湖南大學 工商管理學院
BUSINESS SCHOOL OF HUNAN UNIVERSITY

Université de Lille

Business School of Hunan University

Ordinal Classification with Non-parametric Frontier Methods: Overview and New Proposals

A thesis presented for the degree of Doctor in Economics

Qianying Jin

5 October 2020

Membres du Jury

Rapporteurs :
Prof. Dr. Steve Wenbin Liu, Kent Business School
Prof. Dr. Ignace Van de Woestyne, KU Leuven

Examinateurs :
Prof. Dr. Stéphane Vigeant, Université de Lille (Président du jury)
Associate Prof. Dr. Bing Xu, Edinburgh Business School

Directeurs :
Prof. Dr. Kristiaan Kerstens, CNRS-LEM, IESEG School of Management
Prof. Dr. Zhongbao Zhou, Business School of Hunan University

Acknowledgements

First and foremost, I want to thank my supervisors, Prof. Dr. Zhongbao Zhou and Prof. Dr. Kristiaan H.J. Kerstens. It was Prof. Zhou who introduced me to this wonderful academic world. During the six years from my master to Ph.D., Prof. Zhou has been a supervisor as well as a friend to me. He has given me every freedom to explore different research directions. At the same time, his wealth of knowledge and experience greatly helped me to identify the worthwhile research directions. I am very grateful to Prof. Zhou for his constant and unwavering support and patience! Prof. Kerstens is a supervisor with a keen interest and enthusiasm for doing research. The joy and enthusiasm he has for his research is contagious and motivational to me. I am more than fortunate to have had the opportunity to work with him and learn from him. Prof. Kerstens helped me to come up with my thesis topic and guided me through its development over the last two years. In addition to the guidance around my research topic, Prof. Kerstens has continued to help me enrich my curriculum vitae and build my research career. I am thankful for the excellent example he has provided as a successful economist and an academic researcher. Together, Prof. Zhou and Prof. Kerstens are great supervisors. I appreciate all their contributions of time, ideas, and funding to make my Ph.D. experience productive and stimulating.

My sincerest thanks goes to Prof. Dr. Wenbin Liu, Prof. Dr. Ignace Van De Woestyne, Prof. Dr. Stéphane Vigeant and Dr. Bing Xu for their time and effort invested in reading my work as the jury members of my defense.

I would like to thank the China Scholarship Council (CSC) for funding me to start a new journey in France. Meanwhile, my sincere thanks go to IESEG School of Management. IESEG has provided me with excellent working conditions to realize my Ph.D thesis. Moreover, IESEG accepted me as a Teaching and Research Assistant so that I am financially supported to perform my research projects. I thank as well my university of inscription, l'Université de Lille and l'Ecole Doctorale SESAM.

My sincere thanks also go to my co-authors. Especially, I would like to thank Prof. Wenbin Liu for our joint work on evaluating cardinality constrained portfolios. The inspiring discussions with Prof. Liu helped me a lot in scientific writing and developing critical thinking. I also want to thank Prof. Ignace Van De Woestyne for our joint work on questioning the common convexification strategy in calculating the metafrontier productivity indices.

I have the good fortune to have had great experiences in two different schools and to have being supported by people from both sides. At the Business School of Hunan University, I appreciated the group of people who share the same interests: Helu Xiao, Ling Lin, Yong Jiang, Cenjie Liu, Yanfei Bai, Tiantian Ren, Ximei Zeng, Qing Liu, Shisong Xiao and Minxi Huang. They have contributed immensely to my personal and professional experience. This group has been a source of friendship as well as good advices and collaborations. Meanwhile, I deeply love the dynamic academic environment and the intellectual community at IESEG School of Management. I appreciated the enormous help from every colleague I met at IESEG, especially my Ph.D. colleagues (some are already doctors) and the colleagues in Building C. I especially want to thank Adrian Barragan Diaz, Helen Cocoo, Jenny Tran, Minh Phan, Karina Shitikova and Emilie Buisine for all the beautiful time we spent together, and for the multicultural experience and

perspectives we shared with each other.

Finally, my most special thanks are reserved for my family. I thank my parents, Zugao Jin and Yunlian Liu, for their unconditional love, patience and dedication. Even if sometimes they do not fully understand and accept my choices, they are always there to support me. Further, I owed my older brother, Zicheng Jin, an enormous amount of thankfulness. We grew up together and he always supported me in all my decisions. I am grateful to him for being my brother, for listening to my problems and giving me constructive advices.

Qianying Jin

Lille, 26 June 2020

General Abstract

Following the idea of separating two groups with a hypersurface, the convex (C) frontier generated from the data envelopment analysis (DEA) method is employed as a separating hypersurface in classification. No assumption on the shape of the separating hypersurface is required while using a DEA frontier. Moreover, its reasoning of the membership is quite clear by referring to a benchmark observation. Despite these strengths, the DEA frontier-based classifier does not always perform well in classification. Therefore, this thesis focuses on modifying the existing frontier-based classifiers and proposing novel frontier-based classifiers for the ordinal classification problem.

In the classification literature, all axioms used to construct the C DEA frontier are kept in generating a separating frontier, without arguing their correspondence with the related background information. This motivates our work in Chapter 2 where the connections between the axioms and the background information are explored. First, by reflecting on the monotonic relation, both input-type and output-type characteristic variables are incorporated. Moreover, the minimize sum of deviations model is proposed to detect the underlying monotonic relation if this relation is not priori given. Second, a nonconvex (NC) frontier classifier is constructed by relaxing the commonly used convexity assumption. Third, the directional distance function (DDF) measure is introduced for providing further managerial implications, although it does not change the classification results comparing to the radial measure. The empirical results show that the NC frontier classifier has the highest classification accuracy. A comparison with six classic classifiers also

reveals the superiority of applying the NC frontier classifier.

While the relation of the characteristic variables often suggests consideration of a monotonic relation, its parallel problem of considering a non-monotonic relation is rarely considered. In Chapter 3, a generalized disposal assumption which limits the disposability within a value range is developed for characterizing the non-monotonic relation. Instead of having a single separating frontier, a NC separating hull which consists of several frontiers is constructed to separate the groups. By adding the convexity assumption, a C separating hull is then constructed. An illustrative example is used to test the performance. The NC hull classifier outperforms the C hull classifier. Moreover, a comparison with some existing frontier classifiers also reveals the superiority of applying the proposed NC hull classifier.

Chapter 4 proposes novel frontier classifiers for accommodating different mixes of classification information. To be specific, by reflecting on the monotonic relation, a NC classifier is constructed. If there is a priori information of the substitution relation, then a C classifier is generated. Both the NC and C classifiers generate two frontiers where each envelops one group of observations. The intersection of two frontiers is known as the overlap which may lead to misclassifications. The overlap is reduced by allowing the two frontiers to shift inwards to the extent that the total misclassification cost is minimized. The shifted cost-sensitive frontiers are then used to separate the groups. The discriminant rules are also designed to incorporate the cost information. The empirical results show that the NC classifier provides a better separation than the C one does. Moreover, the proposed DDF measure outperforms the commonly used radial measure in providing a reasonable separation.

Keywords: Nonparametric Frontier; Nonconvex; Non-monotonicity; Cost-Sensitive; Ordinal Classification

Résumé Général

Classification ordinale avec méthodes aux frontières non paramétriques : Aperçu et nouvelles propositions

Suivant l'idée de séparer deux groupes par une hypersurface, la frontière convexe (C) générée par la méthode d'analyse de l'enveloppe des données (DEA) est utilisée pour la séparation dans la classification. Aucune hypothèse sur la forme de l'hypersurface n'est nécessaire si l'on utilise une frontière DEA. De plus, son raisonnement sur l'appartenance est très clair en se référant à une observation de référence. Malgré ces points forts, le classificateur basé sur la frontière DEA n'est pas toujours performant dans la classification. Par conséquent, cette thèse vise à modifier les classificateurs frontaliers existants et à proposer de nouveaux classificateurs frontaliers pour le problème de la classification ordinale.

Dans la littérature, tous les axiomes utilisés pour construire la frontière C de la DEA sont conservés pour générer une frontière de séparation, sans argumenter leur correspondance avec les informations de base correspondantes. C'est ce qui motive notre travail au chapitre 2, où les liens entre les axiomes et les informations de base sont examinés. Tout d'abord, en réfléchissant à la relation monotone, les variables caractéristiques du type d'entrée et du type de sortie sont incorporées. En outre, le modèle de la somme minimale des écarts est proposé pour détecter la relation monotone sous-jacente si cette relation n'est pas donnée

a priori. Deuxièmement, un classificateur de frontière nonconvexe (NC) est construit en assouplissant l'hypothèse de convexité. Troisièmement, la mesure de la fonction de distance directionnelle (DDF) est introduite pour fournir des implications managériales, bien qu'elle ne modifie pas les résultats de la classification par rapport à la mesure radiale. Les résultats empiriques montrent que le classificateur à frontière NC a la plus grande précision de classification. Une comparaison avec six classificateurs classiques révèle également la supériorité de l'application du classificateur à frontière NC.

Alors que la relation des variables caractéristiques suggère souvent la prise en compte d'une relation monotone, le problème parallèle de la prise en compte d'une relation nonmonotone est rarement pris en compte. Au chapitre 3, une hypothèse d'élimination généralisée qui limite l'élimination dans une fourchette de valeurs est développée pour caractériser la relation non monotone. Au lieu d'avoir une seule frontière de séparation, une coque de séparation NC qui se compose de plusieurs frontières est construite. En ajoutant l'hypothèse de convexité, une coque séparatrice C est alors construite. Un exemple illustratif montrent que le classificateur de coques NC surpasse le classificateur C. En outre, une comparaison avec certains classificateurs frontaliers existants révèle également la supériorité de l'application du classificateur de coque NC.

Le chapitre 4 propose de nouveaux classificateurs frontaliers permettant de prendre en compte différentes combinaisons d'informations de classification. En réfléchissant à la relation monotone, un classificateur NC est construit. Si la relation de substitution existe, alors un classificateur C est généré. Les classificateurs NC et C génèrent tous deux deux des frontières où chacun enveloppe un groupe. L'intersection de deux frontières est connue sous le nom de chevauchement, ce qui peut entraîner des classifications erronées. Le chevauchement est réduit en permet-

tant aux deux frontières de se déplacer vers l'intérieur dans la mesure où le coût total de la classification erronée est minimisé. Les frontières déplacées sensibles aux coûts sont alors utilisées pour séparer les groupes. Les règles discriminantes sont également conçues pour intégrer les informations sur les coûts. Les résultats empiriques montrent que le classificateur NC assure une meilleure séparation que le classificateur C. En outre, la mesure de la DDF proposée surpasse la mesure radiale couramment utilisée en fournissant une séparation raisonnable.

Mots clés : Frontière Non-Paramétrique ; Non-Convexe ; Non-Monotonie ; Sensible au Coût ; Classification Ordinale

Contents

List of Figures

List of Tables

CHAPTER 1 General Introduction

Classification is fundamentally a data mining task whereby discriminant rules are developed for assigning an observation to prior-known groups. In order to train a classifier, the data is collected concerning observations with known group memberships. This training data is then used to develop discriminant rules for future classification of the observation whose group membership is unknown. Classic examples of classification applications include medical diagnosis (allocating patients to certain disease classes based on symptoms and lab tests), and credit scoring (accepting or rejecting some credit applications based on their application data).

Mathematical programming (MP) methods for classification emerged in the 1960s, gained popularity in the 1980s, and have been developing dramatically ever since. Most of the MP-based classifiers are nonparametric, which has been cited as an advantage over methods that require assumptions about the distribution of the data (Stam (1997)). One of the earliest linear programming (LP) classifiers was proposed by Mangasarian (1965) where a hyperplane is constructed to separate two groups of data. Studies of LP classifiers in the early 1980s were carried out by Freed and Glover (1981), Stam and Ragsdale (1992) and Bajgier and Hill (1982). Along with the development of the LP classifiers, nonlinear programming classifiers are natural extensions for some of these LP models (Mangasarian (1996),

Stam and Joachimsthaler (1989), Mangasarian, Setiono, and Wolberg (1990) and etc.). Various programming goals developed for deciding the best separation include minimizing the sum of deviations (MSD), minimizing the maximum deviation (MMD), minimizing the sum of interior distances (MSID), the hybrid models and their variants (Joachimsthaler and Stam (1990)). Having the programming goal related to minimizing the number of misclassifications, the mixed-integer programming (MIP) classifiers stand out. Intuitively, all MP-based classifiers have a geometric interpretation where a hypersurface is constructed and expected to provide an optimal separation between the groups. However, the functional form which generates the separating hypersurface is explicitly assumed in the MP-based classifiers. It is not impossible, but very difficult to prescribe a functional form to fit for real applications. In this sense, a nonparametric classifier that provides a data-based piecewise linear frontier receives increasing attention since no assumption on the frontier shape is required.

The data-based nonparametric method refers to the Data Envelopment Analysis (DEA) method proposed by Charnes, Cooper, and Rhodes (1978) which is originally developed for ranking a set of observations. The current application of the DEA methods in classification could be categorized into three types.

In the first type, the efficiencies calculated from the DEA methods are used to separate two groups of observations. Specifically, a cut-off efficiency calculated form the DEA methods is taken as the threshold value that differentiates between two groups in MP-based classifiers, see Sueyoshi and Kirihara (1998), Emel, Oral, Reisman, and Yolalan (2003), Pendharkar and Rodger (2003), Cheng, Chiang, and Tang (2007), Min and Lee (2008), Premachandra, Bhabra, and Sueyoshi (2009), Premachandra, Chen, and Watson (2011), Lu, Lee, and Zou (2012), Paradi and Yang (2014), Malhotra and Tsetsekos (2016), etc. Subsequently, the

layer technique instead of the cut-off efficiency is studied for classification, see Paradi, Asmild, and Simak (2004), Avkiran and Cai (2014). Note that in this type of DEA-based classifiers, groups of observations are assumed to be homogeneous and thus to be evaluated with the same DEA model. To a certain extent, this assumption on homogeneity ignores the essential differences between two groups.

The second type is known as the Data Envelopment Analysis-Discriminant Analysis (DEA-DA) method firstly proposed by Sueyoshi (1999) and subsequently developed by Sueyoshi (2001, 2004). This DEA-DA method is extended to allow for various data types (Jahanshahloo, Lotfi, Balf, and Rezai (2007)); (Lotfi and Mansouri (2008); Boudaghi and Saen (2018); and etc), and a multi-group setting (Sueyoshi (2006)). Empirically, the applications cover the industry (Sueyoshi and Goto (2009a); Sueyoshi and Goto (2009b); Sueyoshi and Goto (2012)), marketing (Hwang, Lin, and Chuang (2007); Farzipoor Saen (2013); Tavassoli, Faramarzi, and Farzipoor Saen (2014)), finance (Sueyoshi and Hwang (2004)) etc. The DEA-DA method is claimed to incorporate a methodological strength of DEA into the DA formulation. However, it is essentially based on the use of goal programming. The geometric illustration for the DEA-DA methods is the same as that of the LP-based classifiers. Rather than using an assumption-free frontier, a linear hyperplane is generated from the DEA-DA methods to separate two groups of observations.

The third type refers to the DEA frontier-based classifiers which are constructed from the standard DEA methods proposed by Banker, Charnes, and Cooper (1984). Troutt, Rai, and Zhang (1996) first propose to use the convex DEA frontier as an acceptability frontier in credit applicant acceptance systems. Different groups of observations are therefore located on the opposite sides of this convex frontier. Without pre-specifying the exact shape of a separating hypersurface, this

convex frontier is piece-wise linear and bounds one group of observations closely. The data-based nonparametric method emphasized in this study refers explicitly to this type of frontier-based nonparametric methods.

Ever since the first application of the DEA frontier in classification proposed by Troutt, Rai, and Zhang (1996), the idea of employing the convex DEA frontier as a separating frontier has been well adapted by proposing alternative objective functions (Seiford and Zhu (1998)), incorporating various data types (e.g., Leon and Palacios (2009), Yan and Wei (2011)) and has been applied in different application areas (e.g., Seiford and Zhu (1998), Pendharkar (2002), Pendharkar, Rodger, and Yaverbaum (1999); Pendharkar, Khosrowpour, and Rodger (2000))).

One common thing in these single frontier classification literature is that the separating frontier is assumed to be convex. To the best of our knowledge, none of the current research has ever left out the convexity assumption. The assumption of convexity is commonly kept in production analysis since it is common in the economic theory. When it comes to the classification problem, the assumption on convexity is accepted without arguing its correspondence with the related background information in classification. This motivates our work in **Chapter 2**. The main objective of Chapter 2 is to relax this convexity assumption and construct a nonconvex separating frontier. This nonconvex frontier is based on the Free Disposal Hull (FDH) approach that has been initially proposed by Deprins, Simar, and Tulkens (1984). It has a monotonous or staircase shape and envelops the data tighter than the convex separating frontier does. Another objective is to develop the frontier-based classifier into a more general form. To be specific, both characteristic variables with monotonically increasing relation and those with monotonically decreasing relation are included. Moreover, a directional distance function measure is introduced so that more managerial information could

be provided.

In the literature, following along the idea of using a single separating frontier, Chang and Kuo (2008) propose to use a pair of DEA frontiers so that these two frontiers each describes a set of observations. The intersection of two frontiers is known as the data-based overlap. McLachlan (1992, p. 16) remarks that classification accuracy depends mostly on how well the discriminant rule can handle observations in the overlap. Therefore, the majority of the subsequent work focus on either eliminating the overlap in the training process (e.g., Kuo (2013)) or further classifying the observations that are located in the overlap (e.g., (Pendharkar (2012); Pendharkar and Troutt (2014); Pendharkar (2011); Pendharkar (2018)).

Although the idea of using the DEA frontier is extended from using a single frontier to using double frontiers, the same axioms on constructing the separating frontier are retained, e.g., the convexity assumption. Therefore, **Chapter 4** firstly intends to build a nonparametric frontier-based classifier which is capable of generating either convex or nonconvex separating frontiers. Apart from relaxing some original assumptions of constructing the frontiers, a novel treatment of overlap is proposed so that the total misclassification cost is minimized for the training process. Specifically, this is achieved by allowing the two frontiers to shift inwards to the extent of achieving a minimum misclassification cost. Furthermore, the discriminant rules are also designed to incorporate the cost information.

In all the existing frontier-based classification literature, the background information on the relation between the characteristic variables and the group label often suggests consideration of a monotonic relation. However, in many applications, the relation between the group label and the characteristic variables could also show non-monotonicity. One example that illustrates the non-monotonic relation is the medical diagnoses, where both high values and low values may indicate

symptoms of certain diseases. While the research on incorporating the monotonic relation is increasing, its parallel problem of considering the non-monotonic relation is rarely taken into account in classification. To the best of our knowledge, the research on dealing with the non-monotonicity in classification is quite limited and not stable (Lam and Choo (1993)). Therefore, the primary interest of **Chapter 3** is to incorporate the non-monotonic variables while developing a nonparametric frontier-based classifier for the classification problem. To characterize the non-monotonic variables, a generalized disposal assumption is developed following the S-disposal assumption proposed by Briec, Kerstens, and Van de Woestyne (2016, 2018). This generalized disposal assumption limits the disposability of a variable within a value range. Correspondingly, a separating hull which consists of several frontiers is constructed to separate the groups of observations.

To sum up, the overarching objective of this thesis is to propose some novel nonparametric frontier-based classifiers for achieving a better classification performance and for more general classification problems. To be specific, the first aim is to study the connections between the commonly used axioms in the DEA methods and the background information in classification. Although the initial inspiration of applying the DEA-based convex frontier is that it provides a tight envelopment without any assumption on the shape of the frontier, it is now essential to provide some theoretical basis so that the commonly used assumptions can be relaxed depending on the specific classification problems. A second design objective is to incorporate the cost information in constructing the nonparametric frontier-based classifiers, not only in generating the separating frontiers but also in designing the discriminant rules. That is, the proposed nonparametric frontier-based classifiers are expected to be cost-sensitive. Finally, the non-monotonic relation which is widely existed but not extensively studied is incorporated while constructing the nonparametric frontier-based classifiers.

References

Avkiran, N., and L. Cai (2014): "Identifying Distress Among Banks Prior to a Major Crisis Using Non-oriented Super-SBM," *Annals of Operations Research*, 217(1), 31–53.

Bajgier, S. M., and A. V. Hill (1982): "An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminant Problem," *Decision Sciences*, 13(4), 604–618.

Banker, R. D., A. Charnes, and W. W. Cooper (1984): "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science*, 30(9), 1078–1092.

Boudaghi, E., and R. F. Saen (2018): "Developing a Novel Model of Data Envelopment Analysis–Discriminant Analysis for Predicting Group Membership of Suppliers in Sustainable Supply Chain," *Computers & Operations Research*, 89, 348–359.

Briec, W., K. Kerstens, and I. Van de Woestyne (2016): "Congestion in Production Correspondences," *Journal of Economics*, 119(1), 65–90.

——— (2018): "Hypercongestion in Production Correspondences: An Empirical Exploration," *Applied Economics*, 50(27), 2938–2956.

Chang, D., and Y. Kuo (2008): "An Approach for the Two-group Discriminant Analysis: An Application of DEA," *Mathematical and Computer Modelling*, 47(9-10), 970–981.

Charnes, A., W. Cooper, and E. Rhodes (1978): "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2(6), 429–444.

Cheng, E. W., Y. H. Chiang, and B. S. Tang (2007): "Alternative Approach to Credit Scoring by DEA: Evaluating Borrowers with Respect to PFI Projects," *Building and Environment*, 42(4), 1752–1760.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring Labor Efficiency in Post Offices," in *The Performance of Public Enterprises: Concepts and Measurements*, ed. by M. Marchand, P. Pestieau, and H. Tulkens, pp. 243–268. North Holland, Amsterdam.

Emel, A. B., M. Oral, A. Reisman, and R. Yolalan (2003): "A Credit Scoring Approach for the Commercial Banking Sector," *Socio-Economic Planning Sciences*, 37(2), 103–123.

Farzipoor Saen, R. (2013): "Using Cluster Analysis and DEA-Discriminant Analysis to Predict Group Membership of New Customers," *International Journal of Business Excellence*, 6(3), 348–360.

Freed, N., and F. Glover (1981): "Simple but Powerful Goal Programming Models for Discriminant Problems," *European Journal of Operational Research*, 7(1), 44–60.

Hwang, S. N., C. T. Lin, and W. C. Chuang (2007): "Stock Selection Using Data Envelopment Analysis-Discriminant Analysis," *Journal of Information and Optimization Sciences*, 28(1), 33–50.

Jahanshahloo, G. R., F. H. Lotfi, F. R. Balf, and H. Z. Rezai (2007): "Discriminant Analysis of Interval Data Using Monte Carlo Method in Assessment of Overlap," *Applied Mathematics and Computation*, 191(2), 521–532.

Joachimsthaler, E. A., and A. Stam (1990): "Mathematical Programming Approaches for the Classification Problem in Two-Group Discriminant Analysis," *Multivariate Behavioral Research*, 25(4), 427–454.

Kuo, Y.-C. (2013): "Consideration of Uneven Misclassification Cost and Group Size for Bankruptcy Prediction," *American Journal of Industrial and Business Management*, 3(08), 708.

Lam, K. F., and E. U. Choo (1993): "A Linear Goal Programming Model for Classification with Non-Monotone Attributes," *Computers & Operations Research*, 20(4), 403–408.

Leon, C. F., and F. Palacios (2009): "Evaluation of Rejected Cases in an Acceptance System with Data Envelopment Analysis and Goal Programming," *Journal of the Operational Research Society*, 60(10), 1411–1420.

Lotfi, F. H., and B. Mansouri (2008): "The Extended Data Envelopment Analysis/Discriminant Analysis Approach of Fuzzy Models," *Applied Mathematical Sciences*, 2(30), 1465–1477.

Lu, S. L., K. J. Lee, and M. L. Zou (2012): "How to Gauge Credit Risk: An Investigation Based on Data Envelopment Analysis and the Markov Chain Model," *Applied Financial Economics*, 22(11), 887–897.

Malhotra, R., and G. Tsetsekos (2016): "Evaluating Loans Using Variable Benchmark Data Envelopment Analysis," *International Journal of Business Intelligence and Systems Engineering*, 1(1), 77–98.

Mangasarian, O. L. (1965): "Linear and Nonlinear Separation of Patterns by Linear Programming," *Operations Research*, 13(3), 444–452.

Mangasarian, O. L. (1996): "Machine Learning via Polyhedral Concave Minimization," in *Applied Mathematics and Parallel Computing*, pp. 175–188. Springer.

Mangasarian, O. L., R. Setiono, and W. H. Wolberg (1990): "Pattern Recognition via Linear Programming: Theory and Application to Medical Diagnosis," in *Proceedings of the Workshop on Large-Scale Numerical Optimization*, pp. 22–31. SIAM.

McLachlan, G. J. (1992): "Discriminant Analysis and Statistical Pattern Recognition," *Wiley Series in Probability and Mathematical Statistics, New York: Wiley, 1992*.

Min, J. H., and Y. C. Lee (2008): "A Practical Approach to Credit Scoring," *Expert Systems with Applications*, 35(4), 1762–1770.

Paradi, J., M. Asmild, and P. Simak (2004): "Using DEA and Worst Practice DEA in Credit Risk Evaluation," *Journal of Productivity Analysis*, 21(2), 153–165.

Paradi, J., and X. Yang (2014): "Data Envelopment Analysis of Corporate Failure for Non-manufacturing Firms Using a Slacks-based Measure," *Journal of Service Science and Management*, 7(04), 277.

Pendharkar, P. (2012): "Fuzzy Classification Using the Data Envelopment Analysis," *Knowledge-Based Systems*, 31, 183–192.

Pendharkar, P. (2018): "Data Envelopment Analysis Models for Probabilistic Classification," *Computers & Industrial Engineering*, 119, 181–192.

Pendharkar, P., M. Khosrowpour, and J. Rodger (2000): "Application of Bayesian Network Classifiers and Data Envelopment Analysis for Mining Breast Cancer Patterns," *Journal of Computer Information Systems*, 40(4), 127–132.

Pendharkar, P., and J. Rodger (2003): "Technical Efficiency-based Selection of Learning Cases to Improve Forecasting Accuracy of Neural Networks Under Monotonicity Assumption," *Decision Support Systems*, 36(1), 117–136.

Pendharkar, P., J. Rodger, and G. Yaverbaum (1999): "Association, Statistical, Mathematical and Neural Approaches for Mining Breast Cancer Patterns," *Expert Systems with Applications*, 17(3), 223–232.

Pendharkar, P. C. (2002): "A Potential Use of Data Envelopment Analysis for the Inverse Classification Problem," *Omega*, 30(3), 243–248.

Pendharkar, P. C. (2011): "A Hybrid Radial Basis Function and Data Envelopment Analysis Neural Network for Classification," *Computers & Operations Research*, 38(1), 256–266.

Pendharkar, P. C., and M. D. Troutt (2014): "Interactive Classification Using Data Envelopment Analysis," *Annals of Operations Research*, 214(1), 125–141.

Premachandra, I., G. Bhabra, and T. Sueyoshi (2009): "DEA as a Tool for Bankruptcy Assessment: A Comparative Study With Logistic Regression Technique," *European Journal of Operational Research*, 193(2), 412–424.

Premachandra, I., Y. Chen, and J. Watson (2011): "DEA as a Tool for Predicting Corporate Failure and Success: A Case of Bankruptcy Assessment," *Omega*, 39(6), 620–626.

Seiford, L., and J. Zhu (1998): "An Acceptance System Decision Rule with Data Envelopment Analysis," *Computers & Operations Research*, 25(4), 329–332.

Stam, A. (1997): "Nontraditional Approaches to Statistical Classification: Some Perspectives on L_p-Norm Methods," *Annals of Operations Research*, 74, 1.

Stam, A., and E. A. Joachimsthaler (1989): "Solving the Classification Problem in Discriminant Analysis via Linear and Nonlinear Programming Methods," *Decision Sciences*, 20(2), 285–293.

Stam, A., and C. Ragsdale (1992): "On the Classification Gap in Mathematical Programming-Based Approaches to the Discriminant Problem," *Naval Research Logistics*, 39(4), 545–559.

Sueyoshi, T. (1999): "DEA-Discriminant Analysis in the View of Goal Programming," *European Journal of Operational Research*, 115(3), 564–582.

——— (2001): "Extended DEA-Discriminant Analysis," *European Journal of Operational Research*, 131(2), 324–351.

——— (2004): "Mixed Integer Programming Approach of Extended DEA–Discriminant Analysis," *European Journal of Operational Research*, 152(1), 45–55.

——— (2006): "DEA-Discriminant Analysis: Methodological Comparison Among Eight Discriminant Analysis Approaches," *European Journal of Operational Research*, 169(1), 247–272.

