A Methodology for the Characterization of Business-to-Consumer E-commerce

by

Alfredo Vellido

December 2000

A Cristina

Morena soy, oh hijas de Jerusalén, pero codiciable

Como las tiendas de Cedar,

Como las cortinas de Salomón.

No reparéis en que soy morena,

Porque el sol me miró.

El Cantar de los Cantares, 1, 5-6

Nuestro misterio, de todos modos, aunque breve como una primavera, es de una considerable intensidad.

Félix de Azúa

Abstract

A Methodology for the Characterization of Business-to-Consumer E-commerce

Alfredo Vellido

This thesis concerns the field of business-to-consumer electronic commerce. Research on Internet consumer behaviour is still in its infancy, and a quantitative framework to characterize user profiles for e-commerce is not yet established. This study proposes a quantitative framework that uses latent variable analysis to identify the underlying traits of Internet users' opinions. Predictive models are then built to select the factors that are most predictive of the propensity to buy on-line and classify Internet users according to that propensity. This is followed by a segmentation of the online market based on that selection of factors and the deployment of segment-specific graphical models to map the interactions between factors and between these and the propensity to buy online.

The novel aspects of this work can be summarised as follows: the definition of a fully quantitative methodology for the segmentation and analysis of large data sets; the description of the latent dimensions underlying consumers' opinions using quantitative methods; the definition of a principled method of marginalisation to the empirical prior, for Bayesian neural networks, to deal with the use of class-unbalanced data sets; a study of the Generative Topographic Mapping (GTM) as a principled method for market segmentation, including some developments of the model, namely: a) an entropy-based measure to compare the class-discriminatory capabilities of maps of equal dimensions; b) a *Cumulative Responsibility* measure to provide information on the mapping distortion and

define data clusters; c) *Selective Smoothing* as an extended model for the regularization of the GTM training.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

ARD: Automatic Relevance Determination

AURP: Area Under the ROC Plot

BNN: Bayesian Neural Networks

BP: Pack-Propagation

CG: Conditional Gaussian (distribution)

CIM: Conditional Independence Maps

CHAID: Chi-Squared Automatic Interaction Detector

EM: Expectation-Maximization

FN: False Negatives

FP: False Positives

FSCL: Frequency Sensitive Competitive Learning

GTM: Generative Topographic Mapping

GVU: Graphics, Visualization & Usability (Center)

KMO: Kaiser-Meyer-Olkin (measure of sampling adequacy)

LDA: Linear Discriminant Analysis

LR: Logistic Regression

MCMC: Markov Chain Monte Carlo

MLP: Multilayer Perceptron

NN: Neural Network

ROC: Receiver Operator Characteristic

RVM: Relevance Vector Machine

SOM: Self-Organizing Map

SMS: Selective Mapping Smoothing

SRT: Single Regularization Term

SVM: Support Vector Machine

TN: True Negatives

TP: True Positives

WWW: World-wide Web

PART 1

Introduction

Cuando la flecha está en el arco, tiene que partir.

Rafael Sánchez Ferlosio

Chapter 1

Introduction and review of the literature.

1.1. Introduction

The area of electronic commerce is complex and multifaceted and its full conceptualization is still emerging. This thesis only concerns the category of consumer-oriented or business-to-consumer electronic commerce (Zwass, 1999), deemed to have considerable market potential for the future. The context of electronic commerce is rich both in hype and hard facts, and interest in it has been fuelled by the potential size of the world-wide market resulting from the exponential growth of Internet adoption. Hype is exemplified by the success of the stock market capitalisation of many companies involved on it, regardless of their performance in terms of profit generation. However there is much factual information to support the expectation of a large market size: According to CommerceNet / Nielsen Media Research (August, 1998), the Internet population was growing at a 2.5% monthly rate, while the online-buyers population was growing at a rate of 8%. Following the Iconocast "Internet at a glance report" (Iconocast, 2000), 38.8 out of 102.1 million U.S. households were online in 1999 (source: Jupiter Communications) and, considering only the 10 top "online spending" countries in the world (August 1999, International Data Corporation), $37.4B were spent in the U.S. alone and $8.4B elsewhere.

Promising as the figures may appear, the business-to-consumer online market has yet to reach the critical mass to ensure its own future success. This is the first necessary requirement for the introduction of innovations, especially in the telecommunications ambit (Rice *et al.*, 1990). For that to happen, technological breakthroughs, such as the merging of Internet with TV broadcasting and mobile telephony, must permeate the market. Nevertheless, new technologies and the growth of Internet adoption might not be sufficient to realize the full potential of the online consumer market. In fact, this exponential growth has also been reported as the reason behind a worsening of online shopping customer service that threatens to put at risk customer loyalty in this marketing channel (Glasner, 1999). This threat exemplifies the necessity to go beyond a merely technology-centred view to carefully explore customer needs and expectations. Public organisations such as the European Association for the Co-ordination of Consumer Representation in Standardization (ANEC) are already trying to promote standardisation of online transaction protocols, based on consumers' opinions, to sustain e-commerce growth (Farquhar and Balfour, 1998). A list of priorities were identified, falling into two broad areas: *security / privacy* and *ease of use*, including aspects of *assurance* and the management of *perception of risk*.

The exploration of online consumers' needs and expectations justifies a data-based analysis of their shopping behaviour, in order to provide a sound empirical basis for business decision making. Most of the existing studies in the area are basically qualitative in nature. Unfortunately, the quantitative research on consumer behaviour online is still in its infancy and there is still no quantitative framework for the analysis of online purchasing behaviour. This thesis aims to provide a framework along the lines described in the next section.

1.2. Goals of the thesis

This thesis aims to put forward a complete quantitative methodology for the exploratory analysis of online consumers' shopping behaviour. This methodology is summarized in figure 1.1 and comprises several stages, corresponding to the parts in which the thesis is organised. These stages are, namely:

- **Latent variable analysis:** A survey data set containing a large number of online behaviour-related variables, from a sample of Internet users' opinions of online shopping and online vendors, is made available for this research. The latent traits behind these consumer opinions are to be extracted, using factor analysis and labeled according to existing literature.
- **Predictive modelling:** The latent factors obtained in the previous part, together with demographic, socio-economic and Web usage information, are to be used to construct a global predictive model of the propensity to buy online.
- **Cluster-based segmentation:** Those factors selected as the best predictors of the propensity to buy online are to be used as bases for the segmentation of the online

consumer market, utilizing a neural network-based model for data clustering and visualization: the Generative Topographic Mapping (GTM).

