ABSTRACT

Title of dissertation: STATISTICAL AND OPTIMAL LEARNING WITH APPLICATIONS IN BUSINESS ANALYTICS

Bin Han, Doctor of Philosophy, 2015

Dissertation directed by: Assistant Professor Ilya O. Ryzhov
Robert H. Smith School of Business

Statistical learning is widely used in business analytics to discover structure or exploit patterns from historical data, and build models that capture relationships between an outcome of interest and a set of variables. Optimal learning on the other hand, solves the operational side of the problem, by iterating between decision making and data acquisition/learning. All too often the two problems go hand-in-hand, which exhibit a feedback loop between statistics and optimization.

We apply this statistical/optimal learning concept on a context of fundraising marketing campaign problem arising in many non-profit organizations. Many such organizations use direct-mail marketing to cultivate one-time donors and convert them into recurring contributors. Cultivated donors generate much more revenue than new donors, but also lapse with time, making it important to steadily draw in new cultivations. The direct-mail budget is limited, but better-designed mailings can improve success rates without increasing costs.

We first apply statistical learning to analyze the effectiveness of several design approaches used in practice, based on a massive dataset covering 8.6 million direct-mail communications with donors to the American Red Cross during 2009-2011. We find evidence that mailed appeals are more effective when they emphasize disaster preparedness and training efforts over post-disaster cleanup. Including small cards that affirm donors' identity as Red Cross supporters is an effective strategy, while including gift items such as address labels is not. Finally, very recent acquisitions are more likely to respond to appeals that ask them to contribute an amount similar to their most recent donation, but this approach has an adverse effect on donors with a longer history. We show via simulation that a simple design strategy based on these insights has potential to improve success rates from 5.4% to 8.1%.

Given these findings, when new scenario arises, however, new data need to be acquired to update our model and decisions, which is studied under optimal learning framework. The goal becomes discovering a sequential information collection strategy that learns the best campaign design alternative as quickly as possible. Regression structure is used to learn about a set of unknown parameters, which alternates with optimization to design new data points. Such problems have been extensively studied in the ranking and selection (R&S) community, but traditional R&S procedures experience high computational costs when the decision space grows combinatorially. We present a value of information procedure for simultaneously learning unknown regression parameters and unknown sampling noise. We then develop an approximate version of the procedure, based on semi-definite programming relaxation, that retains good performance and scales better to large problems. We also prove the asymptotic consistency of the algorithm in the parametric model, a result that has not previously been available for even the known-variance case.

STATISTICAL AND OPTIMAL LEARNING WITH APPLICATIONS IN BUSINESS ANALYTICS

by

Bin Han

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Assistant Professor Ilya O. Ryzhov, Chair/Advisor
Assistant Professor Shawn Mankad
Associate Professor Leonid Koralov
Assistant Professor Yuan Liao
Professor Michael Ball

Acknowledgments

I would like to acknowledge help and support from all the people who have made this thesis possible.

First and foremost I'd like to thank my advisor, Dr. Ilya O. Ryzhov for his invaluable support on my research. Whenever I encounter difficulties, he is always available for guidance and very resourceful on providing ideas and references.

I gratefully acknowledge all the people I collaborated with on my research. For the optimal learning project, I would like to thank Dr. Boris Defourny for his ideas on convex approximation. For the fundraising marketing project I acknowledge Dr. Jelena Bradić for support on statistical theories and Aleksandar Bradić for data processing. I am also grateful to the American Red Cross, Russ Reid Company, and the Wharton Customer Analytics Initiative for providing the opportunity to work with the dataset that forms the basis for the project. I would like to thank Tony DiPasquale, Sharron Silva, and John Wilburn at the American Red Cross for their support on this research.

I would also like to thank my committee members. Dr. Leonid Koralov has helped me on fundamental theories in probability. Dr. Michael Ball has taught me on integer programming. From Dr. Shawn Mankad I have learned many insights and ideas when working together on projects. Finally I would like to thank Dr. Yuan Liao for discussing with me about my dissertation and joining the committee.

I would also like to thank all professors that have helped me on my graduate studies. Dr. Paul Smith has taught me applied statistics and for many times he

has been the “go-to” person when I have questions in statistical modeling. Dr. Erid Slud has taught me mathematical statistics which helps me on understanding many fundamental questions in statistics. Dr. Aravind Srinivasan and Dr. Hal Daumé have helped me on understanding of algorithms, data structure and machine learning.

I would also like to acknowledge Alverda McCoy and Celeste Regaldo for your help and support on administrative issues.

It is impossible to remember all, and I apologize to those I’ve inadvertently left out.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AIC	Akaike's Information Criterion
ANN	Artificial Neural Network
ARC	American Red Cross
AUC	Area Under the Curve
BIC	Bayesian Information Criterion
BLB	Bags of Little Bootstrapping
CKG	Correlated Knowledge Gradient
CV	Cross Validation
DT	Decision Tree
EM	Expectation Maximization
GAM	Generalized Additive Model
GBM	Gradient boosting machine
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
KGUP	Knowledge Gradient with Unknown Precision
KNN	K-Nearest Neighbor
Lasso	Least Absolute Shrinkage and Selection Operator
ML	Maximum Likelihood
OCBA	Optimal Computing Budget Allocation
OLS	Ordinary Least Squares
PCA	Principle Component Analysis
PII	Personally Identifiable Information
ROC	Receiver Operating Characteristic
R&S	Ranking and Selection
SDP	Semi-definite Programming
STAART	Strategy Through Applied Analytics, Research and Testing
SVD	Singular Value Decomposition
SVM	Support Vector Machine
VIF	Variance Inflation Factor
VIP	Value of Information

Chapter 1: Introduction and Overview

The terms *statistics*, *machine learning*, *data mining* and *predictive analytics* have been used interchangeably in both academia and industry. They can all be generally described, however, as *learning from data*. This ambiguity in terminology reflects interdisciplinary collaborations among different fields on tackling the same problem. In the early days, statisticians often encountered data from agriculture, industrial engineering, social science, etc., and these data were relatively small. Recent advances in computing power, data storage and web technology have brought us into an era of *big data*, which have captured interests of researchers from many other domains such as computer science and electronic engineering.

Despite considerable overlap between fields, different terms may imply a different focus or objective. Statistics often aims at explaining the relationship between an outcome of a measurement (quantitative or categorical) and a set of independent variables, or estimating the underlying distribution of given data. Machine learning on the other hand, concentrates on prediction power of the model, i.e., if the learner built from the training data generalizes to new or future unseen data [1]. Data mining focuses more on the practical side of learning, which may indicate many aspects such as data wrangling, integration, exploratory analysis to learning, prediction and data visualization [2]. In this thesis, however, we use the term *statistical learning* to refer to all of these facets.

Statistical learning can usually fall into two categories: *supervised* and *unsupervised* learning [1]. In supervised learning, we try to build a model or learner that relates a response variable (such as risk, revenue, success rate, etc.) to a set of *features* (such as experimental designs, field measurements, subject characteristics, etc). The learner should generalize beyond the training data. To guarantee good prediction performance, one should pay extra attention on the over/under-fitting dilemma, namely the model should fit well enough to the pattern of the training data (low bias) while not being overly complex which picks up errors or randomnesses in the training data, thus being robust to new testing data (low variance), i.e., the bias-variance trade-off [3].

In unsupervised learning, we observe features or covariates but are not given the response variable. The problem is instead using features to describe how data are organized or structured. Such examples include clustering, feature transformation and selection.

The following sections provide a survey of classical and recent work on statistical learning. Discussions in Section 1.1, 1.3, 1.4, 1.5 are directly applicable in Chapters 2. Methods discussed in Section 1.6 are related to Chapter 3.

1.1 Linear Statistical Models

Linear models as well as their generalized versions are widely used in statistical learning and can be considered as one of the most classical approaches in supervised learning. Linear regression and *ordinary least squares* (OLS) can be dated back to the 19th century when Gauss and Legendre first used them on astronomical data [4]. The performance and many good properties of OLS estimates highly depend on assumptions on the errors, e.g., independence, constant variance, zero mean etc. When they are normally distributed, the OLS estimates can be proved to be the *best linear unbiased estimator* (BLUE) [5].

1.1.1 Generalized Linear Models

Generalized linear models (GLMs) [6] incorporate more general distributions (within the exponential family) other than the normal distribution for the response variable, e.g., binomial, Poisson, Gamma as well as compound Poisson-Gamma (Tweedie) distributions [7]. The relation between the mean of response μ and the set of covariates x is described by a link function g, i.e.,

$$g(\mu) = \eta = x^\top \beta,$$

where β is the vector of coefficients. If we write the density of the response in the canonical form of the exponential family

$$f_Y(y; \theta, \phi) = \exp\left(\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right),$$

it is easy to obtain $\mu = b'(\theta)$, or $\theta = (b')^{-1}(\mu)$. The canonical link function is then defined by $g(\mu) = \theta = (b')^{-1}(\mu)$. By using the canonical link function we obtain a sufficient statistic equal in dimension to β in the linear predictor η [6]. Each distribution corresponds to a specific canonical link, e.g., the link for normal distribution is $\eta = \mu$, while for binomial distribution it is $\eta = \text{logit}(\mu) = \log(\mu/(1-\mu))$.

In GLMs the parameters are estimated using the maximum likelihood (ML) approach. Instead of looking at the sum of squared errors, we use *deviance* to assess model fit, which is defined as $2l(y, \phi; y) - 2l(\hat{\mu}, \phi; y)$, i.e., twice the difference between log likelihood of a saturated model (with n parameters) and the model under investigation. Maximizing the likelihood is equivalent to minimizing the deviance. Notice that, for normal distribution, the deviance is exactly the same as the sum of squared error, so GLMs cover least squares regression as a special case. For the binomial distribution, the deviance can be derived as

$$2\Sigma_i \left(y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i) \log((m_i - y_i)/(m_i - \hat{\mu}_i) \right),$$

where i denotes the i-th covariate class and m_i, y_i denote the number of experiments and number of successes in i-th class respectively.

Typical ways of solving the ML problem include *Fisher scoring* [6] or Newton's method. In Fisher scoring the expected Hessian matrix $\mathbb{E}(\frac{\partial^2 l}{\partial\beta_i \partial\beta_j})$ is used, while Newton's method uses the observed value. However, under the canonical link, the Hessian matrix is constant so the two methods coincide.

A further generalization of GLMs are the *generalized additive models* (GAMs).

In GAMs the predictor is transformed using some smooth functions, and the model can be written as

$$g(\mu) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p).$$

Methods such as the backfitting algorithm [8] can be used to estimate the parameters. When using GAMs or any such more complicated models, the problem of over-fitting becomes more prominent and deserves more caution.

1.1.2 Mixed Model

In many cases specific characteristics of the data can determine which statistical model to use. For example, when data are longitudinal, i.e., data entries are grouped by panels or blocks, we should consider the effects from each panel. Examples include transaction data of a set of customers, where the transaction records are grouped by customer ids, or teaching evaluations from a number of high schools, where the teachers are grouped by schools. In such cases the number of panels can be extremely large, which makes it unwise to treat them directly as fixed effects. We might also have limited access to specific information about each panel, such as demographic information of customers or characteristics of schools. One common approach is to use random variables to model the effects for each panel, i.e., the *mixed model* [9]. If we use it under the GLM framework, the resulting model is referred to as a *generalized linear mixed model* (GLMM). The model can be written as

$$\mathbb{E}(y|u) = X\beta + Zu,$$

where X and β are the design matrix and coefficients for the fixed effects, and Z and u are the matrix and variables for the random effects. To obtain the likelihood function we need to compute an integration over the random variables. The problem is typically solved using expectation-maximization (EM) algorithms [10] where random effect terms are treated as missing data and estimated iteratively. When data become large, however, the computation becomes intractable [11].

1.2 Other Supervised Learning Methods

There are many other supervised learning algorithms that have been developed in machine learning and artificial intelligence domains to solve either classification or regression problems. Such learning algorithms focus less on statistical properties of the model but more on its prediction performance. Sometimes the learner is described directly by its prediction outcome. For example, in binary classification problems, the learner can be specified by its decision boundaries, which depict regions on the domain of features where the responses should be labeled as positive or negative.

1.2.1 Decision Tree

Decision trees (DTs) use recursive algorithm to build a tree-like graph or model of features or decisions and their corresponding outcomes [3]. DTs are typically used for classification problems. In DTs, each internal node tests an attribute and each branch corresponds to the attribute value. The leaf node then assigns a classification to the data point. The decision boundaries of DTs are usually shaped as high dimensional rectangles in the feature space.

The top-down order in which features enter the model is determined by information gain. Specifically, define entropy for a sample S as $\text{Entropy}(S) = -p_+ \log p_+ - p_- \log p_-$, where p_+ and p_- are the proportion of positive and negative examples respectively. The information gain is then defined as

$$\text{Gain}(S, A) = \text{Entropy}(S) - \Sigma_{v \in \text{values}(A)} \frac{|S_v|}{S} \text{Entropy}(S_v),$$

where A is a particular feature and S_v is the subset of S having $A = v$.

Hyper parameters in DTs include height of the tree or threshold for minimum number of points allowed in leaf nodes. A typical way of preventing over-fitting is by pruning, which evaluates impact on validation dataset from pruning each possible node (and nodes below it). Nodes are removed greedily according to the improvement on validation set accuracy.

1.2.2 K-nearest neighbor

K-nearest neighbor (KNN) [12] is a type of instance-based learning or lazy learning, meaning that there is no training phase and the learning only occurs when predicting new instances. In classification problem, the label of a new instance is based on labels of k data points that are closest to the data point in interest. The distance metric varies by problems. Examples include Euclidean distances, Manhattan distances, Hamming distances [13], etc. Results can be determined by a majority vote or weighted average based on the distance of each data point to the particular point in interest. KNN can also be used in regression problems. Given a new data point, KNN first select k nearest data points based on certain distance metric, then a local linear regression model is build only using these k data points.

The hyper-parameter k in KNN can be used to specify the variance-bias tradeoff. For smaller k we obtain more complicated decision boundaries in the model, which has lower bias but higher variance. For larger k we obtain simpler model with higher bias but lower variance. A validation dataset is typically used to specify this parameter.

There are several issues in using KNN. The first one is the curse of dimensionality, especially for Euclidean distance metric. In high dimensional space the Euclidean distance between any two data points is almost equivalent, which makes it more difficult to define "neighbors" [14]. Another problem is scaling of each feature, which dramatically influences the distance. Thus one needs to make sure the unit used for each feature represents its actual contributions to the distance. A third problem is correlation among features. In such cases special distance metric needs to be considered, such as Mahalanobis distance with a predefined covariance matrix [15].

1.2.3 Support Vector Machine

Classic *support vector machine* (SVM) uses a hyperplane to separate negative from positive examples in the feature space in classification problems. Thus the learner produces a linear decision boundary for linearly separable samples [16]. The orientation of the plane is determined by maximizing the margin, which is defined as the distance of nearest point to the plane. The two marginal hyperplanes can be written as $w^\top x + b = \pm 1$, i.e., positive examples satisfy $w^\top x + b \geq 1$ and negative examples satisfy $w^\top x + b \leq -1$. The optimization problem can be written as

$$\begin{aligned} &\min \frac{1}{2} \parallel w \parallel^2 \\ &s.t.\ y_i(w^\top x_i + b) \geq 1, \text{ for } i = 1 \ldots n. \end{aligned}$$

The optimization is solved by finding its dual problem using Lagrange multipliers and maximizing the quadratic linear program

$$\begin{aligned} &\max \Sigma_{i=1}^n \lambda_i - \frac{1}{2} \Sigma_{i=1}^n \Sigma_{j=1}^n \lambda_i \lambda_j y_i y_j x_i^\top x_j \\ &s.t.\ \lambda \geq 0 \text{ and } \lambda^\top y = 0, \end{aligned}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)^\top$ denotes the Lagrange multipliers. The second term in the dual objective $y_i y_j x_i^\top x_j$ reveals a particular choice of the *kernel function*. To obtain a non-linear decision boundary for non-linearly separable samples, the kernel function needs to be changed. Examples include the polynomial kernel $(x_i^\top x_j + 1)^d$ or the radial basis kernel $\exp(-\frac{\|x_i - x_j\|^2}{2\sigma^2})$ [17].

1.2.4 Artificial Neural Network

Artificial neural network (ANN) is an example of a "black-box" learner, which is capable of producing any arbitrary decision boundary as needed. Its structure resembles a biological neural network, which consists of interconnected nodes or "neurons". It typically contains three layers: input layer, intermediate layer and output layer. The intermediate layer can contain multiple sub-layers. Each node in the input layer receives an input of a particular feature, and passes the “signal” to all the nodes in intermediate layers. The intermediate nodes receive information from multiple sources and produce certain outputs, governed by the *activation function*. The outputs are then passed to the output nodes to generate predictions [3].

The performance of an ANN learner highly depends on the structure of the network and specification of nodes. Two commonly used neuron types are perceptron and sigmoid. The activation function for perceptron is

$$y = \begin{cases} 1 & x^\top \beta > b \\ \\ 0 & \text{otherwise} \end{cases}$$

The sigmoid activation function is

$$y = \frac{1}{1 + \exp(x^\top \beta)}.$$

Both activation functions are non-linear, which enables the possibility of creating arbitrary decision boundaries. There are many algorithms for estimating the parameters in ANN. Most of the algorithms employ some form of gradient descent, using back-propagation to compute the actual gradients [18].

1.3 Model Validation

The process of statistical modeling often involves iterations between model fitting and model validation. It is often difficult to obtain a good model in one run, but the quality of the model can be improved through an iterative process.

Correctness of statistical assumptions needs to be checked in model validation process. For OLS, residual vs. fitted value/covariates plot validates the homoscedasticity and zero-mean assumptions of the error. QQ-plot provides validation about the normality assumption. Jack-knife plot and Cook's distance plot can be used to check the outliers and influential points respectively [19].

The problem of collinearity occurs when covariates are linearly dependent on each other, i.e., the design matrix is singular or close to singular. In such case the variance of parameter estimates increase dramatically and the model becomes less robust, i.e., small perturbation in the data would lead to large difference in the estimates [20]. The interpretation of estimated parameters also becomes ambiguous. A typical way of detecting collinearity is to use *variance inflation factor* (VIF), defined as

$$v\hat{a}r(\hat{\beta}_j) = \frac{s^2}{(n-1)v\hat{a}r(x_j)} \frac{1}{1-R_j^2},$$

where R_j^2 is the R^2 when x_j is regressed on other covariates.

In GLMs, instead of using residual, we can generate deviance plot in a similar fashion and do F test for significance of the model. The test, however, is only valid asymptotically and should not be relied upon when data are small [6].

Beyond assumption validation, the model performance also needs to be checked, as well as the issue of over-fitting and under-fitting. Out-of-sample testing is usually used in such task. When data are relatively small, cross-validation (CV) can be used to reduce the variance in the testing data. When data are large, however, CV can be computationally expensive and splitting the data into training and testing sets will suffice. The problem of over-fitting is detected when the performance in the training data is significantly better than that in the testing data.

Different metrics can be used to test model performance for various purposes. For binary classification problem, we can plot *receiver operating characteristic* (ROC) curve and compute area under the curve (AUC) if the model output is the probability of being positive (e.g. logistic regression) [21]. When lowering the threshold for classifying new data as positive (thus increasing sensitivity), one would expect to increase the false positive rate (fall-out), so the larger the AUC the better the model. For continuous response variable, *cumulative gain chart* or *Gini index* are sometimes used to assess the model [22]. The cumulative gain chart is obtained by ordering the data by fitted value and cumulatively plot the true response for each tier of the fitted value. It reveals the effectiveness of the predictive model on separating high responses from low responses.

1.4 Model Selection

The problem of model selection arises when we want to build a concise and interpretable model, but are given a large number of variables which are likely to be correlated. The problem are often coupled with the collinearity issue. Below, we discuss several approaches to tackle this issue.

1.4.1 Ridge Regression

Ridge regression [23] uses an l_2 penalty term to address the collinearity issue. The coefficients are estimated by minimizing

$$\Sigma_i (y_i - x_i^\top \beta)^2 + \lambda \Sigma_{j=1}^p \beta_j^2,$$

where λ is the tuning parameter. The solution is given by

$$\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y.$$

Recall that collinearity results in a singular or close to singular $X^\top X$. By adding the term λI, the inversion becomes much less problematic. Thus ridge regression reduces the variance of $\hat{\beta}$. On the other hand, since denominator is increased, the estimates are down-biased. Ridge regression provides a way of reducing the model variance, but unfortunately doesn't do the model selection task.

1.4.2 Stepwise Regression

Stepwise regression is a combination of forward selection and backward elimination, which includes or excludes variables using a series of F-tests or t-tests [24–26]. The algorithm uses thresholds of p-values for entering the model (p_{in}) and leaving the model (p_{out}). In the forward selection step, it adds one variable from the pool of remaining variables with the smallest p-value in t-test for coefficient estimates given existing variables present in the model, and stops when all p-values are greater than p_{in}. In the backward elimination step, it deletes one variable from the existing variables with the largest p-value while satisfying $p > p_{out}$, and stops deleting if all p-values are smaller than p_{out}.

We can see that, the lower the thresholds for p_{in} and p_{out}, the harder it is for variables to enter the model and the easier it is for variables to leave the model. Thus the lower the threshold, the conciser the model and in turn the less the over-fitting problem.

Stepwise regression provides a practical way for model selection. It still, however, searches in a large candidate space and usually involves large number of iterations. When data are large, it becomes more computationally intensive.

1.4.3 Principal Component Analysis

Principal component analysis (PCA) [27] is an unsupervised learning algorithm for model selection. It transforms a set of possibly correlated variables into a new set of linearly uncorrelated variables (orthogonal components). The method is particularly useful when there exists a complex and unclear dependence relation among covariates. In such case directly applying the original covariates to the model provides no clear explanation about the effects of each covariate. Thus it is plausible to generate new features which bear no physical interpretation, since the main task here is to prevent over-fitting and improve prediction performance [28].

The principal components are defined as the vector in the feature space with the largest variance, defined as the sample variance of projections of all data points onto the vector. It is also the eigenvector with the largest eigenvalue in the feature space. Another way of interpreting it is finding the vector such that the sum of squares of distance of all other points to the vector is minimized. In some fields it is also named *singular value decomposition* (SVD) [29].

The number of principal components can be smaller than the original number of features, if we set a minimum threshold on the variance or eigenvalue. Thus it can perform model selection and dimension reduction, but the interpretability of the model or features is lost.

1.4.4 Lasso

Lasso stands for "least absolute shrinkage and selection operator", which uses an l_1 penalty to reduce some of the coefficients down to zero, achieving both model selection and variance reduction [30]. It was first proposed for improving the performance of OLS. The objective can be written as

$$\hat{\beta}_\lambda = \arg\min \left(\parallel Y - X\beta \parallel_2^2 + \lambda \Sigma_{j=1}^{p} |\beta_j| \right),$$

where λ is a hyper-parameter. Notice that the penalty term does not include the intercept. Tuning of λ changes the balance between model bias and model conciseness. It is typically determined in an out-of-sample fashion, e.g., cross-validation. When data are large, however, other metric can be used such as Akaike's Information Criterion (AIC) [31] or Bayesian Information Criterion (BIC) [32].

An extension of the method named "grouped Lasso" can select variables in groups [33], which is particularly useful for categorical variables, where it may be desirable to select all dummy variables all at once. The objective in this case becomes

$$\hat{\beta}_\lambda = \arg\min\left(\| Y - X\beta \|_2^2 + \lambda \Sigma_{g=1}^{G} \| \beta_{I_g} \|_2\right),$$

where I_g is the index set for gth group of variables. The penalty can be considered as one in between l_1 and l_2 penalty. The group Lasso method can also be used together with GLM such as logistic regression [34]. A further extension of Lasso is one combined with mixed effects models [35].

Lasso was first introduced to solve the problem when $N \ll p$. Recent research, however, has also found substantial practical benefits of using Lasso in applications with $N \gg p$ [36].

Similar to ridge regression, Lasso also introduces a down-bias to the estimates. A common practice is to implement a two-stage procedure, which first uses Lasso to perform model selection, and then use the selected feature to refit the model without using any penalties.