Sueyoshi, T., and M. Goto (2009a): "Can R&D Expenditure Avoid Corporate Bankruptcy? Comparison Between Japanese Machinery and Electric Equipment Industries Using DEA–Discriminant Analysis," *European Journal of Operational Research*, 196(1), 289–311.

——— (2009b): "DEA–DA for Bankruptcy-Based Performance Assessment:

Misclassification Analysis of Japanese Construction Industry," *European Journal of Operational Research*, 199(2), 576–594.

Sueyoshi, T., and M. Goto (2012): "Efficiency-based Rank Assessment for Electric Power Industry: a Combined Use of Data Envelopment Analysis (DEA) and DEA-Discriminant Analysis (DA)," *Energy Economics*, 34(3), 634–644.

Sueyoshi, T., and S.-N. Hwang (2004): "A Use of Nonparametric Tests for DEA-Discriminant Analysis: A Methodological Comparison," *Asia-Pacific Journal of Operational Research*, 21(02), 179–195.

Sueyoshi, T., and Y. Kirihara (1998): "Efficiency Measurement and Strategic Classification of Japanese Banking Institutions," *International Journal of Systems Science*, 29(11), 1249–1263.

Tavassoli, M., G. Faramarzi, and R. Farzipoor Saen (2014): "Multi-criteria ABC Inventory Classification Using DEA-Discriminant Analysis to Predict Group Membership of New Items," *International journal of applied management science*, 6(2), 171–189.

Troutt, M., A. Rai, and A. Zhang (1996): "The Potential Use of DEA for Credit Applicant Acceptance Systems," *Computers & Operations Research*, 23(4), 405–408.

Yan, H., and Q. Wei (2011): "Data Envelopment Analysis Classification Machine," *Information Sciences*, 181(22), 5029–5041.

CHAPTER 2

Ordinal Classification with a Single Nonparametric Frontier: The Role of Nonconvexity

2.1 Introduction

The nonparametric frontier method -also known as Data Envelopment Analysis (moniker DEA)- is commonly attributed to Charnes, Cooper, and Rhodes (1978). This method provides a relative efficiency measure for each Decision Making Unit (DMU) by comparing its relative performance to all observed DMUs. In addition to generating a relative efficiency measure that accordingly ranks the DMUs, this linear programming method floats a piecewise linear frontier that envelops all observed DMUs. This frontier provides an extremal relation between a vector of inputs and a vector of generated outputs.

This extremal relation implied by the nonparametric frontier has originally been applied to estimate multiple inputs and multiple outputs production corres-

pondences yielding efficiency measures and productivity indices of various kinds. In production economics, the nonparametric frontier represents the extremal combinations of the outputs that can be produced from the combinations of available inputs. Obviously, this extremal relation between inputs and outputs implied by a nonparametric frontier formally complies with some standard axioms of production required to obtain a valid production model. Thus, the nonparametric frontier offers a reasonable approximation for the theoretical production frontier in applied production analysis. Its envelopment is based on a minimum extrapolation principle in that all empirical observations are used along with extensions of these observations. The extensions are based on some simple axioms about what is considered feasible. Hence, it provides a conservative estimate of theoretical production frontiers.

The idea of empirically estimating production frontiers via these nonparametric methods has been widely applied across economic sectors. In a rather recent survey of the first 40 years of scholarly literature in DEA from the year 1978 till 2016, Emrouznejad and Yang (2018) list about 10300 research articles.

More recently, the utilization of nonparametric frontiers has also crossed disciplinary boundaries to estimate similar extremal relations in finance. For instance, a wide variety of frontier models have been proposed to obtain relative efficiencies of mutual funds (see, Murthi, Choi, and Desai (1997) for a seminal article and Basso and Funari (2016) for a recent survey): some models are directly transposed from production theory, other models include traditional diversification effects related to the modern portfolio theory. In modern portfolio theory, the efficient portfolio frontier is conceived as a set of optimal portfolios that simultaneously yield the highest return for the lowest possible risk, eventually subject to some additional constraints. This extremal relation between return and risk measures

can be interpreted by considering the return measure as an output that needs to be maximised and the risk measure as an input that needs to be minimised. Sengupta (1989) is the first to introduce an efficiency measure into this mean-variance quadratic (hence convex) optimisation problem. The article of Briec, Kerstens, and Jokung (2007) is among the first to extend this mean-variance optimisation problem towards the inclusion of skewness: this requires cubic (hence nonconvex) optimisation problems. With the help of an appropriate efficiency measure, these authors manage to simultaneously maximise return and skewness while minimising the variance. An extension to the higher moment portfolio problem assuming investors comply with mixed risk-aversion behaviour (i.e., a positive preference for odd moments and a negative preference for even moments) is developed in Briec and Kerstens (2010).

When it comes to the ordinal classification problem, a classifier is trained via learning from a number of labeled observations. This classifier is then expected to correctly predict the group membership of a new observation to its maximum level. The ordinal classification problem is important and quite common in practical applications. Examples of this problem includes but are not limited to costumer churn (e.g., De Caigny, Coussement, De Bock, and Lessmann (2019)), bankruptcy prediction (e.g., De Bock (2017)), credit scoring (e.g., Lessmann, Baesens, Seow, and Thomas (2015)), etc. For the mathematical programming (MP) approaches in classification, intuitively the idea is to generate a separating hyperplane (or more generally a separating hypersurface) to separate different groups of observations. For every group, this separating hyperplane or hypersurface is conceived as an envelopment of its observations.

Naturally, the nonparametric frontier method could be a candidate in estimating a separating hyperplane or hypersurface. To the best of our knowledge, the

idea of using a nonparametric frontier has been first introduced in Troutt, Rai, and Zhang (1996). These authors employ a convex nonparametric frontier as an acceptability frontier in credit applicant acceptance systems. The acceptability frontier, known as a separating frontier in this contribution, is generated by enveloping all the accepted applicants. For a new applicant, if it is located within this acceptability frontier, then it is accepted, otherwise rejected. Seiford and Zhu (1998) modify the method in Troutt, Rai, and Zhang (1996) by proposing an alternative objective function. In Leon and Palacios (2009), the convex frontier method is applied to the cases with non-discretionary characteristic variables. Variations in the importance of characteristic variables have been handled by using a preference cone in Yan and Wei (2011). Pendharkar and his coauthors conduct a series of empirical and experimental studies in classification problems with nonparametric frontiers (e.g., bankruptcy prediction (Pendharkar (2002)), mining breast cancer patterns (Pendharkar, Rodger, and Yaverbaum (1999); Pendharkar, Khosrowpour, and Rodger (2000)), etc.).

Despite our ignorance as to the real shape of a separating hypersurface, all existing nonparametric frontier methods stick to the convexity assumption. Convexity assumes that for any two points from one set, the linear combinations of these two points belong to the same set. If the separating hypersurface derived from any discriminant function happens to be convex, then the estimated convex frontier offers a reasonable estimate. However, the estimated convex frontier sometimes appears to be overtly optimistic. For instance, when analysing the superior performance of neural networks over convex frontiers in mining breast cancer patterns, Pendharkar, Rodger, and Yaverbaum (1999, p. 231) claim that one of the reasons is that the frontier method assumes the convexity of acceptable cases, while neural networks relax this assumption. Therefore, the main objective of this contribution is to construct a nonconvex separating frontier to envelop a

group of observations. This nonconvex frontier is based on the Free Disposal Hull (FDH) approach that has been initially proposed by Deprins, Simar, and Tulkens (1984). It has a monotonous or staircase shape and envelops the data tighter than the convex separating frontier does.

Another assumption used for constructing a nonparametric frontier is the free disposability. In the frontier classification literature, the observation is treated as a DMU which has all the characteristic variables as inputs and has a single output of value 1. In this sense, it is always assumed that the classification seeks to select an observation with smaller characteristic values. While in applications, some characteristic variables are favored with bigger values. By connecting the assumption of free disposability with the monotonic relation in classification, this contribution aims at extending the current frontier-based classifiers to a more general form. Specifically, the characteristic variables are categorized into input-type and output-type variables depending on their monotonic relation with the group membership. Different from the logical inputs and outputs that are commonly used in production analysis, mathematical inputs and outputs are defined to represent the characteristic variables with monotonically decreasing and monotonically increasing relations, respectively. Furthermore, in the situation where the monotonic relation is not explicitly given, the Minimize Sum of Deviations (MSD) model proposed by Freed and Glover (1986) is applied to reflect the monotonic relation.

To meet the above two objectives, this contribution is structured as follows. Section 2.2 introduces the basic intuitions of applying a frontier method by using geometrical illustrations. Then, the models and procedures used in constructing nonparametric separating frontiers are presented in Section 2.3. In Section 2.4, two empirical applications are used to show the eventual improvements of our nonpara-

metric separating frontier methods relative to the use of six traditional classifiers found in the literature. Finally, in Section 2.5 this contribution is concluded with a summary of its achievements and a discussion of potential future research topics.

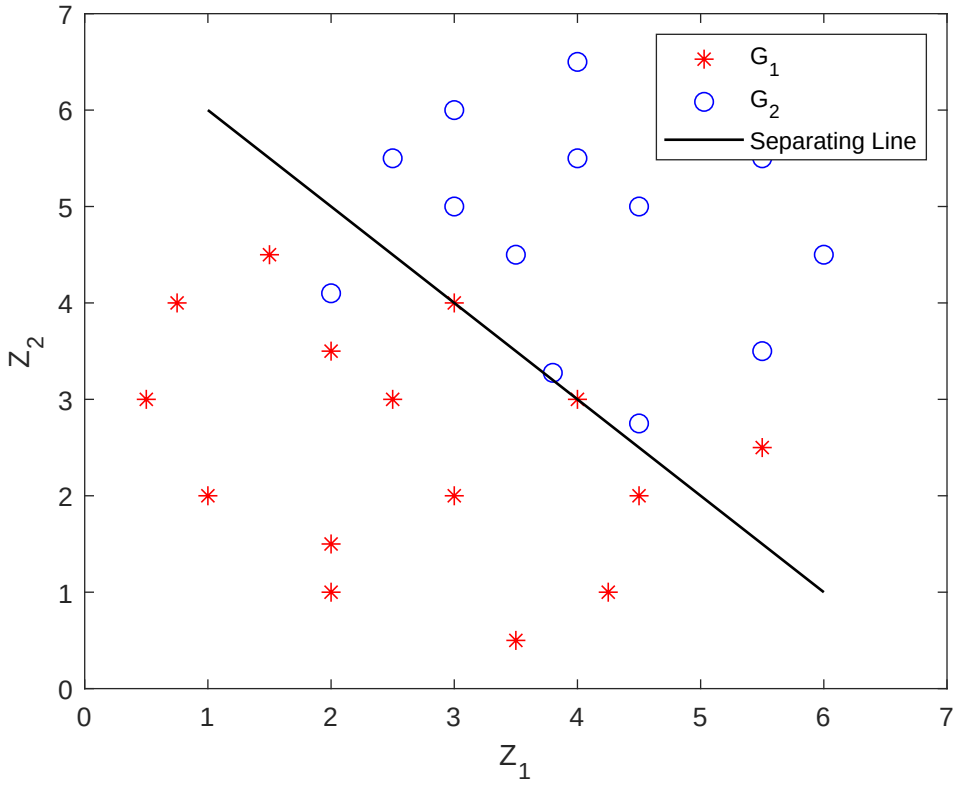
2.2 Basic Intuitions and Geometrical Illustration

In this contribution, an ordinal classification problem with two groups is investigated. Prior information on the labeled observations exists such that they belong to either the bad group (G_1) or the good group (G_2). In real-word scenarios, such as the bankruptcy prediction, the natural ordering is quite clear. That is, bankruptcy companies are labelled to be bad while non-bankruptcy ones are good.

We start with an illustrative example for explaining the general idea of the MP classification methods. A two-dimensional example with two characteristic variables Z_1 and Z_2 is visualized in Figure 2.1 to meet this end. The asterisks in Figure 2.1 represent the training observations from G_1, while the circles represent those from G_2. With the MP classification methods, a separating hyperplane is expected to separate two groups of observations in a best way. Depending on the assumptions made with regard to the discriminant functions, the derived separating hyperplane can be a simple separating line (e.g., see Freed and Glover (1986), Lam, Choo, and Moy (1996), Sueyoshi (2004), among others) or rather a separating curve (e.g., see Silva and Stam (1994), Smaoui, Chabchoub, and Aouni (2009), etc.). In Figure 2.1, the solid line stands for one possible separating line. It is observed that most training observations from G_1 are situated below the separating line, while those from G_2 are mainly situated above the separating line. By learning from the position of the training observations relative to the separating line, the widely used discriminant rule is determined as follows. For an

observation, if it is situated below the separating line, then it is believed to belong to the bad group G_1. Otherwise, it is perceived to belong to the good group G_2.

Figure 2.1: Illustration of a separating line in ordinal classification



Similarly, applying a nonparametric frontier method in classification also aims at finding a piecewise linear separating hypersurface which discriminates the training observations in some best way. Two groups of training observations are supposed to be situated either below or above the piecewise linear separating hypersurface.

Comparing to the traditional MP classification methods mentioned above, there are two modelling advantages by applying a nonparametric frontier method. First, no specific assumption on the shape of the separating hypersurface is required.

The separating hypersurface is a nonparametric frontier generated from available observations and some combination of axioms about what is considered feasible. It is a piece-wise linear frontier which envelops the training observations tightly. Hence, it is expected to provide a better classification.

Second, the nonparametric frontier method is capable of considering the monotonic relationship of the characteristic variables, which is a type of background information. Specifically, if the possibility of belonging to the good group increases with the increase of a characteristic value (while the others are held constant), then the monotonicity of this characteristic variable is increasing. On the contrary, any characteristic variable whose decrease (while the others are held constant) leads to the increase of the possibility of belonging to the good group has a monotonically decreasing relation. For instance in the student admission, the admitting level is monotonically increasing with respect to the student's academic performance. As for the numerical example in Figure 2.1, it is clear that the higher the two characteristic variables are, the more likely an observation is located in the good group. Hence, both characteristic variables here have a monotonically increasing relation.

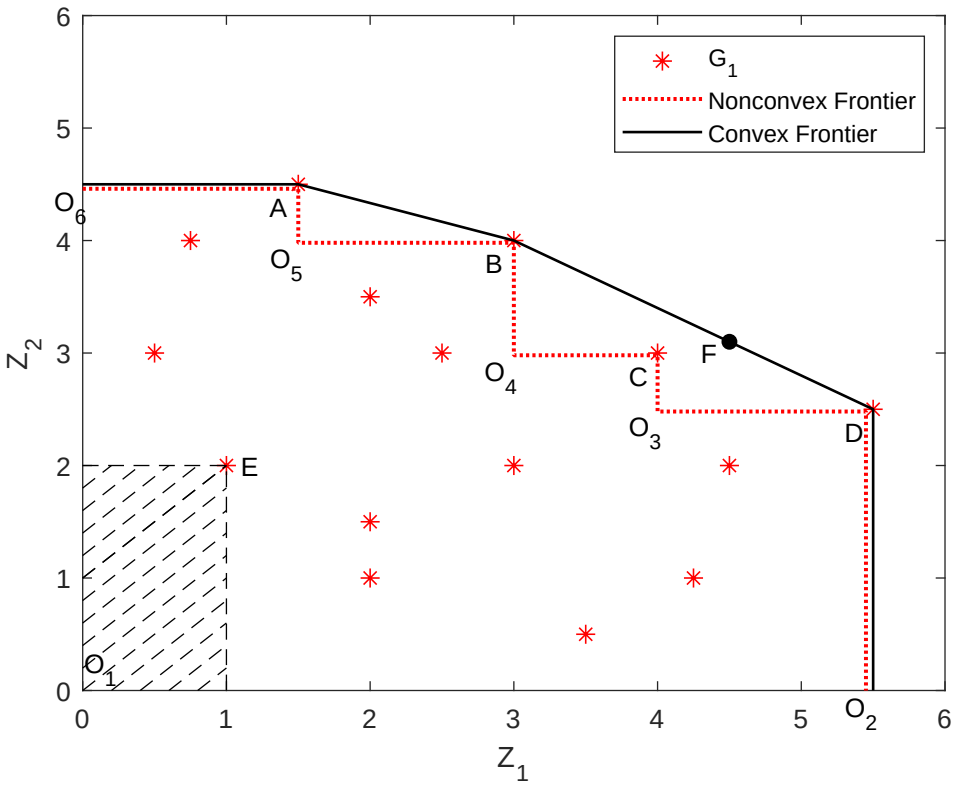
In the following, the general procedure of training a nonparametric frontier classifier is illustrated. First we focus on constructing an attainable set to characterize the base training group G_1. With the same two-dimensional example in Figure 2.1, both the nonconvex (NC) and convex (C) attainable sets of G_1 are discussed. The boundary of the attainable set is known as a separating frontier which envelops one group of observations.

The NC attainable set is derived from the observations in G_1 and the axiom on free disposability. This NC attainable set describes all possible combinations of characteristic values that corresponding observations are believed to belong to

the bad group. In Figure 2.2, the asterisks represent the observations from G_1. First, if an observation has the same characteristic values as one observation from G_1, then it certainly belongs to the attainable set of G_1. Then, with the observed monotonicity, a free disposal area is formed and implies that all observations in this area share the same group membership. Take the training observation E as an example, the shaded area restricted to the third quadrant located below and to the left of E represents the free disposal area of E. All observations located in this area belong to the bad group just as the training observation E does. This is due to the monotonically increasing relation between characteristic variables and group membership in this example. It illustrates that the possibility of belonging to the bad group will increase with the decreasing of two characteristic variables. Put is differently, an observation remains in the bad group as long as its characteristic values are no larger than that of an observation from G_1. Finally, the union of all these free disposal areas derived from the training observations in G_1 constitutes the NC attainable set of G_1. In Figure 2.2, this is the area restricted to the third quadrant located below and to the left of polyline $O_2DO_3CO_4BO_5AO_6$ marked with dotted lines.

In the literature, it is common to have a C attainable set which generates a C frontier. Comparing to the above NC attainable set, the C set is derived by having one additional axiom on convexity. Mathematically, the axiom on convexity implies that for any two observations from one set, the linear combination of these two observations belong to the same set. In classification, this convexity axiom explains a substitution relation between two characteristic variables. When it gets to our numerical example in Figure 2.2, since the observations B and D are in the attainable set of G_1, their linear combination F is also supposed to be in the same attainable set. Note that there is no observation from G_1 that directly dominates F. Comparing to the observation B, the decreasing of Z_2 increases the

Figure 2.2: Nonconvex and convex attainable set of the bad group G_1

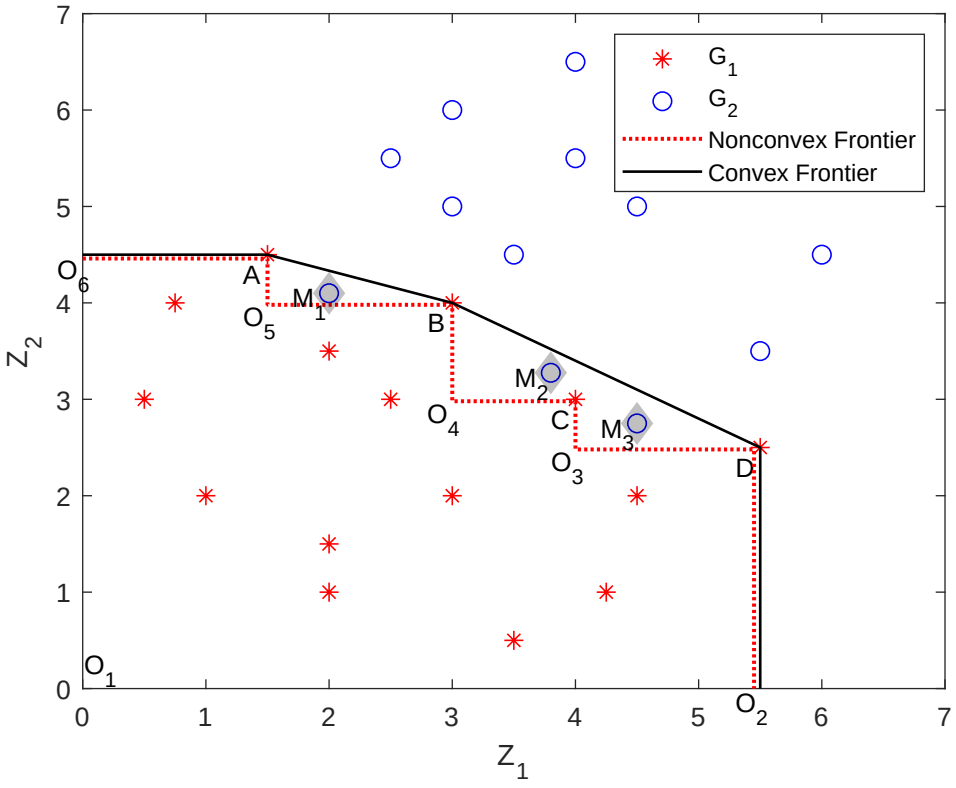


possibility of assigning the observation F to the bad group while the increasing along Z_1 decreases the possibility. However, the substitution relation implies that the decreasing of Z_2 perfectly offsets the increasing of Z_1. Hence, the observation F is supposed to have the same group membership as B does. In applications, there may exist this type of substitution relation. Like in the student admission, the disadvantage in GMAT scores can be compensated by the advantages in SAT to a certain extent. Based on the observations from G_1, a C attainable set of G_1 is obtained with the axioms on free disposability and convexity. In Figure 2.2, this is the area restricted to the third quadrant located below and to the left of polyline O_2DBAO_6 marked with solid lines. Comparing to the NC attainable set, two extra areas, namely, AO_5BA and BO_4CO_3DB, are added due to the convexity. The C

attainable set also describes all possible combinations of characteristic values that corresponding observations are believed to belong to the bad group.

The boundary of the attainable set is known as the frontier. As shown in Figure 2.2, the NC frontier marked by dotted lines has a staircase shape. It consists of four observations, namely, A, B, C and D. The C frontier derived from the C attainable set consists of three observations, namely A, B and D. The observation C is perceived to be dominated by a series of virtual points on the line segment BD, like F. These virtual points are derived from a convex combination of the observations B and D. Clearly, the NC frontier provides a tighter envelopment of the observations than the C one does.

Figure 2.3: Nonconvex and convex separating frontiers



This derived frontier is then used as a separating frontier to separate two groups of observations. For a new observation, if it is located within the separating frontier, then it belongs to the bad group G_1, otherwise it belongs to the good group G_2. As shown in Figure 2.3, the NC separating frontier gives a perfect separation between two groups of observations. That is, the observations from G_1 are situated within the NC separating frontier while those from G_2 are situated beyond the NC frontier. For the C case, although all observations from G_1 are situated within the C separating frontier, not all observations from G_2 are correctly situated beyond the C frontier. Three observations that are actually from G_2 are situated within the C separating frontier. These three misclassified observations M_1, M_2 and M_3 are marked with the faded rhombus in Figure 2.3. These extra misclassifications imply that if there is no clear information on the substitution relation between the characteristic variables, a NC separating frontier is preferred in terms of its conservation comparing to the C frontier.

2.3 Nonparametric Frontier Approaches for Discriminant Analysis

2.3.1 Basic Concepts

Consider an ordinal classification problem with a set of observations. These observations constitute the total sample G to be used for training a classifier. Prior knowledge on the group membership is given such that the observations are exhaustively classified into two groups, namely, the bad group G_1 and the good group G_2. Note that the group G_1 together with the group G_2 form a partition of G, that is $G_1 \cup G_2 = G$ and $G_1 \cap G_2 = \emptyset$.

The observations Z_j $(j = 1, \ldots, n)$ are characterized by K characteristic variables. That is, $Z_j = (z_{1,j}, \ldots, z_{k,j}, \ldots, z_{K,j})$. According to the background information on the relation between the group membership and the characteristic variables, there are two types of monotonic relations.

Definition 2.3.1. *The characteristic variable $z_{k,j}$ has a monotonically decreasing relation, if the following holds true: $Z_{j_1} = (z_{1,j_1}, \ldots, z_{k,j_1}, \ldots, z_{K,j_1}) \in \mathbb{R}^K$ belongs to the good group if there exists an observation $Z_{j_2} = (z_{1,j_2}, \ldots, z_{k,j_2}, \ldots, z_{K,j_2}) \in \mathbb{R}^K$ from the good group for which $z_{k,j_2} > z_{k,j_1}$ and $z_{l,j_2} = z_{l,j_1}$ for all $l \in \{1, \ldots, K\} \setminus \{k\}$.*

Definition 2.3.2. *The characteristic variable $z_{k,j}$ has a monotonically increasing relation, if the following holds true: $Z_{j_1} = (z_{1,j_1}, \ldots, z_{k,j_1}, \ldots, z_{K,j_1}) \in \mathbb{R}^K$ belongs to the good group if there exists an observation $Z_{j_2} = (z_{1,j_2}, \ldots, z_{k,j_2}, \ldots, z_{K,j_2}) \in \mathbb{R}^K$ from the good group for which $z_{k,j_2} < z_{k,j_1}$ and $z_{l,j_2} = z_{l,j_1}$ for all $l \in \{1, \ldots, K\} \setminus \{k\}$.*

Accordingly, the characteristic variables are exclusively categorized into two types. If the possibility of belonging to the good group increases (decreases) with the decrease (increase) of a characteristic value, then it is a characteristic variable with a monotonically decreasing relation, denoted by $X \in \mathbb{R}^m$. Otherwise, if the possibility of belonging to the good group increases (decreases) with the increase (decrease) of a characteristic variable, then it is a characteristic variable with a monotonically increasing relation, denoted by $Y \in \mathbb{R}^s$. The observation Z is then characterized by the characteristic variables X and Y: $Z = (X, Y) \in \mathbb{R}^{m+s}$. Note that $m + s = K$.

In situations where the monotonic relation exists but is not explicitly given,

the MSD model is applied to differentiate the characteristic variables:

$$\begin{aligned} \min_{\alpha_k, s^+_{1,j}, s^-_{2,j}} \quad & \sum_{j \in G_1} s^+_{1,j} + \sum_{j \in G_2} s^-_{2,j} & & \\ s.t. \quad & \sum_{k=1}^{K} \alpha_k z_{k,j} + s^+_{1,j} \leq d - \eta & & j \in G_1 \\ & \sum_{k=1}^{K} \alpha_k z_{k,j} - s^-_{2,j} \geq d & & j \in G_2 \\ & s^+_{1,j} \geq 0, s^-_{2,j} \geq 0, d \text{ and } \alpha_k \text{ are free} & & \end{aligned} \tag{2.3.1}$$

where d is a threshold value and $\alpha_k (k = 1, \ldots, K)$ are weights, both of which are unknown and therefore to be determined. To avoid a trivial solution (where $\alpha_k = 0$ for all k and $d = 0$) and to have a clear separation between two groups, a small number η is introduced (Glover (1990)).

For the observations from the bad group G_1, the weighted average of their characteristic variables, which is $\sum_{k=1}^{K} \alpha_k z_{k,j}$, is supposed to be below the threshold value d. While the weighted average of the characteristic variables of the observations from the good group G_2 is generally above the threshold value. Misclassifications are allowed by having the slacks $s^+_{1,j}$ and surpluses $s^-_{2,j}$. If the optimal value of the weight α^*_k is negative, then the increase (decrease) of this characteristic variable reduces (enlarges) the value $\sum_{k=1}^{K} \alpha^*_k z_{k,j}$. Therefore, it decrease (increase) the possibility of belonging to the good group. On the contrary, if the optimal value of the weight α^*_k is positive, then the increase (decrease) of this characteristic variable increases (decreases) the possibility of belonging to the good group. To sum up, by solving model (2.3.1), the following monotonic relation is defined by the sign of α^*_k:

(i) If $\alpha^*_k < 0$, then characteristic variable k has a monotonically decreasing relation, denoted by x;

(ii) If $\alpha_k^* \geq 0$, then characteristic variable k has a monotonically increasing relation, denoted by y.

2.3.2 Acceptance Possibility Set

The bad group G_1 is used as the base training group to construct the separating frontier. That is, the attainable set is constructed based on the observations from G_1. It describes all possible combinations of characteristic values that corresponding observations belong to the bad group.

In production analysis, a production possibility set (PPS) is used to describe the attainable set in production. For all the combinations of the resources and the products within the PPS, they are attainable (or say producible) under a certain technology. Instead of discussing the attainability in producing, the attainable set in classification describes the attainability in accepting an observation to the bad group. Hence, an acceptance possibility set (APS) is used to describe the attainable set in classification.