- **Graphical Modelling:** Having identified the different consumer segments, it will then be possible to apply Graphical Modelling separately to each segment, to produce graphs representing the inter-relationships between the predictive factors, or segmentation bases, and between these and the propensity to buy online.

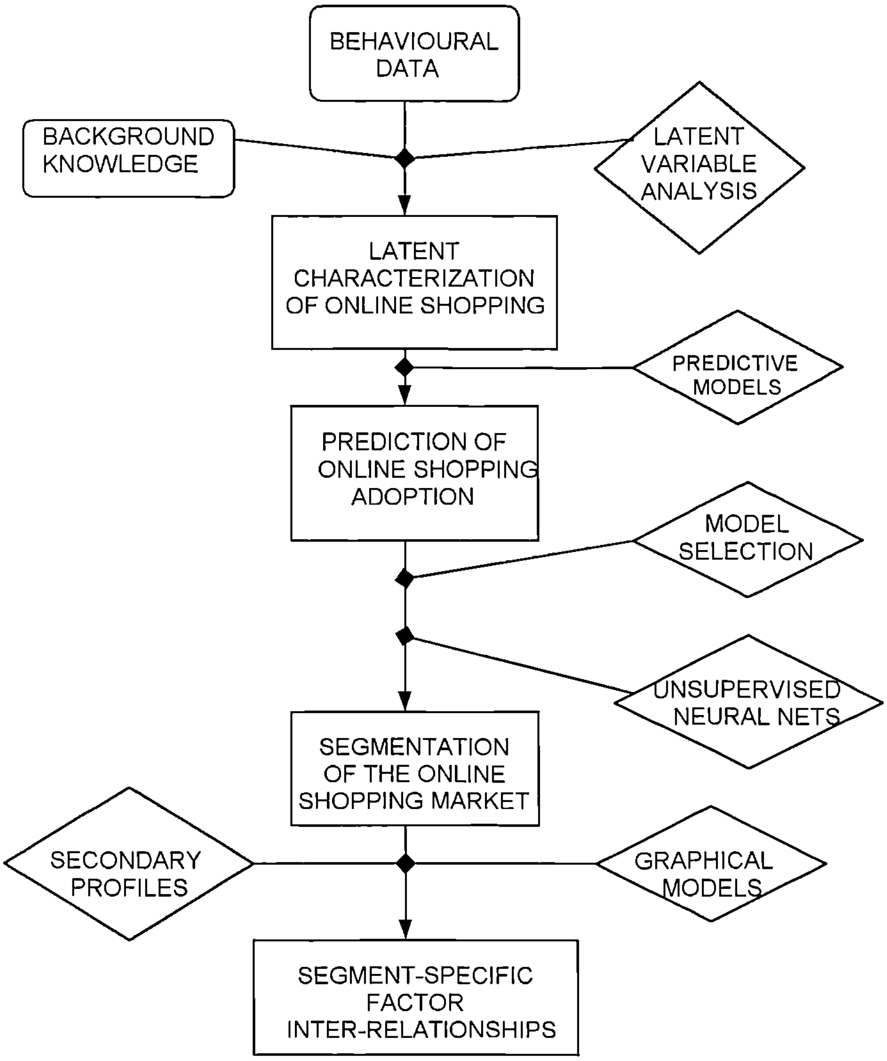

BEHAVIOURAL DATA
BACKGROUND KNOWLEDGE
LATENT VARIABLE ANALYSIS
LATENT CHARACTERIZATION OF ONLINE SHOPPING
PREDICTIVE MODELS
PREDICTION OF ONLINE SHOPPING ADOPTION
MODEL SELECTION
UNSUPERVISED NEURAL NETS
SEGMENTATION OF THE ONLINE SHOPPING MARKET
SECONDARY PROFILES
GRAPHICAL MODELS
SEGMENT-SPECIFIC FACTOR INTER-RELATIONSHIPS

Figure 1.1: *Graphical representation of the methodology put forward throughout this thesis.*

According to this schema, the main goal of the thesis is to provide answers to the following questions:

- *A latent variable structure is to be obtained from the observable variables that refer to the Web users' opinions of the online shopping channel and of the online shopping vendors. Is this structure explainable in terms of key factors identified in previous qualitative studies?*
- *Which of those latent factors contain the necessary information to best predict Web users' propensity to buy on-line?*
- *Are variables, such as age, household income, and Web usage patterns, commonly used in market research, good predictors of that propensity?*
- *To what extent can the propensity to buy online be inferred, using quantitative methods, from a selection of the latent factors shown to be the best predictors of such a propensity and how does it compare with the prediction obtained from the complete set of factors?*
- *Can the e-commerce consumer market be segmented using latent factors as segmentation variables, in such a way that the resulting segments have a sensible managerial interpretation in terms of those factors and in terms of secondary information?*
- *Is the neural network-based unsupervised model proposed in this thesis suitable for the task of segmentation of the online market?*
- *Can graphical modelling provide further insight into the relational structure of the factors for each of the market segments?*

Although not being part of the goals set in this section, the concluding part of this thesis contains suggestions for the integration of the proposed methodology with state-of-the-art personalization Web-technologies.

1.3. Novel contribution

The main novelties provided by this thesis can be summarised as follows:

- Definition of a systematic, quantitative methodology for the characterization and segmentation of the business-to-consumer e-commerce market, with a focus on the modelling of the propensity to buy online.
- Description of the latent dimensions underlying consumers' opinions on on-line shopping and online vendors, using quantitative methods.
- Application of the Automatic Relevance Determination model selection methodology for Bayesian Neural Networks (BNN) to a marketing problem, including a principled method of marginalisation, for BNNs, to the empirical prior to deal with the use of class-unbalanced data sets.
- Qualitative and quantitative defense of the *tandem approach* model to market segmentation, based on the direct relation of the clusters' definition with the propensity to buy on-line, as a profit-optimization process.
- Proposition of the GTM as a principled model for market segmentation and market data visualization. The development of the GTM model includes three novelties: a) The proposition of an entropy-based measure to quantify and compare the class-discriminatory capabilities of trained maps; b) The definition of the *Cumulative*

Responsibility as a measure that simultaneously provides information on the mapping distortion and cluster definition and quantification; c) The definition of the *Selective* GTM *Smoothing* model, which includes multiple regularization terms associated to the basis functions, as part of the model training.