1.4.5 Gradient Boosting

Gradient boosting machine (GBM) is an ensemble learning method built from a group of weak learners [37]. It has gained recent popularity for solving model selection and collinearity problems [38]. GBM can work with any model fit meaure ρ given that it is differentiable. The fitting procedure is typically conducted in a step-wise fashion. Define the negative gradient of ρ as

$$U(\hat{Y}) = -\frac{\partial\rho(Y,\hat{Y})}{\partial\hat{Y}},$$

where $\hat{Y}$ is fitted value for the response Y. At step m, it regresses $U(\hat{Y}^m)$ on one of week learners and obtains $\hat{U}(\hat{Y}^m)$ which is the fitted result from the best learner. Then it updates

$$\hat{Y}^{m+1} = \hat{Y}^m + v\hat{U}(\hat{Y}^m),$$

with v being the step size. The process is then repeated until some convergence threshold.

The weak learners are typically the features, i.e., each feature corresponds to one weak learner. Since they enter the model in a particular order (however, one feature can enter the model multiple times), less important features would be less likely to enter the model, or at least in a much later time. Thus a more concise model is obtained if we stop the process at an earlier time. Similar to Lasso, GBM also produces down-biased estimates.

1.5 Bootstrapping and Bagging

A bootstrap sample is obtained by resampling from the data with replacement. It is possible for one data point to enter the bootstrap sample multiple times, which provides a way to generate new samples from existing data. Bootstrapping has been used to assess many properties of estimators, such as bias, variance, confidence intervals, prediction error, etc [39]. Uniform distribution is typically used in the sampling process, but more general distributions have also been proposed such as importance sampling [40]. We may sample data by panels, or assign different weights on data points, e.g., up-sampling of positive examples when the proportion of positive examples is low.

Bootstrap sample can be generated on the same size-scale as the original dataset, but the computation becomes burdensome for large datasets. One natural approach is small sub-sampling, which generates samples on a smaller scale compared to the original dataset. The statistics literature demonstrates, however, that using a single smaller-order subsample can bias the outcome of model by introducing false positives [41]. This can occur if $M < \frac{N}{5}$ (where N is the sample size and M is the subsample size). Furthermore, a stronger result has been proved that, if only a single subsample is considered, it is virtually impossible to retrieve a sparse set of significant features (that is, the probability of doing so vanishes to zero) [42]. Subsampling introduces additional noise into the problem. Thus, a single subsample may inflate the variance of the estimated coefficients, analogous to how the variance of a classical sample mean is larger when the sample size is smaller. A feature that appears infrequently in the big data may be misrepresented in the subsample.

On the contrary, multiple subsamples can give a more representative picture, i.e., bootstrap aggregating, or *bagging* [41]. Bagging is an ensemble learning method designed to improve stability and accuracy of learn algorithms. It generates multiple learners using multiple bootstrapped samples and them aggregates them to get the final model. For classification problem the aggregation is done by a majority vote, while for regression problem the result is obtained by averaging or weighted averaging. The *bag of little bootstraps* (BLB) is a recent method that combines bagging with subsampling to achieve both computational efficiency and model adequacy [43]. Recent work in statistics [44,45] proves that, if M and S are correctly chosen (where M is the subsample size and S is the number of subsamples), the aggregated results retain theoretical properties such as consistency, and correct bias that may arise with a single subsample.

1.6 Stochastic Optimization

An optimization problem typically consists of an objective function, a set of decision variables and a number of constraints on the variables. Decision variables are tuned so that the objective function is maximized or minimized, while satisfying the set of constraints. The objective function and constraints can be linear (linear programming) [46,47], or the variables can be confined in discrete space such as integers (integer programming) [48,49], or the problem can be about optimizing convex functions over convex sets (convex optimization) [50]. For convex optimization problems the objective function can be differentiable or non-differentiable (non-smooth optimization [51]). For non-smooth optimization there is e.g. the sub-gradient method [52] which utilizes sub-derivatives to solve the problem. If the objective is convex, the sub-gradient method reduces to the steepest descent method [52].

All cases above can be classified as deterministic optimization problems, which use fixed parameters to describe the model. A harder class of optimization problems arises when there is uncertainty in the parameters, i.e., the *stochastic optimization* [53,54]. Many examples fall into this framework:

- In the newsvendor problem, the seller tries to determine the quantity of a daily order of newspapers to satisfy a random demand D, so as to maximize the revenue defined as dollar amount earned by selling the newspaper minus the cost of the order;

- In travel planning, the traveler needs to find a shortest path from a set of alternatives, when the observed traveling time for a particular path is subject to random changes such as traffic lights, weather condition or road constructions;
- In marketing campaign design, the manager wants to determine the best set of designs for the marketing tools (direct-mails, emails or webpages) to raise the maximum amount of revenue, the highest conversion rate or the greatest customer satisfactory. The outcome is highly uncertain.

In all the examples above, our decisions not only determine the value of objective functions, but also observations of unknown parameters. If we consider the distribution of unknown parameters as known, the distribution is fixed and new observations are viewed only as realizations of the underlying distribution. If, however, we are uncertain about the distribution, new observations ought to be used to change our beliefs about the distribution. In such case, one needs to optimize on the learning process as well as the objective function, i.e., how to choose a series of decisions so that the optimal solution can be found as quickly as possible. This is when the optimal learning comes into play [54].

One set of statistical learning methods uses Bayesian statistics to address this problem [55–58]. The initial belief about the random quantities can be represented as a prior distribution. When a new outcome is observed, Bayes' rule can be used to update the belief distribution to obtain the posterior distribution.

Another application of statistical learning in stochastic optimization is modeling complex objective function. One such example is stochastic dynamic programming (or reinforcement learning in the computer science community) [59,60], which addresses the problem of choosing an action given a state which generates a reward and takes us to the new state [54]. The Q-learning method [61] is one candidate for solving such problem, which uses a Q-function to approximate the true utility function. Another way to construct this approximation is by means of a regression model relating the value of a state to a relatively small set of inputs or parameters [62].

Chapter 3 provides a more detailed introduction about applications of optimal learning on ranking and selection (R&S), which is a sub-class of stochastic optimization that focuses on selecting the best alternative from a set of possible choices [63].

Chapter 2: Statistical Learning for Non-profit Fundraising

2.1 Overview

In this work we apply statistical learning on a non-profit fundraising problem for the American Red Cross (ARC). When a major disaster strikes (e.g. Hurricane Katrina or the Haiti earthquake), ARC experiences sharp spikes in one-time donations. These donations are coordinated for immediate disaster relief, as well as a wide variety of "development" programs, such as community disaster preparation, emergency response training, and sustainability efforts. However, fewer than 30% of one-time donors return to give a second time. The unpredictability of donor response limits managers' ability to plan long-term operations for programs that require steady funding [64]. In order to secure a consistent, reliable cash flow, ARC devotes significant efforts toward cultivating one-time "disaster donors" into long-term donors.

Cultivation is largely accomplished through direct mail, which accounts for about 2/3 of the total direct marketing budget in ARC. However, fewer than 50% of these are retained from one year to the next, so it is important to ensure that new donors are always being converted. Simply sending more mail may not be an effective way to achieve this goal [65], and in any case may not be feasible under a fixed budget.

In this work we study the problem of improving conversion and retention rates through the design attributes of the mailings themselves. “Design” can refer to the content of the letter, the presence or absence of a free gift (such as a set of address labels), various methods for setting suggested donation options, and other marketing strategies. From an organizational point of view, a non-profit manager may have records of millions of past communications with donors, but does not have access to detailed demographic information of donors. The operational decision of how to design and target a new fundraiser must be based on the information that is visible to the organization.

Our main contribution is an empirical analysis that identifies design attributes of direct-mail fundraisers that exert a significant impact on donor cultivation and retention. The context for our analysis is provided by a dataset jointly compiled by the Red Cross and Russ Reid Company during 2009-2011, for a cultivation program known as STAART (Strategy Through Applied Analytics, Research, and Testing). The dataset covers $49 million in donations to STAART from over 300,000 donors, with detailed campaign information available for over 8,000,000 individual recorded communications with over 1,000,000 individuals. Specifically, we have records of the characteristics of the outreach strategy used, such as the design or formulation of the mailed appeal; limited characteristics of individual donors, such as their previous donation amounts; and some characteristics of disasters such as their magnitude and location. We also consider *interactions* between design attributes of campaigns, thus accounting for potential heterogeneity of design effects by donor class.

We focus on four important design strategies used by the Red Cross:

- *Supporter cards* that affirms the donor's identity as a Red Cross supporter;
- Different ways of composing the *story*, or the written appeal included in a mailing, such as stories focusing on disaster preparedness or relief efforts;
- Free gift items such as address labels, glow sticks, etc;
- The technique of dynamically generating the suggested donation amounts based on the donor's previous behaviour (e.g. 75%, 100%, or 150% of the donor's most recent donation).

To summarize, we contribute to the literature on non-profit analytics by identifying designs that exert a significant impact on the outcome of a fundraising campaign, as well as key interactions between these designs and various donor segments. To our knowledge, this is the largest study to date on the interactions between *donors*, *disasters*, and *designs*. Our results (e.g., for preparedness stories and gift items) suggest ways in which cultivation campaigns (such as STAART) should be considered differently from other types of fundraisers. These insights lead to clear, simple policy recommendations; we conduct simulations that suggest that these recommendations have significant potential to improve fundraising efficiency.

2.2 Other related work

The subject of charitable donations has been studied in economics, marketing, and public policy. There has been relatively little work focusing on donor cultivation and retention, particularly from the operational perspective; we believe our work to be the first large-scale study of this type. Below, we survey the viewpoints of other communities and contrast them with our own.

Both theoretical and empirical studies of non-profit donations have often focused on the impact of donor income level, donor demographics, and policy decisions such as tax credits, on trends in charitable giving. Empirical work on this topic tends to rely on publicly available surveys on family expenditures [66, 67], which offer detailed income and demographic data on relatively small samples of households. One example of this demographics-oriented approach is a recent study based on the U.S. Panel Study of Income Dynamics, which surveys 5,000 families in the United States [68, 69]. See also [70] for a thorough demographic analysis of the market for charitable donations. Other recent work has presented evidence correlating donation amounts with other factors such as media coverage [71] and trends in the stock market [72].

The operations perspective has mostly considered revenue management and efficient resource allocation [73–75]. A recent work [76] studied how charitable motives can maximize the revenue generated in an auction. The literature also contains a number of theoretical models designed from the donor's point of view, e.g. seeking to optimally allocate resources to maximize a utility function [77, 78]. There is also a great deal of interest in behavioural drivers of donations. For example, how the prestige of a university affects alumni donations [79], how "foot-in-the-door" behaviour (e.g. asking potential donors to fill out a survey before asking them to donate) affects the likelihood of donor response [80], how donor motivation is affected by information about other donors [81], how the propensity of repeated direct mailings to irritate donors and negatively impact retention [82], or how donations are impacted by the promise of matching grants [83, 84]. Donor motivation is another important topic of research [85, 86], but is outside the scope of our study (and our data). We use the data to infer variation between donors, whether it stems from behavioural, demographic, or economic factors.

The literature has considered a number of theoretical econometric models for predicting donation amounts. One widely-used class of such models is known as the Tobit model [87], applied e.g. to investigate the effect of income and estate taxes on donations by [88]. This approach (see also the extension in [68]) is motivated by the particular structure of donations, which are always non-negative, and have a high incidence of zero values, because many surveyed households do not give to charity at all. In our setting, however, most individual donations to the ARC are fairly small, and the organization places high value on the *incidence* of donation (that is, the ability to reliably elicit a response), as opposed to the monetary amount. Additionally, while the organization can distinguish between individual donors, it does *not* have access to personally identifiable information (PII) about them (e.g. income or demographics). We rely on the data to establish the drivers of donor cultivation, treating the PII as an unobservable random effect.

The ARC has been the subject of extensive previous studies, for example, incentives in the context of blood donations [89], or consumer attitudes toward brands that partner with the organization [90]. Behavioural studies such as [91] have used Red Cross donations to provide a context for studying donor motivations.

Logistical issues faced by the organization have been studied e.g. in [92]. To our knowledge, however, our study is the first to focus on the effective design of direct-mail fundraisers, particularly in regard to donor cultivation. Our work is closer to [93], which also studies the incidence of donation for a direct-mail fundraiser (by a different organization), with a dataset covering 48,000 communications with 3,000 households, and a small number of basic design attributes such as the presence of a brochure or payment slip in the mailing. By contrast, we study a massive dataset with over 8 million communications and a rich set of donor and campaign attributes (up to 300 in the largest model we consider).

2.3 Description of STAART data

New donors enter the STAART database by contributing to a specific disaster relief campaign (e.g. after a major disaster), or by responding to an acquisition campaign. These new *acquisitions* subsequently receive mailed appeals encouraging them to continue their support. If a donor responds to such an appeal, he or she is said to be "converted," and is considered a current supporter of the program. Note that, in order to be converted, a donor must give at least two donations: the first donation identifying the donor as an acquisition, and a second donation in response to a conversion attempt. Once converted, a donor is regularly sent several types of mailings, broadly classified as *renewal* (direct appeals for a contribution), *cultivation* (newsletter-like mailings, primarily intended to build a relationship with the donor) and *follow-up* (other intermittent mailings). If the donor does not respond to any of these appeals for a period of 18 months, he or she is reclassified as *lapsed*. Some campaigns have a *generic* type, meaning that they are catch-all fundraisers targeted at all donors.

The Red Cross dataset consists of several large lists that separately catalogue communications, donations, donors, and disasters. We performed data processing to cross-reference and extract information from these lists. These data have a "layered" structure, such that increasingly smaller subsets of the data contain more detailed information (with finer granularity). In all, the dataset records 20.2 million (20.2M) individual communications with 1.3M different donors during 2006-2011. However, most of the information pertaining to fundraiser design is available for 8.6M communications taking place during 2009-2011, and we also have more detailed information about donors for $4.3M$ of these communications. Fewer than 10% of communications result in donations, and we also have multiple layers of data for these gifts. Table 2.1 shows how much of each type of information is available. Below, we describe the data in more detail.

Donors. Each donor is assigned a unique *account* number, so that we can always identify the specific donor with whom any given communication occurred. The *location* of a donor is represented in our study by the U.S. state associated with an account. Limited *affiliation* information is available, e.g. whether the donor is listed with a county or city chapter. At the same time, the Red Cross does *not* have access to personally identifiable information about the donors (such as demographic or income information). This generally holds for the entire non-profit industry.

Communications with donors are classified according to *campaign type*, which

Table 2.1: Amount of data available for various types of information.

Type	Amount	Donors	Content	Raw size	Full size
Communications	20.2M	1.3M	Account ID, location, affiliation Campaign types of communications	1.6 GB	—
	8.6M	1.2M	All design features	1.1 GB	3.0 GB
	4.3M	531K	Donor segmentation information	609 MB	2.4 GB
Gifts	819K	366K	Dates, amounts, payment methods	105 MB	—
	308K	193K	All design features	69 MB	103 MB
	169K	98K	Donor segmentation information	39 MB	110 MB
	89K	87K	Disaster type, magnitude, location	28 MB	49 MB
	6.9K	6.5K	Segmentation+disaster information	2 MB	6 MB

"Raw size" refers to the data obtained directly from the Red Cross; "Full size" also includes additional information derived from the dataset, described in Section 2.4.2.

reflects the status of a donor (acquisition, renewal, lapsed, etc.) at the time of the communication. Donors are further categorized according to *donor class*, a measure of how much they give per donation, which can be low ($\$10-\$99$), medium ($\$100 - \$499$), or high ($\$500 - \$9999$), with some additional classes such as "Haiti-influenced donors" representing connections to a particular disaster. We also have records of donor *recency*, which represents the time since their last donation (e.g. 0-6 months, 6-12 months, etc.). These two pieces of information are jointly referred to as *segmentation information.* Other donor classes may also be defined in connection with specific disasters, e.g. "Haiti-influenced donors" whose first donation followed on the Haiti earthquake.

To understand our results, it is important to bear in mind the specifics of the relationship between campaign type and donor recency. Table 2.2 shows the

Table 2.2: Breakdown of communications by campaign type and donor recency for several important campaign types.

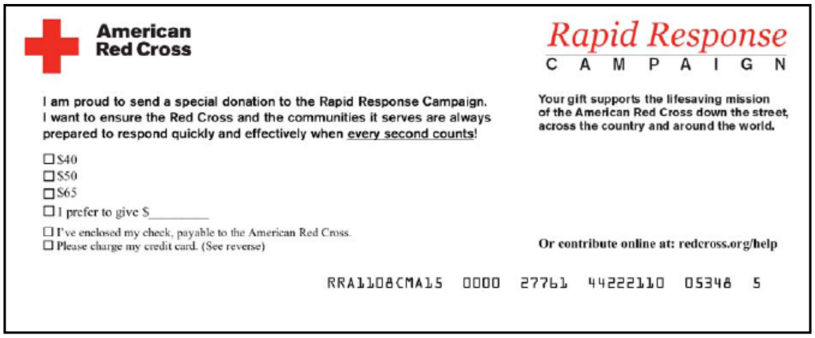
	0-6 mos.	7-12 mos.	13-18 mos.	19-24 mos.	25-36 mos.	37-48 mos.
Acquisition	52192	6313	1562	2657	0	125525
Cultivation	478665	98744	87008	3546	0	0
Lapsed	0	0	22009	17006	38158	0
Renewal	308236	407490	22366	11149	1370	770
Generic	1017854	350986	361570	79511	0	0

number of communications in each recency category, for five important campaign types. All donors in the Acquisition category have only made one disaster donation. As expected, many of them did so within the past six months, reflecting the effort to cultivate recent donors. However, a substantial group of communications in this category were targeted at donors who donated over 36 months ago, demonstrating a late effort by the Red Cross to reach out to donors who had never been cultivated. Among our current supporters (Renewal and Cultivation), many have made their last donation recently, but substantial proportions fall into the 7-12 and 13-18 month categories.

Donations. The *date*, *amount*, and *payment method* of every donation are recorded, as well as the *fund* receiving the donation. In the available data, 60 funds are associated with specific disasters (e.g. Iowa flood, Haiti earthquake, or Tohoku tsunami), for which we have records of *disaster type* (e.g. earthquake, flood, or hurricane), *magnitude* (death toll, cost in millions), and *location* (domestic or international).

Designs. Any given piece of mail is constructed with a set of design features. These include *personalization* of the mailing (inclusion of the donor's name and address), the presence or absence of various *donation options* (e.g. checkboxes for donating \$20 or \$30), and the possible inclusion of *gift items* such as mailing labels or a glowstick. A *supporter card* is included in 5.5% of communications. Cards are sent to the Renewal type, but cover all of the main donor and recency categories within that type; the gift items were sent to Acquisition and Lapsed types, again covering a variety of recency groups. A proportion of 64% of all communications offers the option to donate *online*. In 3.5% of communications, the donor has the option to *choose* the fund that will receive his or her donation. The *formulation* of the appeal is described, e.g. whether the letter mentions a specific disaster (about 40% of communications) or offers a generic story about disaster relief (10%), or whether it emphasizes disaster preparedness (50%). Figure 2.1 shows an example of a mailing with three suggested amounts of \$40, \$50 and \$65. These amounts may also be *dynamically generated*, i.e., the amount is calculated based on the previous donations.

We often have records of multiple communications with the same donor account. Figure 2.2 shows the empirical distribution of the number of communications



American
Red Cross

Rapid Response
CAMPAIGN

I am proud to send a special donation to the Rapid Response Campaign.
I want to ensure the Red Cross and the communities it serves are always prepared to respond quickly and effectively when every second counts!

Your gift supports the lifesaving mission of the American Red Cross down the street, across the country and around the world.

□ \$40
□ \$50
□ \$65
□ I prefer to give \$__________

□ I've enclosed my check, payable to the American Red Cross.
□ Please charge my credit card. (See reverse)

Or contribute online at: redcross.org/help

RRA1108CMA15 0000 27761 44222110 05348 5

Figure 2.1: Example of a rapid response mailing (publicly available [94]).

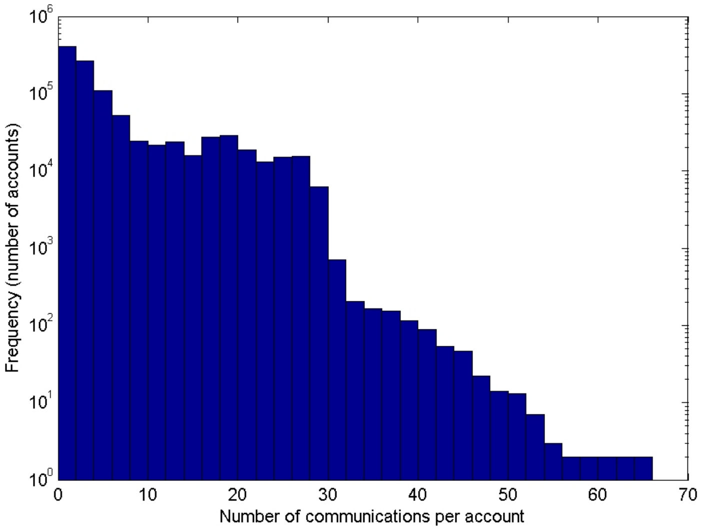


Figure 2.2: Empirical distribution (log-scale) of the number of communications per account.

per account, that is, the frequency of accounts for various numbers of communications. As expected, most accounts have fewer than 10 communications, but there are some with over 60. The decision to continue communicating with a particular donor is influenced by factors in our dataset such as the donor class and recency (high-class, recent donors are targeted more often). Note that class and recency are determined by the donor's most recent donation only; Red Cross analysts believe that this information is sufficient for deciding whether to target a donor. In general, the Red Cross also prefers to target donors who choose to give to *general* funds (such as "Where Our Need Is Greatest") rather than to specific disaster funds; however, this is not a major factor in STAART, since over 91% of all gifts in our data are made to general funds, and only 3.5% of mailings allow a choice of fund.

However, these communications may have different campaign types, reflecting the donor's transition from Acquisition to Renewal, or Renewal to Lapsed. The donor's segmentation information can also change over time. The design features can change with every communication as the organization experiments with new outreach methods. Note also that the outcome of a communication (*success* or *failure*, i.e. whether or not the communication elicited a donation) determines the *amount* of information that we receive. We can only observe detailed fund and disaster information for successful communications, where there is a record of money sent to a particular fund. Because the number of gifts is far smaller than the number of communications, the total volume of fund and disaster information is also relatively small; see Table 2.1 for the exact numbers. To leverage as much of the data as possible, we develop a separate model for each layer of data.

2.4 Methodology

We describe the methodology used to analyze the Red Cross dataset. In Section 2.4.1, we describe the basic statistical model that forms the foundation of our analysis. Section 2.4.2 discusses additional modeling and feature generation. Section 2.5 gives the full technical details of the estimation of the model on the data.

2.4.1 Statistical model and procedure

Let $i \in \{1, ..., I\}$ denote the ith donor account, with I being the total number of accounts. To reflect the longitudinal nature of the data, let the *panel size* N_i be the number of recorded communications with account i, and let $N = \sum_{i=1}^{I} N_i$ be the total number of communications. Also let y_{ij} denote the result of communication $j = 1, ..., N_i$ with account i.

A communication is "successful" ($y_{ij} = 1$) if it results in a donation. Otherwise, the communication is considered to be a failure ($y_{ij} = 0$). We begin with a mixed-effect logistic regression model, which assumes that

$$\mathbb{E}\left(y_{ij} \,|\, b_i\right) = g^{-1}\left(\mathbf{x}_{ij}^T\boldsymbol{\beta} + b_i\right), \tag{2.1}$$

where g denotes the logit link function. The p-vector $\mathbf{x}_{ij}$ represents the attributes of the jth communication with account i. This includes any relevant donor, donation and design information (see Section 2.3, Table 2.1) available for this communication. For example, a particular component x_{ijk} can be equal to 1 if the jth communication with account i included an option to donate online, and 0 otherwise.

The parameter b_i is a random effect [95]; we assume that $b_i \sim \mathcal{N}\left(0, \sigma^2\right)$, where σ^2 represents random variation between panels, and that the individual observations y_{ij} are conditionally independent given b_i. We include random effects in the model for several reasons. First, b_i can be used to represent unobservable variation in donor behaviour, specific to account i, and reduces statistical bias that arises when multiple observations come from the same source. Second, random effects reflect the fact that the donors in the dataset come from a larger population of donors, and the Red Cross continuously communicates with new individuals. Random effects thus allow us to reason about the entire population (and, potentially, new donors). Third, random effects allow for a much more compact model with only a single additional parameter σ^2, whereas adding a fixed effect for each account would add hundreds of thousands of attributes. Finally, modeling b_i as a random variable reflects the organization's considerable uncertainty about individual donor characteristics and behaviour.

