A MINE ALTERNATIVE TO D-OPTIMAL DESIGNS FOR THE LINEAR MODEL

by

AMANDA MARIE BOUFFIER

(Under the Direction of Jonathan Arnold)

ABSTRACT

The Maximally Informative Next Experiment (MINE) criterion was developed for designing large, expensive genomics experiments. Four variations of the MINE method for the linear model were created: MINE-like, MINE, MINE with random orthonormal basis, and MINE with rotation for the linear model. Theorem 1 establishes sufficient conditions for the maximization of a MINE criterion under the linear model. Theorem 2 is established when the MINE criterion is equivalent to the classic design criterion of D-optimality. By simulation under the linear model, we establish that the MINE with random orthonormal basis and MINE with random rotation are faster to discover the true linear relation with p regression coefficients and n observations when p >> n. These two variations also display a lower false positive rate than MINE or MINE-like methods for a least a majority of the experiments.

INDEX WORDS: Linear Model, Maximally Informative Next Experiment (MINE), Model-Guided Discovery, D-optimality

A MINE ALTERNATIVE TO D-OPTIMAL DESIGNS FOR THE LINEAR MODEL

by

AMANDA MARIE BOUFFIER

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2013

A MINE ALTERNATIVE TO D-OPTIMAL DESIGNS FOR THE LINEAR MODEL

by

Amanda Marie Bouffier

Approved:

Major Professor: Jonathan Arnold

Committee: H. B. Schüttler

Thiab Taha

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2013

ACKNOWLEDGEMENTS

First and foremost I would like to acknowledge my advisor, Dr. Jonathan Arnold, for without his knowledge and endless encouragement this thesis would not have been possible. I extend my thanks to my committee members Dr. H. B. Schüttler and Dr. Thiab Taha for their advice and assistance. I would also like to thank my parents and friends for their support and as great source of information. This work was in part supported by NSF DBI-1062213.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Experimental Design methodology has been used in science, statistics, and engineering but its use is focused on model refinement and/or process control. Adaptive control figures predominantly in engineering. The classic work on experimental design focusing on model refinement is summarized by Fisher (1931) and more recently by John (1971) and on response surfaces by Box et al. (2005). The design problem for biological sciences here is distinct from that of classic experimental design in statistics or process control in engineering. This experiment focuses on the distinct goal of model-guided discovery.

A major problem was identified by Kitano (2003) in that model-guided discovery simulations in systems biology constructed from models before experimentation and are combined with the available data to help choose the next best experiment. These methods have been only recently introduced in the life sciences due to the complexity of the models involved and the sparse and noisy nature of genomics data (Dong et al., 2008). However, it is possible today to study these more complex design problems with the increasingly powerful computational resources, allowing for more and more complex models. In 2008, research by Dong and others focused on the discovery of information about the models through a new method called The Maximally Informative Next Experiment (MINE) in the context of nonlinear models (Dong et al., 2008). The MINE method was actually applied in practice to discover the connection between the

biological clock and ribosome biogenesis (Dong et al., 2008). This discovery was later confirmed by Jouffe et al. (2013).

In this thesis, we will use a more simplified situation than in the original MINE paper, namely the linear model, one of the oldest and best studied models in statistics. The focus of this research is two-fold: (1) to investigate a new criterion for discovery in experimental design, MINE, and its relationship to existing design methods and (2) to test the new criteria's validity in a practical sense. In order to examine the first, we prove two theorems. To study the second aspect, a number of variations on choosing the next set of observations are studied over 1000 replicates for each variation in computer simulations. These methods are then compared to determine their operating characteristics. We hope to find at least one method will show promise as a new model discovery method for the linear model.

CHAPTER 2

A MINE ALTERNATIVE TO D-OPTIMAL DESIGNS FOR THE LINEAR MODEL[1]

1. Amanda Bouffier, Bernd Schuttler, and Jonathan Arnold. To be submitted to PLoS Computational Biology.

Abstract

Doing large-scale genomics experiments can be expensive, and so experimenters want to get the most information out of each experiment. To this end the Maximally Informative Next Experiment (MINE) criterion for experimental design was developed. Here we explore this idea in a simplified context, the linear model. Four variations of the MINE method for the linear model were created: MINE-like, MINE, MINE with random orthonormal basis, and MINE with rotation. Each method varies in how it maximizes the MINE criterion. Theorem 1 establishes sufficient conditions for the maximization of the MINE criterion under the linear model. Theorem 2 is established when the MINE criterion is equivalent to the classic design criterion of D-optimality. By simulation under the linear model, we establish that the MINE with random orthonormal basis and MINE with random rotation are faster to discover the true linear relation with p regression coefficients and n observations when p >> n. Specifically in the simulations with 1000 replicates, we have n < 100 and p = 1000 and $\sigma = 0.01$. Lastly, these two variations also display a lower false positive rate than MINE-like method and for a majority of the experiments for the MINE method.

Introduction

The Problem: The researcher wishes to carry out model-guided discovery about a system from a sequence of n experiments. The challenge is that each of the n experiments performed is very expensive, and so at each stage (n+1) it is desirable to design the next experiment to be maximally informative. The approach in which n experiments are to be done sequentially in such a way as to capture the most

information at each stage n about the underlying model is called utilizing the Maximally Informative Next Experiment (MINE). To understand MINE we will consider the linear model, $Y = X\beta + \varepsilon$, where Y is a n x 1 vector of dependent measurements, X is a n x p matrix of p independent variables, each with n measurements, β is a p x 1 parameter vector, and ε is a n x 1 vector of independently and identically distributed normal N(0, σ^2) errors.

The problem has four features. First, there are many parameters and limited data (n << p) so there will be many more unknown parameters than data. In this setting a large sample of variables (p) is to be observed as it is not known in advance which ones are relevant. In fact, typically n ~ 100 while p ~ 10^3-10^6. Second, the X matrix is partitioned into two parts, X = (X',X"), where X' is an n x p' matrix of independent variables that the experimentalist can control and X" is an n x p" matrix of independent variables that cannot be controlled under the conditions p = p' + p" (Lopez-Fidealgo and Garcet-Rodriguez, 2003). The X' matrix will be referred to as the design matrix. For simplicity, we will assume the entire X matrix is made up of X'. Third, the next experiment encompasses stages n+1, …, n+d, where d is the dimension of the experiment. Experiments constitute batches of d observations. The fourth and final feature of the experiment is that each experiment of d observations is very costly, be it time, materials and/or subjects, or financially such as $250,000 per experiment (Dong et. al., 2008). So at each stage n in the overall study, there is a high premium on choosing the best next experiment. The problem is to discover with reasonably high probability the model β in as few steps (n) as possible. We call this a problem in model discovery because what we wish to know is what linear relation can be discovered from

the many variables (p) measured over the time course of the study. Again the number of variables measured is large because it is not known in advance which ones are relevant. The process of discovering the model is cyclical as shown in Figure 1.

This problem makes points of contact with several distinguished problems in statistics and engineering. There are problems in experimental design (Fisher, 1935; Kiefer, 1959; John, 1971; Federov, 1972; Box, Hunter, and Hunter, 2005) leading to model refinement, particularly for sequential designs (Tsay et al., 1976). There are problems in control, as addressed by, for example, with response surfaces (Box and Draper, 1998). The problem of model guided discovery, we will show, is distinct from all of these.

As an example problem for use in model-guided discovery, suppose the researcher wishes to understand human longevity (Poon et al., 2007). The researcher may examine the characteristics of US centenarians. Several thousand variables are measured including genetics, diet, and lifestyle on each centenarian because it is unknown which variable or variables have an effect on longevity. Some of these variables can be controlled, such as diet and lifestyle. Others, like the genes carried by the centenarian, cannot be controlled. In model refinement, the goal is to select a design, X' (diet and lifestyle), to reduce the error in the parameters β by consideration of, for example, $X'^{T}X$ (John, 1971) and its determinant (*i.e*, D-optimality). In process control with the aid of response surfaces, the goal might be to select a design X' to maximize the expected longevity $E(y_i)$ (where E() denotes expectation and y_i denotes the longevity of the ith individual in the study) by manipulating diet and lifestyle. In model-guided discovery the goal is simply to choose a design X'at each stage to

discover the factors that determine longevity with as few centenarians (n) as possible and using limited data, to discover potentially many important variables. Such an engineering approach to extending lifespan has been implemented in the nematode (Sagi and Kim, 2012).

Another example of this framework in systems biology can be seen in the description of genetic networks at a steady state or system in equilibrium (Gardner *et al*., 2003). In this setup a genetic network is approximated to first order by the following linear system:

[1] $$\frac{dx}{dt} = Ax - y$$

Here the column vector x describes the concentration of mRNAs of genes in a network and the y vector describes external perturbations. The A matrix captures the network relationships among the genes and $\frac{dx}{dt}$ is the derivative with respect to time.

The steady state is assumed so that the dynamical system reduces to:

[2] $$y = Ax$$

The problem is to infer the network A. An experiment entails measuring all mRNAs under a particular perturbation y, so several are tested. This setup reduces to a linear regression problem. Such design problems have been considered for nonlinear genetic network models as well (Federov, 1972; Dette and Biedermann, 2003; Dette and Strigul, 2003; Dong *et al.* 2008), but we will not focus on these here.

Model Estimation by the Ensemble Method

A standard approach to estimating the regression coefficients β is the least squares method. This approach reduces to solving the normal equations below for the least squares estimates of the parameters $\hat{\beta}$:

[3] $$X^T X\hat{\beta} = X^T Y$$

The challenge in our problem is that X^TX will not often be of full rank because of collinearity in the independent variables and because there are so few data points relative to the number of parameters (n << p). While the normal equations [3] could be solved by use of a generalized inverse, there are likely to be many solutions that are equally consistent with the data and not one best least squares estimate $\hat{\beta}$ of the parameters in the model. The key is to not find one estimate, but rather an ensemble of estimates consistent with the data Y.

To address this problem the likelihood is consulted at each stage:

[4] $$L(\beta|Y) = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n}[y_i - \sum_{j=1}^{p} X_{ij}\beta_j]^2/2\sigma^2}$$

Since there are so many independent variables with so little data, this surface resembles a golf course with its varied terrain of many hills and sandpits than a mountain or mountain range. However, we can reconstruct the entire likelihood function by Markov Chain Monte Carlo methods (MCMC). By integrating over a standardized L(β) with respect to β and using a particular prior distribution, we can make predictions about the behavior of the system even in the presence of such limited data (Battogtokh *et al*., 2002). So instead of finding one best parameter β to represent the system, instead we construct L(β|Y) or alternatively, the entire posterior distribution with a different prior distribution. These are special cases of the ensemble method, in which

some figure of merit is used to select a distribution of models fitting experimental data (Battogtokh *et al*., 2002), and as a special case the reconstruction of the standardized L(β|Y) is referred to as the ensemble. In this paper the standardized likelihood can be calculated exactly when the variance (σ^2) is known, the case to be used here, and with a Gaussian conjugate prior on β from [4] (Box and Tao, 1992):

[5] $$p(\beta|Y) \propto (\sigma^2)^{-\frac{p}{2}} e^{[-(\beta-\mu_n)^T(X^TX+\Lambda_0)(\beta-\mu_n)]/2\sigma^2}$$

The posterior mean μ_n is described as:

[6] $$\mu_n = (X^TX + \Lambda_0)^{-1}(X^TX\hat{\beta} + \Lambda_0\mu_0)$$

The least squares estimate $\hat{\beta}$ enters into the calculation of the posterior mean μ_n. In the program description section below we replace $X^TX\hat{\beta}$ with X^TY from [3].

Here the precision matrix which determines the prior distribution in [5] is Λ_0 and will be taken as b*I where b is a positive constant and I is the p x p identity matrix. We also refer to this equation [5] as an example of the ensemble p(β|Y). In the past we have used a uniform prior over a finite interval (Battogtokh *et al*., 2002), but the Gaussian prior insures that the integration can be done along all components of the parameter vector, β, over the whole parameter space and can approach that of a noninformative prior distribution by letting diagonal elements of this matrix become small.

Normally the moments of the ensemble would be calculated by MCMC methods (Battogtokh et al., 2002), but from [5] we can obtain the moments of β directly and for example the linear model with known error variance:

[6] $$E_\beta(\beta) = \mu_n$$

[7] $$Var(\beta) = \sigma^2(X^TX + \Lambda_0)^{-1} = C = B^{-1}$$

Only the posterior mean, $\mu_{n,}$ needs to be updated after each experiment to calculate the posterior distribution because the error variance, σ^2, is known.