First, all n_1 observations from G_1 are in the APS. Then, based on the monotonic relation between the characteristic variables and the group membership, a free disposal set denoted by T_j could be derived for every observation Z_j $j = (1, \ldots, n_1)$ from G_1. The monotonic relation says that comparing to an observation $Z_j = (X_j, Y_j)$ from G_1, an observation with more X and less Y is also supposed to belong to the bad group. That is, for $Z_j = (X_j, Y_j)$, $T_j = \{(X, Y) \in \mathbb{R}^{m+s} \mid X \geq X_j \text{ and } Y \leq Y_j\}$. The union of all the free disposal sets of the observations from G_1 constitutes a NC attainable set denoted by T_{NC}. Specifically, T_{NC} depicts the

observations belonging to the bad group as follows:

$$\begin{aligned} T_{NC} &= \bigcup_{j=1}^{n_1} T_j \\ &= \left\{ (X,Y) \in \mathbb{R}^{m+s} \mid \sum_{j=1}^{n_1} \lambda_j X_j \leq X, \sum_{j=1}^{n_1} \lambda_j Y_j \geq Y, \sum_{j=1}^{n_1} \lambda_j = 1, \lambda_j \in \{0,1\} \right\}. \end{aligned} \tag{2.3.2}$$

Based on the above NC APS, having an additional axiom on convexity leads to a C APS. As explained in Section 2.2, the convexity in classification implies a substitution relation among characteristic variables. If the prior information of such a substitution relation is provided, then a C APS is constructed as follows:

$$T_C = \left\{ (X,Y) \in \mathbb{R}^{m+s} \mid \sum_{j=1}^{n_1} \lambda_j X_j \leq X, \sum_{j=1}^{n_1} \lambda_j Y_j \geq Y, \sum_{j=1}^{n_1} \lambda_j = 1, \lambda_j \geq 0 \right\}. \tag{2.3.3}$$

For the NC case, an observation belongs to the bad group if and only if it is located within the free disposal area of one observation from G_1. While for the C case, except for the above situation, if it is located within the free disposal area of a convex combination of two or more observations originally from G_1, this observation is also believed to belong to the bad group. Obviously, $T_{NC} \subseteq T_C$: a NC monotonic hull is a subset of a C monotonic hull. Put differently, the NC APS provides a tighter envelopment of the training observations than the C one does.

In order to simplify the expressions, we use the following notation to stand for the APS of G_1 under the NC and C cases:

$$\begin{aligned} T_\Lambda = \Bigg\{ (X,Y) \in \mathbb{R}^{m+s} \mid & \sum_{j=1}^{n_1} \lambda_j X_j \leq X, \sum_{j=1}^{n_1} \lambda_j Y_j \geq Y, \\ & \sum_{j=1}^{n_1} \lambda_j = 1, \lambda_j \in \Lambda, j = 1, \ldots, n_1 \Bigg\}, \end{aligned} \tag{2.3.4}$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \quad \text{or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0,1\}\}.$$

In the classification literature, all the papers adopt either an input-oriented or an output-oriented radial measure. In this contribution, a directional distance function (DDF) measure is proposed to gauge the relative distance to the frontier. Following Chambers, Chung, and Färe (1998), T_Λ can be represented by the following DDF $D_{\Lambda,g}(Z)$:

$$D_{\Lambda,g}(Z) = \sup\{\delta \in \mathbb{R} \mid Z + \delta g \in T_\Lambda\}. \tag{2.3.5}$$

where $g = (g_X, g_Y) \in \mathbb{R}^m \times \mathbb{R}^s$ represents the projection direction. To be meaningful, $g_{x_i} \leq 0$ for all $i \in \{1, \ldots, m\}$ and $g_{y_r} \geq 0$ for all $r \in \{1, \ldots, s\}$. In this way, the characteristic variables X are non-increased and the characteristic variables Y are non-decreased while increasing the value of δ, which is the favorable behavior.

The assumption on convexity differentiate the NC APS ($T_{\Lambda^{\mathrm{NC}}}$) from the C one ($T_{\Lambda^{\mathrm{C}}}$). However, this does not change the definition of the DDF measure, only the value of the DDF measure may be enlarged. That is, $D_{\Lambda^{\mathrm{NC}},g}(Z) \leq D_{\Lambda^{\mathrm{C}},g}(Z)$. The value of $D_{\Lambda,g}(Z)$ serves as an indicator that positions the observations relative to the boundary of the APS (T_Λ). It is well-defined for all possible observations $Z = (X, Y) \in \mathbb{R}^m \times \mathbb{R}^s$. A non-negative $D_{\Lambda,g}(Z)$ means Z is in the interior of T_Λ. If an observation Z is located outside T_Λ, then $D_{\Lambda,g}(Z)$ becomes negative and this observation is projected onto the frontier in the direction opposite to g.

2.3.3 Separating Frontiers: Nonconvex and Convex

In this subsection, the separating frontiers are constructed for the nonconvex and the convex cases, respectively. With the projecting direction $g = (g_X, g_Y)$ where $g_X \leq 0$ and $g_Y \geq 0$, the following model is used to measure the relative distance of the observation (X_0, Y_0) to the boundary of T_Λ which depicts the bad group.

$$\begin{aligned}
\max_{\lambda_j, \overline{\delta}_\Lambda} \quad & \overline{\delta}_\Lambda \\
s.t. \quad & \sum_{j=1}^{n_1} \lambda_j x_{i,j} \leq x_{i,0} + \overline{\delta}_\Lambda g_{x_i} \qquad \forall i \in \{1, \ldots, m\} \\
& \sum_{j=1}^{n_1} \lambda_j y_{r,j} \geq y_{r,0} + \overline{\delta}_\Lambda g_{y_r} \qquad \forall r \in \{1, \ldots, s\} \\
& \sum_{j=1}^{n_1} \lambda_j = 1 \\
& \lambda_j \in \Lambda \qquad \forall j \in \{1, \ldots, n_1\}
\end{aligned} \tag{2.3.6}$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \text{ or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0, 1\}\}.$$

In the C case, model (2.3.6) is a linear programming (LP) problem, while it involves solving a binary mixed integer program (BMIP) for the NC case. To remedy the computational issue in the NC case, a fast implicit enumeration-based method is proposed by Cherchye, Kuosmanen, and Post (2001) requiring only to compute minima and maxima of lists of ratios. Instead of solving a BMIP model, the following exact solution is obtained for model (2.3.6) under the NC case:

$$\overline{\delta}^*_{\Lambda^{\mathrm{NC}}} = \max_{j=1,\ldots,n_1} \left(\min_{i=1,\ldots,m} \left(\frac{x_{i,j} - x_{i,0}}{g_{x_i}} \right), \min_{r=1,\ldots,s} \left(\frac{y_{r,j} - y_{r,0}}{g_{y_r}} \right) \right). \tag{2.3.7}$$

By solving model (2.3.6) for all observations from G_1, a frontier set defined by FS_Λ is generated. FS_Λ consists of the observations from G_1 that has $\overline{\delta}^*_\Lambda = 0$. Normally, the set FS_Λ under the NC case is different from that under the C case. All frontier observations in $FS_{\Lambda^{\mathrm{C}}}$ could be found in $FS_{\Lambda^{\mathrm{NC}}}$. However, not all frontier observations in $FS_{\Lambda^{\mathrm{NC}}}$ belong to $FS_{\Lambda^{\mathrm{C}}}$, since some frontier observations generated under the NC case are dominated by some convex combinations of the observations. Therefore, $FS_{\Lambda^{\mathrm{C}}} \subseteqq FS_{\Lambda^{\mathrm{NC}}}$.

2.3.4 Separating Frontier based Discriminant Rules

The separating frontier represented by the observations in the frontier set FS_Λ is then used to determine the membership of a new observation. Specifically, the following model is used to calculate the distance of the observation $Z_0 = (X_0, Y_0)$ relative to the separating frontier:

$$\begin{aligned}
\max_{\lambda_j, \delta_\Lambda} \quad & \delta_\Lambda \\
s.t. \quad & \sum_{j \in FS_\Lambda} \lambda_j x_{i,j} \leq x_{i,0} + \delta_\Lambda g_{x_i} && \forall i \in \{1, \ldots, m\} \\
& \sum_{j \in FS_\Lambda} \lambda_j y_{r,j} \geq y_{r,0} + \delta_\Lambda g_{y_r} && \forall r \in \{1, \ldots, s\} \\
& \sum_{j \in FS_\Lambda} \lambda_j = 1 \\
& \lambda_j \in \Lambda && \forall j \in FS_\Lambda
\end{aligned} \tag{2.3.8}$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \text{ or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0, 1\}\}.$$

For observation $Z_0 = (X_0, Y_0)$, the optimal distance measure calculated from

model (2.3.8) is the same as that calculated from model (2.3.6). The difference is that only the observations in the frontier set are used in the left hand side of the inequalities in model (2.3.8). Although $\delta^*_\Lambda = \overline{\delta}^*_\Lambda$ always holds, the decrease in sample size can save some computational time.

If $\delta^*_\Lambda \geq 0$, then it indicates there exists a benchmark that dominates the observation $Z_0 = (X_0, Y_0)$. This benchmark is generated from the left-hand side of the inequality constraints in model (2.3.8) and is represented by $Z_b = (\sum_{j\in FS_\Lambda} \lambda^*_j X_j, \sum_{j\in FS_\Lambda} \lambda^*_j Y_j)$. It is either an observation from the NC frontier set $FS_{\Lambda^{\text{NC}}}$ or a convex combination of the observations from $FS_{\Lambda^{\text{C}}}$. In the case where $\delta^*_\Lambda \geq 0$, the following two inequalities hold: $\sum_{j\in FS_\Lambda} \lambda^*_j X_j \leq X_0$ and $\sum_{j\in FS_\Lambda} \lambda^*_j Y_j \geq Y_0$. It is known that an observation is more likely to belong to the bad group with the increase of variables X and the decrease of variables Y . Comparing to the benchmark Z_b which is from the bad group, the observation $Z_0 = (X_0, Y_0)$ has more X and less Y. Obviously, it should be assigned to the bad group G_1.

On the contrary if $\delta^*_\Lambda < 0$, then the observation Z_0 dominates the benchmark Z_b. That is, $\sum_{j\in FS_\Lambda} \lambda^*_j X_j > X_0$ and $\sum_{j\in FS_\Lambda} \lambda^*_j Y_j < Y_0$. The benchmark is on the boundary of the APS of the bad group and is about to leave the bad group. Comparing to the benchmark Z_b, the observation Z_0 has less X and more Y. Therefore, its possibility of belonging to the bad group is lower than that of the benchmark. Hence, the observation Z_0 is preferred to be assigned to the good group if there is no further information.

To sum up, the membership of the observation Z_0 is determined by the sign of the optimal distance measure δ^*_Λ calculated from model (2.3.8). The discriminant rules are summarized as follows:

(Rule 1) If $\delta^*_\Lambda \geq 0$, then Z_0 belongs to the bad group;

(Rule 2) If $\delta_\Lambda^* < 0$, then Z_0 belongs to the good group;

Note that the membership is solely determined by the sign of the measure rather than the value. Put it differently, the choice of the direction vector does not make a difference in the classification results. In the literature, all the papers adopt either an input-oriented or an output-oriented radial measure. If the direction vector is $g = (-X_0, \vec{0})$ where $\vec{0}$ represents a zero vector and X_0 is assumed to be non-negative, then δ_Λ becomes an input-oriented radial measure. If the direction vector is $g = (\vec{0}, Y_0)$ where and Y_0 is assumed to be non-negative, then δ_Λ becomes an output-oriented radial measure. No matter which direction vector is chosen, the same classification results are obtained.

However, the choice of the direction vector makes a difference in obtaining an applicable benchmark. Take the university admission as an example, the applicant is classified into two categories, namely, admitted and not yet. The academic performance and the language scores are two main characteristic variables to be considered, among others. In the short term, the academic performance could not be easily enhanced while it is more likely to increase the language scores. In this sense, the applicant is interested in knowing a favorable language score to be admitted by universities while maintaining the current academic performance. This could be easily achieved by setting the direction value of the academic performance to be 0.

2.4 Empirical Analysis

2.4.1 Test Setting and Evaluation Criteria

In this section, the performances of the proposed C and NC nonparametric frontier classifiers are tested with two data sets. For the sake of replication, we choose two secondary data sets. First, the performance is measured with a Japanese bank data set which is balanced in the sample size, see Table 1 in Sueyoshi (2001) for the detailed data. Second, an unbalanced data set on the corporate bankruptcy in the US electric power industry is used, see Table 2 in Sueyoshi (2006) for the detailed data.

In addition, the performances of applying the proposed frontier classifiers are compared to that of applying six classic classifiers. Specifically, these classic classifiers are: Logit, Probit, Fisher's linear classifier, Smith's quadratic classifier, neural networks, and decision tree. The detailed description of these six classifiers is available in Sueyoshi (2001).

The hit rate results are reported to show the performance of all listed classifiers. A hit results when an observation emanating from a certain group is assigned to this group by means of the used classification rules. A hit rate is the proportion of the observations that are correctly classified under the used classification rule (see Huberty and Olejnik (2006) for definitions and variations). In this contribution, the apparent hit rate which measures the classification accuracy is used. It is the ratio of the correctly predicted observations to the total sample.

For the choice of the direction vector, we use $g = (-X_0, Y_0)$ for the observation $Z = (X_0, Y_0)$. This ensures that the DDF measure $(D_{\Lambda,g}(Z))$ obtains a

proportional interpretation (see Briec (1997)). Of course, such a percentage interpretation is not indispensable in our classification context where the focus is rather on the sign of the DDF measure. However, for convenience we stick to this proportional distance function. In our classification context with potentially negative inputs and outputs, Kerstens and Van de Woestyne (2011) argue that one can benefit from using a direction vector $g = (-|X_0|, |Y_0|)$ for a given observation $Z = (X_0, Y_0)$ so as to preserve a proportional interpretation.

2.4.2 Balanced Data Set

The balanced data set of 100 observations originates in Sueyoshi (2001) and is related to the Japanese banks. The group labels of these 100 Japanese banks are known a priori. The bottom 50 banks constitute the bad group G_1, while group G_2 contains the remaining top 50 banks. Since group G_1 represents the bad group of the poorly performing banks, it is chosen as the base training group to construct the separating frontier. The performance of banks is characterized by in total seven characteristic variables. Details on the definitions of these characteristic variables are provided in Sueyoshi (2001).

Table 2.1: Characteristic variables for the Japanese bank data set

	X	Y
	Cost-profit ratio	Return on total assets
Index	Bad loan ratio	Equity to total assets
Measures	Loss ratio of bad loans	Return on total domestic assets
		Return on equity

The characteristic variables are differentiated into two categories, as shown in Table 2.1. For the three index measures in the column of the characteristic variables X, the performance of a bank is better when these indexes are lower. For

example, a bank is believed to achieve better performance if it has less bad loans. By contrast, higher values of indexes in the column of the characteristic variables Y contribute to a better performance of the corresponding banks. For instance, the return on total assets shows the profitability of the assets in generating revenue: a higher value of this index implies a higher profitability and hence it indicates a better performance.

Table 2.2: Classification accuracies of various classifiers: Japanese bank data set

	Classifiers	Apparent Accuracy
Frontier-based Classifiers	Frontier - C	90,00
	Frontier - NC	100,00
Classic Classifiers	Logit	93,00
	Probit	93,00
	Fisher's linear classifier	91,00
	Smith's quadratic classifier	85,00
	Neural network	98,00
	Decision tree	93,00

Table 2.2 shows the classification performances of applying all of the above mentioned classifiers. The classification accuracies are reported in the last column. Horizontally, the first block reports the results of the C and NC frontier-based classifiers. The second block contains the results of six classic classifiers. These results are copied from the ones reported in Sueyoshi (2006).

The comparison of the two frontier-based classifiers in Table 2.2 finds that the classification performance is substantially improved by relaxing the convexity assumption. For both C and NC situations, all observations from G_1 are correctly classified. For the C frontier method, 10 observations originally from G_2 are misclassified into G_1. This leads to a classification accuracy of 90% in applying the C method. While the NC frontier method gives a classification accuracy of 100%.

This 100% classification accuracy indicates that two groups of banks are perfectly separated without any misclassification.

Comparing now the frontier-based classifiers with the classic classifiers yields one additional finding. The highest classification accuracy reaches 100% by applying the NC frontier methods. While for the six classic classifiers, the highest classification accuracy is 98% by applying the neural networks. There still leaves two banks misclassified. All the listed classic classifiers fail to provide a perfect separation between the two groups of banks. With respect to this bank data set, the NC separating frontier achieves a better classification performance than the classic classifiers.

In general, we find that the classification performance of the C frontier method can be substantially improved by relaxing the convexity assumption. Among all the listed methods in Table 2.2, the NC frontier classifier achieves the best separation between the two groups of banks.

2.4.3 Unbalanced Data Set

The second real data set used is an unbalanced data set related to the corporate bankruptcy data in the US electric power industry. The data is described in Sueyoshi (2006). In summary, it contains 22 default firms (G_1) and 61 non-default firms (G_2). The cost of misclassifying a default firm into the non-default group is relatively high, hence G_1 is chosen as the base training group to construct the separating frontier. The performance of all the firms is characterized by 13 financial ratios. Details on the definitions of these characteristic variables are provided in Sueyoshi (2006).

The characteristic variables are differentiated into two categories as shown in

Table 2.3: Characteristic variables for the US electric power industry data set

	X	Y
Ratios	Long-term debt to total assets	Cash to total assets
	Return on equity	Working capital to total assets
	Beta	Sales to total assets
		Shareholder equity to total assets
		Net income to total assets
		Retained earning to total assets
		Market to book ratio
		Price over earnings
		Earnings per share
		Share price

Table 2.3. For the three ratios in the column of the characteristic variables X, the probability of default is lower when these ratios are smaller. For example, more long-term debt compared to its total assets implies a higher possibility of being default. While for the ratios in the column of the characteristic variables Y, a higher value contributes to a lower possibility of getting default. For instance, the cash to total assets ratio is used to measure a firm's liquidity or its ability to pay its short-term obligations. The higher this ratio implies a smaller possibility of obtaining a default.

The same apparent hit rate as above is used to measure the classification accuracy. The accuracy results of various classifiers are listed in Table 2.4. The structure of Table 2.4 is similar to that of the Table 2.2.

From the results in Table 2.4, the main findings are very much in line with the above balanced data set. Although the classification accuracy of the C frontier methods is as high as 98.80%, it could be further improved by relaxing the assumption of convexity. Specifically, two groups of electric power firms are perfectly separated with the NC frontier methods.

Table 2.4: Classification accuracies of various classifiers: US electric power industry data set

	Classifiers	Apparent Accuracy
Frontier-based Classifiers	Frontier - C	98,80
	Frontier - NC	100,00
Classic Classifiers	Logit	98,80
	Probit	100,00
	Fisher's linear classifier	96,38
	Smith's quadratic classifier	98,80
	Neural network	100,00
	Decision tree	93,98

Both some of the classic classifiers and the NC frontier classifier achieve the best classification accuracy of 100%. It shows that this data set of the electric power industry is more separable than that of the Japanese banks, in spite of its imbalanced sample size.

To sum up, we find that an imbalance in the relative sizes of groups within the sample does not show significant influence when applying the frontier-based classifiers. The NC frontier methods still achieve a perfect separation between two groups of observations.

2.5 Conclusions

In most MP classification applications, the best functional form of the discriminant function is unknown. The nonparametric frontier methods envelops all observations in a flexible way since its precise shape is determined by the strength of the maintained axioms. In this sense, the piecewise linear envelopment frontier can serve as a separating hypersurface. All observations within the separating frontier

share the same group membership as those from the base training group, while the observations outside this frontier is classified into the opposite group.

This contribution has innovated in two main ways. First, instead of sticking to the convexity assumption, a NC frontier has been used and ends up with a better envelopment of the training observations. Second, a more generalized form of frontier-based classifier is introduced by incorporating characteristic variables with both the monotonically increasing and the monotonically decreasing relations. Moreover, the DDF measure is introduced to provide the alternative benchmark with more flexibility.

The empirical comparison between the NC frontier classifier with the C frontier classifier reveals that the NC frontier offers a tighter envelopment of observations than the C frontier does. Therefore, a perfect separation is obtained by applying a NC frontier for both the Japanese bank data set and the US electric company data set. If there is no prior information on the substitution relation among characteristic variables, then the NC frontier method is by far the best choice among the frontier-based methods. This study also compares the proposed frontier classifiers with six classic classifiers with respect to the same two real data sets. The empirical results show that the NC frontier method outperforms the six listed classifiers.

We end with developing some perspectives for potential future research. First, it is an open question to which extent the existing single frontier methods could be further enhanced for better discrimination by a further relaxation of some of the axioms inherited from production theory. Just as relaxing convexity yields a monotonous frontier instead of a convex piecewise linear frontier, one may wonder whether it is possible to weaken the currently maintained axiom of strong disposal. A recent theoretical attempt to do so is developed in Briec, Kerstens, and Van de

Woestyne (2016) and empirically implemented in Briec, Kerstens, and Van de Woestyne (2018). Second, one may equally wonder to which extent the same ideas can be transposed in the limited literature employing double separating frontiers in a classification setting (e.g, see Sueyoshi (2001), Sueyoshi (2006) and Chang and Kuo (2008)). Third, while we have in this contribution compared the frontier methods to a series of traditional classification methods, it could be interesting to compare the best of the frontier methods to some of the best performing state of the art classification methods (see Lessmann, Baesens, Seow, and Thomas (2015)) to check their relative classification and prediction accuracies.

References

Basso, A., and S. Funari (2016): "DEA Performance Assessment of Mutual Funds," in *Data Envelopment Analysis: A Handbook of Empirical Studies and Applications*, ed. by J. Zhu, pp. 229–287. Springer, Berlin.

Briec, W. (1997): "A Graph-Type Extension of Farrell Technical Efficiency Measure," *Journal of Productivity Analysis*, 8(1), 95–110.

Briec, W., and K. Kerstens (2010): "Portfolio Selection in Multidimensional General and Partial Moment Space," *Journal of Economic Dynamics and Control*, 34(4), 636—-656.

Briec, W., K. Kerstens, and O. Jokung (2007): "Mean-Variance-Skewness Portfolio Performance Gauging: A General Shortage Function and Dual Approach," *Management Science*, 53(1), 135–149.

Briec, W., K. Kerstens, and I. Van de Woestyne (2016): "Congestion in Production Correspondences," *Journal of Economics*, 119(1), 65–90.

——— (2018): "Hypercongestion in Production Correspondences: An Empirical Exploration," *Applied Economics*, 50(27), 2938–2956.

Chambers, R., Y. Chung, and R. Färe (1998): "Profit, Directional Distance Functions, and Nerlovian Efficiency," *Journal of Optimization Theory and Applications*, 98(2), 351–364.

Chang, D., and Y. Kuo (2008): "An Approach for the Two-group Discriminant

Analysis: An Application of DEA," *Mathematical and Computer Modelling*, 47(9-10), 970–981.

Charnes, A., W. Cooper, and E. Rhodes (1978): "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2(6), 429–444.

Cherchye, L., T. Kuosmanen, and T. Post (2001): "FDH Directional Distance Functions with An Application to European Commercial Banks," *Journal of Productivity Analysis*, 15(3), 201–215.

De Bock, K. W. (2017): "The Best of Two Worlds: Balancing Model Strength and Comprehensibility in Business Failure Prediction Using Spline-Rule Ensembles," *Expert Systems with Applications*, 90, 23–39.

De Caigny, A., K. Coussement, K. W. De Bock, and S. Lessmann (2019): "Incorporating Textual Information in Customer Churn Prediction Models based on a Convolutional Neural Network," *International Journal of Forecasting*.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring Labor Efficiency in Post Offices," in *The Performance of Public Enterprises: Concepts and Measurements*, ed. by M. Marchand, P. Pestieau, and H. Tulkens, pp. 243–268. North Holland, Amsterdam.

Emrouznejad, A., and G.-l. Yang (2018): "A Survey and Analysis of the First 40 Years of Scholarly Literature in DEA: 1978–2016," *Socio-Economic Planning Sciences*, 61, 4–8.

Freed, N., and F. Glover (1986): "Evaluating Alternative Linear Programming Models to Solve the Two-Group Discriminant Problem," *Decision Sciences*, 17(2), 151–162.

Glover, F. (1990): "Improved Linear Programming Models for Discriminant Analysis," *Decision Sciences*, 21(4), 771–785.

Huberty, C. J., and S. Olejnik (2006): *Applied MANOVA and Discriminant Analysis*, vol. 498. John Wiley & Sons.

Kerstens, K., and I. Van de Woestyne (2011): "Negative Data in DEA: A Simple Proportional Distance Function Approach," *Journal of the Operational Research Society*, 62(7), 1413–1419.

Lam, K. F., E. U. Choo, and J. W. Moy (1996): "Minimizing Deviations From the Group Mean: A New Linear Programming Approach for the Two-Group Classification Problem," *European Journal of Operational Research*, 88(2), 358–367.

Leon, C. F., and F. Palacios (2009): "Evaluation of Rejected Cases in an Acceptance System with Data Envelopment Analysis and Goal Programming," *Journal of the Operational Research Society*, 60(10), 1411–1420.

Lessmann, S., B. Baesens, H.-V. Seow, and L. C. Thomas (2015): "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research*, 247(1), 124–136.

Murthi, B., Y. Choi, and P. Desai (1997): "Efficiency of Mutual Funds and Portfolio Performance Measurement: A Non-Parametric Approach," *European Journal of Operational Research*, 98(2), 408–418.

Pendharkar, P., M. Khosrowpour, and J. Rodger (2000): "Application of Bayesian Network Classifiers and Data Envelopment Analysis for Mining Breast Cancer Patterns," *Journal of Computer Information Systems*, 40(4), 127–132.

Pendharkar, P., J. Rodger, and G. Yaverbaum (1999): "Association, Statistical, Mathematical and Neural Approaches for Mining Breast Cancer Patterns," *Expert Systems with Applications*, 17(3), 223–232.

Pendharkar, P. C. (2002): "A Potential Use of Data Envelopment Analysis for the Inverse Classification Problem," *Omega*, 30(3), 243–248.

Seiford, L., and J. Zhu (1998): "An Acceptance System Decision Rule with Data Envelopment Analysis," *Computers & Operations Research*, 25(4), 329–332.

Sengupta, J. (1989): "Nonparametric Tests of Efficiency of Portfolio Investment," *Journal of Economics*, 50(1), 1–15.

Silva, A. P. D., and A. Stam (1994): "Second Order Mathematical Programming Formulations for Discriminant Analysis," *European Journal of Operational Research*, 72(1), 4–22.

Smaoui, S., H. Chabchoub, and B. Aouni (2009): "Mathematical Programming Approaches to Classification Problems," *Advances in Operations Research*, 2009, Art. ID 252989.

Sueyoshi, T. (2001): "Extended DEA-Discriminant Analysis," *European Journal of Operational Research*, 131(2), 324–351.

——— (2004): "Mixed Integer Programming Approach of Extended DEA–Discriminant Analysis," *European Journal of Operational Research*, 152(1), 45–55.

——— (2006): "DEA-Discriminant Analysis: Methodological Comparison Among Eight Discriminant Analysis Approaches," *European Journal of Operational Research*, 169(1), 247–272.

Troutt, M., A. Rai, and A. Zhang (1996): "The Potential Use of DEA for Credit Applicant Acceptance Systems," *Computers & Operations Research*, 23(4), 405–408.

Yan, H., and Q. Wei (2011): "Data Envelopment Analysis Classification Machine," *Information Sciences*, 181(22), 5029–5041.

CHAPTER

3

Ordinal Classification with a Nonparametric Separating Hull: The Role of Non-Monotonicity and Nonconvexity

3.1 Introduction

Classification, as a widely discussed topic in the data mining literature, aims at assigning an observation to a predefined group based on its characteristic variables. It is normally achieved by training a classifier based on a set of training observations, then this classifier is able to classify future observations in an automated way. In order to obtain a well-trained classifier, using background knowledge is of fundamental importance in the training process. A common type of such knowledge concerns the monotone relation between the group labels and the characteristic variables: higher values of characteristic variables increase the probability that an

observation belongs to a certain group and vice versa. In student admission, for instance, one would expect the probability of admission to increase with a better academic performance. Since the monotonicity is frequently encountered in applications, several researchers became interested in the incorporation of monotonicity constraints in different classification methods, such as neural networks, decision trees and ensembles (see Cano, Gutiérrez, Krawczyk, Woźniak, and García (2019) for a recent survey).

However, in many applications, the relation between the group labels and the characteristic variables could also show non-monotonicity. Note that the non-monotonicity discussed here is not the non-monotonic data that arises due to noise, or the omission of important predictors (e.g., Feelders and Pardoel (2003); Rademaker, De Baets, and De Meyer (2009)). The interest is on the natural non-monotonic relation bounded with the application itself. For example, in medical diagnoses, both high values and low values may indicate symptoms of certain diseases. Similar example can be found in differentiating healthy firms from the poor ones by debt-to-equity ratio. A firm is perceived to be poor while it has either a very high or a very low debt-to-equity ratio. In such applications, the relation between the characteristic variables and the group label is apparently non-monotonic.