For given β and σ^2, the joint probability of observing y_{ij}, $i = 1, ..., I$, $j = 1, ..., N_i$, can be written as

$$L(\boldsymbol{\beta}, \sigma) = \prod_{i=1}^{I} \int_{-\infty}^{\infty} \prod_{j=1}^{N_i} \left(\frac{e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i}}{1 + e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i}} \right)^{y_{ij}} \left(\frac{1}{1 + e^{\mathbf{x}_{ij}^T \boldsymbol{\beta} + b_i}} \right)^{1 - y_{ij}} \frac{e^{-\frac{b_i^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} db_i \qquad (2.2)$$

where the integral represents an expectation of a conditional probability given b_i. Then,

$$(\boldsymbol{\beta}^*, \sigma^*) = \arg\max_{\boldsymbol{\beta}, \sigma} \log L(\boldsymbol{\beta}, \sigma) \qquad (2.3)$$

represents the maximum-likelihood estimates of $\boldsymbol{\beta}$ and σ. The MLE optimization problem in (2.3) is typically solved using expectation-maximization (EM) algorithms [96] where the random effect terms are treated as missing data and estimated iteratively. However, this approach is intractable in our problem, because (2.2) is a product of I integrals (where I is on the order of 10^6), which cannot be expressed in closed form and must be evaluated numerically. Numerical methods such as Gaussian quadrature may be feasible for small I, N_i or p, but scale very poorly to large data [97, 98].

All else being equal, a concise model with a smaller number of features is preferable. Hundreds of features can be extracted from the Red Cross dataset, including those described in Section 2.3 and the interaction terms discussed later in Section 2.4.2. However, from a managerial viewpoint, it is preferable to focus on a small set of key drivers of campaign success, and from a statistical viewpoint, a smaller model reduces the risk of overfitting and is easier to generalize; also, extra attributes impose additional noise on prediction. To identify the most significant features, we use Lasso for model selection [99] and replace (2.3) by

$$\left(\boldsymbol{\beta}^{*},\sigma^{*}\right)=\arg\min_{\boldsymbol{\beta},\sigma}\left\{-\log L\left(\boldsymbol{\beta},\sigma\right)+\lambda\sum_{k}\left|\beta_{k}\right|\right\}, \tag{2.4}$$

where L is as defined in (2.2). The tuning parameter λ represents a price or penalty incurred if we choose a non-zero value for any β_k included in the final model. The penalty function $\|\boldsymbol{\beta}\|_1 = \sum_k |\beta_k|$ is non-differentiable around zero, causing β_k to shrink exactly to zero if the kth feature is found to be insignificant. See Section 1.4.4 for more discussion. Equation (2.4) balances the need for an accurate model with more predictive power against the need for a compact model with fewer features. By choosing λ carefully, we ensure that non-zero regression coefficients will only be assigned to attributes with a significant impact on model accuracy. See Section 2.5 for additional numerical results demonstrating the benefits of this approach.

The choice of λ is automated as follows. Let $\boldsymbol{\beta}\left(\lambda\right)$ and $\sigma\left(\lambda\right)$ be the choice of $\boldsymbol{\beta}$ and σ that solves (2.4) for given λ. Then, let

$$\mathcal{A}\left(\lambda\right)=\left\{k\,:\,\beta_k\left(\lambda\right)\neq 0\right\}$$

be the set of attributes identified by (2.4) as being significant. The size $\left|\mathcal{A}\left(\lambda\right)\right|$ is the number of features included in this model. We then solve

$$\lambda^{*}=\arg\min_{\lambda}\left\{-2\log L\Big(\boldsymbol{\beta}(\lambda),\sigma\left(\lambda\right)\Big)+\left|\mathcal{A}\left(\lambda\right)\right|\cdot\log N\right\}, \tag{2.5}$$

choosing the penalty term to minimize the well-known Bayesian Information Criterion (BIC) of [100]. Combining BIC with Lasso is a fairly widespread technique, and has been found to yield good practical performance on a variety of problems [101]. Other possible criteria from the literature include Akaike's Information Criterion (or AIC; see [102]) and cross-validation (see e.g. [103]). However, AIC-Lasso tends to include more non-zero predictors than necessary [104]. Furthermore, although the CV criterion is widely used in the literature, it has less theoretical support; a recent work [105] has obtained consistency results only for a restrictive class of models with orthogonal design matrices, which rarely occur with discrete data.

Finally, we remark on the additional challenge of estimating (2.4) on a large dataset. Model selection reduces the feature space, but (2.4) remains computationally prohibitive for large N. In such circumstances, a natural strategy is to take a random sample (of a tractable size) from the data, and use this "subsample" to estimate the model. We adopt this approach; however, the statistics literature [41,45] has demonstrated that using a single subsample can introduce bias into the model, as well as inflate the variance of the estimated coefficients. To mitigate this issue, we draw S small subsamples, thus obtaining S distinct Lasso models. Due to the noise introduced by subsampling, the set $\mathcal{A}(\lambda^*)$ of selected features varies across subsamples. To reduce this variability and ensure that only significant features are selected, we use a "majority vote," i.e., we include the kth feature in our model if it is selected in over 50% of the S Lasso models. Please see the Section 2.5 for the full technical details of this procedure.

Letting $\boldsymbol{\mathcal{A}}^*$ be the set of all such features, we can finally recompute

$$\arg\max_{\boldsymbol{\beta},\sigma} \log L(\boldsymbol{\beta},\sigma),$$

subject to the additional constraint that $\beta_k = 0$ for $k \notin \mathcal{A}^*$, yielding the optimal estimates of the significant regression coefficients. This step is known as "debiasing" or "post-Lasso," and has been demonstrated to eliminate bias from the Lasso estimator and produce more precise confidence intervals in settings such as least squares and quantile regression [106, 107]. Because the computational cost of this step is still prohibitive, we can perform debiasing on a new set of subsamples and average the results to obtain the final coefficients. Please see Section 2.5 for a detailed discussion.

2.4.2 Modeling and feature generation

In addition to the information already present in the data and described in Section 2.3, we generated additional attributes to address important statistical and modeling issues. To study the effectiveness of segment-specific fundraising strategies, we constructed numerous interaction terms and incorporated them into our model. In particular, we considered interactions between design features, such as the presence or absence of dynamic amounts, and donor features such as type (Acquisition, Renewal, etc.) and recency. This allows us to capture segment-specific effects, e.g. strategies that work better with new donors than lapsed donors. Interactions between donor features and the presence or absence of other donation options were also investigated. Model selection becomes crucial when considering interactions, as the number of two-way interaction terms grows quadratically with the number of features. The last column of Table 2.1 shows that the size of the data increases dramatically once the additional features have been generated.

Secondly, we generated additional control covariates to reduce bias due to unobservable correlations between communications. The bias is potentially due to correlation between measured and unmeasured, confounded, or "missing" features of each communication. Most notably, the behaviour of a single donor during the surveyed time period may be subject to time dependencies. A donor is unlikely to maintain the same level of contribution over time; rather one may expect the donor to lapse once the resources he or she has allocated for donation have been exhausted. From a behavioural standpoint, a donor who contributes frequently may simply be more motivated, or place higher value on pro-social activity. We control for the time factor as follows. For communication j with account i, we calculate 1) the number of previous communications with i, prior to the date of communication j; 2) the number of previous *successful* communications with i; 3) the number of previous communications *of the same type* with i; and 4) the number of different funds that have sent requests to the donor thus far. These attributes are included in $\mathbf{x}_{ij}$. Additional information on the time lapsed between donations is provided in the form of the recency attribute.

Another interesting research question deals with the dynamic generation of ask amounts. Donor class information, based on the size of the donor's last donation, allows us to control for the magnitude of the asked amount and separate the effect of dynamic generation from the effect of the precise amount asked for. We also control for any fixed ask amounts (donation options) that appear on the mailing.

We also validated our results through cross analysis, comparing the results of feature selection across different model types. For example, we compared the results of the model in Section 2.4.1, where each observation corresponds to an individual communication, and a different model where an entire campaign is viewed as a single observation, and the response variable is the success rate of the campaign. While the exact values of the estimated coefficients differ between models, the key managerial implications of the results are consistent throughout the study.

2.5 Challenge of massive data

From the point of view of traditional statistics, the logistic regression model described in Section 2.4.1 should always benefit from more data. From a purely theoretical viewpoint, a large sample size N is always a good thing, and theoretical issues arise only when $N < p$. However, in practice, the *estimation* of the model becomes computationally intractable when N is in the millions. We emphasize that the computational challenge arises, not from memory issues (various techniques and software packages, e.g. `biglm` in R, can be used to address that issue), but rather from the estimation of (2.1), which requires us to optimize an expensive, highly non-linear function.

Recall that (2.2) is a product of I integrals, where I is the number of unique donor accounts (over 1M in all). Furthermore, each integrand is a product of N_i logistic functions with a normal density, and thus is highly non-linear and non-convex. None of the I integrals has a closed-form solution; consequently, (2.2) can only be evaluated numerically, e.g. using Monte Carlo integration or Gaussian quadrature. Numerical integration introduces additional error into the evaluation of the likelihood function, and is also expensive for large I and N_i since each integrand must be evaluated multiple times. For these reasons, quadrature methods are infeasible for large problems, leading to both memory and convergence issues for expectation-maximization (EM) algorithms. This issue is well-known in the literature; for example, EM algorithms scale poorly to large datasets [98]. In our experience, the available computational procedures for solving (2.3) with random effects simply stalled, crashed, or otherwise failed to produce meaningful results.

With the advent of increasingly large datasets, the statistics literature has now begun to pay closer attention to large-sample data, where N is very large (in the millions) and p is moderately large (several hundred). Even with a large number of samples, such data may be vulnerable to noise accumulation, spurious correlations, and algorithmic instability [108]. Ideally, statistical methods for such data should be computationally tractable while retaining the theoretical guarantees of classical statistics (such as consistency). In order to scale up to the Red Cross dataset, we synthesize several emerging statistical methodologies, such as small-sample bootstrapping and stability selection, that yield both tractable and rigorous results.

As discussed in Section 1.5, the *Bag of Little Bootstraps* (BLB) method, which combines the idea of bootstrapping and subsampling, can be a useful tool to mitigate these issues. We draw S small subsamples, leading to S distinct, independently estimated Lasso models. Each subsample will produce different results: the number of selected features may vary across subsamples, and the set of selected features itself may vary. However, as we describe below, these results can be aggregated to obtain a single final set of accepted features, regression coefficients, and standard errors.

We separate the estimation procedure into two stages: we perform variable selection first, removing insignificant features to produce a model of reduced size, and then estimate random effects to correct for variation between donors. See e.g. [109] for a theoretical treatment of an approach separating fixed effect and random effect estimation. Both stages use subsampling to address big data issues.

Model selection. We perform subsampling in line with the technique of [43, 44] as follows. For each of S subsamples, we draw M communications with replacement from the complete dataset. The work by [43] recommends setting $M = N^{\gamma}$ for $\frac{1}{2} \leq \gamma < 1$, and obtains robust empirical results for $\gamma = 0.7$. For a dataset with $N = 8.6 \times 10^{6}$ communications, the size of a single subsample is $M \approx 71,500$. With regard to the number of samples, a common technique [99] is to use $S = \frac{N}{M}$, or approximately $S \approx 120$.

We then perform model selection as in Section 2.4.1, replacing N by M in (2.5); however, as long as $M > p$, BIC preserves its theoretical consistency properties, which means that it will still correctly identify significant features [110]. To aggregate the results, we use a version of the stability selection criterion of [111] as follows. Each subsample $s = 1, ..., S$ produces a different solution λ_{s}^{*} of (2.5), and a different acceptance set $\mathcal{A}(\lambda_{s}^{*})$. Intuitively, the kth feature is more likely to be significant if it is selected by a larger number of these subsets. We include the kth feature in our final model if

$$\frac{1}{S}\sum_{s=1}^{S} 1_{\{k\in\mathcal{A}(\lambda_s^*)\}} \geq \rho, \tag{2.6}$$

that is, the proportion of samples in which k is selected exceeds a threshold $\frac{1}{2} < \rho < 1$. Note that the extreme cases $\rho = 0$ and $\rho = 1$ correspond to the union and intersection, respectively, of the sets $\mathcal{A}\left(\lambda_s^*\right)$. Let $\mathcal{A}^*$ be the set of all k for which (2.6) holds.

Estimation. To correct for unobserved variation between donors, it is necessary to refit the random effects model of (2.1) with the additional constraint that $\beta_k = 0$ for $k \notin \mathcal{A}^*$ (as proved in [107], this also corrects bias in the regression coefficients). However, even with this reduction in the size of the model, (2.3) remains prohibitively expensive to compute for the entire dataset. Again, we approach this problem through subsampling. To preserve the longitudinal structure of the large dataset across all subsamples, we now use entire panels as the unit of sampling. We modify the BLB technique to include importance sampling from the empirical distribution of the number of communications per panel (shown in Figure 2.2).

Formally, this is done as follows. Let $M' = I^\gamma$ be the number of donors included in each subsample, and let $S' = \frac{I}{M'}$ be the number of subsamples generated. A single subsample is created by simulating M' realizations of a discrete random variable Z with pmf

$$P\left(Z = i\right) = \frac{N_i}{\sum_{i'=1}^{I} N_{i'}}.$$

Let $Z_1, ..., Z_{M'}$ denote these M' sampled values. For each $m' = 1, ..., M'$, if $Z_{m'} = i$, we add N_i communications $y_{i,1}, ..., y_{i,N_i}$ to the subsample. In this way, a particular panel has a higher probability of being sampled if it contains more communications. Furthermore, if a panel is sampled, we automatically add every communication in that panel to the subsample, thus preserving the longitudinal nature of the data.

It remains to obtain a single set of estimated coefficients from the results of subsampling. We reoptimize (2.3), subject to $\beta_k = 0$ for $k \notin \mathcal{A}^*$, independently on each of the S' new subsamples. Let $\hat{\beta}_{k,s'}$ be the estimated coefficient of feature k returned by (2.1) on subsample $s' \in \{1, ..., S'\}$. We calculate

$$\bar{\beta}_k = \frac{1}{S'} \sum_{s'=1}^{S'} \hat{\beta}_{k,s'}$$

and report this as our final estimate of the effect of feature k. In words, we aggregate the results of subsampling by simply averaging the estimated coefficients across subsamples. Under available consistency results for subsampling, this average should converge to the true coefficient β_k with enough subsamples. Then, we let

$$\hat{\sigma}_k^2 = \frac{1}{S'-1} \sum_{s'=1}^{S'} \left(\hat{\beta}_{k,s'} - \bar{\beta}_k \right)^2 \tag{2.7}$$

be the sample standard error of the regression output across subsamples. We then use $\frac{\bar{\beta}_k}{\hat{\sigma}_k}$ as the relevant t-statistic, with $S' - 1$ degrees of freedom, for the null hypothesis that $\beta_k = 0$. Standard techniques can be used to calculate a p-value.

We briefly discuss the choice of (2.7) to calculate standard errors. Notice that (2.7) is calculated based only on the estimated coefficients in our subsamples, not on the estimated standard errors produced by the regression model within each subsample. A recent work by [112] has argued that these within-subsample standard errors do not contribute to the asymptotic standard error of the aggregated estimator $\bar{\beta}_k$. Moreover, (2.7) over-estimates the true variance, meaning that $\hat{\sigma}_k^2$ will produce conservative confidence intervals [112]. For our purposes, this conservative estimator is sufficient to evaluate the significance of our results.

To summarize, we analyze the massive Red Cross dataset by separating statistical estimation into two stages. The first stage selects the most important features by conducting Lasso-type regularization on each bootstrapped subsample, then aggregating the results with stability selection. The second stage removes all features $k \notin \mathcal{A}^*$, and corrects for the unobserved variation between donors by estimating random effects in this reduced model. In addition to the theoretical advantages of aggregation, we can see from Figures 2.3(a)-2.3(b) that our approach empirically produces more conservative feature sets – there is clearly a small core of features that are "agreed" on by a majority of subsamples, but there are also clear outliers in the "tails" of the histograms that are selected in a very small proportion of subsamples (or in just one subsample).

Numerical illustration. We briefly illustrate the advantages of GLMM-Lasso with subsampling over a rougher but simpler technique, namely ordinary logistic regression (LR), in terms of two standard performance metrics (see, e.g., [113] for details). We compare these methods using 5-fold cross-validation (CV), a common technique in data mining for evaluating the predictive power of a model. First, we compare the deviance residuals achieved by both methods (averaged over the 5 folds in CV). The comparisons are carried out individually on 10 different subsamples, each of size N^γ. (As we discussed earlier, it is always necessary to run models on small subsamples in order to tractably obtain results.) The logistic regression model does not perform any model selection; thus, the results illustrate

Table 2.3: Deviance residuals of GLMM-Lasso with subsampling vs. plain logistic regression, demonstrated on 10 random subsamples.

Subsample	Plain LR	GLMM-Lasso
1	100.2518	12.30325753
2	41.28834	12.61857265
3	57.07833	12.42619023
4	74.61048	12.8426679
5	86.99473	12.73456152
6	73.44249	12.28891669
7	145.982	12.39356766
8	53.33318	12.2563628
9	79.01672	12.31632096
10	30.35353	12.27578767

the benefits of using a more parsimonious model with fewer features.

Table 2.3 presents the results of this comparison. Our model outperforms LR (achieves lower deviance) in each subsample. The results are also much more consistent for the aggregated Lasso model (LR fluctuates more across subsamples), suggesting that there is significant benefit in aggregating over multiple subsamples to reduce variance. Recall also from Figures 2.3(a)-2.3(b) that aggregation leads to more conservative results: by eliminating outlier features that are not selected by a majority of subsamples, we reduce the risk of over-confidently reporting significance.

Next, we compare the area under the ROC curve for both methods. This metric is widely used as a measure of accuracy when the data has binary responses with a small proportion of 1s (as is the case in our application). Results for 10 subsamples are given in Table 2.4. The Lasso model consistently outperforms LR (achieves higher AUC). Furthermore, LR generally has poor predictive power (AUC close to 0.5).

These results are quite consistent with what is known about Lasso in the liter-

Table 2.4: Area under the ROC curve for GLMM-Lasso with subsampling vs. plain logistic regression, demonstrated on 10 random subsamples.

Subsample	Plain LR	GLMM-Lasso
1	0.51984	0.70489
2	0.52615	0.70022
3	0.50530	0.69283
4	0.51175	0.70806
5	0.53812	0.69613
6	0.51591	0.70018
7	0.53985	0.71100
8	0.52010	0.68855
9	0.52511	0.69241
10	0.52399	0.69920

ature. Classical models, such as logistic regression, estimate coefficients for a large number of features that Lasso simply removes from the model. Consequently, any prediction made by such models is subject to a much higher level of noise. Even if a plain LR model were to accurately estimate some of the coefficients, these accurate estimates are essentially drowned out by a large number of inaccurate estimates for other features. This issue, known as "noise accumulation," is quite common; for example, the performance of LR is often no better than random guessing in the presence of noisy data [108]. Furthermore, simple linear models may produce over-inflated standard errors when the data is subject to a high degree of empirical correlation, an issue discussed in Section 2.6.3. In such settings, the p-values produced by LR may themselves be unreliable [114], while Lasso is known to be less vulnerable to this issue. These examples illustrate the benefits offered by model selection in analyzing large datasets.

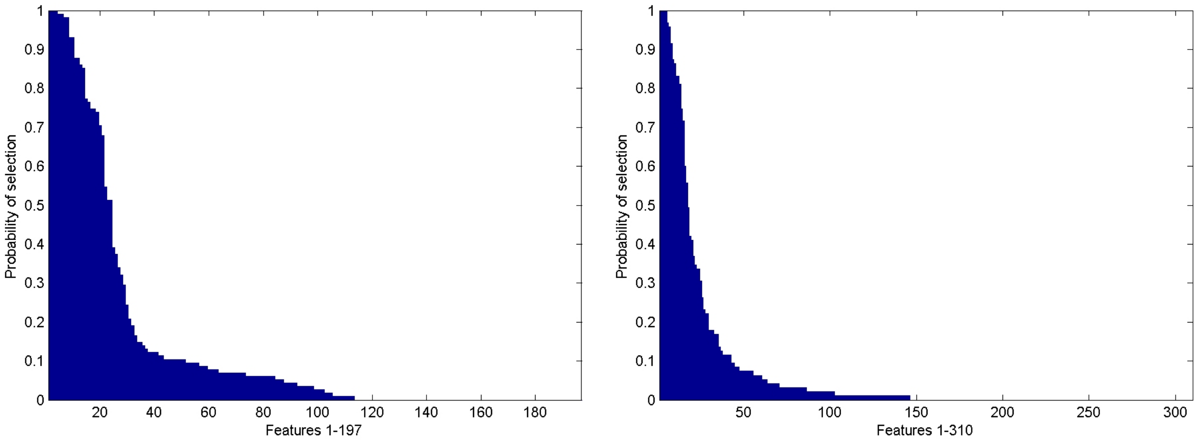
2.6 Results

Sections 2.6.1 and 2.6.2 discuss the logistic regression model described in Section 2.4.1, used to predict the success probability of an individual communication. Sections 2.6.3 and 2.6.4 present additional analysis of campaigns and donations, respectively. Finally, Section 2.6.5 lays out simulation results illustrating the potential of our key insights to improve conversion rates.

2.6.1 Communication-based models: design information

We begin with the model from (2.1), where y_{ij} is the outcome of the jth communication with donor i, and g is the logistic link function. Recall from Table 2.1 that the data have a layered structure. We apply this model to two layers. The first layer uses 8.6M communications from 2009-2011 for which design information is available; the second layer uses 4.3M communications, but adds segmentation information. We do not consider the outermost layer of 20.2M communications, because it lacks the crucial dimension of design information. We also do not consider the innermost layer of 89K communications, because fund information is only available for donations (i.e. successful communications), and is thus unsuitable for predicting the success *probability* of a communication; however, we will return to the issue of fund information in Section 2.6.4.

The first layer, covering 8.6M communications, gives us a total of $p = 197$ features, comprised of campaign types (binary features indicating Acquisition, Cultivation, Lapsed, etc.), design features, donor locations, and the additional covariates



(a) Base model (Section 2.6.1).

(b) Model with segmentation data (Section 2.6.2).

Figure 2.3: Ranking of features, in descending order of the proportion of subsamples where each feature was selected.

and interaction terms described in Section 2.4.2. Due to space considerations, we do not list all of these features here; rather, we list (in Table 2.5) the most significant features identified by the model selection procedure from Section 2.4.1.

We draw $S = 120$ subsamples (see the Section 2.5 for a discussion of how S is calculated) by Monte Carlo sampling with replacement from the large dataset. We run model selection separately on each subsample, and include a feature in the final model if it is selected in a sufficiently high proportion of subsamples. We view this proportion as the empirical probability of selecting a feature. Figure 2.3(a) shows the 197 features ranked in descending order of selection probability. We see that over 40% of these features are not selected in any subsample, suggesting that they can be safely eliminated from our model. Moreover, fewer than 15% of features are selected in over half of all subsamples.[1] Table 2.5 lists the top 20 ranked features,

[1] In this way, we obtain more conservative results by using multiple subsamples rather than just one. Aggregation reduces the risk of over-confidently reporting a feature as being significant, when

Table 2.5: Final estimated coefficients for first layer (8.6M communications).

Rank	Feature	Avg. coefficient	Std. deviation	*p*-value
1	Intercept	-3.8329	0.4636	<1e-30
2	Previous successes	0.6623	0.0889	8.2531e-25
3	$50 option	0.6330	0.0990	1.3843e-14
4	Year 2009	0.7559	0.0950	<1e-30
5	$20 option/generic type	-1.9487	0.3503	1.1276e-10
6	$15 option	-0.3373	0.0681	2.5277e-8
7	Allow choice of fund	-1.8780	0.2834	6.1010e-17
8	Dynamic amt./Renewal type*	0.0810	0.0760	0.1473
9	Dynamic amt./Acquisition type	-2.1242	0.3451	1.0304e-13
10	Dynamic amt./$50 option	4.8331	1.0783	1.0715e-6
11	Dynamic amt./Lapsed type*	-3.0032	1.4797	0.0212
12	Generic story/generic type	-0.9814	0.1294	1.2341e-28
13	Supporter card	0.2881	0.0642	1.0269e-6
14	Donor city indicated	-0.1823	0.0561	4.1723e-4
15	Renewal type	0.2693	0.0891	1.1323e-3
16	Donor city/Acquisition type	0.1729	0.0651	3.8793e-3
17	Preparedness story	0.3406	0.0495	3.0188e-18
18	$50 option/Renewal type	-5.1345	1.1209	3.8085e-7
19	$30 option/$50 option	1.6980	0.2131	3.6301e-37
20	Dynamic amt./Cultivation type	-2.6663	0.6355	4.2092e-6

All features are significant at the 0.01 level except those marked with an asterisk (*).

of which half are interaction terms. The technical details for the computation of coefficients and standard errors are given in the Section 2.5.

