1908). The probable error of a mean. *Biometrika*, 6:1–25.

Sudarsanam, N. (2008). *Ensembles of Adaptive One-Factor-at-a-time Experiments: Methods, Evaluation, and Theory*. PhD thesis, MIT.

Tibshirani, R. (1996). Bias, variance, and prediction error for classification rules. Technical Report, Statistics Department, University of Toronto.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90(2):367–378.

Wald, A. (1947). *Sequential Analysis*. John Wiley & Sons.

Wu, C.-F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, Inc.

Wu, L. S.-Y. (1988). Business planning under uncertainty: Quantifying variability. *The Statistician*, 37:2:141–151.

yan Lin, Z. and xin Zhang, L. (2003). Adaptive desgins for sequential experiments. *Journal of Zhejiang University Science*, 4:214–220.

Yang, C.-H., Huang, C.-C., Wu, K.-C., and Chang, H.-Y. (2008). A novel ga-taguchi-based feature selection method. In *Intelligent Data Engineering and Automated Learning - IDEAL 2008 - 9th International Conference*.

Yang, Z. R. (2005). Bayesian radial basis function neural network. In *Intelligent Data Engineering and Automated Learing - IDEAL 2005*.

Yates, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society*, 2:181–247.

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, 64(325):131–146.

Chapter 3

Reusing Runs

3.1 Introduction

In many industrial and research experiments the experimenter first tests a number of runs to determine if the variable settings are correct and if the setup is functional. These early runs are then discarded and the designed experiment is completed. Instead of throwing away these runs, is there potential for them to be reused in the actual experiment? This chapter advocates one strategy for utilizing these early runs, and thus reducing the length of the overall experiment. If these early runs are arranged in an adaptive-One-Factor-at-a-Time (aOFAT) experiment then in addition to the setup function the experimenter can garner information about the system maximum as well as reduce the total number of runs. Early screening experiments offer the best application to realize improvement. In other words, when the experimenter is trying to determine the important main effects while accepting an alias effect or an unbalanced design to reduce experimental runs. In these situations the early set-up runs may be a significant fraction of the total experimental runs and potential for reuse may be worthwhile. With so few runs there is a possibility that the experiment

may be overly influenced by noise; a measure of this possibility is available in the aOFAT without completing a replicate and would be useful. This chapter will focus on setting up and running these two experiments. The analysis is focused on the number of runs that could be reused and the interactions between the two analysis types. This is one basic way of combining the aOFAT experiment with a statistical experiment. Later chapters (Chapters 7 and 8) will look at combining these data into a single, consistent, model.

3.2 Background

The setup runs in an experiment normally consist of varying each parameter separately to the high and low experimental value. Although this procedure is not widely discussed in the experimental design literature it has been observed in numerous actual experiments. These early runs are traditionally thrown out because they are not necessarily orthogonal or balanced and could lead a traditional regression analysis to incorrect model coefficients. If these early runs have slightly more structure, while being intuitive for the operator, they could be incorporated into some of the follow-up analysis. The experimental runs discussed here are D-Optimal and fractional factorial designs. The D-optimal designs are orthogonal but, not necessarily balanced. The fractional factorial designs are both balanced and orthogonal. An unbalanced design has fewer runs in one factor setting, this could be problematic in systems with heteroscedastic noise. But in most homoscedastic early screening designs the utility of balanced, un-replicated designs may be unnecessary. With few runs, there is insufficient data to estimate parameter variance and the biggest benefit of repeated high and low settings is a better mean estimate. It would be possible to add balance to this design by repeating necessary points as Parker et al. (2007) showed in their analysis. The biggest drawback to having unbalanced data is the inability to use standard analysis tech-

niques. There are some suggestions to utilize approximate methods (Montgomery, 1996), but with modern computational resources it is assumed that the access to exact methods using a general linear model (GLM) and distribution estimates (McCullagh and Nelder, 1989) is possible as in Chapter 7. If there is too little data then creating a GLM has too little resolution and a Gibbs sampler (Chapter 8) could be utilized.

3.3 Initialization

The process for setting up an experiment is usually left to a technician who prepares for useful, accurate data through an iterative trial process. Starting the experimental process with an adaptive experiment is straightforward to the technician as well as useful in estimating the maximum experimental setting. After initially connecting all of the hardware and testing the data collection, the system is run at a few settings to be sure that everything is functioning correctly. These setup runs are not previously planned and serve as a baseline to check the functionality of the system. The suggestion in this chapter is to run though all of the variables that will be used in the experiment and check their high and low settings. The purpose is two-fold, first it is good to validate that the variables are responding and to check that the range is appropriate for the experiment. Second this practice allows one to reevaluate the planned experiment to make sure that each setting is achievable and measurable. There are a two major historical choices for running this setup; a one-factor-at-a-time (OFAT) approach or a fractional factorial approach. The fractional factorial is balanced, orthogonal, and could possible measure interactions, it is the primary suggestion of any statistician. A major drawback is that it obfuscates the results to the technician. Multiple variables are changed with each run and so a problem with the limit, or with the hardware, is difficult to diagnose; it also requires the whole experiment to be completed before any

analysis. Another option is to run an adaptive OFAT (aOFAT) and sequentially change each variable between the high and low settings. The resulting experiment is not balanced or orthogonal; and it is impossible to identify interaction terms. The benefits are simple implementation for the technician and allows real-time diagnosis of problems or mismatched variable settings. The non-adaptive OFAT can be planned in advance but cannot identify the maximal settings nor benefit from interactions.

Running an adaptive experiment also has the benefit that it has a high probability of achieving the highest setting for the system. This will help in testing the extremes of the system settings and validating the high/low settings of the variables. If this added experimental step of an initialization aOFAT is used, one important concern is the number of additional runs required. Some of the aOFAT runs can be incorporated into the subsequent design although determining the number of reused runs is not straightforward.

3.4 Reusing aOFAT runs in a Fractional Factorial

To reuse the runs from this setup aOFAT experiment ($n + 1$ runs) in a fractional factorial experiment ($2^{(n-k)}$ runs), the choice of the selected fraction as well as the aOFAT is important. If the aOFAT starts with a set of conditions that is very different from the final set the multiple changes will increase the number of runs that could be reused in the factorial experiment. The drawback of this starting set is that the aOFAT was so far from the best setting, it was probably unable to take advantage of interactions and would be less likely to achieve the maximum setting.

If the choice of the fraction is made in advance then for a seven run experiment, on average, 10% of the aOFAT runs can be reused with an equal sized fraction. This estimate depends on how far the random starting location was from the final run location, the number

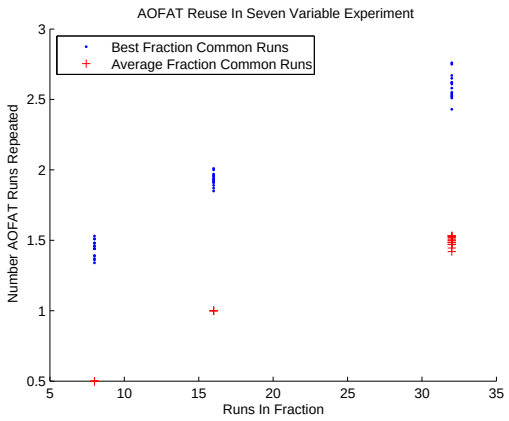

Figure 3-1: Fractional Factorial Run Reuse

of runs exhibits an asymptote as the size of the fraction increases to use n of the $n+1$ runs.

If there is no fraction preference and any fraction is acceptable, then for a seven run experiment, nearly 20% of the runs can be reused with an equal sized fraction. Again, as the size of the fraction increases, the reuse runs asymptotes to a maximum of n runs. This maximum, and not $n+1$, is due to the fact that the variable combinations in the aOFAT can never be completely independent, and thus cannot fit into an orthogonal fractional factorial experiment.

The analysis of the subsequent fractions that were produced after an aOFAT experiment were analyzed using the HPM. This model is well suited to study different fractional factorial designs and the analysis reflects the reality of industrial experiments. Analyzing the results from this model, two fractional-factorial designs that reuse an equal subset of aOFAT runs have no difference in the performance of those fractions to select the maximum setting. The determination of the maximum setting was conducted through an ANOVA analysis of these experiments to select the highest variable setting (Montgomery, 1996). A full linear model was not created because the goal was to select the maximum setting from

the possible experimental points.

Including the aOFAT experimental runs did not influence the outcome of the fractional experiment. There are a couple of potential problems when including the aOFAT experiment in the fractional experiment. First, the aOFAT may limit the selection of fractions to a certain set; and second the non-random run order could limit the experiment to lurking variables. In comparing the results of the best reuse fraction with the remaining fractions using the HPM, there was no difference between the results. This statistical comparison was completed on the 2^{7-4}, 2^{7-3}, and 2^{7-2}, fractions; the results are in Table 3.1. It should be noted that reused aOFAT runs were rarely sequential and the location in the fraction also varied. So while the aOFAT runs are ordered their use in the fraction comes from a random, non-adjacent subset that is used in different locations in the fraction. Although the runs in the fractional factorial experiment are not truly random, they are not ordered and should minimize the effect of lurking variables.

Differences -
Difference = mu (Lv4MaxFrac) - mu (Lv4MinFrac)
Estimate for difference: 0.001455
95% CI for difference: (-0.003278, 0.006187)
T-Test of difference = 0 (vs not =): T-Value = 0.61 P-Value = 0.543 DF = 85
Difference = mu (Lv3MaxFrac) - mu (Lv3MinFrac)
Estimate for difference: 0.008909
95% CI for difference: (-0.012706, 0.030525)
T-Test of difference = 0 (vs not =): T-Value = 0.82 P-Value = 0.415 DF = 85
Difference = mu (Lv2MaxFrac) - mu (Lv2MinFrac)
Estimate for difference: 0.012955
95% CI for difference: (-0.083642, 0.109551)
T-Test of difference = 0 (vs not =): T-Value = 0.27 P-Value = 0.790 DF = 85

Table 3.1: aOFAT Reuse Comparison

The number of runs that can be reused is dependent on the size of the fraction. The relationship is best described by a power function $\text{Reused}_{\text{percent}} = \beta_0 - \beta_1 * 1.1^{-\text{TotalRuns}}$. The asymptote was at 20% and 30% for the average fraction and the best fraction, respectively.

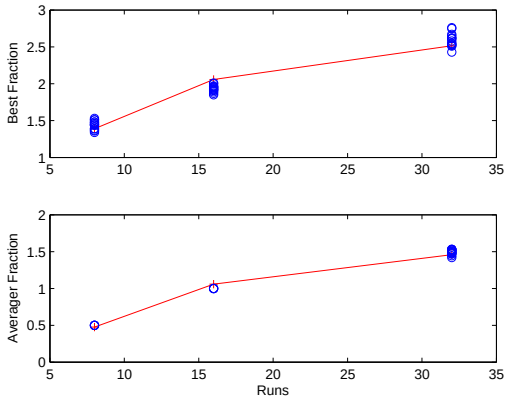

Figure 3-2: Asymptotic Runs Function

To reach 95% of the possible number of reused runs it required 19 and 20 runs for a seven variable experiment.

For reference the β parameters for this model were $\beta_0 = 0.1962$, $\beta_1 = 0.2930$ for the average fraction and $\beta_0 = 0.3304$, $\beta_1 = 0.3360$ for the best fraction. The number of runs that are needed to reach an asymptote can be calculated from this equation. So given to use twenty percent of the best fraction would require 10 runs because $\text{Reused}_{\text{percent}} = 0.3304 - 0.3360 * 1.1^{-10} = 0.20$

3.5 D-Optimal Augmentation

Another augmentation scheme is to use a D-Optimal design to add runs to an aOFAT. Runs are added to a subset of the aOFAT runs that maximize the determinant (hence the D) of the $X^T X$ matrix. We restrict the selections to be orthogonal D-Optimal designs. The use of orthogonal runs minimizes the cross-correlation between variables and greatly aids in interpretation by allowing for more parsimonious models to be constructed. A big difference is that the D-Optimal design is not balanced and so a general ANOVA analysis

can not be used and a regression approach is normally employed.

The orthogonality requirement is particulary appropriate in early screening designs because common aliasing could make creating a follow-up experiment impossible. The selection of the additional D-Optimal designs is done by a selection algorithm. There is no exhaustive search over all of the potential runs as this is practically impossible once there are more than a few variables (approximately seven). A disadvantage to this procedure is similar to the random choice of fractional factorial design, it is not possible for the practitioner to make choices about a desirable aliasing structure.

If the variables have unknown relationships and there is a large number (> 10) of them this aliasing may not be problematic. This is frequently the case for computer experiments. The procedure outlined here is most appropriate for large physical experiments such as turbofan engines, where a screening run is desired. Another option to consider, a space filling design, is not addressed here because it is primarily used to build more complex models, and not for screening experiments.

The selection of a D-Optimal design may not be unique and there are a number of choices for different subsets of the aOFAT experiment. One suggestion is to begin the selection with the latter n runs in the aOFAT and progress forward eliminating the earlier runs. This attempts to include as many of the higher value aOFAT runs as possible. There are other criteria to select the best D-Optimal experiment for the situation and the selection is left to the experimenter.

A final warning is necessary around the use of D-Optimal designs. The creation of these designs is algorithm dependent as in OPTEX program in SAS (Institute, 2000) or cordexch in MATLAB (Math Works, 2007). Because the design space is potentially large an exhaustive search is impossible, or at least impractical, and these algorithms use different sequential optimizers to look for the best points. The risks of those methods are that they

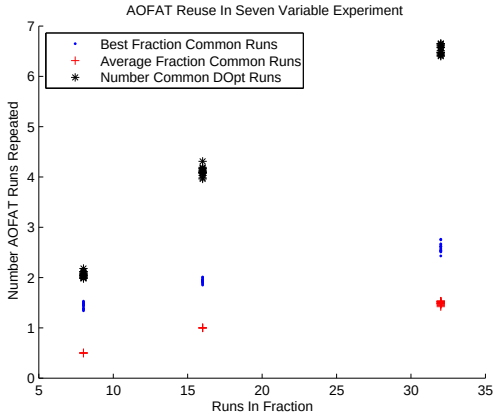

Figure 3-3: D-Optimal Run Reuse

get trapped in local minima or reach a divergent set of possible maxima. Although here the procedure is limited to sets of orthogonal designs there may be multiple solutions for each candidate set. Each of the potential, and equivalent, candidate sets may lead to different system models. As with any experiment it is good practice to follow the guidelines of a experimental statistics book in analyzing the results and iterating as necessary (Wu and Hamada, 2000; Montgomery, 1996; Box et al., 2005).

With this non-balanced procedure, the number of runs for the D-Optimal runs increases the percentage of reusable runs over the fractional factorial. Note that the runs still asymptote to n, the number of variables.

3.6 Maximum Value Choice

One of the benefits from using two experimental methods (aOFAT and a designed experiment) is having two ways of determining the maximum experimental point. The aOFAT model selects the best point based on the last or second to last run. This can be compared with the best predicted experimental point for the fractional factorial model. The analysis

for the fraction can be completed with a traditional ANOVA procedure. Analyzing the D-Optimal experiment requires a regression analysis because the experiment is not balanced.

The Fractional Factorial, D-Optimal, and AOFAT methods were judged on the percentage of times that the maximum experimental value was predicted out of a thousand simulations. The experiment was conducted with seven variables and fractions of 2^{7-3}, 2^{7-4}, and 2^{7-5} runs, the results are averaged over all simulations. The D-optimal experiment used the same number of runs as these fractions. In all of these runs the number of reused runs were maximized. This means that the fraction with the largest number of reused runs was selected; as expected the orthogonal, non-balanced, D-Optimal design had the largest number of common aOFAT runs.

To accurately portray real experiments noise was added to this model as a 0, 5, or 10 times the average effect magnitude times a random number between zero and one. This is a significant amount of variance that accounts for the poor performance of the prediction capabilities of these experiments. It should also be noted that the ability of each of these experiments to predict the maximum is limited because the HPM model has two-way and three-way interactions that cannot be modeled by these reduced run designs.

The results are shown in Figure 3-4. A couple of interesting facts are initially obvious. First the overall performance is quite low, between 20 and 50 percent in predicting the maximum. Again, these are extremely reduced fractional designs and this low performance is expected, but as screening experiments they are still valuable. The second interesting fact is the performance of the aOFAT is comparable to that of the relatively larger fractional factorial and D-Optimal designs. These results are consistent with the previous work on aOFAT also using the HPM model (Frey and Wang, 2005).

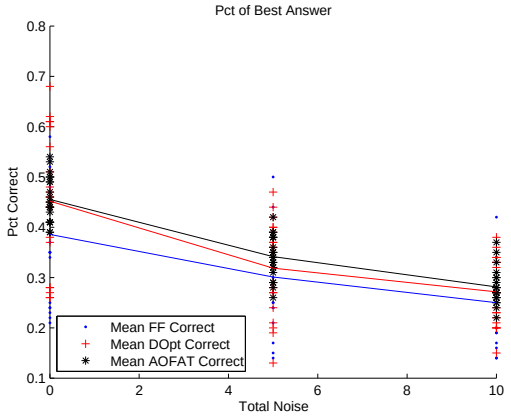

Figure 3-4: HPM percent correct compared with noise

3.7 Sheet Metal Spinning

A specific example was run to demonstrate the run reuse. This example of a sheet-metal spinning process has been used numerous times in the experimental literature; the original data is available in Göbel et al. (2001). The same procedure was run on this data; and the number of reused runs fits the trend seen before. In addition to looking at the number of runs that could be reused, the resulting prediction of the maximal setting was also calculated. This example resulted in an average of only three to five percent correct predictions of the maximum setting. This low fraction is slightly misleading because the values do not change much at the peak. Figure 3-6 shows the average result for the percent of maximum that the experiment predicts. All three predictions are high and make good estimates of the maximum value. As expected, the aOFAT is not dependent on the number of runs in the follow-up experiment.

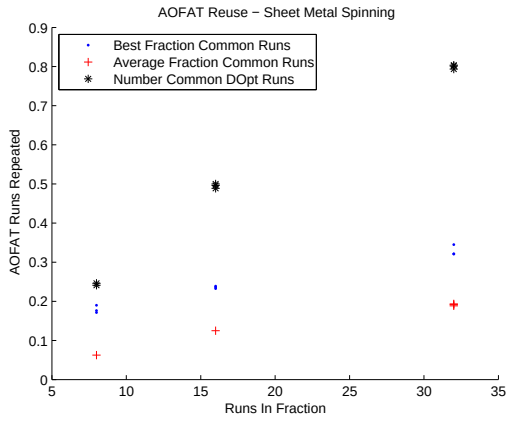

Figure 3-5: Sheet metal spinning repeated runs

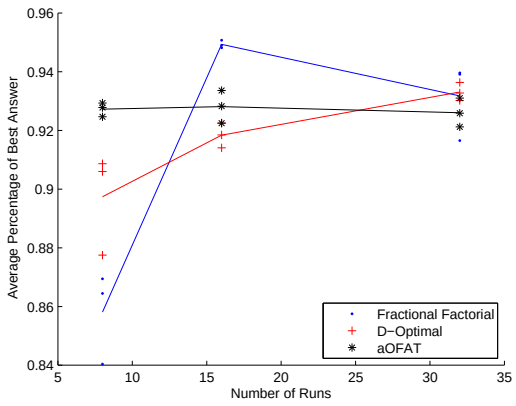

Figure 3-6: Percent of best answer in sheet metal spinning example

3.8 Using Both Models

In addition to being a good practice during experimental setup, there are additional reasons to run both of these experiments. The simplest reason is to serve as insurance. If the experiment fails to run correctly or there are problems such as program cancelation, equipment breakage, or resource limitations then there still exists a good estimate of the best setting. It would also be possible to make a good estimate of the critical variables by looking at the progression of the aOFAT. By looking at the change for each variable and making a correction for possible two-way and three-way interactions, it is possible to get a good estimate of variable effects. This method will be utilized in a later chapter to generate a better variable covariance matrix. In this situation those variable effects could be used to plan follow-up experiments if the first experiment failed.

If this is a production related experiment, a short term improvement could be made by using this setting while waiting for the remainder of the experiment and analysis to be completed. These two different estimates of the maximum have been achieved in different manners and could strengthen or weaken the case for the accuracy of the final model.

3.9 Conclusion

Setting up an experiment through an aOFAT procedure still allows for system understanding while creating potential for run reuse and an independent estimate of the maximum setting. If the maximum number of runs is reused then this extra effort will only cost between $.2n$ and $.5n$ additional runs, depending on the experimental method and total number of runs. The final result is an additional estimate of the maximum that can serve as a temporary stop-gap, insurance to other experimental problems, or as a metric of confidence in

the final model estimates.

This procedure is straight-forward to implement, and selecting the optimal fractional factorial experiment only involves a lookup table. The choice of the cordexch algorithm for finding an orthogonal D-Optimal design is currently very slow and in the cases here took an hour per aOFAT. In applied practice this may be prohibitive and alternative algorithms should be investigated.

Bibliography

Box, G. E. P., Hunter, S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. John Wiley & Sons.

Frey, D. D. and Wang, H. (2005). Towards a theory of experimentation for expected improvement. In *Proceedings of IDETC/CIE*.

Göbel, R., Erdbrügge, M., Kunert, J., and Kleiner, M. (2001). Multivariate optimization of the metal spinning process in consideration of categorical quality characteristics. In *Proceedings of the 1st Annual Meeting of ENBIS, Oslo, Norway*.

Institute, S. (2000). Sas institute inc., sas online doc, version 8. Cary, NC.

Math Works (2007). Matlab. The Math Works, Natick, MA.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall.

Montgomery, D. C. (1996). *Design and Analysis of Experiments*. John Wiley & Sons.

Parker, P. A., Kowalski, S. M., and Vining, G. G. (2007). Construction of balanced equivalent estimation second-order split-plot designs. *Technometrics*, 49:56–66.

Wu, C.-F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, Inc.

Chapter 4

Evolutionary Operation

4.1 Introduction

Evolutionary operation (EVOP) was introduced in the 1950's, popularized by Box and Draper (1969) and grew in use well into the 1980's. The number of academic papers around EVOP has dropped dramatically in recent years although some of the inspired optimization methods continue to flourish. Using recent research on the data regularities in experimental data by Li et al. (2006), the distinction between empirical and scientific improvement will be updated to show that repeated runs are not as detrimental to the system cost and EVOP still has a place in the experimental framework

A suggested framework of single-factor repeated experimentation runs is presented based on computer modeling advances as well as results from adaptive-One-Factor-At-a-Time (aOFAT) experiments (Frey et al., 2003). The method is easy to implement; delivers significant improvement; and incorporates system level considerations. The use of adaptive experiments fits nicely into the overall EVOP process. Taking a larger system view the EVOP is either preceded by, or precedes a traditional statistical experiment. Using an

aOFAT fits well into the framework of the larger system and complements the traditional statistical experiment. If the statistical experiment is run first then the variable order can be selected to maximize the aOFAT result as discussed in Frey and Wang (2005). If the aOFAT precedes the experiment the runs can be reused as shown in Chapter 3 or used to generate a combined model as in Chapters 7 or 8. Using the aOFAT in the EVOP process provides a good method to improve the response while providing a complement to the preceding or proceeding statistical experiment.

4.2 Background

The evolutionary operation procedure was introduced as a production improvement tool that can extend pre-production improvement efforts onto the production floor. The procedure consists of making small variable changes that do not significantly influence product quality. With a sufficient number of these changes, statistical evidence builds to justify making a permanent variable change. This procedure can be thought of as supplying a square-wave between the current and proposed setting of a process variable, or a number of process variables. Although the output remains within performance criteria, given enough time, evidence may accumulate to justify the change. The change justification is based upon a significance test (in most cases a t-test).

The original method defined the goal as searching for scientific feedback to understand the underlying system physics. Although Box and Draper (1969) discussed empirical feedback, their emphasis was on scientific feedback and that is the reference used here. Experimental designs of one to three variables were used repeatedly to drive down the error and improve the manufacturing performance. The method is simple enough to be implemented by manufacturing personnel and accomplished without the need for computer

resources. The process variables under consideration as well as the determination of future experiments is determined by an EVOP committee (Box and Hunter, 1957).

The scientific feedback method differs from optimization or an empirical approach. The goal of scientific feedback is to gather sufficient evidence to be confident in a system model. For empirical evidence the goal is to maximize the system improvement, this could be in terms of profit or another performance metric. The difference in execution of these two goals is the need for run replication.

In implementation, the aOFAT method discussed by Frey et al. (2003) is the same as repeated empirical feedback experiments. Using repeated runs between the settings aligns with the Box and Draper (1969) EVOP procedure for a single factor. It may also be possible to incorporate other models with these single factor EVOP experiments to improve the scientific model while allowing for simple implementation. The use of repeated aOFAT runs is similar to the use of inner noise arrays in Frey and Sudarsanam (2008) when they added a goal of robustness to the experiment.

4.3 Other Models

The traditional EVOP procedure does not use prior system knowledge in the analysis and only requires a variance estimate. There is a suggestion in the end of Box and Draper (1969) that the system knowledge could be used to determine variable transformations. Determining the appropriate variables is part of the responsibility of a committee that organizes the EVOP and they should be aware of the variables used in development. Besides the variable transforms, this near zero starting knowledge for each experiment has the advantage of not making any damaging assumptions but, if the goal is scientific understanding, then the experiments may be inefficient for model exploration. The advantage to scientific

feedback is the general applicability of the knowledge. Other product lines and future developments can draw upon that knowledge to begin with better settings and understanding. The implementation of scientific feedback allows for model refinement in generic manufacturing models and thus better prediction of performance. When the original method was developed in the 1960's these models resided largely in the heads of engineers. The manufacturing advances since then have brought about a profound change in the use of computational power and ubiquity. It is rare to find a manufacturing floor today without a computer, computer controlled operations, and manufacturing simulations.

The goal of EVOP should fit into the larger picture of model improvement and refined understanding of the manufacturing process. Ideally the initial experiments would be performed on a system simulation before being run on the actual processes. These simulations would provide knowledge of the important variables and expected improvement, which would be validated on the actual system. This is different than the initial process set-up with fewer factors investigated and smaller magnitude of changes. When the manufacturing line is initially 'tuned' to run the new product there is normally some experimentation and adjustment to get an acceptable setting. With few runs there are many factors that are insignificant over the noise. These less significant factors could represent significant improvement given greater experimental replication.