Maximally Informative Next Experiment

At each stage n, we choose the next experiment X* by reference to the ensemble p(β|Y) in [5] to infer the unknown but true regression parameters β_0. The new design matrix consists of d rows, and after completion of the next experiment is used to augment X_n to X_{n+d} or $\tilde{X}$. The design of the new experiment is captured in X* and the augmented/updated design for all experiments in $\tilde{X}$. For each member of the ensemble, β, we make a prediction vector about the d outcomes $\hat{Y}^*$ of the next experiment, namely $\hat{Y}^* = X^*\beta$, where β is drawn from the ensemble of models in [5], $\tilde{Y}^*$ is the vector of d predictions for the next experiment, and X* is the design of the next experiment. We choose the new experiment X* such that we have maximum discrimination between the alternatives in the ensemble. If two random models of the ensemble should have correlated predicted responses $\hat{Y}^*$ for experiment X*, the choice of design would be poor as this would not reveal as much information as when two random members should have uncorrelated responses.

One MINE criterion for choice of X^* was developed by use of a microscope analogy (Dong *et al*., 2008). The object in the microscope is β_0. The image under the microscope $\hat{Y}^*$ is mapped onto the object β_0 in the field of the microscope, but the mapping is fuzzy and imperfect with their being uncertainty in $\hat{Y}^*$. Let v_0 be a volume in the object space (*i.e*., the parameter space R^p) under the light microscope where p is the number of parameters, and let v_Δ be the "image difference volume" swept out that is

viewed. For a microscope, the connection between the two volumes is the model of physical optics. In our context here, the connection is the model prediction $\hat{Y}^* = X^*\beta$. Formally, the image difference volume v_Δ is swept out by the image difference vector $\Delta\hat{Y}\ (\beta, \beta', X^*)$ for all pairs of objects (β, β') in (v_0, v_0) or

[8] $$v_\Delta(v_0, v_0', X^*) = \Delta\hat{Y}(v_0, v_0', X^*) = X^*(\beta - \beta')$$

The image difference volume depends explicitly on v_0 and how we "twiddle the dials" on the microscope though X^*. The maximally informative next experiment (MINE) criterion is based on this idea that the more volume in $v_\Delta(v_0, X^*)$, the more detail discerned in v_0. This is achieved by adjusting the data captured in X*.

In order to take advantage of this MINE criterion, we must make a selection of the object volume v_0. The choice is improvised but driven by computational practicality (Dong *et al*., 2008); other choices are possible (Marvel and Williams, 2012). We elect to define a "representative volume" v_Δ swept out by $\Delta\hat{Y}\ (\beta, \beta', X^*)$ when β and β' are drawn randomly as "typical" or average values from the ensemble pair distribution $p(\beta, \beta'|Y) = p(\beta|Y) \times p(\beta'|Y)$, the components given by [5]. The volume is constructed from the variance – covariance ellipsoid of the image difference volume $\Delta\hat{Y}\ (\beta, \beta', X^*)$ and is dependent on the choice of experiment X*. We define the ensemble distribution of $\Delta\hat{Y}\ (\beta, \beta', X^*)$ as:

[9] $$Q_\Delta(\Phi, X^*) := \int_\beta \int_{\beta'} \delta(\Phi - \Delta F(\beta, \beta', X^*))\, p(\beta|Y) \times p(\beta'|Y)$$

The quantity Φ is any point in the $\Delta\hat{Y}\,(\beta,\beta',X^*)$ volume and δ(…) is the Dirac Delta Function. The ensemble distribution in [9] specifies an effective difference volume $v_\Delta\,(v_0,X^*)$ in the image difference space $\Delta\hat{Y}\,(\beta,\beta',X^*)$ by way of the characteristic ellipsoid of this space specified by:

[10] $$D_{ik}(X^*) = COV\left(\hat{Y}_i^*,\hat{Y}_k^*\right) = \int_\Phi \frac{1}{2}[\Phi_i\Phi_k p_{\Delta\hat{Y}}(\Phi,X^*)]$$

The variance – covariance ellipsoid is centered at the origin, Φ= 0 because $p_{\Delta\hat{Y}}(\Phi,X^*)$ is an even function in Φ due to $\Delta\hat{Y}\,(\beta,-\beta',X^*) = -\Delta\hat{Y}\,(\beta,\beta',X^*)$. The variance – covariance ellipsoid has the D-matrix eigenvalues and directions of the half-axes given by the D-matrix eigenvectors. Combining [8], [9], and [10] we have the covariance ellipsoid in terms of the moments of the ensemble:

[11] $$COV\left(\hat{Y}_i^*,\hat{Y}_k^*\right) = E\left(\hat{Y}_i^*\hat{Y}_k^*\right) - E\left(\hat{Y}_i^*\right)E\left(\hat{Y}_k^*\right)$$

Normally these moments could be computed by MCMC methods (Battogtokh et al., 2002) but because of the explicit form in [5] we can evaluate [11] directly from [5] as:

[12] $$X^*(X^TX + \Lambda_0)^{-1}\sigma^2X^{*T} = X^*B^{-1}X^{*T} = X^*CX^{*T}$$

The matrix C is defined to be B^{-1}. A Hilbert Space (HS, *i.e.*, a complete inner product space) formalism is introduced to give a compact form to the MINE criterion. The HS of functions consist of functions defined on the model parameter space, {β: $\beta\,\varepsilon\,R^p$}, for which the covariance is the HS inner product. This inner product is formally defined as:

[13] $$(g|h) = E[g(.)h(.)] - E[g(.)]E[h(.)]$$

The components of the observation vector, $\hat{Y}_i^*$, are represented by:

[14] $$f_i(\beta) := f\left(\beta,\hat{Y}_i^*\right), \quad for\ i = 1,\dots d$$

We can write the covariances in terms of the inner product:

[15] $$D_{ik} = (f_i|f_k)$$

The ensemble standard deviation of the prediction $\hat{Y}_i^*$ is equivalent to the HS vector norm or length denoted by $||\hat{Y}_i^*||$. The norm is defined by $||g|| := (g|g)^{\frac{1}{2}}$.

If the predictions $\hat{Y}_1^*, \dots, \hat{Y}_d^*$ are linearly dependent, then the HS prism is defined by the predictions collapses to a lower dimensional one, and the determinant $\det(D)$vanishes. If the predictions are not linearly dependent, then predictions determine an HS prism whose volume is simply given by the product of their vector lengths, namely $\det(D) = \left(||\hat{Y}_1^*|| \dots ||\hat{Y}_n^*||\right)^2$. In general the predictions are correlated, and we have the Hadamard Inequality:

[16] $$\det(D) \leq \left(||\hat{Y}_1^*|| \dots ||\hat{Y}_n^*||\right)^2$$

The ratio $\frac{\det(d)}{(||\hat{Y}_1^*|| \dots ||\hat{Y}_n^*||)^2}$ can be thought of as a composite measure of the dependence of the predictions and is a function only of the HS angles between predictions.

We are now in a position to introduce a MINE criterion first by introducing the normalized predictions

[17] $$\hat{Z}_i^* = \frac{\hat{Y}_i^*}{\|\hat{Y}_i^*\|}, \; i = 1, \dots, n$$

The normalized covariance matrix or correlation matrix denoted by R is defined by:

[18] $$R_{ik}(X^*) = \left(\hat{Z}_i^* \middle| \hat{Z}_k^*\right) = \frac{D_{ik}(X^*)}{(\|\hat{Y}_i^*\| \, \|\hat{Y}_k^*\|)}$$

$$= E[\hat{Z}^*(.,\hat{Y}_i^*)\hat{Z}^*(.,\hat{Y}_k^*)] - E\left[\hat{Z}^*(.,\hat{Y}_i^*)\right]E\left[\hat{Z}^*(.,\hat{Y}_k^*)\right]$$

This is the correlation matrix among the predictions. We propose the following MINE design criterion V(X*):

[19] $$V(X^*) := \det\big(R(X^*)\big) = \frac{\det\big(D(X^*)\big)}{\left(\|\hat{Y}_1^*\| \ldots \|\hat{Y}_n^*\|\right)^2}$$

This criterion is the squared volume of a prism spanned by the normalized predictions $\hat{Z}_1^*, \ldots, \hat{Z}_n^*$. Such a criterion is advantageous when the predictions are almost but not actually/completely linearly dependent. This is a situation that has been encountered in practice (Dong *et al*., 2008). This MINE criterion from [12] only depends on the ensemble through its variance-covariance matrix and not its mean in [6]. The MINE criterion is also scale-free (Dong *et al*., 2008; Dette, 1997). It clearly differs from the usual model refinement criterion based on $X^{*T}X^*$ *or* X^TX.

In practice the MINE criterion will behave better than $X^{*T}X^*$ for n~100 and p~1000 because its calculation through inverting B is stabilized by Λ_0 in [12], as in Ridge Regression (Draper and Smith, 1966) and will potentially incorporate the data from prior experiments in [12] through the B matrix. Its form also lends itself to optimization for large problems as will be shown under Theorem 1 below and under simulation results later (p >> n).

Maximizing the MINE Criterion

Ideally we would have a necessary and sufficient condition for maximizing the MINE criterion. Here in Theroem 1 we only present a sufficient condition for maximizing the MINE because the necessary condition has not been found.

Theorem 1: If the rows X_i^* of the design matrix X^* are chosen to be $w_i C^{-½}$, where $w_i, \ldots, w_d$ is any orthonormal set, then the MINE criterion $\det(X^* C X^{*T})$ is maximized.

Proof: $V(X^*)$ is maximized when $V(X^*) = 1$. This occurs if and only if $\det(R) = 1$, which is only satisfied if and only if $\left(\hat{Y}_i^* \middle| \hat{Y}_k^*\right) = \delta_{ik}\left\|\hat{Y}_i^*\right\|^2$ from [19], where δ_{ik} is the Kroneker delta. From [13] the inner product can be used to represent the covariances as $D_{ik} = (\hat{Y}_i^* | \hat{Y}_k^*)$. The condition $\det(R) = 1$ is thus equivalent to $D_{ik} = \delta_{ik}\left\|\hat{Y}_i^*\right\|^2$ or equivalently $X_i^* C X_k^{*T} = \delta_{ik}\left\|\hat{Y}_i^*\right\|^2$ (The X_i^* denotes the ith column of X*). We now need to introduce two more equivalencies: $w_i = X_i^* C^{1/2}$ and $w_k^T = C^{1/2} X_k^{*T}$. The fact that any positive semi-definite symmetric matrix, such as C, has a square root gives us the liberty to create such a w_i. Since $X_i^* C X_k^{*T} = \delta_{ik}\left\|\hat{Y}_i^*\right\|^2$, this leads to $w_i w_k^T = \delta_{ik}\left\|\hat{Y}_i^*\right\|^2$, which implies any orthonormal basis can be used for $w_1, \ldots, w_d$.

In particular, the vectors $X_1^*, \ldots, X_d^*$ can be selected as the eigenvectors of C once standardized by $C^{-1/2}$. One efficient route for maximizing the MINE criterion is then simply to compute the eigenvectors and eigenvalues of C or equivalently, to maximize the parallelpiped whose volume is $\det(R)$ (Dong *et al*., 2008) and then to normalize them by C in $w_i = X_i^* C^{-1/2} = X_i^* B^{1/2}$. The choice of normalization still needs to be examined as a model-guided discovery tool. See simulation results for examination of three choices of different orthonormal bases, a (1) normalized eigenvector basis; (2) random basis; (3) normalized eigenvector basis with random rotation.