Although the research on incorporating the monotonic relation is increasing, the parallel research question on how to properly reflect the non-monotonic relation in classification remains valid. To the best of our knowledge, the research on dealing with the non-monotonicity in classification is quite limited. In Lam and Choo (1993), the non-monotonicity is treated by discretizing the non-monotonic variables into several partially monotonic variables. Specifically, each non-monotonic variable is visualized into a one dimension diagram and then the diagram is used

to decide the cut-off values. Different ways to discretize the non-monotonic variables may lead to different classification results. In this contribution, the primary interest is to incorporate the non-monotonic variables while developing a nonparametric classifier for the ordinal classification.

In the literature, the nonparametric classifier is built explicitly based on the Data Envelopment Analysis (DEA) method. There are typically two types of DEA-based classifiers. The first type is primarily based on the use of goal programming (see Sueyoshi (1999, 2001, 2004) for the details) rather than the standard DEA models. Banker, Chang, and Cooper (2002) have argued that researchers should avoid calling goal programming models as DEA methods. The second type is based on the traditional DEA models proposed by Banker, Charnes, and Cooper (1984) which is originally served as a relative efficiency measure for ranking a sample of observations. The DEA-based classifier discussed in this contribution explicitly refers to the second type.

As a data-based method, DEA is widely applied in production analysis and portfolio analysis (Emrouznejad and Yang (2018)), but it is not primarily developed for solving the classification problem. Troutt, Rai, and Zhang (1996) is the first article that employs a DEA-based nonparametric frontier as an acceptability frontier in credit applicant acceptance systems. Following the pioneering work of Troutt, Rai, and Zhang (1996), the DEA-based frontier has been adapted to better represent a separating frontier that distinguishes between two groups of observations (e.g., Seiford and Zhu (1998); Leon and Palacios (2009); Yan and Wei (2011); Pendharkar, Rodger, and Yaverbaum (1999); Pendharkar, Khosrowpour, and Rodger (2000); Pendharkar (2002, 2011) etc.). The idea of using the nonparametric frontier in classification is in line with the general idea of locating a separating hypersurface by mathematical programming (MP) methods. The ma-

jority of the MP methods requires an assumption on the shape of the separating hypersurface. It might be a simple linear separating hyperplane (e.g., see Freed and Glover (1986), Lam, Choo, and Moy (1996), Sueyoshi (2004), among others) or rather a separating hypersurface that is potentially nonlinear (e.g., see Silva and Stam (1994), Smaoui, Chabchoub, and Aouni (2009), etc.). When it comes to the DEA-based classifiers, they do not make any particular assumption on the functional form of the hypersurface. In this sense, the DEA frontier methods are more flexible in terms of closely enveloping the observations.

The DEA-based separating frontier is monotonic and convex. These two characteristics correspond to two types of background information in classification. One type of background information is the monotonic relation between the characteristic variables and the group label. The monotonicity is described by the axiom of free disposability on the monotonic variables. The other type of the background information concerns the substitution relation among the characteristic variables. This relation is reflected by the axiom of convexity. To extend the nonparametric frontier classifier for a wider range of classification problems, the main purpose of this contribution is to construct a nonparametric classifier where some of the current axioms of the DEA classifier are adapted. Specifically, this research is driven by the following two motivations.

First, we consider the standard axiom of free disposability in DEA methods intuitively unappealing for describing the non-monotonic variables, since it amounts to assume that the characteristic variables can be disposed off without any limitation. We suggest to replace this free disposability assumption with a generalized disposal assumption that makes the disposal of the non-monotonic variables only possible within a limited value range. This is a direct extension of the S-disposal assumption which describes the congestion in the input-space (see Briec, Kerstens,

and Van de Woestyne (2016) for a theoretical treatment and see Briec, Kerstens, and Van de Woestyne (2018) for an empirical application). The generalized disposability assumption defined in this contribution is independent of defining a input-type or a output-type variable. It is used to describe the non-monotonic variable which is preferred within a value range rather than monotonically favoring a higher or a lower value.

Second, the axiom of convexity which implies a substitution relation among the characteristic variables is relaxed. If there is no prior information on the relation between the characteristic variables, retaining an assumption of convexity could degrade the classification performance while applying a nonparametric frontier method. Pendharkar, Rodger, and Yaverbaum (1999, p. 231) already claim that the convexity assumption embedded in the DEA-based classifier may be the reason why neural networks outperform the frontier methods in mining breast cancer patterns. However, to the best of our knowledge, all the current frontier classifiers stick to this assumption on convexity. This calls for the development of a nonconvex frontier classifier which is possible with the Free Disposal Hull (FDH) method that has been initially proposed by Deprins, Simar, and Tulkens (1984).

This contribution unfolds as follows. Section 3.2 graphically illustrates the shape of the separating hypersurface when the ordinal classification problem has both monotonic and non-monotonic variables. Instead of having one separating frontier generated from the nonparametric method, a separating hull consists of several separating frontiers is derived to differentiate between the groups of observations. Then, the models and procedures used to construct the nonparametric classifiers are presented in Section 3.3. Specifically, a generalized free disposability assumption is proposed to capture the property of the non-monotonic variables. Then, a dominance adapting directional distance function (DAD) that measures

the distance to the corresponding frontier is defined to accommodate the generalized free disposability assumption. Finally, an algorithm is introduced to predict the membership of an observation with the proposed separating hull. In Section 3.4, an illustrative example is available for comparing the proposed separating hull method with some existing methods. Section 3.5 concludes, discusses limitations, and offers directions for future research.

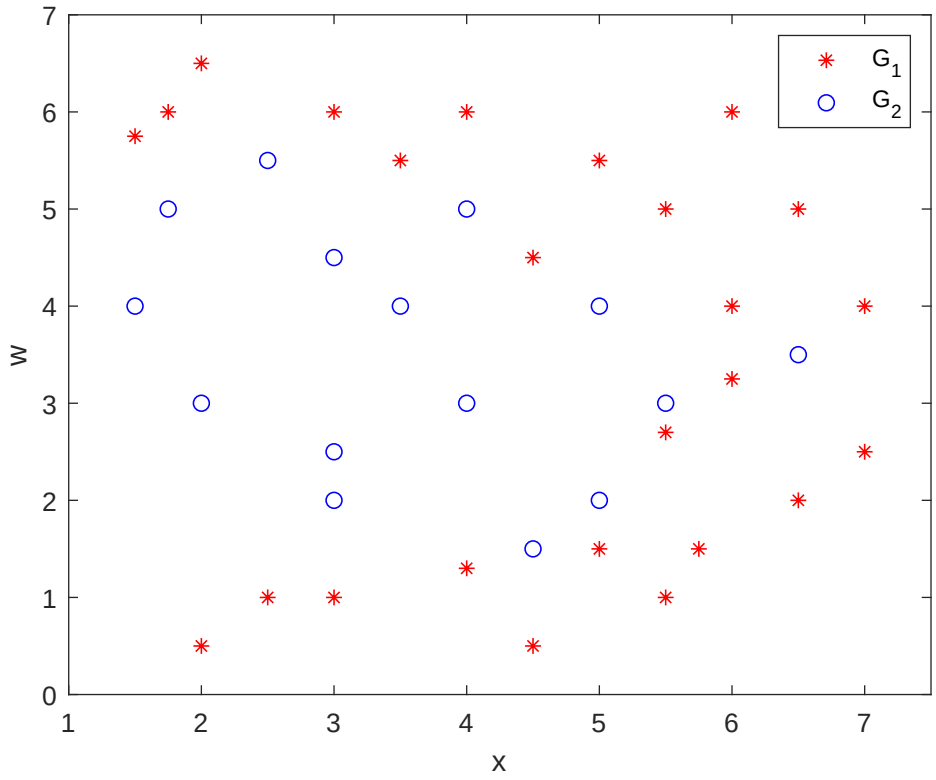
3.2 Basic Intuitions and Geometrical Illustration

An ordinal classification problem with two ordered groups of observations is investigated, namely, a bad group G_1 and a good group G_2. The observations are characterized by both monotonic and non-monotonic characteristic variables. Specifically, the possibility of belonging to the good group increases with the augmentation of the monotonically increasing variable and with the reduction of the monotonically decreasing variable. While for the non-monotonic characteristic variables, an observation belongs to the good group if the corresponding value is located within a certain value range. Both positive and negative deviations from the value range indicate that the observation is more likely belonging to the bad group. Based on the training observations whose membership is a priori known, the nonparametric classifier is trained and expected to be capable of predicting the membership of an observation.

A two-dimensional classification problem with two characteristic variables x and w is illustrated to introduce the intuitive idea of our nonparametric classifier. The characteristic variable x corresponds to a monotonically decreasing variable and w corresponds to a non-monotonic characteristic variable. Thus, the smaller the value of x is, the higher the possibility of classifying an observation to the good

group G_2 is. While for the value of w, the observation is classified into G_2 if it is within a preferred value range. In Figure 3.1, the training observations belonging to G_1 are marked by the asterisks and those from G_2 are marked by the circles. Clearly, these two groups are not easily linearly separable.

Figure 3.1: A classification example with both monotonic and non-monotonic variables

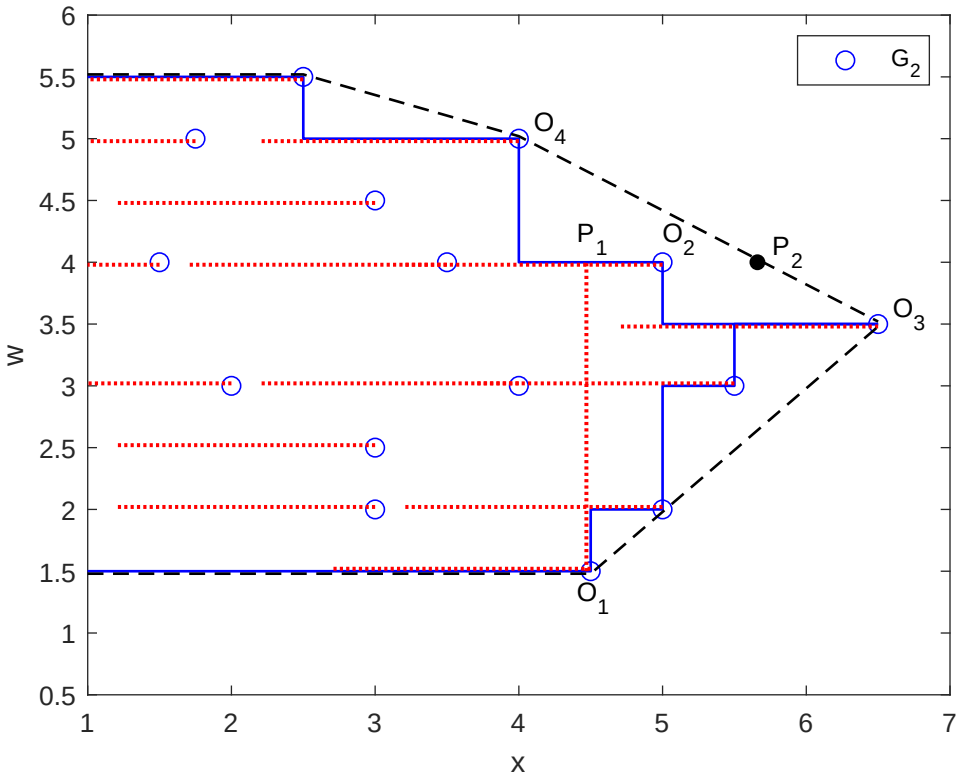


Based on the background knowledge on the monotonic and non-monotonic relations between the characteristic variables and the group label, a nonconvex (NC) attainable set can be constructed to describe all combinations of characteristic values whose corresponding observations belong to the good group G_2. First, if an observation has the same characteristic values as one training observation from G_2, then it certainly belongs to the attainable set of G_2. Then, with the monotonicity

defined for x, an observation belongs to G_2 if its value of x is smaller comparing to that of a training observation from G_2. This generates all the horizontal dotted lines to the left of the training observations in Figure 3.2. Vertically, take x with a value of 4.5 as an example to illustrate the non-monotonicity constraint, the preferred value range of w is [1.5,4]. The lower limit of this value range is determined by the training observation O_1 while the upper limit is decided by point P_1. The point P_1 is on the dotted line generated from the training observation O_2. For every possible value of x, the upper and lower value limits of the non-monotonic characteristic variable w are generated accordingly. Thus, the area to the left of the solid polylines in Figure 3.2 represents the nonconvex (NC) attainable set of G_2.

If an additional assumption on convexity is introduced, a convex (C) attainable set is obtained which is lager or equal comparing to the NC one. The convexity assumption corresponds to the background knowledge on the substitution relation between the characteristic variables. When it gets to our numerical example in Figure 3.2, since the training observations O_3 and O_4 are in the C attainable set of G_2, their linear combination P_2 is also supposed to be in the same attainable set. Note that there is no training observation from G_2 that directly dominates P_2. Horizontally, the value of x of point P_2 is larger than that of the training observation O_2 which has the same value of w. Vertically, the value of w of point P_2 is beyond the preferred value range of w determined solely by the training observations. It is the assumed substitution relation that makes the point P_2 acceptable for the attainable set. In Figure 3.2, this is the area restricted to the left of the dashed lines. The C attainable set also describes all possible combinations of characteristic values whose corresponding observations are believed to belong to the good group G_2. It is constructed by employing two types of background knowledge: one is the monotonicity and non-monotonicity constraints, the other

Figure 3.2: The nonconvex and convex attainable set with characteristic variables x and w

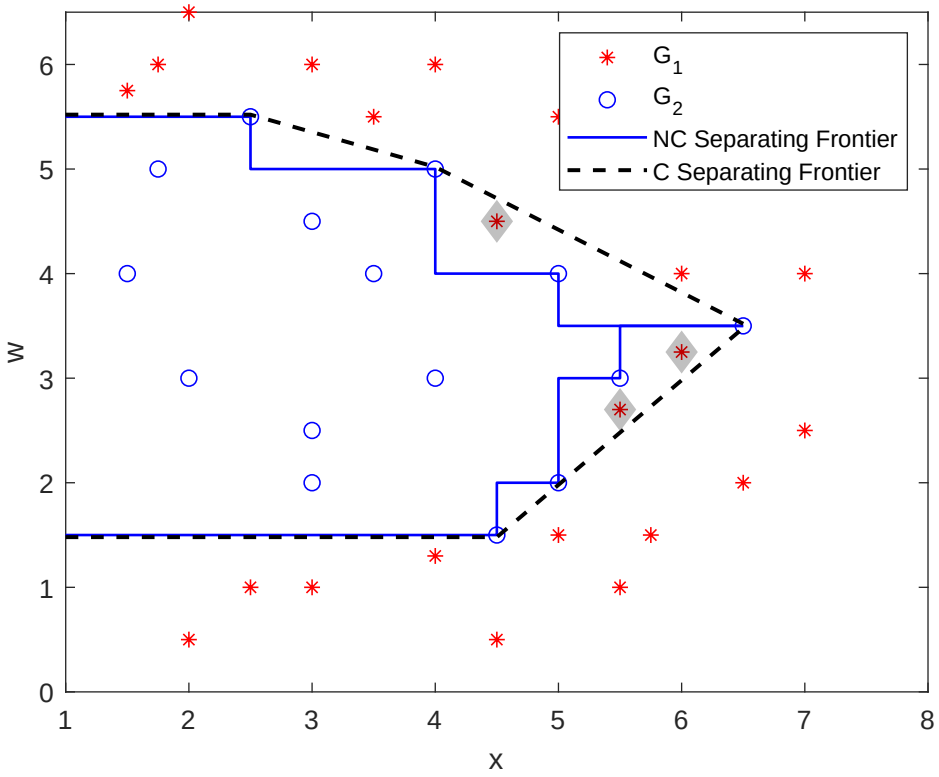


is the substitution relation.

The boundary of the attainable set is called the separation hull which could be either NC or C depending on the assumptions made. Both the NC and the C separating hulls are non-monotonic since there exists non-monotonic characteristic variables. In Figure 3.3, the solid staircase lines represent the NC separating hull and the dashed lines represent the C separating hull. It is observed that all training observations from G_2 which are marked by the circles are located within the separating hull for both the NC and the C cases. For the training observations from G_1, all of them are located beyond the NC separating hull but three of them are located within the C separating hull. That is, these three training observations

marked with the faded rhombuses are misclassified while applying the C separating hull. Clearly, the NC separating hull provides a better separation than the C separating hull, since the NC one envelops the observations tighter.

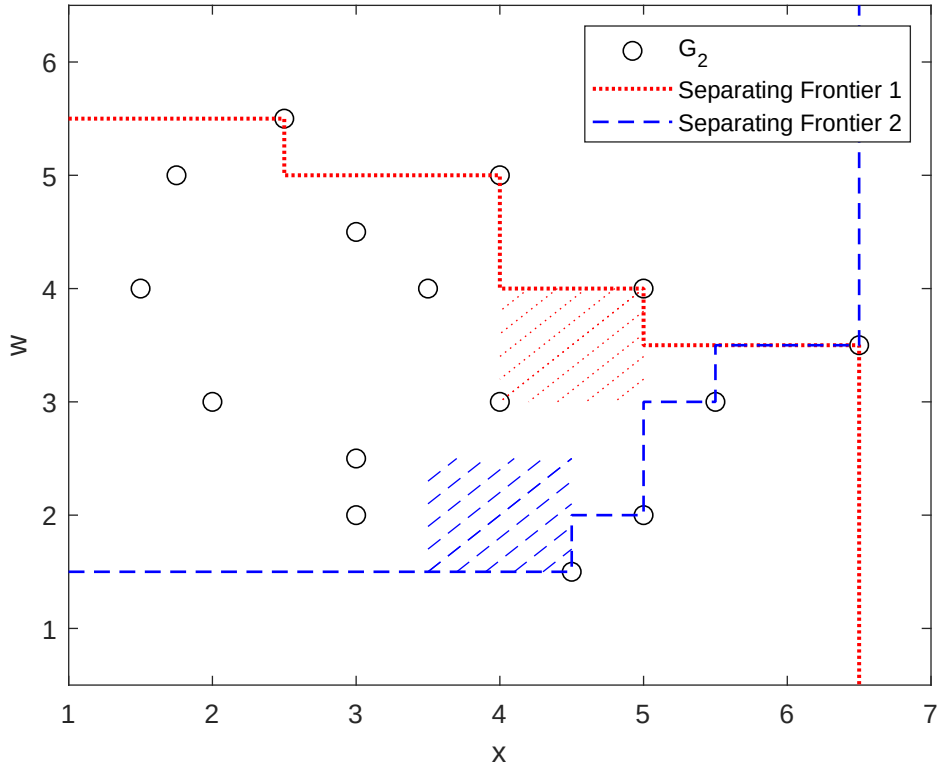
Figure 3.3: The nonconvex and convex separating hulls



As shown in Figure 3.3, predicting the membership of an observation is now transformed into the problem of positioning this observation relative to the derived separating hull. If it is within the separating hull, then it belongs to the good group G_2, otherwise it belongs to the bad group G_1. In our non-parametric classifier, the attainable set of G_2 is represented as the intersection of some subsets. Therefore, the above problem could be further simplified into the question whether the observation is located within corresponding separating frontiers. These separating frontiers jointly determines the preferred value range of the non-monotonic

variables as well as the limits of the monotonic variables.

Figure 3.4: The nonconvex subsets with regard to characteristic variables x and w



Since there is only one non-monotonic variable w in this numerical example, two separating frontiers are enough to decide the upper and lower value limits of w. The NC case is explained in detail as an example in Figure 3.4. We start with the attainable set of G_2 which could be represented as the intersection of two subsets. One subset is the area restricted to the third quadrant located below and to the left of the dotted lines, and the other is the area restricted to the second quadrant located above and to the left of the dashed lines. The boundary of the former subset is called the separating frontier 1, and the latter one is called the separating frontier 2. For an observation with an attainable value of x, the separating frontier 1 gives the upper value limit of its w and the separating frontier 2 determines its

lower value limit. With regard to the attainable value of w, the upper value limit of x is jointly decided by the separating frontier 1 and 2. An observation belongs to G_2 if and only if it is located within both separating frontiers. Note that an observation is located within the separating frontier 1 if it is located below the frontier. On the contrary, an observation that is located above the separating frontier 2 is considered to be within the separating frontier 2.

3.3 Nonparametric Classifier for Cases with Non-Monotonic Characteristic Variables

3.3.1 Basic Concepts

Consider an ordinal classification problem with a set of labelled training observations. These training observations constitute the training sample G to be used for training a classifier. Every training observation is characterized by K characteristic variables. Prior knowledge on the group label is given such that the training observations are exhaustively classified into two groups, namely, the bad group G_1 and the good group G_2. The training observations from G_1 are labelled 1 while those from G_2 are labelled 2. The ordered group label means that the observations with a larger value of the label is perceived to be better. Note that the group G_1 together with the group G_2 form a partition of G. Thus, $G_1 \cup G_2 = G$ and $G_1 \cap G_2 = \emptyset$. A test observation is characterized by the same characteristic variables as the training observations do. The only difference is that the group label of a test observation is unknown and must be predicted by the trained classifier.

With the background information on the relation between the group labels and

the characteristic variables, the K characteristic variables are categorized into two categories and totally three types. The first category is the characteristic variables which are monotonically related to the group labels. Specifically, if the value of the group label increases (decreases) with the decreasing (increasing) of a characteristic variable, then it is known as a monotonically decreasing characteristic variable, denoted by $X \in \mathbb{R}^m$. Otherwise, if the value of the group label increases (decreases) with the increasing (decreasing) of a characteristic variable, then it is known as a monotonically increasing characteristic variable, denoted by $Y \in \mathbb{R}^s$. The second category consists of the characteristic variables that are not monotonically related to the group labels, denoted by $W \in \mathbb{R}^D$. The non-monotonicity is illustrated by the fact that neither a very high nor a very low characteristic value contributes to increasing the group label. Having the non-monotonic variables within a certain value range, corresponding observations are most likely to be labelled 2. To sum up, an observation is denoted by $Z = (X, Y, W) \in \mathbb{R}^m \times \mathbb{R}^s \times \mathbb{R}^D$. Note that $m + s + D = K$.

A direction vector which consists of three components is defined to describe the relation between the characteristic variables and the group label. Specifically, the three components correspond to three types of characteristic variables. To be meaningful, the monotonically decreasing relation of X is depicted by $g_{x_i} < 0$ ($i \in [m]$), where $[m]$ denotes the set $\{1, \ldots, m\}$. The monotonically increasing relation of Y is depicted by $g_{y_r} > 0$ ($r \in [s]$), where $[s]$ denotes the set $\{1, \ldots, s\}$. Finally for the direction vector g_{w_d} ($d \in [D]$), it could be positive if a lower value limit of w_d is to be determined while being negative to have an upper value limit. Remark that $[D]$ denotes the set $\{1, \ldots, D\}$. The direction vector is denoted by $g_p = (g_X, g_Y, g_{W,p}) \in \mathbb{R}^m \times \mathbb{R}^s \times \mathbb{R}^D$ for all $p \in [2^D]$, where $[2^D]$ denotes the set $\{1, \ldots, 2^D\}$. Take the case with $Z = (X, Y, W) \in \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^2$ as an example, four direction vectors are needed to describe the full relation: $g_1 = (-1, 1, 1, 1)$,

$g_2 = (-1, 1, 1, -1)$, $g_3 = (-1, 1, -1, 1)$ and $g_4 = (-1, 1, -1, -1)$. Note that the absolute value of the elements in g_p does not necessarily need to be 1. In this contribution, we use one so that g_p only indicates the monotonic relation without any influence on the value of the efficiency measure. With regard to the direction vector, a positive value always stands for the monotonically increasing relation and a negative one is assigned for describing the monotonically decreasing relation.

3.3.2 Acceptance Possibility Set

An acceptance possibility set (APS) is constructed to describe the attainable set of the good group G_2. It is constituted of all combinations of characteristic values whose group label is no smaller than 2. In the binary case discussed here, the group label of all observations within this APS is 2.

Based on the training observations from the good group G_2, a NC APS is constructed from the axiom of free disposability. For the monotonically decreasing characteristic variable, the axiom of free disposability provides an upper value limit for X. That is, comparing to a training observation $Z = (X, Y, W)$ which is from G_2, an observation $\hat{Z} = (\hat{X}, Y, W)$ is also labelled 2 if $\hat{X} \leq X$. For the monotonically increasing characteristic variable, the axiom of free disposability provides a lower value limit for Y. That is, comparing to a training observation $Z = (X, Y, W)$ that is from G_2, an observation $\hat{Z} = (X, \hat{Y}, W)$ is also labelled 2 if $\hat{Y} \geq Y$. Rather than requiring either an upper or a lower value limit, both limits are needed to classify an observation based on its non-monotonic variable W. Specifically, after fixing the value of X and Y, an observation is labelled 2 if and only if every w_d is within a preferred value range. For all w_d ($d \in [D]$), the preferred value range is bounded by an upper value limit and a lower value limit. By assuming that w_d has a monotonically decreasing relation with the group label,

the upper value limit of w_d is obtained. While a lower value limit of w_d is calculated by assuming that w_d has a monotonically increasing relation. Different from the monotonic characteristic variables, a generalized free disposability is defined for the non-monotonic characteristic variables.

Before constructing the APS, we first define a generalized free disposability to characterize both the monotonic and non-monotonic characteristic variables. $2^{[D]}$ denotes the set of all subsets of $[D]$. Remark that $\emptyset \in 2^{[D]}$ by definition.

Definition 3.3.1. *The APS of G_2 satisfies the generalized disposal assumption if the following holds true: if for every $I_p \in 2^{[D]}$ there exists a training observation $Z = (X, Y, W)$ from G_2 with $\hat{X} \leq X$, $\hat{Y} \geq Y$ and $\hat{W} \geq_p W$, then the observation $\hat{Z} = (\hat{X}, \hat{Y}, \hat{W})$ also belongs to G_2.*

where

$$\hat{W} \geq_p W \iff \begin{cases} \hat{w_d} \geq w_d & \text{if } d \in I_p; \\ \hat{w_d} \leq w_d & \text{else.} \end{cases} \tag{3.3.1}$$

Note that $I_p \in 2^{[D]}$ exhausts the possible combinations of the monotonic relations with regard to the D non-monotonic characteristic variables. Take the case with $Z = (X, Y, W) \in \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^2$ as an example, $I_p \in \{\{1, 2\}, \{1\}, \{2\}, \emptyset\}$.

For every $I_p \in 2^{[D]}$, a sub-APS denoted by $T_{NC,p}$ can be derived corresponding to the direction vector $g_p = (g_X, g_Y, g_{W,p}) \in \mathbb{R}^m \times \mathbb{R}^s \times \mathbb{R}^D$:

$$\begin{aligned} T_{NC,p} = \{(X, Y, W) \in \mathbb{R}^{m+s+D} \quad | & \sum_{j=1}^{n_2} \lambda_j X_j \geq X, \sum_{j=1}^{n_2} \lambda_j Y_j \leq Y, \\ & \sum_{j=1}^{n_2} \lambda_j W_j \geq_p W, \sum_{j=1}^{n_2} \lambda_j = 1, \lambda_j \in \{0, 1\}\}. \end{aligned} \tag{3.3.2}$$

The NC APS of G_2 is the intersection of all sub-APS $T_{NC,p}$ generated from the

possible g_p $p \in [2^D]$. That is,

$$T_{NC} = \bigcap_{p \in [2^D]} T_{NC,p} \tag{3.3.3}$$

Based on the above NC APS of G_2, having an additional axiom on convexity leads to a C APS. As explained earlier, the convexity in classification implies a substitution relation among characteristic variables. If the prior information of such a substitution relation is provided, then a C APS is constructed as follows:

$$T_C = \bigcap_{p \in [2^D]} T_{C,p} \tag{3.3.4}$$

where

$$\begin{aligned} T_{C,p} = \{(X, Y, W) \in \mathbb{R}^{m+s+D} \quad & | \sum_{j=1}^{n_2} \lambda_j X_j \geq X, \sum_{j=1}^{n_2} \lambda_j Y_j \leq Y, \\ & \sum_{j=1}^{n_2} \lambda_j W_j \geq_p W, \sum_{j=1}^{n_2} \lambda_j = 1, \lambda_j \geq 0\}. \end{aligned} \tag{3.3.5}$$

Obviously, $T_{NC,p} \subseteq T_{C,p}$: a NC monotonic hull is a subset of a C monotonic hull. Furthermore, $T_{NC} \subseteq T_C$ also holds. Put differently, the NC APS provides a tighter envelopment of the training observations than the C one does.