Additional models complicate the analysis. Running a larger experimental design in the computer model could then be validated by a final EVOP experiment. There are techniques to merge these computational and physical experiments such as Qian and Wu (2008) which will be explored in later chapters. The complexity and scope of the EVOP experiment should take into account these additional resources. Running a two or three factor experiment as suggested in Box and Draper (1969) may be excessive and a single factor experiment such as Box et al. (2005) may be just as informative. The single factor ex-

periment is suggested here for the simplicity in execution while providing information to more complete computer models. While the single factor experiment is not able to predict interactions Frey and Wang (2005) showed that it has a high probability of benefiting from them.

4.4 Run Size

The criteria for selecting the run size is dependent on the size of the effect and the amount that the variable is changed. The normal test for detecting these differences is the student-t statistic (Box et al., 2005). In Box and Draper (1969) the use of the normal significance tests is preferred based on a standard deviation from a number of EVOP cycles (> 15). Another perspective is that the run size will be dependent on the amount of acceptable variance that can be introduced into the system without detriment to the output. Taking the approach of a system view, determining the run number based on the acceptable increase in variance seems most appropriate (a similar analysis is performed in Box and Draper (1969, pg. 211)).

Given e_i as the estimate of the effect at iteration i then the ratio of that to the variance, $e_i^2/\sigma_e^2\,\chi_p^2$ follows a Chi-Squared distribution. Given the actual values E_i, and assuming no interactions then this follows a non-central Chi-Squared distribution $\chi_p^2(\sum(E_i^2/\sigma_e^2))$. Given the probability of type-I error (incorrectly including a significant effect) at α and the type-II error (missing a significant effect) at β, these probabilities can be used to solve for the sample size.

The overall variance can be estimated as $\sigma^2 + 1/4\sum(E_i^2)$. If the standard deviation is set to change by $k \cdot \sigma$ and using the fact that $\sigma_e^2 = 4 * \sigma^2/n2^p$, then the non-central parameter

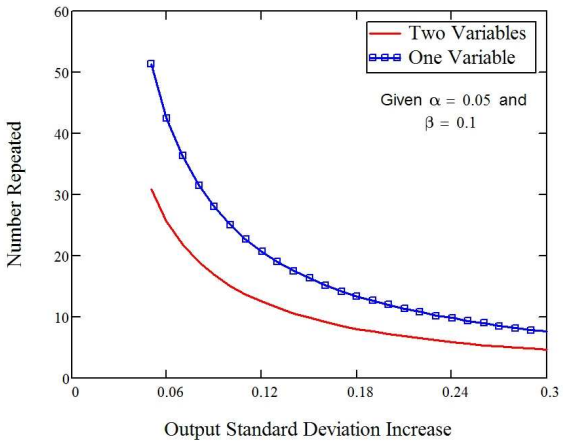

Figure 4-1: EVOP Terms Repeated

can be calculated as:

$$\sum(E_i^2/\sigma_e^2) = n2^p(k^2 - 1) \tag{4.1}$$

Setting the two errors equal to each other it is possible to solve for the minimum number of samples.

$$\chi_p^{2^{-1}}(1-\alpha) = \chi_p^2(n2^p(k^2-1))^{-1}(\beta) \tag{4.2}$$

This estimate is accurate if the interactions are insignificant, and will provide a good estimate of the required runs. The result is shown in Figure 4-1, the number of runs decreases dramatically as the acceptable standard deviation increases. With two variables the number of points repeated decreases by 40 percent due to the shared variance estimate.

Run randomization is implicit in these results. Sets of runs between the two settings are conducted with random order. The suggestion in Box and Draper (1969) that the randomization is not critical is proven in Box (1954) for serial correlation. The foundation of that paper is a correlation between runs and a wide variance for sets of runs. Generalizing those results is cautioned as the tri-diagonal correlation matrix is a full matrix for two settings.

The serial correlation from run to run also affects the correlation between sets of runs. Before using a regular repeating experimental pattern the assumption that the sets of runs are independent needs to be verified. Additionally, latent variables could complicate the experiment and may even lead to erroneous conclusions. The maximum inferential power requires run randomization.

Ideally the knowledge gained from the EVOP is utilized to improve a manufacturing model. If there is no model to improve, or the knowledge gained will not be reused, an empirical feedback, or an optimization goal is more appropriate. If the EVOP committee wants to use scientific feedback the experiments should explicitly take advantage of, and benefit from, any current manufacturing model. The planning committee should minimally utilize variable sensitivity analysis along with any previous tuning results. Further, these models could assist in variable selection, interaction estimation, range and variance prediction, and output estimation. An efficient method of extracting useful data out of a computer model is through computer experimentation for one example methodology see Santner et al. (2003). The result should be a candidate list of likely important parameters. This list should be augmented by the practitioners knowledge of potential opportunities and non-optimal parameters. Running a scientific EVOP on these parameters may reveal surprising interactions and improve the effect precision. In addition to improving production, this information is used to improve and update the model. It is these model improvements that are the most valuable to continually improving the performance of the organization.

There are also many methods to incorporate the computer model and the experimental data to make a dual predictive model for that particular system. The best known are Kennedy and O'Hagan (2001); Qian et al. (2006), if the computer model is also stochastic then the approach of Qian and Wu (2008) works well. These methods rely on a Bayesian approach of combining both types of data to produce better prediction. Although specific

to a particular production line, these models can be used to suggest better operating conditions, and quantify model deviations.

4.5 Comparison to Optimization

Production improvement differs from true numerical optimization in a number of important ways. The terms improvement and optimization have been used loosely here but the difference is important. In production improvement, the objective function is unknown and changing, the number of input variables is not fixed, and the range of input variables is not fixed. Optimization requires an objective metric over which to maximize (or minimize), using a fixed, known, set of variables. In the production world the precise objective function can change periodically as the production rate, material cost, overhead burden, and corporate profit needs change. Thus for each cycle of the EVOP different criteria may be used to measure success. The number of input variables is not fixed, and given a desirable improvement direction it may be feasible to add variables that can help with that improvement. For example if a particular temperature increase improves performance then it may be deduced to add other temperatures from a range of other locations. Finally, it may be possible to change the range of each of the variables if an improvement is noted. This could be as simple as changing a process sensor (with a higher temperature rating) to changing the mechanics of the system (inductive heaters from ceramic). The difference to an optimization procedure is evident in the details of the implementation, and the complete system is critical to the improvement. This reinforces the importance of an EVOP committee to have a system perspective and continually monitor and react to the changing environment.

4.6 Empirical Improvement

Empirical based improvement or feedback has also been referred to as ‘idiot’ feedback by Box and Draper (1969). The negative connotation about this improvement methodology is that afterwards, although it may yield an improved setting, there is no additional knowledge about the system model. The experimenter must decide if the effort to get a more complete model outweighs the cost of experimentation or delaying the implementation of improvements. The models that are generated maybe limited in time and scope to the particular problem at hand and may or may not be valid in a more general future problem. Only if the model has general utility could it be reused and the resulting improvement could have multiplicative benefit to the organization. An experimenter should also consider the risks of following an empirical feedback plan where many of the changes are detrimental.

There are two general methods for gathering feedback. First getting multiple data points for any change, and thus gaining statistical confidence in the scientific foundation of that change; or second reacting to every data point to make as many variable changes as possible, thus making many more changes. Box and Draper (1969) proves that the most profitable method is to utilize a single data point to make a decision on a variable setting. This analysis is aligned with the published aOFAT technique in Frey et al. (2003), although with differences for the amount of noise in the system. The original analysis does not consider other costs associated with making a production change including retraining, updating manuals, or changing production drawings. The final cost may also include some of the risks that occur during a transition such as extra scrap or lower productivity. The suggestion here is to utilize scientific feedback to improve production models and gain greater benefit to the organization. This section attempts to show the potential loss for using scientific feedback versus empirical or optimization feedback. The surprising conclusion, using more accurate

effect distribution is that the incremental cost of replicates is small. If a validation run is made for the other system considerations before implementing empirical feedback then the cost of continuing on to statistical significance is minimal, the mathematical details follow.

Given that effects have an exponential probability distribution (Li and Frey, 2005) and that the ability to detect the effect follows a normal distribution a monetary loss can be calculated for switching one variable. In choosing a loss function the calculation is not dependent on the number of variables examined or on the total number of experiments.

The frequently used cumulative normal distribution is historically used to estimate the probability of detection:

$$F(\zeta) = \int_{\zeta}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \tag{4.3}$$

Research into variable distribution points to an exponential distribution with $\lambda \approx 0.007$, from Li and Frey (2005):

$$f(\zeta; \lambda) = \lambda e^{-\lambda\zeta} \tag{4.4}$$

Given experimental noise σ, along with a cutoff value, ξ and n runs an estimate of the loss in delaying any variable change can be determined. Given a large possible number of changes K.

$$L = \sum_{i_1}^{K} n + i \cdot \int_{-\infty}^{\infty} u f(|u|; .007) F(\frac{\sqrt{n}(\xi - u)}{\sigma}) \tag{4.5}$$

This summation can be expanded, and the value of σ can be approximated as $1/1.2 \cdot \lambda$, as determined by Box and Draper (1969) as a large variance relative to the variable changes.

$$L = K \cdot n + \frac{K^2 + K}{2} \cdot \int_{-\infty}^{\infty} u f(|u|; .007) F(\frac{\sqrt{n}(\xi - u)}{\sigma}) \tag{4.6}$$

Depending on the value of K this loss increases as the number of samples increases; this

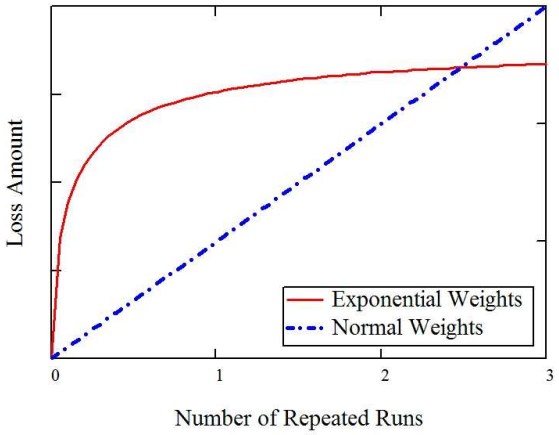

Figure 4-2: Repeated Runs

can be seen in Figure 4-2. Both the traditional and this new analysis show monotonically increasing loss with additional runs. This result led to the conclusion that the fewest possible number of runs before switching maximizes the profit or minimizes the loss. The major assumption built into this analysis is that the expected improvement is centered around zero. This means that any change has equal probability of making an improvement or causing a detriment, this seems like the most pragmatic situation, as Box and Draper (1969) also concluded. In this analysis the ξ value was chosen to minimize the loss given any number of runs, in this case ($\lambda = .007, \sigma = 1/1.2 \cdot \lambda$) the value is close to zero.

If no runs are repeated a large number of the changes will be incorrect, this is outweighed by the correct changes, and will minimize the total expected loss. The difference from the original analysis is the exponential distribution shows a faster asymptote towards a fixed loss. The exponential distribution has more effects near zero; these benefit from a few additional runs. The normal distribution has enough weight in the tails that the best strategy is to get through as many as possible to find these big effects. In general, both results show that the strategy to achieve the greatest gain in a fixed period of time is to get through as

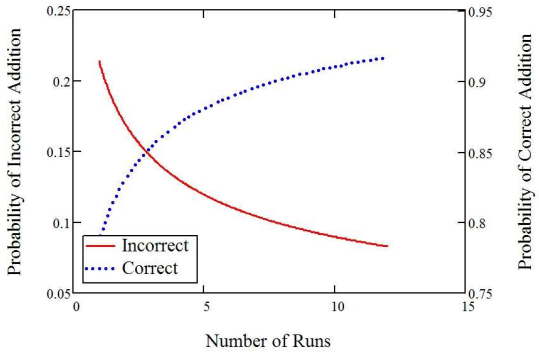

Figure 4-3: Single Variable Probability Incorrect

many variables as possible. However, there is a difference in the benefit of a few repeated runs in selecting the winner. Most pragmatic managers are not willing to make frequent changes to their operation without some evidence that the benefits would outweigh the risk. There are additional risks with the single run strategy of seeing a nonlinear response, a change in variance, or an uncontrollable condition.

As a practical suggestion for empirical improvement, the cost of making small change in profit should be compared with the confidence with additional runs as shown in Figure 4-3. If the cost and probability of negative effects is minimal then the single run strategy might be best. If the costs may be significant or the profit difference is not practically significant then additional runs should be considered.

The ability of empirical improvement to reach a better solution with exponentially distributed variables and a normally distributed confidence has been shown. Adding a second run increases the loss (or reduces the profit) by 7.4%, this change should be weighed against reducing the probability of an incorrect variable by 20.7%. If the variance is the same, this compares with the original method which increases the risk of loss to 15.5%; and it reduces the probability of an incorrect variable by 21.6%. Determining the number of repeated runs

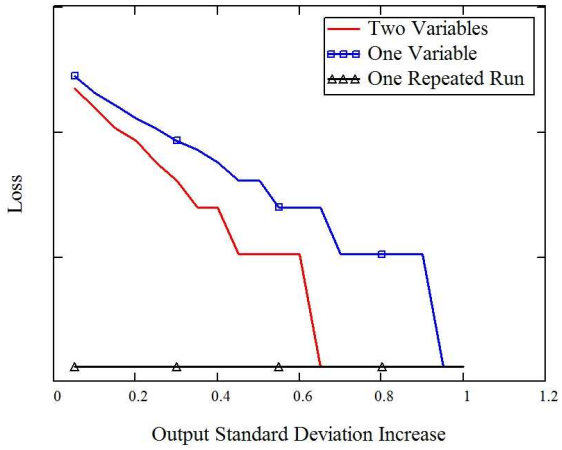

Figure 4-4: Single Variable Loss Standard Deviation

should be based on the manufacturing environment as well as the managerial tolerance of risk. The empirical feedback remains the best strategy to achieve the greatest gain in minimal time and these results show that the cost of repeating runs is not as severe as initially proposed. If the acceptable system standard deviation increase is high enough, the loss due to the repeated runs is minimal as seen in Figure 4-4. If there is a one standard deviation acceptable increase then the loss is comparable to one repeated run. Considering the organizational risk tolerance and the acceptable amount of variance, the empirical and scientific feedback may overlap.

4.7 Additional Runs

Box and Draper (1969) suggests experimental sizes of two to three variables, that are repeated until achieving effect significance. There are conflicting ideas when considering the amount of noise in an experiment. Reducing noise calls for additional runs, this is traditionally viewed as a reduction of $\sigma/\sqrt{n}$ in the standard deviation. In the experimental case

of $2-3$ variables, to reduce the variance by half requires 2^{n+1} extra runs while the aOFAT only requires $2 * (n+1)$, optimally. This number of extra runs indicates homoscedastic conditions that are rarely seen in actual experiments. In practice, there is series correlation between the variables when deciding on the number of runs. One way to capitalize on the improvement and stay close to the minimum number of variables is to adjust stochastically. A procedure might be arranged like this:

1. Choose between two-variable settings (High and Low) randomly with probability based on the number of points that already exist $p_{\text{high}} = 1 - \frac{\text{Num. High Points}}{\text{Total Points}}$.
2. Run the experiment
3. Complete a paired t-test between the two settings.
4. At probability equal to t-value (t) run another experiment changing the next variable, with the current variables at their current expected maximum setting.
5. Choose a threshold (.9) over which the t value is significant and the experiment is advanced to the next variable.

This procedure is similar to a Gibbs sampler on a uniformly distributed random variable. Gelman et al. (2003) gives a good description of why this procedure generates a long term stationary distribution that mirrors the variable importance probability. A key difference in the usual Gibbs implementation is the short run duration. In this case we only run for the number of variables, in most Gibbs conditions the sample number approaches the thousands. In the limit of very few runs Tanner and Wong (1987) has shown that the direction of the runs is still correct.

When this procedure is run against the traditional factorial experiment the results are similar for resolution and power. For example, when this procedure was run against a half-

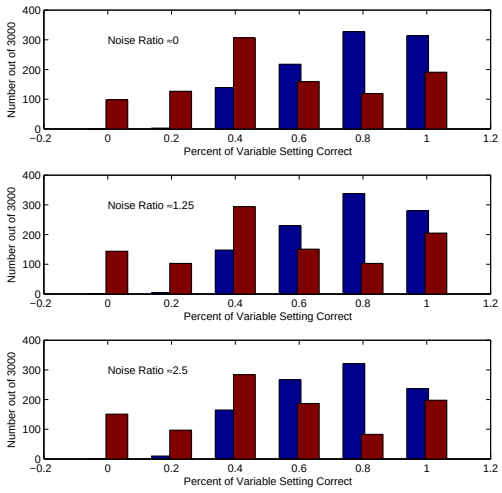

Figure 4-5: Response with noise variation, aOFAT is blue on the left and Fractional Factorial red on the right.

fractional factorial to get the same resolution for six variables, each experiment required the same number of runs. Afterwards the adaptive procedure had an estimate of the maximum but the fractional experiment could also create a model of the system. The situation changed when a series of experiments were made in a typical EVOP framework. In this case two sequential full-factorial experiments of three variables were compared with a statistically significant aOFAT experiment. Both of these experiments required approximately the same number of runs ($\approx 12 - 16$). For this comparison both procedures considered predictions for each variable of high, low, or unknown. This unknown prediction utilized the t-test for the aOFAT and variable significance (F-test) for the factorial experiment.

The results of the experiment for different levels of noise are shown in Figure 4-5. These results are consistent with the work of Frey et al. (2003) - at low levels of noise the aOFAT procedure has a higher likelihood of selecting the best setting. At higher levels of noise the procedures are comparable.

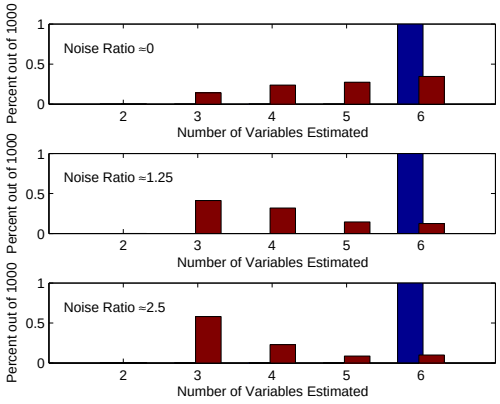

Figure 4-6: Number of modeled variables, aOFAT is blue on the left and the Fractional Factorial red on the right.

The reason for the disparity in performance is due to the sequential nature of the experiment. As suggested in Box and Draper (1969) deciding on the next sequence of runs should be made by committee. If the probability of interaction is high for variables between factorial experiments then this procedure does a poor job of estimating the maximum. While the number of runs was comparable, a committee may have chosen a better arrangement given the six variables under investigation. Using the model building approach by result significance led to fewer important variables in the factorial experiment compared with the aOFAT t-test approach. The reduction in the number of variables is dependent on the amount of noise added to the system. This can be seen in Figure 4-6, as the amount of noise increases the number of significant variables decreases.

The aOFAT methodology offers a less intensive approach to determining improved manufacturing conditions. Through sequential measurements and straight forward t-tests, there is a high likelihood of selecting the best operating conditions. This result is better than running a series of factorial experiments on the same variables. The best situation is to rely on the accumulated experience to make good variable selections and implement

them through an EVOP committee. In this well run situation an aOFAT is also the quickest way to the preferred conditions. As Frey and Wang (2005) showed, if the variable order is known then an aOFAT will benefit from interactions and offers the quickest path to check every variable.

4.8 Conclusion

Evolutionary operation (EVOP) is a statistical method for process improvement during manufacturing. Utilizing small repeated experiments the operating condition can reach more preferred conditions. The foundation for this method should be to refine the manufacturing models and system understanding. With the increase in computational and simulation power more manufacturing processes have accurate models that assist in the design and parameter settings. The validation and verification of these models is challenging and run size limitations may yield unacceptable meta-models (Irizarry et al., 2001).

Evolutionary operation can improve these models while improving the current manufacturing system. The cost of this improvement strategy is an increase in short-term production variation. A six-sigma production facility is designed for a 1.5 sigma long-term shift. If a fraction of this margin is used to improve the process it can result in better future models, cost savings and quality improvement.

The suggested feedback mechanism here is empirical aOFAT experiments that are statistically significant. With more accurate effect distribution information, gathering statistically significant feedback increased the loss by 7.4% for two runs versus one, compared with 15.5% with the historic normal distribution. Additional repeated runs have an even smaller profit reduction and should be used in context with the organizational risk tolerance and change cost. The small difference between empirical and optimization feedback

drives model based experimentation that can offer long term corporate wide benefit at little increased cost. It has been shown that running repeated factorial experiments has potential accuracy and variable size drawbacks compared with a Gibbs based aOFAT sampling technique.

The use of evolutionary operation has a place in the manufacturing environment to improve production as well as validate models. Sequential Gibbs-based aOFAT experiments offer a practical and efficient way to implement empirically-based EVOP.

Bibliography

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, ii. effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 25:484–498.

Box, G. E. P. and Draper, N. R. (1969). *Evolutionary Operation: A Statistical Method for Process Improvement*. John Wiley & Sons, Inc.

Box, G. E. P. and Hunter, J. S. (1957). Multi-factor experimental designs for exploring response surfaces. *Annuals of Mathematical Statistics*, 1:195–241.

Box, G. E. P., Hunter, S., and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*. John Wiley & Sons.

Frey, D. D., Englehardt, F., and Greitzer, E. M. (2003). A role for "one-factor-at-a-time" experimentation in parameter design. *Research in Engineering Design*, 14:65–74.

Frey, D. D. and Sudarsanam, N. (2008). An adaptive one-factor-at-a-time method for robust parameter design: Comparison with crossed arrays via case studies. *ASME Journal of Mechanical Design*.

Frey, D. D. and Wang, H. (2005). Towards a theory of experimentation for expected improvement. In *Proceedings of IDETC/CIE*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Irizarry, M. D. L. A., Wilson, J. R., and Trevino, J. (2001). A flexible simulation tool for manufacturing-cell design, ii: response surface analysis and case study. *IIE Transactions*, 33:837–846.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 63:425–464.

Li, X. and Frey, D. D. (2005). A study of factor effects in data from factorial experiments. In *Proceedings of IDETC/CIE*.

Li, X., Sudarsanam, N., and Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, 11(5):32–45.

Qian, P. Z. G. and Wu, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50:192–204.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, C. F. J. (2006). Building surrogate models based on detalined and approximate simulations. *ASME Journal of Mechanical Design*, 128:668–677.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–.

Chapter 5

Sequential Simplex Initialization

5.1 Introduction

An improved sequential simplex starting routine is presented based on adaptive-One-Factor-at-a-Time (aOFAT) experimentation (Frey et al., 2003). The adaptive $k + 1$ points as a starting simplex improves convergence as well as reduces the number of iterations. The proposed method generates an initial simplex by adjusting each parameter by a small delta sequentially and leaving any parameter change that brings the function closer to its target. This initialization is permitted in the original Nelder-Mead procedure (Nelder and Mead, 1965) with the only limitation that any initial simplex is non-degenerate. In addition to the change in the starting simplex, the delta is adjusted to account for an increased distance between experimental points and the centroid. The proposed delta adjustment is based on the probabilistic variable selection which sets the step equal to that of the old routine. A suite of 35 test routines provided by Moré et al. (1981) is used to demonstrate the effectiveness of this change in improving convergence and reducing the number of iterations.

5.2 Background

The original simplex procedure is from Spendley et al. (1962) it provided a sequential unconstrained optimization procedure that is geometrically based. This procedure was limited by a fixed step size and was quickly replaced by the variable step size procedure of Nelder and Mead (1965). Although the procedure is now over forty years old, it still is seen in numerous applications. Both MATLAB and Mathematica use the routine in fminsearch and NMinimize, respectively. The routine is also presented in the book Numerical Recipes as the amoeba routine (Press et al., 2007).

The exact details of the numerical procedure and its convergence is not discussed here because there are many good references available (Press et al., 2007; Walters et al., 1991; Lagarias et al., 1998) although, a short outline of the procedure is provided for familiarity. The sequential simplex starts with $k+1$ initial data points arranged in a geometric simplex pattern, and the function evaluation at those points. There are five steps to iterate the procedure, in this case given for a function minimization:

1. **Order.** Put the $k+1$ points in descending order of their function values f_i. Ties may be broken by looking at the index value (Lagarias et al., 1998).

2. **Reflect.** Compute a reflection point $x_r = (1+\alpha)\bar{x} - \alpha x_{k+1}$. Note that $\bar{x}$ only includes points up to k. If the new value falls within the current values, $f_1 \leq f_r \leq f_k$, then iterate. Nelder and Mead use $\alpha = 1$

3. **Expand.** If $f_r < f_1$, thus it is a new minimum, expand the simplex by calculating an expansion point $x_e = \gamma x_r + (1-\gamma)\bar{x}$. If $f_e > f_r$ then accept the expansion point and iterate otherwise, accept f_r and iterate. Nelder and Mead used $\gamma = 2$.

4. **Contract.** If $f_r \geq f_{k+1}$, it is the worst point $x_c = \beta x_{k+1} + (1-\beta)\bar{x}$, if $f_k \leq f_r < f_{k+1}$

then $x_c = \beta x_r + (1 - \beta)\bar{x}$. Nelder and Mead used $\beta = .5$. If $f_c \leq f_{k+1}$ then accept f_c and iterate.

5. **Shrink.** If $f_c > f_{k+1}$ then the contraction failed and all points except for x_1 should be replaced by $x_i \leftarrow \delta x_i + (1 - \delta)x_1$. Nelder and Mead used $\delta = .5$, but other authors suggest $\delta = .9$ (Barton and Ivey, 1996).

This procedure is gradient free and determines future points only based on the rank order of the values. It has been shown that this procedure does ultimately converge to a minimizer for general (non-convex) functions of one dimension (Lagarias et al., 1998). There is still work remaining as to why this procedure works so well in practice. For example, there is no known function in $\Re^2$ for which the procedure always converges to a minimizer. A number of degenerate situations have been demonstrated where this algorithm does not converge, which may be dependent on the starting simplex (McKinnon, 1998).

5.3 Initializing the Simplex

In previous work on this iterative procedure an initial simplex is often assumed and the generation of those initial points has not been very well studied in the literature. Spendley et al. (1962) begin the procedure with a regular *k*-dimensional simplex. A simplex of dimension *k* can be defined as the convex hull of a set of *k+1* affine independent points in Euclidean space of dimension *k* or higher. The regular simplex is a regular polytrope, and so all points are separated by a common edge length. Although the Nelder-Mead algorithm nominally starts with a regular simplex successive simplices do not remain regular due to the Expand and Contract steps.