- Definition of an e-commerce market segment-specific independence structure using graphical models.

1.4. Review of the literature

The review process involves a compilation of the literature concerning most of the problems dealt with in this thesis. First, a general survey of business applications of neural networks is provided. It focuses on applications for market segmentation and neural networks in e-commerce. Secondly, some literature on e-commerce, relevant to the thesis, is briefly summarised. Most of this literature is described in more detail as part of the corresponding chapters.

1.4.1. Neural networks in business

Although neural networks originated in mathematical neurobiology, the rather simplified practical models currently in use have moved steadily towards the field of statistics, sacrificing neurological rigour for mathematical expediency. Writing a fully comprehensive survey of business applications of neural networks is, at present, impracticable due to the range of applications and their number. For that reason, the focus is on the period of time 1992-98, during which the statistical context within which neural networks are now applied evolved towards a principled and practical framework (Vellido *et al.*, 1999a).

1.4.1.1. Neural network models

Figure 1.2 summarises the neural network models utilised in the papers reviewed by Vellido *et al.* (1999a). This figure offers a singular feature: 74 out of 93 papers rely on the use of the Feedforward Multilayer Perceptron (MLP) trained by back propagation (BP). It also shows that only one study resorts to single cell Networks (Glorfeld and Hardgrave, 1996), and barely 14 papers accomplish their work by means of unsupervised techniques, mainly Self-Organizing Maps.

The predominance of the supervised Feedforward MLP trained by BP with gradient descent search clashes with the fact that it is generally considered a sub-optimal and rather inefficient technique (Piramuthu *et al.*, 1994; Lenard *et al.*, 1995). Even though, such a model has become a standard of operation, even a byword for supervised Neural Networks. Only a few authors opt for unsupervised models and for most of them (Martin-del-Brio and Serrano-Cinca, 1993; Chen *et al.*, 1995; Serrano-Cinca, 1996; Lewis *et al.*, 1997; Back *et al.*, 1997; de Bodt *et al.*, 1998; Petersohn, 1998; Kiviluoto, 1998b; Brockett *et al.*, 1998; and Ha and Park, 1998) the choice is the Self-Organizing Map and its variants.

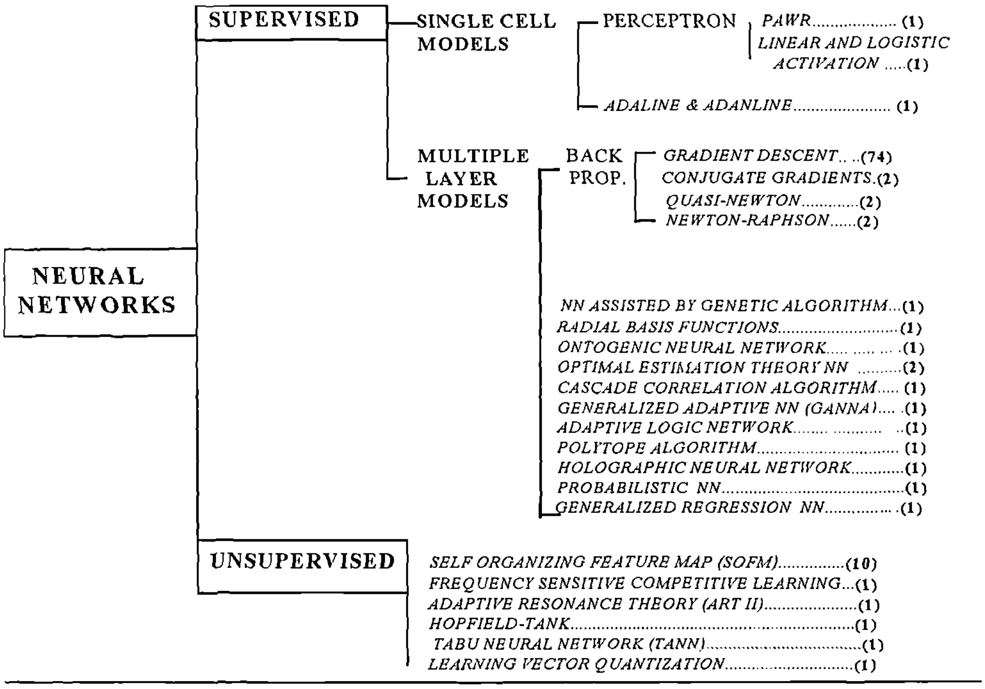

NEURAL NETWORKS
SUPERVISED
SINGLE CELL MODELS
PERCEPTRON
PAWR.................. (1)
LINEAR AND LOGISTIC ACTIVATION .....(1)
ADALINE & ADANLINE.................... (1)
MULTIPLE LAYER MODELS
BACK PROP.
GRADIENT DESCENT.. ..(74)
CONJUGATE GRADIENTS.(2)
QUASI-NEWTON............(2)
NEWTON-RAPHSON......(2)
NN ASSISTED BY GENETIC ALGORITHM...(1)
RADIAL BASIS FUNCTIONS.........................(1)
ONTOGENIC NEURAL NETWORK.... ....(1)
OPTIMAL ESTIMATION THEORY NN ....(2)
CASCADE CORRELATION ALGORITHM..... (1)
GENERALIZED ADAPTIVE NN (GANNA).... .(1)
ADAPTIVE LOGIC NETWORK....... ....(1)
POLYTOPE ALGORITHM.............................. (1)
HOLOGRAPHIC NEURAL NETWORK...........(1)
PROBABILISTIC NN.......................................(1)
GENERALIZED REGRESSION NN............. .(1)
UNSUPERVISED
SELF ORGANIZING FEATURE MAP (SOFM)..............(10)
FREQUENCY SENSITIVE COMPETITIVE LEARNING...(1)
ADAPTIVE RESONANCE THEORY (ART II)....................(1)
HOPFIELD-TANK..(1)
TABU NEURAL NETWORK (TANN)..............................(1)
LEARNING VECTOR QUANTIZATION..........................(1)

Figure 1.2: *Summary of neural network models applied in the reviewed studies. It shows that there are "de facto" standards for the implementation of neural networks in business applications. The methods differ mostly in their approach to optimisation rather than functionality, as seen, for instance, in the rightmost column corresponding to the MLPs.*

1.4.1.2. Main contributions of neural network techniques

The rationale behind this sub-section and its summary, *table 1.1*, is to extract the main contributions of neural networks to the reviewed papers (Vellido *et al.*, 1999a), avoiding cumbersome detail.