In order to simplify the expressions, we use the following notation to stand for the sub-APS under the NC and C cases:

$$\begin{aligned} T_{\Lambda,p} = \{(X, Y, W) \in \mathbb{R}^{m+s+D} \quad & | \sum_{j=1}^{n_2} \lambda_j X_j \geq X, \sum_{j=1}^{n_2} \lambda_j Y_j \leq Y, \\ & \sum_{j=1}^{n_2} \lambda_j W_j \geq_p W, \sum_{j=1}^{n_2} \lambda_j = 1, \lambda_j \in \Lambda, j \in [n_2]\}. \end{aligned} \tag{3.3.6}$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \quad \text{or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0,1\}\}.$$

Then, T_Λ stands for the APS of G_2. Specifically, $T_{\Lambda^{\mathrm{NC}}} = \bigcap_{p\in[2^D]} T_{\Lambda^{\mathrm{NC}},p}$ corresponds to the NC case while $T_{\Lambda^{\mathrm{C}}} = \bigcap_{p\in[2^D]} T_{\Lambda^{\mathrm{C}},p}$ corresponds to the C case .

Following Chambers, Chung, and Färe (1998), the frontier of the sub-APS could be represented by the following dominance adapting directional distance function (DAD) $D_{\Lambda,p}(Z)$:

$$D_{\Lambda,p}(Z) = \sup\{\delta \in \mathbb{R} \mid Z + \delta(-g_p \circ v) \in T_{\Lambda,p}\}. \tag{3.3.7}$$

where $\circ$ represents the Hadamard product, also known as the element-wise product. The projection direction vector $-g_p \circ v$ is characterized by the direction vector $-g_p$ and the scaling vector v which is non-negative. The projection direction is opposite to $g_p = (g_X, g_Y, g_{W,p})$ which represents the monotonic relations of the characteristic variables. In this way, the monotonically decreasing characteristic variables are increased and the monotonically increasing ones are reduced while increasing the value of δ, which is the favorable behavior.

There are different choices possible for the scaling vector v in practical applications. In our classification context with potentially negative inputs and outputs, a common choice is using $v = (|X_0|, |Y_0|, |W_0|)$ for an observation $Z = (X_0, Y_0, W_0)$. This ensures that the DAD measure ($D_{\Lambda,p}(Z)$) obtains a proportional interpretation (see Briec (1997) and Kerstens and Van de Woestyne (2011)). Of course, such a percentage interpretation is not indispensable in our classification context where the focus is rather on the sign of the DAD measure. However, for convenience we stick to this proportional distance function.

The assumption of convexity differentiates the sub-APS $T_{\Lambda^{\mathrm{NC}},p}$ from $T_{\Lambda^{\mathrm{C}},p}$. However, this does not change the definition of the DDF measure, only the value of the DAD measure may be enlarged. That is, $D_{\Lambda^{\mathrm{NC}},p}(Z) \leq D_{\Lambda^{\mathrm{C}},p}(Z)$. $D_{\Lambda,p}(Z)$ serves as an indicator that positions the observations relative to the boundary of the sub-APS $(T_{\Lambda,p})$. It is well-defined for all possible observations $Z = (X, Y, W) \in \mathbb{R}^m \times \mathbb{R}^s \times \mathbb{R}^D$. A non-negative $D_{\Lambda,p}(Z)$ means z is in the interior of $T_{\Lambda,p}$. If the observation Z is located beyond $T_{\Lambda,p}$, then $D_{\Lambda,p}(Z)$ becomes negative and it is projected onto the frontier in the direction of g_p.

Eventually, it is the boundary of the APS of G_2 that constitutes the separating frontier to differentiate between two groups of observations. Derived from the relation between the APS and the sub-APS, the final DDF measure is calculated by the following:

$$D_\Lambda(Z) = \min_{p \in [2^D]} D_{\Lambda,p}(Z) \tag{3.3.8}$$

3.3.3 Models for Calculating the Nonconvex and Convex Separating Hulls

With respect to a specific $g_p = (g_X, g_Y, g_{W,p}) \in \mathbb{R}^m \times \mathbb{R}^s \times \mathbb{R}^D$, the following model is solved for the observation $Z_0 = (X_0, Y_0, W_0)$. Note that $g_X < 0$ and $g_Y > 0$ always hold while $g_{W,p}$ varies for different p.

$$
\begin{aligned}
\max_{\lambda_j, \delta_{\Lambda,p}} \quad & \delta_{\Lambda,p} \\
s.t. \quad & \sum_{j=1}^{n_2} \lambda_j x_{i,j} + s_i^+ = x_{i,0} + \delta_{\Lambda,p}|x_{i,0}| && \forall i \in [m] \\
& \sum_{j=1}^{n_2} \lambda_j y_{r,j} - s_r^- = y_{r,0} - \delta_{\Lambda,p}|y_{r,0}| && \forall r \in [s] \\
& \sum_{j=1}^{n_2} \lambda_j w_{d,j} - s_d^- = w_{d,0} - \delta_{\Lambda,p}|w_{d,0}| && \forall d \in I_p \\
& \sum_{j=1}^{n_2} \lambda_j w_{d,j} + s_d^+ = w_{d,0} + \delta_{\Lambda,p}|w_{d,0}| && \forall d \in [D] \setminus I_p \\
& \sum_{j=1}^{n_2} \lambda_j = 1 \\
& \lambda_j \in \Lambda && \forall j \in [n_2]
\end{aligned}
\tag{3.3.9}
$$

where

$$
\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \text{ or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0,1\}\}.
$$

In the C case, model (3.3.9) is a linear programming (LP) problem, while it involves solving a binary mixed integer program (BMIP) for the NC case. To remedy the computational issue in the NC case, a fast implicit enumeration-based method is proposed by Cherchye, Kuosmanen, and Post (2001) requiring only to compute minima and maxima of lists of ratios. Instead of solving a BMIP model, the following exact solution is obtained for model (3.3.9) under the NC case:

$$
\begin{aligned}
\delta^*_{\Lambda^{\mathrm{NC}},p} = \max_{j\in[n_2]} \Bigg(& \min_{i\in[m]} \left(\frac{x_{i,j} - x_{i,0}}{|x_{i,0}|} \right), \min_{r\in[s]} \left(\frac{y_{r,0} - y_{r,j}}{|y_{r,0}|} \right), \\
& \min_{d\in I_p} \left(\frac{w_{d,0} - w_{d,j}}{|w_{d,0}|} \right), \min_{d\in[D]\setminus I_p} \left(\frac{w_{d,j} - w_{d,0}}{|w_{d,0}|} \right) \Bigg)
\end{aligned}
\tag{3.3.10}
$$

By solving model (3.3.9), the optimal values of $\delta_{\Lambda,p}$ are obtained. $\delta^*_{\Lambda,p}$ measures

the proportional distance of the observation $Z_0 = (X_0, Y_0, W_0)$ to the boundary of $T_{\Lambda,p}$. The s_i^{+*}, s_r^{-*}, s_d^{+*} and s_d^{-*} are the slacks and surpluses. Only if $\delta^*_{\Lambda,p}$ and all the slacks and surpluses equal zero, then the observation $Z_0 = (X_0, Y_0, W_0)$ is strongly efficient. Otherwise if only $\delta^*_{\Lambda,p} = 0$ holds for the observation $Z_0 = (X_0, Y_0, W_0)$, it is considered to be weakly efficient. The possible situations for the distance measure under a specific $g_p = (g_X, g_Y, g_{W,p}) \in \mathbb{R}^m \times \mathbb{R}^s \times \mathbb{R}^D$ are the following:

(s.1) If $\delta^*_{\Lambda,p} > 0$, then the observation is inefficient and located within $T_{\Lambda,p}$;

(s.2) If $\delta^*_{\Lambda,p} < 0$, then the observation is super-efficient and located outside $T_{\Lambda,p}$.

(s.3) If $\delta^*_{\Lambda,p} = 0$ and the slacks and surpluses satisfy $s_i^{+*} = s_r^{-*} = s_d^{+*} = s_d^{-*} = 0$, then the observation is strongly efficient.

(s.4) If $\delta^*_{\Lambda,p} = 0$ and not all slacks and surpluses equals 0, then the observation is weakly efficient.

By calculating $\delta^*_{\Lambda,p}$ for the training observations from G_2, the separating frontier p is represented by the set of the strongly efficient training observations which satisfy (s.3). This frontier set of the separating frontier p is represented by $FS_{\Lambda,p}$.

With the $\delta_{\Lambda,p}$ calculated for every possible g_p where $p \in [2^D]$, the final distance measure of the observation $Z_0 = (X_0, Y_0, W_0)$ to the boundary of T_Λ is derived from:

$$\delta^*_\Lambda = \min_{p \in [2^D]} \delta^*_{\Lambda,p}. \tag{3.3.11}$$

The ultimate efficiency δ^*_Λ measures the nearest proportional distance to the boundary of T_Λ. For the observation $Z_0 = (X_0, Y_0, W_0)$, if its δ^*_Λ is positive, then for $\forall p$, it has $\delta^*_{\Lambda,p} > 0$. This observation is within the attainable set T_Λ. If its δ^*_Λ is negative, then there exists at least one $p \in [2^D]$ that corresponds to a negative

$\delta^*_{\Lambda,p}$. This observation $Z_0 = (X_0, Y_0, W_0)$ is located outside the attainable set T_Λ. If $\delta^*_\Lambda = 0$, then for $\forall p$, it has at least one $\delta^*_{\Lambda,p} = 0$ and has other $\delta^*_{\Lambda,p}$ being positive. This observation $Z_0 = (X_0, Y_0, W_0)$ is on the boundary of the attainable set T_Λ, also on the strongly efficient frontier of at least one $T_{\Lambda,p}$. Note that if the observation $Z_0 = (X_0, Y_0, W_0)$ is on the weakly efficient frontier of one $T_{\Lambda,p}$ but not the strongly efficient one, then it is not on the boundary of T_Λ. To be specific, if $\delta^*_{\Lambda,p} = 0$, then the observation $Z_0 = (X_0, Y_0, W_0)$ is weakly efficient. For being strongly efficient, all constraints have to be satisfied with equalities while model (3.3.9) achieves the optimum for this observation.

By calculating δ^*_Λ for the training observations from G_2, the aggregate separating hypersurface can be represented by the set of the training observations with $\delta^*_\Lambda = 0$. FS_Λ is used to denote this frontier set. Note that $FS_\Lambda = \underset{p\in[2^D]}{\cup} FS_{\Lambda,p}$. That is, the aggregate separating hypersurface is the boundary of the intersection of all 2^D sub-APSs. If there exist only nonmonotonic characteristic variables, then the aggregate separating hypersurface is a closed hull. Normally, the set FS_Λ under the NC case is different from that under the C case. All frontier observations in $FS_{\Lambda^{\mathrm{C}}}$ could be found in $FS_{\Lambda^{\mathrm{NC}}}$. However, not all frontier observations in $FS_{\Lambda^{\mathrm{NC}}}$ belongs to $FS_{\Lambda^{\mathrm{C}}}$, since some frontier observations generated under the NC case are dominated by some convex combinations of the training observations. Therefore, $FS_{\Lambda^{\mathrm{C}}} \subsetneqq FS_{\Lambda^{\mathrm{NC}}}$.

3.3.4 Separating Hull based Discriminant Rules

The aggregate separating hull represented by the training observations in FS_Λ is used to label an observation denoted by $Z_0 = (X_0, Y_0, W_0)$. In this contribution, the sign of the final distance measure matters more than the exact value (see supra). Therefore, there is no need to solve model (3.3.9) for 2^D times if a stop

criterion is met. Before introducing the algorithm, the following model is used to calculate the proportional distance of the observation $Z_0 = (X_0, Y_0, W_0)$ to the separating frontier p:

$$
\begin{aligned}
\max_{\lambda_j, \hat{\delta}_{\Lambda,p}} \quad & \hat{\delta}_{\Lambda,p} \\
s.t. \quad & \sum_{j \in FS_{\Lambda,p}} \lambda_j x_{i,j} + s_i^+ = x_{i,0} + \hat{\delta}_{\Lambda,p} |x_{i,0}| && \forall i \in [m] \\
& \sum_{j \in FS_{\Lambda,p}} \lambda_j y_{r,j} - s_r^- = y_{r,0} - \hat{\delta}_{\Lambda,p} |y_{r,0}| && \forall r \in [s] \\
& \sum_{j \in FS_{\Lambda,p}} \lambda_j w_{d,j} - s_d^- = w_{d,0} - \hat{\delta}_{\Lambda,p} |w_{d,0}| && \forall d \in I_p \\
& \sum_{j \in FS_{\Lambda,p}} \lambda_j w_{d,j} + s_d^+ = w_{d,0} + \hat{\delta}_{\Lambda,p} |w_{d,0}| && \forall d \in [D] \setminus I_p \\
& \sum_{j \in FS_{\Lambda,p}} \lambda_j = 1 \\
& \lambda_j \in \Lambda && \forall j \in FS_{\Lambda,p}
\end{aligned}
\tag{3.3.12}
$$

where

$$
\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \text{ or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0, 1\}\}.
$$

By solving model (3.3.12), the optimal distance measure $\hat{\delta}^*_{\Lambda,p}$ is derived. The only difference between model (3.3.9) and model (3.3.12) is that less training observations from G_2 are used in the left hand side of the inequalities. In particular, only the frontier observations in $FS_{\Lambda,p}$ are used in model (3.3.12). Although $\delta^*_{\Lambda,p} = \hat{\delta}^*_{\Lambda,p}$ always holds, the decrease in sample size can save some computational time.

In the following, the algorithm for labeling every test observation $Z_0 = (X_0, Y_0, W_0)$ is designed:

Step 1: Initialize p as 1 and v_0 as an empty vector.

Step 2: If $p \leq 2^D$, then go to Step 3, otherwise go to Step 6.

Step 3: Generate $\hat{\delta}^*_{\Lambda,p}$, s_i^{+*}, s_r^{-*}, s_d^{+*} and s_d^{-*} by solving model (3.3.12).

Step 4: Set $v_p = v_{p-1} \cup \hat{\delta}^*_{\Lambda,p}$,

(a) If $\hat{\delta}^*_{\Lambda,p} > 0$, then go to Step 5.

(b) If $\hat{\delta}^*_{\Lambda,p} < 0$, then go to Step 6.

(c) If $\hat{\delta}^*_{\Lambda,p} = 0$ and all slacks and surpluses equal zeros, then go to Step 6.

(d) If $\hat{\delta}^*_{\Lambda,p} = 0$ and not all slacks and surpluses equal zero, then go to Step 5.

Step 5: Set $p = p + 1$, then go to Step 2.

Step 6: $\hat{\delta}^*_{\Lambda} = \min v_p$.

Step 7: Decide the group label based on the sign of $\hat{\delta}^*_{\Lambda}$.

(a) If $\hat{\delta}^*_{\Lambda} \geq 0$, then the observation is labelled 2 and belongs to the good group.

(b) If $\hat{\delta}^*_{\Lambda} < 0$, then the observation is labelled 1 and belongs to the bad group.

3.4 An Illustrative Example

The proposed frontier-based classifiers are applied to an illustrative example where the characteristic variables are non-monotonic. For the sake of replication, we choose a secondary data set provided by Pendharkar (2011).

This data set is characterized by two characteristic variables. It is generated from simulations of the normal distributions with means -1, 0 and 1. The standard deviations for all distributions equal one. The examples that were generated from normal distributions with means of -1 and 1 are labelled as belonging to group 1, and the examples that were generated from normal distributions with means of 0 are labelled as belonging to group 2. In total 60 observations are generated and 30 each belong to either the training or the test sample. For both the training and the test sample, 20 observations belong to group 1 and the other 10 are from group 2. The detailed observations and their group labels are reported in Table 1 in Pendharkar (2011).

Figure 3.5: A plot of the training data set

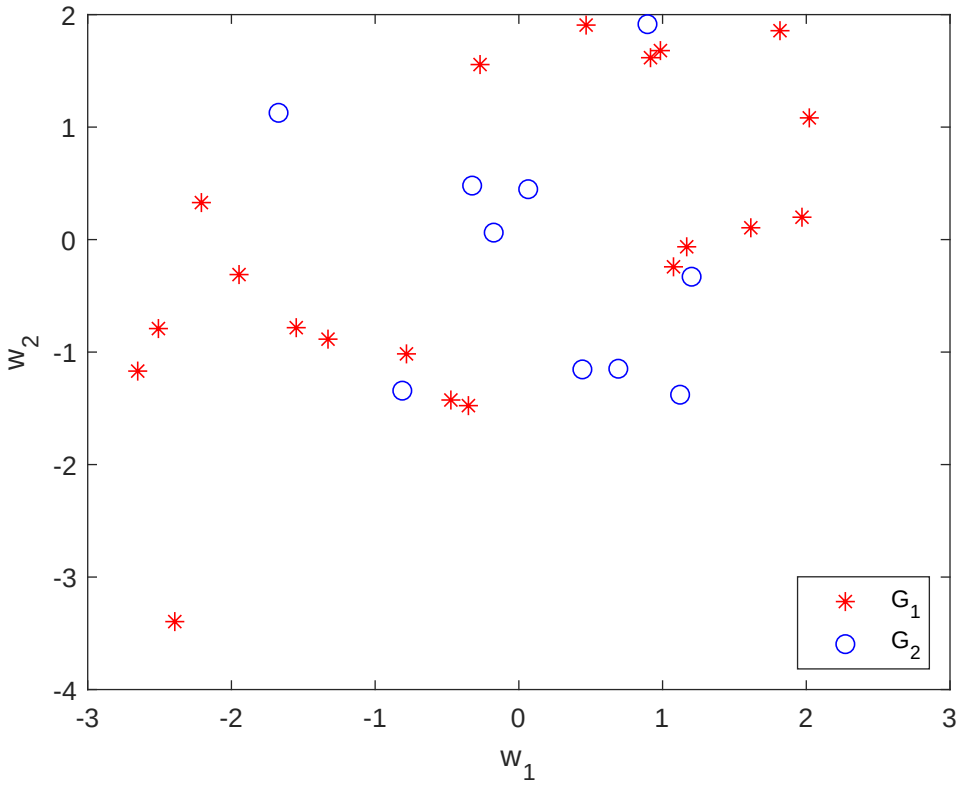


Figure 3.5 illustrates the plot of the training observations. The asterisks stands

for the training observations from the bad group G_1, and circles are the training observations from the good group G_2. It can be seen that the characteristic variables of the simulated data is non-monotonically related with the group labels. Neither a too big value nor a too small value is favored by the observations from G_2. In addition, it is worth noting that this simulated data set contains observations with negative characteristic values and two groups are not linearly separable. In Pendharkar (2011), a radial basis function is used to map a higher dimensional space where the negative characteristic values are converted into non-negative ones and the non-linearly separable data is transformed into a linearly separable data. Specifically, the two-dimensional data set is transformed into a three-dimensional one. However, there is no need for pretreatment of the original data set while applying our frontier-based classifier.

The performance of applying the proposed frontier-based classifiers is compared to that of applying the hybrid radial basis function network-data envelopment analysis (RBFN-DEA) neural network listed in Pendharkar (2011). Depending on the relative misclassification costs of Type I or II errors, three RBFN-DEA methods are examined. Specifically, these three methods are called the no Type II error DEA model (NTIIEM), the no Type I error DEA model (NTIEM) and the nearest neighborhood DEA approach (NNA).

To compare the performance on the test sample, a series of measures based on the confusion matrices are calculated for both the proposed frontier-based methods and the methods displayed in Pendharkar (2011). The confusion matrix is defined in Table 3.1. In this example, the observations from G_2 are considered to be the positive ones while those from G_1 are the negative ones. The cell marked by true negative (TN) records the number of correctly predicted observations from G_1. The cell marked by false positive (FP) records the number of the observations

which are from G_1 but are predicted to be in G_2. The cell marked by false negative (FN) records the number of the observations which are from G_2 but are predicted to be in G_1. The cell marked by true positive (TP) records the number of correctly predicted observations from G_2.

Table 3.1: A confusion matrix

	Predicted Group = 1	Predicted Group = 2
Actual Group = 1	TN	FP
Actual Group = 2	FN	TP

Table 3.2 summarizes the prediction performance from applying the five listed classifiers. Among these methods, the first three are from Pendharkar (2011) and the other two are proposed in this contribution. The first four columns report the number of test observations under each situation. The overall accuracy is defined by (TP+TN)/(TP+TN+FP+FN); it illustrates the percentage of correctly predicted observations regardless of the group. The precision is defined by calculating TP/(TP+FP); this gives the proportion of the observations that are classified into the good group G_2 and are actually from G_2. High precision relates to the low false positive rate. The recall is defined by calculating TP/(TP+FN); it gives the proportion of the observations that are actually from the good group G_2 and are correctly classified. High recall relates to the low false negative rate. The value of F1 score is calculated by $2 \times \text{precision} \times \text{recall} \div (\text{precision} + \text{recall})$; this F1 score takes both false positives and false negatives into account. The values of the last four columns are reported in percentages. A high percentage indicates a good performance.

The comparison of the five listed classifiers in Table 3.2 shows that the NC frontier-based classifier outperforms the other four classifiers. It has the highest overall accuracy, the highest precision and also the highest F1 measure. In addi-

Table 3.2: A summary of the prediction performance on the test sample for all listed classifiers

	TP	TN	FP	FN	Accuracy	Precision	Recall	F1
NTIIEM	10	3	17	0	43.33	37.04	100.00	54.05
NTIEM	4	16	4	6	66.67	50.00	40.00	44.44
NNA	7	13	7	3	66.67	50.00	70.00	58.33
Convex Hull	9	8	12	1	56.67	42.86	90.00	58.06
Nonconvex Hull	8	14	6	2	73.33	57.14	80.00	66.67

tion, its percentage of the recall measure is not bad. The highest recall percentage of 100% is achieved by applying the NTIIEM. However, the results of the NTIIEM show that although all positive observations are correctly predicted, most of the negative observations are misclassified to be positive. If there is a skewed emphasis on correctly predicting the positive cases, then this method could be a potential choice. Other than that, the NC frontier-based classifier is the best choice based on the test data.

In addition, the classification accuracy of our frontier-based classifiers is compared with that of the well known architectures of other neural networks used to solve the classification problems. Among these neural networks are feed forward neural network (FFNN) using error backpropagation learning algorithm (Rumellhart (1986)), and a probabilistic neural network (PNN) proposed by Specht (1990). The detailed experimental setting of these two methods can be traced in Pendharkar (2011).

The best accuracy that the listed neural network method could achieve is 66.67 % while applying the NTIEM or the NNA. By applying the NC frontier-based classifier, the accuracy is improved by 6.66% (73.33 %-66.67 %). The results in Table 3.3 indicate that the NC frontier-based classifer performs well compared to the listed neural network models. Moreover, the comparison between the results of

Table 3.3: Accuracy results of different methods on the test sample

	Accuracy
NTIIEM	43.33
NTIEM	66.67
NNA	66.67
FFNN	63.66
PNN	43.44
Convex Hull	56.67
Nonconvex Hull	73.33

the C and NC frontier-based classifiers implies that there is seemingly no substitution relation existing in this example. This confirms that if there is no prior information on the substitution relation among the characteristic variables, then the NC frontier-based classifier is a conservative and a better choice than the C frontier-based classifier.

3.5 Conclusions

While the background information on the relation between the characteristic variables and the group label often suggests the consideration of a monotonic relation, its parallel problem of considering a non-monotonic relation is rarely taken into account in classification. We consider the classification problem in a more general formulation where both monotonic and non-monotonic relations are incorporated. Different from the standard disposal assumption used for describing a monotonic relation, a generalized disposal assumption which limits the disposability within a value range is defined for characterizing the non-monotonic relation. Accordingly, a dominance adapting directional distance (DAD) function which accommodates the generalized disposability notion is developed for measuring the distance of an

observation to the corresponding NC separating frontier. A NC separating hull consists of these NC separating frontiers is used to predict the membership of a new observation. If a new observation is located within this NC separating hull, then it belongs to the good group, otherwise it belongs to the bad group. We design an algorithm to simplify the procedures of predicting the membership of a new observation.

We also analyze our nonparametric classifier in a commonly used C setting. We argue that only if there exists the additional background information on the substitution relation among the characteristic variables, then a C nonparametric classifier is preferred. Otherwise, a NC classifier is more conservative and provides better classification performance than the C classifier.

We have applied the proposed nonparametric classifiers to a nonlinearly separable binary classification problem. The NC classifier outperforms the C classifier in terms of the overall accuracy, the precision and the F1 measure. It confirms our argument of applying a C classifier only after detecting a substitution relation. Moreover, the proposed NC classifier is shown to outperform some existing DEA-based classifiers in terms of several commonly used criteria.

We end with developing some perspectives for potential future research. First, the non-monotonic relation is also frequently encountered in performance evaluation. For instance, an increase in age is efficiency-improving for relatively young farmers, but is efficiency-impeding for relatively senior farmers. The effect of considering this type of non-monotonic relation is explored by Wang (2002) in a parametric setting. A straightforward extension is to modify the efficiency measure defined in this contribution so that the non-monotonic relation can be examined in a nonparametric setting. Second, the current separating hull classifier can be extended to handle classification problems with multiple groups.

References

Banker, R. D., H. S. Chang, and W. W. Cooper (2002): ""Small Sample Properties of ML, COLS and DEA Estimators of Frontier Models in the Presence of Heteroscedasticity" by AN Bojanic, SB Caudill and JM Ford, European Journal of Operational Research 108, 1998, 140–148: A Comment," *European Journal of Operational Research*, 136(2), 466–467.

Banker, R. D., A. Charnes, and W. W. Cooper (1984): "Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science*, 30(9), 1078–1092.

Briec, W. (1997): "A Graph-Type Extension of Farrell Technical Efficiency Measure," *Journal of Productivity Analysis*, 8(1), 95–110.

Briec, W., K. Kerstens, and I. Van de Woestyne (2016): "Congestion in Production Correspondences," *Journal of Economics*, 119(1), 65–90.

——— (2018): "Hypercongestion in Production Correspondences: An Empirical Exploration," *Applied Economics*, 50(27), 2938–2956.

Cano, J.-R., P. A. Gutiérrez, B. Krawczyk, M. Woźniak, and S. García (2019): "Monotonic Classification: An Overview on Algorithms, Performance Measures and Data Sets," *Neurocomputing*, 341, 168–182.

Chambers, R., Y. Chung, and R. Färe (1998): "Profit, Directional Distance Functions, and Nerlovian Efficiency," *Journal of Optimization Theory and Applications*, 98(2), 351–364.

Cherchye, L., T. Kuosmanen, and T. Post (2001): "FDH Directional Distance Functions with An Application to European Commercial Banks," *Journal of Productivity Analysis*, 15(3), 201–215.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring Labor Efficiency in Post Offices," in *The Performance of Public Enterprises: Concepts and Measurements*, ed. by M. Marchand, P. Pestieau, and H. Tulkens, pp. 243–268. North Holland, Amsterdam.

Emrouznejad, A., and G.-l. Yang (2018): "A Survey and Analysis of the First 40 Years of Scholarly Literature in DEA: 1978–2016," *Socio-Economic Planning Sciences*, 61, 4–8.

Feelders, A., and M. Pardoel (2003): "Pruning for Monotone Classification Trees," in *Advances in Intelligent Data Analysis V*, ed. by M. R. Berthold, H.-J. Lenz, E. Bradley, R. Kruse, and C. Borgelt, pp. 1–12, Berlin, Heidelberg. Springer Berlin Heidelberg.

Freed, N., and F. Glover (1986): "Evaluating Alternative Linear Programming Models to Solve the Two-Group Discriminant Problem," *Decision Sciences*, 17(2), 151–162.

Kerstens, K., and I. Van de Woestyne (2011): "Negative Data in DEA: A Simple Proportional Distance Function Approach," *Journal of the Operational Research Society*, 62(7), 1413–1419.

Lam, K. F., and E. U. Choo (1993): "A Linear Goal Programming Model for Classification with Non-Monotone Attributes," *Computers & Operations Research*, 20(4), 403–408.

Lam, K. F., E. U. Choo, and J. W. Moy (1996): "Minimizing Deviations From the Group Mean: A New Linear Programming Approach for the Two-Group

Classification Problem," *European Journal of Operational Research*, 88(2), 358–367.

Leon, C. F., and F. Palacios (2009): "Evaluation of Rejected Cases in an Acceptance System with Data Envelopment Analysis and Goal Programming," *Journal of the Operational Research Society*, 60(10), 1411–1420.

Pendharkar, P., M. Khosrowpour, and J. Rodger (2000): "Application of Bayesian Network Classifiers and Data Envelopment Analysis for Mining Breast Cancer Patterns," *Journal of Computer Information Systems*, 40(4), 127–132.