In the algorithmic implementation the suggestion for getting the initial simplex points begins with a single starting point P_0 and e_i unit vectors. The remaining k initial points just represent small orthogonal deviations from that point, calculated from Equation 5.1 as suggested by Press et al. (2007).

$$P_i = P_0 + \Delta e_i \tag{5.1}$$

The Δ's could either be a single value for all variables or specific values for each direction. The MATLAB (Math Works, 2007) implementation only uses this freedom when $P_0 = 0$ and sets $\Delta = .00025$, and for all other values $\Delta = \delta \cdot P_0 \cdot e_i^T$ and $\delta = 0.05$, or a change of 5% of the current variable value. This is noted in the code as a suggestion of L.Pfeffer at Stanford (further reference could not be found).

This widely used starting procedure does not generate a regular simplex. This can be shown in two dimensions when three points form a right triangle and not the regular 2-simplex of an equilateral triangle, this is also true for higher dimensions. Having a regular simplex is not required because given any non-degenerate (volume≠0) starting simplex all following simplicies are also non-degenerate; for the proof see Lagarias et al. (1998). This implies that any non-degenerate simplex may be a starting point and will not affect the algorithms degeneracy.

5.4 Proposed Improvement

The method proposed here to improve this initialization consists of first, a better choice of starting conditions and second, choosing the step-size based on the distance to the reflected point. Initially, the variables are changed in order, as before, but, if a change yields an improvement then the remaining variable changes progress from this point. This procedure

is identical to performing an aOFAT experiment (Frey et al., 2003). These $k+1$ runs would 'aim' the simplex in the most likely direction of improvement. Given that this is a hill-climbing algorithm, this would ideally decrease the number of additional runs. An added benefit may be that a directed starting simplex will move away from cyclic, or stalling points. There are essentially no theoretical results for the sequential simplex in dimensions greater than two, and so better initialization may help avoid the problems pointed out by McKinnon (1998) and Hall and McKinnon (2004).

The simplex procedure is geometric, and the next trial point is based on the distance from the current worst point to the centroid of the remaining points. To match the traditional algorithm's distance the increase Δ would have to be set dependent on the number of x-variables k as follows:

$$\Delta \propto \frac{1}{k}\sqrt{k^2 + \frac{1}{k} - 1} \tag{5.2}$$

The traditional procedure is not a regular simplex and so this value is the weighted average of the origin and the orthogonal points. This distance asymptotes to one, or towards the desired delta. With the aOFAT procedure this distance correction is more probabilistic. If the probability of any variable making a positive change is p (nominally assumed to be 0.5), and given k variables, then the expected distance value can be given by the following.

$$\Delta \propto \frac{1}{k}\sqrt{\frac{p \cdot k^3}{3} - p \cdot k^2 - \frac{p \cdot k}{3} + p + k^2 + \frac{1}{k} - 1} \tag{5.3}$$

This is the weighted distance for the largest and smallest points in the simplex. For each variable change there is a p probability for making a change and moving the centroid, as p goes to zero the results are the same as in Equation 5.2.

The original method asymptotes while the modified approach does not. Most implementations do not include this asymptote based on the number of variables. The reasoning

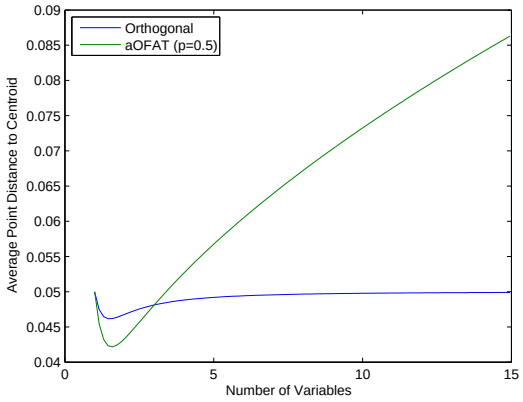

Figure 5-1: Distance to Centroid

can be seen in Figure 5-1; after about five variables the value does not change substantially. The modified approach will have to take into account the number of variables as the delta continues to grow without asymptote.

The starting step-size will be modified before the algorithm runs based on the number of variables. The step-size could be modified *in-situ* based on the acceptance of a variable but, if we only want to run each setting once it could only modify the subsequent variables. This would add a dependency to the algorithm based on the order of the variables, something to avoid for this generic solver.

5.5 Improvement Considerations

It is possible to say very little about this procedure without some assumptions about the function over which it is applied. Given a strictly convex function in two dimensions with bounded level sets and coefficients $\alpha = 1$, $\gamma = 2$, and $\beta = 1/2$, Lagarias et al. (1998) showed that given simplicies (Δ_n) generated at the n^{th} iteration of the algorithm the limits are as follows:

$$\lim_{n\to\infty} \mathrm{vol}(\Delta_n) = 0 \tag{5.4}$$

and

$$\lim_{n\to\infty} \mathrm{diam}(\Delta_n) = 0 \tag{5.5}$$

The convergence is dependent on the volume and diameter change at each step. It is not possible to determine if a greater or lesser volume or diameter will improve convergence at each step but that the overall convergence is sensitive to volume and diameter changes.

For both starting simplicies the ratio of the volume change is the same, it is only dependent on α, γ and β. This is true because the amount that a point changes is proportional on the distance between that point and the centroid, which balances out the smaller volume change for points further from the centroid. Although the change in volume is the same and thus the rate of convergence at that point, the volume of the two initial simplicies are both proportional to Δ^k. The volumes are similar until the difference in the delta's becomes large. As seen in Figure 5-2, the modified starting simplex has a reduced starting volume that may increase the number of iterations although the rate of volume change is the same and so this should not affect the convergence.

This smaller volume is a tradeoff to keep the distance of the initial simplex move the same as the original routine. Each move, either expand or contract, is dependent on the distance from the centroid to the reflection point. The original algorithm had two possible values for that distance either the origin or any other point. In the proposed algorithm we used the final point and the middle point in the aOFAT simplex to calculate an average delta. This is a simplification, although the smallest point is in the middle, the largest step may also be one of the first two points. If the first few variable changes are accepted versus the final few, then the centroid is far from the start and close to the end, and the biggest

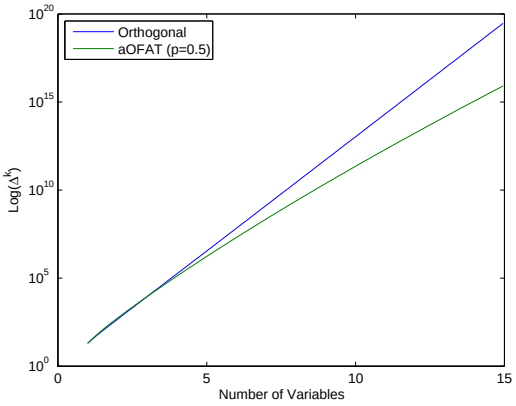

Figure 5-2: Volume of Simplex

step is in the first point. On the other hand the final variable is only changed once and so the final point is the furthest from the centroid of the k^{th} variable. These two counteracting effects are compared by the percentage of times this point is greater than the last point, and the amount that it is greater. The relationship between the percentage of other points as well as the error is shown in Figure 5-3. This gives a weighted error between 0.7% and 3.7%, for the proposed method depending on the number of variables. The second problem is taking a weighted average of the middle (smallest) point and last (largest) point does not reflect the distribution of these variables. If the beginning point is larger than the final point the weighted average is too small and underestimates the average distance. Because these two errors are both small and occur in opposite directions they are not included in the proposed model.

5.6 Test Cases

The aOFAT starting condition as well as the step-size change were implemented in MATLAB by changing the current fminsearch routine and run against the standard test suite by

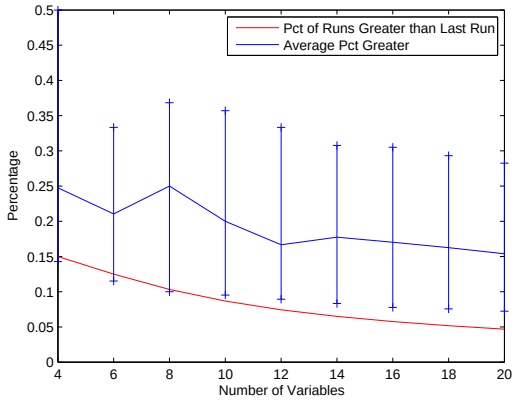

Figure 5-3: Centroid Distance Comparison

Moré et al. (1981). These 35 functions were designed to test the extremes of unconstrained optimization and have been used as a metric for changes to the Nelder-Mead procedure a number of times (Nazareth and Tseng, 2002; Price et al., 2002). The procedure was run with a maximum number of iterations of 10^5, a maximum of 10^8 function evaluations, a tolerance of 10^{-12} on the output and a tolerance of 10^{-9} on the x values, the results are shown in Table 5.1.

These results show a benefit for the new method. One proposed comparison metric has been the $\log(\hat{f} - f)$ (Barton and Ivey, 1996). The original method was -13.37 with the proposed method -14.46. Although the change was only made in the starting $k + 1$ simplex points, this yielded an improvement to the accuracy of the final results. Two of the test problems that originally did not converge now converged correctly with the modified procedure.

Looking at the runs with a similar metric $\log(n)$ shows the improvement with the new procedure. For this calculation the two problems that reach the maximum number of iterations were left out (10 and 16). When the procedures reached different local minima or

failed to converge those problems were also not counted in the run metric (problems 18, 20, 21, 25, and 26). Removing these problems favored the original routine because there is no penalty for not converging or using the maximum number of iterations. The original routine had a log-run value of 2.73 and the modified routine 2.72. In the test problems this represents a 2% savings in runs or an average of nine fewer iterations.

In this difficult test suite of functions the improved convergence is evident to a greater degree than the iteration decrease. This is attributed to the challenge of this problem set, and the fact that without a good starting direction it is easy to get trapped in cyclic or stalling situations. In the majority of smoother, and more realistic, applications it is predicted that the decrease in runs may be larger. The proposed routine did not lead to any major decreases in performance. The modified routine sacrificed some of the possible run reduction by making the initial step sizes similar. If instead, the volumes were maintained, the step size would have increased lowering the number of iterations.

The code for this modified routine is available from the Matlab file exchange website (http://www.mathworks.com/matlabcentral/fileexchange/).

5.7 Conclusion

An improved sequential simplex starting routine based on adaptive-One-Factor-at-a-Time (aOFAT) experimentation was proposed. Starting with this new simplex improved the eventual convergence (on two of the 35 test cases) as well as reduced the total number of iterations by 2%. The proposed method generates an initial simplex by adjusting each parameter sequentially and leaving any parameter change that brought the function closer to its target. This starting simplex is permitted in the original Nelder-Mead as long as it is non-degenerate. In addition to the change in the starting simplex the delta is adjusted to

account for an increased distance between experimental points and the centroid. This delta adjustment based on the probabilistic variable selection and decreases the initial volume of the simplex. A suite of 35 test routines provided by Moré et al. (1981) is used to demonstrate the effectiveness of this change in improving convergence and reducing the number of iterations.

Bibliography

Barton, R. R. and Ivey, Jr., J. S. (1996). Nelder-mead simplex modifications for simulation optimization. *Management Science*, 42(7):954–973.

Frey, D. D., Englehardt, F., and Greitzer, E. M. (2003). A role for "one-factor-at-a-time" experimentation in parameter design. *Research in Engineering Design*, 14:65–74.

Hall, J. A. J. and McKinnon, K. I. M. (2004). The simplest examples where the simplex method cycles and conditions where expand fails to prevent cycling. *Mach. Pragram.*, 100:133–150.

Lagarias, J. C., Reeds, J. A., Wright, M. H., and Wright, P. E. (1998). Convergence properties of the nelder-mead simplex method in low dimensions. *SIAM Journal of Optimization*, 9(1):112–147.

Math Works (2007). Matlab. The Math Works, Natick, MA.

McKinnon, K. I. M. (1998). Convergence of the nelder-mead simplex method to a nonstationary point. *SIAM Journal of Optimization*, 9(1):148–158.

Moré, J. J., Garbow, B. S., and Hillstrom, K. E. (1981). Testing unconstrained optimization software. *ACM Transactions on Mathematical Software*, 7(1):17–41.

Nazareth, L. and Tseng, P. (2002). Guilding the lily: a variant of the nelder-mead algorithm based on golden-section search. *Computational optimization and Applications*, 22(1):133–144.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Journal of Computation*, 7:308–313.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). *Numerical Recipes: The Art of Scientific Computing (3rd Edition)*. Cambridge University Press.

Price, C. J., Coope, I. D., and Byatt, D. (2002). A convergent variant of the nelder-mead algorithm. *Journal of Optimization Theory and Applications*, 113(1):5–19.

Spendley, W., Hext, G. R., and Himsworth, F. R. (1962). Sequential application of simplex design in optimisation and evolutionary operation. *Technometrics*, 4(4):441–461.

Walters, F. H., Lloyd R. Parker, J., Morgan, S. L., and Deming, S. N. (1991). *Sequential Simplex Optimization*. CRC Press.

			Minimum Function Value			Total Iterations	
Num	Name	Order	Actual	fminval	Modified	Original	Modified
1	Rosenbrock	2	0	8.85E-20	6.85E-20	123	130
2	Freudenstein and Roth	2	48.9842†	48.98425	48.98425	95	96
3	Powell	2	0	1.03E-27	2.07E-27	419	379
4	Brown	2	0	3.43E-19	9.30E-20	183	208
5	Beale	2	0	8.72E-21	1.52E-20	91	88
6	Jennrich and Sampson	2	124.362	124.3622	124.3622	78	71
7	Helical valley	3	0	1.84E-19	5.74E-19	205	177
8	Bard	3	8.21487E-03	8.214877E-03	8.214877E-03	182	178
9	Gaussian	3	1.12793E-08	1.127933E-08	1.127933E-08	130	124
10	Meyer	3	87.9458	87.94586	87.94586	100000‡	100000‡
11	Gulf	3	0	3.64E-29	4.21E-29	1806	1513
12	Box	3	.0755887†	7.558874E-02	7.558874E-02	307	226
13	Powell	4	0	1.29E-34	5.09E-35	670	660
14	Wood	4	0	1.41E-18	7.22E-19	405	526
15	Kowalik and Osborne	4	3.07506E-04	3.075056E-04	3.075056E-04	247	259
16	Brown and Dennis	4	85822.2	85822.20	85822.20	100000‡	354
17	Osborne 1	5	5.46489E-05	5.464895E-05	5.464895E-05	696	531
18	Biggs	6	0	5.66E-03†	2.455E-22	705	1100
19	Osborne 2	11	4.01377E-02	4.013774E-02	4.013774E-02	3534	3014
20	Watson	20	0	3.98E-03∗	3.22E-03∗	2214	2404
21	Extended Rosenbrock	10	0	5.37∗	3.63E-18	9103	17466
22	Extended Powell	10	0	1.29E-34	5.09E-35	670	660
23	Penalty	4	2.24997E-05	2.249978E-05	2.249978E-05	826	623
24	Penalty II	4	9.37629E-06	9.376293E-06	9.376293E-06	2299	2433
25	Variably Dimensioned	10	0	1.25∗	1.11∗	4861	5523
26	Trigonometric	10	0	2.80E-05∗	4.22E-05∗	2187	2188
27	Brown Almost Linear	10	0	1.73E-20	5.62E-20	3730	4897
28	Discrete BV	10	0	1.91E-19	6.93E-20	1355	1150
29	Discrete Integral	10	0	7.11E-18	7.09E-18	1320	1518
30	Broyden Tridiagonal	10	0	1.94E-17	2.00E-17	1350	1277
31	Broyden Banded	10	0	2.68E-17	1.32E-16	1388	1513
32	Linear Full Rank	10	10	10.0	10.0	1679	1958
33	Linear Rank 1	10	4.634146341	4.63E+00	4.63E+00	386	389
34	Linear Rank 1 with 0's	10	6.135135135	6.14E+00	6.14E+00	378	409
35	Chebyquad	9	0	3.06E-19	1.13E-18	2494	1801

† Solution converged to local minima
∗ Solution failed to converge
‡ Maximum iterations reached

Table 5.1: Unconstrained Optimization Test Functions

Chapter 6

Mahalanobis Taguchi Classification System

The use of adaptive experimentation can be extended beyond the traditional experimental domains. In this situation, historic use of highly fractionated orthogonal arrays created an opportunity to benefit from adaptive variable selection. The goal of this chapter is to present a classification system that incorporates adaptive experimentation for variable selection. The probable exploiting of interactions and very few runs make up for an inability to build a model and accommodate potential non-random effects. Analyzing images, or other data processing and statistical learning techniques provide unique challenges, as well as numerous tools. The background, techniques, and direction of this area of research will not be discussed here and the interested reader should see Hastie et al. (2001). The idea presented in this chapter has been expanded with an additional example and further discussion in Foster et al. (2009).

6.1 Introduction

The Mahalanobis Taguchi System (MTS) is a pattern analysis technique, which is used to make accurate predictions in multidimensional systems. This methodology has continuously evolved through the research effort led by Genichi Taguchi. This system has found industrial use as a data analytic approach that can be used to classify multiple systems. Examples have been given in medical diagnostics, inspection systems, sensor syetms, and even marketing applications (Taguchi and Jugulum, 2002).

The Mahalanobis distance (MD), which was introduced by a well-known Indian statistician P.C. Mahalanobis, measures distances of points in multidimensional spaces. The Mahalanobis distance has been extensively used in several areas, like spectrographic and agricultural applications. This distance is proved to be superior to other multidimensional distances like Euclidean distance because it takes correlations between the variables into account. In MTS the Mahalanobis distance (actually, a modified form of the original distance) is used to represent differences between point and pattern groups in quantitative terms. It can also be used to classify different objects in multidimensional systems. If this distance is above a certain threshold then the data point is not part of that data set that belongs to normal or reference group. The Mahalanobis distance is a multiple of the Hotelling T^2 that has been used in the statistics literature for many years. It is frequently used to identify statistical outliers as in Hawkins (1980). Here , the signal-to-noise (S/N) ratios are used to determine the accuracy of the Mahalanobis distance with respect to predictions or classification.

To compute the distance one first has to calculate the mean vector (μ) and the covariance matrix (**K**) of the training population (this is usually referred to as normal or reference

group). The distance for any sample in the space (**f**) is given by a scalar:

$$D = \frac{1}{n}(\mathbf{f} - \mu)^T \mathbf{K}^{-1}(\mathbf{f} - \mu) \tag{6.1}$$

The sample vector (**f**) is comprised of a number of features or variables that are important to the classification.

To begin with all the features or variables that may be important for pattern analysis are included. Usually, the number of features is large so the next step is to use orthogonal arrays (OAs) and S/N ratios to determine the reduced set of important features or variables.

The basic steps in MTS can be summarized as follows:

Stage I: Construction of a Measurement Scale

- Select a Normal group or reference group with suitable features or variables and observations that are as uniform as possible.
- Use this group as a base or reference point of the scale.

Stage II: Validation of the Measurement Scale

- Identify the conditions outside the reference group.
- Compute the Mahalanobis distances of these conditions and check if they match with decision-maker's judgment.
- Calculate S/N ratios to determine accuracy of the MTS system.

Stage III: Identify the Useful Variables (Developing Stage)

- Find out the useful set of variables using Orthogonal arrays and S/N ratios.

Stage IV: Future Diagnosis with Useful Variables

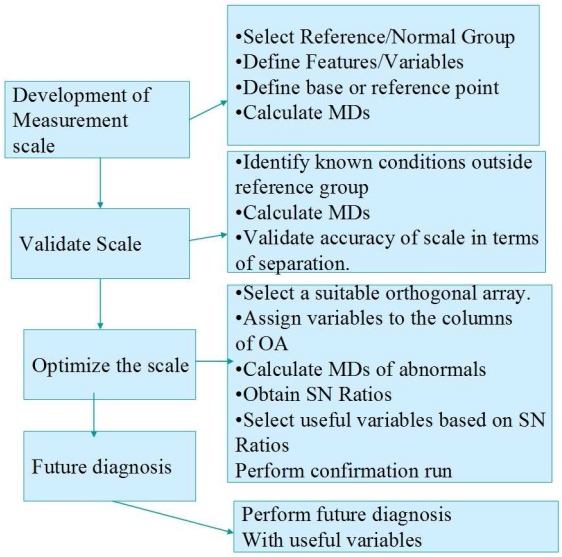

Figure 6-1: Steps in MTS

- Monitor the conditions using the scale, which is developed with the help of the useful set of variables. Based on the values of Mahalanobis distances, appropriate corrective actions can be taken.

Figure 6-1 is another presentation of the different steps in MTS (Foster et al., 2009). From the steps it is clear that role of orthogonal arrays are prominent in MTS analysis. Each experimental run in the orthogonal array design matrix uses a subset of variables; the resulting S/N ratios of these subsets are calculated using the distances from the reference group and S/N ratios are then used to determine the best variables.

The selection procedure using the OA is to run the entire matrix and then use a variable addition procedure to determine if any variable should be included. At the end of the procedure the appropriate subset of variables has been selected that give the maximum S/N ratio. Typically in MTS, either larger-the-better type or dynamic type S/N ratios are used. But this work is restricted to the larger-the-better type, which is given by:

$$\eta_j = -10 \log_{10} \sum_{i=1}^{n} \frac{1}{\left(\frac{D_{ji}}{D_{jj}}\right)^2} \tag{6.2}$$

Where the sample (D_{ji}) is the Mahalanobis distance to classification i from a population n for each of the j classifications. This S/N ratio maximizes the distance between the different classifications. Given j classifications the distances for all of the permutations are added together to form a composite S/N ratio for the choice of variables and the test population. For complete identification all permutations need to be considered, and thus are added together.

The comparison of an orthogonal array (OA) search method will be made with adaptive One-Factor-At-a-Time (aOFAT) and forward search selection procedure.

6.1.1 Description of Experimentation Methodology

Each string of variables can be between 10-50 individual variables long. Thus a complete run of all variable combinations yields 2^{10} to 2^{50} experimental runs, excessive for all but the simplest of simulations. To overcome this limitation reduced factor experimentation is normally used.

The OA is a fractional factorial experimental design technique where for the entire experiment, any two variables will have each possible combination run an equal number of times. Only symmetrical designs of strength two are considered. An orthogonal array $OA(N, 2^{N-1})$ is the same as a Level-III 2^{k-p} fractional factorial design.

aOFAT is compared with two-level, strength 2 symmetrical orthogonal arrays, and with a forward selection algorithm. The forward selection algorithm was proposed by Abraham and Variyath (2003) as an alternative to the OA in an attempt to decrease the computation time in variable selection.

In the forward selection algorithm, each individual variable is arranged by its contribution to the output. Then each variable is combined in descending order of importance until the change in the S/N ratio is insignificant.

6.1.2 Image Classification System

In many contexts, it is necessary to classify an image into one of several categories despite noise and distortion of the image. Some applications of such a capability include:

- Target recognition in autonomous military applications
- Matching evidence from a crime scene with a database
- Searching image databases via samples of images rather than keywords
- Classifying medical diagnostic scans

The system here classifies gray-scale* representations of fine art prints. Given a small bitmap, the goal is to classify it from a comprehensive database. For purposes of this study, four well known portraits were chosen: Da Vinci's 'Mona Lisa', Whistler's 'Portrait of the Artist's Mother', Peale's 'Thomas Jefferson', and Van Gogh's 'Self Portrait with Bandaged Ear'. The low resolution bitmaps (32 X 32) used in the study are depicted in the top row of Figure 6-2.

In practice, if one were given an image to identify, it would likely be affected by various types of noise. The image may have been taken by a camera and the possibility exists that the image will be out of focus. The image may have been broadcast and so there may exist some degree of either white noise or 'snow' superimposed upon it. The image

*In gray-scale, a value of zero represents black while a value of 255 represents white. All the integers between are smoothly varying shades of gray between those extremes

may have been scanned into a computer and therefore it is possible for the image to be framed off-center. Further, it may be desirable to correctly identify the image without prior knowledge of whether the image is a negative or a print. To simulate such noise conditions, the following operations in the following order were performed on each image to be classified:

1. The image was blurred by convolving the image with a pixel aperture whose size varies randomly among 3, 4, and 5 pixels square.

2. The image was superposed with 'snow' by switching each pixel to white with probability 0.75.

3. The position of the image in the 'frame' was shifted by either -2, -1, 0, 1, or 2 pixels with equal probability. The shift was made both horizontally and vertically but the amount of the shift in the x and y directions were probabilistically independent.

4. The images were transformed into a negative with probability 0.5.

Examples of the effects of these noises are depicted in Figure 6-2. The first row contains bitmaps of all four portraits without noise. Below each portrait are three noisy versions of the same portrait. The degree of noise is intended to be severe enough to make classification of the images difficult.

6.2 Feature Extraction Using Wavelets

Wavelets were chosen for this application to extract features from the images and create the variables. The goal of this section is to provide enough background to allow the reader to understand the case study. The treatment will therefore be qualitative. For a more detailed mathematical introduction to wavelets in engineering, the reader may wish to read

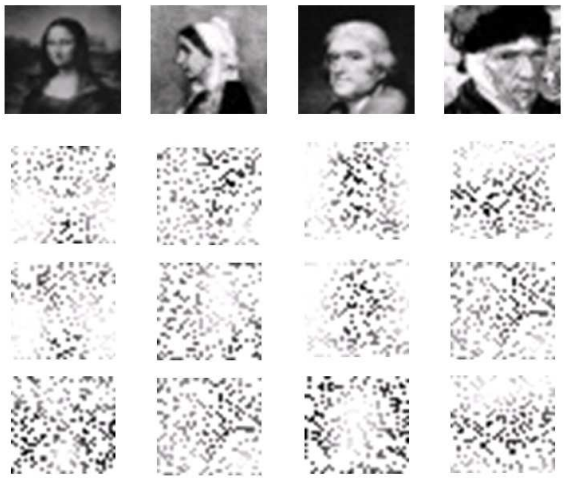

Figure 6-2: Fine art images before and after application of noise

Williams and Amaratunga (1994), or specifically concerning images Williams and Amaratunga (1993).

A wavelet transform is a tool that cuts-up data, functions, or operators into different frequency components with a resolution matched to its scale (Daubechies, 1992). Therefore, wavelets are useful in many applications in which it is convenient to analyze or process data hierarchically on the basis of scaling.