Most of the papers in which different techniques are compared report that neural networks outperform (statement 1, *table 1.1*), or perform similarly (statement 2), to the rest of the methods. Some authors are concerned about what are seen as "wild claims" surrounding the application of neural networks, and concerns are raised about “computer scientists who are venturing into areas which are normally the province of statisticians and forecasters [...] applying neural networks to statistical problems which are outside their primary areas of expertise” (Chatfield, 1993, 1995). Other authors affirm that “the statistical tests used by researchers in Management Science and Econometrics have been almost ignored” (Curry and Morgan, 1996). On the other hand, Refenes (1994) defends neural networks as something much more serious than a “passing fad”, arguing in favour of the cross-fertilisation between research areas, as “there is much to be gained for both computer scientists and statisticians by not treating forecasting, classification and pattern recognition problems as exclusive provinces”.

As inferred from statement 5 in *table 1.1*, neural networks are a powerful analytical tool on their own, but their performance is expected to benefit from the assistance of other techniques, the combination with them or its inclusion as part of more complex models. In other words, we can *hybridise* them, use them in a *mixture* of models, or *integrate* them in a more general schema.

Applications of neural networks in “real world” scale are scant, as *table 1.1* reveals. Even though quite a few companies in diverse areas have nurtured neural network prospective studies, not so many make use of neural networks in every-day-business life (Borowsky, 1995; and Robins, 1993a,b). Moreover, data suppliers are usually private companies operating in a tough competitive environment, reticent to share any expert knowledge (Some examples in Altman *et al.*, 1994; Dasgupta *et al.*, 1994; and Desay *et al.*, 1996).

MAIN CONTRIBUTIONS OF NEURAL NETWORKS	Total
1. Neural networks yield better results than other techniques when compared to them.	**45**
2. Neural networks yield similar results to other techniques when compared to them.	**14**
3. Neural networks yield worse results than other techniques when compared to them.	**3**
4. One type of neural networks outperform others.	**18**
5. An integrated/mixture/hybrid model, including neural networks improves the results of the study.	**21**
6. Neural networks are deemed as promising for future developments of the application.	**26**
7. Neural networks are shown to offer new insights into the application.	**29**
8. Neural networks are utilised in a "real-world" case of the application.	**9**

Table 1.1: *Main contributions of neural network techniques to their applications.*

1.4.1.3. Frequently quoted advantages and disadvantages of neural networks

Neural network methodologies are being applied to a whole spectrum of business related disciplines. Which are the reported advantages of this technique and which are the shortcomings that researchers have found through the path of their work?: *Table 1.2* provides a summary of both. Those comments too specifically related to the type of application, linked to commercial software packages or inherent to the model have been avoided.

The advantages

Despite the self-explanatory nature of the table, it is worth stressing that, globally, advantages have been more frequently remarked than disadvantages (132 quotations versus 94). Three advantages have been pointed out more than 20 times: The suitability of neural networks to handle incomplete, missing or noisy data; being a non-parametric method, not requiring any *a priori* assumptions about the distribution and/or mapping of the data; and their demonstrated capability to map any complex non-linearity and/or approximate any continuous function. Given the behavioural nature of the data analysed in this thesis, the latter feature is likely to become specially relevant.

The disadvantages

Table 1.2 shows how the lack of explanatory capabilities is considered as the main shortcoming of the application of neural networks. The adduced incapacity to identify the relevance of independent variables and to generate a set of rules to express the operation of the model, makes neural networks to be usually deemed as “black boxes”. Both problems are addressed in turn.

Throughout the review (Vellido *et al.*, 1999a), the lack of formal techniques for non-linear methods to assess the relative relevance of independent variables (Tam and

Kiang, 1992; Jo *et al.*, 1997) is frequently stressed. As a matter of fact, many studies resort to selection techniques associated with linear methods, so that nothing prevents for different sets of variables to be more relevant in non-linear terms. Nevertheless,

ADVANTAGES of NEURAL NETS	Total	DISADVANTAGES of NEURAL NETS	Total
NNs are able to learn any complex non-linear mapping / approximate any continuous function.	**31**	NNs lack theoretical background concerning explanatory capabilities / NNs as "black boxes"	**28**
As non-parametric methods, NNs do not make *a priori* assumptions about the distribution of the data / input-output mapping function.	**30**	The selection of the Network topology and its parameters lacks theoretical background / It is still a "trial and error" matter.	**21**
NNs are very flexible with respect to incomplete, missing and noisy data / NNs are "fault tolerant"	**29**	Neural Networks learning process can be very time-consuming	**11**
Neural Network models can be easily updated / are suitable for dynamic environments.	**15**	Neural Networks can overfit the training data, becoming useless in terms of generalisation.	**10**
NNs overcome some limitations of other statistical methods, while generalizing them.	**15**	There is no explicit set of rules to select a suitable NN paradigm / learning algorithm.	**8**
Hidden nodes, in feed-forward supervised NN models can be regarded as latent / unobservable variables.	**5**	NNs are too dependant on the quality / amount of data available.	**6**
NNs can be implemented in parallel hardware, increasing their accuracy and learning speed.	**4**	NNs can get stuck in local minima / narrow valleys during the training process.	**5**
Neural Networks performance can be highly automated, minimizing human involvement.	**3**	NN techniques are still rapidly evolving and they are not reliable / robust enough yet.	**3**
NNs are specially suited to tackle problems in non-conservative domains.	**3**	NNs lack classical statistical properties. Confidence intervals and hypothesis testing are not available.	**2**

Table 1.2: *Main reported advantages and disadvantages of neural network techniques.*

several authors make use of alternative methods: Genetic Algorithms, as a type of unstructured search, are increasingly being used to assist neural networks in the task of variable selection; an example in this review is provided by Back *et al.* (1996). Decision Trees have been used by Lee *et al.* (1996) for the same purpose. In this thesis, an inherently non-linear and statistically sound method for variable selection, in the context of Bayesian feed-forward neural network training, will be deployed.