Managerial insights. The results provide immediate insights into the effectiveness of supporter card, formulation of appeal and gift items on campaign success. Feature 13 shows that the presence of a supporter card exerts a positive impact on the odds of success for an individual communication (as hypothesized). Features 12 and 17 suggest that a generic story (that is, a story describing disaster relief

in fact it may be an outlier whose apparent significance is only due to the additional noise induced by subsampling.

without reference to a specific disaster) performs poorly, while a preparedness story has a positive effect. Also, no gift items (mailing labels or glowsticks) are selected in the final model, suggesting that these gifts do not exert a significant effect on the outcome. The model also does not select features representing the inclusion of a personal disaster preparedness checklist, or a photograph depicting people being helped.

The results suggest that mailings are more effective when they focus on disaster *preparedness*, rather than on post-disaster relief efforts. This result is somewhat surprising, as it runs counter to the conventional wisdom (even among some prospective donors, surveyed by [115]), that more visceral, "emotive" imagery translates to more donations. Empirical studies such as [116] and [117] have also found evidence that emotive content is more likely to elicit donations.

One possible explanation is that donor *cultivation* programs, such as STAART, are targeted at a specific audience, whose characteristics differ from that of the broader pool of prospective donors. In this light, an interesting connection can be made to a study of hospice donations by [117], which found evidence that emotively designed webpages led to a higher overall volume of online donations. However, among 239 donors who submitted in-depth information about their motivations, the authors also identified a group of 101 donors, who had a history of giving to charity, and who also reacted better to informative rather than emotional content. Similar results were observed by [118] for donations to a homeless shelter.

The donors in the STAART dataset have all made at least one donation prior to their inclusion in the program, and thus all have at least some experience making donations. Unsurprisingly, we also see that a longer history of giving (represented by feature 2) contributes positively to the outcome. However, more informative, preparedness-oriented content contributes an additional positive effect, while more visceral stories and even visual images of people being helped appear to have no impact, or even a negative one. Additionally, a supporter card emphasizing the donor's identity as a conscientious supporter also contributes positively. Thus, we do not argue that preparedness-oriented appeals will attract more donations in every setting; rather, we argue that such appeals are more effective in the specific setting of cultivating and retaining donors.

The results also provide some initial insight into the last question in Section 2. Recall that dynamic generation of ask amounts occurs in 44% of all communications, making it a significant component of the organization's strategy. Table 2.5 suggests that the impact of this strategy is dependent on the campaign type. The effect appears to be positive for the Renewal type, representing current supporters of the program, and negative for the Acquisition and Lapsed types, representing new and lapsed donors, respectively. However, features 8 and 11 have high standard errors. Furthermore, Table 2.2 shows that donors in these types exhibit substantial heterogeneity with regard to their recency. We examine this issue in more detail in Section 2.6.2, where we consider a smaller but richer layer of the dataset.

Table 2.6: Final estimated coefficients for second layer (4.3M communications).

Rank	Feature	Avg. coefficient	Std. deviation	p-value
1	Intercept	-3.2204	0.0336	<1e-30
2	Previous successes	0.1298	0.0065	6.6399e-29
3	Year 2010	-0.4265	0.0253	6.6727e-25
4	$15 option/$20 option	-2.0118	0.1816	1.5356e-16
5	0-6 mos. recency/Low donor class*	0.0814	0.0398	0.0226
6	Supporter card	0.5744	0.0449	3.0471e-19
7	Generic story	-0.7580	0.0519	6.9461e-22
8	Haiti-influenced donors	-0.3055	0.0389	3.9877e-11
9	Dynamic amt./0-6 mos. recency	0.1100	0.0307	3.4019e-4
10	Acquisition type/0-6 mos. recency	0.7042	0.1317	7.1230e-7
11	Preparedness story	0.4060	0.0359	6.5999e-17
12	37-48 mos. recency	-0.2995	0.1124	4.9178e-3
13	Specific disaster story	-0.5899	0.0336	7.9810e-26
14	13-18 mos. recency	-0.2742	0.0498	3.8371e-7
15	Allow choice of fund/0-6 mos. recency*	0.0878	0.1579	0.2902
16	Dynamic amt./Renewal type	-0.1774	0.0535	7.7752e-4
17	Renewal type/Low donor class	0.2041	0.0543	1.9305e-4

All features are significant at the 0.01 level except those marked with an asterisk (*).

2.6.2 Communication-based models: design and donor information

The next layer of data uses 4.3M communications, but adds segmentation information in the form of two groups of features representing donor class and recency. After adding interaction terms between these new features and the design information available from before, the total number of possible features in our model is $p = 310$. However, Figure 2.3(b) shows that, once again, only a small number of these features is consistently identified as significant. Indeed, only 17 features were selected in at least 50% of subsamples; these are listed in Table 2.6 with aggregated estimates, standard errors, and p-values.

Managerial insights. Most importantly, Table 2.6 corroborates our previous findings in Section 2.6.1. Both supporter cards (feature 6) and preparedness-oriented stories (feature 11) carry significant positive effects. Unsurprisingly, a specific disaster story works better than a generic disaster story (features 7 and 13). What is more surprising (but in line with our interpretation from before) is that both specific and generic stories carry negative effects. Additionally, a special class of donors whose first contribution was influenced by the Haiti disaster (feature 8) exhibits a significant negative effect. We note that the Haiti disaster received a great deal of media attention, and thus this donor pool may contain more impulsive donors who value emotive over informative content, and may be unlikely to convert into regular supporters.

Dynamic amounts present a more complex issue. We see that this strategy now appears to produce a negative effect when applied to the Renewal type (feature 16), which seems to contradict the findings of Table 2.5. At the same time, the same strategy produces a positive effect for the "0-6 mos. recency" category, which contains donors from multiple campaign types (Table 2.2). Feature 10 also suggests interactions between campaign type and donor recency. To clarify this issue, we ran a version of the model in which features 9, 10, and 16 were replaced with three-way interaction terms. The estimated coefficients for these features are given in Table 2.7. For the other features in the model (carried over from Table 2.6), the estimated coefficients changed slightly in magnitude, but kept the same signs as before, so we omit them out of space considerations.

The results indicate that dynamic amounts are effective for recent donors (0-6 mos. recency) who have not yet converted (Acquisition type). Earlier, in Table 2.5, the strategy appeared to work poorly on the Acquisition type; however, Table 2.2 shows that 2/3 of the communications in this type actually targeted donors with a very long recency (37-48 mos.). From Table 2.6, we see that these donors, unsurprisingly, do not respond, leading to an overall negative effect on the Acquisition type. However, if we consider only those unconverted donors whose first donation was made in the past six months, we see that dynamic amounts have a significant positive effect.

By contrast, the strategy exhibits a negative effect for the Renewal type, particularly on less recent donors (7-12 month recency). For recent new donors, who have made their first disaster donation within the past six months, there is an opportunity to "strike while the iron is hot" by offering them the chance to replicate their behaviour, this time in the role of "Red Cross supporter" rather than "disaster donor." However, for donors who have made a second donation and already converted into the STAART program, this approach is no longer effective.

We note that all dynamic amounts in use by the Red Cross use a scale that includes, or is close to, 100% of a donor's most recent donation. We do not argue against all possible dynamic scales. However, the evidence suggests that it is

Table 2.7: Final estimated coefficients and standard deviations for three-way interactions.

Feature	Avg. coefficient	Std. deviation	*p*-value
Dynamic amt./Renewal type/0-6 mos. recency*	-0.0880	0.0543	0.0549
Dynamic amt./Renewal type/7-12 mos. recency	-0.2177	0.0645	6.4953e-4
Dynamic amt./Acquisition type/0-6 mos. recency	0.7822	0.1284	4.0931e-8
Dynamic amt./Generic type/0-6 mos. recency*	0.0463	0.0370	0.1079

Features are significant at the 0.01 level unless marked with an asterisk (*).

unrealistic to use dynamic amounts that essentially ask a donor to maintain the same donation amount in the long term. It may be useful to experiment with other scales that, for example, use much lower ask amounts for longer donation histories. The main managerial implication of our results is that the best chance to influence donors to repeatedly give the same amount occurs at the very beginning of their donation history, before conversion occurs.

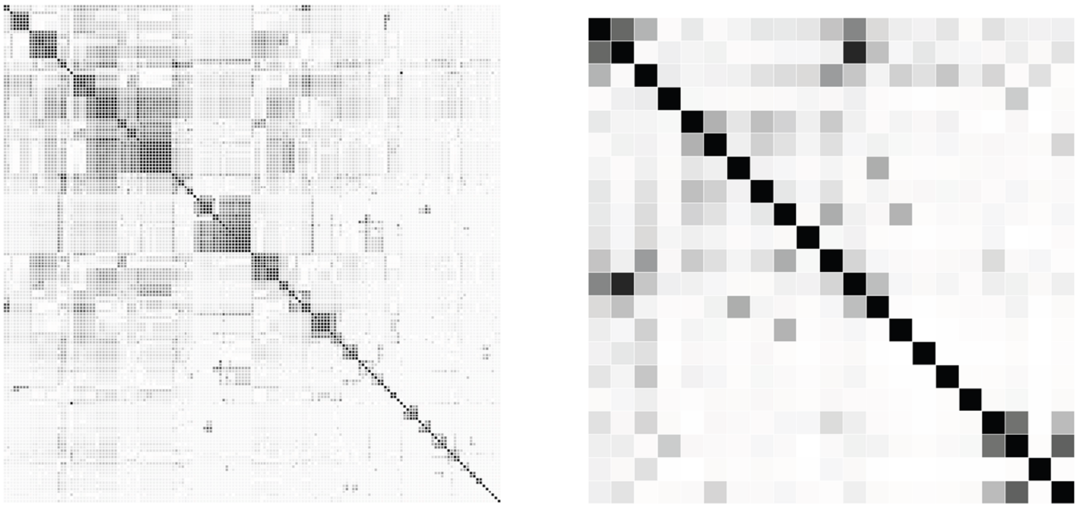
Finally, we briefly note that both Tables 2.5 and 2.6 indicate that, all else being equal, there were fewer donations in 2010 than 2009. This is particularly clear in Table 2.6, where "Year 2010" has a strong negative effect, while no other year is even selected. It is interesting to note that the stock market performed especially poorly in 2009. [72] found a positive correlation between stock market performance and charitable donations, but observed that the effect appears to be lagged; under this model, poor stock market performance in 2009 would be expected to lead to fewer donations in 2010, providing a possible explanation for the pronounced negative effect of 2010 in Table 2.6.

2.6.3 Campaign-based models

We constructed a different set of models that considered the data in Sections 2.6.1 and 2.6.2 from another viewpoint. These models aggregate the set of communications by campaign. Again, we use the logistic model in (2.1), but now y_{ij} represents the success rate of campaign i on the jth donor segment. Success rate is expressed as the ratio of the number of successful communications (resulting in donations) to the total number of communications in the campaign. For example, if the organization mailed 100 copies of a letter to a certain class of donors, and received 6 gifts in response, the success rate of the letter is 0.06 for that donor segment. A simple modification of (2.1) allows us to consider continuous-valued observations between 0 and 1. The main purpose of this analysis is to corroborate the results obtained in Sections 2.6.1 and 2.6.2 and demonstrate that similar results emerge without the subsampling techniques described in the Section 2.5. Our discussion focuses on the second model (Section 2.6.2, segmentation information included); we show that the relationships observed in this model also hold when the data are reorganized for campaign-centric analysis.

The attributes $\mathbf{x}_{ij}$ of the ith campaign and jth segment are largely the same as in Sections 2.6.1 and 2.6.2, and interaction terms are constructed as in Section 2.4.2. However, we are not able to include features that are not in one-to-one correspondence with campaign segments. For example, donor location varies on the level of individual communications, as the same letter can be mailed to people in different states. With recency and donor class included in the model, the total number of features was $p = 157$, with $I = 60$ panels (campaigns) and $N = 952$ campaign segments in all. The size N_i of each campaign ranges up to 132 segments.

The relatively small size of this dataset allows for a tractable analysis on its entirety, without the need for small subsamples. At the same time, aggregation across campaigns leads to the new problem of inflated empirical correlation, stemming from the relatively small magnitude of N relative to p. Figure 2.4(a) shows empirical correlations between all 157 features; in particular, the dark blocks visible



(a) Heatmap of 157 features.

(b) Heatmap of 21 selected features.

Figure 2.4: Empirical correlations of (a) all features and (b) selected features. Darker colours represent heavier correlation (closer to 1).

in Figure 2.4(a) show that, for certain groups of features, the empirical correlations are close to 1. This issue complicates statistical analysis, as the matrix $\mathbf{x}^T\mathbf{x}$ is not invertible. Random effect models are even more sensitive to correlation, as the number of panels is even smaller than the sample size. Model selection solves this issue by reducing the number of features from 157 to just 21, not counting the intercept. Figure 2.4(b) shows that these selected features exhibit much lighter correlation. Table 2.8 shows the final results for this model, ranking the selected features by p-value.

Managerial insights. It is most relevant to compare Tables 2.6 and 2.8, because they both include segmentation information. With this in mind, we see that Table 2.8 reproduces our key findings from Section 2.6.2. Most crucially, we observe identical insights on relief vs. preparedness: both generic (feature 9) and specific (feature 6) disaster stories carry negative effects. By contrast, preparedness-

Table 2.8: Final estimated coefficients for the campaign-based model with segmentation information.

Rank	Feature	Estimate	Std. error	p-value
1	(Intercept)	-3.12390	0.07434	<1e-30
2	$15 option/$20 option	-1.40281	0.12573	3.1995e-29
3	0-6 mos. recency/Low donor class	0.55805	0.07598	1.0679e-13
4	13-18 mos. recency	-0.40337	0.06540	3.4145e-10
5	7-12 mos. recency/Low donor class	0.50571	0.08343	6.8060e-10
6	Specific disaster story	-0.38501	0.07145	3.5228e-08
7	Card	0.66038	0.18892	2.3262e-04
8	Includes specific fund/37+ mos. recency	-0.79300	0.26354	1.3062e-03
9	Generic story	-0.33655	0.13033	4.9400e-03
10	Preparedness story	0.21504	0.08808	7.3436e-03
11	Dynamic amt./0-6 mos. recency	0.22205	0.10655	0.0187
12	Dynamic amt./Lapsed type	-0.86614	0.43200	0.0227
13	Allow choice of fund/0-6 mos. recency	0.61133	0.33262	0.0328
14	0-6 mos. recency/Haiti-influenced donors	0.18883	0.11368	0.0484
15	Followup type	0.25435	0.17342	0.0707
16	Includes specific fund/7-12 mos. recency	-0.50035	0.44334	0.1292
17	0-6 mos. recency/High donor class	0.17362	0.16390	0.1445
18	Dynamic amt./Renewal type	-0.18834	0.20609	0.1814
19	19-24 mos. recency/Haiti-influenced donors	0.15064	0.21575	0.2419
20	Renewal type/Low donor class	-0.13402	0.19576	0.2482
21	Option to donate online/High donor class	0.11607	0.19923	0.2809
22	Renewal type	-0.01884	0.15227	0.4522

"Std. error" refers to the usual statistical standard error of an estimated coefficient. The first 14 features were significant at the 0.05 level.

oriented campaigns (feature 10) have significantly higher success rates. Furthermore, supporter cards (feature 7) continue to contribute to campaign success.

We also considered a campaign-centric version of the model from Section 2.6.1 (design information, but no segmentation). We do not give the full details here for space considerations, as they mostly repeat our previous discussion. However, we briefly note that this model produced the same results with regard to preparedness vs. relief, as well as supporter cards.

2.6.4 Gift-based models

We also considered another set of models incorporating donation and disaster information. That is, we link an incoming donation to a specific disaster with attributes such as type (flood, earthquake, etc.), location (foreign or domestic), and magnitude (e.g. death toll). Each donation is associated with a monetary amount and a payment method, which we have also not discussed up to this point. We mostly obtain the negative result that disaster-specific information does not significantly affect donations, and include the discussion below for completeness.

Donation and disaster information is difficult to include in our previous models, because it can only be observed for successful communications (those that result in donations). Russ Reid has confirmed that there is no way to connect an unsuccessful communication to a specific fund. While the literature offers models for handling "donations" of size zero, our situation is more complicated because we also observe additional information (extra features) when a communication succeeds that is not observable if the communication fails. We chose to conduct a separate analysis that is confined to gifts only, removing all unsuccessful communications. That is, we return to equation (2.1), but now define y_{ij} to be the dollar amount of the jth *gift* contributed by the ith account. The link g is chosen to be the identity function, corresponding to a linear regression model.

Fewer than 5% of all communications are successful, which drastically reduces the model size. Table 2.9 shows the sizes of models fit to different layers of data; no model is large enough to require subsampling. For space considerations, we do not

Table 2.9: Sizes of gift-oriented models.

Layer	Size	Features	Interactions	Total	No. selected	0.05 level
1: All gifts, 2009-2011	309,451	96	33	129	52	34
2: Segmentation only	168,588	104	163	267	68	51
3: Disaster only	89,529	103	104	207	23	17
4: Segmentation+disaster	6,908	108	228	336	23	14

list the full results from all four models here, but we highlight the main points from all four models. Note that, due to the smaller size of these models, fewer features are statistically significant at the 0.05 level. The last column of Table 2.9 counts these features for each layer.

Disaster attributes appeared to have no significant impact on donation amounts. Of the 14 features in Layer 4 that were statistically significant at the 0.05 level, only one involved a disaster attribute. This was the interaction "Earthquake/High donor class." It is unsurprising that the high donor class should have a strong positive correlation with donation amount, as this class includes the largest gifts, up to $9999. As for the earthquake attribute, we note that it applies to the Haiti disaster, which was widely publicized and led to a high volume of donations.

Of the 34 statistically significant features in Layer 1, 18 corresponded to different donor locations (represented by U.S. state). Potentially, this suggests that there may be regional differences in donation amounts (though not in campaign success rates; see Sections 2.6.1-2.6.3). However, as segmentation and disaster information was added in Layers 2-4, the number of statistically significant locations shrunk to 6/51 in Layer 2, 9/17 in Layer 3, and just 2/14 in Layer 4. On this basis, we argue that the impact of campaign design and donor segmentation is much greater, and more important for policy decisions, than the possible impact of regional differences.

Design attributes seemed to have relatively little impact on donation amounts. Layer 4 contains only a single significant feature involving such an attribute, the interaction "Dynamic amt./High donor class," with a significant positive effect. However, there are several significant features involving the high donor class, all with strong positive effects, suggesting that the effect is more likely due to the fact that donors in the high donor class simply give more money to begin with.

Managerial insights. Most of the statistically significant features selected in Layers 1-4 were related to donor attributes such as recency and low/medium/high class, in contrast with our results from Sections 2.6.1-2.6.3. This suggests that a well-designed appeal may get more donors to respond (increasing the campaign success rate or the probability of receiving a donation from a particular donor), but the *amounts* of their donations are largely determined by immanent donor characteristics such as the donor class. Of course, this result should be considered in the specific context of *cultivation* campaigns.

2.6.5 Simulation results

We conducted a simulation study to quantify the potential benefits of the insights in Sections 2.6.1-2.6.4. By simulating donors, we can compare historical fundraising strategies with our recommended ones. It is difficult to evaluate our recommendations based purely on the historical data, since the data represent the actual outcomes of a particular set of design choices used in the past, and there is no way to redo those same communications with a different set of designs.

For our simulations, we randomly sampled 10,000 donors who received mailings during the first six months of 2009. For the ith donor account, we randomly generate a value $\hat{b}_i$ from the random effect distribution estimated in our statistical analysis. These values are viewed as fixed in the subsequent procedure. Then, for the jth historical communication with account i within the six-month period, we simulate an outcome $\hat{y}_{ij}$ from a Bernoulli distribution satisfying $\mathbb{E}\left(\hat{y}_{ij}\right) = g^{-1}\left(\mathbf{x}_{ij}^T\bar{\boldsymbol{\beta}} + \hat{b}_i\right)$, where g is the logit link function and $\bar{\boldsymbol{\beta}}$ is a vector of the final estimated coefficients from Section 2.6.2 (including the three-way terms).

The vector $\mathbf{x}_{ij}$ can now be modified to reflect different fundraising strategies. The *historical strategy* simply consists of setting the elements of $\mathbf{x}_{ij}$ equal to their historical values. The *new strategy* uses the following rules. First, a supporter card is always included in the first communication with donor i that is of the Renewal type, but not in any other communications with that donor. Second, dynamic amounts are only applied to new, unconverted donors, as discussed in Section 2.6.2. Third, generic stories are never used; preparedness and specific stories are each used 50% of the time (we assumed that the organization may prefer to use a variety of stories, even if one type works better than another). Fourth, gift items are never included. For the first communication with a donor, descriptive features such as recency are always set to their historical values.

In this way, we can generate donors with realistic features, as well as outcomes for two versions of the same communication that have different design features. We updated time-dependent features dynamically for both strategies. For example, for the ith donor, we store a counter representing the number of successful communications with that donor. The counter is incremented if $\hat{y}_{ij} = 1$ (for the particular strategy used) and used to set the "previous successes" feature for the next communication with the donor. Likewise, if $\hat{y}_{ij} = 1$, the recency of donor i is reset to "0-6 months" for the next communication.

On average, the sampled donors receive 12-13K mailings in the six-month period. The 99% confidence intervals for the success rates achieved by the two strategies are 0.0539 ± 0.0006 for the historical strategy, and 0.0812 ± 0.0009 for the new strategy. The new strategy significantly improves the success rate. Additionally, when $\hat{y}_{ij} = 1$, we can plug $\mathbf{x}_{ij}$ into the models in Section 2.6.4 to obtain predicted donation amounts for the simulated successes (this also allows us to dynamically update the donor class features for the next communication with donor i). On average, the historical strategy collects \$61,231 ± \$1,167 in revenues, while the new strategy collects \$99,559 ± \$1,453 (99% confidence intervals reported for both strategies).

We should note several grounds for caution in interpreting this comparison. First, all success probabilities are calculated from the estimated model, although this model was calibrated using a massive volume of historical data. Second, the simulations assume that, in both scenarios, the Red Cross sent the same number of mailings to the same donors. However, this may actually cause the simulations to under-report the improvement achieved by the new strategy: in practice, if the organization were to receive more donations, it would also tend to communicate with those donors more frequently, thus creating an opportunity for still more successes.

Ultimately, while the precise numerical improvement achieved by the new strategy reflects the assumptions made in our simulations, these results suggest that the managerial insights from Section 2.6 can be translated into a few simple design rules that offer significant potential for improving retention rates, even under a pre-specified number of communications with each donor.

2.7 Conclusion

This chapter presents a data-driven study of disaster donor cultivation, using a massive dataset from the American Red Cross, to formulate several models of fundraising success. The results of this analysis lead to the following managerial insights for managers at the Red Cross and other non-profits who work on *cultivation* of first-time donors into regular supporters:

1. *Relief vs. preparedness.* The influence of emotive imagery on charitable behaviour is widely recognized. However, in donor *cultivation*, there is evidence to suggest that preparedness-based appeals are much more effective than relief-based appeals. More broadly, this suggests that non-profits may benefit from more informative (rather than emotive) content in programs that focus on cultivation.

2. *Supporter cards.* A small card affirming a donor's identity as a Red Cross supporter appears to exert a significant positive impact on cultivation efforts. We recommend the inclusion of this item as a standard component of STAART mailings, perhaps in the first communication with a donor. We believe that

non-profits in general can improve donor retention by using such techniques to reinforce the identity of potential supporters.

3. *Gift items.* On the other hand, other gift items, such as emergency lights and address labels, appear to have no significant effect on cultivation. From the evidence, we conclude that these items can be eliminated as a cost-saving measure.

4. *Dynamic amounts.* The strategy of dynamic amount generation individualizes the ask amounts in a mailed appeal, based on each donor's previous donation history. Essentially this strategy encourages a donor to maintain an earlier level of contribution. The evidence suggests that this works on very recent first-time donors who have not yet been converted, but may actually be counterproductive with current and lapsed supporters. In such cases, it may be better to use a few standard ask amounts, or substantially scale down the dynamic amounts.