To demonstrate that the wavelet's property of cutting up data based on scale is useful in image processing, let us consider the effect of wavelet transforms on the image of the Mona Lisa. Wavelet coefficients from a 32 X 32 gray-scale bitmap of the Mona Lisa (on the left in Figure 6-3) were extracted using a two dimensional wavelet transform based on the Daubechies four coefficient wave filter. These wavelet coefficients are represented by a 32X32 matrix. The entire set of coefficients was used to reconstruct the image using an inverse wavelet transformation (the image second from the left in Figure 6-3). One can see that this reconstruction preserves essentially all of the detail of the original bitmap. To generate the next image, we discarded all but the 16 X 16 coefficients in the upper left then

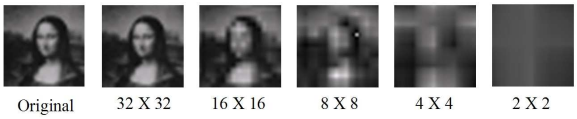

Figure 6-3: The Mona Lisa reconstructed from its wavelet transform after all but the N X N coarsest levels of scale have been discarded

padded the matrix with zeros back to 32 X 32, and reconstructed the image. The resulting image (third from the left in Figure 6-3) reveals that the first 16 X 16 elements contain information describing the rough features of the original image. This process was repeated by removing more elements of the wavelet coefficients resulting in successively coarser images.

The ability of wavelets to cut up an image on the basis of scale has made them very useful in image compression. By discarding wavelet coefficients below a certain threshold, the amount of information to be stored or transmitted can be significantly reduced without significantly degrading the perceived quality of the image. This strategy succeeds because the features that allow people to identify an image tend to be characterized by length scales. The overall proportion and balance of Van Gogh's portrait is very different from that of the Mona Lisa. Thus, the two portraits can be distinguished on the basis of features with medium length scales. However, it is also quite possible to distinguish the two paintings on the basis of features on a much smaller scale. The style of the brush strokes in Van Gogh's portrait is very different from that of the Mona Lisa; most people could distinguish the two paintings with only a one inch square sample of the original paintings.

The properties of wavelets that make them useful for compressing images also make them useful for recognizing images in the presence of noise. When snow is superimposed on an image, it will tend to disrupt the finest details so that the information at that scale may actually hamper recognition. Similarly, the coarsest levels of resolution may contain

very little information useful for image recognition. The image on the right in Figure 6-3 is uniformly gray. This shows that the painting has uniformly distributed patterns of dark and light at the coarsest level, but this is a property of most fine art because people appreciate paintings that appear balanced. Therefore, the coarsest levels of wavelet coefficients may not be useful in distinguishing the Mona Lisa from other portraits. It is possible that the features that best allow one to distinguish the Mona Lisa from other fine art prints (especially in the presence of noise) are found at intermediate scales.

Given the power of wavelets in extracting key features of an image based on a hierarchy of scales, they were selected for this image recognition system. The matrix of wavelet coefficients were used to construct the Mahalanobis distances and compare the three experimentation routines. Each image was 32 X 32 the wavelet transform was also 32 X 32. To reduce the vector size and, considering that the art medium is more interesting at larger scales, only the first 8 X 8 matrix were used and the rest zero padded before the inversion. This vector was then 64 bytes long, the last byte was also removed to give a convenient length of 63 bytes, the same length as a traditional orthogonal array.

6.3 Comparing Results of the Different Methods

Each method was trained with a set of noisy pictures. After the training routine each of the three routines produced a vector of the ideal variables for identification. These ideal vectors were then applied to another set of noisy pictures, and the results compared.

The aOFAT performed with the highest average identification percentage, and utilized an average training time. It was able to take advantage of the two-factor and higher interactions and the noise was not sufficient to effect the results. The aOFAT scaled well with a reduction in the number of training images.

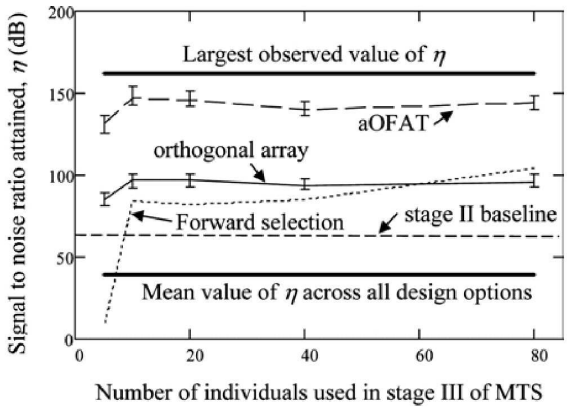

Figure 6-4: Results of the three search methods for the image classification

In the OA they were able to utilize the first and second order interactions but not take advantage of higher level interactions. The OA chose a point that was less optimal than the aOFAT because it did not include anything greater than two-level interactions. Even though the OA was run with multiple arrays it might prove to be advantageous to run some that focus on the two- and greater level interactions because of their importance.

The third method, the forward search, was the most efficient to run, and proved to be equal to the OA when the number of individual was greater than fifty. This method was highly dependent on a strong hierarchy of effects that was not as evident in this problem. In situations with a large hierarchical bias it would perform well at low computational cost.

As the number of classifications grow the routine should show an improvement of a similar magnitude to that shown in the reduction of the training population. It may be possible, using non-wavelet routines, to reduce the dependence on the higher level interactions, but the current experimentation shows that realistic problems have higher level interactions, and low noise. This situation is the ideal application of an aOFAT experiment.

6.4 Conclusion

There is an appropriate place for an adaptive experiment combined with classification techniques, here the Mahanalobis-Taguchi Strategy (MTS). When the number of classification variables are too numerous to enumerate all possibilities, choosing the best sub-set is similar to the maximum seeking experiment that aOFAT has demonstrated utility. This adaptive experiment is used as a variable screening procedure as part of a complete classification framework. Within published classification procedures such as linear regression, logistic regression, and discriminate analysis the use of an adaptive selection experiment can improve results and reduce the computational burden.

Compared with the other available methodologies such as OA or forward search, aOFAT is shown to yield a better result. aOFAT produces S/N ratios that are significantly greater than the other routines while incurring similar experimental cost. As Daniel (1973), and other experimentalists agree; in most experimentation too much time is spent on unimportant and uninteresting regions, aOFAT is a technique that focuses interest into the important and interesting areas and then allows for sub-set analysis. More information about this particular application including more examples is available in Foster et al. (2009).

Bibliography

Abraham, B. and Variyath, A. M. (2003). Discussion of 'a review and analysis of the mahalanobis-taguchi system'. *Technometrics*, 45:22–24.

Daniel, C. (1973). One-at-a-time plans (the fisher memorial lecture, 1971). *Journal of the American Statistical Association*, 68:353368.

Daubechies, I. (1992). Ten lectures on wavelets. In *CBMS-NSF Regional Conference Series, SIAM*.

Foster, C., Frey, D., and Jugulum, R. (2009). Evaluating an adaptive one-factor-at-a-time search procedure within the mahalanobis taguchi system. *International Journal of Industrial and Systems Engineering*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hawkins, D. (1980). *Identification of Outliers*. Chapman and Hall.

Taguchi, G. and Jugulum, R. (2002). *The Mahalanobis Taguchi Strategy: A Pattern Technology System*. John Wiley and Sons.

Williams, J. R. and Amaratunga, K. (1993). Matrix and image decomposition using wavelets. In *Proceedings MAFELAP '93, Eighth International Conference on the Mathematics of Finite Elements, Brunel, England*.

Williams, J. R. and Amaratunga, K. (1994). Introduction to wavelets in engineering. *International Journal of Numerical Methods in Engineering*, 37:2365–2388.

Chapter 7

aOFAT Integrated Model Improvement

An overall purpose to the procedures discussed in this thesis is to increase the overall utility of experimentation by combining statistical methods with adaptive experiments. One goal may be to utilize the aOFAT experiment combined with future experimental data to build a composite model. There is one specific method that will be investigated here and will benefit from our prior knowledge within the aOFAT. A subsequent chapter will investigate a general method that can build models from adaptive experiments without assistance from data regularities or other application specific information. Utilizing the aOFAT experiment, outside of superficially comparing the results, is important in leveraging the experimental cost to improve the system and enhance resultant models.

7.1 Introduction

Experiments can be used for a variety of purposes including optimization, model development, factor identification, and robustness exploration. The academic approach is to use experiments to build a model followed by model optimization and validation as in Wu and

Hamada (2000). This contrasts the stated objective of most industrial optimization experiments as in Montgomery (1996) or Myers and Montgomery (2002). The purpose here is to combine these two activities with two specific experiments to achieve an optimum, followed by the creation of a parametric model. Both of these individual experiments have numerous approaches and different techniques, the challenge is to benefit from the first activity in completing the second. Providing both the optimum as well a parametric model is pragmatic in that many times the optimum is found to be insufficient in some unforeseen aspect and a more complete model is needed. Finding an optimum or near-optimum initially is also desirable as many designed experiments are left unfinished when equipment fails, priorities change, or budgetary limits are met. In addition to the precautionary, designing an experiment to seek out an optimal point initially may create savings by using that point while the remainder of the experiment is run. The savings could be substantial and with high likelihood, no further changes may be needed. The use of designed experimentation has been championed by a few firms although the development of the techniques comes from the statistical community. This combined technique bridges the gap between the intuition of the practitioner and the statistical framework.

7.2 Background

The traditional classification of the different types of experiments by Wu and Hamada (2000) are: treatment comparisons, variable screening, response surface exploration, system optimization, and system robustness. These classifications are based on developed techniques while, in practice, industrial experiments are run to meet a specific objective, perhaps to improve a product or to eliminate a defect. These objectives normally requires a number of traditional experiments, first a variable screening experiment may be used to

determine the important factors, followed by a rough system optimization experiment to move around in the design space and a final response surface experiment for higher order effects. The noise variables may need to be addressed through a specific robustness experiment to finalize the setting. If there is financial or scheduling pressure an initial experiment may be used to determine an immediate setting that can then be adjusted when the larger experiment is complete. It is also possible that there will be a decision to end the experiment early if a satisfactory setting is found in the initial runs. Experiments may also end early if the test unit fails, or the project has budgetary or scheduling problems. Getting useful knowledge out of those incomplete experiments is difficult and may be impossible.

The procedure outlined here is targeted for a dual target of optimization with a goal of building a system model for alternative setting options or robustness studies. The first stage of traditional optimization is to decide on an experimental design. The number of runs determines the number of parameters that can be estimated. As described in Chapter2, given $n+1$ experimental runs it is possible to, at most, estimate n model parameters. Larger experiments are frequently used to estimate two-way interactions $X_1 \cdot X_2$ or three-way interactions $X_1 \cdot X_2 \cdot X_3$ as well as to understand system noise and error. Most experiments are design to be balanced with equal number of high and low settings, orthogonal between the different variables, and finally, run in a random order to try and minimize time dependent noise effects. Because of the sensitivity of noise, most designed experiments are much larger than necessary compared with the maximum parameters that can be estimated, some alternatives to this inefficiency will be addressed.

7.3 Procedure

To begin the initial optimization search an adaptive One-Factor-At-a-Time (aOFAT) experiment is performed. This is an adaptive optimization procedure that has been recently described in the literature by Frey and Sudarsanam (2008) and Frey et al. (2006). This procedure utilized here is as follows, an initial random variable setting is run. Then sequentially through each variable a single change is made and that new setting run. If the result is improved then the new variable setting remains, if not, it is returned to its original value. This experimentation technique requires $n+1$ runs, one for the initial setting and one for each variable. Although this procedure has been discounted in a number of books such as Wu and Hamada (2000), it has shown to be effective in achieving an optimum under normal levels of noise and a typical ratio of interactions to main effects. If the noise is too high or if there are too many significant interactions, then a more traditional approach may be more effective. The other potential problem is an absence of run randomization and any time or order dependency could lead to poor results. In a study of 113 published experiments, this method had a very high likelihood of producing the optimal setting compared with other alternative procedures using a similar number of runs (Frey et al., 2006).

To quantify this improvement we will use a hierarchical probability model (HPM) that was constructed by Li and Frey (2005) using the aforementioned 113 industrial experiments, and described in detail in Chapter 2. This HPM creates a response that mimics one of the 113 original experiments; it can be used to gauge the initial improvement of an aOFAT experiment over the best possible variable setting. The biggest influence to this response is in the pure error which is defined here as a ratio to the factor effects (FE). Even with large amounts of experimental error an aOFAT experiment yields 90% of the possible improvement as shown in Figure 7-1. A ratio of 0.2 is typically found in experiments.

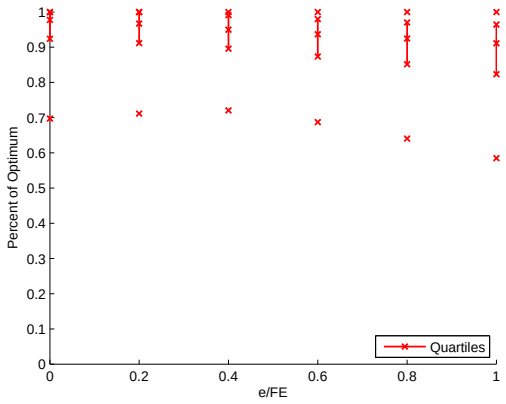

Figure 7-1: aOFAT Percentage Improvement

This technique requires $n + 1$ experimental runs and provides a good method for determining the optimal variable settings. In addition to searching for an optima the other outcome of this experiment is an estimate of each variable's importance. The challenge with using aOFAT results is the significant probability of exploiting interactions as well as main effects, which are not possible to estimate with only $n + 1$ runs. To make a more accurate estimate of the importance of each variable, probable interaction effects will be removed.

Using these 113 experiments Frey and Wang (2005) have looked at the expected improvement for each variable x_k for completing an aOFAT experiment; this expected value given n runs is shown in Equation 7.1

$$E(Y(\Delta x_k) = \sqrt{\frac{2}{\pi}}\left(\frac{\sigma^2_{ME} + \sigma^2_{INT}(n-2k+1)}{\sqrt{\sigma^2_{ME} + (n-1)\sigma^2_{INT} + \frac{\sigma^2_\varepsilon}{2}}} + \cdots \right.$$
$$\left. \frac{\sigma^2_{INT}(k-1)}{\sqrt{\sigma^2_{ME} + \sigma^2_{INT}\left(n - k^{1/\pi \arctan\left(\sigma_{INT}/\sqrt{\sigma^2_{ME}+(n-2)\sigma^2_{INT}+\frac{\sigma^2_\varepsilon}{2}}\right)+\frac{1}{2}}\right) + \frac{\sigma^2_\varepsilon}{2}}} \right) \tag{7.1}$$

Where the standard deviation of the main effects, interaction terms, and the error are given by σ_{ME}, σ_{INT}, and σ_ε respectively. Some assumptions are made about the size of the different effects, which are based on the results seen in the industrial experiments. Assuming $\sigma_{INT} = \sigma_{ME}/3$, $\sigma_\varepsilon = \sigma_{ME}/4$, and because we are only interested in the relative influence of each variable $\sigma_{ME} = 1$. This reduction leads to a reduced form shown in Equation 7.2.

$$E(Y(\Delta x_k)) = \frac{.889 - .178k + .089n}{\sqrt{.11n + .92}} + \frac{.089k - .089}{\sqrt{.11n - .11k^{.32 \arctan\left(.33/\sqrt{.11n+.81}\right)+.5} + 1}} \tag{7.2}$$

This expected improvement information can be used in the covariance matrix of a Gaussian process that will model the follow-up experiment. The interaction information is needed in addition to the response because the later variables are more likely to benefit from interaction effects than earlier ones. The complexity of this equation is normally unnecessary with a small number of runs, and a linear approximation will be used instead. With seven variables and thus eight experimental runs the expected improvement of a linear

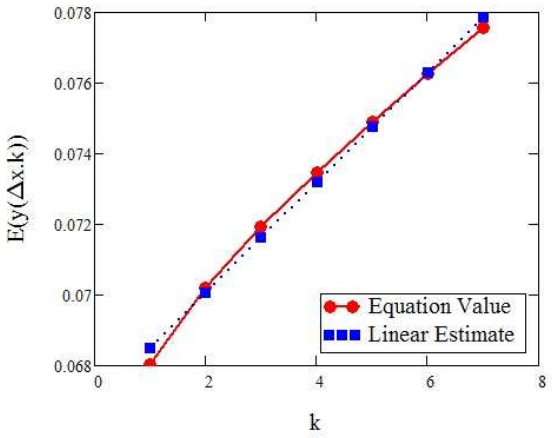

Figure 7-2: Expected Improvement Comparison

estimate is compared with the equation as shown in Figure 7-2.

This linear approximation is described by the slope of this line. Increasing the variables decreases the slope by $O(n^2)$ and thus approaches zero rapidly with a large number of variables. To predict the slope for a specific number of variables n, the log-log plot yields this relationship, $S = 0.081 \cdot n^{-1.842}$, as can be seen in Figure 7-3. Now, after running the aOFAT experiment, an optimal or near optimal point is known as well as the relative contribution of each variable with the interactions ignored.

The second, follow-up, experiment used here is an orthogonal array (OA) based experiment that was introduced in Chapter 2 and also used in Chapter 6. An OA is a set of linearly independent run columns for each variable. Each column is orthogonal to the other columns in the set and so can estimate the main effects easily. Depending on the design and the size of the OA it can also estimate a number of interactions. The choice of designs are Plackett and Burman (1946) designs, and are well known in the statistical literature and introduced in Chapter 2. The designs can be easily constructed and are of length $N = 4k, k = 1, 2, \ldots$ where N is not a power of 2. The Plackett-Burman designs

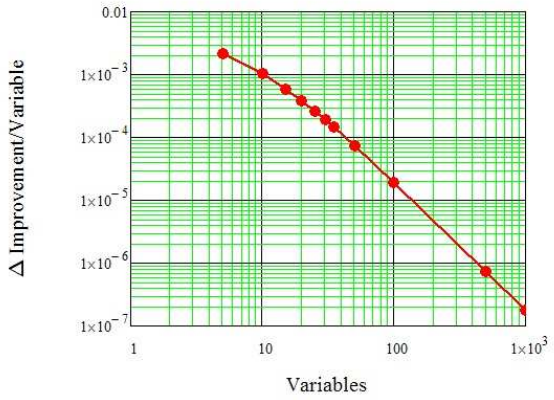

Figure 7-3: Additional Variable Slope

have a useful property - if there are only a few significant terms the remaining columns can estimate interactions. This design property along with the prior information available from the aOFAT will be useful in the Bayesian analysis.

After completing the aOFAT experiment followed by the Plackett-Burman experiment the data collection procedure is complete. At this point it should be noted that there are other experimental methodologies to collect the data including running repeated aOFAT experiments, or other types of designed experiments including fractional-factorial, D-Optimal, A-Optimal, and minimum abberation designs. These methodologies may be more appropriate given a particular area of application or understanding of the underlying physics. The following analysis is more general than the Plackett-Burman design and any experimental design could be substituted.

7.4 Analysis

The Plackett-Burman OA design will create the foundation for a model estimate. It is not possible to use traditional analysis by combining both sets of experimental runs into a

large matrix, because this matrix would be singular. This is due to the fact that the aOFAT matrix is singular. Removing the singularity can be accomplished by removing runs from the aOFAT, but this normally requires removing half of the runs, and then leads to little improvement.

The analysis method used here is a Bayesian procedure that is modified from the procedure of Joseph (2006), that is based on what is referred to as an empirical Bayesian analysis. Additional information on the mathematics behind empirical Bayesian analysis are available in Chapter 2.

Given a linear estimate $F = X^T \cdot \mu + \epsilon(X)$ where X consists of the k most important variables, and the error is a Gaussian process, $\epsilon \sim \mathcal{GP}(0, \sigma_k^2 \mathbf{\Psi})$ without loss of generality we can say $F = X^T \cdot \mu + X^T \cdot \beta$ where $\beta \sim \mathcal{GP}(0, \sigma_k^2 \mathbf{\Psi})$. The $\mathbf{\Psi}$ term is the correlation matrix. The most frequent correlation functions are product or exponential correlations.

$$\mathbf{\Psi}(X_1, X_2) = \prod_{i=1}^{p} \Psi_i(X_{1i}, X_{2i}) \tag{7.3}$$

This correlation function looks more simple than the traditional exponential function because the experimental values here are assumed to be only -1 or 1. In this function p is the number of runs for a full factorial experiment. Here it is assumed that Ψ is stationary for all p and that our variables are $\in (-1, 1)$, so $\Psi_i(\vec{X_1}, \vec{X_2}) = \Psi_i(|X_{1i} - X_{2i}|/2)$ that has only two values $\Psi_i(0) = 1$ and $\Psi_i(1)$. This differs from the traditional empirical analysis but, is consistent with the approach. In Chapter 8 the full exponential correlation function will be employed because the data is not from a designed experiment.

The expected values and variances can be determined from these definitions.

$$\mathsf{E}(f) = \mu \tag{7.4}$$

$$\mathsf{Var}(f) = \sigma_k^2 \mathbf{\Psi}_p \tag{7.5}$$

and

$$\begin{aligned} \mathsf{E}(\beta) &= \mathsf{E}(\mathbb{X}_p^{-1} \cdot (F - \mathbb{X} \cdot \mu)) \\ &= 0 \end{aligned} \tag{7.6}$$

$$\begin{aligned} \mathsf{Var}(\beta) &= \mathsf{Var}(\mathbb{X}_p^{-1} \cdot (F - \mathbb{X}\mu)) \\ &= \mathbb{X}_p^{-1} \sigma_o^2 \mathbf{\Psi}_p (\mathbb{X}_p^{-1})^T \\ &= \sigma_k^2 \mathbb{X}^T \mathbf{\Psi} \mathbb{X} \end{aligned} \tag{7.7}$$

This last expression can be simplified using the structure of the product correlation function. A full-factorial experiment can be defined in a recursive fashion where $\mathbb{X}_0 = 0$ and additional terms defined by:

$$\mathbb{X}_i = \begin{pmatrix} \mathbb{X}_{i-1} & -\mathbb{X}_{i-1} \\ \mathbb{X}_{i-1} & \mathbb{X}_{i-1} \end{pmatrix} \tag{7.8}$$

And noting that the last column is half negative followed by half positive, with the remaining columns identical between the halves, both haves are correlated by the $\Psi(1)$ value. The

entire product correlation equation can be expressed as:

$$\mathbf{\Psi}_i = \begin{pmatrix} \mathbf{\Psi}_{i-1} & \Psi(1)\mathbf{\Psi}_{i-1} \\ \Psi(1)\mathbf{\Psi}_{i-1} & \mathbf{\Psi}_{i-1} \end{pmatrix} \tag{7.9}$$

Substituting these into the variance of β:

$$\sigma_k^2 \mathbb{X}^T \mathbf{\Psi} \mathbb{X} = \frac{\sigma_k^2}{2^{2^p-1}} \cdot (1 + \Psi(1)) \cdot \begin{pmatrix} \mathbb{X}_{p-1}^T \mathbf{\Psi}_{i-1} \mathbb{X}_{p-1} & 0 \\ 0 & r_p \cdot \mathbb{X}_{p-1}^T \mathbf{\Psi}_{i-1} \mathbb{X}_{p-1} \end{pmatrix} \tag{7.10}$$

where $r_p = \frac{1-\Psi_p(1)}{1+\Psi_p(1)}$ and defining $\tau^2 = \frac{\sigma_k^2}{\prod_{i=1}^{p}(1+r_i)}$ we know that $\mathbb{X}_0^T \mathbf{\Psi}_0 \mathbb{X}_0 = 1$ so

$$\mathsf{Var}(\beta) = \tau^2 \cdot \mathbb{R} \tag{7.11}$$

where $\mathbb{R}$ is the diagonal matrix for the variables in p:

$$\mathbb{R} = \begin{pmatrix} 1 & & & & & \\ & r_1 & & & & \\ & & r_2 & & & \\ & & & \ddots & & \\ & & & & r_1 \cdot r_2 & \\ & & & & & \ddots \end{pmatrix} \tag{7.12}$$

This matrix from the product correlation function, has two properties hierarchy and heredity from Wu and Hamada (2000), that are often discussed in the experimental literature. Hierarchy is defined as having largest factors as main effects, followed by smaller two-way interactions, and smaller three-way interactions. In this matrix $r_i < 1$ and so this

property holds true for the covariance matrix. Heredity is defined as a property where a significant main effect is more likely to have interactions that are also significant. This property is also apparent in this matrix, if r_i is large then interactions with r_i will also be large.

Given that we are estimating the parameters in this model from a reduced run set, there are too many parameters in $\mathbb{R}$. Here we will reduce the model by including a predetermined weight vector, w.

$$r_i = r * w_i \tag{7.13}$$

This still makes $\prod(1 + r_i)$ unique and not reducible, and if we have fixed w_i such that $\max(w_i) = 1$ then it is necessary to only determine a single parameter $\hat{r}$. In the original work Joseph (2006) used a single value $r_i = R$, thus the properties of hierarchy and heredity hold but the variables are all weighed equally. The experimental matrix was used as a posterior to this information to create a model. One drawback to this approach is that all of the variables are weighted equally and so the data has to be sufficient for the posterior estimate to change. In the experimental work by Li and Frey (2005) it was found that variables are exponentially distributed and so a uniform assumption of Joseph (2006) would require substantially more data to reach the same posterior accuracy.

Taking the approach of Robbins (1956) that greater effort used to create a better prior model will benefit the overall performance of the resulting estimate. The aOFAT experiment was used to estimate the variable ranking (as well as estimate the maximum). This is incorporated into the w_i weight variable is from Equation 7.2 where w_0 is set to the mean value of w_i. There is no estimate of the error of the aOFAT variable weights, so the error around the w_i's is unknown. To control this affect w_i is reduced as $\arg\max_{r\in[0,1)}$ approaches 1.0. The influence of the w_i's drives $r \to 1.0$ then the w_i's are iteratively reduced, by setting

$w = w^{0.9}$. This can be justified by noting that large r values are driven by large disparities between the weights and the experimental values. This covariance shrinkage maintains the hierarchical and heredity variable properties while reducing undue influence of the aOFAT error.