Several techniques for automatic rule extraction from trained artificial neural networks are reviewed in Andrews and Diederich (1996). In this review, only Setiono *et al.* (1998) contribute an explicit neural network model for rule extraction. Furthermore, several attempts have been made to integrate neural networks and Expert Systems: a synergistic effect between them is expected, as "Expert Systems are characterised by the capability of explaining its own reasoning process" (Deng, 1993). Some examples in this review are: Deng (1993), Jung and Burns (1993), Zeleznikov *et alia* (1996), and Willems and Brandts (1997); a comprehensive account of these models is provided by Schocken and Ariav (1994).

1.4.1.4. Neural networks in market segmentation

The application of neural networks to market segmentation is a new and promising research area: As remarked in Dasgupta *et al.* (1994), "Neural Network models have seldom been evaluated with respect to market response applications". Only a few papers, sparsely published, have broached this problem.

Customer grouping is a usual challenge for a wide variety of marketers: Davies *et al.* (1996) analyse how customers develop positive or negative attitudes towards bank Automatic Teller Machines (ATMs): different consumer groups are expected to show different expectations of this service so that a clustering strategy should be beneficial for the bank. Also in the financial area, Dasgupta *et al.* (1994) characterise potential individual-level customer segments in terms of lifestyle variables (deemed to be "substantially more difficult" than profiling those segments based on demographic or geographic characteristics). "Consumer groups with a markedly different pattern of benefits sought should be considered to be 'natural' segments in the market": such is the point of departure for Mazanec (1992) to discriminate tourists within a *benefit* approach (what tourists expect from their holidays). A highly practical case study is developed by Balakrishnan *et al.* (1996) who, using coffee brands switching probabilities derived from scanner data at a sub-household level, carry out a six-segment classification study. Setiono *et al.* (1998) resort to a rule-extraction Neural Network to target companies for IT promotion. Finally, Fish *et al.* (1995) inspect the possibility of clustering managers-customers purchasing from a firm.

The inspection of the neural networks utilised in these papers reveals some surprising features. Even though unsupervised models would appear to suit the clustering strategies better, only one out of the six studies reviewed resorts to them: The *Frequency Sensitive Competitive Learning* (FSCL) algorithm, a Vector Quantization technique, used in Balakrishnan *et al.* (1996). The remaining papers make use of supervised feed-forward MLPs trained by Back Propagation and Gradient Descent

(Davies *et al.*, 1996; Fish *et al.*, 1995; Dasgupta *et al.*, 1994; Setiono *et al.*, 1998), or similar alternatives (Mazanec, 1992). Two of the studies do not validate the results with a hold-out sample. The justification in Davies *et al.* (1996) stems from the use of the weight matrices to assess the existence of segments. In Mazanec (1992), the reason for the lack of testing appears to be the reduced size of the data sample, together with the pre-clustering process involved. No study accomplishes a validation beyond a simple split into independent training and test subsets: generalisation seems to be secondary with respect to cluster interpretability. The rather small data sample sizes (no study gets to use more than one thousand examples) reaffirm the relevance of this feature.

Neural networks are compared to other statistical models: Linear Discriminant Analysis in Fish *et al.* (1995), Dasgupta *et al.* (1994), and Mazanec (1992); Logistic Regression in Fish *et al.* (1995) and Dasgupta *et al.* (1994). Only Balakrishnan *et al.* (1995) go beyond "competition" to try a "collaboration" scheme: The FSCL algorithm is used to provide the "clustering seeds" for the K-Means algorithm to refine the segments: such hybridisation leads to an important improvement of the results.

1.4.1.5. Neural networks in e-commerce

Several market research companies have reported the use of neural network techniques for support in the area of electronic commerce(e.g., Decider e-Commerce™ from Neural Technologies®, OptiMatch™ from Neural Inc.®, etc.) This research reveals that no academic application of neural networks to the area has been reported. Even though, some authors have echoed the possibilities of these techniques in the field: Wallin (1999) proposes the use of artificial neural networks for consumer's commercial behaviour personalization or, in other words, for user modelling. From a different perspective, Scharl and Brandtweiner (1998) discuss the use of neural networks in the context of intelligent mobile agents, which promise to "radically change inherent characteristics of electronic commerce".

1.4.2. Literature on e-commerce

A brief summary of some of the studies in the area of e-commerce that will be used as references for the experiments throughout the thesis is provided next.

1.4.2.1. Factors underlying online shopping behaviour

Research on Internet consumer behaviour is still very much in its infancy and only a few authors have started to map the structure of the factors influencing consumers' attitudes towards on-line shopping.

Jarvenpaa and Todd's (1996/97) seminal study describes a salient structure for consumers' attitudes towards online shopping. It consists of four main groups of factors, inspired by traditional retail patronage, and adapted to the Web shopping context. The factors can be summarized as:

- ***Product perceptions***, including the dimensions of*price, quality* and *variety.*

- ***Shopping experience,*** described as a mixture of *effort, compatibility* and *playfulness.*
- ***Customer service,*** including *responsiveness, assurance, reliability, tangibility* and *empathy.*
- ***Consumer risk,*** split into *economic, social, performance* and *personal* risks.

In another important study, Hoffman *et al.* (1999) describe, from the point of view of consumers' information privacy concerns, some other factors that can greatly influence on-line shopping adoption: *Environmental control* and *secondary use of information control*:

- ***Environmental control*** is defined as the "consumer's ability to control the action of other people in the environment during a market transaction". It implies *security* and *economic risk* and it is proportional to the level of *anonymity.*
- ***Secondary use of information control*** is defined as "the consumer's ability to control the dissemination of information related to or provided during such transactions or behaviours to those who were not present".

1.4.2.2. Prediction of online purchasing behaviour

There is a scarcity of quantitative research addressing the selection of variables influencing online shopping behaviour and the prediction of the propensity to buy online. To the best of the authors' knowledge, this has only been accomplished in three main studies.