Model Refinement

A traditional approach to choosing the design X^* (in contrast to MINE) is to choose X* to maximize some simple function of the variance-covariance matrix of the parameter estimates, β, such as the determinant, to create a D-optimal design (Kiefer, 1959; Fedorov, 1971). Consider then the augmented design matrix $\tilde{X}$ which is not only a function of the current design X but includes the possible design of the new experiment X^*. This means:

[20] $$\tilde{X} = (X^T, X^{*T})^T = (X, X^*)$$

Under ordinary model refinement, we wish to minimize some simple function of the variance-covariance matrix of β, such as:

[21] $\det\left(\tilde{X}^T\tilde{X} + \Lambda_0\right)^{-1}$ or equivalently maximize $\det\left(\tilde{X}^T\tilde{X} + \Lambda_0\right)$

This can be written explicitly in terms of the new experiment with the identity:

[22] $$\tilde{X}^T\tilde{X} = X^TX + X^{*T}X^*$$

The model refinement criterion is then to maximize $\det(A)$ where:

[23] $$A = X^TX + \Lambda_0 + X^{*T}X^*$$

The derivative of det(A) with respect to each component of X^* can be computed from:

[24] $$\nabla \det(A) = tr(Adj(A))\nabla A$$

Here the ∇ is the gradient with respect to X^*, tr refers to the trace and Adj denotes the Adjoint. A necessary condition for maximizing the det(A) is for $\nabla \det(A) = 0$. This max determinant (max det) problem is closely related to solutions to an affine formulation of this max det problem, and the problem is most closely related to the *analytic centering problem* (Vandenberghe *et al*., 1998). These authors cast the search for D-optimality in design as a convex optimization problem with the max det problem

linear in X^* and with linear inequality constraints (Vandenberghe *et al*., 1998). The linearity in X^* is achieved by constructing X^* from a set of rows (or designs) that are known in advance. The optimization problem is then reduced to determining how often each row (design) is used. Here we do not know the rows in advance.

MINE can produce a D-optimal Design.

Kiefer and Wolfowitz (1960) established that D-optimal designs are equivalent to mini-max designs, which minimize the maximum of the expected loss associated with each possible design. It is natural to ask whether or not there is any such relation between D-optimal designs and MINE. While the model refinement procedure appears to start from an entirely different criterion than MINE, it is possible to establish a relation between these different kinds of optimal designs by imposing the same constraints on the respective optimization problems. When we do this, we can establish:

Theorem 2 (Equivalence Theorem of D-optimality and MINE): The MINE procedure in Theorem 1 is D-optimal in the sense that X^* maximizes det(A) subject to the constraint $X_j^* C X_i^{*T} = 1$ where $A = X^T X + \Lambda_0 + X^{*T} X^*$.

Proof: In order for MINE and a D-optimal solution to be directly comparable they need to be maximized subject to the same constraints on X^*. So we maximize det(A) subject to the following constraint from [12]:

[25] Max det(A) subject to $X_j^* C X_i^{*T} = 1$

The constraint insures the columns of X^* are an orthonormal basis.

We can introduce a related criterion $\tilde{G}(X^*)$ as:

$$\det(B + X^{*T}X^*) = \det(B^{½})\det(B^{½})\det(B^{-½}(B + X^{*T}X^*)B^{-½})$$

$$= \det(B)\det(I + B^{-½}X^{*T}X^*B^{-½})$$

$$= \det(I + W^TW) = \tilde{G}(X^*) \qquad \text{where} \quad W = X^*B^{-½}$$

Note that $\det(A) = \tilde{G}(X^*)/\det(B)$. So maximizing det(A) is the same as maximizing $\tilde{G}(X^*)$. The maximization problem in [25] is equivalent to:

[26] $$Max\ \tilde{G}(X^*) = \det(I + W^TW)\ subject\ to\ w_jw_i = 1\ for\ j = 1, \ldots d$$

The constant d is the number of observations in the new experiment. We can think of the original optimization problem as equivalent to determining the best set of normalized vectors w_j.

From [26], the constraints imply that $tr(W^TW) = d$ where d is the dimension of the next experiment (*i.e.*, the number of observations in the next experiment). We also have the trace being the sum of the eigenvalues of W^TW.

[27] $$Tr(W^TW) = \sum_{v=1}^{d}\lambda_v Tr(W^TW) = \sum_{v=1}^{d}\lambda_v = d$$

Constraint [27] implies a constraint on the eigenvalues in [27], but not the converse. To finish the proof we will first maximize $G(X^*)$ subject only to [27] reminiscent of Vandenberghe et al. (1998).

We will then show the solution of this max det problem can also be chosen to satisfy all of the constraints in [26].

[28] $$W^TW = \sum_{j=1}^{d} w_j^T w_j\ where\ d < p$$

Since d < p, we can choose at least p-d orthonormal vectors u_v such that:

[29] $$W^TWu_v = 0 \qquad v = p - d + 1, .., p$$

We choose these u_v vectors also to be orthogonal to w_1, …, w_d. We will call this subspace of the parameter space as the unexplored subspace. Note that while $rank(W^TW) \leq d$, but dimension(W^TW) = p (that is, p > d here). This implies that p-d eigenvalues are zero. (This implies that for p-d eigenvalues, say for v=p-d+1,…,p, are zero. As an example, if p = 1000 and d = 10, then the last 990 of the eigenvalues are zero. We have that:

[30] $$\lambda_v = 0 \; for \; v = p - d + 1, \dots, p.$$

These degenerate eigenvalues in [30] are associated with the unexplored subspace. This fact along with the determinant being the product of the eigenvalues implies from [26]:

[31] $$G(X^*) = \prod_{v=1}^{d}(1 + \lambda_v)$$

Now maximize $G(X^*)$ with respect to λ_1, …, λ_d only subject to constraint [27] using the method of Lagrange multipliers (with multiplier Φ). We find that:

[32] $$\lambda_1 = \lambda_2 = \cdots = \lambda_d = \frac{G(X^*)}{\Phi} - 1 \qquad \text{and} \qquad \sum_{v=1}^{d} \lambda_v = d$$

These two imply that $\lambda_v = 1$ for v = 1, …, d on the explored subspace of the parameter subspace.

The result of maximizing with respect to the eigenvalues is that W^TW is diagonalized with only the first d diagonal elements being 1. The maximum value of $G(X^*)$ is 2^d from [31].

From here, if we choose w_1, …, w_d to be an orthonormal set such that $w_j w_k^T = \delta_{jk}$, then we have $W^TWw_j = w_j$ for all j = 1, …, d. Thus w_j is an eigenvector of W^TW with eigenvalue $\lambda_j = 1$ for j = 1, …, d. All constraints in [26] are satisfied for the solution to the max det problem with [27].

Choice of prior distribution

The choice for the prior mean vector μ_0 is reasonably taken as zero since most of the independent variables are not expected to have an effect on the dependent variable y. The only question is the choice of b specifying the precision matrix in $B = (X^T X + \Lambda_0)$ where $\Lambda_0 = bI$ (and specifies the prior). Dumouchel and Jones (1994) provide one argument to select b with an idea to making the design robust to violations of linear model assumptions. We will suggest another approach.

Let $X^T X$ have eigenvalues λ_i with corresponding orthonormal eigenvectors u_i. Then we can write $X^T X$ and B as:

[33] $X^T X = \sum \lambda_i u_i u_i^T$

[34] $B = \sum (\lambda_i + b) u_i u_i^T$

[35] $X^T X u_i = \lambda_i$

[36] $u_i^T u_j = \delta_{ij}$

The eigenvalues of B are λ_i+b and have the same eigenvectors as $X^T X$. We can now introduce a new variable:

[37] $p_i := u_i^T \beta$

We can loosely think of the uncertainty or standard deviation of the β-vectors (across the ensemble) in the u_i direction as:

[38] $\sigma_i := \sigma(p_i) = \beta_i^{-½} = 1/(\lambda_i + b)^{½}$

In the absence of any experimental data ($\lambda_i = 0$), the uncertainty in the u_i direction should reduce to:

[39] $\sigma_i(prior) = 1/b^{½}$

We would expect the uncertainty without experimental constraints (of data) to exceed the uncertainties with data or that:

[40] $\sigma_i(prior) \gg \min(\sigma_i)$ or equivalently $b \ll \max(\lambda_i)$

The maximum eigenvalue of X^TX provides an upper bound on b. This one would be satisfied, for example, if b were chosen to be equivalent to the weight of one observation in X'X. The matrix X'X would quickly dominate. The next constraint is more stringent, and so it is not necessary to check that [40] is satisfied.

Another constraint on b arises from the requirement that the true regression coefficients not be too far out in the tails of the prior distribution; otherwise the data through X^TX will never find the true regression coefficients. We can think of the prior distribution as equivalent to a fishing-net. We want this net to be well cast to catch the fish.

Introduce β_{max} = max(|true β_k|) = max (|E (β_k)|) where the expectation is taken over the ensemble. Then the b-value should be chosen so that

[41] $\sigma_i(prior) \gg \beta_{max}$

This is equivalent to requiring:

[42] $b \ll 1/\beta_{max}^2$

So with the prior data, X^TX, and some idea of the magnitude of the regression coefficients, there are constraints on the prior as specified by the precision matrix and hence b. These constraints are satisfied in the simulations to follow. As an example, if the largest magnitude of a regression coefficient were 50, then b << $1/50^2$ or b << 0.0004. Since we set all variables and know the largest regression coefficient's

magnitude is 50, we set b = 0.0001. This is a tighter constraint than the first. We also do not want b to be too small to allow C in [12] to still be inverted as in ridge regression.

Program description for implementing MINE in the linear model

The program MINE to implement the above procedures is written in JAVA under version 1.6 and utilizes the version 5 of the Jama library (The Mathworks 1998). The program must be called with seven arguments: the input file name, the number of loops, a name for the output file, the alpha significance level, a starting seed, a run number for the seed, and an identifier for which method is to be used. There are four variants on the MINE method described below in this section and summarized in Figure 2. The input file must be a simple text file with the following information each separated with a blank line: real β_0 components each on a separate line, the p-value, the μ_0 or initial guess vector with each value on a separate line, and finally the prior's precision matrix where each value on a single row is separated by a space. However, this program was made before the introduction of the b variable ($\Lambda_0 = bI$). So this prior input is then multiplied by b= 0.0001. The number of loops argument determines how many experiments (with 10 observations per experiment) are calculated not including the pilot. For example, given 100 loops this program will generate 1010 observations. The rationale for both a starting seed and a run number is to have a random seed for each run but a retrievable number instead of simply using 1 seed through the total number of runs. This was also helpful for running the scripts calling this program.

Following the initialization of many variables (including but not limited to the matrices and vectors to store the X matrix, Y matrix and ε, the number of variables, p,

and the significance matrix) and the set up of data from the input file, the main part of the program begins. Since the first experiment has no data on which to base a design, this pilot experiment is randomly generated. The program is designed to handle a variety of data. If the number of variables is over 50, then the first experiment is selected to have 10 observations; otherwise, the first experiment has between five and nine observations. The number of pilot observations varies between 0 and 10 as determined by the number of variables. Each value in the pilot is created by generating a random number between 0 and 10 and then dividing it by the p value. The set is not then normalized nor orthonogonal. After the pilot experiment is generated, the random number generator is changed depending on which method is used in order to allow for the same pilots but the rest of the numbers generated being different. From here the first set of the dependent Y vector is calculated along with the associated error vector, ε. With the initial data generated the real calculations can begin.

Each loop first consists of calculating the posterior mean [6] and then calculating the significance of the regression parameters in β with the cycle exiting here if the number of loops reaches the target number of experiments. The mean or μ_n is calculated by [6]. Although in this simulation we have the true values β_0, we use $X^T Y$ in [6] rather than using the real value or the previous mean. The significance subroutine takes the most recently calculated mean μ_n and the whole X matrix. It solves for the posterior variance-covariance matrix of β with [7] with the X matrix and multiplies this by σ^2. The z-value of each β is calculated with each individual mean value divided by the square root of the diagonal of the above matrix. The p-value is then calculated using this z-value. These p-values are then sorted from largest to smallest. The resulting

values are checked using a Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to decide which components of β are significant. The calculation for p-value is done using the algorithm from Press et al. (1992, p. 221). From here, depending on the specific method in Figure 2, the new observations are calculated and finally the new Y vector is calculated by $Y = X^*\beta + \varepsilon$ and the cycle repeats in Figure 1. The error vector (ε) or error values are created by generating a standard normal random variable value and multiplying it by σ (which was set by the input file).

MINE-like method. The simple naïve *MINE-like method* takes the ten eigenvectors of the C matrix in [7] associated with the ten largest C-eigenvalues using a common subroutine to generate the C matrix and simply uses these eigenvectors to generate the next X*.

MINE Method. The *MINE method* simply uses the eigenvectors associated with the C matrix as above with the ten largest eigenvalues and multiplies the corresponding eigenvectors by the square root of the B matrix to obtain the next X*.