Given $y|\beta \sim \mathcal{N}(\mathbb{X}\vec{\mu} + \mathbb{X}\vec{\beta}, \sigma)$ and without enough information we consider σ to be small compared with the β variance- $\beta \sim \mathcal{N}(0, \tau^2\mathbb{R})$. By applying the properties of the normal distribution we can determine y.

$$y \sim \mathcal{N}(\mathbb{X}\vec{\mu}, \sigma_k^2\mathbf{\Psi}) \tag{7.14}$$

The log-likelihood of this distribution can be used to determine r from Sargan (1964):

$$\begin{aligned} l = \text{constant} - \frac{1}{2}\log\det(\sigma_k^2\mathbf{\Psi}) - \\ \frac{1}{2}(\vec{y} - \mathbb{X}\vec{\mu})^T(\sigma_k^2\mathbf{\Psi})^{-1}(\vec{y} - \mathbb{X}\vec{\mu}) \end{aligned} \tag{7.15}$$

which yields: $\hat{r} = \arg\max_{r\in[0,1)} l$

A stepwise addition procedure is used to add variables to $\vec{\mu}$. And from the distributions above β can be estimated:

$$\begin{aligned} p(\hat{\beta}|y) \sim \mathcal{N}(\mathbb{R}\mathbb{X}^T\mathbf{\Psi}^{-1}(\vec{y} - \mathbb{X}\vec{\mu})\frac{\tau^2}{\sigma_k}, \\ \tau^2(\mathbb{R} - \frac{\tau^2}{\sigma_k}\mathbb{R}\mathbb{X}^T\mathbf{\Psi}^{-1}\mathbb{X}\mathbb{R})) \end{aligned} \tag{7.16}$$

To determine the variables to add we can look at the interval of β. The interval is given by $\hat{\beta}_i \pm \Phi^{-1}(1 - \alpha/2)$ where Φ^{-1} is the inverse normal distribution, if this interval does not contain 0 then it would be a credible addition. This can also be expressed such that the

absolute value of the normalized score must be greater than $\Phi^{-1}(1-\alpha/2)$.

$$t_i = \left| \frac{\hat{\beta}}{\text{diag}(\sqrt{\text{Var}(\hat{\beta}|y)})} \right| \tag{7.17}$$

After choosing the most probable variable to add, μ_k and σ_k need to be found by substituting the new value:

$$\hat{\mu}_k = (\mathbb{X}^T \mathbf{\Psi}^{-1} \mathbb{X})^{-1} \mathbb{X}^T \mathbf{\Psi}^{-1} y \tag{7.18}$$

and

$$\hat{\sigma}_k^2 = \frac{1}{n}(\vec{y} - \mathbb{X}m\vec{u})^T \mathbf{\Psi}^{-1} (\vec{y} - \mathbb{X}m\vec{u}) \tag{7.19}$$

There is another stopping condition used in literature, the traditional R^2 value (multiple correlation coefficient).

$$R_k^2 = 1 - \frac{(\vec{y} - \mathbb{X}\vec{\mu})^2}{(\vec{y} - \vec{\mu})^2} \tag{7.20}$$

The t_i values are criticized as underestimating the variance and thus overestimating the confidence, and including too many variables. This is because the $\hat{\sigma}_k$ predictor is a biased estimate of the true mean squared prediction error. The R^2 estimate has another criticism that it always increases with added variables, and thus also includes too many variables. There is a correction for t_i (Zimmerman and Cressie, 1992) but it is not used here because, for the general linear model, this error has been shown to be asymptotically insignificant (Prasad and Rao, 1990). The use of an adjusted-R^2 is also not used here because the overfitting estimate based solely on the number of predictors versus the number of data points is misleading by not including the influence of the covariance matrix. The forward selection procedure is a frequently used method for variable addition. Other options include a back-

ward elimination, stepwise, all subsets (for $n \lesssim 7$), or other algorithmic best subsets. The C_p statistic was not used because iterating all possible combinations was not possible with $n = 11$. There are many good procedures available to determine the important variables, in the examples selected here we were limited by our imposed number of runs. The number of coefficients was maximum for the number of runs and so less dependent on the adding criteria. These routines were developed to limit the extra variables suggested important by a predicted residual sum of squares (PRESS) approach. In these examples we are limited by the maximum amount of information and so that limit is not applicable.

To summarize: the procedure initially has no variables. First estimate $\hat{r}$, determine the largest significant t_i, add that to the model by finding $\hat{\mu}_k$ and $\hat{\sigma}_k^2$. Repeat this procedure as long as the new $\hat{t}_i$ is significant, the $PRESS$ statistic is decreasing, or the maximum number of variables is reached.

7.5 Results

Three examples of this augmented method are presented, the first uses the Hierarchical Probability Model (HPM) introduced in Chapter 2. This model has a significant probability of two and three-way interactions, and stretches the use of the Plackett-Burman designs in detecting interactions. The second example is drawn from an analytic model presented by Wu and Hamada (2000) to show the challenge in identifying confounded variables. The third example is a physical experiment of a wet-clutch design presented originally in Lloyd (1974). For each of these examples the primary focus is on comparing the results in model building, and not the optimization search.

7.5.1 Hierarchical Probability Model (HPM)

Using the HPM generated model, a dual approach with an aOFAT followed by a 12 run Plackett-Burman design was compared with a 24 Run Plackett-Burman design. Both of these designs had 11 variables of interest and were run 200 times total with four different randomly generated HPM models. The PRESS (Prediction Sum of Squares from Chapter 2) statistic was used to compare the selected models. The results of both methods are shown in Figure 7-4. The larger experiment is able to generate slightly smaller PRESS values while the dual method uses fewer variables. The comparison statistic was run on all of the points in the full factorial experiment. This experiment was run with both the β significance criteria as well as the R^2 criteria and they both gave similar results. Because these models are so limited, the performance limitation is the number of experimental runs. Given the limited number of runs the dual method performs well compared to the larger method. There are two additional cases that will be investigated, first adding runs to both experiments and second, running the same sized second experiment. As the PRESS statistic shows in Figure 7-4, the run limitation indicates more terms are necessary to fit the model. The experiment lacks sufficient resolution to completely fit the best model. The runs could be increased either through a fractional-factorial experiment or larger Plackett-Burman design. Although they both yield similar results, here Plackett-Burman designs of 32 and 48 runs were used. In total the dual method has four fewer runs. The result with these larger run matrices is shown in Figure 7-5. The use of the aOFAT runs reduces the runs in the model while still achieving a similar PRESS statistic. It is expected that there is a limitation to adding more variables through the covariance matrix. This forced ranking of the inputs limits the number that can be added to the model. As the number of runs grow a reduced correlation matrix can increase the influence of these few runs on the final result.

Weighing the prior experiments is used to reduce the influence on the runs as a ratio of the number of initial runs $k + 1$ to the runs in the experiment n.

$$w_i = w_i^{\frac{k+1}{n}} \tag{7.21}$$

The result from this weighing is shown in Figure 7-6; the dual method has a reduced variance but a difference in the number of terms is still seen due to the difference in experiment size. The goal behind this methodology is for screening experiments and not large run experimentation. The weights influence the entire correlation matrix and lacks sufficient support for this rank in the entire experiment. As the ratio of data that determines the correlation structure is small compared to the run information the assumption of accuracy is no longer valid. The covariance matrix with a single r is justified in Joseph (2006) for a constant correlation coefficient; he indicates that assigning different weights can only be justified by knowing the relative weight of some effects. It is not assumed here that we 'know' the relative weights only that the guess is appropriate given the data. As the relative amount of data grows the weight differences are reduced.

In practice there are many initial or set-up runs that are normally discarded before the screening experiment is run. These runs can be used to help influence the covariance matrix that is followed by the actual experiment. If the run sizes are identical then the covariance matrix will improve the outcome. The result of using a 24 run Plackett-Burman design for both systems is shown in Figure 7-7. The dual method reduces the PRESS using the same number of variables. This higher performance for the dual method is expected and has utilized runs that are normally discarded. One caution when using these methods is that the extra runs need to reflect the correlation between the input variables for the experiment. Different variable ranges and locations should be corrected, as necessary.

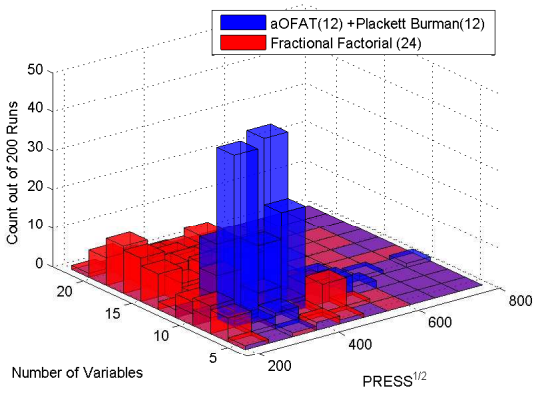

Figure 7-4: Hierarchical Probability Model (HPM) Comparison

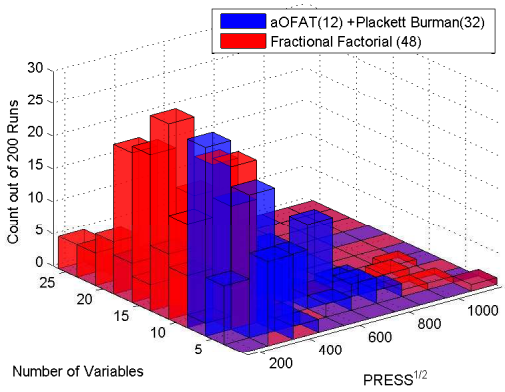

Figure 7-5: HPM Large Experiment

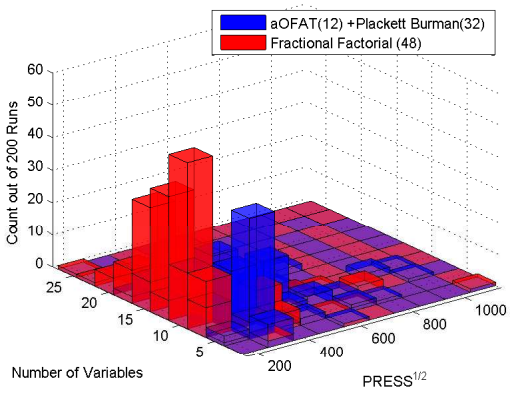

Figure 7-6: HPM Weighted Large Experiment

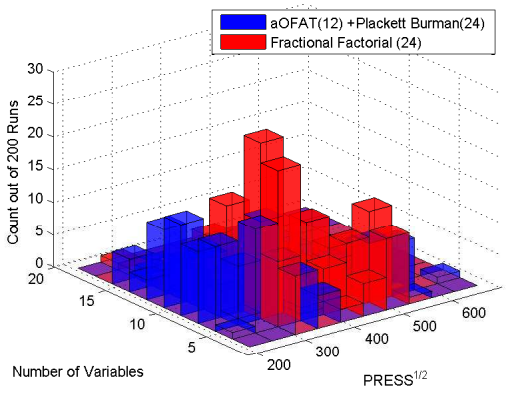

Figure 7-7: HPM Same Second Experiment Size

7.5.2 Analytic Example

The second example attempts to identify an analytic model presented by Wu and Hamada (2000, pg. 362). This analytic example is used to demonstrate a difficulty in evaluating experiments with complex aliasing. The model is $y = 2A + 4C + 2BC - 4CD + \epsilon$ where $\epsilon \sim \mathcal{N}(0, .5)$. The objective is to correctly identify this four variable model in an experiment with 11 variables (A-K). In the original analysis the models (C,CD,A) and (C,CD,BC), both which contain three of the four correct variables, and no incorrect variables were found to explain the data well. They also identified three other three-variable models that only have one correct variable, and three two-variable models with only one correct variable. The conclusion was that the analyst may find many equally plausible models. Here a more automatic procedure is presented based on Bayesian priors. A similar model to this was also used by Joseph (2006) to demonstrate his approach to Bayesian analysis. The objective is to match the performance of both of these previous methods using a dual approach consisting of an aOFAT and a 12-run Plackett-Burman experiment. To keep the number of runs comparable a comparison will use a 24-run Plackett-Burman experiment.

A comparison of the Bayesian analysis to the procedure given in Wu and Hamada (2000) is presented in Joseph (2006) and will not be repeated here. Each procedure was run two hundred times on different random sets of data. All of the variables were permuted before each run, so variable order was not significant. The criteria for adding variables is critical to the performance. The PRESS statistic was used here, and as long as it decreased variables were added.

There were two main competing models, these two areas can be seen in Figure 7-8. The goal is to have a small PRESS statistic with few model variables. The dual method was able to leverage the correlation information to add variables that resulted in a better

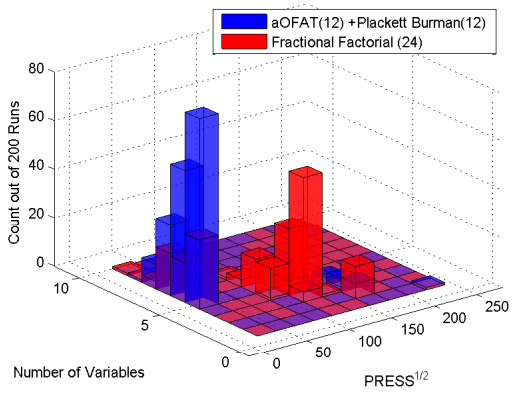

Figure 7-8: Wu and Hamada (2000) Analytical Experiment

model. The high aliasing in this experiment led to many equally compelling model options. A more informed covariance structure improved the probability that the correct selections were being made.

One complicating aspect of this selection is that models with an average of two extra variables better fit the data compared with models with fewer variables. The complete model would have been A, B, C, D, BC, CD, where B and D are extraneous variables. These additional variables are used to reduce the noise component. In real systems there is observed a regularity of inheritance where a significant interaction component normally has significant main effect. In this situation adding those components, even if superfluous, reduces the cross-validated PRESS error. This performance is similar to the predicted performance by Wu and Hamada (2000) while automatically selecting the model. If the modeler would like to actively participate in model selection the relative choice of important variables could be done outside of the physical experiment.

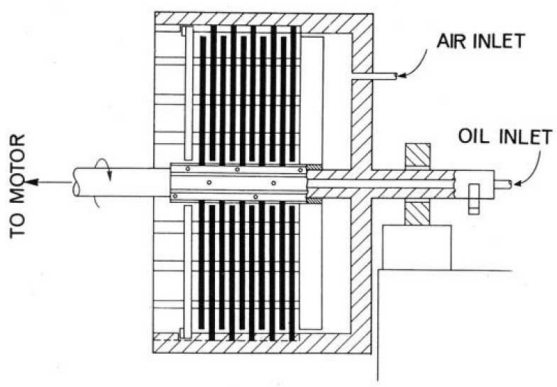

Figure 7-9: Wet Clutch Example

7.5.3 Wet Clutch Experiment

The final experiment used the results from a full-factorial wet-clutch experiment for analysis from Lloyd (1974). A wet clutch is used to disengage two shafts, an example is shown in Figure 7-9. For this particular experiment there were seven variables of interest, oil flow (A), pack clearance (B), spacer plate flatness (C), friction material grooving (D), oil viscosity (E), material friction (F), and rotation speed (G). The original experiment was created to optimize and improve the design of wet clutches.

Because this was an actual experiment there is no exact answer, and the true model is unknown. One "solution" was generated by using a Bayesian analysis on all of the runs from the full-factorial experiment. The significance level of the Bayesian analysis is set to 1%. This gives the model of A, C, D, E, F, G, BC, BD, BG, CE, CF, CG, DE, EF, FG.

For this non-replicated experiment another method of analysis is Lenth's method (Lenth, 1989). The main effects and important two-way interactions was provided by Li et al. (2006) as A, B, C, E, G, AD, AG, BD, BG, CD, CG, DE, and EG. The difference is primarily in the fact that Li et al. (2006) included three-way and four-way interactions in his

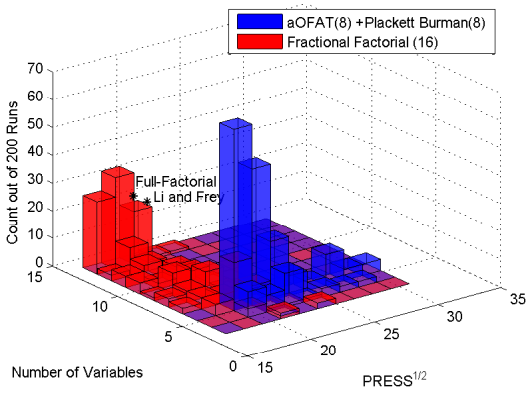

Figure 7-10: Wet Clutch Comparison

analysis, although the model only includes main and two-way interactions.

The results for this system are similar to the two previous examples. The dual approach is able to perform well against the larger model although due to the experiment size it identifies fewer terms as shown in Figure 7.5.3. The number of significant terms is surprising and followup experiments would have to decide on the number of parameters to include. The larger model was able to predict a greater percentage of the important variables and did not show the typical bimodal characteristic of the dual approach. In addition to the PRESS statistic, β significance and the R^2 procedure, an adjusted-R^2 calculation was also used and did not change the results.

7.6 Conclusion

In this chapter a method to augment current experimentation techniques through a dual approach was demonstrated. The initial experiment is an adaptive One-Factor-At-a-Time (aOFAT) search for the preferred setting followed by a supersaturated designed experiment. The initial aOFAT procedure finds the optimum result with 90% confidence and provides

covariance information. This experiment is followed by a highly saturated two-level experiment, in this case a Plackett-Burman design. The two results are combined through an empirical Bayesian procedure that utilizes hierarchical and heredity system characteristics. An adjustment improves the results when the two experiments differ in size. When faced with an industrial problem that requires both an optimum determination as well as a parametric model this dual approach can maximize the utility of each experimental run while accurately meeting both requirements. It is not necessary to select a optimum seeking experiment at the expense of a model building experiment.

Bibliography

Frey, D. D. and Sudarsanam, N. (2008). An adaptive one-factor-at-a-time method for robust parameter design: Comparison with crossed arrays via case studies. *ASME Journal of Mechanical Design*.

Frey, D. D., Sudarsanam, N., and Persons, J. B. (2006). An adaptive one-factor-at-a-time method for robust parameter design: Comparison with crossed arrays via case studies. In *Proceedings of 2006 ASME International Design Engineering Technical Conference and Computers and Information in Engineering Conference, DETC2006*.

Frey, D. D. and Wang, H. (2005). Towards a theory of experimentation for expected improvement. In *Proceedings of IDETC/CIE*.

Joseph, V. R. (2006). A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48:219–229.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics*, 31(4):469–473.

Li, X. and Frey, D. D. (2005). A study of factor effects in data from factorial experiments. In *Proceedings of IDETC/CIE*.

Li, X., Sudarsanam, N., and Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, 11(5):32–45.

Lloyd, F. A. (1974). Parameters contributing to power loss in disengaged wet clutches. *SAE Preprints*, 740676:10.

Montgomery, D. C. (1996). *Design and Analysis of Experiments*. John Wiley & Sons.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methoology*. Wiley.

Plackett, R. L. and Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33:305–325.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85:163–171.

Robbins, H. (1956). An empirical bayesapproach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1*.

Sargan, J. D. (1964). Three-stage least-squares and full maximum likelihood estimates. *Econometrica*, 32:77–81.

Wu, C.-F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, Inc.

Zimmerman, D. L. and Cressie, N. A. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics*, 42:27–43.

Chapter 8

Combining Data

This chapter expands the utility of adaptive experimentation to situations where two experiments are run on different systems. The two systems under experimentation may have different costs, timing, or quality. A frequent application is when one system is a computer experiment and the other a physical model. Finite element analysis (FEA) and computational fluid dynamics (CFD) are two examples of computer software that have good relative comparative value but have difficulty predicting absolute values. A small number of physical experiments are needed to correctly place the scale and bias of these computer estimates. These situations create unique challenges to experimenters, in selecting the best experiment for both conditions, as well as appropriate methodologies for combining the data. Here the focus will be on situations where the goal is to maximize the response while building the best model of the physical system.

8.1 Background

The foundation for this work is the ability to combine experiments from different sources. This is an area of active research and the procedure used here is a Bayesian Hierarchical Gaussian Process model similar to the one described in Qian and Wu (2008). This procedure was started in Kennedy and O'Hagan (2000), when they looked at combining two deterministic computer models. The real world experimental noise could not be included. In Kennedy and O'Hagan (2001) an extension was made to include physical models in addition to the computer models. These activities are known by different names including computer model calibration and surrogate model building. The most recent additions have been a model combination in a Bayesian framework (Qian et al., 2006; Reese et al., 2004). The Qian and Wu (2008) approach is generally applicable and could be applied to two computer models, a physical and computer model, or two physical models. The investigation here will focus on one physical model and one computer model. The only difference to two computer models is the inclusion of a noise term in the low quality model.

With the different costs of the low-quality and high-quality process the goal is to minimize the number of high-quality runs while getting the most accuracy in the combined model. Two different procedures will be used to create the set of high-quality run points- a standard all-variable procedure and an adaptive method that utilizes the results from the previous runs. The process used to combine the two data sets will be covered in detail before getting to the procedure specifics.

8.2 Process

The output of this technique is a conversion from a lower accuracy computer model with a bias and scale error to a higher accuracy physical model. The end result is a combined model that is tuned to that particular physical model. The generalization of this model to other physical instances should be evaluated carefully. The assumption behind this approach is that the computer model captures the general process characteristics but may be inaccurate for particular values or scale. Correcting the computer model based on physical points could be done by standard regression, however the problem is complicated by the disparate size of the computer experiment compared with the physical model. The underlying physics also may have complex interactions and few data points. One popular approach is to view the model as a hierarchical Gaussian random field model:

$$\hat{Y}_c(X) = F^T * \beta + \epsilon(X) \tag{8.1}$$

Where $\epsilon(\cdot)$ is a Gaussian random process with zero mean and variance equal to σ_c^2 and correlation function $R(\cdot|\theta_c)$. Where F is the input matrix, either a column of ones for an intercept model or a matrix of $F(x_i) = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1 \ldots n$ for a linear model. The inclusion of the linear effects assists in estimating the correlation coefficients as the number of runs grow. The reason behind this is clarified by looking at the likelihood estimate:

$$l = -\frac{1}{2}[n \log \sigma_z^2 + \log(\det(R)) + (y - f\beta)^T R^{-1}(y - f\beta)/\sigma^2] \tag{8.2}$$

As the number of runs grows the $(y - f\beta)$ term dominates the likelihood and the estimation of the coefficients of R is proportionally less accurate. Adding the linear terms reduces this error making the calculation significantly easier. Joseph et al. (2008) found that many

physical systems follow this linear effect property between the inputs and outputs.

The last consideration here is the correlation function. To be able to draw statistical conclusions from the gathered data some assumptions need to be made about the underlying process. Here it is assumed that the random process is stationary, thus for any time and spatial offset the cumulative distribution function (CDF) remains unchanged. Given the particular underlying function set ω from the population of possible functions Ω the output Y can be expressed as a function:

$$Y(x, \omega) = Y(x \in R^k; \omega \in \Omega) \tag{8.3}$$

Specifically, the assumption of second-order stationary (or identical CDF's) is used to estimate the model. Second-order or strong stationary requires that the first and second moments are time (and spatially) invariant. This results in ω as a particular realization of an outcome in Ω, that gives $E(Y(x)) = \mu$ for all $x \in R$. This condition requires that, for some function $C(\cdot)$, the covariance matrix satisfies:

$$Cov\{Y(x_1), Y(x_2)\} = C(x_1 - x_2) \tag{8.4}$$

In the implementations here, the function is also isotropic and is only dependent on $||x_1 - x_2||$. Given the process stationary requirement is a popular choice of correlation functions is the Gaussian correlation function. Bochner (1955) shows that any correlation function can be written in the form:

$$R(h) = \int_{R^d} \cos(h^T w) dF(w) \tag{8.5}$$

where F is a finite, positive, symmetric function. If the Gaussian distribution ($N(0, 2\theta^2)$) is

used for F then the following can be shown:

$$R(h) = \int_{-\infty}^{\infty} \cos(hw) \frac{1}{\theta\sqrt{2\pi}\sqrt{2}} \exp -w^2/\theta^2 4dw \tag{8.6}$$

$$= \exp(-(h/\theta)^2) \tag{8.7}$$

This function is a specific implementation of a larger family of correlation functions known as the power exponential correlation functions:

$$R(h) = exp(-|h/\theta|^p) \tag{8.8}$$

The choice of $p = 2$ gives the Gaussian function, although $p = 1$ has also been well-studied. The main choice of a correlation function corresponds to the desirable smoothness. Deciding between the different options should be made based on the underlying process. There are numerous definitions of continuity or smoothness but the general view is that as $p \to 2$ and the scale parameter $\theta \to 1.0$ the smoothness increases. Here, because $p = 2$, the only changes in smoothness will be due to changes in the correlation parameters θ. There is one other correlation function that should be mentioned for completeness. The Matérn correlation function was introduced by Matérn (1960). The choice of the t-distribution as F leads to the Matérn family of correlation functions.

$$R(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}} (\frac{2\sqrt{\nu}|h|}{\theta})^{\nu} K_{\nu}(\frac{2\sqrt{\nu}|h|}{\theta}) \tag{8.9}$$

Where K_{ν} is the modified Bessel function of order ν. As $\nu \to \infty$ the Matérn correlation function becomes the Gaussian correlation function. The additional parameter ν gives this correlation function tremendous flexibility in adjusting the smoothness. This parameter is specifically called the smoothness because the function is continuously differentiable up to

order $\nu - 1$.

The choice of high smoothness is a conservative choice without additional information about the system under investigation, and is popular in the literature (Santner et al., 2003).

8.3 Hierarchical Two-Phase Gaussian Process Model

This implementation of a Gaussian process model begins with a low accuracy (and low resource) model Y_c from the previous section. The output of this model is the input to the second phase.

$$\hat{Y}_p(x) = \rho(x)\hat{Y}_c(x) + \delta(x) + \epsilon(x) \tag{8.10}$$

This model takes in the $\hat{Y}_c$ model and makes a correction for scale (ρ) and for bias (δ). Both of these parameters are also Gaussian process (GP), $\rho = GP(\rho_0, \sigma_\rho^2, \theta_\rho)$ and $\delta = GP(\delta_0, \sigma_\delta^2, \theta_\delta)$. The hierarchical aspect of this model is in selecting the distributions for the model parameters, β, σ^2, and θ for each Gaussian Process. The choice of a prior distribution is important in the final sampling procedure. As pointed out by Gelman et al. (2003) the improper choice of priors can lead to misleading results. The priors that are used here are of a standard class. With a known mean and an unknown variance the likelihood for a n-vector of y observations is given as a $N(y|\mu, \sigma^2)$:

$$p(y|\sigma^2) \propto \sigma^{-n} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2) \tag{8.11}$$

$$= (\sigma^2)^{-n/2} \exp(-\frac{n}{2\sigma^2} \nu) \tag{8.12}$$

where ν is the known parameter:

$$\nu = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)^2 \tag{8.13}$$

The unknown parameters follow a conjugate prior distribution of the inverse gamma:

$$p(\sigma^2) \propto (\sigma^2)^{-(\alpha+1)}\exp(\beta/\sigma^2) \tag{8.14}$$

$$= \frac{\gamma^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-(\alpha+1)}\exp(\beta/\sigma^2) \tag{8.15}$$

Where $\Gamma()$ is the Gamma function. The α and β parameters are known as the hyper-parameters, and this is what leads to the hierarchical designation. These hyper-parameters will be chosen before running the simulation, and are ideally chosen with some knowledge of the system. After the variance is determined the mean parameters are drawn from a normal distribution. If the assumed mean is incorrect then this prior is no longer valid, and a different model is required.