In the first of them (Jarvenpaa and Todd, 1996/97), four main salient factors constitutive of online shopping behaviour were related, using linear regression, to two dependent variables: *attitudes toward shopping* and *intention toward shopping*. In the second study (Novak and Hoffman, 1997), data from the GVU's 7th WWW users survey were used to investigate the relationship between the concept of *flow construct* and consumer attitudes on-line, as well as in other traditional marketing channels. The variable selection was accomplished by significance analysis of the interaction parameters in logistic regression, and the dependent variables were *search* and *purchase* behaviour for different product categories. Bellman *et al.*, in a recent study (1999), explicitly address the issue of the identification of the factors influencing online buying behaviour. Stepwise variable selection in a logistic regression model was used in order to find the variables, from the raw data, that best predicted actual purchases. The following variables, in decreasing order of relevance, were found to be the best predictors: "Looking for product information on the Internet", "Number of months online", "Number of e-mails received per day", "Work on the Internet in their offices every week", "Read news online at home every week", "Total household working hours", "Click on banners" and "Agree that Internet improves productivity". The logistic regression model predicted with a 66% of accuracy whether the survey respondents bought anything online or not.

1.4.2.3. Market segmentation in the area of electronic commerce

Market segmentation has been highlighted as one of the important and necessary avenues of research needed in the field of electronic commerce (Chang, 1998) and is an issue that Internet vendors cannot ignore (O'Connor and O'Keefe, 1997). Despite this, only a few studies have addressed the subject: Gordon and De Lima-Turner (1997) examined how Internet users make a trade-off among attributes associated with Internet advertising policy, within the framework of a theory of *social contract* previously applied to direct mail marketing.

The inherently world-wide structure of the Internet matches naturally with an international market segmentation perspective. McDonald (1996) explored the Internet usage motivations of consumers of several countries, aiming to find segments that transcended national boundaries.

Hoffman *et al.* (1996), in a general study of the baselines for the commercial development of the Internet, carried out some basic segmentation according to the cross-classification of Internet and Web usage frequencies. The resulting segments were profiled using demographic characteristics and computer usage behaviour.

1.5. Publications related to this chapter

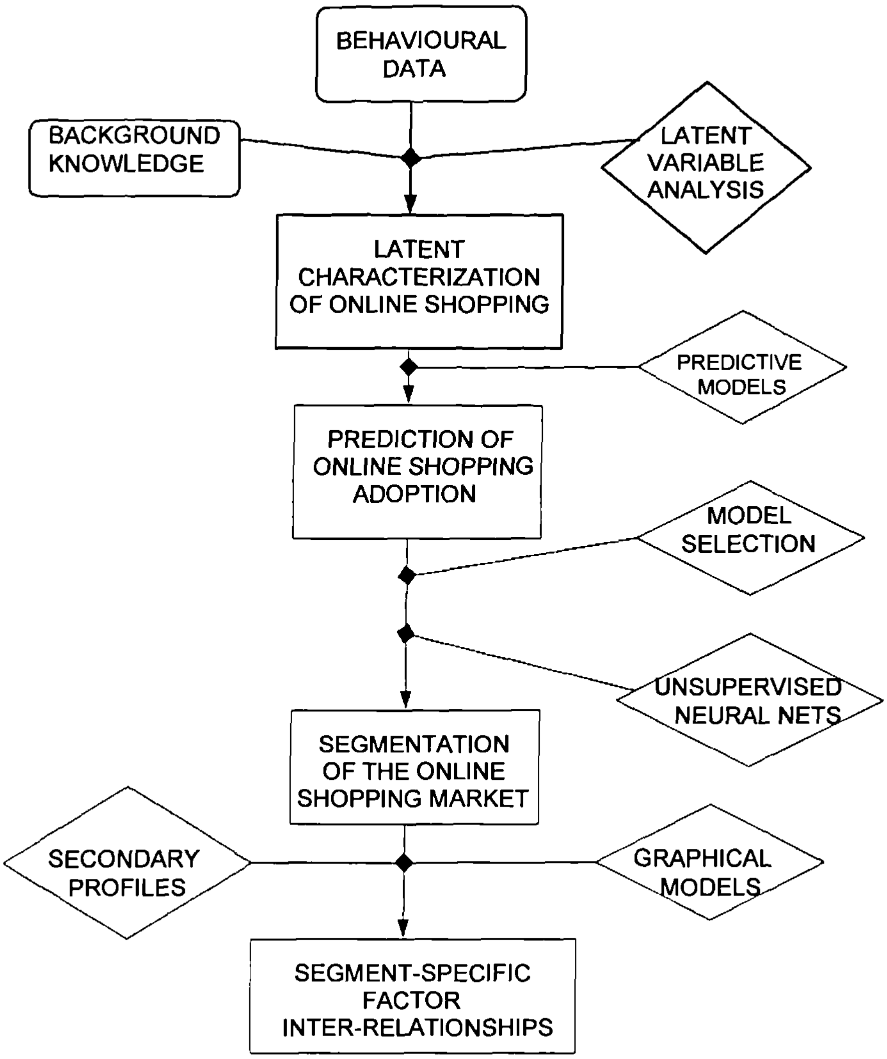
Vellido, A., Lisboa, P.J.G, and Vaughan, J. (1999) Neural networks in business: a survey of applications (1992-1998). *Expert Systems with Applications,* **17(1),** 51-70.

PART 2

Latent variable characterization of the online shopping market

The subterranean miner that works in us all, how can one tell whither leads his shaft by the ever shifting, muffled sound of his pick?

"Moby Dick". Herman Melville


BEHAVIOURAL DATA
BACKGROUND KNOWLEDGE
LATENT VARIABLE ANALYSIS
LATENT CHARACTERIZATION OF ONLINE SHOPPING
PREDICTIVE MODELS
PREDICTION OF ONLINE SHOPPING ADOPTION
MODEL SELECTION
UNSUPERVISED NEURAL NETS
SEGMENTATION OF THE ONLINE SHOPPING MARKET
SECONDARY PROFILES
GRAPHICAL MODELS
SEGMENT-SPECIFIC FACTOR INTER-RELATIONSHIPS

Preface

The exponential growth of Internet adoption is continuously raising the expectations about the potential size of the worldwide market for e-commerce. Nevertheless, the online market potential is unlikely to be fully realized without a wider exploration of the consumers' needs and expectations. This wider exploration justifies a data based analysis of online customer shopping behaviour, in order to provide a sound empirical basis for business decision making.

In this part of the thesis we aim to find a parsimonious non-observable description of the Internet users' opinions of online shopping and online vendors, using latent variable analysis techniques. The resulting latent factors will be interpreted according to existing literature in the field. This factor description, explainable in meaningful and managerially operative terms, is expected to be a valuable tool in the hands of e-commerce market analysts, but also the foundation for the rest of the analyses carried out in this thesis.