A major challenge of donor cultivation, and a limiting factor of this study, is the relative lack of information on donors. Previous work on donor behaviour has drawn from surveys and policy studies, which cover a relatively small number of individuals, but provide detailed information on income, demographics, donor motivation, and other relevant factors. At the same time, while these attributes are valuable in understanding the economic and behavioural drivers of donations, they are unobservable to organizations like the Red Cross during *operational* decisions. We have formulated recommendations to help non-profit managers to improve cultivation campaigns based on information that they have available at the time the decision is made. To our knowledge, this is the first study to adopt a data-driven approach to this problem. We believe this to be an important contribution to the study of non-profit donations.

Chapter 3: Optimal Learning with Combinatorial Feature Selection

3.1 Overview

Many applications in business analytics and operations research exhibit a feedback loop between statistics and optimization. First, historical data are used to fit regression models that relate a performance metric of interest to a set of user-specified design inputs. Second, the decision-maker chooses a new set of inputs, guided by their estimated effects in the regression model. This decision is then implemented in the field and a response is observed; this response, in turn, becomes a new data point used to improve the regression model, and the process is repeated. Statistical estimation thus alternates with optimization: in the first step, the design inputs are treated as fixed (taken from historical data) and regression coefficients are estimated, and in the second step, the coefficients are treated as fixed and the regression features become decision variables that determine the next data point.

Consider how these issues arise in non-profit fundraising discussed in Chapter 2. A manager at the American Red Cross is tasked with designing a monthly or quarterly fundraising campaign using direct-mail. The manager are still facing the problem of choosing from a large number of features, such as the story type, matching grants [84], including or excluding a free gift item, a photograph, a supporting card, etc.

The manager's objective is to choose a set of designs that maximizes the response rate (or proportion of mailings that elicit a gift). The model may include interaction terms as in Chapter 2, and the chosen design may be subject to constraints (for example, only one type of story may be used). Thus, the manager's problem can be formulated as a binary integer program, where the variables represent decisions to include different design features.

The challenge in this problem stems from the fact that the regression model available to the manager is subject to uncertainty. Moreover, after the campaign is implemented, its outcome can be treated as a new observation and added to the available data. Upon refitting the model, we may obtain a different set of coefficients, leading to a different IP and possibly a different decision for the next campaign. Thus, it may be suboptimal to implement the decision suggested by the current model; likewise, a decision that appears to be suboptimal may in fact be much better than the current coefficients indicate. In this work, we consider design strategies that anticipate the effect of new information on the model and quantify the economic value of this information. Our goal is to integrate the statistical and optimization components of the problem and make decisions that have high *potential* to be optimal, or to improve the quality of the statistical model.

In the simulation literature, optimal information collection problems are widely studied in the context of ranking and selection or R&S [119, 120]. In R&S, there is a finite set of alternatives (e.g., combinations of design features), each of which has an unknown value (the mean response rate for a campaign with those designs). The decision-maker has a limited experimental budget for collecting information about individual alternatives. We wish to allocate the budget efficiently, in order to identify the true best alternative as quickly as possible. Bayesian statistics can be used in R&S [56] to model our evolving beliefs about the value of each alternative; one advantage of Bayesian models is that they allow "correlated beliefs" [121,122] for modeling relationships between different alternatives (for example, two combinations with multiple common elements). Correlated beliefs can also be extended to a linear regression framework [58].

However, in the problem we consider, the number of alternatives grows combinatorially with the number of regression features, leading to high computational costs for many standard classes of R&S algorithms:

- Value of information procedures (VIPs) calculate an expected improvement criterion for each alternative [123]. In our setting, this calculation would require us to either enumerate every alternative, or solve a nonlinear, nonconvex binary IP. While [124] proposes a VIP specifically for a parametric learning model (such as our regression model), the parametric structure is only used to reduce the storage cost, not the computational cost, and the VIP still enumerates every alternative.

- Indifference-zone methods [63, 125] are often based on the idea of sequentially screening the set of alternatives; in each stage, we collect some number of observations from every alternative in our set, and then screen out those alternatives that fail a certain statistical test. However, in non-profit fundraising, the number of possible combinations of designs is much greater than the total experimental budget, making it problematic to test large numbers of alternatives even once.

- Optimal computing budget allocation (OCBA) methods [126, 127] determine a proportion of the budget to be allocated to each alternative. These methods can be implemented sequentially, but usually do not allow correlated beliefs; [128] considers correlated sampling distributions, but requires the correlation structure to be known rather than learned over time. Furthermore, OCBA methods do not consider parametric belief models.

In this work, we develop information collection algorithms for regression problems with combinatorial feature selection. The learning process is modeled by a version of Bayesian linear regression that allows the noise variance to be unknown (as is certainly the case in non-profit fundraising). Using VIPs as a foundation, we first derive an explicit form for the expected improvement criterion in the context of this model, and prove the asymptotic optimality of the VIP in this setting. However, due to the combinatorial decision space, the cost of computing expected improvement remains high, motivating additional algorithmic developments. We create a convex approximation, based on optimal quantization and semidefinite programming relaxation, to the nonconvex expected improvement problem. The computational complexity of this problem is polynomial in the number of features, not the number of alternatives. Finally, we solve a small binary IP to round the solution. As a result, we obtain decisions with high value of information much more quickly than if we were to compute the expected improvement criterion exactly for every alternative.

Some recent work has touched on similar issues. For example, [129] studies a general framework for learning in combinatorial optimization (e.g., subset selection). However, the model and algorithmic approach in this work rely on the ability to individually collect information about each element of the chosen subset during a single experiment, which enables the assumption of independent beliefs about the elements. By contrast, in regression, we only observe a single scalar response for the chosen set of features, thus creating correlations. Another recent work by [130] has considered SDP relaxations for VIPs, but assumes known sampling noise as well as a continuous and highly structured decision space. [124] considers learning in linear regression, but assumes a generic finite set of alternatives, and does not consider the challenges arising from combinatorial feature selection. Finally, one stream of research [131, 132] applies branch-and-bound techniques to discrete simulation optimization, but treats the objective function as a black box, without the additional parametric structure afforded by regression.

This work makes the following contributions. 1) We introduce a conjugate Bayesian learning model with unknown variance and derive the expected improvement criterion in this setting, thus generalizing the results for the known-variance model in [124]. 2) We prove the asymptotic consistency of the VIP in the parametric model, a result that has not previously been available for even the known-variance case. 3) We propose two additional algorithmic developments to improve the computational efficiency of the VIP. First, we show how the computation is simplified in the special case where the binary decision is not subject to any additional linear constraints. Second, we provide an approximate algorithm, based on SDP relaxation, that mitigates the difficulties of computing expected improvement in the general case. 4) We provide numerical results showing the added value of the new VIP and its approximate variant over several benchmarks, as well as the computational savings afforded by the SDP approximation.

Finally, we briefly note that, while we use the non-profit fundraising application for motivation throughout this work, the algorithmic and theoretical results apply to a broader problem class, in which regression and optimization alternate as decisions are made over time. For example, [133] estimates a linear regression model for the quality of a cancer treatment as a function of drug dosage. Given such a model, an optimization problem can be solved to create a treatment for a new clinical trial. We observe that, in practice, the outcome of each new trial would feed back into the regression model, giving rise to an information collection problem similar to the one considered here.

3.2 Model

In Section 3.2.1, we formulate an integer programming model for the feature selection problem, using the Red Cross fundraising application to provide motivation. In Section 3.2.2, we give a Bayesian learning model used to update a set of beliefs about the regression parameters in the presence of unknown sampling noise.

3.2.1 Regression-based optimization

Consider the linear regression model

$$\eta = \varphi^\top \beta + \varepsilon, \tag{3.1}$$

where η is a response variable, φ is a vector of r features, and ε is a zero-mean noise term. We assume that $\varphi \in \{0,1\}^r$, that is, all of the features are binary. Each component φ_i represents the presence or absence of a particular design input. In traditional regression, we would be given a fixed set of observations $\eta^1, ..., \eta^n$ and corresponding feature vectors $\varphi^1, ..., \varphi^n$, and our task would be to estimate β. Suppose, however, that β has already been estimated and our task is to choose inputs that maximize the mean of the next observation. We then solve, for fixed β, the binary integer program

$$\begin{aligned} V(\beta) = \max\nolimits_{\varphi} \quad & \beta^\top \varphi \\ \text{s.t.} \quad A\varphi &= h \\ \varphi &\in \{0,1\}^r \end{aligned} \tag{3.2}$$

where A and h represent constraints on the allowable inputs.

In the context of the non-profit fundraising application, campaign performance can be evaluated in terms of the success rate $y \in (0,1)$, or the proportion of mailings that elicit donations. A transformation $\eta = \text{logit}(y)$ enables us to apply linear regression. See Section 2.3 for a detailed description of design features and Section 2.6 for a subset of features we selected based on historical data for potential key drivers.

The linear constraints $A\varphi = h$ may come from multiple-choice decisions, e.g., $\varphi_i + \varphi_j = 1$ if i and j represent two possible story types. More importantly, they may come from interactions between attributes, which are common in applications of linear regression. For example, the combined effect of a disaster preparedness story with the Renewal campaign type may be greater than the sum of the individual effects of these features. Then, if φ_i and φ_j represent the respective decisions to include a preparedness story and target the Renewal type, our model will include an additional binary variable φ_k with the requirement $\varphi_k = \varphi_i \varphi_j$. This constraint may be linearized by including

$$\begin{aligned} \varphi_k &\leq \varphi_i \\ \varphi_k &\leq \varphi_j \\ \varphi_i + \varphi_j - 1 &\leq \varphi_k \end{aligned}$$

among the constraints in (3.2). By adding slack variables, these constraints can be converted into the form $A\varphi = h$. Note that such slack variables will still satisfy the binary constraints.

Finally, for notational convenience, we let $\Phi = \{\varphi \in \{0,1\}^r | A\varphi = h\}$ denote the feasible region of (3.2). Then, $K = |\Phi|$ is the number of feasible decisions. Note that K depends exponentially on r if most of the attributes are controllable by the decision-maker. In non-profit fundraising, r may be fairly small; for example, [134] identifies 10-20 significant features that impact donor cultivation. The deterministic IP in (3.2) may thus be relatively easy to solve, but the dimension of information collection substantially complicates the problem, as will be shown in the following sections.

3.2.2 Bayesian model for information collection

In practice, the regression coefficients β are subject to considerable uncertainty, particularly for newer mailing designs for which extensive historical data are not available. We model this uncertainty using a Bayesian prior distribution on β. The parameters of the distribution will change as new information is collected, leading to improved solutions of (3.2). We also allow the variance of the noise term ε to be unknown (which is certainly the case in non-profit fundraising), and include our uncertainty about the noise into the model.

Returning to (3.1), we assume that $\varepsilon \sim \mathcal{N}\left(0, \frac{1}{\rho}\right)$, where ρ is an unknown *measurement precision*. We further assume that $\rho \sim Gamma\,(a^0, b^0)$, where the prior parameters a^0, b^0 are pre-specified by the decision-maker. Finally, we assume that the conditional distribution of β given ρ is multivariate normal with mean vector θ^0 and covariance matrix $\frac{1}{\rho}\Sigma^0$, where θ^0, Σ^0 are also user-specified.

Suppose that our budget allows us to conduct N experimental campaigns (e.g., on small groups of donors) before committing to a final estimate of β (e.g., for a large-scale campaign). For $n = 0, 1, ..., N-1$, the $(n+1)$st campaign is characterized by the feature vector $\varphi^n \in \Phi$, and η^{n+1} denotes the outcome of the campaign. The noise terms $\varepsilon^1, ..., \varepsilon^N$ are assumed to be i.i.d., as is standard in linear regression. Let $\mathcal{F}^n$ denote the sigma-algebra generated by $\varphi^0, \eta^1, \varphi^1, \eta^2, ..., \varphi^{n-1}, \eta^n$. The following result shows that the conditional distribution of (β, ρ) given $\mathcal{F}^n$ remains multivariate normal-gamma for all n, and provides a fast recursive update for the parameters. The proof is given in the Appendix.

PROPOSITION 1. *Suppose that the conditional distribution of (β, ρ) given $\mathcal{F}^n$ is multivariate normal-gamma with parameters $(\theta^n, \Sigma^n, a^n, b^n)$. Then, the conditional distribution of (β, ρ) given $\mathcal{F}^{n+1}$ is multivariate normal-gamma with parameters*

$$\theta^{n+1} = \theta^n + \frac{\eta^{n+1} - (\varphi^n)^\top \theta^n}{1 + (\varphi^n)^\top \Sigma^n \varphi^n} \Sigma^n \varphi^n, \tag{3.3}$$

$$\Sigma^{n+1} = \Sigma^n - \frac{\Sigma^n \varphi^n (\varphi^n)^\top \Sigma^n}{1 + (\varphi^n)^\top \Sigma^n \varphi^n}, \tag{3.4}$$

$$a^{n+1} = a^n + \frac{1}{2}, \tag{3.5}$$

$$b^{n+1} = b^n + \frac{(\eta^{n+1} - (\varphi^n)^\top \theta^n)^2}{2(1 + (\varphi^n)^\top \Sigma^n \varphi^n)}. \tag{3.6}$$

Note that (3.3)-(3.4) are identical to the recursive least-squares update [62, Sec. 9.3] in frequentist statistics. Thus, the sequence of posterior mean vectors obtained from the Bayesian model is precisely the sequence of least-squares estimators obtained after each new data point. Furthermore, it follows from the properties of the normal distribution that, for any $\varphi \in \Phi$, we have $\varphi^\top \beta \sim \mathcal{N}\left(\varphi^\top \theta^n, \frac{1}{\rho} \varphi^\top \Sigma^n \varphi\right)$ conditionally given $\mathcal{F}^n$ and ρ. Thus, the Bayesian model characterizes our uncertainty about the value of every feasible alternative, but incurs a cost of $\mathcal{O}\left(r^2\right)$ to store and update.

Finally, we can state the objective of the information collection problem. We wish to create an adaptive policy that will design campaigns based on the most recent information. Let Π be the set of all functions $\pi : \mathbb{R}^r \times \mathbb{S}^r_+ \times \mathbb{R} \times \mathbb{R}$ mapping

(θ, Σ, a, b) to a decision $\varphi \in \Phi$. The optimal policy solves the problem

$$\sup_{\pi \in \Pi} \mathbb{E}^{\pi} V\left(\theta^N\right), \tag{3.7}$$

where $\mathbb{E}^{\pi}$ denotes a conditional expectation given $\varphi^n = \pi\left(\theta^n, \Sigma^n, a^n, b^n\right)$ for all n. Essentially, (3.7) evaluates a policy in terms of its ability to guide us to a more favourable solution of (3.2) using the final regression estimates θ^N. The problem in (3.7) is a finite-horizon dynamic program with a multi-dimensional and continuous state space; thus, while the optimal policy can be characterized using Bellman's equation [135], it is computationally intractable. In the following, we develop efficient heuristics for this problem.

3.3 The KGUP algorithm for combinatorial feature selection

In this section, we propose a VIP for combinatorial feature selection in parametric models. Section 3.3.1 derives a closed-form expression for the value of information in this setting. The consistency of the VIP is proved in Section 3.3.2. For the moment, we consider the exact form of the VIP, which requires enumeration of every alternative; algorithmic improvements will be presented in the following section.

3.3.1 Derivation of the KGUP algorithm

Value of information procedures are based on the insight that, while (3.7) is intractable in general, it may admit a closed-form solution for $N = 1$. This solution yields the alternative that would be optimal to implement, if this were the last experiment with no additional chances to collect information. Such an alternative maximizes the well-known expected improvement criterion [136], given for fixed $\psi \in \Phi$ by

$$v_{\psi}^{KG,n} = \mathbb{E}^n \left[\max_{\varphi \in \Phi} (\varphi^\top \theta^{n+1}) | \varphi^n = \psi \right] - \max_{\varphi \in \Phi} (\varphi^\top \theta^n), \tag{3.8}$$

where $\mathbb{E}^n[\cdot] = \mathbb{E}[\cdot | \mathcal{F}^n]$, i.e., the conditional expectation given $\mathcal{F}^n$. Since (3.8) represents the marginal value of a single measurement of ψ, it is also known as the knowledge gradient [135] or KG, a name that we also adopt in this work.

Notice from (3.3) that, at stage n, the posterior mean θ^{n+1} is unknown, but the uncertainty in this vector derives only from a scalar quantity η^{n+1}. All other quantities in (3.3)-(3.6) are known at time n. The following result characterizes the conditional distribution of θ^{n+1} given $\mathcal{F}^n$ in terms of a scalar random variable.

PROPOSITION 2. *The predictive distribution of η^{n+1} given $\mathcal{F}^n$ and $\varphi^n = \psi$ is given by*

$$\eta^{n+1} \sim t \left(2a^n, \psi^\top \theta^n, \frac{b^n (1 + \psi^\top \Sigma^n \psi)}{a^n} \right),$$

which denotes a univariate Student's t-distribution with mean $\psi^\top \theta^n$, scale parameter $\frac{b^n}{a^n}(1 + \psi^\top \Sigma^n \psi)$, and $2a^n$ degrees of freedom.

Proof. The characteristic function of η^{n+1} given ρ is

$$\begin{aligned}\mathbb{E}(e^{it\eta^{n+1}}|\rho) =&\mathbb{E}(\mathbb{E}(e^{it\eta^{n+1}}|\rho,\beta)|\rho)\\ =&\mathbb{E}(e^{it\psi^\top\beta-\frac{1}{2}\frac{1}{\rho}t^2}|\rho)\\ =&e^{-\frac{1}{2}\frac{1}{\rho}t^2}\mathbb{E}(e^{it\psi^\top\beta}|\rho)\\ =&e^{-\frac{1}{2}\frac{1}{\rho}t^2}e^{it\psi^\top\theta^n-\frac{1}{2}\psi^\top\frac{1}{\rho}\Sigma^n\psi t^2}\\ =&e^{it\psi^\top\theta^n-\frac{1}{2}\frac{1}{\rho}(1+\psi^\top\Sigma^n\psi)t^2}.\end{aligned}$$

Consequently, we can write the density of η^{n+1} as

$$\begin{aligned}p(\eta^{n+1}) =&\int\limits_{\mathbb{R}^+} g(\eta^{n+1}|\rho)g(\rho)d\rho\\ =&\int\limits_{\mathbb{R}^+}\sqrt{\frac{\rho}{2\pi(1+\psi^\top\Sigma^n\psi)}}\exp\left(-\frac{\rho(\eta^{n+1}-\psi^\top\theta)^2}{2(1+\psi^\top\Sigma^n\psi)}\right)\frac{(b^n)^{a^n}}{\Gamma(a^n)}\rho^{a^n-1}\exp(-b^n\rho)d\rho\\ =&\frac{\Gamma(a^n+\frac{1}{2})}{\Gamma(a^n)}\frac{1}{\sqrt{2\pi(1+\psi^\top\Sigma^n\psi)b^n}}\left(1+\frac{\left(\eta^{n+1}-\psi^\top\theta^n\right)^2}{2(1+\psi^\top\Sigma^n\psi)b^n}\right)^{-a^n-\frac{1}{2}}.\end{aligned}$$

Define $\tilde{\eta}^{n+1}=\sqrt{\frac{a^n}{(1+\psi^\top\Sigma^n\psi))b^n}}(\eta^{n+1}-\psi^\top\theta^n)$. Then, the pdf of $\tilde{\eta}^{n+1}$ is given by

$$p\left(\tilde{\eta}^{n+1}\right)=\frac{\Gamma(\frac{2a^n+1}{2})}{\Gamma(\frac{2a^n}{2})(\pi 2a^n)^{\frac{1}{2}}}\left(1+\frac{(\tilde{\eta}^{n+1})^2}{2a^n}\right)^{-\frac{2a^n+1}{2}},$$

which is the pdf of the standard Student's t-distribution with $2a^n$ degrees of freedom. Thus

$$\eta^{n+1}=\sqrt{\frac{b^n(1+\psi^\top\Sigma^n\psi)}{a^n}}\tilde{\eta}^{n+1}+\psi^\top\theta^n$$

follows the desired distribution. □

As in [122], we observe an analogy with classical frequentist statistics. When the noise variance is known [124], a similar result can be derived where the scalar random variable is normally distributed. However, when the variance is unknown, we use Student's t-distribution instead.

Let T_{s^n} denote a random variable following a standard Student's t-distribution with $s^n = 2a^n$ degrees of freedom. Using (3.3) with Proposition 2, we can rewrite θ^{n+1} as

$$\theta^{n+1} = \theta^n + \Sigma^n \psi \sqrt{\frac{b^n}{a^n \left(1 + \psi^\top \Sigma^n \psi\right)}} T_{s^n}. \tag{3.9}$$

Combining (3.8) and (3.9), we can derive a new formulation of the KG quantity, given by

$$v_\psi^{KG,n} = \mathbb{E}^n \left(\max_{\varphi \in \Phi} p_\varphi^n + q_\varphi^n(\psi) T_{s^n} \right) - \max_{\varphi \in \Phi} p_\varphi^n, \tag{3.10}$$

where $p_\varphi^n = \varphi^\top \theta^n$ and

$$q_\varphi^n(\psi) = \varphi^\top \Sigma^n \psi \sqrt{\frac{b^n}{a^n \left(1 + \psi^\top \Sigma^n \psi\right)}}. \tag{3.11}$$

Now, observe that the quantity inside the expectation in (3.10) is a piecewise linear function with slopes represented by the quantities $q_\varphi^n(\psi)$. We sort the pairs $\{p_\varphi^n, q_\varphi^n(\psi)\}_{\varphi \in \Phi}$ and relabel them as $\{p_i^n, q_i^n\}_{i=1}^K$ such that the values q_i^n are in ascending order. Standard techniques (see Section 5.3 of [54]) can be applied to obtain a sequence $\{c_i^n\}_{i=1}^K$ of breakpoints satisfying $j = \arg\max_i \left(p_i^n + q_i^n t\right)$ if and only if $t \in \left[c_{j-1}^n, c_j^n\right)$ (this may involve removing some dominated alternatives that never achieve the argmax for any value of t). Finally, the analysis of [122] can be applied to obtain

$$
\begin{aligned}
v_{\psi}^{KG,n} &= \sum_{i=1}^{K-1} \left(q_{i+1}^{n}(\psi) - q_{i}^{n}(\psi)\right) \mathbb{E}\left[(T_{s^n} - |c_i^n|)^{+}\right] \\
&= \sum_{i=1}^{K-1} \left(q_{i+1}^{n}(\psi) - q_{i}^{n}(\psi)\right) \left(\frac{s^n + (c_i^n)^2}{s^n - 1} g_{s^n}(|c_i^n|) - |c_i^n| \left(1 - G_{s^n}(|c_i^n|)\right)\right)
\end{aligned}
\tag{3.12}
$$

where $g_s(\cdot)$ and $G_s(\cdot)$ are the pdf and cdf, respectively, of the standard Student's t-distribution with s degrees of freedom.

We define a new VIP, called the *Knowledge Gradient with Unknown Precision* (KGUP), which chooses the design of the $(n+1)$st campaign to be

$$\psi^{KGUP,n} = \arg\max_{\psi \in \Phi} v_{\psi}^{KG,n}. \tag{3.13}$$

In this way, the value of information approach is extended to learn linear regression coefficients in the presence of unknown noise variance. The procedure strikes a balance between the estimated value of a design φ (given by $\varphi^{\top}\theta^{n}$) and our uncertainty about that design, which depends on the posterior variances of the coefficients of the features included in the design as well as their correlations with other features (which may not be included).

In R&S, we would typically solve (3.13) by evaluating $v_{\psi}^{KG,n}$ for every $\psi \in \Phi$. Currently available algorithms for this computation incur a cost of $\mathcal{O}(K \log K)$ for each ψ [54]. Thus, the overall computational complexity is $\mathcal{O}(K^2 \log K)$. Because K grows exponentially in the number of attributes r, that is, $K \sim 2^r$, this translates to a complexity of $\mathcal{O}(r4^r)$, which can be prohibitively expensive. Furthermore, it may be difficult to exploit the binary structure of φ since (3.12) is highly nonlinear and nonconvex in the belief parameters. We will return to these important algorithmic issues in the next section after demonstrating the consistency of the procedure.

3.3.2 Consistency of the KGUP algorithm

In this section we show a form of consistency for the KGUP procedure, namely that $Var\left(\varphi^\top \beta \,|\, \mathcal{F}^n\right) \to 0$ for every feasible $\varphi \in \Phi$, which implies that we asymptotically obtain perfect information about every alternative. We also show that, if the feasible region Φ contains r linearly independent vectors, this is equivalent to the statistical consistency of θ^n in the classical sense. However, if some of the regression coefficients are irrelevant to the feasible decisions (e.g., if we have a constraint $\varphi_i = 0$ for some i), it may not be necessary to learn their exact values. It is also possible to have pathological constraints such as $\varphi_i = \varphi_j$ that would cause identifiability issues for any regression model. The assumption on Φ excludes these cases; in practice, one could use model selection techniques [108] on historical data to eliminate such irrelevant features prior to beginning the information collection problem.