The final parameters that must be determined are the correlation parameters. To determine the final distribution all of the individual probabilities are combined:

$$p(\beta, \sigma^2, \theta) = p(\beta, \sigma^2)p(\theta) = p(\beta|\sigma^2)p(\sigma^2)p(\theta) \tag{8.16}$$

Determining the $p(\theta)$ is challenging as it is independent of the scale and location parameters. The first choice is to integrate directly given information on β and σ^2.

$$p(\hat{Y}|Y) = \int\int\int p(\beta, \sigma^2, \theta|Y)d\beta d\sigma^2 d\theta \tag{8.17}$$

Drawing samples from a distribution of that complexity was only found feasible if the

priors for β and σ^2 were uninformed and improper. That direction led to problems with improper posterior distributions. As computational power increases sampling from this complex distribution may be feasible, but at nearly double the resources there may be some alternative options. Handcock and Stein (1993); Santner et al. (2003) both looked at this integration for systems of dimension two and found that a plug-in predictor has about 90% of the variance of this full Bayesian approach.

A plug-in estimate of θ is, in most cases, a Maximum Likelihood Estimate (MLE) of θ given the data. Zimmerman and Cressie (1992) showed for a kringing surface (or any Gaussian process) that the plug-in predictor underestimates the true variance. This situation is most problematic when θ is small. The amount of the underestimation is shown by Prasad and Rao (1990) to be asymptotically negligible for general linear models. The plug-in procedure is used here and caution is due when interpreting the variance estimates. If variance estimates are critical, Zimmerman and Cressie (1992) provide a correction that reduces the bias of the estimator.

In the situation here the following likelihood estimate is provided:

$$\begin{aligned} p(\theta_c, \theta_\rho, \theta_\delta | Y_c, Y_p, \beta, \rho, \delta) \propto \quad & p(\theta_c, \theta_\rho, \theta_\delta) \cdot \int_{\sigma_c^2, \sigma_\rho^2, \sigma\delta^2} \int_{\beta, \rho_0, \delta_0} p(\beta, \rho_0, \delta_0, \sigma_c^2, \sigma_\rho^2, \sigma\delta^2) \cdot \\ & p(Y_c, Y_p | \beta, \rho_0, \delta_0, \sigma_c^2, \sigma_\rho^2, \sigma_\delta^2, \theta_c, \theta_\rho, \theta_\delta) \\ & d(\beta, \rho_0, \delta_0) d(\sigma_c^2, \sigma_\rho^2, \sigma_\delta^2) \end{aligned} \tag{8.18}$$

Instead of expanding this into the full MLE form and then taking the integrals the reader is referred to the Appendix of Qian and Wu (2008). Before getting to the details of the MLE the prior distribution for θ still needs to be determined. The previous priors were determined to yield a proper posterior distribution, but for these variables the MLE makes that difficult. With the Gaussian correlation function used here, the unknown parameter θ

follows a Gaussian distribution so a proper prior is an inverse gamma distribution. With the MLE the prior distribution may be informative and dominate the results. The MLE results are always checked for a dominate prior and the variance of the prior distribution is increased as needed. Another approach is to use an uninformed prior, $p(\theta) = c$, this is discouraged as the resulting MLE may result in improper posterior estimates.

The lists of prior distributions include:

$$p(\sigma_c^2) \sim IG(\alpha_c, \gamma_c) \tag{8.19}$$

$$p(\sigma_\rho^2) \sim IG(\alpha_\rho, \gamma_\rho) \tag{8.20}$$

$$p(\sigma_\delta^2) \sim IG(\alpha_\delta, \gamma_\delta) \tag{8.21}$$

$$p(\beta|\sigma_c^2) \sim N(u_c, \nu_c \mathbf{I}_{k+1}\sigma_c^2) \tag{8.22}$$

$$p(\rho_0|\sigma_\rho^2) \sim N(u_\rho, \nu_\rho\sigma_\rho^2) \tag{8.23}$$

$$p(\delta_0|\sigma_\delta^2) \sim N(u_\delta, \nu\sigma_\delta^2) \tag{8.24}$$

$$\theta_c \sim IG(a_c, b_c) \tag{8.25}$$

$$\theta_\rho \sim IG(a_\rho, b_\rho) \tag{8.26}$$

$$\theta_\sigma \sim IG(a_\delta, b_\delta) \tag{8.27}$$

Because β includes linear terms it is of length $k+1$, where k is the number of x variables. The power exponential correlation function requires k terms so θ_c, θ_ρ, and θ_σ are all of length k. These are all of the hyper-parameters that need to be specified for the model. Using these hyper-parameters we can determine the conditional distribution.

Given the general model: $Y = F \cdot \beta$, and $p(\beta|\sigma) \sim N(u, v\sigma)$ solving for $p(\beta|Y)$:

$$\begin{aligned} p(\beta|y) &= p(Y|\beta)p(\beta) && (8.28)\\ &\sim \exp(\frac{1}{2\sigma}(Y - F\beta)^T R^{-1}(y - F\beta)) \cdot \exp(\frac{1}{2\sigma v}(u - \beta)^2) && (8.29)\\ &\sim \exp(\frac{1}{2\sigma} * (\beta^T(F^T R^{-1} F + 1/v)\beta + (u/v + F^T R^{-1} y)^T \beta)) && (8.30) \end{aligned}$$

This is a multivariate normal distribution, substituting:

$$\begin{aligned} \Sigma^{-1} &= (F^T R^{-1} F + 1/v)\frac{1}{\sigma} && (8.31)\\ p &= (u/v + F^T R^{-1} Y)\frac{1}{\sigma} && (8.32) \end{aligned}$$

The final distribution is $\beta \sim N(\Sigma p, \Sigma)$. This will be used for the distributions of β, ρ_0, and δ_0.

$$\begin{aligned} p(\beta|\cdot) \sim \quad & N([\tfrac{1}{v_c} I + F^T R_c^{-1} F]^{-1}(u/v + F^T R_c^{-1} Y), \\ & [\tfrac{1}{v_c} I + F^T R_c^{-1} F]^{-1}\sigma_c^2) \end{aligned} \tag{8.33}$$

Where R_c is the power exponential correlation matrix using θ_c that is found by a maximum likelihood estimate later in this section.

To simplify the equations the convention of Qian and Wu (2008) will be used. $\tau = \sigma_\delta/\sigma_\rho$ and $M = AR_\rho A + \tau R_\delta$ where A is a diagonal matrix with $\hat{Y}_c(x_p)$ on the diagonals and R_ρ and R_δ are the correlation matrices of θ_ρ and θ_δ.

$$p(\rho_0|\cdot) \sim N(\frac{u_\rho/\nu_\rho + \hat{Y}_c(x_p)M^{-1}(Y_p - \delta_0 1_{np})}{1/\nu_\rho + \hat{Y}_c(x_p)^T M^{-1}\hat{Y}_c(x_p)}, \frac{\sigma_r ho^2}{1/\nu_\rho + \hat{Y}_c(x_p)^T M^{-1}\hat{Y}_c(x_p)}) \tag{8.34}$$

$$p(\delta_0|\cdot) \sim N(\frac{u_\delta/(\nu_\delta\tau) + 1_{np}M^{-1}(Y_p - \rho_0\hat{Y}_c(x_p))}{1/(\nu_\delta\tau) + 1_{np}^T M^{-1}1_{np}}, \frac{\sigma_\rho^2}{1/(\nu_\delta\tau) + 1_{np}^T M^{-1}1_{np}}) \tag{8.35}$$

The conditional distributions on the remaining terms combine an inverse gamma and a normal distribution. Given an inverse gamma, $p(\sigma^2) \sim IG(\alpha, \gamma)$, and a normal $p(y|\mu, \sigma^2) \sim N(y|\mu, \sigma^2)$, with the continuing assumption that μ is known, then:

$$\begin{aligned} p(y|\sigma^2) &\propto (\sigma^2)^{-n} \exp(\frac{1}{2\sigma^2}(Y-\mu)^T(Y-\mu)) \\ p(\sigma^2) &\propto (\sigma^2)^{-(\alpha+1)} \exp(\gamma/\sigma^2) \end{aligned} \tag{8.36}$$

Combining these:

$$\begin{aligned} p(y|\sigma^2)p(\sigma^2) &\propto (\sigma^2)^{-(\alpha+1-n/2)} \exp(\frac{(Y-\mu)^T(Y-\mu)}{2\sigma^2} + \gamma/\sigma^2) \\ &\propto IG(\alpha + n/2, \gamma + (Y-\mu)^T(Y-\mu)/2) \end{aligned} \tag{8.37}$$

Applying this to the remaining variables the conditional distributions are:

$$\begin{aligned} p(\sigma_c^2|\cdot) \quad \sim \quad & IG(\frac{n_c}{2} + \frac{k+1}{2} + \alpha_c, \\ & \frac{(\beta_c - u_c)^T(\beta_c - u_c)}{2\nu_c} + \frac{(Y_c - F\beta)^T R_c^{-1}(Y_c - F\beta)}{2} + \gamma_c) \end{aligned} \tag{8.38}$$

$$\begin{aligned} p(\sigma_\rho^2|\cdot) \quad \sim \quad & IG(\frac{n_p}{2} + \frac{1}{2} + \alpha_\rho + \alpha_\delta, \\ & \frac{(\rho_0 - u_\rho)^2}{2\nu_\rho} + \gamma_\rho + \gamma_\delta + \\ & (Y_p - \rho_0 \hat{Y}_c(x_p) - \delta_0 1_{n_p})^T M^{-1}(Y_p - \rho_0 \hat{Y}_c(x_p) - \delta_0 1_{n_p})/2) \end{aligned} \tag{8.39}$$

The last conditional distribution is for τ, the simplification ($\tau = \sigma_\delta/\sigma_\rho$) leads to an irregular form:

$$\begin{aligned} p(\tau|\cdot) \quad \propto \quad & \frac{1}{\tau^{\alpha_\delta + 3/2}} * \\ & \exp(-\frac{1}{\tau}(\frac{\gamma_\delta}{\sigma_\rho^2} + \frac{(\delta_0 - u_\delta)^2}{2\nu_\delta \sigma_\rho^2})) \frac{1}{\sqrt{\det(M)}} \\ & \exp(-(Y_p - \rho_0 \hat{Y}_c(x_c) - \delta_0 1_{n_c})^T M^{-1}(Y_p - \rho_0 \hat{Y}_c(x_c) - \delta_0 1_{n_c})/2\sigma_\rho^2) \end{aligned} \tag{8.40}$$

After expanding all of the integrals and substituting the simplifications, the final likeli-

hood equations is:

$$\begin{aligned} L &= p(\theta_c, \theta_\rho, \theta_\delta) \int_\tau \tau^{-(\alpha_\delta + 3/2)} 1/\sqrt{\det(A)} \\ &\quad \frac{1}{\sqrt{\det(R_c)}} \frac{1}{\sqrt{\det(M)}} \cdot \frac{1}{\sqrt{DE}} \\ &\quad (\gamma_c + \frac{4C - B^T A^{-1} B}{8})^{-(\alpha_l + n_c/2)} \\ &\quad \cdot (\gamma_\rho + \gamma_\delta/\tau + \frac{4 * EG - F^2}{8E})^{-(\alpha_\rho + \alpha_\delta + n_p/2)} d\tau \end{aligned} \tag{8.41}$$

where:

$$A = \nu_c^{-1} I + F_c^T R_c^{-1} F_c \tag{8.42}$$

$$B = -2\nu^{-} 2\beta_0 - 2F_c^T R_c^{-1} Y_c \tag{8.43}$$

$$C = \frac{1}{\nu_l} \beta_0^t \beta_0 + Y_c R_c^{-1} Y_c \tag{8.44}$$

$$D = \nu_\rho^{-1} + \hat{Y}_c(x_p)^T M^{-1} \hat{Y}_c(x_p) \tag{8.45}$$

$$\begin{aligned} T &= (\nu_\rho^{-1} + \hat{Y}_c(x_p)^T M^{-1} \hat{Y}_c(x_p))(1_{n_p}^T M^{-1} 1_{n_p}) - \\ &\quad (\hat{Y}_c(x_p)^T M^{-1} 1_{n_p})^2 \end{aligned} \tag{8.46}$$

$$\begin{aligned} U &= -2[(\nu_\rho^{-1} + \hat{Y}_c(x_p) M^{-1} \hat{Y}_c(x_p))(1_{n_p}^T M^{-1} Y_p) - \\ &\quad (u_\rho \nu_\rho^{-1} + \hat{Y}_c(x_p)^T M^{-1} Y_p)(\hat{Y}_c(x_p) M^{-1} 1_{n_p})] \end{aligned} \tag{8.47}$$

$$\begin{aligned} V &= (\nu_\rho^{-1} + \hat{Y}_c(x_p)^T M - 1 \hat{Y}_c(x_p))(u_\rho^2 \nu_\rho^{-1} + \\ &\quad Y_p^T M^{-1} Y_c) - (u_\rho \nu_\rho^{-1} + \hat{Y}_c(x_p)^T M^{-1} Y_p)^2 \end{aligned} \tag{8.48}$$

$$E = (\nu_\delta T)^{-1} + TD^{-1} \tag{8.49}$$

$$F = -2u_\delta (\nu_\delta \tau)^{-1} + UD^{-1} \tag{8.50}$$

$$G = u_\delta^2 (\nu_\delta \tau)^{-1} + VD^{-1} \tag{8.51}$$

This problem can be separated for θ_c and $(\theta_\rho, \theta_\delta)$.

$$\hat{\theta}_c = \max_{\theta_c} p(\theta_c) \frac{1}{\sqrt{\det(R_c)}} \frac{1}{\sqrt{\det(A)}} (\gamma_c + \frac{4C - B^T A^{-1} B}{8})^{-(\alpha_c + n_c/2)} \tag{8.52}$$

This equation can be solved using a standard nonlinear optimization algorithm. Due to the sensitivity of the prior distribution and the discontinuous properties near zero a log transform is normally performed on L. The robust Nelder and Mead (1965) sequential simplex was found to provide good convergence although it was more resource intense compared with the quasi-Newton Broyden (1970); Fletcher (1970); Goldfarb (1970); Shanno (1970) (BFGS) method.

The the second part still has the integration:

$$\begin{aligned} \hat{\theta}_\rho, \hat{\theta}_\sigma &= \max_{\theta_\rho, \theta_\delta} \int_\tau p(\theta_\rho) p(\theta_\delta) \\ &\quad \tau^{-(\alpha_\delta + 3/2)} \frac{1}{\sqrt{\det(M)}} \frac{1}{\sqrt{DE}} \cdot (\gamma_\rho + \gamma_\delta/\tau + \frac{4 * EG - F^2}{8E})^{-(\alpha_\rho + \alpha_\delta + n_p/2)} d\tau \end{aligned} \tag{8.53}$$

There are a number of ways to solve this integration, the method used here and by Qian and Wu (2008) is the Sample Average Approximation (SAA) method of Ruszczynski and Shapiro (2003). The procedure is used to determine the expected value of a function by drawing values from a specific distribution. The goal is to begin by finding a suitable distribution for $\tau^{-(\alpha_\delta + 3/2)}$:

$$\tau^{-(\alpha_\delta + 3/2)} \propto \frac{2^{(\alpha + 1/2)}}{\Gamma(a)} \tau^{-(\alpha + 1/2 + 1)} * \exp(-2/\tau) * \exp(2/\tau) \tag{8.54}$$

$$= p(\tau) * \exp(2/\tau) \tag{8.55}$$

Given $\tau \sim IG(\alpha + 1/2, 2)$ then using the SAA method:

$$\int_{\tau} p(\tau) f(\tau) \approx \frac{1}{S} \sum_{s=1}^{S} f(\tau) \tag{8.56}$$

And the function $f(\tau)$ for this summation is:

$$f(\tau) = p(\theta_\rho) p(\theta_\delta) \exp(2/\tau) \frac{1}{\sqrt{\det(M)}} \frac{1}{\sqrt{DE}} \cdot (\gamma_\rho + \frac{\gamma_\delta}{\tau} + \frac{4 * EG - F^2}{8E})^{-(\alpha_\rho + \alpha_\delta + n_p/2)} \tag{8.57}$$

and putting everything together:

$$\hat{\theta}_\rho, \hat{\theta}_\sigma = \max_{\theta_\rho, \theta_\delta} \frac{1}{S} \sum_{s=1}^{S} f(\tau^{<s>}) \tag{8.58}$$

where $\tau^{<s>}$ is a vector of s independent draws from the inverse gamma distribution - $IG(\alpha_\delta + 1/2, 2)$. This method has been shown to be is asymptotically accurate in Shapiro and Nemirovski (2005). To solve this equation the Nelder and Mead (1965) sequential simplex was used, since the BFGS quasi-Newton method failed frequently when the determinant was close to zero.

8.4 Simulation Procedure

The procedure under evaluation is the use of the statistical procedure outlined above to combine two data sets. The first low quality data set is drawn from a space-filling Latin hypercube. The second data set is either a adaptive-One-Factor-at-a-Time (aOFAT) or a traditional star pattern run from a high-quality experiment. Both of these procedures minimizes the number of runs to adjust every variable. Each of the high-quality points is also run in the low-quality model, this improves the convergence by requiring fewer augmented

points. The procedure is as follows:

1. Create Artificial Response Surface (Krigifed Surface)
2. Generate Gibbs Draws from Conditional Distributions
3. Generate Metropolis Draws for the irregular distributions
4. Generate Metropolis Draws from predicted distribution (data augmentation)
5. Check Convergence and repeat if necessary

The details of the krigified Surface and the convergence checks are provided in a subsequent section. In this section the details of the Gibbs sampling, the Metropolis-within-Gibbs and the data augmentation approach will be discussed.

The Gibbs algorithm (Geman and Geman, 1984; Casella and George, 1992) is a method to implement Markov Chain Monte Carlo (MCMC) sampling. MCMC sampling requires sequential draws from an approximate distribution that is corrected as the chain progresses. Each sampling step is only dependent on the previous step, making it a Markov Chain. Each draw is designed to get the distribution closer to the target distribution. The Gibbs algorithm divides the update into a sampling vector, in this case $\psi = (\beta, \rho_0, \delta_0, \sigma_c, \sigma_\rho, \tau, \hat{Y}_p)$. This vector is updated in random order using the current values of the vector until the update is made. As the length of this chain grows it approaches the desired stationary distribution. Two variables, τ and $\hat{Y}_p$ cannot be sampled from a conditional posterior distribution. These two variables will be sampled through a Metropolis draw. This algorithm is an acceptance/rejection method based on a random walk. A random draw is made around the current point from a selected *jumping distribution*. The probability of both the new point and the current point are calculated and if the ratio is greater than a uniform random draw on $[0, 1]$ then the new point is accepted. The target acceptance rate is around

0.44 in one dimension, for τ, and 0.23 in multiple dimensions, for $\hat{Y}_p$. The reasoning behind these acceptance rates and further information about the Gibbs, Metropolis, and Metropolis-within-Gibbs can be found in Gelman et al. (2003).

The points for the $\hat{Y}_p$ predictions are calculated through data augmentation. The method used here was presented by Tanner and Wong (1987) for determining the posterior distribution when the parameter distributions are still being determined. Although they claim that the parameters posterior modes could be used, the highly correlated structure in this situation required continued sampling of the parameters from their converged distribution. This approach has advantages over the first method in Qian and Wu (2008) in that the predicted values are available at the end of the simulation without any further calculation. A question arises for this method- should it be included in the Gibbs loop or in a subsequent calculation? At any point in a Gibbs update there are some parameter values that are correct and some that are incorrect. Because the updated values are not used in any other parameter of the Gibbs process this update can be made at any time, including afterwards or before. If these values are used in any other step then this would have to be randomized to guarantee a reversible chain and convergence towards the stationary distribution

There is a probability that the Gibbs and metropolis algorithms may not reach the stationary distribution. This is problematic if there are two disparate regions of the distribution with similar probabilities. To detect these issues and other anomalies Gelman and Rubin (1992) suggests that running multiple sequences from an over-dispersed starting condition and measuring convergence is critical.

8.5 Convergence

Convergence of the MCMC algorithm is challenging to assess. Brooks and Gelman (1998a) describe many methods and problems in measuring convergence. The historic choice is to monitor the the trend of a single simulation. Although logically congruent, Gelman et al. (2003) shows that it is extremely difficult to distinguish convergence if the trend is extremely slow. Another method that is less ambiguous is to compare many parallel MCMC simulations. Gelman et al. (2003) propose taking a ratio of the total variance to the within simulation variance. Given m parallel simulations each with length n the simulation draws are ψ_{ij} $(i = 1, \ldots, n;\ j = 1, \ldots, m)$. The between ($B$) and within ($W$) variances can be calculated.

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\frac{1}{n} \sum_{i=1} n\psi_{ij} - \bar{\psi})^2 \tag{8.59}$$

$$\bar{\psi} = \frac{1}{n * m} \sum_{i} \sum_{j} \psi_{ij} \tag{8.60}$$

$$W = \frac{1}{m} \frac{1}{n-1} \sum_{j=1}^{m} (\psi_{ij} - \bar{\psi_i})^2 \tag{8.61}$$

$$\bar{\psi_i} = \frac{1}{n} \sum_{i} \psi_{ij} \tag{8.62}$$

The posterior variance estimate is a weighted average of W and B, and the ratio of that to the within variance gives the monitoring factor.

$$\hat{R} = \sqrt{\frac{\frac{n-1}{n} W + \frac{1}{n} B}{W}} \tag{8.63}$$

This is referred to as the Gelman-Rubin statistic (Gelman and Rubin, 1992) or the Potential Scale Reduction Factor (PSRF). The convergence of this statistic to 1.0 avoids the pitfalls of

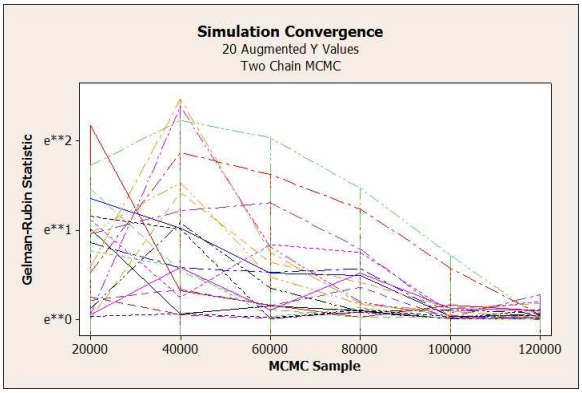

Figure 8-1: MCMC Convergence

visual techniques. The drawback is the convergence is only in the limit ($n \to \infty$). A sample of this convergence can be seen in Figure 8-1. Note that R is not monotonically decreasing with additional simulations. This is not unexpected but undesirable, and the particular simulation used here is prone to that situation. First, the Metropolis-within algorithm has a variance adjustment parameter. As that parameter is adjusted, the acceptance rate of the Metropolis algorithm changes and the variance changes. Second, half of our parameters have an inverse gamma distribution. The MCMC chains may not visit the tails enough, so a small visit to the tail increases the between variance substantially.

To improve the convergence a number of options exist. First the convergence properties could be measured from the model parameters and not the augmented data. These parameters converge faster and then posterior sampling for the augmented data could be performed at the mode. There are a couple of problems with this first, using the mode would eliminate the complex scale and bias transitions, decreasing the accuracy of the model. Second, the posterior distribution of the augmented data given the posterior mode of an earlier parameter simulation assumes that the distribution is degenerate with mass located at its mode. This assumption is very significant by reducing the correlation influence and variance esti-

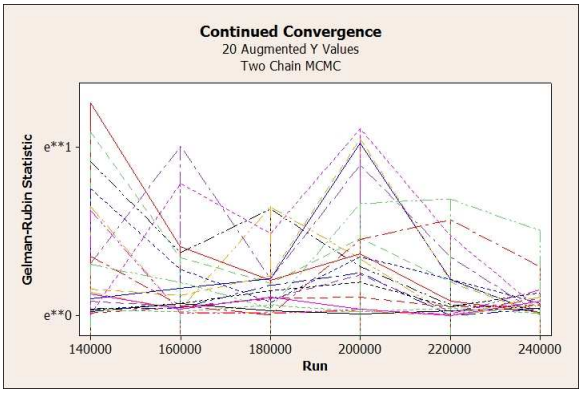

Figure 8-2: MCMC Convergence Continued

mates.

Given the general Gaussian process model:

$$\hat{Y} = f_0 \cdot \hat{\beta} + r_0^T R^{-1}(y - F\hat{\beta}) \tag{8.64}$$

$$\hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} y \tag{8.65}$$

In this situation the *y*'s are the predicted points from another model that is dependent on the distributions of σ_c and β. σ_c is a random variable with an Inverse Gamma distribution, and β has a normal distribution. Reducing these to point estimates does not reflect the long tail of σ_c or β making both inaccurate. Gibbs (1997) goes into greater detail on this influence.

Another way to demonstrate this problem. When the model in Figure 8-1 is run for an additional 120,000 simulations with the Gelman statistic calculated, the results are not consistent. The result of this is shown in Figure 8-2, and for these additional 120,000 runs the Metropolis-within algorithms had fixed variance parameters.

For this simulation, the lack of convergence can be addressed through a number of

methods. First the initial sample needs to be dispersed. This is more challenging than initially expected. If the samples are too dispersed then the first Gibbs sample drives all values toward the distribution mean and now everything is under dispersed. Another way of viewing this is that the autocorrelation for the different Markov chains affects the location of the point estimates while the correlation between chains at any particular location better reflects the final distribution.