Chapter 2

Characterization of the latent dimensions underlying online shopping behaviour

2.1. Introduction

There is still no quantitative framework for the analysis of online consumer shopping behaviour. Thus far, little formal research has been devoted to the investigation of the underlying factors characterising consumer behaviour online and the existing studies are mostly qualitative and exploratory in nature. The present chapter provides the first part of such a quantitative framework of analysis, based on well established statistical principles, aiming to address this issue and begin to fill the gap of quantitative research that exists in the area.

The modeling of data in high dimensional spaces is likely to be affected by problems of curse of dimensionality and overfitting (leading to a lack of generalizability of the models). Furthermore, the results of the analyses based on a parsimonious description of the data, stemming from dimensionality reduction procedures, are expected to be more interpretable and operational in managerial terms. There are two main approaches to dimensionality reduction: either the selection of a subset of the original, observable variables, or the generation of a group of new variables, non-observable or latent, as a combination of the original ones (Hand, 1997). Factor analysis, the technique for latent variable extraction used in this study, follows the latter approach. The factor analysis of the data can be further justified on the basis of the following arguments (Green and Krieger, 1995):

- It can help to overcome some of the limitations associated with survey data: Presence of noise, poorly measured variables, inadequate selection of survey items in terms of balance across studied constructs.
- The resulting factor structure can be interpreted in terms which are not explicit in the observable data, whilst more operative in the business context.

The proposed framework explores online shopping behaviour, from a consumer-centered point of view, by characterizing the latent structure of online consumers' opinions and expectations. The success of the application would entail the provision of a, yet unknown, parsimonious latent description of the Internet user opinions of on-line shopping.

The question that this chapter aims to investigate can be formally stated as follows:

> *A latent variable structure is to be obtained from the observable variables that refer to the Web users' opinions of the online shopping channel and of the online shopping vendors. Is this structure explainable in terms of key factors identified in previous qualitative studies?*

The data utilized in this study consists of questionnaires about Internet users' opinions of Web vendors and on-line shopping. The completed questionnaires are publicly available from the 9th GVU's WWW users survey, produced by the Graphics, Visualization & Usability Center (Kehoe and Pitkow, 1998). This chapter is organized as follows: First, the existing literature in the area will be reviewed in some detail. Next, an overview of the data used in the study will be provided. This will be followed by the extraction and interpretation of the latent variables and a section of conclusions.

2.2. Previous research on factors influencing online shopping adoption

Research on Internet consumer behaviour is still very much in its infancy and only a few authors have started to map the structure of the factors influencing consumers' attitudes towards on-line shopping.

Jarvenpaa and Todd's (1996/97) seminal study, based on an open-ended survey with a sample of 220 primary household shoppers, describes a salient structure for consumers' attitudes towards online shopping. It consists of four main groups of factors, inspired by traditional retail patronage, and adapted to the Web shopping context. The factors can be summarized as:

- ***Product perceptions***, including the dimensions of *price, quality* and *variety.*
- ***Shopping experience*** described as a mixture of *effort, compatibility* and *playfulness*. In the online context, *effort* is more mental than physical, and can involve ease of use and ease of placing and cancelling orders. *Compatibility* refers

to the consumers' lifestyle and shopping habits. Finally, *playfulness* can be described in an Internet context, by making use of the *flow construct* concept (Hoffman and Novak, 1997).

- ***Customer service*** includes *responsiveness, assurance, reliability, tangibility* and *empathy. Responsiveness* "concerns how well prepared merchants are to meet the diverse needs of shoppers during the different phases of shopping". *Assurance* is "the degree to which the service provider instills confidence in customers". *Reliability* is "the degree to which the service provider can be counted on to deliver on his or her promises". The concept of *tangibility* refers to the ability of the vendor to replace the real product with an appealing and information-rich virtual substitute. Finally, *empathy* is defined as the degree to which the vendor is able to adapt to the individual needs of the consumer.
- ***Consumer risk*** is split into *economic, social, performance* and *personal* risks. The *economic risk* "stems from the possibility of monetary loss associated with buying a product". *Personal risk* has to be understood in terms of the concept of *environmental control* (Hoffman *et al.*, 1999). *Performance risk* involves the "perception that a product or service may fail to meet expectations". Finally, *Social risk* is concerned with both the consumer's self-perception, and that of their peers.

Some of the factors described in these groups can be explored further. The marketing concepts of product and product perception have evolved rapidly in recent times, becoming increasingly *information-based* (Brannback, 1997). Considering the success of *high-involvement, information-rich,* products online (Schwartz, 1997) a further dimension, the *information richness* of products, can be added to those described in the study by Jarvenpaa and Todd (1996/97). The concept of *information richness* can not be dissociated from the factor of *tangibility*. In turn, this lack of *tangibility*, at the point of purchase, heightens the uncertainty of purchasing something which has yet to be physically experienced, and removes the potential for consumers to utilize price-quality assessment as a risk reduction strategy to alleviate both *economic* and *performance risk*. This highlights the fact that some of the factors described in the study by Jarvenpaa and Todd (1996/97) overlap in a Web context. For instance *Assurance*, again part of the ***customer service*** group, has to be related to the ***consumer risk*** group. *Economic risk* and *performance risk* are also closely related, since there is a potential economic loss involved in an on-line purchasing situation whereby payment and delivery are not concurrent and what was purchased may not be what was expected.

In another important study, Hoffman *et al.* (1999) describe, from the point of view of consumers' information privacy concerns, some other factors that can greatly influence on-line shopping adoption: *Environmental control* and *secondary use of information control*:

- *Environmental control* is defined as the "consumer's ability to control the action of other people in the environment during a market transaction". It implies *security* and *economic risk* and it is proportional to the level of *anonymity*.

- *Secondary use of information control* is defined as "the consumer's ability to control the dissemination of information related to or provided during such transactions or behaviours to those who were not present".

The findings of these mainly qualitative studies are to be used as a benchmark for the identification and description of the factors stemming from the experiments in section 2.4.