MINE with Random Orthonormal Basis. In MINE with a random orthonormal basis a set of ten random orthonormal vectors is generated and then standardized by the square root of the B matrix ($B^{1/2}$). First, the ten vectors are created from using the random orthonormal set subroutine. Then each individual vector is multiplied by the square root of the B matrix to obtain the next X*.

MINE with Random Rotation. The MINE with a random rotation method first finds all the eigenvectors in the C matrix as above but selects the set of all degenerate vectors instead of simply the ten largest. Then the method creates a random orthonormal array of QxQ where Q is the number of degenerate eigenvectors to multiply the degenerate matrix with. This is used to rotate the degenerate eigenvector set. The first ten are then multiplied by the square root of B matrix and used to obtain the next X*.

The randomly generated orthonormal set used in both the MINE with a random orthonormal basis and MINE with random rotation is done by first generating a single vector of random Gaussian values. The vector is then normalized to a unit vector. This vector is used as the basis for generating more vectors generated in the same method and is made orthogonal using a modified Gram-Schmidt (MGS) algorithm (Bjork and Paige, 1992).

To obtain the square root of the B matrix, first C is calculated. Then the square root of C is solved by using the Singular Value Decomposition, $V^T DV$, where the V matrix here is the eigenvectors in column form and the D matrix is a diagonal matrix with the square root of the eigenvalues on the diagonals. This is then inverted by the method in the Jama package (The Mathworks, 1998) to get the square root of B matrix.

The output file contains first how long it took to complete the run after listing the number of variables. The file then lists the entire X matrix of each observation as a single row. Then in a single line, the Y matrix is listed next, followed by all posterior mean values μ_n starting with the input prior mean with each calculation being a single

line. The next matrix is a binary 1 and 0 scoring the mean value as significant with a 1 and not significant as a 0, with each position exactly corresponding to the posterior mean matrix. The next line lists the estimates of the ε vector components for each of the Y vector's values. The rest of the output file is used for checking purposes and includes the P-values for each significance test, the variable considered the most significant, the posterior variance of each β component, the diagonals for the X^TX matrices, and the X^TX eigenvalues.

Simulation Results

There are four variations on the MINE procedure examined here and defined in the previous section. The similarities and differences in the pathways of the four methods are summarized in Figure 2. All four methods employ the same subroutines for the majority of their implementation but differ in the way each particular method chooses the next experiment or set of observations to use, as described in the program description above.

The first method is called the naïve MINE or MINE-like method because this version does not incorporate the B matrix required in Theorem 1. This simpler method only calculates the eigenvectors of C and uses the eigenvectors with the ten largest eigenvalues (according to the algorithm) to define the next experiment X*. The second method, MINE, takes the MINE-like method and simply multiplies the chosen eigenvectors by $B^{½}$. The comparison of the MINE and MINE-like method allows us to assay the importance of the standardization in the MINE procedure.

The third method, called MINE with random orthonormal basis, does not use calculated eigenvectors as suggested Theorem 1 from, for example, matrix C. Instead, the third method creates a set of random orthonormal vectors and standardizes this basis by $B^{1/2}$. The final method tested is called MINE with random rotation. This method combines the previous methods by taking the chosen set derived in the MINE method but rotates the set by a random orthonormal basis before standardizing using a modified Gram-Schmidt Algorithm (Bjorck and Paige, 1998).

With each of the four methods the same set of 1000 components of the true β_0 were used. This β_0 only had ten components that were nonzero. The order of the nonzero components was not changed in the list of 1000 components. These were the first ten values of β_0 and were as follows: 11, -36, -26, 9, 33, -50, -45, 15, 3, and 17. The program was run with 1000 replicates for each of the four methods. Each replicate had a unique pilot experiment (consisting of ten observations) but these pilot experiments were the same for each method, allowing for a stronger comparison of the four methods. Each individual run had a unique random seed so that the rest of the replicate run would be unique. The error σ in the linear model used was 0.01, and all of the mean values (μ_0) were initialized to zero.

There were a number of criteria considered for comparing the MINE methods. These criteria include (1) identifying the nonzero values of β_0 by using a Benjamini-Hochberg multiple test correction (Benjamini and Hochberg, 1995) at a 1% significance level, (2) identifying the correct sign and value for the nonzero components of β_0, and (3) determining the number of false positives.

The first criterion discerns if the methods correctly identify the nonzero values of β_0 as being significant or successful discovery as given by the test described above. A method is considered better or more successful the fewer experiments completed. There are four figures provided (3-6) to display this criterion and all iterations are shown.

The MINE-like method seems to perform the poorest in this criterion (Figure 3) in that successful discovery was very late. However, being satisfied with a lower percentage of correctly included independent variables (say 7 out of 10) allows for more replicates meeting this criterion. The method began mostly (over 50% of the replicates) identifying 7 of 10 at the 87^{th} experiment and only at the 90^{th} experiment did over 90% of the replicates identify 7 of 10 of the true components of β_0. For over 90% of the replicates to identify all nonzero values of β_0 required 97 experiments.

The other three methods performed significantly better than the MINE-like. The MINE method performed almost twice as fast in this criterion as the MINE-like (Figure 4). For example, over 50% identified 7 of 10 at the 45^{th} experiment and over 90% at experiment 50. Only 63 experiments were required for over 90% of the replicates to identify 9 of 10 experiments. However, to get all nonzero values of β_0 required much more time. Sixty-six experiments were needed to get over 50% and 83 experiments before over 90% of the replicates considered significant.

Both the MINE with random orthonormal basis and the MINE with rotation performed almost identically (Figures 5 and 6). Both identify 90% of the nonzero values of β_0 in over 900 replicates at the 47^{th} experiment. However, attempting to identify all ten nonzero components of β_0 in all samples requires much more data, similar to the MINE

method. It takes more than 80 experiments for both of these methods to identify all nonzero values of β_0 in over 800 of the replicates.

The second criterion involved identifying the correct sign and value for the nonzero beta values. This is evaluated by methods described earlier. Figures 7-10 depict an average value for the first 20 values where the first ten are the nonzero and the second are zero and are shown as a comparison. As previously discussed, early detection is important.

As in the previous criteria the MINE-like method performed poorly (Figure 7). After the pilot experiment the nonzero β_0 values have the correct sign identified and never change sign though the full run of experiments; however, the experimental β_0 values do not reflect the components of the true vector β_0 until the final experiment. Also interestingly the values increase slowly until about experiment 85 when the absolute value of each β component drastically increases towards the true value.

MINE performs similarly in pattern to the MINE-like method. MINE does just as well at sign detection, with no real variable ever offering the incorrect sign (Figure 8). The pattern of the MINE is less gradual than the MINE-like but features a slow growth then a sudden spike and approaches the asymptote of the true value. The MINE reaches the slope change between experiments 50-55 and so it takes the remaining 45-50 experiments to arrive at the plateau of the real value.

The other two methods also outperform the MINE-like method. Again, in this criterion the MINE with random orthonormal basis and the MINE with rotation perform almost identically (Figures 9 and 10). Sign identification seems to be the easiest

criterion as these two also perform flawlessly here. Unlike the previous two, these methods seem to have a very linear pattern in the values of the β_0 nonzero components.

The next criterion involves comparing the false positives of each method (Figure 11). This observes if any of the β_0 that are actually zero are considered significant. For comparison, the average number of incorrectly identified values, averaged over all simulations for each method, is shown.

Again, we see similar patterns where the MINE-like performs poorest. Initially, it looks like it is performing adequately, since up until experiment 60 there are zero false positives. Since we have to wait until experiment 85 for any reasonable amount of success, we find that the false positives are beyond 400 and at some of the highest peaks compared to the other three methods. The only way this method could possibly be considered better is that it drops off faster at the final observed experiment but since this point is a worst case, this point should not be reached.

The MINE method has similar pattern to the MINE-like, again performing in a similar scaled manor. At the point where information could be accepted for the nonzero values, around experiment 55, the false positive rate is around 300 which would allow for about 70% of the variables to be eliminated. If more experiments are performed, the number identified peaks just under 460 during experiments 73-84 afterwards it gradually begins dropping.

The MINE with random orthonormal basis and MINE with random rotation again display almost identical results. However, in contrast to the MINE-like and MINE methods, these two have a more linear growth of false positives, especially after the first 15 and before the last 15 experiments. Due to the data being displayed on a single

graph, the similarities are more observable with the two lines eclipsing each other. During the most optimal selection periods between experiments 20-45, the false positives do not go over 225. This allows for a greater than 75% reduction in variables. These two do however peak ever so slightly higher with the average reaching just under 465 but a much later experiment and are only lower during a 20-30 experiment window.

Discussion

In 2008 a key problem in systems biology was solved as identified by Kitano (2002) with a new methodology called MINE (Dong *et al*., 2008). The MINE methodology is used to integrate several cycles of modeling and experiments to yield discoveries about the underlying process being studied. The result of the application of the MINE methodology was new insights into the relation of the clock to ribosome biogenesis (Dong *et al*., 2008, Jouffe *et al*., 2013). This new approach to model-guided discovery has sparked a flurry of developments in MINE methodology (Donahue *et al*., 2010; Marvel and Williams 2012; Liepe *et al*., 2013). It is natural to ask how this new experimental design methodology of MINE is related to classical experimental design criteria and whether or not we can validate MINE mathematically as a discovery tool when there are many parameters and sparse, noisy data ($p >> n$). A natural place to validate this new MINE tool is in the framework of the oldest and mostly widely used statistical model, the linear model.

One of the consequences of the work here is to establish another view of one MINE procedure. When the same constraints are imposed on MINE and the D-optimality criterion, then the MINE procedure discussed here is D-optimal under the

linear model. The effect of minimizing the determinant of the correlation matrix of the predictions is equivalent to minimizing the determinant of the variance-covariance matrix of the parameter estimates as described in detail in the Equivalence Theorem 2. We suspected this would be the case from the application of the MINE procedure in systems biology, where the application of the MINE procedure appeared to decrease the estimated error variance σ^2 over time (Dong *et al*., 2008). In the language of the microscopy analogy, maximizing the volume observed under the microscope by choice of experiment is equivalent to reducing the ellipsoid of variation in the optical field of the parameter space. It is this key relation that Marvel and Williams exploit to address Kitano's problem (Marvel and Williams 2012).

Having shown the MINE procedure in practice is useful for discovery (Dong *et al*., 2008), it is natural to ask how MINE performs in a simpler setting of the linear model. We explored its performance in four methods. In this simpler setting, where we can actually calculate the ensemble directly without resorting to using Markov Chain Monte Carlo as used in nonlinear systems (Yu et al., 2007), we can solve the associated optimization problem of MINE in Theorem 1 in a way that may suggest new approaches to MINE in nonlinear models. The result of Theorem 1 was the realization that the maximization of the MINE criterion here is defined up to an orthonormal basis of the data space. There are a variety of different bases that could be selected. Theorem 1 also calls for a standardization of the basis. This standardization does prove important as we see upwards of a 50% improvement in some criteria between the MINE-like and the MINE methods.

First, it was important to see the two more similar methods (MINE with random orthonormal basis and MINE with random rotation) performed incredibly similarly. Second, these two proved better in all of the criteria at almost all experiments. These allowed for the earliest detection, during which provided the closest to actual values on all variables, and provided the fewest false positives for a larger sample reduction. The only area at which these two methods were out performed was in the number of experiments needed for most of the simulations to identify the real values given that initial detection had begun. The MINE method once 10% of the simulations began detecting these values was able to reach 90% more quickly. Though this region was smaller for the MINE method, the other two were not only able to reach or arrive at the 10% quicker but generally complete (get over 90%) quicker.

A third consequence of this work is to open up a new convex programming problem that is closely tied to the max det problem so thoroughly analyzed by Boyd and co-workers (Vandenberghe *et al.*, 1998). The argument here in the max det problem is quadratic in the design parameters with linear inequality constraints potentially as opposed to an affine argument. An open question is whether or not this new problem is a convex programming problem. If so, then much of the machinery developed by Boyd and coworkers could be developed for the problem here. We have illustrated the use of the convex programming procedure in our discussions in this work.

In conclusion, we feel that the MINE discovery tool has opened up many exciting design problems that will transform the way scientists now integrate theory and experiment in a number of areas beyond systems biology (Townsend and Lopez-Giraldez, 2010; Townsend and Leuenberger, 2011).