Because of this autocorrelated walk, each chain may visit a low probability location for a disproportionate amount of time. This increases the between variation and not the within variation, and can explain the divergence. The utility of this statistic is highly dependent on the dispersion of the initial chains. Originally the problem was an inability to diagnose convergence, that has now been substituted for a problem of setting up disperse enough initial conditions. The use of the PSRF has been criticized in non-normal conditions. Brooks and Gelman (1998b) presents a number of alternative metrics, with the main suggestion a range metric, but other order metrics were suggested. They show that an average range metric can have too large a variance within chain, yielding an over-optimistic convergence statistic. The proposed method in this work extends this idea in two directions. First instead of using a range or standard deviation estimate, a more robust statistic of the Median Absolute Deviations (MAD) or S_n or Q_n (Rousseeuw and Croux, 1993) is used and second the predicted values will substitute for additional chains. A big disadvantage of using a variance or range estimate is when the distributions are not symmetric; the estimate is bias and can be influenced by a few low probability points. MAD is a good metric that has a 50% breakdown point (i.e. 50% of the data could be incorrect or arbitrarily large before the MAD metric was influenced), but it is symmetric and has a discontinuous influence function (the amount of change given a change in a single data point). S_n and Q_n both are more appropriate with non-symmetric distributions although Q_n has a smooth influence

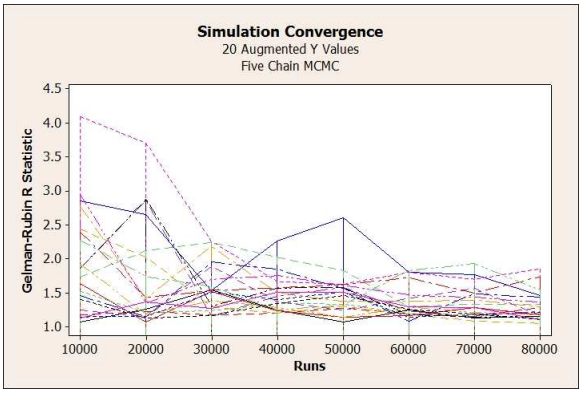

Figure 8-3: MCMC Convergence $\hat{R}$

function.

$$Q_n = d\{|x_i - x_j|; i < j\}_{(k)} k = \binom{h}{2} \approx \binom{n}{2}/4h = [n/2] + 1 \tag{8.66}$$

The Q_n statistic is the k^{th} order statistic of the $\binom{n}{2}$ inter-point distances, where k is approximately the number of half of the data points. This can be combined as Brooks and Gelman (1998b) did with other values into an order PSRF $\hat{R}$ value.

$$\hat{R}_Q = \frac{\frac{1}{mn} Q_n(i \in mn)}{\frac{1}{m*(n)} Q_n(Q_n(i \in n) \in m)} \tag{8.67}$$

A comparison of the two metrics is shown in Figures 8-3 and 8-4. Note that both show an artificial convergence at the same number of runs. The new metric is an improvement as it does not have a centered parameter and is solely a dispersion measure. It is better suited to non-symmetric distributions, like the ones here. Unfortunately, the computation time of the two metrics differs. The Rousseeuw and Croux (1993) algorithm for Q_n takes $O(n \log n)$ (an algorithmic improvement over expected $O(n^2)$) versus the variance calculation at $O(n)$.

An extension was suggested in Brooks and Gelman (1998b) to reduce the multiple

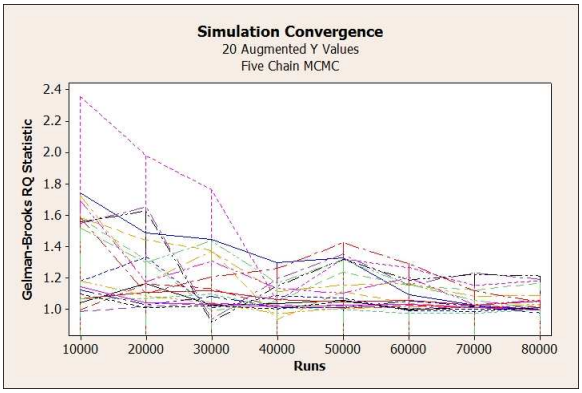

Figure 8-4: MCMC Convergence $\hat{R}_Q$

PSRF metrics to a single number. This was not used here because a slope characteristic could be used to determine convergence. The suggestion that Brooks and Gelman (1998b) gave was to be sure that the variances (both within and between) had settled down. They complete this through a graph of the variances. In the multivariate case shown here, if convergence has not been reached if one PSRF is increasing in value. This statement suggests a relationship between the number of estimated values and the number of simulated chains in identifying convergence.

The benefit of the methodology used here is the comparison between the variance ratios of within chains to between chains. This is a useful statistic in judging convergence as long as the starting points are over-dispersed. The number of required chains is an open problem. Gelman (1995) suggests that they should be sufficient (> 10). The number used here is three but, the convergence of multiple predicted values increases the actual number of unique starting points. The convergence of each of these points proceeds uniquely; and although not exactly equivalent to separate chains they provide a useful additional criteria. The biggest difference is that each predicted value uses the same values for the other parameters. The overall location in that parameter space is the same but each point is

in a unique part of that space.

The criteria for convergence is twofold. First each predicted parameter has to have a $R_Q < 1.2$; and second, the direction for each predicted value must decrease. Thus there are 60 unique starting locations that must all be near convergence and continuing on a convergent path. An indication of convergence for this metric offers sufficiency in all test cases. This was tested on 10 different krigified surfaces by doubling the final number of runs to check for any lack of convergence and non was found. Further investigation may show that this metric is too conservative and requires excessive runs, that issue will not be dealt with here.

8.6 Krigifier (Trosset, 1999)

Generating test cases to compare the different methodologies is difficult. The previous Hierarchical Probability Model (HPM) methodology that was used in other chapters only includes linear and interaction terms. This methodology is designed to work outside of the linear framework and is better suited to space filling designs. Data from real-world deterministic processes are noisy. This noise originates from many sources including the data-collection process, lurking variables, numerical roundoff, and process instability. This correlated deterministic signal could be approximated by a stochastic correlated signal. The process selected to generate these much less intense stochastic signals is the kringing procedure. This method was first developed by geostatisticians for interpolating a number of data points with a specific stochastic process (Wackernagel, 2002). The parameters for the stochastic process are first estimated and then used to fit the observed data. This process is extremely flexible, which is convenient to fit a wide variety of data but can have a frustrating number of parameters. To simplify the process here the underlying function is

a general second-order linear function. This was chosen to provide a maximum location, or a ridge, as suggested as a frequent function seen in experimental design (Myers and Montgomery, 2002). The noise was created using a stationary Gaussian process. The correlation function was a power-type function with $k = 1$; this yields the absolute value of the differences. This was selected over the more traditional $k = 2$ because the surfaces were noisier and Trosset (1999) suggests more realistic.

The procedure comes from Trosset (1999):

1. Create underlying quadratic trend
2. Create stationary Gaussian Process
3. Use Latin Hypercube to generate random points, $x_1, \ldots, x_n$
4. Generate $y_1, \ldots, y_n$ from the quadratic function
5. Interpolate $y_1, \ldots, y_n$ from the Gaussian Process to generate the noise
6. Sum the noise and trend terms to get the final $y_1, \ldots, y_n$ values

This process is used twice, once to generate the low accuracy and a second time to create the high accuracy data. The noise is zero for the low accuracy data versus a third of the signal for the high accuracy experiment.

8.7 Results

Both methods were run 250 times with different random krigified surfaces. Twenty random low accuracy points were generated for y_c using a latin-hypercube sampling method in seven dimensions. The eight high accuracy points y_p were generated with either method

and then also fit with the low accuracy points . This simple problem in seven dimensions took approximately 60 minutes on an Amazon-EC2 High-CPU Medium instance machine from Amazon Web Services (2008). Further parallelization is possible as the chains are currently run in series, but the computing resources would have to be increased.

After the chains converged for all of the y_p values the mode was used as the predicted value. The final results were normalized and the absolute error calculated. Each surface was randomly generated and so some had greater variance, and a greater range than others. Additionally, the star runs were started at a random point, which may have been close to the maximum point already.

To compare the results between the two starting conditions a regression line was fit to the data. A robust regression procedure was employed because of the large variance between the different krigified surfaces. The advantage of a robust fit was a tolerance for outliers. The robust fit procedure was an iteratively re-weighted least squares method using a bi-square weighing function.

The results are shown in Figure 8-5. The general outcome is as expected, there is a more negative slope for the aOFAT method compared with the with the star initialization. On average the aOFAT procedure moved to conditions of greater value, and thus made more accurate predictions around the maximum. If the aOFAT started at a ridge or peak then the runs were identical to a star procedure at that same location so the difference between the lines should not be too extreme.

If only the maximum for each run is compared, and not all of the runs, then this effect is highlighted even more in Figure 8-6. The star procedure began in a random location and so had a probability of starting at the maximum value and resulting in a lower error than the aOFAT.

This procedure could be used in situations where two competing objectives of system

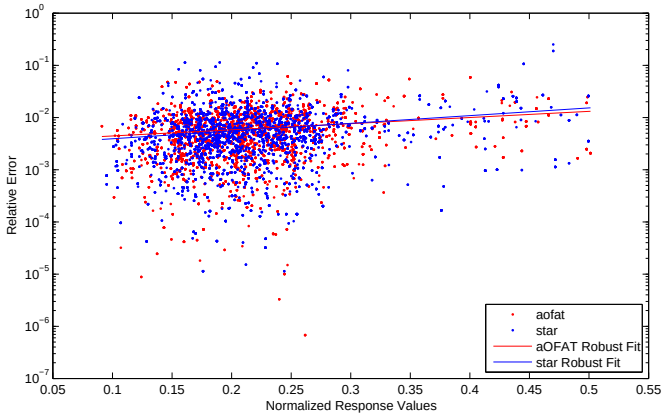

Figure 8-5: Prediction error

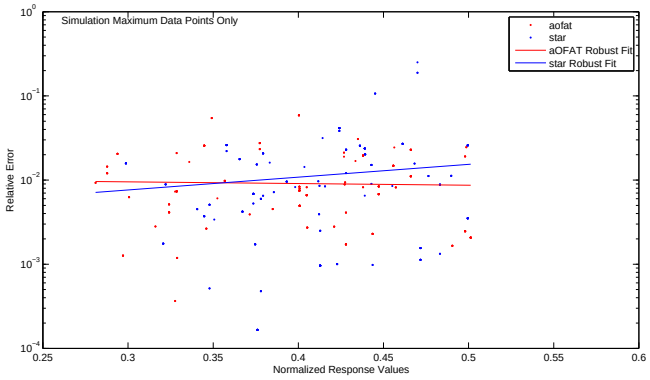

Figure 8-6: Prediction error for run maximum only

maximization and model parametrization are desired. The aOFAT method would build the model with a bias towards finding optimal points. Runs beyond the initial aOFAT runs presented here could be determined using a number of procedures such as Williams et al. (2000), Santner et al. (2003), or Currin et al. (1991). The appropriate total number of runs has been identified by both of these authors as an area of current research. There are few arguments that the minimum number of runs should be less than the total number of variables and this experiment is an appropriate method to initialize an experiment to prepare for further runs.

The procedure did not use a pairwise comparison as a time savings to implement the procedure on a number of different machines simultaneously, and thus required more runs. Future studies could compare some additional methodologies. One procedure could be to use a highly fractionated designed experiment. This was not addressed in this case because previous chapters of this thesis and Frey et al. (2003) looked at that comparison. Future challenges exist to define a subsequent experiment that continues to build the model after the $n + 1$ runs are complete. One direction that Currin et al. (1991) pursued is for each additional run to be selected to maximize the expected entropy reduction. A simple modification to get this result would be to change the entropy calculation from:

$$H(x) = E(-\log p(x)) - \log dx \tag{8.68}$$

to:

$$H(x) = E(-\log(y(x) * p(x))) - \log dx \tag{8.69}$$

which would be the same as maximizing the selection of $|y * \sigma||$. Currin et al. (1991) states this is the same as minimizing the weighted posterior variance of the unknowns.

8.8 Conclusion

Combining the experimental results from two different systems is a new and critical problem. In this work a method was presented to use aOFAT experiments for physical experiments combined with latin hypercube computer experiments. A new metric of convergence was presented, as well as a technique for using value predictions instead of additional chains. It was shown that the aOFAT methodology creates a model that is biased towards accuracy at the maximum values. This method is effective in creating a good model around the system values of interest. The implementation potential ranges from physical and analytical models to different computer models or even human expert opinions. The Bayesian technique presented in this chapter is one method that has proven useful in a number of previous problems. There are different approaches to combine two experiments but, all methods require some initial high-cost experimental points where the aOFAT methodology provides good experimental value while focusing on the maximum.

Bibliography

Amazon Web Services (2008). Amazon elastic computing cloud. http://aws.amazon.com/ec2/.

Bochner, S. (1955). *Harmonic Analysis and the Theory of Probability*. University of California Press, Berkeley.

Brooks, S. and Gelman, A. (1998a). Some issues in monitoring convergence of iterative simulations. In *Interface 1998 - Computing Science and Statistics*.

Brooks, S. P. and Gelman, A. (1998b). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms. *Journal of the Institute of mathematics and Its Applications*, 6:76–90.

Casella, G. and George, E. I. (1992). Explaining the gibbs sampler. *The American Statistician*, 46:167174.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functionss, with applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, 86 No 416:953–963.

Fletcher, R. (1970). A new approach to variable metrick algorithms. *Computer Journal*, 13:317–322.

Frey, D. D., Englehardt, F., and Greitzer, E. M. (2003). A role for "one-factor-at-a-time" experimentation in parameter design. *Research in Engineering Design*, 14:65–74.

Gelman, A. (1995). *Practical Markov Chain Monte Carlo*, chapter Inference and Monitoring Convergence, pages 131–143. London: Chapman and Hall.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511.

Geman, S. and Geman, D. J. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gibbs, M. N. (1997). *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, Cambridge University.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means. *Mathematics of Computation*, 24:23–26.

Handcock, M. S. and Stein, M. L. (1993). A bayesian analysis of kriging. *Technometrics*, 35:403–410.

Joseph, R. V., Hung, Y., and Sudjianto, A. (2008). Blind kriging: A new method for developing metamodels. *ASME Journal of Mechanical Design*, 130:1–8.

Kennedy, M. C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87:1133–1152.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society*, Ser. B, 63:425–464.

Matérn, B. (1960). *Spatial Variation*. PhD thesis, Meddelanden fran Statens Skogsforskningsinstitut.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methoology*. Wiley.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Journal of Computation*, 7:308–313.

Prasad, N. G. N. and Rao, J. N. K. (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85:163–171.

Qian, P. Z. G. and Wu, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50:192–204.

Qian, Z., Seepersad, C. C., Joseph, V. R., Allen, J. K., and Wu, C. F. J. (2006). Building surrogate models based on detalined and approximate simulations. *ASME Journal of Mechanical Design*, 128:668–677.

Reese, C. S., Wilson, A. G., Hamada, M., Martz, H. F., and Ryan, K. J. (2004). Integrated analysis of computer and physical experiments. *Technometrics*, 46:153–164.

Rousseeuw, P. J. and Croux, C. (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88:1273–1283.

Ruszczynski, A. and Shapiro, A., editors (2003). *Stochastic Programming Handbooks in Operations Research and management Science*. Elsevier, Amsterdam.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer.

Shanno, D. F. (1970). Conditioning of quasi-newton methods for function minimization. *Mathematics of Computation*, 24:647–656.

Shapiro, A. and Nemirovski, A. (2005). *Continuous Optimization: Current Trends and Applications*, chapter On Complexity of Stochastic Programming Problems, pages 111–144. Springer, New York.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–.

Trosset, M. W. (1999). The krigifier: A procedure for generating pseudorandom nonlinear objective functions for computational experimentation. Interim Report 35, Institute for Comptuer Applications in Science and Engineering, NASA Langley Research Center.

Wackernagel, H. (2002). *Multivariate Geostatistics*. Springer, New York.

Williams, B. J., Santner, T. J., and Notz, W. (2000). Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica*, 10:1133–1152.

Zimmerman, D. L. and Cressie, N. A. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics*, 42:27–43.

Chapter 9

Conclusions

This work focused on combining adaptive experiments with designed statistical experiments. Each of the techniques involved using adaptive-One-Factor-at-a-Time (aOFAT) experiments, as well as other standard statistical methodologies. Run reuse from a prior adaptive experimentation was the initial area addressed. The adaptive experiment cannot be preplanned and so the potential run reuse in the subsequent experiment is stochastic. A number follow-up experimental options were investigated. First, the use of a traditional fractional factorial design in the follow-up experiment where the fraction was pre-selected or based on the greatest reuse. Depending on the number of variables and size of fraction, the number of runs reused asymptotes to approximately twenty percent of the total aOFAT runs. This run reuse was demonstrated on a number of actual experiments as well as surrogate experiments. The second area of investigation was non-balanced D-optimal designs to increase run reuse. As suggested in Wu and Hamada (2000), a fully orthogonal non-balanced D-optimal design is a good alternative to a fractional factorial design. This change dramatically improved run reuse to fifty percent, and fits in the framework of planning the design after an initial aOFAT is complete.

In addition to investigating the number of reused runs, the independence of the resultant maximum estimates was also demonstrated. Running an adaptive experiment before a statistical experiment creates an opportunity for run reuse while providing an independent maxima estimate and some response information.

The adaptive experimental approach could also be used on the manufacturing floor. The method of evolutionary operation (EVOP) was revisited with a focus on utilizing adaptive experimentation. The alignment of this continuous improvement technique with the sequential maximization nature of an aOFAT provides a useful combination. Box and Draper (1969) concluded that the use of this methodology was na ive. This conclusion is challenged by investigating actual system responses and showing a place for sequential adaptive experiments. Instead of using small fractional factorial experiments, repeated single steps in an adaptive procedure was shown to be more robust to initial and continued variable selection. Because of the stochastic nature of the repeated procedure a modified Gibbs sampler was introduced to minimize the additional runs while converging to a better variable setting. An offshoot of this procedure is the use of an adaptive experiment in computational unconstrained function maximization.

The modified sequential simplex procedure was originally developed for evolutionary operation. Although, this ranked-based geometric procedure was used frequently in the 1970's and 1980's, it was replaced by more complex derivative-based methods. More recently it has returned to popularity with the increased use of computer simulations. As a robust method it is able to handle discontinuities and noise at the cost of more function evaluations. There are implementations of the simplex in most numerical programs for unconstrained optimization. The typical initial setup is based on changing one variable at a time. This was improved by adding an adaptive element and performing an aOFAT initially. In this situation the aOFAT procedure was changed to align the geometric cen-

ter to that of the non-adaptive method. Through the adaptive procedure and the step-size improvement, the overall convergence is increased and the number of function evaluations was reduced. The adaptive procedure is aiming the simplex, and thus reducing the distance to the improved operating conditions. This improvement was demonstrated on a test suite for numerical optimization.

Outside of the optimization another issue faced in computational methods is variable selection. Using the Mahalanobis-Taguchi Strategy (MTS), data classification is based on a statistical distance. One hurdle to using this system is in selecting the best variables for classification. Traditionally orthogonal arrays are used to select the best variables. This method can be improved by using an aOFAT experiment for variable selection. This procedure was specifically applied to an image classification system where the variables of interest are the coefficients of a wavelet transform. In this case the addition of variables adds to the computational load of the classification system. It is important to add the minimum number of variables while maximizing their usefulness.

To further the benefit of running an aOFAT experiment along with a statistical experiment, methods to combine both data are investigated. Combining two different types of data was approached in a Bayesian framework. The use of a correlated Gaussian random variable to make a posterior prediction has been used successfully by Joseph (2006). Part of this methodology is to use a correlation matrix for the input variables. Instead of using a larger experiment the information was divided between an early aOFAT experiment to create the correlation matrix followed by a highly aliased Plackett-Burman design. This goal is to combine the relative strengths of both of these procedures. The aOFAT can be used to create a variable ranking while the aliased design is able to efficiently define the model. A procedure to define the correlation matrix was created that benefits from published data regularities and variable distributions. This method performs equivalently to

using an uninformed correlation matrix and a larger experimental design. The procedure was demonstrated on a number of published examples as well as surrogate functions.

The last aspect of adaptive experiments was to combine experiments of different accuracy. Combining computational and physical experiments is one example of these different accuracies. The use of an adaptive experiment uses a minimum number of runs while likely having points near the maximum. A new method of calculating convergence was presented as well as a procedure to maximize each simulated Markov chain. The result was a procedure that provides a good model using both data types that is more accurate near the maximum values.

9.1 Future Work

Demonstrating the potential of applied adaptive experiments should open up greater opportunities for their application in the overall experimental process. This work specifically focused on aOFAT experiments but, there are other adaptive methodologies which could be investigated. One area of investigation is to find an adaptive procedure that can also be used outside of solely function maximization. Modifying Soból (1990) sequences to be adaptive from the previous information may be one possibility.

The use of the Bayesian framework to combine multiple models is a current area of investigation. The application is slow and incompatible with larger data sets, finding faster methods for data combination would leverage greater opportunities for the method in industrial practice. Creating an application as a web-based service is one possibility to overcome the computational limitations.

9.2 Summary

The goal of this work was to create a foundation for the integration of adaptive experimentation and statistical experimentation in practice. Simple techniques were presented for running the setup experiment and getting some benefit from those runs. This continues to the factory floor where evolutionary operation was improved and simplified with adaptive experiments. A numerical maximization procedure was improved through a better starting approach, and a classification procedure was shown to benefit from an adaptive parameter selection technique. The final area focused on using data from an adaptive experiment and a traditional experiment. First, the covariance calculation was improved to yield more accurate and smaller models with the same number of runs. Second, incorporating data from two different sources was shown to benefit from one adaptive experiment. The overriding goal for all of these procedures is to extend the framework for adaptive techniques to a greater audience and provide tools necessary for application.

Bibliography

Box, G. E. P. and Draper, N. R. (1969). *Evolutionary Operation: A Statistical Method for Process Improvement*. John Wiley & Sons, Inc.

Joseph, V. R. (2006). A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48:219–229.

Soból, I. M. (1990). Sensitivity estimates for non-linear mathematical models. *Matematicheskoe Modelirovanie*, 2:112–118.

Wu, C.-F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, Inc.

Appendix A

Adaptive Human Experimentation

The ability to understand the variance of an engineering system is historically done in a design, build and test cycle (Pahl and Beitz, 1995). Newer technology has pushed the envelope with computer simulation and virtual experimentation, but state-of-the-art variance prediction is limited due to necessary simplifying computational and mathematical assumptions and by model inadequacy (Petroski, 1994). These assumptions limit the model fidelity and can lead to unforseen, and early, product failures. There have been improvements in greater statistical experimentation (the six-sigma process (Creveling et al., 2003) and designed experimentation (Wu and Hamada, 2000)), and more complex mathematical modeling. Even with these methods, predicting failures early in the design process is challenging. First, mathematical or computer models are incomplete, leading to underlying assumptions that cannot test the true variance of the system. Second, early in the process there are no physical prototypes to validate the computer models or conduct robustness experiments. Moveover, the adequacy of any initial prototypes in reflecting the final design as made is a large unknown. The current best method is to depend on expert estimates and historic data to predict the future potential of alternative designs. This extrapolation has its

limitations as Petroski (1994) discusses.

Humans are superior to computers in creative solutions, making loose associations, thinking dynamically, and bringing in unique perspectives. Computers are better at organization, statistical computation, data storage and retrieval, and mechanistic processing. This chapter discusses the possibility of combining the benefits of both of these systems and apply them to early process mechanical design and simulation problems.

A.1 Layout

It may be possible to improve the value, and quality, of predictive models in accurate system estimation by using distributed human knowledge combined with statistical data analysis techniques. Combining the tacit knowledge of a significant number of different viewpoints is known to yield better estimates in other disciplines (Surowiecki, 2004) this has not been applied to systematically exploring system characteristics. Additionally, correct use of designed experiments within this distributed knowledge can lead to more powerful statistical estimates. A similar approach, although to business problems, has been explored in a recent MIT thesis by Tang (2006).

There are three levels of models for this combined system, first the model of the actual mechanical system under investigation, second the combined model that has been created from the lower fidelity models (using one of the previous methods mentioned in this thesis) and third the model of the interactions of the individuals and their interpretation, biases, and previous knowledge. The most challenging for future research is this third model, it is needed to explore the important aspects of combining human knowledge. Ideally, the fidelity of this model should be sufficient to understand group cognitive ability when solving these problems. A number of different model types could be explored to find one

that best represents this situation. To validate this model, experiments could be created that are based on academic environments and industrial settings. The long term research benefit of this combined human performance model will be to understand the potential of this technique as a tool to improve robustness, discover its application limitations, and create guidelines for use.

The experiments need to be built in a manner consistent with current research in human psychology, expert and leadership studies, and designed experimentation. It is important to be able to distinguish able users, identify problems and guessing, and provide reasonable judgment bases.

Research should focus on different aggregation techniques to deliver a capable model based on distributed knowledge. There are many options to combine opinions and create accurate models of the system variation with respect to the variance in opinions. Questions of interest include how to weigh the different opinions, how to create an accurate model of variance, and how to disassociate the system from the observer variance and to what degree does the model represent the system versus human variance. The result of this model can then be used as a surrogate system model, be used to plan experiments, and to validate existing results.

It will be necessary to create a tool that interacts with users, performs the calculations and returns these combined opinion models. The output from this tool will be used to train the combined human knowledge model.

A.2 Background

Combining the distributed power of human computation has been demonstrated in numerous applications (Barr and Cabrera, 2006; Westphal et al., 2005; Gentry et al., 2005). Some

applications include games, like the ESP Game (Ahn and Dabbish, 2005), others are focused on scientific knowledge, like the Stardust@Home (Westphal et al., 2005) while others are interested in making money, like Amazon's Mechanical Turk.

There have been initial investigations into the statistical and game theoretic aspects of these interactions (Gentry et al., 2005). This previous work focused on the comparison to distributed computing and security/cryptology issues. There has been little progress in exploring the statistical nature of these systems (other than cheater detection) and better incorporation of human psychological and physiological aspects.

Group interactions have been modeled as cooperative or a Pareto optimum, non-cooperative or Nash formulations, or supervisor/subordinate or Stackelberg formulations. In early design modeling influences can include educational background, corporate reporting department, interest area, or other motivation such as recent conversations, fatigue, or even attitude. It is not feasible to understand all of the influences of each individual but, ideally the aggregation techniques filter these out and reach a coherent model that predicts the human model performance. The results from these models may be compared to the performance of quality teams. Teams debate the merits of different models and frame the problem correctly and deliver quality predictions. The problem with this ideal behavior is that it is difficult to see in much of the corporate bureaucracy (Schon, 1995). The more anonymous method proposed here is more congruent with corporate performance metrics but cannot be used on the breadth of problems that a diverse, well functioning team could. The objective is targeted to frameworks where DOE's would be applicable. (Shih et al., 2006) argues that decision making through confrontational, and not individual cognition, yields high value through discussion and competition. But (Otto and Wood, 2001) argue that the drawbacks to this confrontation not encountered individually (or in the low pressure on-line environment) include the difficulties with team decisions including individual

dominance, misdirected focus, or a rushed time-frame. The methods proposed here ideally address these issues by offering an alternative modeling technique that is predicated on the idea that the general population is correct.

There has been research utilizing humans in a supervisory role in computer experiments and less as the subject of the experimentation. These architectures utilize important supervisor aspects of humans along with computer and analytical 'agents' the majority of this literature is in the AI community (Khosla et al., 2004). This differs from the research here as the role of the human is as a computational unit, not as a supervisor.

