INTERACTION ANALYSIS OF THE HHP MULTI-DIMENSIONAL MODEL

by

YULIA MALITSKAIA

An honors thesis submitted in partial fulfillment

of the requirements for the degree of

Bachelor of Science

Undergraduate College

Leonard N. Stern School of Business

New York University

May 2013

Professor Marti G. Subrahmanyam

Faculty Advisor

Professor William H. Greene

Thesis Advisor

Contents

Abstract

Advancements in technology have made gathering large arrays of data faster and more efficient. Mining these data sets for useful information is a quickly growing field of research that has had impacts on many areas of application. This trend is particularly represented in the Heritage Health Prize (HHP) competition, which aims to support the development of an efficient algorithm for predicting the number of days patients will spend in the hospital given their medical history. The current winning approaches aim to maximize HHPs performance error and therefore apply blending techniques based on multiple data-mining solvers. The interpretation of such composite models, however, is obscure. As a result, in this project we focus on the development of a scalable approach for the comprehensive analysis of interaction effects in the context of transparent econometric models using the R and LIMDEP econometrics packages.

1 Introduction

Advancements in technology have made gathering large collections of data faster and more efficient. Therefore, organization and analysis of big data is a quickly growing field of research, and the development of this technology has impact on nearly every application domain. Companies realize that by more efficiently exploiting accumulated data, they can potentially increase productivity rates, profitability, make better decisions, and differentiate themselves from competitors (Barton and Court, 2012). One of the main objectives of our study is the analysis of interactions within the complex models. This research is especially meaningful and interesting in the context of real data which can be gathered from data prediction competition platforms like Kaggle.

The Heritage Health Prize (HHP) competition represents one of the current Kaggle-based challenges. It aims to support the development of an efficient algorithm for predicting the number of days patients will spend in the hospital given their medical history. With this information, providers expect to create new care plans that will help minimize unnecessary hospitalization. The competition began two years ago and is scheduled to end on April 4, 2013 when the $3 million prize will be awarded. Every six months, the HHP judging panel evaluates the success of the competing entries and the top two contestants submit papers explaining their approaches.

Focused primarily on minimizing the performance error, the past and current winning approaches apply blending techniques similar to those exploited by past competitions like the Netflix Prize (Töscher et al., 2009). In this scenario, multiple models are each trained to predict the desired outcome and then combined to form the final result. For the first milestone, both the first and second place teams blended the predictions of 20 models. The second place winner solely used stochastic gradient descent, while the first place winner employed a wider range of techniques that included gradient boosting machines, neural networks, bagged trees, and linear models. Subsequently, despite over 900 entrants competing, the same teams won again during the second milestone. In this iteration, the second place team expanded to include 27 models by adding combined

gradient descent and gradient boosting machines. The first place team, however, utilized a mixture of 79 models, which now also included additive groves and multivariate adaptive regression splines. Using a now consolidated set of main features, blending algorithms then won again in the third milestone.

The goal of our research was to develop a scalable approach for a comprehensive analysis of interaction effects. While the blending of algorithms is the primary competition approach, the best score among single algorithms has been achieved with the gradient boosting machine (Friedman, 2001). As its name implies, this technique relies on boosting, the observation that a combination of multiple weak learners can result in an approach that is strongly correlated with the true classification. In the case of the gradient boosting machine (GBM), these weak learners are based on trees that automatically incorporate interactions during the training procedure. Yet even with this single algorithm approach, the composite model is obscure, making interpretation difficult. As a result, the project focuses on the domain of transparent econometrics models (Greene, 2012) and the extension of the corresponding econometrics packages, such as LIMDEP, with a dynamic programming procedure for screening and preselecting the significant interaction effects of multidimensional models. The report outlines the scope and results of this approach conducted in the context of the HHP data model.