Chapter 3

Conclusion

The Maximally Informative Next Experiment (MINE) was first proposed and used in 2008 (Dong *et al*., 2008). The idea of using model driven discovery is not new, but it is still being explored in new environments and subjects. With the advances in genetics and life sciences and the improvements in computational power, new design problems can be explored using the MINE criterion. Here we establish a basis under the most simplified setting for this new criterion by both exploring (1) how it relates to previously studied model design methods, such as D-optimality and (2) showing its performance. For example, we establish an equivalence theorem between MINE as a discovery tool and the model refinement criterion of D-optimality. Here we provide the foundation in the linear model where we can do more explicit calculations, as opposed to resorting to Markov Chain Monte Carlo Methods needed in the original setting of nonlinear models in which MINE method was proposed (Dong et al., 2008). Therefore, we can now begin to explore more variations on the MINE method for nonlinear models. The subject of experimental design is also expanded based on the linear model explored here to include the design question of model-guided discovery. With these few variants we were able to show two of the four were mostly better than the others in this limited situation. However, not all options of in the explored variants were studied. Some of these include (1) varying the ordering of the nonzero regression coefficients in the vector β_0 of true regression coefficients as well as their magnitude and magnitude of range, i.e. including

both larger and smaller components of β_0, (2) varying any number of other variables, (3) the error variance σ^2 or (4) the prior distribution parameter b in [42] or (4) even looking at the effect of the error terms ε. Studying these additional options will help fully determine the extent of the MINE model in the linear setting as a viable discovery tool and help set a base for use in non-linear settings under varying parameters.

Chapter 4

References

Battogtokh, D, DK Asch, ME Case, J Arnold, & HB Schüttler 2002. An ensemble method for identifying regulatory circuits with special reference to the *qa* gene cluster of *Neurospora crassa*, *PNAS USA* **99**: 16904-16909

Benjamini Y, & Y Hochberg 1995. Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Roy Stat. Soc. Ser. B* **57**: 289-300.

Bjorck, A & CC Paige 1992. Loss and recapture of orthogonality in the modified Gram Schmidt algorithm. *SIAM J. Matrix Anal. Appl*. **13**: 176-190

Box, EP & NR Draper 1998. *Evolutionary Operation*. Wiley, NY

Box, EP, JS Hunter, & WG Hunter 2005. *Statistics for Experimenters*. Wiley, NY

Box, EP & GC Tiao 1992. *Bayesian Inference in Statistical Analysis*. Wiley, NY

Dong, W, X Tang, Y Yu, R Nilsen, J Griffith, J Arnold & H-B Schüttler 2008. Systems biology of the clock in *Neurospora crassa*. *PLoS ONE* **3**(8): e3105 (28 pages). doi:10.1371/journal.pone.0003105*

Dette, H 1997. Designing experiments with respect to standardized optimality criteria. *J. Roy. Stat. Soc B* **59**: 97-110

Dette, H & S. Biedermann 2003. Robust and efficient design for the Michaelis-Menten Model. *J. Am. Stat. Assoc.* **98**: 679-686

Dette, H & N Strigul 2003. Efficient design of experiments in the Monod model. *J. Roy. Stat. Soc. B* **65**: 725-742

Donahue, MM, GT Buzzard & AE Rundell 2010. Experiment design through dynamical characterization of non-linear systems biology models through sparse grids. *IET Systems Biology* **4**: 239-262

Draper, N & H Smith 1966. *Applied Regression Analysis*, 2nd Edition, Wiley, NY, NY

DuMouchel, W & B Jones 1994. A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics* **36**: 37-47

Fedorov, VV 1971. *Theory of Optimal Experiments*. Academic Press, NY, NY

Fisher, RA 1935. *The Design of Experiments*. Oliver and Boyd, London

Gardner, TS, D Di Bernardo, D Lorenz, & JJ Collins 2003. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**: 102-105

Goldberg, MA 1972. The derivative of a determinant. *Am. Math Monthly* **79**: 1124-1125

John, PWM 1971. *Statistical Design and Analysis of Experiments*. MacMillan, NY

Jouffe, C, G Cretenet, L Symul, E Martin, F Atger, F Naef, & F Gachon 2013. The circadian clock coordinates ribosome biogenesis. *PLoS Biology* **11**, e1001455

Kiefer, J 1959. Optimum Experimental Designs. *J. Roy. Stat. Soc B* **21**: 272-319

Kiefer, J & J Wolfowitz 1960. The equivalence of two extremum problems. Canad. J. Math 12: 363-365

Kitano, H 2002. Systems biology: a brief overview. *Science* **295**: 1662-1664

Liepe, J, S Filippi, M Komorowski, & MPM Stumpf 2013. Maximizing the information content of experiments in systems biology. *PLoS Computational Biology* **9**: e1002888

Lopez-Fidalgo, J & SA Garcet-Rodriguez 2004. Optimal experimental designs when some independent variables are not subject to control. *J. Am. Stat. Assoc.* **99**: 1190-1199

Marvel, SW & CM Williams 2012. Set membership experimental design for biological systems. *BMC Systems Biology* **6**: 21

Poon, LW, M Jazwinski, RC Green, JL Woodard, P Martin, WL Rodgers, MA Johnson, D Hausman, J Arnold, J, A Davey, MA Batzer, WR Markesbery, M Gearing, IC Siegler, S Reynolds, and J Dai (2007). Methodological considerations in studying Centenarians: lessons learned from the Georgia Centenarian Studies. In *Annual Reviews of Gerontology and Geriatrics: Biopsychosocial Approaches to Longevity, Vol. 27.* LW Poon and TT Perls (ed.s), Springer-Verlag, NY, NY. pp. 231-264

Press, WH, SA Teukolsky, WT Vetterling, & BP Flannery 1992. *Numerical Recipes in C*, 2nd Edition, Cambridge University Press, NY, NY

Sagi, S & SK Kim 2012. An engineering approach to extending lifespan in C. elegans. *PLoS Genetics* **8**(6) e1002780

The MathWorks, Inc. and the National Institute of Standards and Technology *JAMA: A Java Matrix Package* [Java reference library] version 1.0.2 August 1998 Retrieved from: http://math.nist.gov/javanumerics/jama/

Townsend, JP &C Leuenberger 2011. Taxon sampling and the optimal rates of evolution for phylogenetic inference. *Systematic Biology* **60**: 358-365

Townsend, JP &F Lopez-Giraldez 2010. Optimal selection of gene and ingroup taxon sampling for resolving relationships. *Systematic Biology* **59**: 446-457

Tsay, J-Y (1976). On the sequential construction of D-optimal designs. *J. Am. Stat. Assoc.* **71**: 671-674

Vandenberghe, L, S Boyd, & S-P Wu 1998. Determinant maximization with linear matrix inequality constraints. *SIAM J. Matrix Analysis and Applications* **19**: 499-53

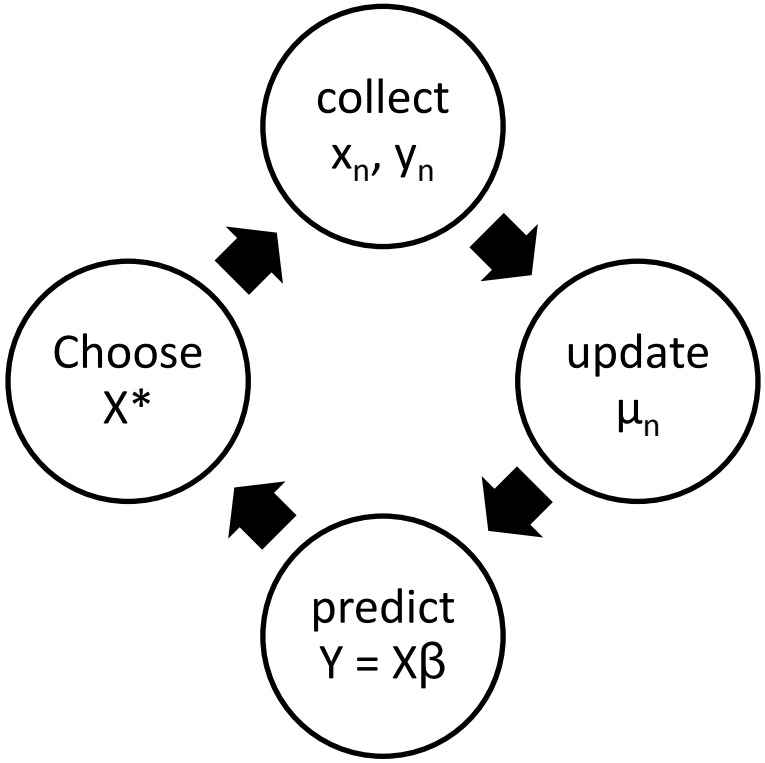

Figure 1: Cycle of MINE discovery – Simplified Computing Life Paradigm

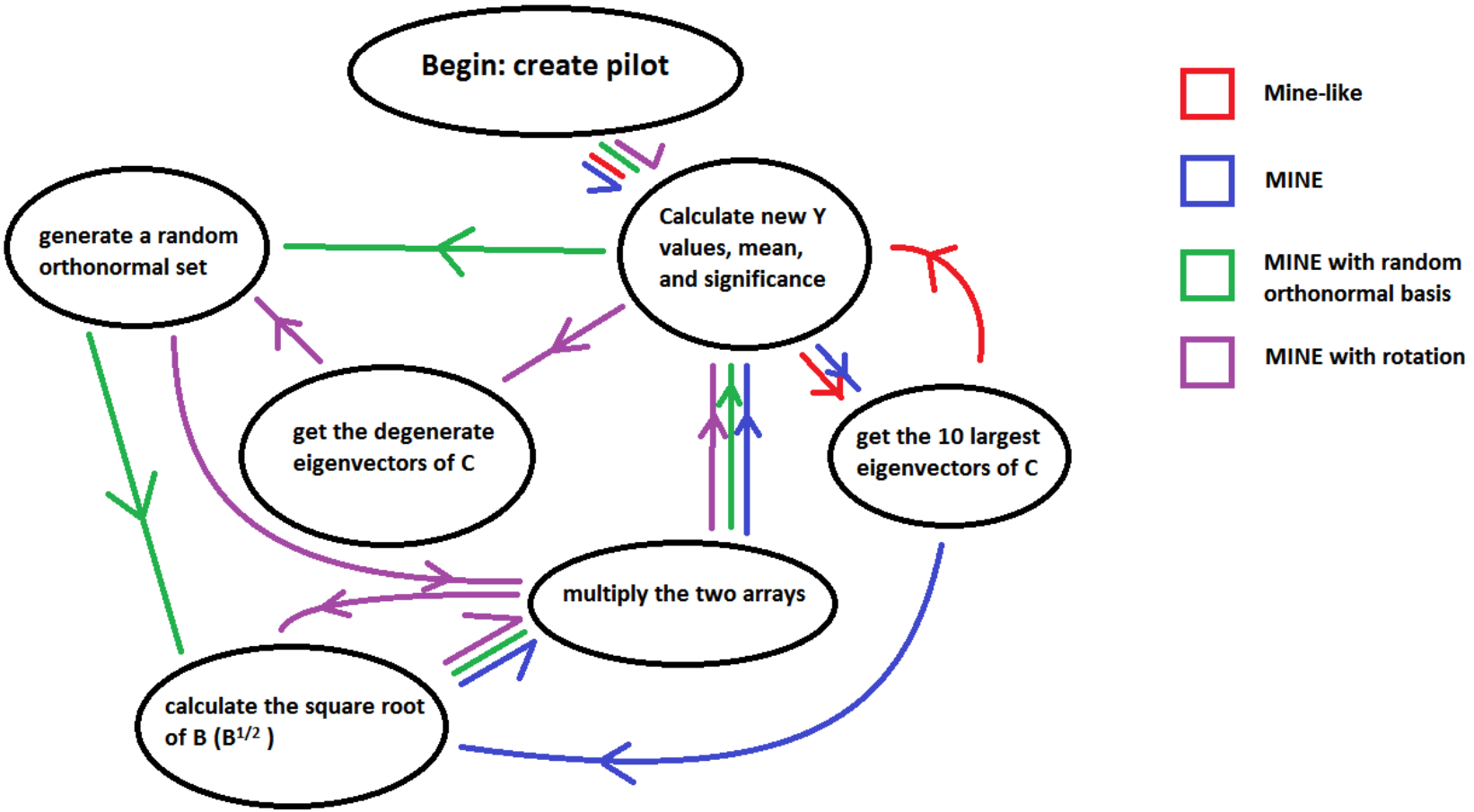

Figure 2: Visual representation of the pathways for each MINE method.