There is a large literature around emergent intelligence (Bonabeau et al., 1999), and while it may be possible that the group solves problems impossible for each individual, thus exhibiting collective intelligence, the group interactions in this case are not as important as seen in swarm intelligence. This could be investigated by looking at the importance of the aggregation process as well as when individuals are presented with alternative opinions. It will be critical to determine the decision making structure, either by simple voting (as seen in most collective intelligence systems) or through a more complex aggregation mechanism (Torra and Narukawa, 2007).

A.3 Potential Research

The research could extend the modern computational and analysis design paradigm to include the human as an integrated part of the system. A model of this new system could be created and validated through human experimentation. Some good possible models include agent based modeling and decision field theory (Busemeyer and Townsend, 1993).

The experiments are an integral part of this research. Investigations should focus on the methodology to create valid distributed experiments that are able to utilize the best

of human expertise, psychology, and designed experimentation. These experiments will require the creation of a tool that can generate validation data as well as benefit the company and user to entice participation. To ensure that this methodology is valid across potential design information users both academic and industrial examples will be sought.

Building on the foundations of statistical experimental designs (Wu and Hamada, 2000) and expertise tests (Klein, 1998) an experimental system can be created with checks for consistency and accuracy. Insight from the experimentation itself may also be possible, there may be additional biases explored and some unforseen pitfalls discovered.

The experiments focus on designer, or human, intuition. This direction faces a number of obstacles including understanding the problem, absolute or comparative analysis, reaching conclusions for multi-attributes, and the effects of teamwork.

During these experiments attempts will be made to investigate designer biases, inconsistency, and feedback delay. Some of these effects are well documented but others, especially when dealing with distributed teams, have not been studied.

A.4 Work

The research could be initiated though a number of human experiments. The best options are computer, or web, based studies to solicit the input from designers in a number of problems. Three proposed studies are presented here but, this is just a suggested layout and there are many other options.

The first study could investigate variable choices for experimental design, this area is called intuition and variable decision. Choosing variables for a designed experiment is difficult and the result could determine the effectiveness of the experimental run. Ideally variables are important, independent and inexpensive. Poor choices lead to experiments

that are challenging to run, excessively large, and nearly impossible to interpret. This study will focus on understanding the variables of interest by asking a number of individuals. The variables discovered could be classified into one of four groups: those that are everyone agrees to being important, those that are agreed to being unimportant, those that are disputed but are unimportant, and finally those that are disputed and important. Creating an experiment that is able benefit from this knowledge will reduce time and effort while producing rich data and useful results. These data will be gathered through the web and combined using some expert based hierarchy. The expertise for the users will be determined through a combination of known answers as well as some cluster analysis. The individuals fall into specific groups and are classed together. This classification along with some known questions will be used to grade the classes and weigh the individual inputs.

The second study could investigate differing expert rankings. This would be an attempt to self-regulate and learn about the participant expertise. This study will maximize the natural cognitive ability through pairwise comparisons and simple evaluation.

The third and final implementation of these experiments will be extended to greater design evaluation. These designs will not just be evaluated based on performance but also in robustness and originality. Problems that can be presented in this manner are difficult to test, complex, or from a variety of domains, as in mechatronic problems.

This system will use standard experimentation formulation from Montgomery (1996); Wu and Hamada (2000) to pose the problem to the human computer and then return the result. By using the humans the result should be creative, original, and intelligent and the computer should help maintain that the response is unbiased, quick, and universal. The result will directly benefit from the participant diversity and create a network of users eager to experiment with their new designs and see the designs of others. To help include the participants in the process there will be some visual cues to help them realize the status of

each of the projects.

A.5 Previous Work

In addition to the articles and books mentioned above the researchers listed here are also active in this area -

Gerd Gigerenzer - Adaptive Behavior and Cognition - Max Plank Institute - He explores the simple heuristics that are used every day to help us succeed. There are certain inherent biases when dealing with human intuition that need to be understood and avoided to achieve maximum results.

Norman Johnson - LANL - Symbiotic Intelligence Project. He created a system that uses internet and human actors to solve complex problems by creating networks of these simple actors. He uses the theory of evolutionary biology to advance individual solutions and kill off under performing solutions. They use the self-organizing nature of the agents to create these networks and organize solutions.

Luis von Ahn - Human Computation - He created CAPTCHA's and a number of games that are based on the idea of an underlying computation behind a game environment (espgame and peekaboom). The idea that computers can do certain calculations that cannot be completed easily (or ever) by a computer.

A.6 Potential Contribution

Utilizing distributed human knowledge to tackle design problems will create early models that are quick to create and give an accurate system performance estimate. This technology will foster greater creativity, earlier design iteration, and a greater confidence in the result.

Feedback from a diversity of sources, all with different opinions provide powerful potential to improve designs and validate opinions. This feedback, combined with an appropriate statistical methodology can improve the design process and increase the effectiveness of the designer.

The algorithms presented in the previous chapters focused on utilizing an adaptive experiment in addition to a traditional experiment. One potential adaptive experiment is to use human knowledge to determine variable importance, create covariance matrices, or to create composite models of a more expensive or complex experiment, all three of these methods are presented in previous chapters of this thesis.

Bibliography

Ahn, L. V. and Dabbish, L. (2005). Esp: Labeling images with a computer game. In *American Association for Artifical Intelligence Spring Symposium - Technical Report*.

Barr, J. and Cabrera, L. F. (2006). Ai gets a brain. *Queue*, 4(4):24–29.

Bonabeau, E., Dorigo, M., and Theraulaz, G. (1999). *Swarm intelligence : from natural to artifical systems*. New York : Oxford University Press.

Busemeyer, J. R. and Townsend, J. T. (1993). Decision field theory : A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, 100(3):432–459.

Creveling, C. M., Slutsky, J. L., and Antis, D. (2003). *Design for Six Sigma in technology and product development*. Prentice Hall.

Gentry, C., Ramzan, Z., and Stubblebine, S. (2005). Secure distributed human computation. In *EC '05: Proceedings of the 6th ACM conference on Electronic commerce*, pages 155–164, New York, NY, USA. ACM Press.

Khosla, R., Lai, C., and Mitsukura, Y. (2004). Human-centered multiagent distirbuted architecture for knowledge engineering of image processing applications. *International journal of pattern recognition & Artifical intelligence*, 18(1):33–62.

Klein, G. (1998). *Sources of Power: How People Make Decisions*. MIT Press.

Montgomery, D. C. (1996). *Design and Analysis of Experiments*. John Wiley & Sons.

Otto, K. N. and Wood, K. L. (2001). *Product Design: Techniques in Reverse Engineering and New Product Development*. Prentice Hall.

Pahl, G. and Beitz, W. (1995). *Engineering Design: A Systematic Approach*. Springer. translated by Ken Wallace, Lucitenne Blessing and Frank Bauert.

Petroski, H. (1994). *Design paradigms : case histories of error and judgment in engineering*. Cambridge University Press.

Schon, D. (1995). *The Reflective Practitioner: How Professionals Think in Action*. Ashgate Publishing.

Shih, S.-G., Hu, T.-P., and Chen, C.-N. (2006). A game theory-based approach to the analysis of cooperative learning in design studios. *Design Studies*, 27(6):711–722.

Surowiecki, J. (2004). *The wisdom of crowds : why the many are smater than the few and how collective wisdom shapes business, economies, societies, and nations*. New York : Doubleday.

Tang, V. (2006). *Corporate Decision Analysis: An Engineering Approach*. PhD thesis, Massachusetts Institute of Technology.

Torra, V. and Narukawa, Y. (2007). *Modeling Decisions : Informatino Fustion and Aggregation Operators*. Springer.

Westphal, A. J., Butterworth, A. L., Snead, C. J., Craig, N., Anderson, D., Jones, S. M., Brownlee, D. E., Farnsworth, R., and Zolensky, M. E. (14-18 March 2005). Stardust@home: A massively distributed public search for interstellar dust in the stardust interstellar dust collector. In *Thirty-Sixth Lunar and Planetary Science Conference; Houston, TX*.

Wu, C.-F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. John Wiley & Sons, Inc.

Appendix B

Replacing Human Classifiers: A Bagged Classification System

This application of human classifiers demonstrates an area of human computation and a method for aggregation. This early research method could benefit from individual adaptive experiments and a broad overall aggregation technique. This initial study focuses on automating a human classification process. The goals are to: improve classification consistency, assign confidence level for each automated classification, and have no increase in workload throughout the implementation. The proposed method uses multiple bagged classification trees, initially for the individual classifications and then applied to the combined group.

Each human classifier trains a separate bagged classification tree. An estimate of the classifier confidence is created and shown to be accurate. These individually trained classifiers are combined through a group decision algorithm. The 76% reduction in work allows the workers to train an additional classification tree on the most difficult cases. This additional tree is used in a weighted combination with the previous trees to improve the estimate

and reduce the workload.

This procedure is straight-forward and the results, classification plus confidence, are easily explainable to the human classifiers. This procedure is demonstrated on U.S. Post Office zip-code data, showing the ease of implementation and improvement, but could be used on a variety of classification problems.

B.1 Introduction

In most classification schemes the training data is assumed to be correct, and the goal of the classifier is to emulate that data, in many situations that correctness assumption is invalid. A more realistic case is when humans are classifying images, in this case numerical zip codes, and are only about 85% accurate. There are a number of humans performing this task in parallel, with each zip code being read once and each human differing in their accuracy.

With the same number of person-hours, the goal is to implement an automated system that improves throughput while maintaining classification accuracy. The procedure starts with training a bagged tree classifier for each individual. This individually trained tree will then be used to reduce that individual's work load. To maintain the current accuracy a confidence estimate is created for the classifier and all low-confidence images are reviewed by the individual. The confidence estimate is created uniquely for each classifier and is based on that specific human trainer.

After separate individual classifier systems are created for each human classifier the predictions are then compared and integrated in a decision algorithm. The low confidence predictions are returned to the human classifiers for a better classification. These returned and reclassified images are used to train an additional classifier, eventually to be added to

the decision algorithm.

In this demonstration some of the typical classification problems are not present. All of the training and validation cases come from a uniform distribution of zip code numbers. The training is done in a short time period and the noise is mainly driven by a forced short, and random, cycle time.

This final process is straight-forward, and easy to explain to the human classifiers, and allows them to focus on devising better, and more consistent classification rules or procedures, for the difficult cases.

B.2 Classifier Approach

These handwritten images were from LeCun and Cortes (2008). The set used here consists of 10,000 test images that are 24x24 pixels in size examples are shown in Figure B-1. All of the test cases were randomized (in the set they are in order) and a small subset of 100 used for each of the human classifiers. As with most real world human classification systems each person has a different level of ability and a different training set.

The image inputs to this system were translated to input variables through a 2D discrete wavelet transform. To avoid some of complexities with converting images to wavelets a simple Haar wavelet was used (Hubbard (1998)). All of the 784 resulting coefficients are used as variables for the discrimination. There are many more complex transforms that have been used on this data set with success as in LeCun et al. (1998). A more complex, and accurate, transform is unnecessary because the biggest effect on the accuracy is the ability of the human classifiers. The benefits of using a wavelet transform include the quick speed, tolerance for noise, and general applicability.

A classification tree is a method that consists of making hard divisions in a variable to

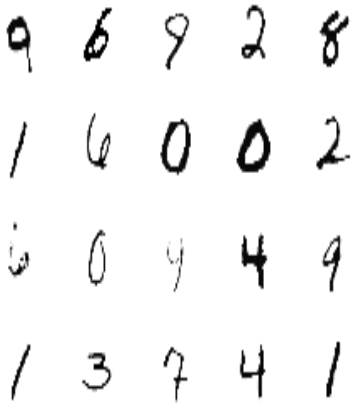

Figure B-1: Four Example Zip Codes. Five number images from the 10,000 possible images

maximize the purity of each final branch. For each variable a split will be made that creates a division where each of the branches is more similar, in this case has more similar zip code numbers are grouped together. This iterative process begins by trying every variable and then selecting the variable that makes the biggest improvement. After a selection is made then the process is repeated on each of the sub-trees. The process is stopped when each of the final decision nodes is of the same class (purity) or has too few cases.

Because there are 784 variables it is inefficient and inaccurate to build one large tree, so a large number of smaller trees were combined in a technique called bagging. Bagging has been discussed in numerous different areas such as Breiman (1996a) and Breiman (1996b) and Tibshirani (1996). The individual trees were pruned minimally to avoid singular nodes but, as suggested in the literature, full optimal pruning was not used.

Individual trees were built with a small number of random input variables chosen from the 784 available wavelet variables. Five Variables was selected as a good starting point and used throughout the selection process. By choosing less than one percent of the input values the cross-correlations would be minimized which is important considering the nature of the

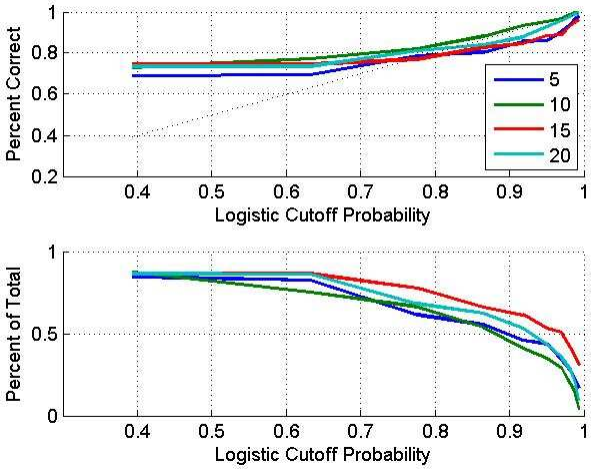

Figure B-2: Variables Per Tree. Given 5, 10, 15, and 20 variables for each tree the accuracy in percent correct is compared with the logistic probability in the top panel. The bottom panel shows the percent of data less than the logistic probability .

wavelet transform. For other situations this may have been too few. Figure B-2 shows the changes in accuracy as the number of variables changes (for a single human model) there is an increase until 15 variables per tree. With the unknown difference in the different humans and the fact that the number of trees will not be fixed, five was determined to be sufficient here although future investigations could search for a more optimal number of variables.

Each individual tree is created from a bootstrap sample equal to the original data size, 100 in this case. The number of individual trees was not fixed but determined based on an estimate of error. This estimate was a smoothed out-of-bag (OB) error as given by Breiman (1996b). The remaining data points that we not used in the bootstrap (≈37%) are fit using the classification tree, and added to a running tally for each image. The guess for any image is the mode of all of the guesses, or if there is a tie it is the most recent guess. The error for that run, r_b, is given by the sum of the errors for all of the images. With only a small number of training images this error may be quite erratic and so is smoothed. The function used to smooth is $e_b = p * e_{b-1} + (1 - p) * r_b$ where p is a variable, in this case 0.75.

Additional trees were added or 'bagged' as long as this error decreases.

In addition to providing a stopping condition the OB samples were also used to fit a logistic regression model. This is a new technique to estimate the confidence of that particular classification tree. The choice of logistic regression provides a probability that can be easily understood by the classifier in the final analysis. In many classification methods it is not straightforward to make accurate confidence estimates, k-NN, Naive Bayes, Neural Networks, and SVM all provide misleading numbers (Delany et al. (2005)). Because ensemble techniques (with the right functions) are unbiased in their limit, they can accurately estimate confidence to the prediction as shown by Breiman (1996b). In the tree methodology the margin parameter has been found to be an accurate and quick confidence estimate. The margin is the difference between the top vote receiving class and the next class. So after all of the trees vote in a particular classifier, the normalized difference between the top two is the margin. As compared with a range, standard deviation, median absolute difference or squared error, it has been found to be extremely effective and very easy to calculate.

Using this margin parameter from the OB samples a logistic regression model was fit to the error. With the small training sets, a minimum of five incorrect images were required to estimate the two logistic parameters, β_0 and β_1. To reach a better estimate of these parameters, they were based on 10-fold run over the number of trees. The logistic model confidence estimate had low discrimination against the training data as can be seen in Figure B-3 but, worked very well against the true values as can be seen in Figure B-2 and Figure B-4. The margin is able to differentiate good variable choices from guesses, or erroneous choices, accurately. There are two metrics to evaluate these confidence estimates. First, if the confidence estimate is 80%, then it should reflect that it is correct on 80% of the images. The second metric is the ability for the confidence estimate to accurately predict

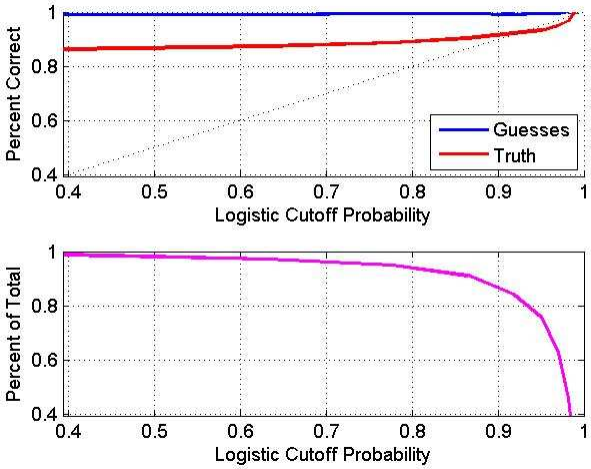

Figure B-3: Confidence Estimate on Training Data. The relationship between the error on the training data and the logistic probability is given in the top panel. The percentage of the data less than the logistic probability is given in the bottom panel.

the greatest percent of the population, the greater percentage of accurate values the better, and the fewer images that needs to be re-evaluated.

To demonstrate this property more clearly the entire data set of 10,000 numbers was passed through the trees for a particular classifier and the results are compared with the logistic confidence estimate. The accuracy of the probability estimate is within 2% until p=65%.

This procedure was run with three different individuals, and their results compared. Individually, each human classifier performed evaluations on a separate subset of the data, and, as indicated above, five random X's were chosen for a variant number of trees. Individually, the classifiers were all very similar performing at accuracies of 85.1%, 85.4%, and 94.3%. All evaluators are using the same input system and have similar distractions and time pressures. If a subset of data from the three different human classifiers is randomly combined and used to train a classifier, that classifier had the expected combined performance.

B.3 Combining Classifiers

Using the individual logistic confidence estimates, each human evaluator would be able reduce the number of evaluations necessary (at their same performance) by $p_r = 55.0\%$, 38.9%, and 36.2%. This can be seen in Figure B-4 at 85.1%, 85.4%, and 94.3% for Chad, Helen, and Jon respectively. Without decreasing performance, the individual could reduce their work load by this number of evaluations using their classifier but, because this is a group process some additional reductions can be made through a decision algorithm. First if all three automated classifiers are in agreement then those can be classified with very high probability. In a sample case of 1000 never seen before test images, we had 26% of the total in this category, at an accuracy of 96%. This high percentage of cases in agreement is due to the marginal probabilities near 91%. Given c classifiers the number falling into this first class is $p_1 * .91^c$. This is a higher percentage, and a higher marginal probability than initially expected but, can be explained by the fact that some images are easily classified, and agreed upon.

The second decision method to combine the classifications is through confidence based voting. Due to the fact that the human classifiers do not have equal performance probabilities, this voting is done sequentially and the best classifier gets the final vote. The probability of each individual classifier contributing is $p_c = p_r - p_m * p_r^{n-1}$, assuming identical reduction probabilities and n judges. If two other classifiers agree then the combined probability is calculated and may outweigh the other classifier. As the individual reduction probability, p_r, increases it reduces the group load but, if the individual reduction increases too much the group is not able to benefit from other members and thus the actually workload increases, as shown in Figure B-5. The range presented here is small because as the ability of each classifier changes it is expected that the marginal probability also changes.

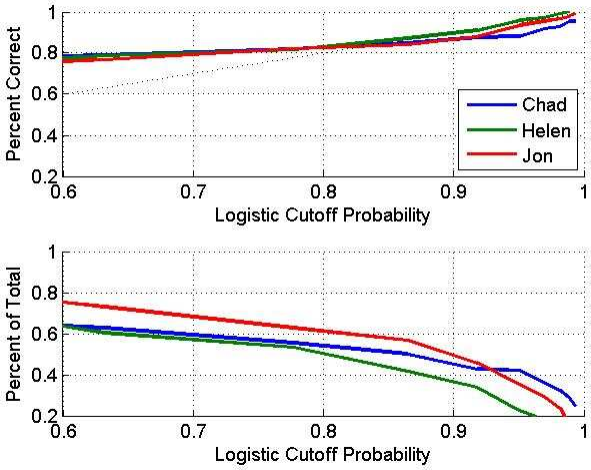

Figure B-4: Different Human Classifiers. The relationship between the logistic probability and the accuracy for all 10,000 images is given in the top panel for three different classifiers and their combined estimate. The bottom panel shows the percentage of the population less than the logistic probability

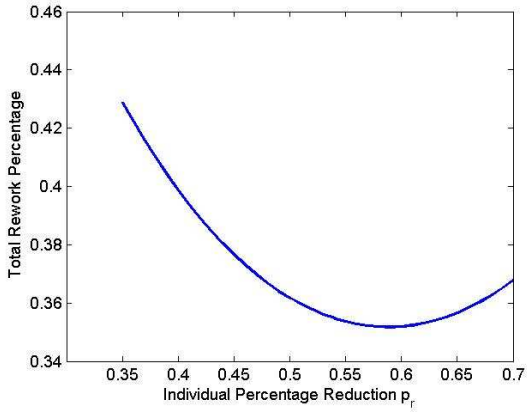

Figure B-5: Percentage Rework. This plot is based on a marginal probability for the logistic parameter of 0.85 and three judges. The individual percentage reduction p_r is on the horizontal axis and the percentage rework is on the vertical axis.

This combined classification system is explained as a weighted as a voting method. Each automated classifier has a class vote and a confidence. The decision is to go with the highest confidence, either in a single classifier, or if more than one agree, it could be the combination. For B classifiers, this combination can be expressed as:

$$\text{class} = \arg\max_{i \in class} \left(\sum P_B^i \right)$$

Using this equation, it is possible to scale up the classifiers very easily. Each individual classifier is developed to the nuances of their human trainer and only combined in a final group decision algorithm or 'meeting'. This parallels an effective human process, with more objective confidence measures.

After the group decision meeting, the 24% of the original images remain. The human classifiers have their workload reduced by 76%. The total workload for the three here, requires less than one of the original workers. These additional human resources could be used to re-evaluate some of the images and to improve the process.

To improve the overall process with these extra resources a statistical technique called boosting introduced by Shapire (1990) is used here. Generally, the concept is to run the points through an initial classifier, and then those points that are incorrectly identified are used to train an additional classifier with increased weight. This weighted training can extend many levels. There are a number of algorithms that have shown this can be more effective than general bagging approaches that are employed initially. The drawback of boosting is in this re-weighted training, it may suffer from over-fitting, or extreme weighting. This image recognition problem had high, and inconsistent, human image recognition error and over-fitting was deemed problematic. Each human based classifier was built using the more robust bagged classification technique, while combining these classification

trees was found to benefit from a boosting approach.

After all of the automated classifiers were complete and the decisions made, the humans completed the final classifications on the remaining 24% of the images. With three human classifiers it was possible to have each human read each image and create a classification. This additional data was then combined into another automated classifier. Thus this final classifier was trained with cases that had low-confidence in the other classifiers. It was also the first to use redundancy in the trainers to improve the quality of the training set.

This final classifier is considered as a boosted classifier that offers an exponential weight that is combined with the other base classifiers. This classifier is combined in a slightly different manner than the previous ones. Because it is exclusively trained on the errors of the other classifiers it has a greater weight to settle disputes. The added weight was $\alpha = (1 - err)/err$, this is the same weighting technique as the popular AdaBoost routine (Hastie et al. (2001)). In this case $\alpha \approx 1.2$, and is low mostly due to the few training runs after only one round image analysis. Future runs would be used to continue to advance the training of this classifier, and increase its weight.

Even in this early stage of improvement this classifier can be combined with the other three. This classifier is added first in the sequence, and because α is near one it has almost the same weight as the other classifiers. The classification accuracy remained near 81% although the percentage rework dropped from 24% to 20%. Future runs would continue to refine this classifier until the number of runs equaled the other classifiers. After this point the additional runs would be targeted at creating another classifier for improvement. This is aligned with the literature on the boosting methodology.

B.4 Conclusion

This work focused on automating a human classification process as demonstrated through U.S. Post Office zip-code data. The goals were to: improve classification consistency, assign confidence level for each automated classification, and have no increase in workload throughout the implementation. The method used multiple bagged classification trees, initially for the individual classifications and then applied to the combined group. The scope of the classifier is increased by the use of a margin based logistic regression confidence parameter. Individual tree confidence parameters accurately predicted the performance against the population and could be combined accurately.

The individual classifiers use bagged classification trees based on five random variables in a standard Haar wavelet transform of the images. Each of these human based classifiers is aggregated through a voting with confidence procedure to decide the classification. The accuracy was selected to be at 80% and the automated classifiers reduced the workload by 76%.

After the individual classifiers were complete, additional classifications were made on the remaining 24% of the images. These most difficult images had new classifications performed by all of the human classifiers. The results are used to build another bagged classification tree, this classification tree was combined in a weighted manner similar to a statistical boosting method. The results with this new method maintained the accuracy at 80% and reduced the workload by 78%

This procedure was clear and straight-forward to implement and the results reduced the workload greater than expected, the classification plus confidence concept was easy to explain to the human classifiers, as well as solidly founded in current statistical procedures. The use of a voting decision system mimicked the current human system and was found to

enhance the total understandability and effectiveness of the system.

Bibliography

Breiman, L. (1996a). Bagging predictors. *Machine Learning*, 26(4):123–140.

Breiman, L. (1996b). Out-of-bag estimation. Web, ftp://ftp.stat.berkeley.edu/pub/users/breiman/OOBestimation.ps.Z.

Delany, S. J., Cunningham, P., Doyle, D., and Zamolotskikh, A. (2005). Generating estimates of classification confidence for a case-based spam filter. In *Proceedings of the 6th International Conference on Case-based Reasoning (ICCBR)*.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer.

Hubbard, B. B. (1998). *The world according to wavelets*. A K Peters, Ltd.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE, 86(11):2278:2324*.

LeCun, Y. and Cortes, C. (2008). *The MNIST Database of handwritten digits*. http://yann.lecun.com/exdb/mnist/.

Shapire, R. E. (1990). The strength of weak learability. *Machine Learning*, 5(2):197–227.

Tibshirani, R. (1996). Bias, variance, and prediction error for classification rules. Technical Report, Statistics Department, University of Toronto.