The subsequent section provides a brief overview of relevant econometric methods, conventional feature selection methods and the extensions that must be accounted for when working with interactions. Section 4 introduces the step-by-step procedure of the proposed approach. In Section 5, we overview the HHP data model. Section 6 demonstrates this approach in the context of the HHP application, compares numerical results with two alternative approaches and discusses the encountered issues and the corresponding extensions. And finally, Section 7 concludes with a summary and future work.

2 Econometric Models

Econometrics is an active field of research that establishes a consistent estimation framework encompassing parametric, semi-parametric, and nonparametric approaches and applying them to a wide spectrum of real-world problems. Yet, the HHP competition introduces new challenges for actual economics models involving panel data and count data.

In econometrics, panel data is analyzed as a combination of both cross-sections and time-series. This allows a researcher to use the panel data framework to simultaneously evaluate the dynamic and heterogeneous effects across cross-sections. In the HHP model, however, each patients history is specific to that individual and therefore inconsistent with the conventional panel data structure. As a result, all the HHP contestants average monthly records into one-year cross-section data models. Nonetheless, the panel data approach is still valuable for analyzing heterogeneity across age cohorts and evaluating the patients effects on age groups as seen in the following formula:

$$y_{ai} = \boldsymbol{x'}_{ai}\boldsymbol{\beta} + c_a + \epsilon_{ai}$$

where we set a to stand for age groups and i represents the patients. In each age group there is an uneven distribution of patients and therefore we created an unbalanced panel data model. The variable c_a is known as the age-specific heterogeneity effect, and is the focus of the fixed and random effects models. The fixed effects model assumes c_a is correlated with the independent features and treats this effect by adding age-specific constants into the regression model. Contrary to the fixed effects approach, the random effects model considers the independent variables to be exogenous, and c_a is included as a combination of a constant and group-specific heterogeneous random effects.

Both fixed and random effects models have their own pros and cons. A benefit of the fixed effects model is that it makes a realistic assumption about correlation. On the other hand, it does not allow the analysis of time invariant variables and may require the inclusion of many group-

```
+---------------------------------------------------------------------+
|             Test Statistics for the Regression Model                |
+---------------------------------------------------------------------+
|       Model            Log-Likelihood    Sum of Squares  R-squared  |
|(1)  Constant term only   -53403.82792       18138.15458     .00000  |
|(2)  Group effects only   -51703.71995       17344.93239     .04373  |
|(3)  X - variables only   -51335.06679       17177.55916     .05296  |
|(4)  X and group effects  -50570.40566       16835.52483     .07182  |
+---------------------------------------------------------------------+
|                         Hypothesis Tests                            |
|         Likelihood Ratio Test           F Tests                     |
|         Chi-squared   d.f.   Prob         F   num   denom  P value  |
|(2) vs (1)   3400.22      9  .0000    386.33     9   76028   .00000  |
|(3) vs (1)   4137.52     45  .0000     94.44    45   75992   .00000  |
|(4) vs (1)   5666.84     54  .0000    108.87    54   75983   .00000  |
|(4) vs (2)   2266.63     45  .0000     51.09    45   75983   .00000  |
|(4) vs (3)   1529.32      9  .0000    171.52     9   75983   .00000  |
+---------------------------------------------------------------------+
```

Figure 1: LIMDEP test statistics for the fixed effects model.

specific dummy parameters into the model. Alternatively, the random effects model only adds one parameter, provides efficient estimation of coefficients, and supports the analysis of time-invariant variables. One of the major cons, however, is its uncorrelation assumption being too strong for practical applications.