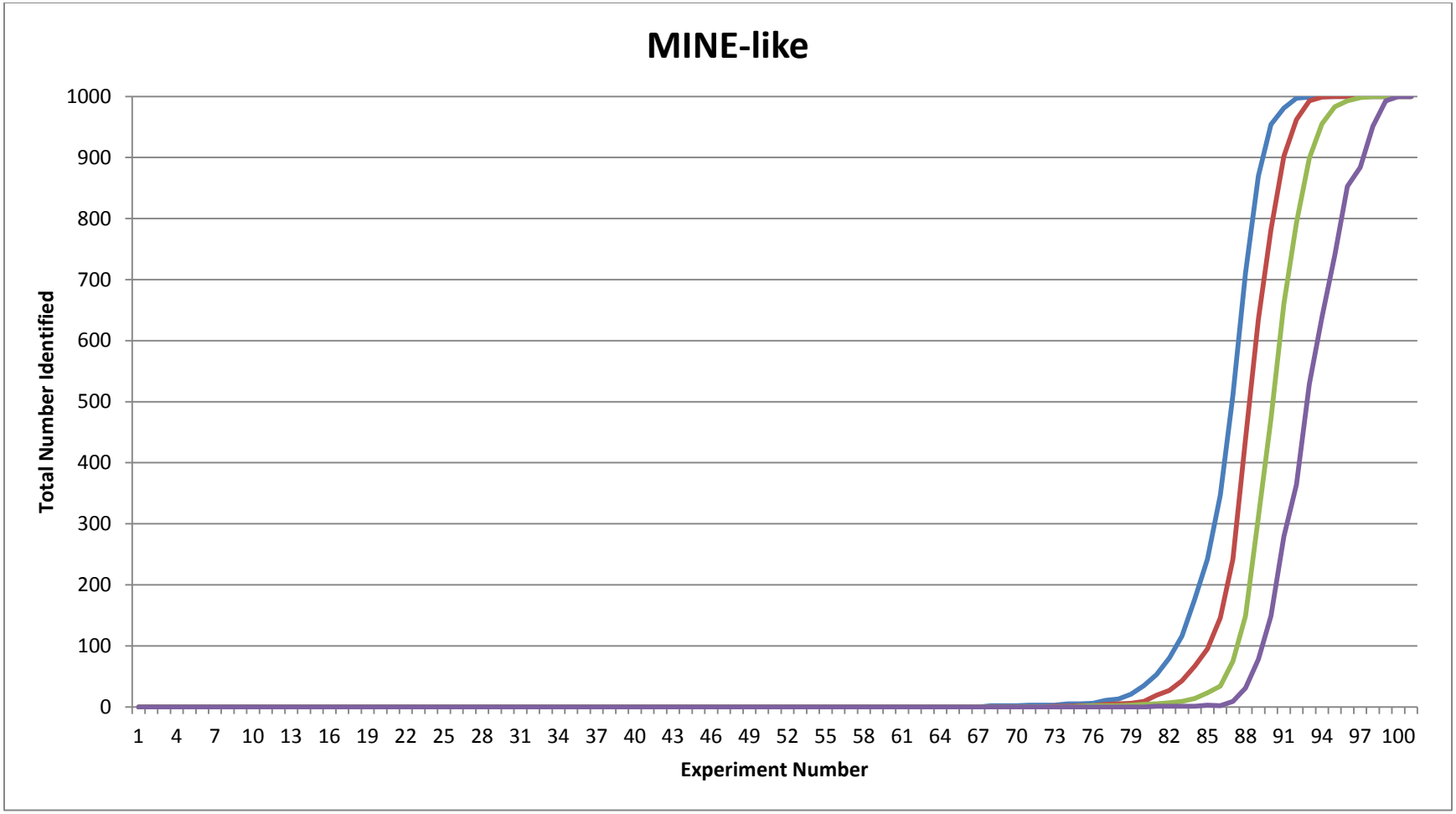

Figure 3: Graph of significant nonzero regression coefficients for the MINE-like method. The graph identifies the number of replicates (y-axis) with a varying percentage of the nonzero β_0 components as significant as a function of the number of experiments (x-axis) for the MINE-like method. Blue corresponds to 70% correctly identified, red to 80%, green to 90% and purple to 100%.

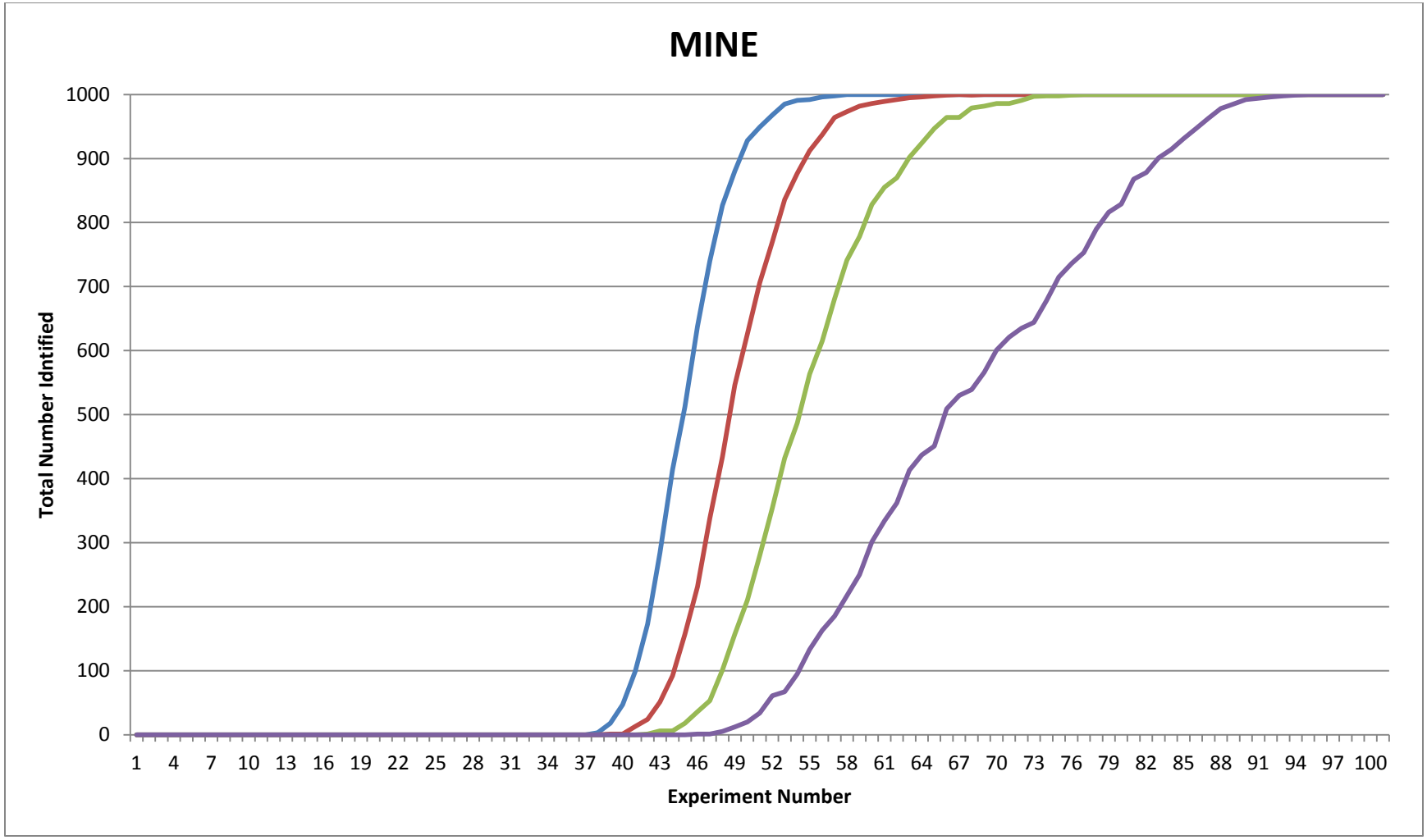

Figure 4: Graph of significant nonzero regression coefficients for the MINE method. The graph identifies a number of replicates (y-axis) with a varying percentage of the nonzero β_0 components as significant as a function of the number of experiments (x-axis) for the MINE method. Blue corresponds to 70% correctly identified, red to 80%, green to 90% and purple to 100%.

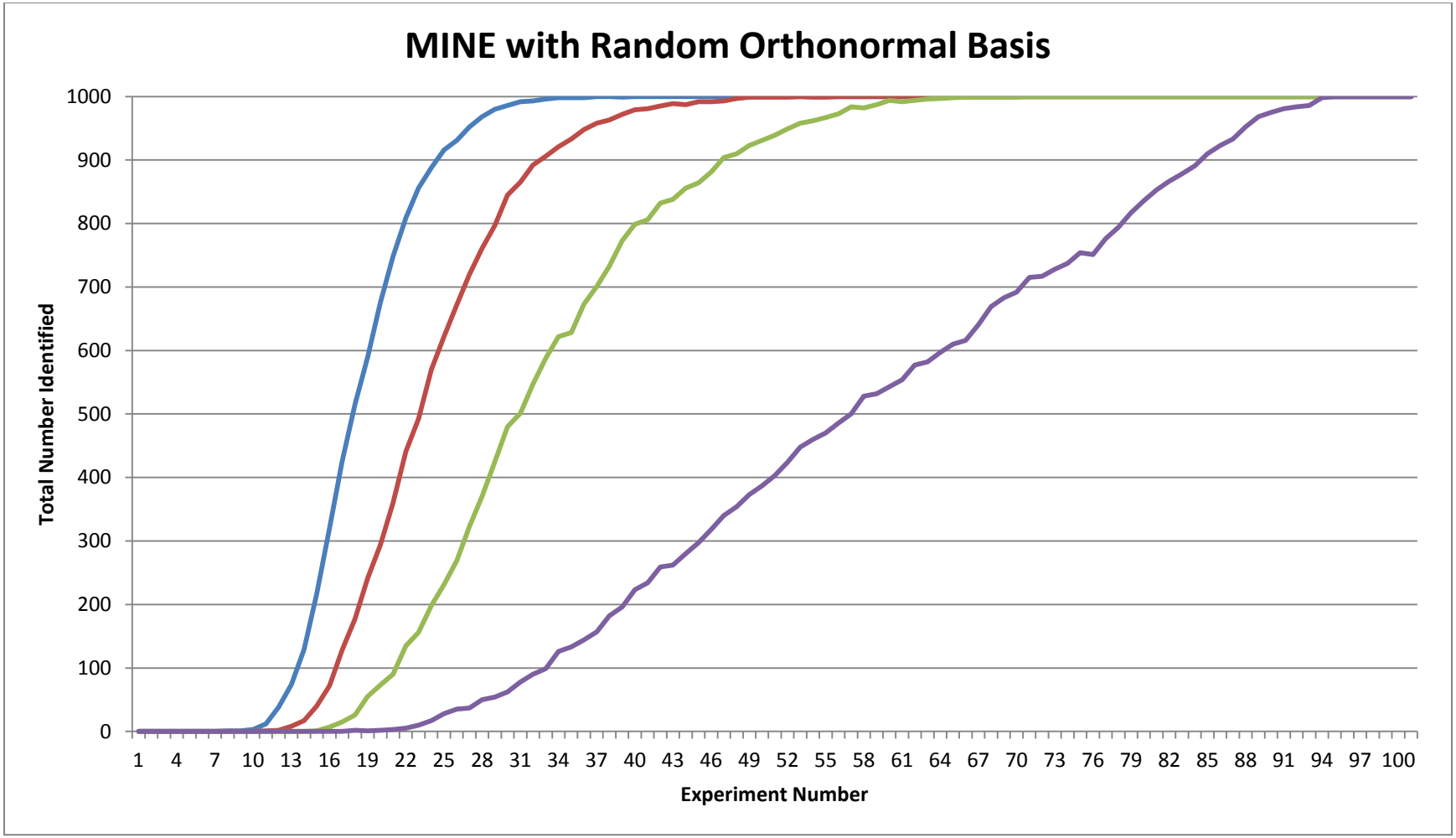

Figure 5: Graph of significant nonzero regression coefficients for the MINE with random orthonormal basis method. The graph identifies a number of replicates (y-axis) with a varying percentage of the nonzero β_0 components as significant as a function of the number of experiments (x-axis) for the MINE with random orthonormal basis method. Blue corresponds to 70% correctly identified, red to 80%, green to 90% and purple to 100%.