Pendharkar, P., J. Rodger, and G. Yaverbaum (1999): "Association, Statistical, Mathematical and Neural Approaches for Mining Breast Cancer Patterns," *Expert Systems with Applications*, 17(3), 223–232.

Pendharkar, P. C. (2002): "A Potential Use of Data Envelopment Analysis for the Inverse Classification Problem," *Omega*, 30(3), 243–248.

Pendharkar, P. C. (2011): "A Hybrid Radial Basis Function and Data Envelopment Analysis Neural Network for Classification," *Computers & Operations Research*, 38(1), 256–266.

Rademaker, M., B. De Baets, and H. De Meyer (2009): "Loss Optimal Monotone Relabeling of Noisy Multi-Criteria Data Sets," *Information Sciences*, 179(24), 4089–4096.

Rumellhart, D. E. (1986): "Learning Internal Representations by Error Propagation," *Parallel distributed processing*, 1, 318–362.

Seiford, L., and J. Zhu (1998): "An Acceptance System Decision Rule with Data Envelopment Analysis," *Computers & Operations Research*, 25(4), 329–332.

Silva, A. P. D., and A. Stam (1994): "Second Order Mathematical Programming Formulations for Discriminant Analysis," *European Journal of Operational Research*, 72(1), 4–22.

Smaoui, S., H. Chabchoub, and B. Aouni (2009): "Mathematical Programming Approaches to Classification Problems," *Advances in Operations Research*, 2009, Art. ID 252989.

Specht, D. F. (1990): "Probabilistic Neural Networks," *Neural Networks*, 3(1), 109–118.

Sueyoshi, T. (1999): "DEA-Discriminant Analysis in the View of Goal Programming," *European Journal of Operational Research*, 115(3), 564–582.

——— (2001): "Extended DEA-Discriminant Analysis," *European Journal of Operational Research*, 131(2), 324–351.

——— (2004): "Mixed Integer Programming Approach of Extended DEA–Discriminant Analysis," *European Journal of Operational Research*, 152(1), 45–55.

Troutt, M., A. Rai, and A. Zhang (1996): "The Potential Use of DEA for Credit Applicant Acceptance Systems," *Computers & Operations Research*, 23(4), 405–408.

Wang, H.-J. (2002): "Heteroscedasticity and Non-monotonic Efficiency Effects of a Stochastic Frontier Model," *Journal of Productivity Analysis*, 18(3), 241–253.

Yan, H., and Q. Wei (2011): "Data Envelopment Analysis Classification Machine," *Information Sciences*, 181(22), 5029–5041.

CHAPTER

4

Ordinal Classification with Double Nonparametric Frontiers: The Role of Nonconvexity and Misclassification Costs

4.1 Introduction

A classification aims to determine whether an observation belongs to a particular group by evaluating a set of characteristic values. It is known as an ordinal classification problem when the groups are ordered. As an important and widely studied topic, its applications includes but are not limited to costumer churn (e.g., De Caigny, Coussement, De Bock, and Lessmann (2019)), bankruptcy prediction (e.g., De Bock (2017)), credit scoring (e.g., Lessmann, Baesens, Seow, and Thomas (2015)), etc. Numerous techniques and methods have been proposed, such as statistical methods, support vector machines, artificial neural networks, decision trees and ensemble classifiers. A comprehensive review of statistical and data mining

techniques used for classification can be found in Kotsiantis, Zaharakis, and Pintelas (2007).

Fisher's linear discriminant function (Fisher (1936)) and Smith's quadratic discriminant function (Smith (1946)) are popular statistical approaches to solve the classification problem under the assumption of multivariate normality and variance-covariance homogeneity. In case the above assumptions are violated, the performance of mathematical programming (MP) methods has been proven superior to the former methods for classification purposes in many studies (e.g., Bajgier and Hill (1982), Freed and Glover (1986), Stam and Ragsdale (1992), Silva and Stam (1994), Smaoui, Chabchoub, and Aouni (2009), etc.). In the MP classifiers, one or several hypersurfaces which bound the groups of observations are used to separate two groups of observations. In most applications, the nonlinear hypersurface which bounds the observations tighter provides a better separation than the linear hypersurface does. However, the nonlinear MP classifier requires to specify an assumption on the nonlinear functional form which eventually generates the separating hypersurface. It is not impossible, but very difficult to prescribe such a nonlinear function to fit for a real application. In this sense, a nonparametric classifier which provides a data-based piecewise linear frontier may well receive increasing attention since no assumption on the frontier shape is required.

In previous nonparametric studies on classification, there are typically two types of classifiers which are related to the Data Envelopment Analysis (DEA) models. The first type is essentially based on the use of goal programming (see Sueyoshi (1999, 2001, 2004) for the details) rather than the standard DEA models. It is known as the Data Envelopment Analysis-Discriminant Analysis (DEA-DA) method and has been well developed by Sueyoshi (2006), Jahanshahloo, Lotfi, Balf, and Rezai (2007), Lotfi and Mansouri (2008), Sueyoshi and Goto (2010), etc.

Banker, Chang, and Cooper (2002) have argued researchers to avoid calling the goal programming models standing in for the DEA methods. The second type is based on the traditional DEA models proposed by Charnes, Cooper, and Rhodes (1978) which is originally proposed for ranking a set of observations. The DEA-based classifier discussed in the following explicitly refers to the second type which generates a piece-wise linear frontier.

Troutt, Rai, and Zhang (1996) first propose to use the DEA frontier as an acceptability frontier in credit applicant acceptance systems. The DEA frontier provides a convex (C) envelopment of the given observations. Without pre-specifying the exact shape of a separating hypersurface, the DEA frontier is piece-wise linear and bounds the observations closely. Ever since the first application of DEA methods in classification proposed by Troutt, Rai, and Zhang (1996), the idea of employing the C DEA frontier as a separating frontier has been well adapted by proposing alternative objective functions (Seiford and Zhu (1998)), incorporating various data types (e.g., Leon and Palacios (2009), Yan and Wei (2011)) and has been applied in different application areas (e.g., Seiford and Zhu (1998), Pendharkar (2002), Pendharkar, Rodger, and Yaverbaum (1999); Pendharkar, Khosrowpour, and Rodger (2000))). In the above methods, there is only one single separating frontier trained from a certain group of observations and then used to differentiate between two groups of observations. It works well when the two groups of observations can be clearly separated.

However, groups of observations in most applications are found to have overlaps. An observation located in the overlap indicates that there is no clear cut way to determine the group in which the observation should be classified. In order to capture the overlap which is the main source of misclassifications, the idea of using double separating frontiers is proposed. In Chang and Kuo (2008),

a pair of DEA frontiers is constructed to correspondingly envelop two groups of observations. While the two frontiers each describes a set of observations, their intersection is known as the data-based overlap area. After explicitly defining the overlap, the question follows is to determine the membership of the observations that are located in the overlap.

One potential way is to remove the overlap as early as in the training process. This is achieved by using the stratified DEA method proposed by Zhu (2003) where the separating frontiers could be shifted inwards layer by layer. In the paper of Chang and Kuo (2008), the overlap is completely eliminated by removing the same number of layers from both groups. In order to account for the uneven misclassification costs, an asymmetric-stratified DEA method is proposed in Kuo (2013) so that the number of layers removed from two groups could be different and is determined by achieving the minimal total misclassification cost. In this way, the overlap is completely eliminated. However, the cost is to potentially misclassify the training observations located in the area of overlap.

Another alternative choice is to report the overlap as it is in the training process. In the prediction process, further discriminant rules are designed to classify the test observations located in the overlap. In this case, there is no misclassification allowed in the training process, which is the opposite extreme of the first case where the overlap is completely eliminated. Furthermore, none of the existing research directly uses their radial efficiency results to determine the membership of the observations located in the overlap. After defining the overlap, the membership is always decided by incorporating other methods, e.g., membership functions (Pendharkar (2012)) or interaction or MSD method (Pendharkar and Troutt (2014)). The asymmetric misclassification costs are also incorporated in designing the discriminant rules, e.g., using a cost-sensitive nearest neighbourhood

approach (Pendharkar (2011)), using probabilistic DEA techniques (Pendharkar (2018)), among others.

Regardless of various treatments on the overlap, one common thing in the frontier classification literature is assuming that the separating frontier is convex. To the best of our knowledge, none of the current research has ever questioned and left out the convexity assumption. The only exception is that when analysing the superior performance of neural networks over convex frontiers in mining breast cancer patterns, Pendharkar, Rodger, and Yaverbaum (1999, p. 231) claim that one of the reasons could be that the frontier method assumes the convexity of acceptable cases, while neural networks relax this assumption. The assumption of convexity is commonly kept in production analysis since it is common in the economic theory. When it comes to the classification problem, the assumption on convexity is accepted without arguing its correspondence with related background knowledge in classification.

To address the above shortcomings, the overarching objective of this study is to propose a novel nonparametric frontier-based classifier which aims at achieving the minimal misclassification cost. It is developed based on two design goals that overcome the shortcomings of the existing approaches. First, we intend to explore the connections between the axioms used in the nonparametric analysis and the background information in ordinal classification. Although the initial inspiration of applying the DEA-based C frontier is that it provides a tight envelopment without any assumption on the shape of the frontier, it is now necessary to provide some theoretical basis so that the commonly used assumptions could be relaxed depending on the applications. A second design objective is to develop a nonparametric frontier-based classifier which is cost-sensitive (CS) and inherently designed to minimize the total misclassification cost.

To this end, we present a new methodological framework for building the CS frontier-based classifier that accommodates asymmetric misclassifications costs under various mixes of prior-known background information. First, depending on the prior-known background information, the nonparametric frontier-based classifier is capable of generating either C or nonconvex (NC) separating frontiers. The background information of a monotonic relation between the characteristic variables and the group membership corresponds to the axiom of free disposal. This gives rise to the NC frontiers generated from the Free Disposal Hull (FDH) approach (Deprins, Simar, and Tulkens (1984)). The background information on the substitution relation between the characteristic variables corresponds to the axiom of convexity. Only if there are certainties about both the monotonicity and the substitutability, it is then reasonable to apply the C frontiers generated from the DEA-based classifier. Second, apart from exploring the theoretical basis for constructing the frontiers, the classifier is designed to be CS. The cost information is involved both during the training process and in designing the discriminant rules. Instead of eliminating the overlap completely in the training process, the overlap is minimized to the extent that the total misclassification cost is minimized. In addition, the overlap is reduced by excluding the observations point by point rather than removing the complete layer of frontier observations. In this case, partial overlap is allowed if the additional shift could not reduce the total misclassification cost any more. When it comes to the predicting process, the discriminant rules are designed to incorporate the cost information as well. Moreover, it is shown that the choice of the direction of the directional distance function (DDF) matters when predicting the observations in the overlap. To illustrate the proposed framework, a graduate admission example is used for graphically showing the classification results.

This contribution is structured as follows. In Section 4.2, the groups of obser-

vations are characterized by the acceptance possibility sets (APSs) based on the axioms corresponding to the background information. Both the NC and C APSs are constructed. Rather than focusing on a single type of data like most papers do, both input-type and output-type characteristic variables are incorporated in our classifier. The constructions of the envelopment frontiers which bound the corresponding APSs and the separating frontiers after shifting inwards are introduced in Section 4.3. An algorithm is designed to shift the frontiers point by point so that the total misclassification is minimized. A graduate admission example is illustrated to show the differences between the C and NC frontiers as well as showing the results of the shifting algorithm. In Section 4.4, the DDF measure based discriminant rules which incorporates the asymmetric misclassification costs are introduced. The drawbacks of the commonly used radial measure are illustrated with the graduate admission example. Finally, in Section 4.5 this contribution is concluded with a summary of its achievements and a discussion of potential future research topics.

4.2 Acceptance Possibility Set

Consider a binary classification problem with a set of training observations which are characterized by some characteristic variables. The training observations are exhaustively classified into two groups based on the prior information on their memberships. Meanwhile, those characteristic variables are expected to fully grasp the property that could differentiate the observations from one group to another. By learning from these two groups of training observations, a classifier is trained and should be able to predict the membership of a test observation where the data of the same characteristic variables is collected.

For classification problems like differentiating between bankruptcy and non-bankruptcy firms, or diagnosing patients from healthy people, there exists a natural order between two groups. The naturally favored group, e.g., the group of non-bankruptcy firms, is known as the good group, relatively the bankruptcy firms which are unfavored belong to the bad group. In this contribution, the training observations from the bad group constitute the bad training sample set which is denoted by G_1. Correspondingly, the good training sample set that consists of the training observations from the good group is denoted by G_2. Note that $G_1 \cap G_2 = \emptyset$.

Except the prior information on the group membership, another common background information is about the monotonic relation of the characteristic variables. That is, the K characteristic variables which characterize the observations are differentiated into two types depending on their monotonic relation is increasing or decreasing. Specifically, if the possibility of belonging to the good group increases (decreases) with the abatement (augment) of a characteristic variable, then it is known as a variable with the monotonic decreasing relation. The set of these monotonically decreasing variables is denoted by $X \in \mathbb{R}^m$. If the possibility of belonging to the good group increases (decreases) with the augment (abatement) of a characteristic variable, then it is defined as a variable with the monotonic increasing relation. The set of these monotonically increasing variables is denoted by $Y \in \mathbb{R}^s$. To sum up, an observation $Z \in \mathbb{R}^K$ is explicitly characterized by $Z = (X, Y) \in \mathbb{R}^{m+s}$.

While applying a nonparametric frontier-based method, an acceptance possibility set (APS) is introduced to describe the property of a certain group. It is a concept derived from the production possibility set (PPS) which is well-known in production analysis. A PPS contains all combinations of resources and products

that are producible under certain technology. Correspondingly, an APS consist of all combinations of characteristic values that corresponding observations are presumed to belong to a certain group. Both the APS of the bad group and that of the good group are constructed based on their training observations and some axioms.

We start with constructing the APS of the bad group based on the n_1 training observations from G_1 . The background information on the monotonic relation corresponds to a frequently used dominance assumption on the training observations from G_1. The monotonic relation states that the possibility of belonging to the bad group increases with the augment of X and the abatement of Y. That is, if an observation with more X and less Y than a training observation from G_1, it is then being dominated and believed to belong to the bad group. For a training observation $Z_j = (X_j, Y_j) \in G_1$, a free disposal set denoted by $T_{j,1}$ could then be represented by $T_{j,1} = \{(X, Y) \in \mathbb{R}^{m+s} \mid X \geq X_j \text{ and } Y \leq Y_j\}$. The union of all the free disposal sets of the training observations from G_1 constitutes a nonconvex (NC) APS denoted by $T_{NC,1}$. Specifically, $T_{NC,1}$ depicts the observations belonging to the bad group as follows:

$$\begin{aligned} T_{NC,1} &= \bigcup_{j=1}^{n_1} T_{j,1} \\ &= \left\{ (X, Y) \in \mathbb{R}^{m+s} \mid \sum_{j=1}^{n_1} \lambda_j X_j \leq X, \sum_{j=1}^{n_1} \lambda_j Y_j \geq Y, \sum_{j=1}^{n_1} \lambda_j = 1, \lambda_j \in \{0, 1\} \right\}. \end{aligned} \tag{4.2.1}$$

Similarly, the APS of the good group could be constructed from the n_2 training observations from G_2. However, the same background knowledge on the monotonic relation indicates an opposite dominance relation comparing to that defined in G_1. The monotonic relation states that the possibility of belonging to the good group increases with the abatement of X and the augment of Y. That is, a training

observation from G_2 remains in the good group if it starts decreasing its X and increasing Y, since the observation after the change is being dominated. That is, the corresponding free disposal set denoted by $T_{j,2}$ for the training observation $Z_j = (X_j, Y_j) \in G_2$ is represented by $T_{j,2} = \{(X, Y) \in \mathbb{R}^{m+s} \mid X \leq X_j \text{ and } Y \geq Y_j\}$. The union of all these free disposal sets derived from the training observations from G_2 constitutes a NC APS denoted by $T_{NC,2}$. Specifically, $T_{NC,2}$ depicts the observations belonging to the good group as follows:

$$\begin{aligned} T_{NC,2} &= \bigcup_{j=1}^{n_2} T_{j,2} \\ &= \left\{ (X, Y) \in \mathbb{R}^{m+s} \mid \sum_{j=1}^{n_2} \lambda_j X_j \geq X, \sum_{j=1}^{n_2} \lambda_j Y_j \leq Y, \sum_{j=1}^{n_2} \lambda_j = 1, \lambda_j \in \{0, 1\} \right\}. \end{aligned} \tag{4.2.2}$$

Comparing to the set $T_{NC,1}$ in (4.2.1), the inequity symbols are reversed in set $T_{NC,2}$ in (4.2.2). That is, the same monotonic relation of the characteristic variables generates opposite axioms on free disposal for the bad group and the good group. In this sense, the bad group is bounded by the best performed training observations, while the good group is bounded by the worst performed training observations.

When it comes to the axiom on convexity, it does not make a difference between two groups. The convexity in classification implies a substitution relation among the characteristic variables. Note that a convex (C) APS is only preferred if the prior information of such a substitution relation is provided. By adding this additional axiom, the above NC APSs are transformed into the following C ones :

$$T_{C,1} = \left\{ (X, Y) \in \mathbb{R}^{m+s} \mid \sum_{j=1}^{n_1} \lambda_j X_j \leq X, \sum_{j=1}^{n_1} \lambda_j Y_j \geq Y, \sum_{j=1}^{n_1} \lambda_j = 1, \lambda_j \geq 0 \right\}. \tag{4.2.3}$$

$$T_{C,2} = \left\{ (X,Y) \in \mathbb{R}^{m+s} \mid \sum_{j=1}^{n_2} \lambda_j X_j \geq X, \sum_{j=1}^{n_2} \lambda_j Y_j \leq Y, \sum_{j=1}^{n_2} \lambda_j = 1, \lambda_j \geq 0 \right\}. \tag{4.2.4}$$

In order to simplify the expressions, we use the following notation to stand for the APS of the bad group and the APS of the good group under both the NC and C cases:

$$T_{\Lambda,1} = \left\{ (X,Y) \in \mathbb{R}^m \times \mathbb{R}^s \mid \sum_{j=1}^{n_1} \lambda_j X_j \leq X, \sum_{j=1}^{n_1} \lambda_j Y_j \geq Y, \right. \\ \left. \sum_{j=1}^{n_1} \lambda_j = 1, \lambda_j \in \Lambda, j = 1, \ldots, n_1 \right\}, \tag{4.2.5}$$

$$T_{\Lambda,2} = \left\{ (X,Y) \in \mathbb{R}^m \times \mathbb{R}^s \mid \sum_{j=1}^{n_2} \lambda_j X_j \geq X, \sum_{j=1}^{n_2} \lambda_j Y_j \leq Y, \right. \\ \left. \sum_{j=1}^{n_2} \lambda_j = 1, \lambda_j \in \Lambda, j = 1, \ldots, n_2 \right\}, \tag{4.2.6}$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_j \geq 0\}, \quad \text{or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_j \in \{0,1\}\}.$$

Following Chambers, Chung, and Färe (1998), $T_{\Lambda,1}$ and $T_{\Lambda,2}$ can be represented using the following directional distance function (DDF), respectively:

$$D_{\Lambda,g_1}(Z) = \sup\{\delta \in \mathbb{R} \mid Z + \delta g_1 \in T_{\Lambda,1}\}. \tag{4.2.7}$$

$$D_{\Lambda,g_2}(Z) = \sup\{\delta \in \mathbb{R} \mid Z + \delta g_2 \in T_{\Lambda,2}\}. \tag{4.2.8}$$

where $g_1 = (g_{X,1}, g_{Y,1})$ and $g_2 = (g_{X,2}, g_{Y,2})$ represents the projection directions of

the bad group and that of the good group, respectively.

The value of the DDF measure illustrates that a training observation from G_1 (G_2) remains to belong to the bad group (the good group) after it changes its characteristic variables along the direction of g_1 (g_2) by δ. Obviously, the projection direction of the group is in accordance with its own dominance relation defined. To be meaningful, $g_{x_i,1} < 0$ for all $i \in [m]$ and $g_{y_r,1} > 0$ for all $r \in [s]$, where $[m]$ denotes the set $\{1, \ldots, m\}$ and $[s]$ denotes the set $\{1, \ldots, s\}$. In this way, the characteristic variables X are reduced and the characteristic variables Y are increased while increasing the value of δ, which is the favorable behavior for characterizing the bad group. On the contrary, $g_{x_i,2} > 0$ for all $i \in [m]$ and $g_{y_r,2} < 0$ for all $r \in [s]$ which is the favorable behavior for characterizing the good group.

The assumption on convexity differentiates the NC APS from the C one. However, this does not change the definition of the DDF measure, only the value of the DDF measure may be enlarged. That is, $D_{\Lambda^{\mathrm{NC}},1}(Z) \leq D_{\Lambda^{\mathrm{C}},1}(Z)$. Likewise, $D_{\Lambda^{\mathrm{NC}},2}(Z) \leq D_{\Lambda^{\mathrm{C}},2}(Z)$. The DDF measure serves as an indicator that positions the observations relative to the boundary of corresponding APS. It is well-defined for all possible observations $Z = (X, Y) \in \mathbb{R}^m \times \mathbb{R}^s$. A non-negative DDF measure means the observation Z is in the interior of the APS. If the observation Z is located beyond the APS, then its DDF measure becomes negative and it is projected onto the frontier in the direction opposite to its defined g_1 or g_2.

4.3 Models for Generating Double Separating Frontiers

4.3.1 Nonparametric Models for Generating the Envelopment Frontiers

In the C frontier classification literature, a pair of piece-wise hypersurfaces are generated to best separate the observations. These hypersuefaces are the envelopment frontiers of two groups of training observations. In this subsection, the construction of both the C and NC envelopment frontiers are introduced.

With the projecting direction $g_1 \in \mathbb{R}^m_- \times \mathbb{R}^s_+$, the following model is used to measure the distance of the observation $Z_0 = (X_0, Y_0)$ to the boundary of $T_{\Lambda,1}$ which depicts the bad group:

$$\begin{aligned}
\max_{\lambda_{j,1},\hat{\delta}_{\Lambda,1}} \quad & \hat{\delta}_{\Lambda,1} \\
s.t. \quad & \sum_{j=1}^{n_1} \lambda_{j,1} x_{i,j} \leq x_{i,0} + \hat{\delta}_{\Lambda,1} g_{x_i,1} && \forall i \in [m] \\
& \sum_{j=1}^{n_1} \lambda_{j,1} y_{r,j} \geq y_{r,0} + \hat{\delta}_{\Lambda,1} g_{y_r,1} && \forall r \in [s] \\
& \sum_{j=1}^{n_1} \lambda_{j,1} = 1 \\
& \lambda_{j,1} \in \Lambda && \forall j \in [n_1]
\end{aligned} \tag{4.3.1}$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_{j,1} \geq 0\}, \text{ or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_{j,1} \in \{0,1\}\}.$$

In the C case, model (4.3.1) is a linear programming (LP) problem, while it involves solving a binary mixed integer program (BMIP) for the NC case. To remedy the computational issue in the NC case, a fast implicit enumeration-based method is proposed by Cherchye, Kuosmanen, and Post (2001) requiring only to compute minima of lists of ratios. Instead of solving a BMIP model, the following exact solution is obtained for model (4.3.1) under the NC case:

$$\hat{\delta}^*_{\Lambda^{\mathrm{NC}},1} = \max_{j=1,\ldots,n_1} \left(\min_{i=1,\ldots,m} \left(\frac{x_{i,j} - x_{i,0}}{g_{x_i,1}} \right), \min_{r=1,\ldots,s} \left(\frac{y_{r,j} - y_{r,0}}{g_{y_r,1}} \right) \right). \tag{4.3.2}$$

By solving model (4.3.1) for all the training observations from G_1, a frontier set defined by $\widehat{FS}_{\Lambda,1}$ is generated. Specifically, $\widehat{FS}_{\Lambda,1} = \{j \in G_1 | \hat{\delta}^*_{\Lambda,1} = 0\}$. Normally, the set $\widehat{FS}_{\Lambda,1}$ under the NC case is different from that under the C case. All frontier observations in $\widehat{FS}_{\Lambda^{\mathrm{C}},1}$ could be found in $\widehat{FS}_{\Lambda^{\mathrm{NC}},1}$. However, not all frontier observations in $\widehat{FS}_{\Lambda^{\mathrm{NC}},1}$ belong to $\widehat{FS}_{\Lambda^{\mathrm{C}},1}$, since some frontier observations generated under the NC case are dominated by some convex combinations of the training observations. Therefore, $\widehat{FS}_{\Lambda^{\mathrm{C}},1} \subsetneqq \widehat{FS}_{\Lambda^{\mathrm{NC}},1}$.

The training observations in this frontier set $\widehat{FS}_{\Lambda,1}$ dominate all other possible observations that belong to the bad group. The envelopment frontier formed by $\widehat{FS}_{\Lambda,1}$ bounds $T_{\Lambda,1}$. All training observations from G_1 are located within this envelopment frontier.

Similarly, with the projecting direction $g_2 \in \mathbb{R}^m_+ \times \mathbb{R}^s_-$, the following model is employed to measure the distance of the observation $Z_0 = (X_0, Y_0)$ to the boundary of $T_{\Lambda,2}$ which depicts the good group:

$$
\begin{aligned}
\max_{\lambda_{j,2},\hat{\delta}_{\Lambda,2}} \quad & \hat{\delta}_{\Lambda,2} \\
s.t. \quad & \sum_{j=1}^{n_2} \lambda_{j,2} x_{i,j} \geq x_{i,0} + \hat{\delta}_{\Lambda,2} g_{x_i,2} \qquad \forall i \in [m] \\
& \sum_{j=1}^{n_2} \lambda_{j,2} y_{r,j} \leq y_{r,0} + \hat{\delta}_{\Lambda,2} g_{y_r,2} \qquad \forall r \in [s] \\
& \sum_{j=1}^{n_2} \lambda_{j,2} = 1 \\
& \lambda_{j,2} \in \Lambda \qquad \forall j \in [n_2]
\end{aligned} \tag{4.3.3}
$$

where

$$
\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_{j,2} \geq 0\}, \text{ or (ii) } \Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_{j,2} \in \{0,1\}\}.
$$

Note that under the NC case, the solution for model (4.3.3) is as follows:

$$
\hat{\delta}^*_{\Lambda^{\mathrm{NC}},2} = \max_{j=1,\ldots,n_1} \left(\min_{i=1,\ldots,m} \left(\frac{x_{i,j} - x_{i,0}}{g_{x_i,2}} \right), \min_{r=1,\ldots,s} \left(\frac{y_{r,j} - y_{r,0}}{g_{y_r,2}} \right) \right). \tag{4.3.4}
$$

By solving model (4.3.3) for all the training observations from G_2, a frontier set defined by $\widehat{FS}_{\Lambda,2}$ is generated. That is, $\widehat{FS}_{\Lambda,2} = \{j \in G_2 | \hat{\delta}^*_{\Lambda,2} = 0\}$. Likewise, $\widehat{FS}_{\Lambda^{\mathrm{C}},2} \subseteqq \widehat{FS}_{\Lambda^{\mathrm{NC}},2}$. The training observations in $\widehat{FS}_{\Lambda,2}$ dominate all other possible observations that belong to the good group. Corresponding envelopment frontier bounds $T_{\Lambda,2}$. Therefore, all training observations from G_2 are located within this envelopment frontier.

In an ideal situation where observations are well characterized, these two envelopment frontiers are expected to completely separate two groups of training observations. That is, all training observations from G_2 should be located beyond the envelopment frontier 1. Similarly, all training observations from G_1 should be

located beyond the envelopment frontier 2.

However, in real applications, there often arises the situation where some observations are located within both envelopment frontiers. Put is differently, these observations are located in the intersection of two APSs which is $T_{\Lambda,1} \cap T_{\Lambda,2}$. By solving model (4.3.3) for the training observation $Z_j \in G_1$, if it has $\hat{\delta}^*_{\Lambda,2} \geq 0$, then it is a training observation located in the overlap. The set of the training observations which are from G_1 but are situated in the overlap is represented by $\widehat{O}_{\Lambda,1} = \{j \in G_1 | \hat{\delta}^*_{\Lambda,2} \geq 0\}$. Similarly, the training observation $Z_j \in G_2$ which has $\hat{\delta}^*_{\Lambda,1} \geq 0$ is also located in the overlap. The set of these training observations is represented by $\widehat{O}_{\Lambda,2} = \{j \in G_2 | \hat{\delta}^*_{\Lambda,1} \geq 0\}$. Note that $\widehat{O}_{\Lambda,1} \cup \widehat{O}_{\Lambda,2} \subset T_{\Lambda,1} \cap T_{\Lambda,2}$.