For convenience, we also assume in this analysis that $\Sigma^0_{i,i} > 0$ for all i, that is, we do not have perfect information about any feature. Such features are automatically uncorrelated with any other features, and thus have no impact on the KG quantity. If φ_i is a slack variable, as discussed in Section 3.2.1, we can model $\theta^0_i = 0$ and $\Sigma^0_{i,i} = 0$, and exclude the ith feature from the subsequent analysis.

PROPOSITION 3. *For any sampling policy,* $\lim_{N\to\infty} \frac{a^N}{b^N} = \rho$ *a.s.*

Proof. From Proposition 1, we know that $\frac{a^n}{b^n} = \mathbb{E}^n(\rho)$. It follows from Theorem

V.4.7 of [137] that the process $\left(\frac{a^n}{b^n}\right)_{n=0}^{\infty}$ is a uniformly integrable martingale and converges a.s. to $\mathbb{E}\left(\rho \,|\, \mathcal{F}^{\infty}\right)$. Thus, it remains to show that ρ is $\mathcal{F}^{\infty}$-measurable.

Since the set Φ is finite, any policy π must measure at least one alternative $\bar{\varphi}$ infinitely often as $N \to \infty$. Note that $\bar{\varphi}$ may depend on the sample path, but is measurable with respect to $\mathcal{F}^{\infty}$. The sample variance of all η^{n+1} for which $\pi\left(\theta^n, \Sigma^n, a^n, b^n\right) = \bar{\varphi}$ converges a.s. to the true variance $\frac{1}{\rho}$. It follows that ρ is measurable with respect to $\mathcal{F}^{\infty}$, whence $\mathbb{E}^n\left(\rho\right) \to \rho$ a.s. $\square$

From martingale arguments, it also follows that $\left(\theta^n, \Sigma^n\right)$ have finite a.s. limits under any policy, since it is possible to write

$$\left(\theta^n, \frac{b^n}{a^n - 1}\Sigma^n + \theta^n \left(\theta^n\right)^{\top}\right) = \mathbb{E}^n \left(\beta, \beta\beta^{\top}\right) \tag{3.14}$$

and apply Theorem V.4.7 of [137] to find that $\frac{b^n}{a^n-1}\Sigma^n$ converges. Since Proposition 3 also implies that $\frac{b^n}{a^n-1} \to \frac{1}{\rho}$ a.s., it follows that Σ^n converges. However, the limit here may now depend on the policy. We now bring the KGUP policy back into the discussion and write

$$h_s\left(p, q\right) = \mathbb{E}\left(\max_{\varphi\in\Phi} p_{\varphi} + q_{\varphi} T_s\right) - \max_{\varphi\in\Phi} p_{\varphi}$$

for fixed $s > 1$ and $p, q \in \mathbb{R}^K$. The next two results study the properties of h_{s^n} for $s^n = 2a^n$ and $a^n = a^0 + \frac{n}{2}$ as in (3.5).

PROPOSITION 4. *Suppose that $\left(p^n, q^n\right)$ converges to a finite limit in $\mathbb{R}^K \times \mathbb{R}^K$. Then, the sequence $\left\{\max_{\varphi\in\Phi} p^n_{\varphi} + q^n_{\varphi} T_{s^n}\right\}_{n\geq 0}$ is uniformly integrable.*

Proof. From p. 75 of [137], the componentwise maximum of finitely many uniformly integrable sequences is uniformly integrable. Since both $\{p^n\}$ and $\{q^n\}$ are bounded, it remains to show that $\{T_{s^n}\}$ is uniformly integrable. We choose $s^0 = 2a^0 > 1$ so that each T_{s^n} has finite expectation. Consider the pdf g_{s^n} of the standard Student's t-distribution with s^n degrees of freedom. Because the tails of g_s become lighter with larger s, there exists a value $t^n > 0$ such that $g_{s^n}(t^n) = g_{s^0}(t^n)$ with $g_{s^n}(t) < g_{s^0}(t)$ for $t > t^n$ and $g_{s^n}(t) > g_{s^0}(t)$ for $0 < t < t_n$. Note that $s^n \to \infty$ and g_∞ is the standard normal pdf since T_{s^n} converges in distribution to a standard normal random variable as $s^n \to \infty$. Consequently, $t^n \to t^\infty$ where $g_\infty(t^\infty) = g_{s^0}(t^\infty)$.

Thus, $\tilde{t} = \sup_{n\geq 1} t^n$ is finite. For $M > \tilde{t}$, we have

$$\sup_n \mathbb{E}\left(T_{s^n} 1_{\{T_{s^n} > M\}}\right) \leq \mathbb{E}\left(T_{s^0} 1_{\left\{T_{s^0} > M\right\}}\right),$$

whence

$$\begin{aligned}\lim_{M\to\infty} \sup_n \mathbb{E}\left(|T_{s^n}| 1_{\{|T_{s^n}| > M\}}\right) &= 2 \lim_{M\to\infty} \sup_n \mathbb{E}\left(T_{s^n} 1_{\{T_{s^n} > M\}}\right) \\ &\leq 2 \lim_{M\to\infty} \mathbb{E}\left(T_{s^0} 1_{\left\{T_{s^0} > M\right\}}\right).\end{aligned}$$

The limit in the second line is equal to zero since T_{s^0} has finite expectation. □

PROPOSITION 5. *Suppose that (p^n, q^n) converges to a finite limit (p^∞, q^∞). The following statements are equivalent:*

1. $\lim_{n\to\infty} h_{s^n}(p^n, q^n) = 0$.

2. *There exists a constant ℓ such that $\lim_{n\to\infty} q^n_\varphi = \ell$ for all φ.*

Proof. Assume that 1) holds. Since $s^n \to \infty$, we know that T_{s^n} converges in distribution to a standard normal random variable Z. Note, however, that we are only interested in taking expectations over the distribution of T_{s^n} for the purpose of computing h_{s^n}. By Skorokhod's representation theorem [137, Corollary III.5.9], there exist random variables $\bar{T}_{s^n}$ and $\bar{Z}$ that have the same distribution as T_{s^n} and Z, but with $\bar{T}_{s^n} \to \bar{Z}$ almost surely. Thus,

$$\lim_{n\to\infty} \max_{\varphi\in\Phi} p^n_\varphi + q^n_\varphi \bar{T}_{s^n} = \max_{\varphi\in\Phi} p^\infty_\varphi + q^\infty_\varphi \bar{Z} \tag{3.15}$$

almost surely. From Proposition 4, it follows that the convergence in (3.15) also holds in L^1, whence

$$\lim_{n\to\infty} \mathbb{E}\left(\max_{\varphi\in\Phi} p^n_\varphi + q^n_\varphi \bar{T}_{s^n}\right) = \mathbb{E}\left(\max_{\varphi\in\Phi} p^\infty_\varphi + q^\infty_\varphi \bar{Z}\right).$$

Since $\bar{T}_{s^n}$ has the same distribution as T_{s^n}, it follows that

$$\begin{aligned}
\lim_{n\to\infty} h_{s^n}(p^n, q^n) &= \mathbb{E}\left(\max_{\varphi\in\Phi} p^\infty_\varphi + q^\infty_\varphi \bar{Z}\right) - \max_{\varphi\in\Phi} p^\infty_\varphi \\
&= \sum_{i=1}^{K-1} \left(q^\infty_{i+1} - q^\infty_i\right)\left(f\left(-\left|c^\infty_i\right|\right) - \left|c^\infty_i\right| F\left(-\left|c^\infty_i\right|\right)\right) \\
&= 0, \qquad (3.16)
\end{aligned}$$

where the functions f, F are the standard normal pdf and cdf [121], the values q^∞_i are obtained by sorting q^∞_φ in increasing order, and c^∞_i are the breakpoints of the piecewise linear function $t \mapsto \max_i p^\infty_i + q^\infty_i t$ as discussed previously. It can be shown that the function $z \mapsto f(-z) - zF(-z)$ is strictly positive, whence (3.16) implies $q^\infty_i = q^\infty_{i+1}$ for all i, as required.

Now, assume that 2) holds. In this case,

$$\lim_{n\to\infty} \max_{\varphi\in\Phi} p^n_\varphi + q^n_\varphi \bar{T}_{s^n} = \left(\max_{\varphi\in\Phi} p^\infty_\varphi\right) + \ell \bar{Z}$$

almost surely. Applying Proposition 4 again, we obtain $h_{s^n}(p^n, q^n) \to 0$ as required.

□

We now connect these results to the KGUP policy. First, we demonstrate that $v^{KG,n}_{\psi} \to 0$ for all $\psi \in \Phi$, which is then shown to imply statistical consistency of the regression estimators θ^n. Note that, due to the parametric structure of the problem, this does not necessarily require every ψ to be measured infinitely often, as it does in R&S [135]; in fact, as Corollary 1 shows, we may not need to measure some alternatives at all.

PROPOSITION 6. *Suppose that $\psi \in \Phi$ is measured infinitely often. Then, $\varphi^\top \Sigma^N \psi \to 0$ for all $\varphi \in \Phi$ and $v^{KG,N}_{\psi} \to 0$ a.s.*

Proof. For fixed N, let $N_\psi = \sum_{n=0}^N 1_{\{\varphi^n = \psi\}}$ be the number of times ψ is measured by time N. Define $\bar{\Sigma}^N$ by the equation

$$\left(\bar{\Sigma}^N\right)^{-1} = \left(\Sigma^0\right)^{-1} + \sum_{n=0}^N 1_{\{\varphi^n \neq \psi\}} \varphi^n \left(\varphi^n\right)^\top .$$

By the matrix inverse lemma, it follows that

$$\Sigma^N = \left(\left(\bar{\Sigma}^N\right)^{-1} + N_\psi \psi \psi^\top\right)^{-1} = \bar{\Sigma}^N - \frac{N_\psi \bar{\Sigma}^N \psi \psi^\top \bar{\Sigma}^N}{1 + N_\psi \psi^\top \bar{\Sigma}^N \psi}.$$

Consequently,

$$\begin{aligned} \psi^\top \Sigma^N \psi &= \psi^\top \bar{\Sigma}^N \psi - \frac{N_\psi \left(\psi^\top \bar{\Sigma}^N \psi\right)^2}{1 + N_\psi \psi^\top \bar{\Sigma}^N \psi} \\ &= \frac{\psi^\top \bar{\Sigma}^N \psi}{1 + N_\psi \psi^\top \bar{\Sigma}^N \psi}, \end{aligned}$$

which vanishes to zero as $N_\psi \to \infty$. By the Cauchy-Schwarz inequality,

$$\left(\varphi^\top \Sigma^N \psi\right)^2 \leq \left(\varphi^\top \Sigma^N \varphi\right)\left(\psi^\top \Sigma^N \psi\right) \leq \left(\varphi^\top \Sigma^0 \varphi\right)\left(\psi^\top \Sigma^N \psi\right),$$

implying that $\varphi^\top \Sigma^N \psi \to 0$ for all $\varphi \in \Phi$. Proposition 5 then implies that $v^{KG,n}_{\psi} \to 0$ a.s. □

COROLLARY 1. *If $\psi_1, \psi_2, \ldots, \psi_k$ are measured infinitely often, then $v_{\psi}^{KG,n} \to 0$ almost surely for any $\psi \in span(\psi_1, \psi_2, \ldots \psi_k)$.*

Proof. By Proposition 6 we have $\varphi^\top \Sigma^n \psi_i \to 0$ almost surely for all $\varphi \in \Phi$ and all $1 \leq i \leq k$. This implies $\varphi^\top \Sigma^n \psi \to 0$ almost surely for all $\varphi \in \Phi$. By Proposition 5, we then have $v_{\psi}^{KG,n} \to 0$ almost surely. □

PROPOSITION 7. *Suppose that φ^n is selected using the KGUP policy for all n. Then, for all $\psi \in \Phi$, $\lim_{n\to\infty} v_{\psi}^{KG,n} = 0$ almost surely.*

Proof. We prove this statement by contradiction. Fix ω and let $A_\omega \subseteq \Phi$ be the set of all $\psi \in \Phi$ for which $v_{\psi}^{KG,n}(\omega)$ does not converge to zero. Suppose that A_ω is non-empty. Then, Proposition 6 implies that $A_\omega \neq \Phi$ and also that any $\psi \in A_\omega$ has only been measured finitely many times on the sample path ω.

Since $|A_\omega|$ is finite, we can find a large enough N_1 such that, if $n > N_1$, then $\varphi^n(\omega) \notin A_\omega$. Furthermore, there exists some ε such that, for any N, there exists $n > N$ satisfying $\min_{\psi \in A_\omega} v^{KG,n}(\omega) > \varepsilon$. At the same time, for this ε, there also exists N_2 such that, for all $n > N_2$, $\max_{\psi \notin A_\omega} v_{\psi}^{KG,n}(\omega) < \varepsilon$. Consequently, there exists $n > \max(N_1, N_2)$ for which

$$\min_{\psi \in A_\omega} v^{KG,n}(\omega) > \varepsilon > \max_{\psi \notin A_\omega} v_{\psi}^{KG,n}(\omega).$$

Thus, any alternative in A_ω is preferable to any alternative not in A_ω at this time. However, since $n > K_1$, the KGUP policy must select an alternative not in A_ω, contradicting the definition of the policy. We conclude that $\lim_{n\to\infty} v_{\psi}^{KG,n} = 0$ a.s. for all $\psi \in \Phi$. □

THEOREM 1. *Suppose that φ^n is selected using the KGUP policy for all n. Then, $\psi^\top \Sigma^n \psi \to 0$ and $Var\left(\psi^\top \beta \,|\, \mathcal{F}^n\right) \to 0$ almost surely for all $\psi \in \Phi$.*

Proof. Consider a fixed ω. Propositions 5 and 7 imply that, for fixed $\psi \in \Phi$, there exists $\ell(\omega)$ such that $\varphi^\top \Sigma^n(\omega)\psi \to \ell(\omega)$ for all $\varphi \in \Phi$. We now show that $\ell(\omega)$ does not depend on ψ. By Proposition 7, we have $v^{KG,n}_{\psi_1}(\omega) \to 0$ and $v^{KG,n}_{\psi_2}(\omega) \to 0$ for any $\psi_1, \psi_2 \in \Phi$. Suppose that $\psi_1 \neq \psi_2$. By Proposition 5, we have

$$\varphi^\top \Sigma^\infty(\omega)\psi_1 = \ell_1(\omega), \qquad \varphi^\top \Sigma^\infty(\omega)\psi_2 = \ell_2(\omega)$$

for all $\varphi \in \Phi$.

Now, fix some φ. Proposition 7 implies that $v^{KG,n}_{\varphi}(\omega) \to 0$. It then follows from Proposition 5 that there exists some $\ell(\omega)$ such that $\psi^\top \Sigma^\infty(\omega)\varphi = \ell(\omega)$ for all ψ. Therefore, $\ell_1(\omega) = \ell_2(\omega) = \ell(\omega)$.

Furthermore, as $N \to \infty$, there is at least one alternative that is measured infinitely often on the sample path ω. Combining Proposition 6 with the above results, we obtain $\ell(\omega) = 0$. Thus, $\psi^\top \Sigma^n(\omega)\psi \to 0$ for all $\psi \in \Phi$, as required. It then follows from (3.14) that the conditional variance vanishes to zero as well. □

COROLLARY 2. *Suppose that the decision space Φ contains r linearly independent vectors. Then, the regression coefficients θ^n are consistent under the KGUP policy.*

Proof. Combining Theorem 1 with Corollary 1 implies that $\Sigma^n \to 0$ almost surely. It follows that the largest eigenvalue of Σ^n converges to zero. From (3.3)-(3.4), recall that θ^n is identical to the least-squares estimator. It is well-known [138, 139] that $\lambda_{\max}(\Sigma^n) \to 0$ is necessary and sufficient for consistency. □

3.4 Computational enhancements for the KGUP algorithm

We now present two approaches for computing the KG quantity in (3.10) with lower computational complexity. First, we consider a special case where the regression features are independent (i.e., the constraints $A\varphi = h$ are not present), and derive a direct calculation of the KG quantity that does not require direct computation of breakpoints. Second, we develop a convex approximation, based on SDP relaxation, for the general case. The approximation leads to much faster computations in practice.

3.4.1 The KGUP$_2$ algorithm for independent features

Suppose that the regression features are independent, and the constraints $A\varphi = h$ are removed from (3.2), with the possible exception of $\varphi_0 = 1$ for the intercept. In this case, all of the attributes are directly controllable (except for the intercept), and $K = 2^{r-1}$. By exploiting the binary structure of the decision variables, we obtain the following result.

PROPOSITION 8. *Suppose that Φ is the set of all possible combinations of $r-1$ controllable features. Then,*

$$v_{\psi}^{KG,n} = \sum_{j\geq 1, u_j^n \neq 0} \theta_j^n G_{s^n}\left(\left|\frac{\theta_j^n}{u_j^n(\psi)}\right|\right) + \frac{s^n (u_j^n(\psi))^2 + (\theta_j^n)^2}{(s^n - 1)|u_j^n(\psi)|} g_{s^n}\left(\left(\frac{\theta_j^n}{u_j^n(\psi)}\right)^2\right) - (\theta_j^n)^+, \tag{3.17}$$

where $u_j^n(\psi) = (\Sigma_{j\cdot}^n)^\top \psi \sqrt{\frac{b^n}{a^n(1+\psi^\top \Sigma^n \psi)}}$, and $\theta^n = (\theta_0^n, \theta_1^n, \ldots, \theta_{r-1}^n)^\top$.

Proof. Since φ is a binary vector and we control $(\varphi_1, \varphi_2, \ldots, \varphi_{r-1})$, the maximum of $\varphi^\top \theta^n$ is simply obtained by letting $\varphi_j = 1$ when $\theta_j^n \geq 0$ and $\varphi_j = 0$ otherwise. Then, (3.8) can be rewritten as

$$v_\psi^{KG,n} = \mathbb{E}^n \left[\sum_{j \geq 1} (\theta_j^{n+1})^+ | \varphi^n = \psi \right] - \sum_{j \geq 1} (\theta_j^n)^+. \tag{3.18}$$

Using (3.9) and (3.18), we obtain

$$\begin{aligned} v_\psi^{KG,n} =& \sum_j \mathbb{E}_\psi^n \left[(\theta_j^n + u_j^n T_{s^n})^+ \right] - (\theta_j^n)^+ \\ =& \sum_{j \geq 1, u_j^n > 0} \int_{-\frac{\theta_j^n}{u_j^n}}^{\infty} (\theta_j^n + u_j^n t) g_{s^n}(t) dt + \sum_{j \geq 1, u_j^n < 0} \int_{-\infty}^{-\frac{\theta_j^n}{u_j^n}} (\theta_j^n + u_j^n t) g_{s^n}(t) dt - \sum_{j \geq 1, u_j^n \neq 0} (\theta_j^n)^+. \end{aligned} \tag{3.19}$$

The conclusion follows after simple rearrangements of the terms in (3.19). □

Note that (3.17) is an exact calculation of the KG quantity, but does not require us to sort slopes or compute breakpoints, thus eliminating a cost factor of $\mathcal{O}(K \log K)$. The alternative with the largest KG factor can now be found at a cost of $\mathcal{O}(K)$, a dramatic improvement in efficiency over the general form in Section 3.3.1, although K may still be large.

3.4.2 The KGUP$_3$ algorithm for combinatorial feature selection

The derivation of the KGUP$_2$ algorithm assumes that every feature is directly controllable by the decision-maker, that is, any combination of zeros and ones is allowed. However, this is usually not the case in non-profit fundraising or regression in general, since the model can include interactions between attributes, as discussed in Section 3.2.1. For the general case, we develop a convex approximation of the KG quantity by first applying a quantization procedure to approximate the expectation over the distribution of T_s, and then applying an SDP relaxation to the resulting problem. The SDP can be solved to fixed precision using interior-point methods, whose complexity is polynomial in r [50].

Optimal quantization. Note that the second term of the KG quantity in (3.10) is independent of ψ, and the decision $\psi^{KGUP,n}$ made by the KGUP policy only depends on the first term. Thus, for convenience, we omit the second term and redefine

$$v_{\psi}^{KG,n} = \mathbb{E}^n \left(m(T_{s^n}) \right), \tag{3.20}$$

where $m\left(t\right) = \max_{\varphi \in \Phi} p_{\varphi} + q_{\varphi}\left(\psi\right) t$ for fixed p, q, ψ. Since m is a maximum of linear functions of t, it is convex in t.

The *Voronoi quantizer* for T_s is the function q^{vor}: $\mathbb{R} \to \{t_1, \ldots, t_J\}$ defined as

$$q^{vor}(T_s) = \sum_{j=1}^{J} t_j 1_{C(t_j)}(T_s),$$

where $\{C(t_j)\}_{1 \leq j \leq J}$ is a Borel partition of $\mathbb{R}$ with $C(t_j) = \{t \in \mathbb{R} : |t - t_j| \leq |t - t_{j'}|, j' = 1, \ldots, J\}$. Because T_s is one-dimensional and unimodal, for any fixed J there exists a unique $q^{vor}(\cdot)$ and corresponding sequence $t^J = \{t_j\}_{1 \leq j \leq J}$ minimizing the *quadratic quantization error* [140] given by

$$D_J^{T_s,2} = \mathbb{E}|T_s - q^{vor}(T_s)|^2 = \int_{\mathbb{R}} \min_j |t_j - t|^2 g_s(t) dt.$$

We know $m(t)$ is Lipschitz continuous in t. Letting L be its modulus, we write

$$\begin{aligned}
|\mathbb{E}^n m(T_s) - \mathbb{E}^n m(q^{vor}(T_s))| &\leq \mathbb{E}^n |m(T_s) - m(q^{vor}(T_s))| \\
&\leq L \mathbb{E}^n |T_s - q^{vor}(T_s)| \\
&\leq L \sqrt{D_J^{T_s,2}}.
\end{aligned}$$

Define $\{w_j\}_{1\leq j\leq J}$ as $w_j = G_s\left(\frac{t_j+t_{j+1}}{2}\right) - G_s\left(\frac{t_{j-1}+t_j}{2}\right)$, with the convention $t_0 = -\infty$ and $t_{J+1} = \infty$. Then $\mathbb{E}^n m(q^{vor}(T_s)) = \sum_j w_j m(t_j)$.

Thus, by finding the quantization sequence t^J that minimizes $D_J^{T_s n,2}$, we can approximate the KG quantity in (3.20) by

$$\hat{v}^J = \sum_{j=1}^{J} w_j m(t_j). \tag{3.21}$$

Newton's method has been used to compute the quantization sequence for several distributions, e.g., the standard normal distribution [141]. Here, we use a similar approach to find the quantization sequence for T_s. We first compute the gradient $\mathrm{d}(D_J^{T_s,2})$ and Hessian matrix $\mathrm{d}^2(D_J^{T_s,2})$ of the quadratic quantization error. Starting from some $t^{old} \in \mathbb{R}^J$, we compute

$$t^{new} = t^{old} - \left[\mathrm{d}^2(D_J^{T_s,2})(t^{old})\right]^{-1} \cdot \mathrm{d}(D_J^{T_s,2})(t^{old})$$

and iteratively replace t^{old} by t^{new} in order to find the zero of $\mathrm{d}(D_J^{T_s,2})$ in $\mathbb{R}^J$. In the Appendix, we present tables of the optimal quantizations for $J = 5$, $J = 10$, and $s = 3, ..., 20$; in our numerical experiments, we found that $J = 10$ produces competitive performance.

Semidefinite relaxation. We now describe the SDP relaxation of the problem in (3.13). The algorithm solves this problem for every n with a different set of inputs $(\theta^n, \Sigma^n, a^n, b^n)$. For convenience, we drop the superscripts from these quantities and show the computation for a generic, fixed (θ, Σ, a, b).

By (3.11), (3.21), and the definition of m, we have

$$\max_{\psi\in\Phi} \hat{v}^J = \max_{\psi\in\Phi} \max_{\varphi^1,...,\varphi^J\in\Phi} \sum_{j=1}^{J} w_j \left(\varphi^j\right)^\top \left(\theta + \Sigma t_j d_\psi\right), \tag{3.22}$$

where

$$d_\psi = \frac{\psi}{\sqrt{\frac{a}{b}(1+\psi^\top \Sigma \psi)}}.$$

We now reformulate this result to obtain the objective function for the SDP.