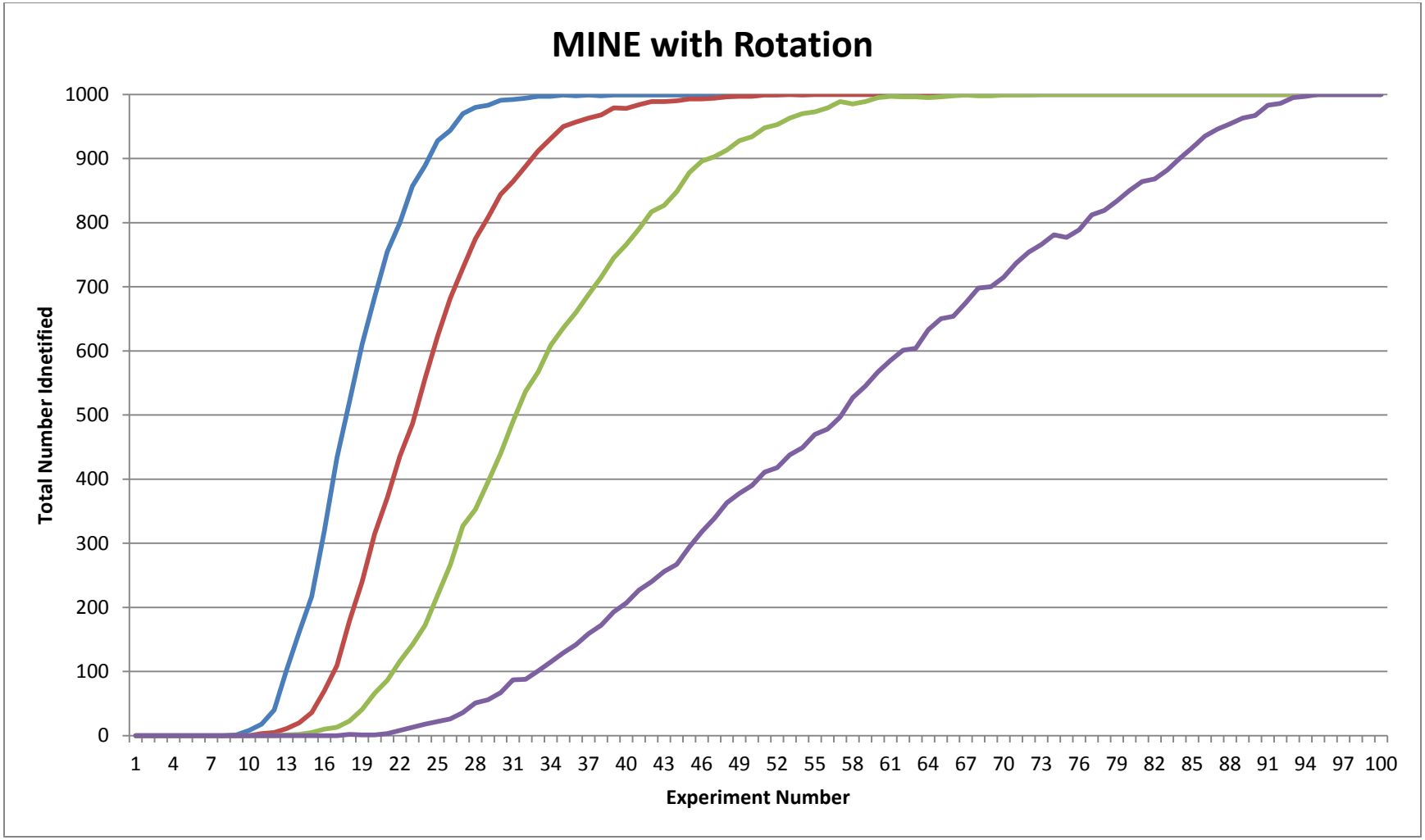

Figure 6: Graph of significant nonzero regression coefficients for the MINE with rotation method. The graph identifies a number of replicates (y-axis) with a varying percentage of the nonzero β_0 components as significant as a function of the number of experiments (x-axis) for the MINE with rotation method. Blue corresponds to 70% correctly identified, red to 80%, green to 90% and purple to 100%.

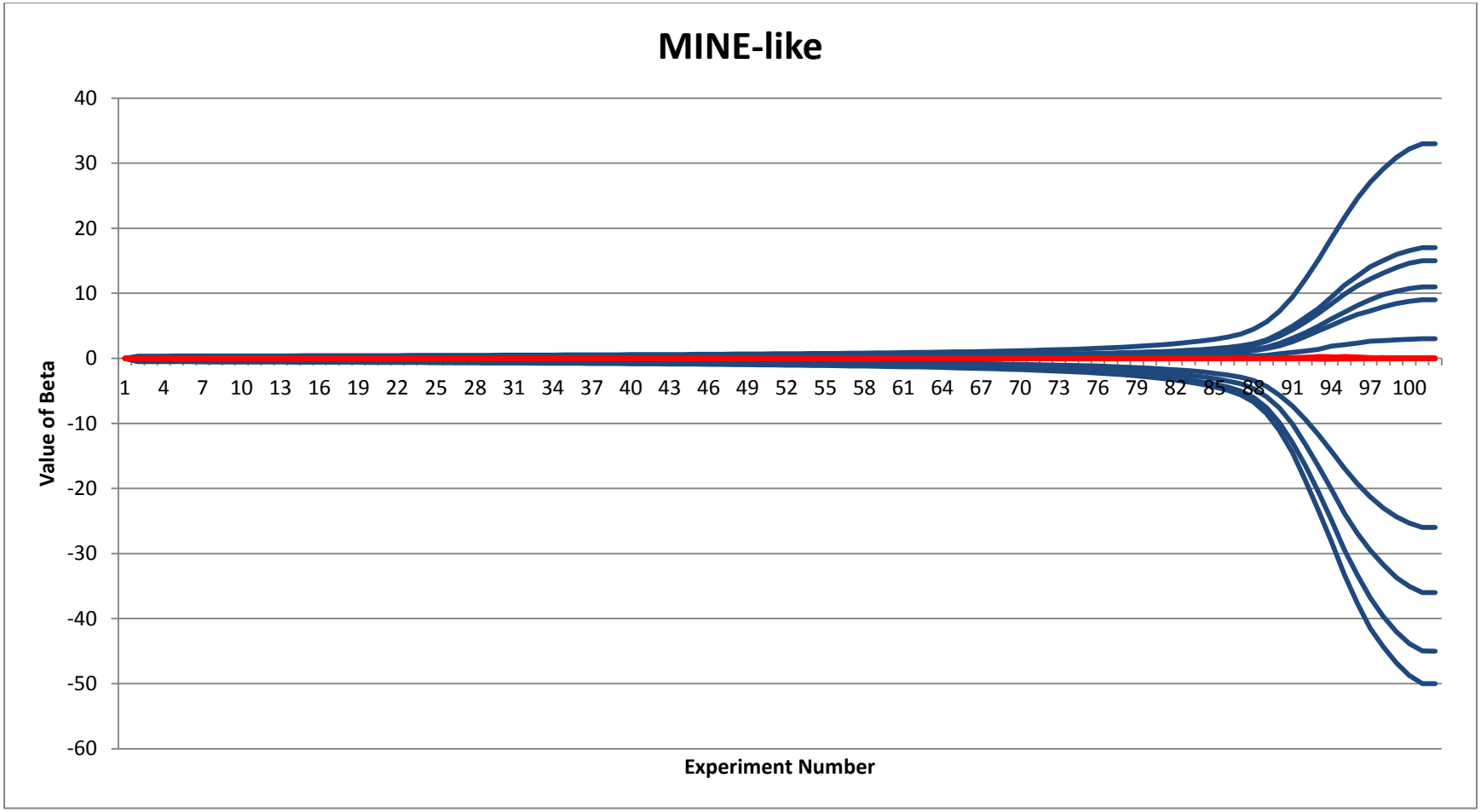

Figure 7: Posterior means of the first 20 regression coefficients for the MINE-like method as a function of the number of experiments. This is averaged over all simulations with 10 zero (in red) and 10 nonzero (in blue). The first ten (red) are truly nonzero.

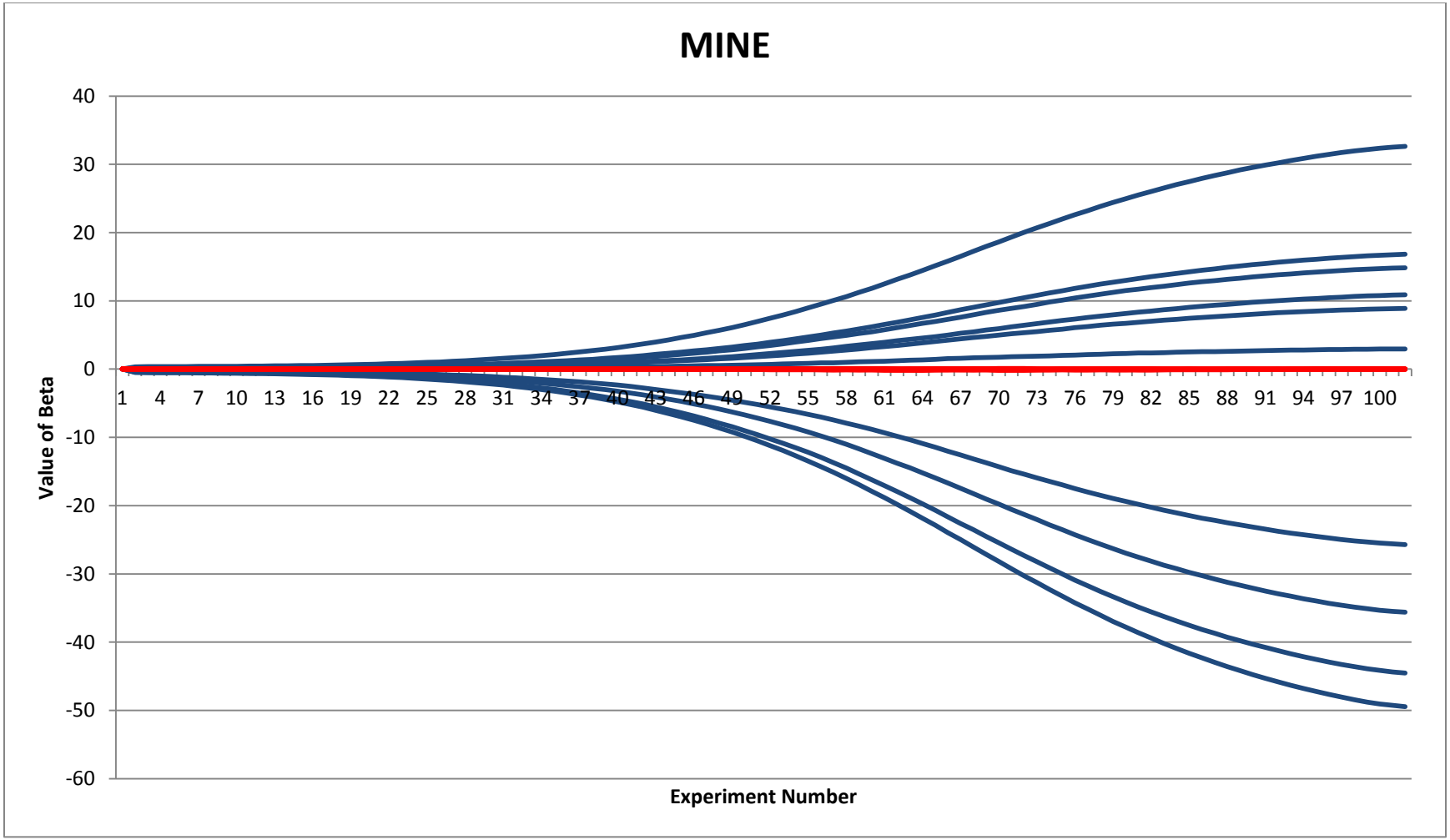

Figure 8: Posterior means of the first 20 regression coefficients for the MINE method as a function of the number of experiments. This is averaged over all simulations with 10 zero (in red) and 10 nonzero (in blue). The first ten (red) are truly nonzero.