The larger the overlap is, normally the worse the classification ability of a classifier has. The NC APS is mathematically smaller than the C one, correspondingly its NC frontier provides a tighter envelopment of the training observations than the C one does. In this sense, the NC frontier is naturally perceived to have a better performance in separating two groups of observations.

4.3.2 Case Study and Double Envelopment Frontiers

In this subsection, we illustrate how the envelopment frontiers are constructed from the training observations by using a simple example. This illustrative example concerns the graduate business school admission decision-making from a large university in the Eastern US. For the ease of illustration and visualization, two characteristic variables are used. One is the standardized graduate management admission test (GMAT) which matches the property of a monotonically increasing variable. In order to have a representative of a monotonically decreasing variable, the other characteristic value is chosen to be the difference between 4 and the value

of the undergraduate grade point average (GPA). Table 4.1 gives the detailed data which is derived from the original admission data in Pendharkar (2012). The 13 rejected training observations from G_1 are used to construct the envelopment frontier that bounds $T_{\Lambda,1}$. G_2 which consists of 16 accepted training observations is used to generate the envelopment frontier that bounds $T_{\Lambda,2}$.

Table 4.1: The graduate admissions decision data

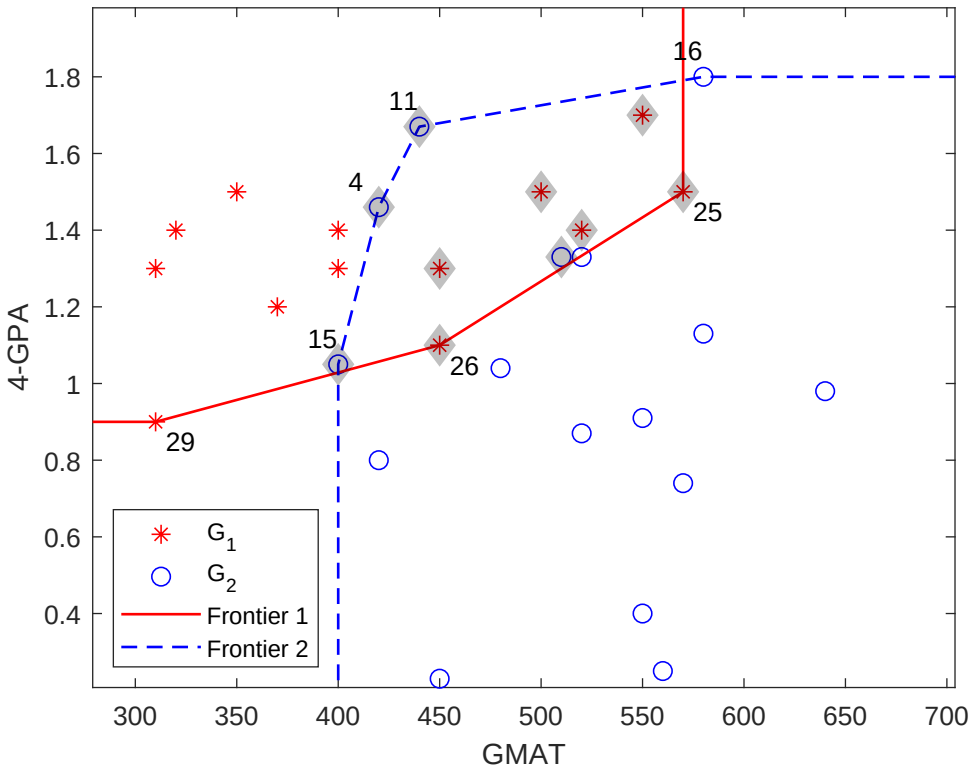
Obs. Number	GMAT	4-GPA	Decision	Obs. Number	GMAT	4-GPA	Decision
1	640	0.98	Accepted	17	310	1.3	Rejected
2	550	0.91	Accepted	18	350	1.5	Rejected
3	510	1.33	Accepted	19	400	1.3	Rejected
4	420	1.46	Accepted	20	370	1.2	Rejected
5	560	0.25	Accepted	21	450	1.3	Rejected
6	550	0.4	Accepted	22	500	1.5	Rejected
7	580	1.13	Accepted	23	520	1.4	Rejected
8	420	0.8	Accepted	24	550	1.7	Rejected
9	450	0.23	Accepted	25	570	1.5	Rejected
10	520	1.33	Accepted	26	450	1.1	Rejected
11	440	1.67	Accepted	27	320	1.4	Rejected
12	480	1.04	Accepted	28	400	1.4	Rejected
13	520	0.87	Accepted	29	310	0.9	Rejected
14	570	0.74	Accepted				
15	400	1.05	Accepted				
16	580	1.8	Accepted				

The frontier set is generated by solving model (4.3.1) for 13 rejected training observations. In the NC case, four training observations have the result of $\hat{\delta}^*_{\Lambda,1} = 0$. That is, $\widehat{FS}_{\Lambda^{\mathrm{NC}},1} = \{23, 25, 26, 29\}$. In the C case, observation 23 is dominated by the convex combination of observation 25 and observation 26. Therefore, the C frontier set $\widehat{FS}_{\Lambda^{\mathrm{C}},1}$ consists of only three training observations. That is, $\widehat{FS}_{\Lambda^{\mathrm{C}},1} = \{25, 26, 29\}$. Then, by solving model (4.3.3) for 16 accepted training observations, the frontier set $\widehat{FS}_{\Lambda,2}$ could be generated. Both the NC and C frontier sets consist

of four accepted training observations: namely 4,11,15 and 16.

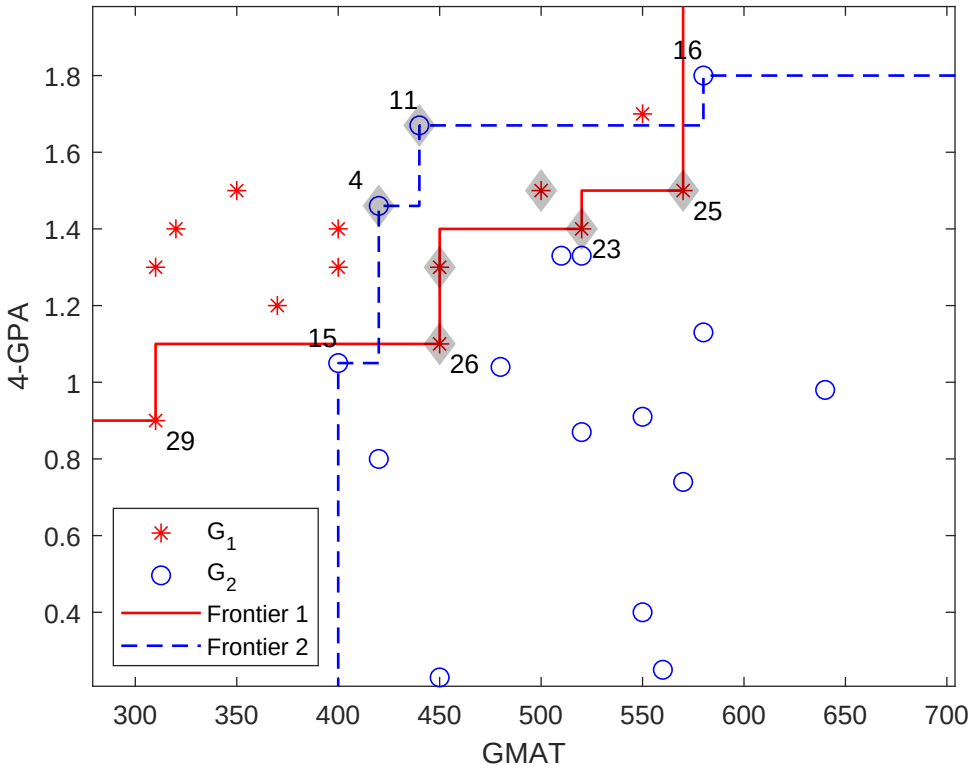
In Figure 4.1 and Figure 4.2, the envelopment frontiers formed by the derived frontier sets are displayed. The training observations from G_1 are marked with asterisks while those from G_2 are marked with circles. The envelopment frontier formed by $\widehat{FS}_{\Lambda,1}$ is labelled frontier 1 and marked by solid lines. And the envelopment frontier formed by $\widehat{FS}_{\Lambda,2}$ is labelled frontier 2 and marked by dashed lines. The training observations that are located in the overlap are marked with the faded rhombus. These observations are located both above the frontier 1 and below the frontier 2.

Figure 4.1: Double convex envelopment frontiers



In the C case which is displayed in Figure 4.1, there are totally 10 training

Figure 4.2: Double nonconvex envelopment frontiers



observations located in the overlap. Specifically, the numbers of observations from G_1 and G_2 that are located in the overlap are $n(\widehat{O}_{\Lambda^{\mathrm{C}},1}) = 6$ and $n(\widehat{O}_{\Lambda^{\mathrm{C}},2}) = 4$, respectively. Note that $n(\dots)$ represents the cardinality of a vector. While for the NC case in Figure 4.2, 7 training observations are located in the overlap, consisting of 2 originally accepted observations and 5 originally rejected ones. That is, $n(\widehat{O}_{\Lambda^{\mathrm{NC}},1}) = 5$ and $n(\widehat{O}_{\Lambda^{\mathrm{NC}},2}) = 2$. Obviously, less training observations are located in the overlap under the NC case than that under the C case.

4.3.3 Double Cost-Sensitive Separating Frontiers

Since the existence of overlap is doomed to cause ambiguity in classification, the researchers tried various methods to eliminate the overlap. One common way is by removing the first several layers of frontiers so that the intersection is completely removed (e.g., Chang and Kuo (2008), Kuo (2013)). This is realized by using the stratified DEA model proposed by Zhu (2003). Instead of removing a complete layer of frontier points, the above envelopment frontiers are proposed to be adjusted point by point so as to achieve a minimum misclassification cost.

Typically, there exist two types of misclassifications. The confusion matrix of the graduate admission example is displayed in Table 4.2 as a detailed illustration of the misclassifications. One type of the misclassifications is the false positive (FP) which means that an accepted observation is predicted to be rejected. The other is known as the false negative (FN) where a case is accepted which should have been rejected.

Table 4.2: The confusion matrix for the graduate admission example

	Predicted Rejected	Predicted Accepted
Actual Rejected	True Positive	False Negative
Actual Accepted	False Positive	True Negative

The costs of misclassifications are prior-known information or decided by the decision makers. The relative cost of having a FP and a FN are represented by c_{FP} and c_{FN}, respectively. In the graduate admission example, the costs of having a FP and a FN are about the same. However, in empirical situations like issuing a loan, the costs of having a FP or a FN could be quite unequal. The misclassification of FN could be potentially costly (since, e.g., this customer may default on his/her loan) while FP implies an opportunity cost. Therefore, it is

more reasonable to minimize the total misclassification cost than minimizing the number of misclassified observations.

As long as there exists an overlap which means $\widehat{O}_{\Lambda,1} \cup \widehat{O}_{\Lambda,2} \neq \emptyset$, the total misclassification cost might be further reduced by shifting the original envelopment frontiers. Before introducing the algorithm to generate the modified frontiers, a general model is introduced to calculate the relative distance of an observation $Z_0 = (X_0, Y_0)$ to a pair of frontiers:

$$
\begin{aligned}
\max_{\lambda_{j,1},\lambda_{j,2},\bar{\delta}_{\Lambda,1},\bar{\delta}_{\Lambda,2}} \quad & \bar{\delta}_{\Lambda,1} + \bar{\delta}_{\Lambda,2} \\
s.t. \quad & \sum_{j\in J_1} \lambda_{j,1} x_{i,j} \leq x_{i,0} + \bar{\delta}_{\Lambda,1} g_{x_i,1} && \forall i \in [m] \\
& \sum_{j\in J_1} \lambda_{j,1} y_{r,j} \geq y_{r,0} + \bar{\delta}_{\Lambda,1} g_{y_r,1} && \forall r \in [s] \\
& \sum_{j\in J_1} \lambda_{j,1} = 1 \\
& \sum_{j\in J_2} \lambda_{j,2} x_{i,j} \geq x_{i,0} + \bar{\delta}_{\Lambda,2} g_{x_i,2} && \forall i \in [m] \\
& \sum_{j\in J_2} \lambda_{j,2} y_{r,j} \leq y_{r,0} + \bar{\delta}_{\Lambda,2} g_{y_r,2} && \forall r \in [s] \\
& \sum_{j\in J_2} \lambda_{j,2} = 1 \\
& \lambda_{j,1} \in \Lambda && \forall j \in J_1 \\
& \lambda_{j,2} \in \Lambda && \forall j \in J_2
\end{aligned}
\tag{4.3.5}
$$

where

(i) $\Lambda \equiv \Lambda^{\mathrm{C}} = \{\lambda_{j,1} \geq 0, \lambda_{j,2} \geq 0\}$, or (ii) $\Lambda \equiv \Lambda^{\mathrm{NC}} = \{\lambda_{j,1} \in \{0,1\}, \lambda_{j,2} \in \{0,1\}\}$.

In model (4.3.5), the pair of frontiers is formed by the observations in J_1 and

J_2, respectively. By solving model (4.3.5) for all the training observations from G_1, the following exclusive sets are derived which accounts for potential three types of misclassifications:

$$\overline{O}_{\Lambda,1} = \{j \in G_1 | \bar{\delta}^*_{\Lambda,1} \geq 0 \text{ and } \bar{\delta}^*_{\Lambda,2} \geq 0\} \tag{4.3.6}$$

$$\overline{Mis}_{\Lambda,1} = \{j \in G_1 | \bar{\delta}^*_{\Lambda,1} < 0 \text{ and } \bar{\delta}^*_{\Lambda,2} \geq 0\} \tag{4.3.7}$$

$$\overline{Gap}_{\Lambda,1} = \{j \in G_1 | \bar{\delta}^*_{\Lambda,1} < 0 \text{ and } \bar{\delta}^*_{\Lambda,2} < 0\} \tag{4.3.8}$$

where $\overline{O}_{\Lambda,1}$ represents the set of the training observations from G_1 that are located in the overlap. The $\overline{Mis}_{\Lambda,1}$ consists of the training observations originally from G_1 but are predicted to belong to the good group. Finally, the training observations in $\overline{Gap}_{\Lambda,1}$ are those located beyond both frontiers, hence it is recorded as the gap area.

Similarly, by solving (4.3.5) for all training observations from G_2, three corresponding sets are derived and are also mutually exclusive:

$$\overline{O}_{\Lambda,2} = \{j \in G_2 | \bar{\delta}^*_{\Lambda,1} \geq 0 \text{ and } \bar{\delta}^*_{\Lambda,2} \geq 0\} \tag{4.3.9}$$

$$\overline{Mis}_{\Lambda,2} = \{j \in G_2 | \bar{\delta}^*_{\Lambda,1} \geq 0 \text{ and } \bar{\delta}^*_{\Lambda,2} < 0\} \tag{4.3.10}$$

$$\overline{Gap}_{\Lambda,2} = \{j \in G_2 | \bar{\delta}^*_{\Lambda,1} < 0 \text{ and } \bar{\delta}^*_{\Lambda,2} < 0\} \tag{4.3.11}$$

After illustrating the potential types of misclassifications, the total misclassification cost could be detailed into the following:

$$\begin{aligned} \overline{C}_\Lambda &= \overline{CO}_\Lambda + \overline{CM}_\Lambda + \overline{CG}_\Lambda \\ &= c_{FN} \times n(\overline{O}_{\Lambda,1}) + c_{FP} \times n(\overline{O}_{\Lambda,2}) \\ &\quad + c_{FN} \times n(\overline{Mis}_{\Lambda,1}) + c_{FP} \times n(\overline{Mis}_{\Lambda,2}) \\ &\quad + c_{FN} \times n(\overline{Gap}_{\Lambda,1}) + c_{FP} \times n(\overline{Gap}_{\Lambda,2}) \end{aligned} \tag{4.3.12}$$

where $\overline{CO}_\Lambda$, $\overline{CM}_\Lambda$ and $\overline{CG}_\Lambda$ represent the total cost of having observations in the overlap, having misclassified observations and having observations in the gap, respectively.

While having $J_1 = \widehat{FS}_{\Lambda,1}$ and $J_2 = \widehat{FS}_{\Lambda,2}$, solving model (4.3.5) is equivalent to solving the model (4.3.1) and model (4.3.3). That is, the derived optimal $\bar{\delta}^*_{\Lambda,1}$ is equal to $\hat{\delta}^*_{\Lambda,1}$ and likewise $\bar{\delta}^*_{\Lambda,2} = \hat{\delta}^*_{\Lambda,2}$. Note that in this case, there only exists the possibility of having training observations located in the overlap. The training observations which are not located in the overlap are correctly classified with the double envelopment frontiers. The total cost is then represented by $\widehat{C}_\Lambda = \widehat{CO}_\Lambda = c_{FN} \times n(\widehat{O}_{\Lambda,1}) + c_{FP} \times n(\widehat{O}_{\Lambda,2})$.

The following algorithm is designed to generate the CS separating frontiers which minimizes the misclassification cost of the training observations.

Step 1: Initialize $d_1 = 1$, $d_2 = 1$,$FS_{\Lambda,1} = \widehat{FS_{\Lambda,1}}$, $FS_{\Lambda,2} = \widehat{FS_{\Lambda,2}}$, and $C_\Lambda = \widehat{C}_\Lambda$.

Step 2: If $d_1 > n(FS_{\Lambda,1})$, then go to Step 3, otherwise $d_2 = d_2 - 1$ and go to Step 4.

Step 3: If $d_2 > n(FS_{\Lambda,2})$, then go to Step 10, otherwise go to Step 5.

Step 4: If $n(FS_{\Lambda,1}) = 1$, then $J_1 = G_1 \setminus FS_{\Lambda,1}$, $J_2 = FS_{\Lambda,2}$ and go to Step 6, otherwise $J_1 = FS_{\Lambda,1} \setminus \{FS_{\Lambda,1}(d_1)\}$, $J_2 = FS_{\Lambda,2}$ and go to Step 6.

Step 5: If $n(FS_{\Lambda,2}) = 1$, then $J_1 = FS_{\Lambda,1}$ and $J_2 = G_2 \setminus FS_{\Lambda,2}$, otherwise $J_1 = FS_{\Lambda,1}$ and $J_2 = FS_{\Lambda,2} \setminus \{FS_{\Lambda,2}(d_2)\}$.

Step 6: Solve model (4.3.5) for the training observations in $G_1 \cup G_2$ and calculate the total misclassification cost $\overline{C}_\Lambda$.

Step 7: If $\overline{C}_\Lambda < C_\Lambda$, then set $C_\Lambda = \overline{C}_\Lambda$, $FS_{\Lambda,1} = J_1$, $FS_{\Lambda,2} = J_2$ and go to Step 8, otherwise go to Step 9.

Step 8: If $\overline{CO}_{\Lambda} = 0$, then go to Step 10, otherwise set $d_1 = d_2 = 1$ and go to Step 2.

Step 9: Set $d_1 = d_1 + 1$ and $d_2 = d_2 + 1$, then go to Step 2.

Step 10: End.

The algorithm will check the frontier observations point by point to see if the exclusion of one frontier observation contributes to the reduction of the total misclassification cost. The algorithm stops if the total misclassification cost of the training sample reaches the minimum or if there is no more overlap existed. By running the above algorithm, two final frontier sets are derived to form the CS separating frontier, namely $FS_{\Lambda,1}$ and $FS_{\Lambda,2}$. The CS separating frontiers are then used to predict the membership of a new observation.

The same graduate admission example is applied to show the final CS separating frontier derived by running the above algorithm. The costs of having a FP and a FN are set to be the same, which is 1. In Table 4.3, the detailed procedures for generating the NC frontiers is displayed.

The final NC cs separating frontiers are showed in Figure 4.3. It is observed that by allowing the training observations 4 and 11 to be misclassified, the overlap is completely eliminated. Except these two frontier observations, all other training observations are situated on the opposite sides of the CS separating frontiers. The total misclassification cost is reduced from 7 to 2.

For the convex case, the total misclassification cost which is originally 10 could also be further reduced by running the proposed algorithm. The exclusion of observations 4, 11 and 26 contributes to a total misclassification cost of 5. The overlap could not be fully eliminated. Comparing to the NC case, the total misclassification cost is still higher although being significantly reduced.

Table 4.3: Specific procedures of running the algorithm for generating the NC frontiers

Steps	d_1	d_2	J_1	J_2	$FS_{\Lambda^{NC},1}$	$FS_{\Lambda^{NC},2}$	$C_{\Lambda^{NC}}$
1	1	1			{23,25,26,29}	{4,11,15,16}	7
2-4-6-7	1	0	{25,26,29}	{4,11,15,16}			
9	2	1					
2-4-6-7	2	0	{23,26,29}	{4,11,15,16}			
9	3	1					
2-4-6-7	3	0	{23,25,29}	{4,11,15,16}			
9	4	1					
2-4-6-7	4	0	{23,25,26}	{4,11,15,16}			
9	5	1					
2-3-5-6-7	5	1	{23,25,26,29}	{11,15,16}			
9	6	2					
2-3-5-6-7	6	2	{23,25,26,29}	{4,15,16}	{23,25,26,29}	{4,15,16}	5
8	1	1					
2-4-6-7	1	0	{25,26,29}	{4,15,16}			
9	2	1					
2-4-6-7	2	0	{23,26,29}	{4,15,16}			
9	3	1					
2-4-6-7	3	0	{23,25,29}	{4,15,16}			
9	4	1					
2-4-6-7	4	0	{23,25,26}	{4,15,16}			
9	5	1					
2-3-5-6-7-8-10	5	1	{23,25,26,29}	{3,15,16}	{23,25,26,29}	{15,16}	2

Figure 4.3: Shifted nonconvex frontiers by excluding some frontier observations

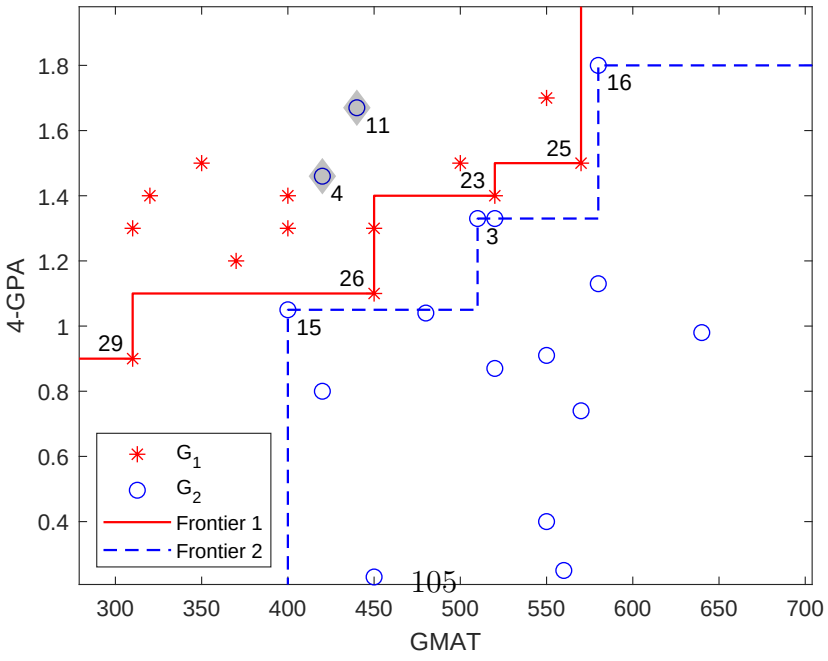
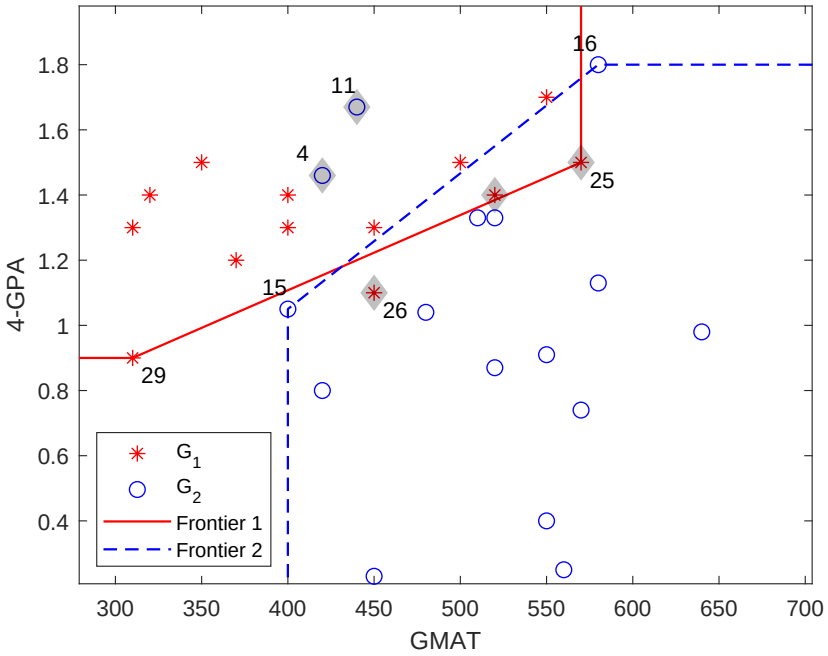


Figure 4.4: Shifted convex frontiers by excluding some frontier observations



4.4 Double-Frontier Based Discriminant Rules

The above CS separating frontiers are then used to predict the membership of a test observation which is characterized by the same characteristic variables. Specifically, these two separating frontiers are formed by the training observations in $FS_{\Lambda,1}$ and $FS_{\Lambda,2}$. The following model is used to calculate the distance of the observation $Z_0 = (X_0, Y_0)$ relative to two separating frontiers:

$$
\begin{aligned}
\max_{\lambda_{j,1},\lambda_{j,2},\delta_{\Lambda,1},\delta_{\Lambda,2}} \quad & \delta_{\Lambda,1}+\delta_{\Lambda,2} \\
s.t. \quad & \sum_{j\in FS_{\Lambda,1}} \lambda_{j,1}x_{i,j} \le x_{i,0}+\delta_{\Lambda,1}g_{x_i,1} && \forall i\in[m] \\
& \sum_{j\in FS_{\Lambda,1}} \lambda_{j,1}y_{r,j} \ge y_{r,0}+\delta_{\Lambda,1}g_{y_r,1} && \forall r\in[s] \\
& \sum_{j\in FS_{\Lambda,1}} \lambda_{j,1} = 1 \\
& \sum_{j\in FS_{\Lambda,2}} \lambda_{j,2}x_{i,j} \ge x_{i,0}+\delta_{\Lambda,2}g_{x_i,2} && \forall i\in[m] \\
& \sum_{j\in FS_{\Lambda,2}} \lambda_{j,2}y_{r,j} \le y_{r,0}+\delta_{\Lambda,2}g_{y_r,2} && \forall r\in[s] \\
& \sum_{j\in FS_{\Lambda,2}} \lambda_{j,2} = 1 \\
& \lambda_{j,1}\in\Lambda && \forall j\in FS_{\Lambda,1} \\
& \lambda_{j,2}\in\Lambda && \forall j\in FS_{\Lambda,2}
\end{aligned}
\tag{4.4.1}
$$

where

$$\text{(i) } \Lambda \equiv \Lambda^{\mathrm{C}}=\{\lambda_{j,1}\ge 0,\lambda_{j,2}\ge 0\}, \text{ or (ii) } \Lambda\equiv\Lambda^{\mathrm{NC}}=\{\lambda_{j,1}\in\{0,1\},\lambda_{j,2}\in\{0,1\}\}.$$

There are four possible combinations of $\delta^*_{\Lambda,1}$ and $\delta^*_{\Lambda,2}$ which imply different membership information.

(s.1) If $\delta^*_{\Lambda,1}\ge 0$ and $\delta^*_{\Lambda,2}<0$, then the observation Z_0 belongs to the bad group;

(s.2) If $\delta^*_{\Lambda,1}<0$ and $\delta^*_{\Lambda,2}\ge 0$, then the observation Z_0 belongs to the good group;

(s.3) If $\delta^*_{\Lambda,1}<0$ and $\delta^*_{\Lambda,2}<0$, then the observation Z_0 is in the gap area;

(s.4) If $\delta^*_{\Lambda,1}\ge 0$ and $\delta^*_{\Lambda,2}\ge 0$, then the observation Z_0 is in the overlap area.

The first two situations are quite clear. In situation 1, the observation Z_0 is located within the separating frontier 1 and beyond the separating frontier 2. Therefore, the only valid membership information supports it to be belonging to the bad group. In situation 2, the observation Z_0 is located within the separating frontier 2 and beyond the separating frontier 1. Therefore, the only valid membership information supports it to be belonging to the good group. However, for situation 3 and 4, the currently information is not enough for determine its membership clearly.