PROPOSITION 9. *Define*

$$C_j = \frac{1}{2}\begin{bmatrix} 0 & \theta^\top & 0^\top \\ \theta & 0 & t_j\Sigma \\ 0 & t_j\Sigma & 0 \end{bmatrix}, \quad Z_j = \begin{bmatrix} 1 \\ \varphi^j \\ d_\psi \end{bmatrix}\begin{bmatrix} 1 \\ \varphi^j \\ d_\psi \end{bmatrix}^\top = \begin{bmatrix} Z_j^{11} & Z_j^{1\varphi} & Z_j^{1d} \\ Z_j^{\varphi 1} & Z_j^{\varphi\varphi} & Z_j^{\varphi d} \\ Z_j^{d1} & Z_j^{d\varphi} & Z_j^{dd} \end{bmatrix}.$$

Then,

$$\max_{\psi\in\Phi} \hat{v}^J = \max_{\psi\in\Phi} \max_{\varphi^1,\ldots,\varphi^J\in\Phi} \sum_{j=1}^J w_j tr(C_j Z_j)\,. \tag{3.23}$$

Proof. We evaluate

$$\begin{aligned} \mathrm{tr}(C_j Z_j) &= \frac{1}{2}\mathrm{tr}\left(\begin{bmatrix} 0 & \theta^\top & 0^\top \\ \theta & 0 & t_j\Sigma \\ 0 & t_j\Sigma & 0 \end{bmatrix}\begin{bmatrix} 1 & (\varphi^j)^\top & (d_\psi)^\top \\ \varphi^j & \varphi^j(\varphi^j)^\top & \varphi^j d_\psi^\top \\ d_\psi & d_\psi(\varphi^j)^\top & d_\psi d_\psi^\top \end{bmatrix}\right) \\ &= (\varphi^j)^\top\theta + \mathrm{tr}(t_j\Sigma d_\psi (\varphi^j)^\top) \\ &= (\varphi^j)^\top\theta + t_j(\varphi^j)^\top \Sigma d_\psi. \end{aligned} \tag{3.24}$$

The conclusion follows from comparing (3.23) and (3.24) with (3.22). □

We observe that C_j is a constant matrix and Z_j is a positive semidefinite matrix with rank 1. This holds for all $j \in \{1,\ldots,J\}$. The problem in (3.23) is similar to an SDP, but has the following nonlinear constraints:

1. The rank-1 constraint on Z_j;

2. The binary constraints on ψ and φ^j;

3. The nonlinear constraints on d_ψ transformed from ψ.

To formulate (3.23) as an SDP, we relax the nonlinear constraints using a set of linear constraints. We first drop the rank-1 constraint on Z_j, then relax the binary constraints as $\psi, \varphi^j \in [0,1]^r$, and finally develop a set of linear constraints on d_ψ from $A\psi = h$.

From $A\psi = h$, we have $\psi^\top A^\top A \psi = h^\top h$. Thus

$$d_\psi = \frac{\psi}{\sqrt{\psi^\top \left(\frac{a}{b}\left(\frac{A^\top A}{h^\top h} + \Sigma\right)\right)\psi}} = \frac{\psi}{\sqrt{\psi^\top P \psi}},$$

where $P = \frac{a}{b}\left(\frac{A^\top A}{h^\top h} + \Sigma\right)$. It follows that

$$d_\psi^\top P d_\psi = 1. \tag{3.25}$$

We define $Y = d_\psi d_\psi^\top$, whence (3.25) is equivalent to

$$\operatorname{tr}(PY) = 1. \tag{3.26}$$

By definition, Y is symmetric and positive semidefinite, and we also require Y to be non-negative:

$$Y_{i,j} \geq 0, \quad 1 \leq i, j \leq r. \tag{3.27}$$

Next, we obtain a bound on the elements of Y by letting

$$\delta = \min_{\psi \in [0,1]^r, A\psi = h} \psi^\top P \psi,$$

whence

$$\operatorname{diag}(Y) \leq 1_r/\delta, \tag{3.28}$$

where 1_r is an r-vector of ones. By convention, there is no upper bound if $\delta = 0$. The quantity δ can be easily found by solving a small quadratic program with linear constraints and r variables, which can be done efficiently using a convex programming solver.

Following [130], we add another constraint on φ that can strengthen the relaxations. Given $\xi \in \mathbb{R}^r$ with $\xi_i > 0$ for each i, we define $\zeta = \sup_{\varphi \in \Phi} \xi^\top \varphi$. Then, for any $\varphi \in \Phi$ we have

$$\varphi\varphi^\top \preceq \zeta \mathrm{Diag}(\varphi)\mathrm{Diag}(\xi)^{-1}, \tag{3.29}$$

where $\mathrm{Diag}(z)$ denotes the diagonal matrix with elements z_i. The value of ζ can be found by solving a small IP. If the IP cannot be solved to optimality, the best available upper bound on ζ should be used, based on, e.g., the optimality gap returned by the IP solver.

Combining (3.26), (3.27), (3.28), (3.29) and the linear constraints on φ^j, we formulate (3.23) as the SDP

$$\max_{Y, Z_1, \ldots, Z_J} \sum_{j=1}^{J} w_j \mathrm{tr}(C_j Z_j)$$

subject to

$$\begin{aligned}
Z_j &\succeq 0, \\
Z_j^{11} &= 1, \\
AZ_j^{\varphi 1} &= h, \\
0 \leq Z_j^{\varphi 1} &\leq 1, \\
AZ_j^{\varphi\varphi} A^\top &= hh^\top, \\
Z_j^{\varphi\varphi} &\geq 0, \\
Z_j^{\varphi\varphi} &\preceq \zeta \mathrm{Diag}(Z_j^{\varphi 1}) \mathrm{Diag}(\xi)^{-1}, \\
Z_j^{dd} &= Y, \\
Y &\geq 0, \\
\mathrm{tr}(PY) &= 1, \\
\mathrm{diag}(Y) &\leq 1_r/\delta,
\end{aligned}$$

for $j = 1, ..., J$.

After solving this SDP, we obtain the matrix Y. With $\mathrm{rank}(Y) = 1$, $\frac{d_\psi}{||d_\psi||}$ is equivalent to the unique normalized eigenvector of Y. After relaxing the rank-1 condition, we can approximate $\frac{d_\psi}{||d_\psi||}$ by the normalized eigenvector corresponding to the largest eigenvalue of Y. We define this eigenvector as v and let $\tilde{v} = \frac{v}{\max_j(v_j)}$, which satisfies $\tilde{v} \in [0,1]^r$. Since ψ and d_ψ only differ by a scaling factor, we can interpret $\tilde{v}$ as a fractional approximation for the binary $\psi \in \{0,1\}^r$. To recover a binary ψ, we can perform a rounding procedure by solving the small IP given by

$$\begin{array}{rrcl} \min_{\psi,z} & 1_r^\top z & & \\ \text{s.t.} & \psi - \tilde{v} & \leq & z, \\ & \tilde{v} - \psi & \leq & z, \\ & A\psi & = & h \\ & \psi & \in & \{0,1\}^r. \end{array} \tag{3.30}$$

This problem projects the solution $\tilde{v}$ onto $\{0,1\}^r$ by minimizing the L_1-norm of the difference between ψ and $\tilde{v}$. While any IP can potentially be difficult to solve, note that the linear IP in (3.30) is substantially simpler than the nonlinear, nonconvex IP defined by (3.13). In practice, the SDP procedure provides considerable computational savings over the basic version of KGUP that enumerates every alternative.

3.5 Numerical experiments

We now study the practical performance of the KGUP policy and the SDP relaxation. All policies are evaluated through simulation: first, simulated experiments allow a fair comparison between two policies given identical starting conditions; second, by simulating the underlying true values of the alternatives, we are able to quantify how well a policy *could* have done if it had made different decisions; third, simulations allow us to perform large numbers of experiments and identify statistically significant distinctions between policies. However, we use the non-profit fundraising application to provide a realistic context for the simulations. The starting prior parameters θ^0, Σ^0 are taken from the results of the empirical analysis in [134], which were estimated based on historical data. In the following, we consider the effects of different a^0, b^0 (beliefs about the sample variance) on performance. The value of r can be adjusted to experiment with different problem sizes by simply fixing some of the features (e.g., we may constrain ourselves to the Acquisition campaign type while varying the designs and donor segments).

The time horizon in the fundraising application is typically small, and we consider $N = 20$ in our experiments (e.g., 20 monthly campaigns). For each policy we consider, we simulate 100 macro-replications in order to obtain statistically significant results. Within each macro-replication, we generate a single "true" value of ρ from the prior $Gamma\left(a^0, b^0\right)$, and then generate a set of "true" coefficients $\beta \sim \mathcal{N}\left(\theta^0, \frac{1}{\rho}\Sigma^0\right)$. However, these true values are not observed by any policy when making decisions. Each policy chooses a design ψ^n according to some decision rule and observes $\eta^{n+1} \sim \mathcal{N}\left(\beta^\top \psi^n, \frac{1}{\rho}\right)$. After N measurements, the performance of the policy π is evaluated by letting ψ^π be the solution that optimizes $V\left(\theta^N\right)$ and computing the normalized opportunity cost

$$C^{\pi,N} = \frac{\max_{\varphi \in \Phi} \varphi^\top \beta - \left(\psi^\pi\right)^\top \beta}{\max_{\varphi,\psi \in \Phi} (\varphi^\top \beta - \psi^\top \beta)}. \tag{3.31}$$

The denominator in (3.31) confines $C^{\pi,N}$ to be in $[0, 1]$. We also compute the *precision estimation error*, defined as $|\rho - \frac{a^n}{b^n}|$.

The literature on simulation optimization typically considers settings where the value of a single decision or alternative is only observable from a black box, without the additional structure imposed by the regression model. However, a few benchmarks are available. We compare the following policies:

Knowledge gradient with unknown precision (KGUP). The exact version of the policy calculates (3.13) by enumerating the alternatives. We include this policy in order to evaluate the loss incurred by the SDP relaxation. Our experiments consider problems where the enumeration can still be performed.

KGUP with SDP relaxation (KGUP$_3$). The SDP relaxation is computed as described in Section 3.4.2. We used $J = 10$ in the quantization procedure; see the Appendix for the values of t^J.

Knowledge gradient with correlated beliefs (CKG). We implement the policy of [124], a VIP designed for regression models, but with a learning mechanism that assumes known sampling variance (i.e., that $\rho = \frac{a^0}{b^0}$). Like KGUP, the CKG policy enumerates the alternatives (no faster version is available).

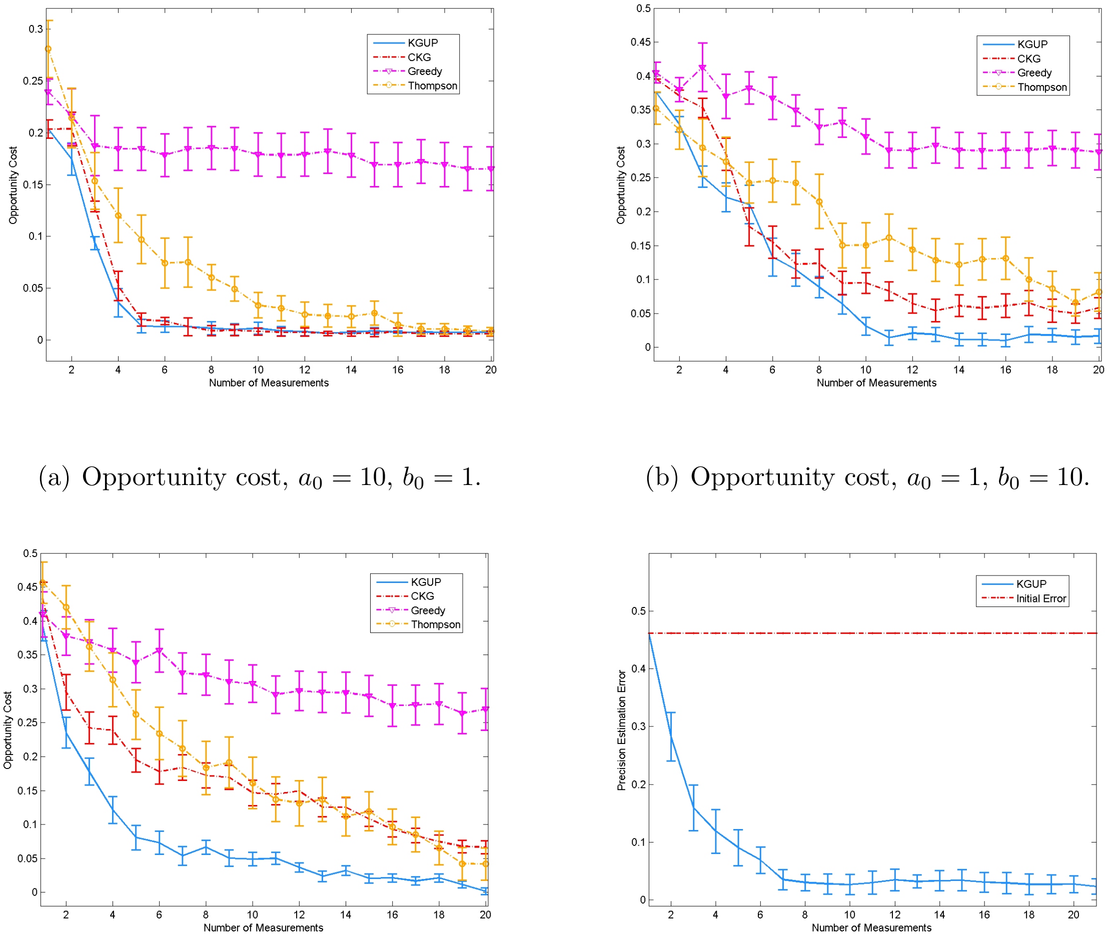
Greedy policy (Greedy). The greedy heuristic implements the argmax of $V(\theta^n)$ at time n, simply replacing the unknown coefficients with their current point estimates. The decision can thus be obtained by solving a small IP.

Thompson sampling (Thompson). The Thompson sampling policy has attracted recent attention [142] because it is easy to implement and enjoys theoretical guarantees on the rate of convergence in some settings. In our problem, the policy first draws a single sample $\hat{\beta}^n$ from the time-n posterior distribution, namely the marginal distribution of β given the normal-gamma parameters $(\theta^n, \Sigma^n, a^n, b^n)$. Then, the policy implements the argmax of $V\left(\hat{\beta}^n\right)$. Thus, it is very similar to a greedy policy, but uses a random sample instead of a point estimate, thus promoting more exploration.

It is important to note that, in the following experiments, we keep r relatively small (so that K is a few hundred or thousand) because one of our main benchmarks (CKG) works by enumerating alternatives, making it computationally expensive to run multiple sample paths. Furthermore, we wish to evaluate the SDP relaxation against the exact KGUP policy, which also enumerates the decision space. However, toward the end of this study, we show that the SDP relaxation provides substantial computational savings in a problem with over $100,000$ alternatives.

In the first experiment, we focus on the value of modeling unknown variance, so KGUP$_3$ is omitted from the comparison. We let $r = 10$, where the first nine features are directly controllable and the tenth is an interaction term. Thus, $K = 2^9 = 512$. Figures 3.1(a)-3.1(c) illustrate $C^{\pi,N}$ as a function of N for different policies π under different starting conditions for a^0, b^0. We average performance over 100 macro-replications and report 95% confidence intervals. The KGUP policy significantly outperforms CKG for small $\frac{a^0}{b^0}$ (Figs. 3.1(b) and 3.1(c)), when the sampling variance tends to be large and exhibits more variation between instances. However, when $\frac{a^0}{b^0}$ is large, there is much less variation in ρ between macro-replications and CKG is competitive with KGUP. The Thompson policy yields similar performance to CKG overall. Finally, Figure 3.1(d) shows how the sampling precision ρ is learned over time by KGUP; we see that learning occurs rapidly within the first few measurements. The other policies are omitted from Figure 3.1(d), but we observed that they learn the variance at about the same rate as KGUP (except CKG, which assumes that the variance is known).

In the second experiment, we consider the performance loss incurred by using KGUP$_3$ to approximate the KG computation instead of enumerating the alterna-



(a) Opportunity cost, $a_0 = 10$, $b_0 = 1$.

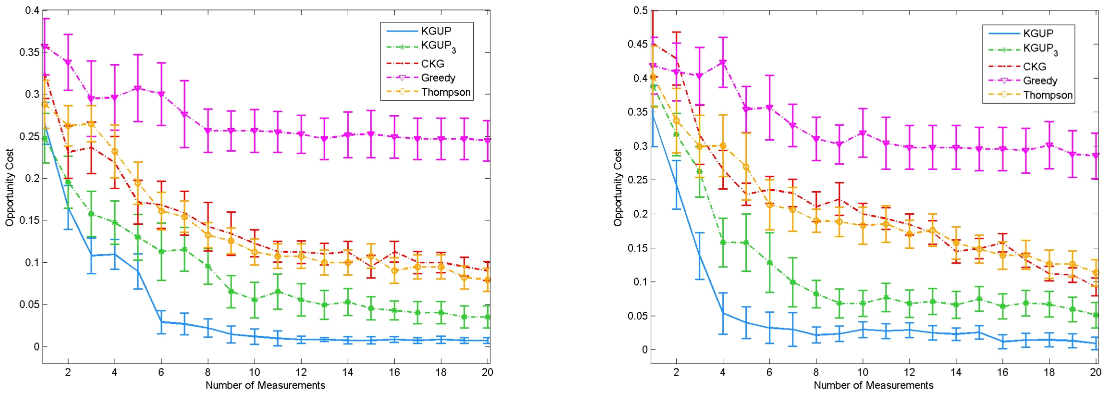
(b) Opportunity cost, $a_0 = 1$, $b_0 = 10$.

(c) Opportunity cost, $a_0 = 0.5$, $b_0 = 1$.

(d) Opportunity cost, $a_0 = 0.5$, $b_0 = 1$.

Figure 3.1: Averaged opportunity cost and precision estimation error over time.

tives. We test the performance of the SDP relaxation for two problem sizes. First, we consider the problem from the first experiment with $K = 512$; second, we consider a larger problem with 3 more independent features and 3 more interaction terms, so that $K = 4096$. The prior for ρ is set to be $a^0 = 0.5$, $b^0 = 1$ in all cases. Figure 3.2 shows that the KGUP$_3$ policy continues to outperform CKG, despite the fact that CKG is still allowed to enumerate the alternatives. This suggests that the



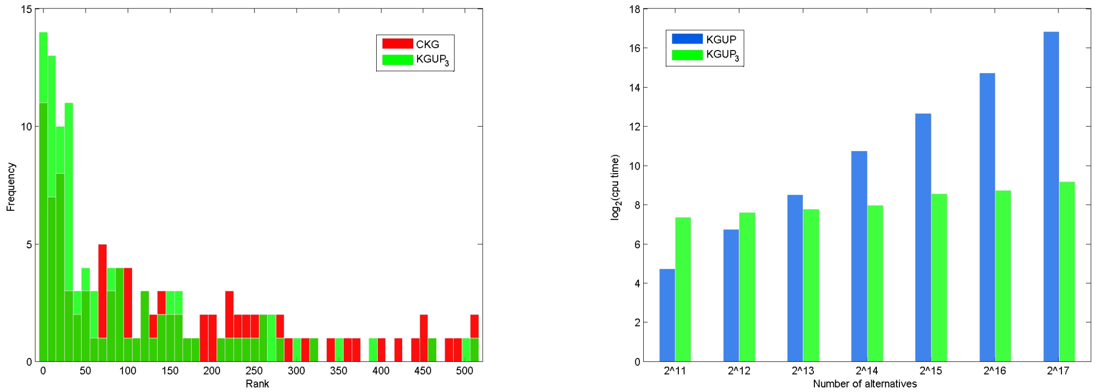
(a) Opportunity cost, $K = 512$.

(b) Opportunity cost, $K = 4096$.

Figure 3.2: Averaged opportunity cost over time.

SDP relaxation provides a reasonably good approximation to KGUP.

Finally, we note that both CKG and KGUP_3 can be viewed as approximate versions of KGUP, in the sense that they both seek to identify alternatives with high value of information, but do not calculate that value exactly in the unknown-variance setting. To test the quality of these approximations, we randomly generate 100 priors for β using $\theta^0 \sim N(0_r, I_{r\times r})$ and $\Sigma^0 = (s + s^\top)(s + s^\top)$, where $s_{i,j} \ \sim N(0,1), \forall 1 \leq i, j \leq r$. The feasible regions for the test problems remain the same as before. For each of these 100 priors, we compute the approximate value of information using KGUP_3 (or CKG), and find the alternative φ' that maximizes this quantity. We then rank the alternatives according to their precise values of information (as computed by KGUP) and see how highly φ' places in that ranking. Figure 3.3(a) shows that KGUP_3 produces a better approximation to the KGUP policy than CKG; over 50% of the alternatives chosen by KGUP_3 have values of information ranked in the top 50 (out of 512).



(a) Empirical distribution of value of information ranks.

(b) Averaged $\log_2$(CPU time) for one iteration.

Figure 3.3: Accuracy and efficiency assessment of KGUP$_3$ using simulated priors for β.

We also compare the computational complexity of KGUP and KGUP$_3$ for various problem sizes. Figure 3.3(b) reports the computational costs (in log-scale) for increasing values of r where we successively add a new independent feature into each problem. The computational cost of KGUP increases exponentially (linear increase in logarithm), whereas the cost of KGUP$_3$ grows much more slowly. This indicates that KGUP$_3$ will run much faster than KGUP (or CKG) when the number of alternatives is large. We note that, for $K = 2^{17}$, a single iteration of KGUP takes about 30 hours to run, whereas KGUP$_3$ takes under 10 minutes.

3.6 Conclusion

We have proposed a framework for information collection in regression-based optimization where we have the ability to select features from a combinatorial space. Such problems arise in applications of business analytics where statistical estimation alternates with decision-making, and we may engage in a limited amount of experimentation to learn about the uncertainty in the statistical model. In particular, this challenge arises in the problem of designing a fundraising campaign for a non-profit organization.

We derived a value of information policy for parametric (regression-based) beliefs with unknown sampling variance. This policy improves upon an existing policy that assumes known variance; however, in practice, neither may be practical due to the high computational cost of enumerating a combinatorial set of alternatives. For this purpose, we have proposed an improved algorithm, based on SDP relaxation, that exhibits significant computational savings in problems with large numbers of alternatives. We believe that this approach provides significant value for learning in regression-based optimization with large decision spaces.