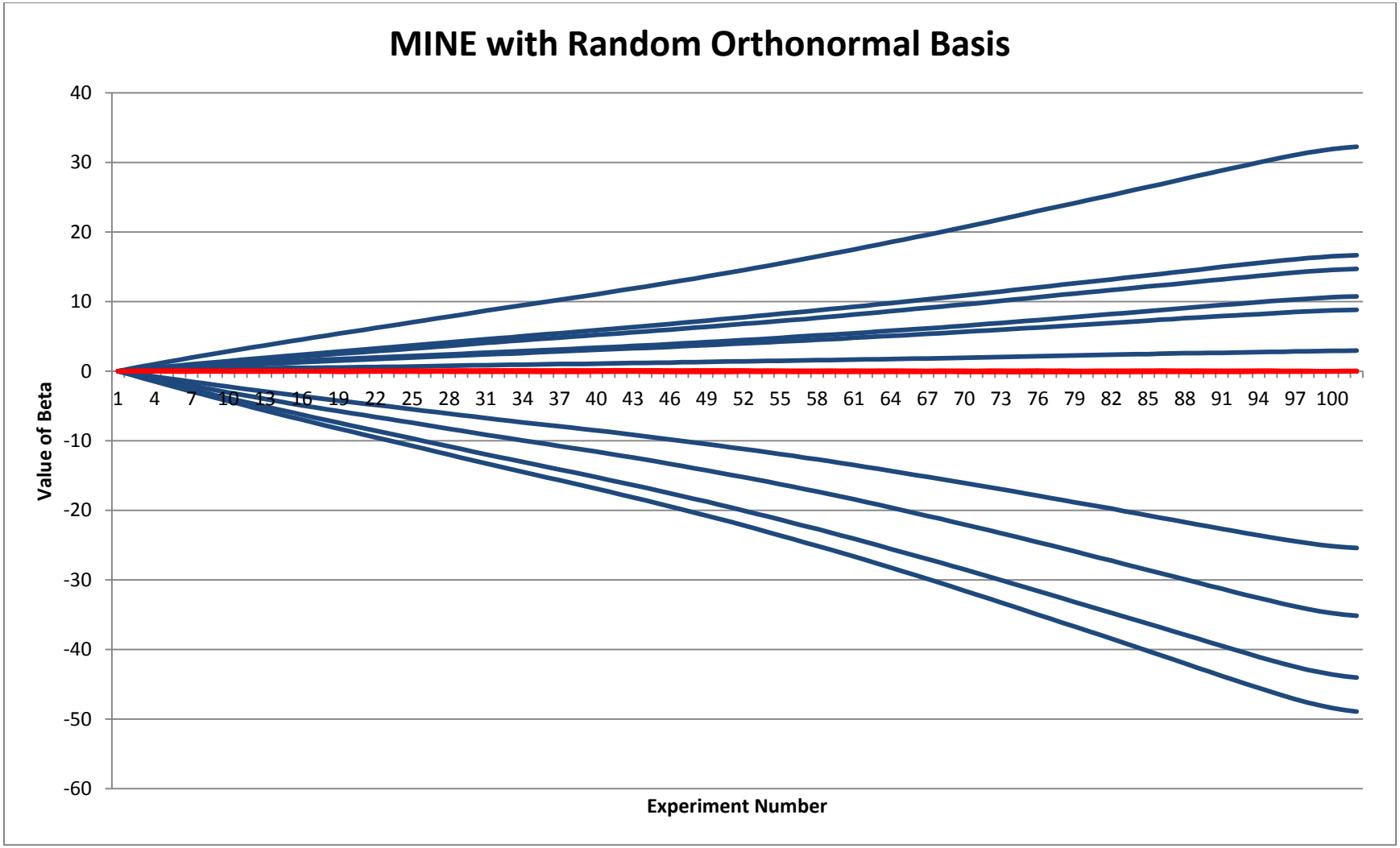

Figure 9: Posterior means of the first 20 regression coefficients for the MINE with random orthonormal basis as a function of the number of experiments. This is averaged over all simulations with 10 zero (in red) and 10 nonzero (in blue). The first ten (red) are truly nonzero.

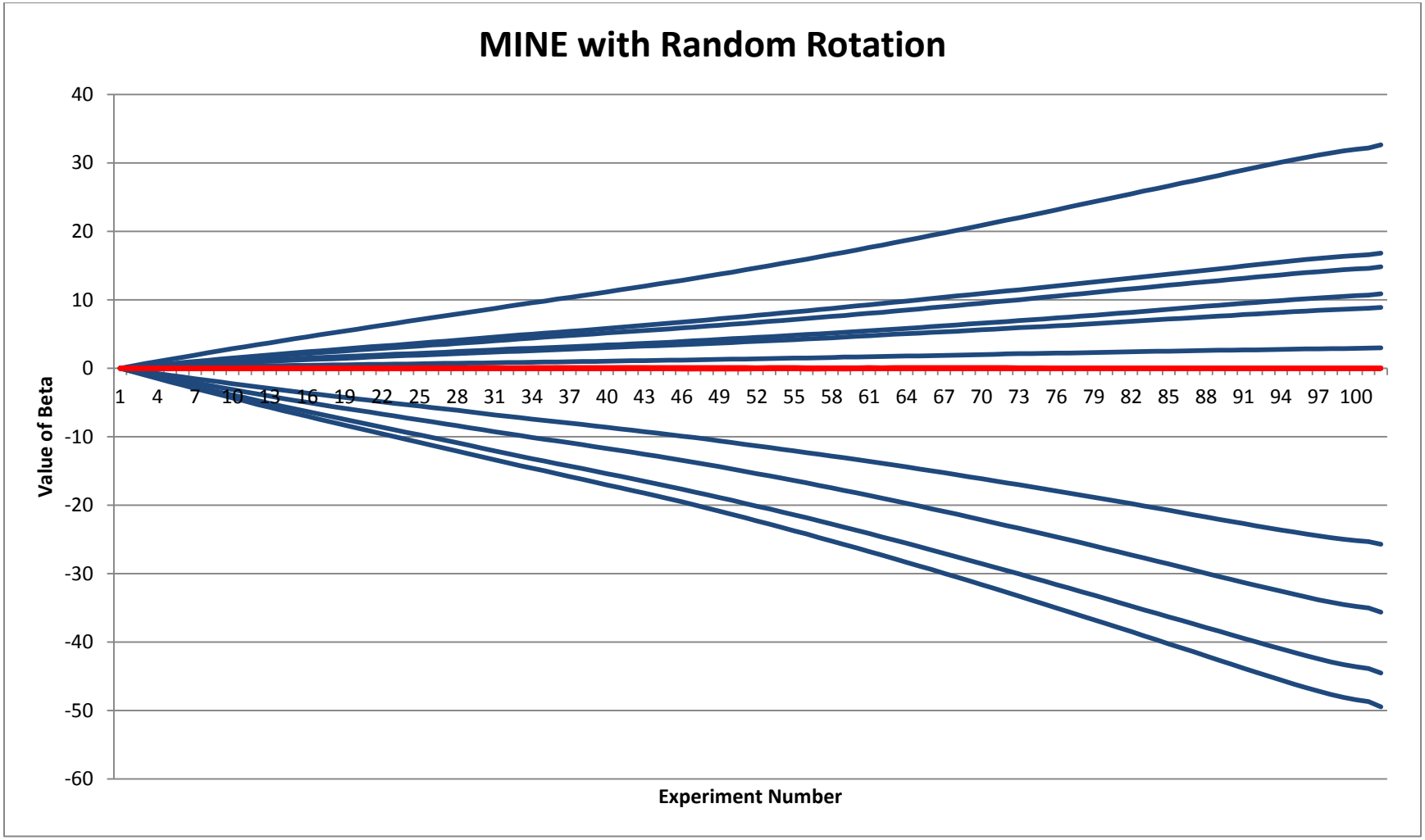

Figure 10: Posterior means of the first 20 regression coefficients for the MINE with rotation as a function of the number of experiments. This is averaged over all simulations with 10 zero (in red) and 10 nonzero (in blue). The first ten (red) are truly nonzero.

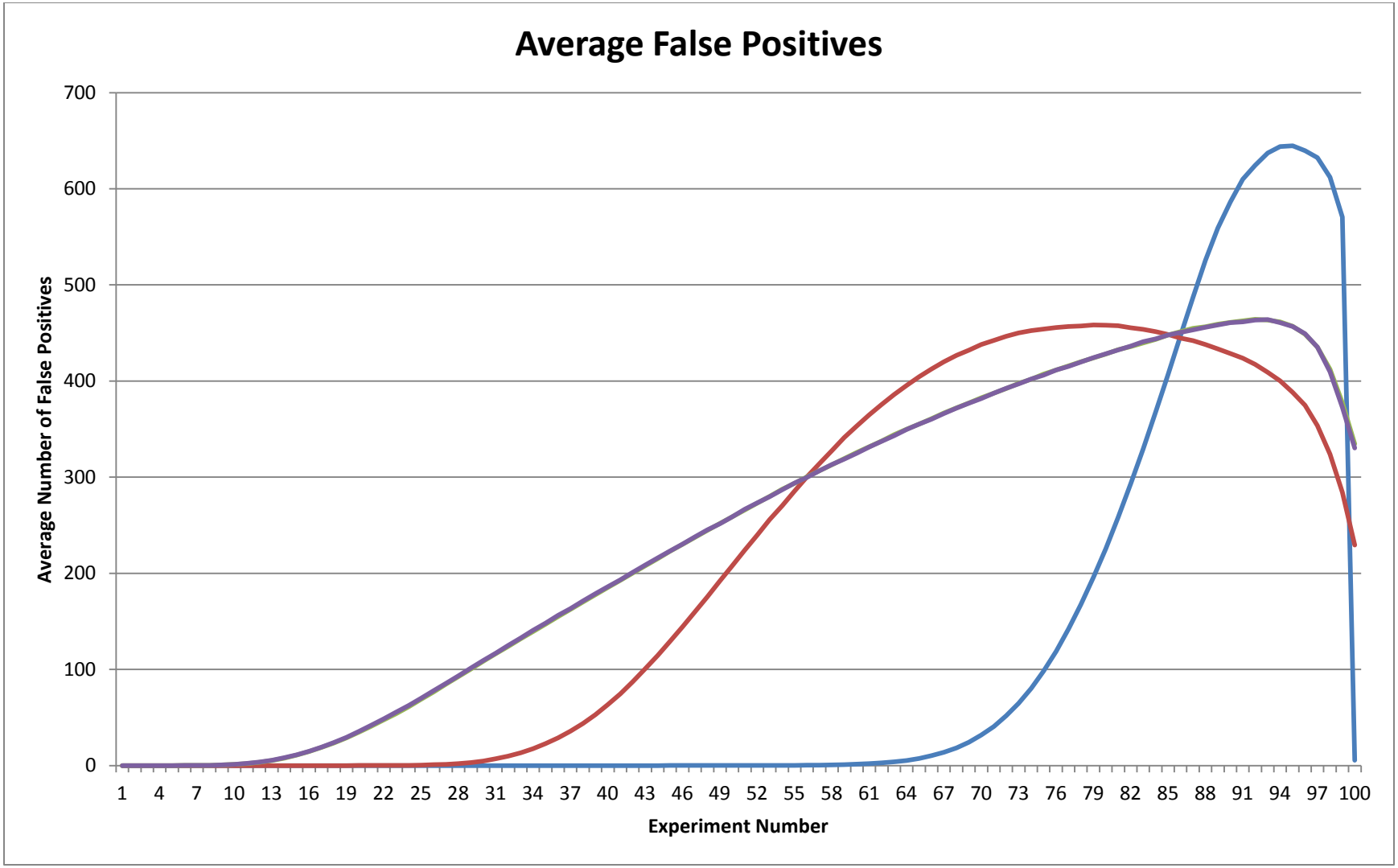

Figure 11: The number of false positives as a function of the number of experiments. These numbers are averaged over all simulations for each method. Blue corresponds to MINE-like, red to MINE, green to MINE with random orthonormal basis, and purple to MINE with random rotation. The final two overlap almost exactly which is why the green line is not visible.