For the ambiguous situation 3 and 4, a conservative and honest way is to report the situation as it is. Alternatively, the membership of the observations in such ambiguous situations could be inferred by comparing the relative distance values. For the observation that satisfies the situation 3, the closer the test observation is located to a separating frontier, the more similarities it is supposed to share with the corresponding group. Hence, it is perceived to belong to the group whose separating frontier is closer to the test observation. Furthermore, by incorporating the misclassification costs, if $0 > \delta^*_{\Lambda,1}/c_{FN} \geq \delta^*_{\Lambda,2}/c_{FP}$ holds, then this observation belongs to the bad group. On the contrary, if $\delta^*_{\Lambda,1}/c_{FN} < \delta^*_{\Lambda,2}/c_{FP} < 0$ holds, then this observation belongs to the good group. For the observation that satisfies the situation 4, the opposite rule is assumed. The closer the test observation is located to a separating frontier, the higher possibility that this observation is going to leave the corresponding group. Therefore, it should be classified into the group whose separating frontier is farther away. That is, if $\delta^*_{\Lambda,1} \times c_{FN} \geq \delta^*_{\Lambda,2} \times c_{FP} \geq 0$ holds, then this observation belongs to the bad group. On the contrary, if $0 \leq \delta^*_{\Lambda,1} \times c_{FN} < \delta^*_{\Lambda,2} \times c_{FP}$ holds, then this observation belongs to the good group.

Apparently, for the situation 1 and situation 2, the membership is predicted simply by the sign of the distance measure. While for the two ambiguous situ-

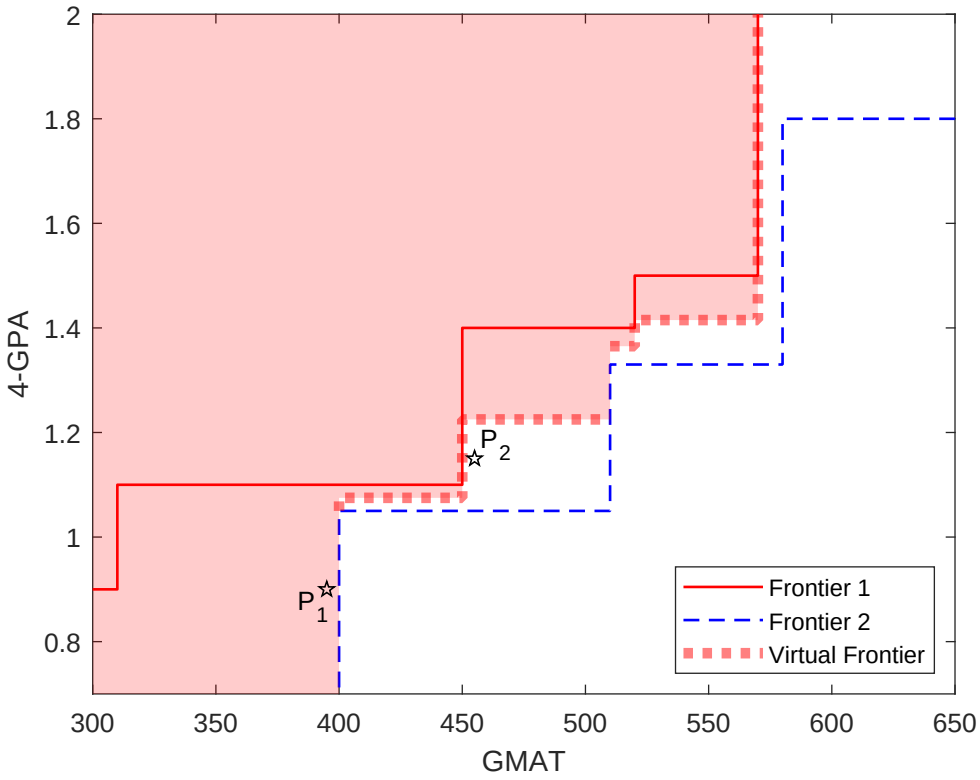
ations which are situation 3 and situation 4, the value of the distance measure which measures the closeness to the separating frontier is particularly important for predicting the membership. This implies that the choice of the direction vectors is of vital importance while applying the double frontier methods.

In all the C frontier classification literature, either an input-oriented or an output-oriented radial measure is applied. That is, for the test observation $Z_0 = (X_0, Y_0)$, the two direction vectors under the input-oriented radial case are $g_1 = (-|X_0|, \mathbf{0})$ and $g_2 = (|X_0|, \mathbf{0})$, respectively. Note that $\mathbf{0}$ represents a vector whose components are all zeros. This direction vector allows the test observation to reach the separating frontiers by changing its input-type characteristic variables. While for the output-oriented radial case, the two direction vectors are $g_1 = (\mathbf{0}, |Y_0|)$ and $g_2 = (\mathbf{0}, -|Y_0|)$, respectively. In this case, the test observation is only allowed to change its output-type characteristic variables in order to reach the separating frontiers.

With the graduate admission example, we show that the commonly used radial measures are not the best choice for the classification. Take the input-oriented case as an example, the allowed changes are increasing or decreasing along the vertical axis in Figure 4.5 and Figure 4.6. The vertical axis represents the characteristic variable of 4-GPA which is the smaller the better. In both figures, the solid lines represent the separating frontier 1 which bounds the bad group, while the dashed lines represent the separating frontier 2 which bounds the good group. By applying the input-oriented radial measure, the bad group is represented by the shaded area which is located above and to the left of the dotted lines. Correspondingly, the area restricted to the fourth quadrant located below and to the right of the dotted lines represents the good group. There is no doubt that the observations located above the separating frontier 1 belong to the bad group. The observations located

below the separating frontier 2 belong to the bad group. The area worths further investigating are the gap and overlap.

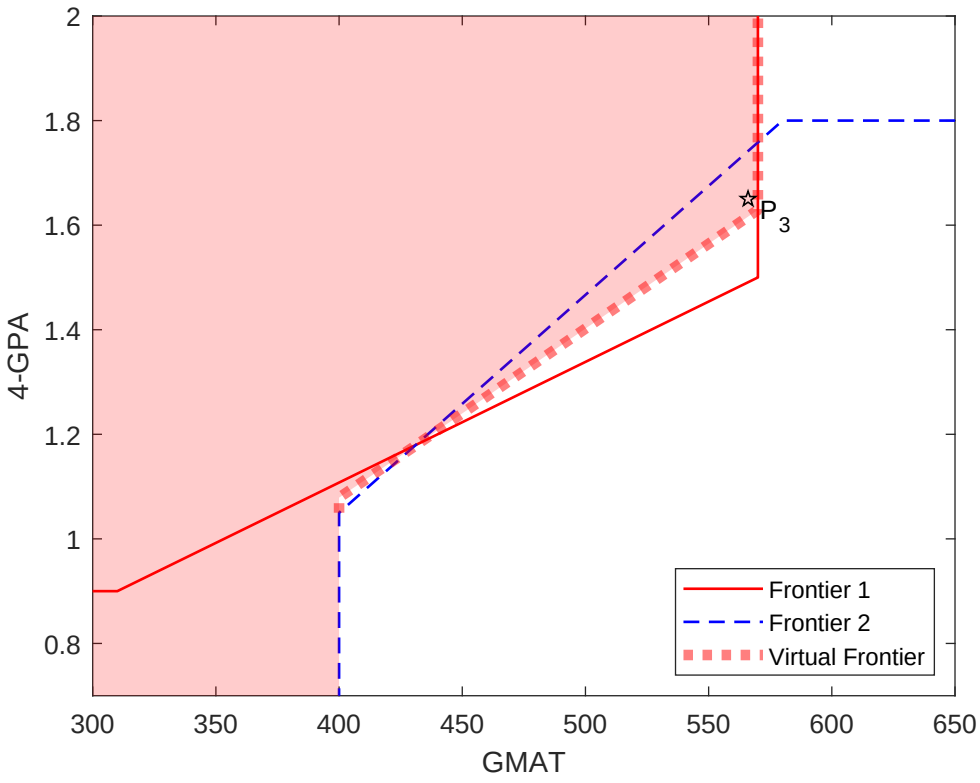
Figure 4.5: The diagram of the separating frontiers with a radial measure under the NC case



We first look into the NC case which is displayed in Figure 4.5. There is no overlap for this illustrative example under the NC case. P_1 and P_2 which are marked by the pentagrams are two observations located in the gap area. With the input-oriented measure, the distance of the observations relative to the separating frontier is measured by vertically projecting to the frontier. For observation P_1, the vertical distance to the separating frontier 1 is finite while that to the separating frontier 2 is infinite. It is therefore considered to belong to the bad group, although it is located just next to the separating frontier 2. For observation P_2, its vertical

distance to the separating frontier 1 is larger than that to the separating frontier 2. Hence, it is predicted to belong to the good group while it is located just next to the separating frontier 1.

Figure 4.6: The diagram of the separating frontiers with a radial measure under the C case



Under the C case of this illustrative example in Figure 4.6, there also exist unreasonable predictions for the observations located in the gap and overlap. We focus on the observations that are located in the overlap. Observation P_3, for example, is located vertically closer to the separating frontier 2 comparing to separating frontier 1. However, it is actually situating next to the separating frontier 1. If there is prior information on that only the input-type characteristic variables are adjustable, then this input-oriented measure makes sense. Otherwise,

the choice of this measure may give rise to unreasonable situations like observations P_1, P_2 and P_3.

Among the various choices possible for the direction vector in practical applications, the DDF measure which obtains a proportional interpretation (see Briec (1997)) is used in this contribution. To be specific, $g_1 = (-|X_0|, |Y_0|)$ is used for projecting $Z_0 = (X_0, Y_0)$ to the separating frontier 1 while $g_2 = (|X_0|, -|Y_0|)$ is used for projecting it to the separating frontier 2. Note that in the classification context with potentially negative characteristic variables, the absolute value is used for preserving a proportional interpretation (Kerstens and Van de Woestyne (2011)).

Figure 4.7: The diagram of the separating frontiers with a proportional DDF measure under the NC case

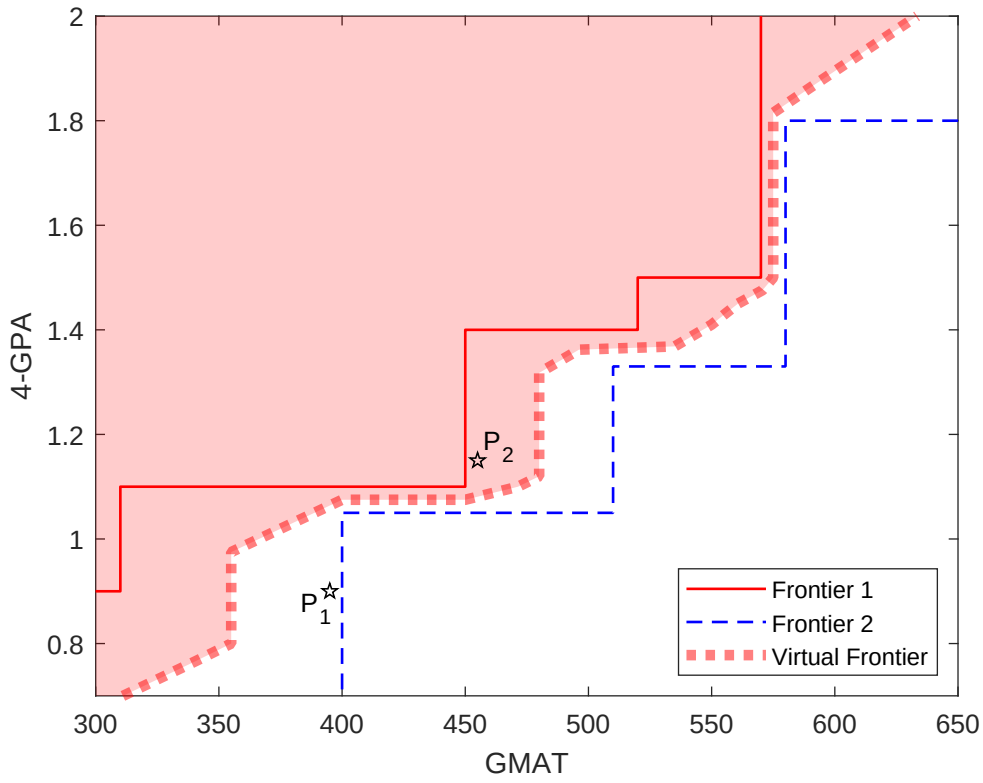
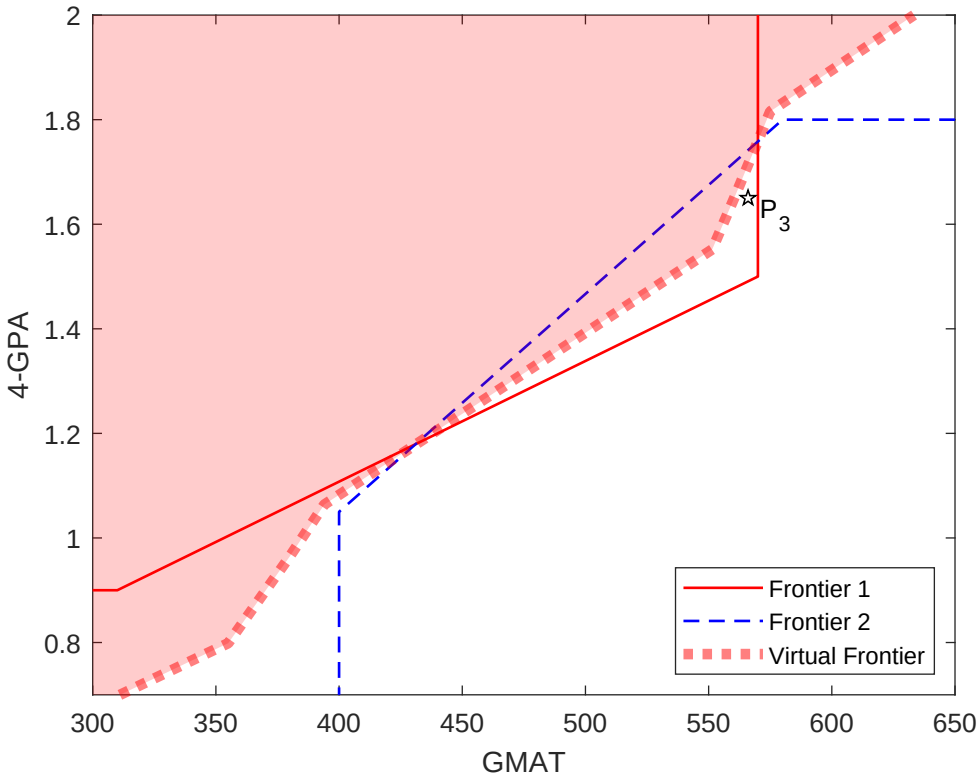


Figure 4.8: The diagram of the separating frontiers with a proportional DDF measure under the C case



With the proportional direction distance measure, the membership of the observations located in the overlap and gap are more properly defined. The results of the graduate admission example are displayed in Figure 4.7 and Figure 4.8. The same marks are used as those in Figure 4.5 and Figure 4.6. The observation P_1 is now predicted to belong to the good group and the observation P_2 belongs to the bad group. For the observation P_3 located in the overlap, it is now believed to belong to the good group.

From the comparison between the input-oriented radial measure and the proportional DDF measure, the latter is apparently more reasonable. Thus, the values of the directional DDF measure are used to predict the membership of an obser-

vation Z_0. To sum up, the following classification rules which incorporate the misclassification costs are used:

(R.1) If $\delta^*_{\Lambda,1} \geq 0$ and $\delta^*_{\Lambda,2} < 0$, then Z_0 belongs to the bad group;

(R.2) If $\delta^*_{\Lambda,1} < 0$ and $\delta^*_{\Lambda,2} \geq 0$, then Z_0 belongs to the good group;

(R.3) If $0 > \delta^*_{\Lambda,1}/c_{FN} \geq \delta^*_{\Lambda,2}/c_{FP}$, then Z_0 belongs to the bad group;

(R.4) If $\delta^*_{\Lambda,1}/c_{FN} < \delta^*_{\Lambda,2}/c_{FP} < 0$, then Z_0 belongs to the good group;

(R.5) If $\delta^*_{\Lambda,1} \times c_{FN} \geq \delta^*_{\Lambda,2} \times c_{FP} \geq 0$, then Z_0 belongs to the bad group;

(R.6) If $0 \leq \delta^*_{\Lambda,1} \times c_{FN} < \delta^*_{\Lambda,2} \times c_{FP}$, then Z_0 belongs to the good group;

4.5 Conclusions

The nonparametric frontier-based classifier is a good choice in estimating the separating hypersurfaces whose shape is mostly unspecified in applications. However, in the DEA-based classification literature, the assumptions made to construct an efficient frontier are blindly copied to estimate the separating frontier. There is a lack of correspondence between the axioms implied by a nonparametric frontier and the background information known for characterizing a separating frontier. This leads to a need for building the connection between the commonly used assumptions and the prior known background knowledge information.

This study proposes a novel method for accommodating different mixes of background knowledge information and asymmetric misclassification costs. By reflecting the background information on the monotonic relation, a NC classifier is constructed with the assumption on free disposability. If there is prior information

on the substitution relation, then a C classifier is generated with an additional assumption on convexity. The graduate admission data shows that the NC classifier has a tighter envelopment than the C one does. The overlap under the NC case is therefore smaller than the C one does. Then, a moderate way is proposed to shift the frontier inwards so that the misclassification cost which is generated by minimizing the overlaps. Furthermore, the discriminant rules are also designed to incorporate the cost information. With the graduate admission data, it is shown that the choice of the measure matters while applying a double frontier method. Specifically, the proposed proportional DDF measure outperforms the commonly used radial measure in providing a reasonable separation.

Several limitations can be identified relating to the presented approach. First, the empirical validation shows that the choice of the direction vector matters in improving the classification performance. In this study, a proportional direction measure is proved to be more favorable than a radial measure. A further work could be investigating the choice of the direction vector and explore the best projection direction. Second, there is a clear overlap and gap area defined under the double-frontier methods which essentially indicates further information needed. In this study, the observations located in the overlap and gap are further classified by comparing their distances to the frontiers. The alternative choices could be the nearest neighbourhood approach like Pendharkar (2011) did. Third, both the NC and C classifiers proposed could be extended from two groups to multiple groups (see Pendharkar and Troutt (2011), Wu, An, and Liang (2011)).

References

Bajgier, S. M., and A. V. Hill (1982): "An Experimental Comparison of Statistical and Linear Programming Approaches to the Discriminant Problem," *Decision Sciences*, 13(4), 604–618.

Banker, R. D., H. S. Chang, and W. W. Cooper (2002): ""Small Sample Properties of ML, COLS and DEA Estimators of Frontier Models in the Presence of Heteroscedasticity" by AN Bojanic, SB Caudill and JM Ford, European Journal of Operational Research 108, 1998, 140–148: A Comment," *European Journal of Operational Research*, 136(2), 466–467.

Briec, W. (1997): "A Graph-Type Extension of Farrell Technical Efficiency Measure," *Journal of Productivity Analysis*, 8(1), 95–110.

Chambers, R., Y. Chung, and R. Färe (1998): "Profit, Directional Distance Functions, and Nerlovian Efficiency," *Journal of Optimization Theory and Applications*, 98(2), 351–364.

Chang, D., and Y. Kuo (2008): "An Approach for the Two-group Discriminant Analysis: An Application of DEA," *Mathematical and Computer Modelling*, 47(9-10), 970–981.

Charnes, A., W. Cooper, and E. Rhodes (1978): "Measuring the Efficiency of Decision Making Units," *European Journal of Operational Research*, 2(6), 429–444.

Cherchye, L., T. Kuosmanen, and T. Post (2001): "FDH Directional Distance Functions with An Application to European Commercial Banks," *Journal of Productivity Analysis*, 15(3), 201–215.

De Bock, K. W. (2017): "The Best of Two Worlds: Balancing Model Strength and Comprehensibility in Business Failure Prediction Using Spline-Rule Ensembles," *Expert Systems with Applications*, 90, 23–39.

De Caigny, A., K. Coussement, K. W. De Bock, and S. Lessmann (2019): "Incorporating Textual Information in Customer Churn Prediction Models based on a Convolutional Neural Network," *International Journal of Forecasting*.

Deprins, D., L. Simar, and H. Tulkens (1984): "Measuring Labor Efficiency in Post Offices," in *The Performance of Public Enterprises: Concepts and Measurements*, ed. by M. Marchand, P. Pestieau, and H. Tulkens, pp. 243–268. North Holland, Amsterdam.

Fisher, R. A. (1936): "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7(2), 179–188.

Freed, N., and F. Glover (1986): "Evaluating Alternative Linear Programming Models to Solve the Two-Group Discriminant Problem," *Decision Sciences*, 17(2), 151–162.

Jahanshahloo, G. R., F. H. Lotfi, F. R. Balf, and H. Z. Rezai (2007): "Discriminant Analysis of Interval Data Using Monte Carlo Method in Assessment of Overlap," *Applied Mathematics and Computation*, 191(2), 521–532.

Kerstens, K., and I. Van de Woestyne (2011): "Negative Data in DEA: A Simple Proportional Distance Function Approach," *Journal of the Operational Research Society*, 62(7), 1413–1419.

Kotsiantis, S. B., I. Zaharakis, and P. Pintelas (2007): "Supervised Machine Learning: A Review of Classification Techniques," *Emerging Artificial Intelligence Applications in Computer Engineering*, 160, 3–24.

Kuo, Y.-C. (2013): "Consideration of Uneven Misclassification Cost and Group Size for Bankruptcy Prediction," *American Journal of Industrial and Business Management*, 3(08), 708.

Leon, C. F., and F. Palacios (2009): "Evaluation of Rejected Cases in an Acceptance System with Data Envelopment Analysis and Goal Programming," *Journal of the Operational Research Society*, 60(10), 1411–1420.

Lessmann, S., B. Baesens, H.-V. Seow, and L. C. Thomas (2015): "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," *European Journal of Operational Research*, 247(1), 124–136.

Lotfi, F. H., and B. Mansouri (2008): "The Extended Data Envelopment Analysis/Discriminant Analysis Approach of Fuzzy Models," *Applied Mathematical Sciences*, 2(30), 1465–1477.

Pendharkar, P. (2012): "Fuzzy Classification Using the Data Envelopment Analysis," *Knowledge-Based Systems*, 31, 183–192.

Pendharkar, P. (2018): "Data Envelopment Analysis Models for Probabilistic Classification," *Computers & Industrial Engineering*, 119, 181–192.

Pendharkar, P., M. Khosrowpour, and J. Rodger (2000): "Application of Bayesian Network Classifiers and Data Envelopment Analysis for Mining Breast Cancer Patterns," *Journal of Computer Information Systems*, 40(4), 127–132.

Pendharkar, P., J. Rodger, and G. Yaverbaum (1999): "Association, Stat-

istical, Mathematical and Neural Approaches for Mining Breast Cancer Patterns," *Expert Systems with Applications*, 17(3), 223–232.

Pendharkar, P. C. (2002): "A Potential Use of Data Envelopment Analysis for the Inverse Classification Problem," *Omega*, 30(3), 243–248.

Pendharkar, P. C. (2011): "A Hybrid Radial Basis Function and Data Envelopment Analysis Neural Network for Classification," *Computers & Operations Research*, 38(1), 256–266.

Pendharkar, P. C., and M. D. Troutt (2011): "DEA Based Dimensionality Reduction for Classification Problems Satisfying Strict Non-satiety Assumption," *European Journal of Operational Research*, 212(1), 155–163.

——— (2014): "Interactive Classification Using Data Envelopment Analysis," *Annals of Operations Research*, 214(1), 125–141.

Seiford, L., and J. Zhu (1998): "An Acceptance System Decision Rule with Data Envelopment Analysis," *Computers & Operations Research*, 25(4), 329–332.

Silva, A. P. D., and A. Stam (1994): "Second Order Mathematical Programming Formulations for Discriminant Analysis," *European Journal of Operational Research*, 72(1), 4–22.

Smaoui, S., H. Chabchoub, and B. Aouni (2009): "Mathematical Programming Approaches to Classification Problems," *Advances in Operations Research*, 2009, Art. ID 252989.

Smith, C. A. (1946): "Some Examples of Discrimination," *Annals of Eugenics*, 13(1), 272–282.

Stam, A., and C. Ragsdale (1992): "On the Classification Gap in Mathematical Programming-Based Approaches to the Discriminant Problem," *Naval Research Logistics*, 39(4), 545–559.

Sueyoshi, T. (1999): "DEA-Discriminant Analysis in the View of Goal Programming," *European Journal of Operational Research*, 115(3), 564–582.

——— (2001): "Extended DEA-Discriminant Analysis," *European Journal of Operational Research*, 131(2), 324–351.

——— (2004): "Mixed Integer Programming Approach of Extended DEA–Discriminant Analysis," *European Journal of Operational Research*, 152(1), 45–55.

——— (2006): "DEA-Discriminant Analysis: Methodological Comparison Among Eight Discriminant Analysis Approaches," *European Journal of Operational Research*, 169(1), 247–272.

Sueyoshi, T., and M. Goto (2010): "Measurement of a Linkage among Environmental, Operational, and Financial Performance in Japanese Manufacturing Firms: A Use of Data Envelopment Analysis with Strong Complementary Slackness Condition," *European Journal of Operational Research*, 207(3), 1742–1753.

Troutt, M., A. Rai, and A. Zhang (1996): "The Potential Use of DEA for Credit Applicant Acceptance Systems," *Computers & Operations Research*, 23(4), 405–408.

Wu, J., Q. An, and L. Liang (2011): "A Modified Super-efficiency DEA Approach for Solving Multi-groups Classification Problems," *International Journal of Computational Intelligence Systems*, 4(4), 606–618.

Yan, H., and Q. Wei (2011): "Data Envelopment Analysis Classification Machine," *Information Sciences*, 181(22), 5029–5041.

Zhu, J. (2003): *Quantitative Models for Performance Evaluation and Benchmarking: Data Envelopment Analysis with Spreadsheets and DEA Excel Solver*, vol. 51. Springer Science & Business Media.

CHAPTER 5 General Conclusion

This thesis has been aiming at modifying some existing nonparametric frontier classifiers. It was done by building the connections between the axioms applied in the nonparametric methods and the background information in classification so that some commonly used axioms could be relaxed depending on the applications. Furthermore, we aimed at proposing novel nonparametric frontier classifiers by incorporating some essential background information into the classification, e.g., the cost information and the non-monotonic relation between characteristic variables. In what follows, we highlight some results and the main contributions.

Chapter 2 has innovated in two main ways. First, the convex frontier-based classifier is modified into a more generalized form by the inclusion of both the characteristic variables with a monotonically increasing relation and those with a monotonically decreasing relation. If the monotonic relation is not priori given, a linear discriminant analysis model named the Minimize Sum of Deviations model is applied to reflect the relation based on the observations. Apart from the consideration of the monotonic relations, the directional distance function measure is introduced to give further information on alternative improvements. Second, in-

stead of sticking to the convexity assumption, a nonconvex frontier has been used and ends up with a better envelopment of the training observations.

The classification problem was explored in a more general formulation in **Chapter 3** where both monotonic and non-monotonic relations were incorporated. Different from the standard disposal assumption used for describing the monotonic relation, a general disposal assumption which limits the disposability within a value range was defined for characterizing the non-monotonic relation. Accordingly, a dominance adapting directional distance (DAD) function was developed for measuring the distance of an observation to the separating frontier. A separating hull consists of these dominance adapting separating frontiers was used to predict the membership of a new observation. We designed an algorithm to simplify the procedures of predicting the membership of a new observation. Both the convex and the nonconvex classifiers were introduced in this setting. We argue that only if there exists the additional background knowledge on the substitution relation among the characteristic variables, then a convex nonparametric classifier is preferred. Otherwise, a nonconvex classifier is more conservative and provides better classification performance than the C classifier. The proposed nonparametric classifiers were examined with a nonlinearly separable binary classification problem. The NC classifier outperformed the C classifier in terms of the overall accuracy, the precision and the F1 measure. It confirmed our argument of applying a C classifier only after detecting a substitution relation. Moreover, the proposed NC classifier was shown to outperform some existing DEA-based methods in terms of several commonly used criteria.

The nonparametric classifiers proposed in **Chapter 4** consisted of two separating frontiers which explicitly described the overlap. The two separating frontiers were adjustable so that the total misclassification cost could be minimized.

Furthermore, the discriminant rules were also designed to incorporate the cost information. Similarly, both the convex and nonconvex classifiers were proposed. The graduate admission data showed that the NC classifier has a tighter envelopment than the C one does. The overlap under the NC case was therefore smaller than the C one does. With the same graduate admission data, it was shown that the choice of the measure matters while applying a double frontier method. Specifically, the proposed proportional DDF measure outperforms the commonly used radial measure in providing a reasonable separation.