Appendix A: Chapter 3 Supplements

A.1 Proof of Proposition 1

Assume that (β, ρ) follows a multivariate normal-gamma distribution with parameters (θ, Σ, a, b). The joint density is given by

$$\begin{aligned} p\left(\beta, \rho \,|\, \theta, \Sigma, a, b\right) &= p\left(\beta \,|\, \rho, \theta, \Sigma\right) p\left(\rho \,|\, a, b\right) \\ &= \left(\frac{\rho}{2\pi}\right)^{\frac{r}{2}} \left|\Sigma\right|^{-\frac{1}{2}} e^{-\frac{\rho}{2}(\beta-\theta)^\top \Sigma^{-1}(\beta-\theta)} \frac{b^a}{\Gamma\left(a\right)} \rho^{a-1} e^{-b\rho} \end{aligned}$$

where Γ is the gamma function. Let $\eta \sim \mathcal{N}\left(\varphi^\top \beta, \frac{1}{\rho}\varphi^\top \Sigma \varphi\right)$ be the observation corresponding to the chosen feature vector φ. From Bayes' rule [143], we know that $p\left(\beta, \rho \,|\, \eta\right)$ is proportional to $p\left(\beta, \rho\right) q\left(\eta \,|\, \beta, \rho\right)$. We then write,

$$\begin{aligned} p\left(\beta, \rho\right) q\left(\eta \,|\, \beta, \rho\right) &= \left(\frac{\rho}{2\pi}\right)^{\frac{r}{2}} \left|\Sigma\right|^{-\frac{1}{2}} e^{-\frac{\rho}{2}(\beta-\theta)^\top \Sigma^{-1}(\beta-\theta)} \frac{b^a}{\Gamma\left(a\right)} \rho^{a-1} e^{-b\rho} \sqrt{\frac{\rho}{2\pi}} e^{-\frac{\rho}{2}\left(\eta - \phi^\top \beta\right)^2} \\ &= \left(\frac{\rho}{2\pi}\right)^{\frac{r}{2}} \left|\Sigma\right|^{-\frac{1}{2}} e^{-\frac{\rho}{2}\left[(\beta-\theta)^\top \Sigma^{-1}(\beta-\theta) + \left(\eta - \varphi^\top \beta\right)^2\right]} \frac{b^a}{\Gamma\left(a\right)} \rho^{a+\frac{1}{2}-1} e^{-b\rho} \frac{1}{\sqrt{2\pi}}. \end{aligned}$$

Define

$$\begin{aligned} \theta' &= \theta + \frac{\eta - \varphi^\top \theta}{1 + \varphi^\top \Sigma \varphi} \Sigma \varphi, \\ \Sigma' &= \Sigma - \frac{\Sigma \varphi \varphi^\top \Sigma}{1 + \varphi^\top \Sigma \varphi}. \end{aligned}$$

By completing the square for β, and using the matrix inversion lemma to observe that $\Sigma' = \left(\Sigma^{-1} + \varphi\varphi^{\top}\right)^{-1}$, we obtain

$$(\beta - \theta)^{\top} \Sigma^{-1} (\beta - \theta) + \left(\eta - \varphi^{\top}\beta\right)^{2} = (\beta - \theta')^{\top} (\Sigma')^{-1} (\beta - \theta') + \frac{\left(\eta - \varphi^{\top}\theta\right)^{2}}{1 + \varphi^{\top}\Sigma\varphi}.$$

It follows that

$$p(\beta, \rho)\, q(\eta \,|\, \beta, \rho) = \left(\frac{\rho}{2\pi}\right)^{\frac{r}{2}} |\Sigma'|^{-\frac{1}{2}} e^{-\frac{\rho}{2}(\beta-\theta')^{\top}(\Sigma')^{-1}(\beta-\theta')} \frac{b^{a}}{\Gamma(a)} \rho^{a+\frac{1}{2}-1} e^{-\rho\left(b + \frac{\left(\eta-\varphi^{\top}\theta\right)^{2}}{2(1+\varphi^{\top}\Sigma\varphi)}\right)} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{|\Sigma'|}{|\Sigma|}}.$$

Letting $a' = a + \frac{1}{2}$ and

$$b' = b + \frac{\left(\eta - \varphi^{\top}\theta\right)^{2}}{2\left(1 + \varphi^{\top}\Sigma\varphi\right)},$$

we obtain

$$p(\beta, \rho)\, q(\eta \,|\, \beta, \rho) \quad \propto \quad \left(\frac{\rho}{2\pi}\right)^{\frac{r}{2}} |\Sigma'|^{-\frac{1}{2}} e^{-\frac{\rho}{2}(\beta-\theta')^{\top}(\Sigma')^{-1}(\beta-\theta')} \frac{(b')^{a'}}{\Gamma(a')} \rho^{a'-1} e^{-\rho b'},$$

which is precisely the normal-gamma density with parameters $(\theta', \Sigma', a', b')$ calculated according to the desired updating equations.

A.2 Tables of Voronoi quantizations for the Student's t-distribution

In this appendix, we present tables of two Voronoi quantizations t^{J} of the standard Student's t-distribution with varying degrees of freedom s. The quantization does not exist when $s = 1$, as the mean of the distribution is undefined in this case, or when $s = 2$, since this corresponds to infinite variance. As s increases, the quantization approaches that of the standard normal distribution, which we also include.

s	t_1	t_2	t_3	t_4	t_5
3	-5.6124	-1.5520	0.0000	1.5520	5.6124
4	-3.4130	-1.1977	0.0000	1.1977	3.4130
5	-2.7943	-1.0636	0.0000	1.0636	2.7943
6	-2.5065	-0.9929	0.0000	0.9929	2.5065
7	-2.3406	-0.9493	0.0000	0.9493	2.3406
8	-2.2327	-0.9197	0.0000	0.9197	2.2327
9	-2.1569	-0.8982	0.0000	0.8982	2.1569
10	-2.1008	-0.8820	0.0000	0.8820	2.1008
11	-2.0576	-0.8693	0.0000	0.8693	2.0576
12	-2.0233	-0.8591	0.0000	0.8591	2.0233
13	-1.9954	-0.8507	0.0000	0.8507	1.9954
14	-1.9722	-0.8436	0.0000	0.8436	1.9722
15	-1.9527	-0.8377	0.0000	0.8377	1.9527
16	-1.9361	-0.8325	0.0000	0.8325	1.9361
17	-1.9217	-0.8281	0.0000	0.8281	1.9217
18	-1.9091	-0.8242	0.0000	0.8242	1.9091
19	-1.8980	-0.8207	0.0000	0.8207	1.8980
20	-1.8882	-0.8176	0.0000	0.8176	1.8882
∞	-1.7241	-0.7646	0.0000	0.7646	1.7241

Table A.1: Optimal quantizations with $J = 5$.

s	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
3	-22.3881	-7.3823	-3.2540	-1.4892	-0.4392	0.4392	1.4892	3.2540	7.3823	22.3881
4	-7.8620	-3.7060	-2.0375	-1.0601	-0.3315	0.3315	1.0601	2.0375	3.7060	7.8620
5	-5.3116	-2.8550	-1.6854	-0.9116	-0.2900	0.2900	0.9116	1.6854	2.8550	5.3116
6	-4.3417	-2.4886	-1.5190	-0.8364	-0.2682	0.2682	0.8364	1.5190	2.4886	4.3417
7	-3.8426	-2.2862	-1.4222	-0.7912	-0.2549	0.2549	0.7912	1.4222	2.2862	3.8426
8	-3.5409	-2.1581	-1.3590	-0.7610	-0.2459	0.2459	0.7610	1.3590	2.1581	3.5409
9	-3.3396	-2.0697	-1.3144	-0.7394	-0.2394	0.2394	0.7394	1.3144	2.0697	3.3396
10	-3.1959	-2.0052	-1.2813	-0.7231	-0.2345	0.2345	0.7231	1.2813	2.0052	3.1959
11	-3.0884	-1.9560	-1.2558	-0.7105	-0.2306	0.2306	0.7105	1.2558	1.9560	3.0884
12	-3.0049	-1.9173	-1.2355	-0.7005	-0.2276	0.2276	0.7005	1.2355	1.9173	3.0049
13	-2.9382	-1.8860	-1.2190	-0.6922	-0.2250	0.2250	0.6922	1.2190	1.8860	2.9382
14	-2.8837	-1.8601	-1.2053	-0.6853	-0.2229	0.2229	0.6853	1.2053	1.8601	2.8837
15	-2.8384	-1.8385	-1.1937	-0.6795	-0.2212	0.2212	0.6795	1.1937	1.8385	2.8384
16	-2.8001	-1.8200	-1.1839	-0.6746	-0.2196	0.2196	0.6746	1.1839	1.8200	2.8001
17	-2.7673	-1.8042	-1.1753	-0.6702	-0.2183	0.2183	0.6702	1.1753	1.8042	2.7673
18	-2.7389	-1.7904	-1.1679	-0.6665	-0.2172	0.2172	0.6665	1.1679	1.7904	2.7389
19	-2.7141	-1.7782	-1.1613	-0.6631	-0.2161	0.2161	0.6631	1.1613	1.7782	2.7141
20	-2.6922	-1.7675	-1.1555	-0.6602	-0.2152	0.2152	0.6602	1.1555	1.7675	2.6922
∞	-2.3451	-1.5913	-1.0578	-0.6099	-0.1996	0.1996	0.6099	1.0578	1.5913	2.3451

Table A.2: Optimal quantizations with $J = 10$.

Bibliography

[1] T. Hastie, J. Friedman, and R. Tibshirani. *The elements of statistical learning.* Springer, 2009.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in knowledge discovery and data mining. *the MIT Press*, 1996.

[3] T. M. Mitchell. *Machine Learning.* Burr Ridge, IL: McGraw Hill, 1997.

[4] S. M. Stigler. *The history of statistics: the measurement of uncertainty before 1900.* Harvard University Press, 1986.

[5] A. C. Rencher and G. B. Schaalje. *Linear models in statistics.* John Wiley & Sons, 2008.

[6] P. McCullagh and J. A. Nelder. *Generalized linear models.* Chapman & Hall, New York, 2nd edition, 1999.

[7] G. K Smyth and B. Jørgensen. Fitting tweedie's compound poisson model to insurance claims data: Dispersion modelling. *Astin Bulletin*, 32(01):143–157, 2002.

[8] T. J. Hastie and R. J. Tibshirani. *Generalized additive models*, volume 43. CRC Press, 1990.

[9] Charles E. McCulloch and Shayle R. Searle. *Generalized, linear, and mixed models.* John Wiley & Sons, Inc., New York, 2000.

[10] G. McLachlan and T. Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[11] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *Signal Processing, IEEE Transactions on*, 42(10):2664–2677, 1994.

[12] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[13] R. W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[14] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is nearest neighbor meaningful? In *Database TheoryICDT99*, pages 217–235. Springer, 1999.

[15] P. C. Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[16] C. J.C. Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[17] R. Burbidge and B. Buxton. An introduction to support vector machines for data mining. *Keynote papers, young OR12*, pages 3–15, 2001.

[18] R. Hecht-Nielsen. Theory of the backpropagation neural network. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, pages 593–605. IEEE, 1989.

[19] J. J. Faraway. *Linear models with R*. CRC Press, 2014.

[20] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[21] J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[22] W. Gersten, R. Wirth, and D. Arndt. Predictive modeling in automotive direct marketing: tools, experiences and open issues. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 398–406. ACM, 2000.

[23] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[24] N. R. Draper, H. Smith, and E. Pownell. *Applied regression analysis*, volume 3. Wiley New York, 1966.

[25] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

[26] M.A. Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, 1:191–203, 1960.

[27] K. Person. On lines and planes of closest fit to systems of points in space. *philosophical magazine*, 2(6):559–572, 1901.

[28] I. Jolliffe. *Principal component analysis.* Wiley Online Library, 2002.

[29] G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420, 1970.

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

[31] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer, 1998.

[32] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[33] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[34] L. Meier, S. Van De Geer, and P. Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[35] J. Schelldorfer, P. Bühlmann, G. DE, and S. VAN. Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

[36] C. Rudin, D. Waltz, R. N. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. N. Gross, B. Huang, S. Ierome, et al. Machine learning for the new york city power grid. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):328–345, 2012.

[37] J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[38] T. Hothorn, P. Bühlmann, T. Kneib, M. Schmid, and B. Hofner. Model-based boosting 2.0. *The Journal of Machine Learning Research*, 11:2109–2113, 2010.

[39] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap.* CRC press, 1994.

[40] R. Srinivasan. *Importance sampling: applications in communications and detection.* Springer Science & Business Media, 2002.

[41] P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, 2002.

[42] J. Bradic. Support recovery via weighted maximum-contrast subagging. *Preprint*, 2015.

[43] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society*, B76(4):795–816, 2014.

[44] A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. The big data bootstrap. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1759–1766, 2012.

[45] J. Bradić. Support recovery via weighted maximum-contrast subagging. *Arxiv preprint arXiv:1306.3494v3*, 2014.

[46] G. Hadley and G. Hadley. *Linear programming*, volume 4. Addison-Wesley Reading, MA, 1962.

[47] G. B. Dantzig. *Linear programming and extensions.* Princeton university press, 1998.

[48] R. S. Garfinkel and G. L. Nemhauser. *Integer programming*, volume 4. Wiley New York, 1972.

[49] L. A. Wolsey. *Integer programming*, volume 42. Wiley New York, 1998.

[50] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, UK, 2004.

[51] J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis.* Springer Science & Business Media, 2001.

[52] N. Z. Shor, K. C. Kiwiel, and A. Ruszcayski. *Minimization methods for non-differentiable functions.* Springer-Verlag New York, Inc., 1985.

[53] C.S. James. Introduction to stochastics search and optimization, 2003.

[54] W. B. Powell and I. O. Ryzhov. *Optimal learning.* John Wiley & Sons, Inc., Hoboken, New Jersey, 2012.

[55] S. S. Gupta and K. J. Miescke. Bayesian look ahead one-stage sampling allocations for selection of the best population. *Journal of Statistical Planning and Inference*, 47(3):229–244, 1996.

[56] S. E. Chick. Subjective probability and bayesian methodology. *Handbooks in Operations Research and Management Science*, 13:225–257, 2006.

[57] K. J. Miescke. Bayes sampling designs for selection procedures. In S. Ghosh, editor, *Multivariate, Design, and Sampling.* M. Dekker, New York, 1999.

[58] T. Minka. Bayesian linear regression. Technical report, Citeseer, 2000.

[59] C. J. C. H. Watkins. *Learning from delayed rewards.* PhD thesis, University of Cambridge England, 1989.

[60] A. G. Barto. *Reinforcement learning: An introduction.* MIT press, 1998.

[61] C. JCH Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[62] W. B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. John Wiley & Sons, 2nd edition, 2011.

[63] S.-H. Kim and B. L. Nelson. Recent advances in ranking and selection. In *Proceedings of the 39th conference on Winter simulation: 40 years! The best is yet to come*, pages 162–172. IEEE Press, 2007.

[64] R. M. Tomasini and L. N. Van Wassenhove. From preparedness to partnerships: case study research on humanitarian logistics. *International Transactions in Operational Research*, 16:549–559, 2009.

[65] J. Meer and H. S. Rosen. The ABCs of charitable solicitation. *Journal of Public Economics*, 95(5):363–371, 2011.

[66] A. Jones and J. Posnett. Charitable donations by UK households: evidence from the Family Expenditure Survey. *Applied Economics*, 23(2):343–351, 1991.

[67] H. Kitchen. Determinants of charitable donations in Canada: a comparison over time. *Applied Economics*, 24(7):709–713, 1992.

[68] S. Brown, M. N. Harris, and K. Taylor. Modelling charitable donations to an unexpected natural disaster: Evidence from the US Panel Study of Income Dynamics. *Journal of Economic Behavior & Organization*, 84(1):97–110, 2012.

[69] PSID. U.S. Panel Study of Income Dynamics. `http://psidonline.isr.umich.edu/`, 2012.

[70] J. A. List. The market for charitable giving. *Journal of Economic Perspectives*, 25(2):157–180, 2011.

[71] P. H. Brown and J. H. Minty. Media coverage and charitable giving after the 2004 tsunami. *Southern Economic Journal*, 75(1):9–25, 2008.

[72] J. A. List and Y. Peysakhovich. Charitable donations are more responsive to stock market booms than busts. *Economics Letters*, 110:166–169, 2011.

[73] F. de Véricourt and M. S. Lobo. Resource and revenue management in nonprofit operations. *Operations Research*, 57(5):1114–1128, 2009.

[74] N. Privett and F. Erhun. Efficient funding: Auditing in the nonprofit sector. *Manufacturing & Service Operations Management*, 13(4):471–488, 2011.

[75] R. W. Lien, S. M. R. Iravani, and K. R. Smilowitz. Sequential resource allocation for nonprofit operations. *Operations Research*, 62(2):301–317, 2014.

[76] P.T.L. Leszczyc and M.H. Rothkopf. Charitable motives and bidding in charity auctions. *Management Science*, 56(3):399–413, 2010.

[77] S. T. Yen. An econometric analysis of household donations in the USA. *Applied Economics Letters*, 9(13):837–841, 2002.

[78] C. Landry, A. Lange, J. A. List, M. K. Price, and N. G. Rupp. Toward an understanding of the economics of charity: Evidence from a field experiment. *Quarterly Journal of Economics*, 121(2):747–782, 2006.

[79] D. B. Arnett, S. D. German, and S. D. Hunt. The identity salience model of relationship marketing success: The case of nonprofit marketing. *Journal of Marketing*, 67(2):89–105, 2003.

[80] B. M. Fennis, L. Janssen, and K. D. Vohs. Acts of benevolence: A limited-resource account of compliance with charitable requests. *Journal of Consumer Research*, 35(6):906–924, 2009.

[81] J. Shang, A. Reed, and R. Croson. Identity congruency effects on donations. *Journal of Marketing Research*, 45(3):351–361, 2008.

[82] M. Van Diepen, B. Donkers, and P. H. Franses. Does irritation induced by charitable direct mailings reduce donations? *International Journal of Research in Marketing*, 26(3):180–188, 2009.

[83] D. Karlan and J. A. List. Does price matter in charitable giving? Evidence from a natural field experiment. *American Economic Review*, 97(5):1774–1793, 2007.

[84] D. Karlan, J. A. List, and E. Shafir. Small matches and charitable giving: evidence from a natural field experiment. *Journal of Public Economics*, 95:344–350, 2011.

[85] J. Andreoni. Philanthropy. In S.-C. Kolm and J. M. Ythier, editors, *Handbook on the Economics of Giving, Reciprocity and Altruism*, volume 2, pages 1201–1269. Elsevier, 2006.

[86] E. Schokkaert. The empirical analysis of transfer motives. In S.-C. Kolm and J. M. Ythier, editors, *Handbook on the Economics of Giving, Reciprocity and Altruism*, volume 1, pages 127–181. Elsevier, 2006.

[87] R. H. Lankford and J. H. Wyckoff. Modeling charitable giving using a Box-Cox standard Tobit model. *The Review of Economics and Statistics*, 73(3):460–470, 1991.

[88] G. Auten and D. Joulfaian. Charitable contributions and intergenerational transfers. *Journal of Public Economics*, 59(1):55–68, 1996.

[89] N. Lacetera, M. Macis, and R. Slonim. Will there be blood? Incentives and displacement effects in pro-social behavior. *American Economic Journal: Economic Policy*, 4(1):186–223, 2012.

[90] B. A. Lafferty, R. E. Goldsmith, and G.T.M. Hult. The impact of the alliance on the partners: A look at cause–brand alliances. *Psychology & Marketing*, 21(7):509–531, 2004.

[91] D. Ariely, A. Bracha, and S. Meier. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *The American Economic Review*, 99(1):544–555, 2009.

[92] A. J. Pedraza-Martinez and L. N. Van Wassenhove. Vehicle replacement in the International Committee of the Red Cross. *Production and Operations Management*, 22(2):365–376, 2013.

[93] J. R. Bult, H. Van der Scheer, and T. Wansbeek. Interaction between target and mailing characteristics in direct marketing, with an application to health care fund raising. *International Journal of Research in Marketing*, 14(4):301–308, 1997.

[94] WCAI. Cultivating disaster donors: A WCAI research opportunity sponsored by Russ Reid and the American Red Cross. `http://www.wharton.upenn.edu/wcai/files/Russ_Reid-ARC_Webinar.pdf`, 2012.

[95] N. M. Laird and J. H. Ware. Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[96] G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions (2nd ed.)*. Wiley-Interscience, 2008.

[97] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Transactions on Signal Processing*, 42(10):2664–2677, 1994.

[98] A. T. Karl, Y. Yang, and S. L. Lohr. Computation of maximum likelihood estimates for multiresponse generalized linear mixed models with non-nested, correlated random effects. *Computational Statistics & Data Analysis*, 73:146–162, 2014.

[99] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer series in Statistics, New York, NY, 2001.

[100] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[101] H. Wang, G. Li, and C.-L. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, B69(1):63–78, 2007.

[102] H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, 1973.

[103] L. Breiman and P. Spector. Submodel selection and evaluation in regression. The x-random case. *International Statistical Review*, 60(3):291–319, 1992.

[104] H. Zou, T. Hastie, and R. Tibshirani. On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.

[105] D. Homrighausen and D. J. McDonald. The lasso, persistence, and cross-validation. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[106] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.

[107] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[108] J. Fan, F. Han, and H. Liu. Challenges of big data analysis. *National Science Review*, 1(2):293–314, 2014.

[109] Y. Fan and R. Li. Variable selection in linear mixed effects models. *The Annals of Statistics*, 40(4):2043–2068, 2012.

[110] Y. Zhang, R. Li, and C.-L. Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105(489):312–323, 2010.

[111] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society*, B72(4):417–473, 2010.

[112] B. Efron. Estimation and accuracy after model selection. *Journal of the American Statistical Association*, 109(507):991–1007, 2014.

[113] M. Smithson and E. C. Merkle. *Generalized linear models for categorical and continuous limited dependent variables*. Chapman & Hall/CRC, 2014.

[114] R. L. Schaefer. Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation*, 25(1-2):75–91, 1986.

[115] B. Breeze and J. Dean. Pictures of me: user views on their representation in homelessness fundraising appeals. *International Journal of Nonprofit and Voluntary Sector Marketing*, 17(2):132–143, 2012.

[116] R. Bennett and R. Kottasz. Emergency fund-raising for disaster relief. *Disaster Prevention and Management*, 9(5):352–360, 2000.

[117] R. Bennett. Impulsive donation decisions during online browsing of charity websites. *Journal of Consumer Behaviour*, 8(2-3):116–134, 2009.

[118] W. D. Diamond and S. Gooding-Williams. Using advertising constructs and methods to understand direct mail fundraising appeals. *Nonprofit Management and Leadership*, 12(3):225–242, 2002.

[119] S.-H. Kim and B. L. Nelson. Selecting the best system. In S. Henderson and B. Nelson, editors, *Handbooks of Operations Research and Management Science, vol. 13: Simulation*, pages 501–534. North-Holland Publishing, Amsterdam, 2006.

[120] S.-H. Kim. Statistical ranking and selection. In *Encyclopedia of Operations Research and Management Science*, pages 1459–1469. Springer, 2013.

[121] P. I. Frazier, W. B. Powell, and S. Dayanik. The knowledge-gradient policy for correlated normal rewards. *INFORMS Journal on Computing*, 21(4):599–613, 2009.

[122] H. Qu, I. O. Ryzhov, and M. C. Fu. Ranking and selection with unknown correlation structures. In C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, editors, *Proceedings of the 2012 Winter Sirnulation Conference*. Institute of Electrical and Electronics Engineers, Inc., 2012.

[123] S. E. Chick, J. Branke, and C. Schmidt. Sequential sampling to myopically maximize the expected value of information. *INFORMS Journal on Computing*, 22(1):71–80, 2010.

[124] D. M. Negoescu, P. I. Frazier, and W. B. Powell. The knowledge gradient algorithm for sequencing experiments in drug discovery. *INFORMS Journal on Computing*, 23(3):346–363, 2011.

[125] L. J. Hong and B. L. Nelson. A brief introduction to optimization via simulation. In *Winter Simulation Conference*, pages 75–85. Winter Simulation Conference, 2009.

[126] C.-H. Chen, J. Lin, E. Yücesan, and S. E. Chick. Simulation budget allocation for further enhancing the efficiency of ordinal optimization. *Discrete Event Dynamic Systems*, 10(3):251–270, 2000.

[127] C. H. Chen, C. Chun-hung, et al. *Stochastic simulation optimization: an optimal computing budget allocation*. World scientific, 2010.

[128] M. C. Fu, J.-Q. Hu, C.-H. Chen, and X. Xiong. Simulation allocation for determining the best design in the presence of correlated sampling. *INFORMS Journal on Computing*, 19(1):101–111, 2007.

[129] S. Modaresi, D. Saure, and J. P. Vielma. Learning in combinatorial optimization: What and how to explore. Technical report, Submitted for publication, 2013.

[130] B. Defourny, I. O. Ryzhov, and W. B. Powell. Optimal informational blending with measurement in the l2 sphere. *Submitted for publication*, 2013.

[131] V. I. Norkin, G. C. Pflug, and A. Ruszczyński. A branch and bound method for stochastic global optimization. *Mathematical programming*, 83(1-3):425–450, 1998.

[132] W. L. Xu and B. L. Nelson. Empirical stochastic branch-and-bound for optimization via simulation. *IIE Transactions*, 45(7):685–698, 2013.

[133] D. Bertsimas, A. O'Hair, S. Relyea, and J. Silberholz. An analytics approach to designing clinical trials for cancer. *Submitted for publication*, 2013.

[134] I. O. Ryzhov, B. Han, and J. Bradic. Cultivating disaster donors using data analytics. *Management Science*, 2015.

[135] P. I. Frazier, W. B. Powell, and S. Dayanik. A knowledge gradient policy for sequential information collection. *SIAM Journal on Control and Optimization*, 47(5):2410–2439, 2008.

[136] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

[137] E. Çinlar et al. *Probability and stochastics*, volume 261. Springer Science & Business Media, 2011.

[138] F. Eicker et al. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, 34(2):447–456, 1963.

[139] H. Drygas. Weak and strong consistency of the least squares estimators in regression models. *Probability Theory and Related Fields*, 34(2):119–127, 1976.

[140] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions*. Springer-Verlag, Berlin, Germany, 2000.

[141] G. Pages and J. Printems. Optimal quadratic quantization for numerics: the gaussian case. *Monte Carlo Methods and Applications*, 9(2):135–166, 2003.

[142] D. Russo and B. Van-Roy. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

[143] M. H. DeGroot. *Optimal statistical decisions*, volume 82. John Wiley & Sons, 2005.