Our panel data analysis began with the Breusch and Pagan Lagrange Multiplier test. This statistic is calculated using the least squares residual values and tests the null hypothesis of "no effects" against the alternative of "some effects". In the case of the analysis of HHPs Primary Condition Group features, the Lagrange Multiplier is 57,296 according to LIMDEP. This implies that the null hypothesis should be rejected. Selection between the fixed and random effects models is usually based on one of two approaches: the Hausman test and the Wu variable addition test. The Hausman test determines whether the delta difference between the fixed and random effects models is significant. This statistic is based on the Wald test and the null hypothesis states that when the delta equals zero both models coefficients have been consistently estimated. Rejection of this test implies the selection of the fixed effects model. The Hausman test, however, has several technical difficulties. For example, in our case the difference of the covariance matrices had a negative root, meaning that the test could not be computed. As a result, we also considered the

Wu variable addition test which expands the random effects model with group means. This test, however, failed again because the group means introduced collinearity into the regression model. Therefore, to determine the significance of the age specific constants, we conducted the F Tests in the context of the fixed effects model. The results can be seen in Figure 1, motivating us to include the age groups as dummy variables in the final dataset.

In addition to the feature selection, another aspect of our study was associated with the selection of the models addressing the count data of the HHP competition. Econometrics traditionally offers a broad spectrum of corresponding approaches for this type of data. In our preliminary independent study we focused on a comparative analysis of the following regression models:

- Linear Regression: a multivariate least square approach
- Poisson Regression: a count data regression model derived from the Poisson distribution
- Negative Binomial: an alternative model to the Poisson approach for allowing over-dispersion.
- Hurdle: an alternative method to Zero Inflation that uses a left-truncated count component like Poisson or Negative Binomial regression for the positive values and an alternate binomial model to decide on whether an instance is a zero or a larger count.
- Zero-Inflation: in addition to the NB or Poisson models, this approach pays extra attention to the zeros in the count data. The zeros are assumed to be generated by two separate distributions, a point mass at zero and count distribution for the remaining data.

From the results documented in the independent study, the negative binomial hurdle performed the best, having calculated the smallest log likelihood and RMSE results. In a typical setting, a Poisson regression would represent the doctor visits, however, the excessive over-dispersion factor makes the model a bad fit for our data since the mean and over-dispersion is captured by a single

parameter. Negative binomial, on the other hand, has an additional fitting parameter which helps resolve over-dispersion.

This initial evaluation, however, was not representative of the competition, which was judged based on the log of the days spent in the hospital. This log-transformation dramatically changed the relative performances, resulting in the log-linear regression having the best RMSE among our econometric models. The reason lies in the utility function of the Jensen's Inequality. In this case, the regression on the logs minimizes the influence of the zeros, and there is a lower penalty for an incorrect prediction.

The independent study established the dataset and identified the best candidate model for the log-transformed data. However, there was still a mismatch in the performance between the log-linear regression and the gradient boosting machine, the best single model submitted in the competition. For our project we decided to extend the independent study to determine a method to minimize this gap. Our assumption was that by adding interaction features into the linear regression, the performance gap between the econometrics model and data-mining approach would be minimized.

3 Feature Selection Methods

The linear regression approach provides a common platform for considering both the main effects and interactions of an arbitrary order. For example, the HHP problem can be expressed through the following set of ordinary log-linear equations:

$$\log(\text{dih} + 1) = \sum^{c_0} a_i x^i + \sum \sum b_{ij} x_i x_j + \dots \tag{1}$$

where dih is the days in hospital, x_i are explanatory variables derived from the medical records of patients, while a_i and b_{ij} are coefficients of the main and interaction features respectively. Despite the fact that all variables in these equations are defined uniformly, interactions exponentially

increase the dimensionality of the task. This in turn introduces various issues, such as model overfitting and complex correlations among covariates. As a result, the development of multi-dimensional data reduction algorithms becomes a necessary prerequisite for the decisive analysis of interaction effects.

The majority of feature selection methods can be divided into two primary categories: backward and forward selection. These approaches rely on different criteria for choosing the variables, but overall the objective is to decrease the dimensionality of the model to the point where the prediction error is minimized. Backward stepwise selection, for example, begins with the entire set of features and aims to eliminate insignificant features one at a time. The significance of the parameters is determined by statistical tests like the z-scores or F-tests. Unlike its counterpart, forward selection iteratively adds features to a previously empty set. There are a number of ways this can be done in practice (Hastie et al., 2009): forward stepwise selection, forward stagewise selection, and Least Angle Regression (LARS). The latter is the most efficient approach built on the integration of several techniques (Efron et al., 2004). Like forward stepwise selection, LARS incrementally adds those features that are most correlated with residuals calculated from the previous steps. However, LARS's step size is shrunk similarly to forward stagewise selection, allowing the implementation of the lasso penalty criteria (Tibshirani, 1994). In contrast with the stagewise approach, this step is optimal and determined by the strength of the next competing feature.

Bringing the interaction effects to the conventional selection methods introduces the new issue associated with the consideration of the hierarchical relationship among interacting covariates. Chipman (1996) considered two major variants of the heredity principles. Strong heredity principle, otherwise known as the principle of marginality (Nelder, 1977), states that a two-factor interaction should be activated together with both its main effects. Alternatively, under weak heredity, only one of the main features needs to be active. In the case of the backward approach, the interaction selection does not require the special treatment, since all main features are already included in the active set. Adding interactions into forward selection, and especially the LARS approach, trig-

gered the development of several extensions. One of them, for example, was the generalized LARS algorithm proposed by Yuan et al. (2009). In this algorithm, the heredity principle was preserved by adding a dependent set of main features associated with each interaction. Consequently, this dependent set was used for estimation of the new measure, average predictability, that generalized the original correlation-based method.

The heredity-complaint extensions of the selection methods do not resolve the scalability issue associated with the multi-dimensional models. This project addresses this problem by proposing a scalable approach that allows to integrate the different variants of backward and forward selection methods into a generic dynamic programming procedure.

4 Approach

The proposed approach is derived from two major techniques: divide and conquer and feature screening. Following its name, the divide and conquer approach divides the entire set of features into multiple smaller subsets and then conquers the subsets by screening the most significant interactions. The screening techniques were successfully employed in previous works, for example by Fan and Lv (2008), as a fast method for selecting significant main features in a high-dimensional model. Our approach extends this algorithm to incorporate the heredity principle and uses this method as a preliminary step for reducing the number of significant interactions in a subset. Afterwards, the model is built incrementally by combining subsets from the bottom up until we reach one set composed of the most significant interactions. In this procedure, each subset is consistent and complete containing the relevant main and interaction features, maintaining the principle of marginality at every stage of the approach.

Our approach is recursive and can be described by an initial and intermediate step. In the initial step, the set of main features is divided into smaller subsets, A_S, that consist of groups of related main features, x_i.

$$A_S = \{x_i, i \in S\} \tag{2}$$

For each subset, S, of A_S, the algorithm includes a composite filtering process based on three techniques for selecting the significant within-set interactions of the group:

$$B_S^{(0)} = \{x_i x_j; i, j \in S, p < 0.001\} \tag{3}$$

$$C_S^{(0)} = \{B_S^{(0)}, n_S > N_S\} \tag{4}$$

$$J_S^{(0)} = \{C_S^{(0)}, p < 0.01\} \tag{5}$$

The filtering procedure begins with pairwise selection, $B_S^{(0)}$, an approach based on linear regression that filters out the least significant interactions, $x_i x_j$, using the t-test's p-value, p. Next, the cluster-based selection, $C_S^{(0)}$, focuses on eliminating the influential rare events by bounding the minimum number of observations, N_S, associated with a particular interaction. The N_S differs for each subset and is determined by the cross-validation procedure. In our research it varied from 50 to 1000. Lastly, joint selection, $J_S^{(0)}$, applies one of the backward or forward selection algorithms and picks significant interactions based on their criteria. In order to choose the most reliable $x_i x_j$, the joint selection is built from the intersection of interactions that appear in each cross-validation fold. Throughout all these steps, all the original main features are maintained and therefore the strong heredity principle is always satisfied. At the completion of the initial step, each subset group, $G_S^{(0)}$, represents the combination of main features and the selected within-set interactions:

$$G_S^{(0)} = \{A_S, J_S^{(0)}\} \tag{6}$$

With the initial subsets complete, the approach starts the incremental procedure of building the

model by combining the subsets of the lower levels from the previous steps. For example, to create the new group at the k^{th} step, the procedure combines two subsets, $G_P^{(k-1)}$ and $G_R^{(k-1)}$, and then follows the three-step filtration process:

$$B_{P,R}^{(k)} = \{x_i x_j; i \in P, j \in R, p < 0.001\} \tag{7}$$

$$C_{P,R}^{(k)} = \{B_{P,R}^{(k)}, n_{P,R} > N_{P,R}\} \tag{8}$$

$$J_{P,R}^{(k)} = \{J_P^{(k-1)}, J_R^{(k-1)}, C_{P,R}^{(k)}, p < 0.01\} \tag{9}$$

Both of the first two parts, $B_{P,R}^{(k)}$ and $C_{P,R}^{(k)}$, are similar to the initial step. The joint selection, $J_{P,R}^{(k)}$, however, differs in that it runs a regression on a subset of variables that include $C_{P,R}^{(k)}$ as well as the interactions from the previous groups, $J_P^{(k-1)}$ and $J_R^{(k-1)}$. Therefore, in our approach the interactions are not fixed and can be filtered out in subsequent steps. As a result, the intermediate group is built from the main features of both previous groups and the jointly selected subset of within and between-set interactions.

$$G_{P,R}^{k} = \{A_P, A_R, J_{P,R}^{(k)}\} \tag{10}$$

In multi-dimensional models, the higher intermediate groups and final model can be achieved via different paths. For example, the group G_{abc} can be built from three different pairs, $\{\mathrm{G}_a, \mathrm{G}_{bc}\}$, $\{\mathrm{G}_b, \mathrm{G}_{ac}\}$ and $\{\mathrm{G}_a, \mathrm{G}_{bc}\}$. Because of the random element of the cross-validation procedure, these paths can lead to different sets of interactions. Thus, the collection of multiple final and intermediate models naturally form an ensemble of solvers that can be processed with the standard blending techniques.

5 Data Model

The Heritage Health Prize Competition contains four years of patient medical history and the corresponding days in hospital (DIH) values for the three years. This heterogeneous data is divided into four groups:

- Member: a categorical summary of the patient consisting of MemberID, age and sex.
- Claims: the most significant medical history on the patient for a given year. There are many sub-categorical variables such as place of service, procedure group, and vendor.
- RX: a history of the number of prescription drugs filled by days since first service.
- Lab: a history of unique laboratory tests by days since first service.

All but the first of these groups contain categorical variables with more than two values. Since these cannot be directly used in a regression model, the standard procedure of representing these values is through a set of dummy variables. The age, for example, is divided into 10 groups: 0-10, 11-20, etc. A binary variable is then used to mark the appropriate range of the patient. The predictors of the claim data, however, required additional treatment since patients could have multiple claims (rows) throughout the year. Merging the corresponding entries into a single row transformed these binary dummy features into count variables for each member as shown in Table 1.

Column	Number of Features	Column	Number of Features
Age At First Claim	9	Places	8
Sex	1	Length Of Stay	11
Claims Truncated	1	Days Since First Seen	13
Days In Hospital	16	Primary Condition Group	45
Providers	14700	Charlson Index	6
Vendors	6388	Procedures	17
Primary Care Physician	1360	Suppressed Length Of Stay	2
Primary Care Physician (last claim)	1360	Drug Count	7
Specialties	12	Lab Count	10

Table 1: List of features of the HHP model (Mestrom, 2011).

6 Results

To test our proposed methodology, we focused on the most significant subset of features as observed by the competitors of the Heritage Health Prize. For the initial step, the features were divided into 4 groups associated with homogenous features such as the primary condition group.

	Description	LM with Main Features		LM with Interactions		GBM
		Features	5-CV	Interactions	5-CV	5-CV
G_1	Age groups, gender, etc.	11	0.4663	9	0.4657	0.4658
G_2	Number of claims, vendors, etc.	8	0.4684	4	0.4683	0.4680
G_3	Primary Condition Group	45	0.4682	17	0.4663	0.4659
G_4	Specialties, Places, Procedures	37	0.4674	18	0.4667	0.4654

Table 2: **Initial Step**: Comparison of linear models (LM) with and without interaction features to the gradient boosting machine (GBM). The table presents experiments performed on separate subsets of features. The columns show the number of features remaining after filtering and the root mean square error (RMSE) after performing 5-fold cross validation.

Table 2 contains the results of the initial step produced by the three different approaches. We evaluate the models using root mean square error (RMSE), the same measure employed for the HHP competition. The table also includes the number of features employed by the linear models. The results show that for each subset, adding interactions based on our approach improves the performance of the linear model (LM). The number of interactions for each group depends on the parameter N_S from (4) which is determined through the 5-fold cross-validation procedure. It is important to note here that in the HHP competition, the difference between the most competitive approaches is usually around 0.001. This makes the improvements we observe, quite significant.

Following our approach, the initial groups are further combined into intermediate groups by adding between-group interactions. One such intermediate step is summarized in Table 3.

	LM with Main Features			LM with Interactions			GBM	
	Features	5-CV	Leader Board	Interactions	5-CV	Leader Board	5-CV	Leader Board
$G_{1,2}$	19	0.4630	0.4702	14	0.4616	0.4702	0.4617	0.4702
$G_{1,3}$	56	0.4627	0.4697	36	0.4605	0.4680	0.4606	0.4667
$G_{2,3}$	53	0.4665	0.4713	11	0.4641	0.4690	0.4639	0.4681
$G_{1,4}$	48	0.4620	0.4692	31	0.4604	0.4686	0.4601	0.4671
$G_{2,4}$	45	0.4664	0.4716	8	0.4658	0.4709	0.4649	0.4691
$G_{3,4}$	82	0.4653	0.4703	19	0.4633	0.4674	0.4608	0.4648

Table 3: **Intermediate Step**: Comparison of the linear models (LM) with and without interaction features to the gradient boosting machine (GBM). The table presents experiments performed on a collection of merged subsets of features. The columns show the number of features remaining after filtering and the root mean square error (RMSE) after performing 5-fold cross validation.

According to the cross-validation results, the approach consistently improves the LM and performs comparably to GBM. However, on the leaderboard GBM outperformed the LM with interactions which indicates some overfitting in the latter algorithm. This issue is associated with the high sensitivity of interactions to data irregularity and can be solved by integrating the standard bagging techniques (Breiman, 1996) into joint selections steps (5) and (9). Currently, these steps are based on the backward selection approach. In order to incorporate bagging the backward approach can be replaced with forward-stagewise regression (Hastie et al., 2009), the linear-regression version of the GBM algorithm.

	LM with Main Features			LM with Interactions			GBM	
	Features	5-CV	Leader Board	Interactions	5-CV	Leader Board	5-CV	Leader Board
G_{final}	101	0.4597	0.4673	119	0.4559	0.4653	0.4557	0.4634

Table 4: **Final Step**: Comparison of the linear models (LM) with and without feature interactions to the gradient boosting machine (GBM). All three models are trained on all features. The columns show the number of features remaining after filtering and the root mean square error (RMSE) after performing 5-fold cross validation. The table also presents the performance of each model as evaluated by the HHP leader board evaluation.

The final results are summarized in Table 4. They confirm that interaction analysis based on our approach improves the performance of the linear model. On the other hand, the linear model with selected interactions still does not compete with GBM. This discrepancy can be explained by the differences of the two models. The GBM approach divides the dataset into clusters and calculates the average response of the within-cluster observations. On the other hand, the linear model deals with low-dimensional correlation analysis using the whole dataset. Furthermore, the linear model does not consider the asymmetry of the interacted features. For example, a patient with a record of 5 acute myocardinal infarction's (AMI) and 2 emergency visits is not necessarily going to spend the same amount of days in the hospital as a different person with a record of 2 AMIs and 5 emergency visits. This issue, however, can be addressed in our approach by dividing the range of independent features into intervals and adding interactions between dummy variables associated with these intervals.

7 Summary

This paper presents an efficient and scalable approach for the interaction analysis of the linear multi-dimensional models. It is developed to address the scalability issue of the conventional backward and forward selection methods by accommodating them into a consistent procedure based on the combination of the divide and conquer algorithm and multi-step screening techniques. The approach is considered in the context of the Heritage Health Prize competition. The numerical results clearly demonstrate a marked prediction improvement in comparison with the ordinary linear model. Moreover, the project identifies several directions for future extensions.

We consider that the gap between the linear model and GBM can be further narrowed, if not surpassed entirely, through two complimentary approaches. First, as revealed from our experiments, the current variant is sensitive to irregularity which in turn leads to overfitting. To resolve this issue, standard bagging techniques can be integrated into the joint selection steps of our algorithm. This is similar to the way GBM iteratively adds new trees. The second modification to the proposed approach would address the difference between the linear and GBM models. The linear model works with the low-dimensional correlation analysis treating the dataset as a whole. The GBM, however, assumes that not all observations are homogeneous, and instead clusters the dataset into subgroups and independently computes the response of each of these clusters. This difference in approaches can be resolved through partitioning the range of each independent variable into intervals and introducing the associated dummy binary variables. Therefore, an interaction between multiple dummy variables can be interpreted as a generation of a cluster.

8 Acknowledgements

I would like to thank my supervisor, Professor William H. Greene, whose help, insightful discussions, courses, and guidance were instrumental in the success of this project. Furthermore, I want to acknowledge Professor Marti G. Subrahmanyam, whose organization and guest lectures made the Stern Honors program such a rewarding experience. Finally, my gratitude goes to Dr. Yuri Malitsky for helping create the data mining framework I used to explore and test my hypotheses.

References

D. Barton, D. Court, Making Advanced Analytics Work For You, Harvard Business Review (2012) 79–83.

A. Töscher, M. Jahrer, R. Bell, the BigChaos Solution to the Netflix Grand Prize, 2009.

J. H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, Annals of Statistics 29 (5) (2001) 1189–1232.

W. H. Greene, Econometric Analysis, Harlow: Pearson Education, 2012.

T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), New York: Springer-Verlag, 2009.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least Angle Regression, Annals of Statistics 32 (2) (2004) 407–499.

R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, Journal of the Royal Statistical Society B 58 (1) (1994) 267–288.

H. Chipman, Bayesian Variable Selection with Related Predictors, Canadian Journal of Statistics 24 (1) (1996) 17–36.

J. A. Nelder, A Reformulation of Linear Models, Journal of the Royal Statistical Society 140 (1) (1977) 48–77.

M. Yuan, V. R. Joseph, H. Zou, Structured Variable Selection and Estimation, The Annals of Applied Statistics 3 (4) (2009) 1738–1757.

J. Fan, J. Lv, Sure Independence Screening for Ultrahigh Dimensional Feature Space, Journal of the Royal Statistical Society 70 (5) (2008) 849–911.

W. Mestrom, my Milestone 1 Solution to the Heritage Health Prize, 2011.

L. Breiman, Bagging Predictor, Machine Learning 24 (2) (1996) 123–140.