2.3. Description of the data

This study is based upon publicly available data from the web-based *9th GVU's WWW User Survey* (Kehoe and Pitkow, 1998), and, more specifically, its *"Internet Shopping (Part 1) Questionnaire"*. From its first two questions, regarding *"general opinion of using the WWW for shopping as compared to other means of shopping"* and *"opinion of Web-based vendors compared to vendors in other forms of shopping"*, 44 items were selected (to be identified, from this point, as data set A, and fully described in the appendices). The size of the data sample utilized is 2180 individuals.

GVU's WWW user surveys are characterized by the self-selection of its respondents, which entails a risk of biased samples that would not reflect the real Internet population. Unfortunately, random sampling is not possible in a decentralized medium such as the Internet. The comparison with other WWW user surveys reveals a bias in the number of respondents toward experienced Internet users, but also very well balanced demographics. This bias is not necessarily an undesirable feature of the survey: as stated in Rogers (1998), “for many marketing needs, this bias is exactly what is desired of the data”. This can be understood in the light of some of the results of the survey, indicating that the more experienced Internet users spend more money on-line and, therefore, represent an interesting target segment, whose characterization would prove useful to marketers.

2.4. Exploration of the latent variable structure

The research question outlined in the introduction will be answered in this section: is it possible to obtain a sensible latent structure from the data utilized in this study?

The latent dimensions underlying Web users’ opinions with regard to online shopping, are extracted using the technique of factor analysis. In this way, high-dimensional data sets are described in the most parsimonious way, becoming more manageable and operative; furthermore, the latent variables will reveal underlying attitudes towards online shopping that would not be evident otherwise. The factor extraction is accomplished using the Maximum Likelihood method, as it is the only systematic and theoretically sound procedure capable of producing a generalised likelihood ratio test of goodness of fit (Krzanowski, 1996) that can be used as a guidance for the selection of the number of factors of the final model. The factors are rotated using the *Varimax* procedure to a final orthogonal structure. It is generally agreed that where a simple structure with orthogonal factors can be found, this is the most efficient method. The Kaiser-Meyer-Olkin (KMO) measures of sampling adequacy fall between the categories of "middling" and "meritorious", according to the ranks devised by Kaiser (1974); the validity of the factor model is therefore supported.

The rotated factor structure for *data set A* is shown in *table 2.1*: Only loadings over (or close to) 0.3 are shown. Also, the cumulative variance results are shown in figure 2.1. The maximum-likelihood goodness of fit test suggests a 9-factor solution. It has to be borne in mind that one of the goals of this chapter is to compare this factor solution stemming from a quantitative analysis of the data, with factors mentioned in previous qualitative research. Therefore, we now attempt to describe this 9-factor solution according to the concepts in Hoffman *et al.* (1999) and Jarvenpaa and Todd (1996/97), summarized in section 2.2.

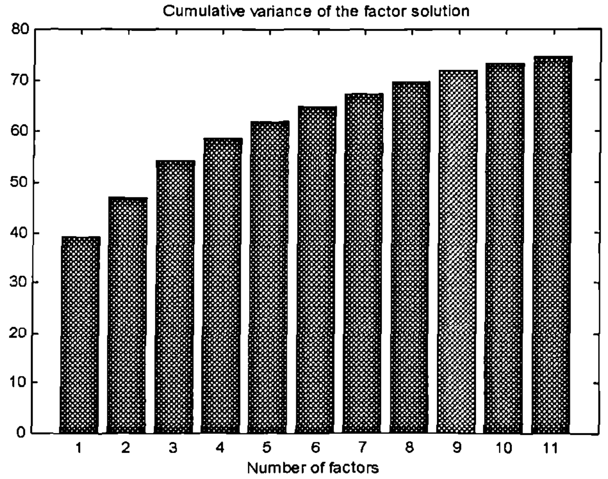


Figure 2.1: *Cumulative variance of the factor solution. The choice of the optimal number of factors is determined by the maximum-likelihood goodness of fit test, which tests for the significance of the residuals after each additional factor is extracted. The smallest number of factors with non-significant residuals in this case is 9.*

Factor 1 is general in terms of loading variables, but has a definite interpretation as *shopping experience,* in terms of *compatibility, control* over the shopping experience

Rotated Factor Matrix

	Factor								
	1	2	3	4	5	6	7	8	9
easyfindv						.581			
easypay									.340
infoupdat			.470						
eascomp			.381			.388			
accessop			.604						
riskypay		-.673							
gatherinfo	.324		.301						
receitime							-.587		
quickinfo			.311			.604			
timedelivr							.526		
easplaceo									.595
cuserv&a			.334				.440		
quicplacor									.585
moreinfo			.455			.376			
easret&ref							.500		
expertopin			.647						
effenhanc	.692								
trustsafe		.717							
easylearn					.444				
payaccex				-.466					
shoprisky		-.759							
itemselect	.501					.311			
afforequip				.635					
improimag								.683	
trustinfo		.620							
shopvisibl									
compsitua	.570	.327							
shopcontrl	.662								
shopsafe	.321	.768							
mentaleffr					-.385				
affrpayfee				.476					
quickshop	.506								
prestige								.702	
shopcomp	.503								
afforequip				-.495					
shopabilit	.608								
easygene					.606				
betterpric	.432								
shopclear					.505				
tryshop					.324				
fitstylesho	.662	.325							
highprofile								.350	
shoprodu	.715								
experimnt									

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

Table 2.1: Rotated factor structure for data set A. The consistency of this factor solution was tested by splitting the data sample into three mutually exclusive sub-samples. The resulting factor solutions were consistent with the overall results, in the sense that loadings had overwhelmingly the same structure for each sub-sample as for the complete data set.

and *convenience.* ***Factor 2*** is very homogeneous and refers to *consumer risk* perception from the point of view of *environmental control*, involving perceptions of *trust* and *security*. ***Factor 3*** turns out in the shape of *customer service*, along the dimensions of *responsiveness* and *empathy;* emphasis is placed in the *information richness* provided by the vendor, which also relates to aspects of *product perception.* ***Factor 4*** is undoubtedly perception of *affordability,* concerning the expenses associated with the necessary equipment, connection, etc. Interestingly, further reductions of the factor model (not reported in this study) indicate that this factor would merge with ***factor 5***, which loads in items that describe *effort* as *ease of use* within the *shopping experience* dimension, thus revealing a correlation between the perceptions of economic effort and user effort. ***Factor 6*** complements *factor 3* above, as it focuses on *information richness* in terms of *product perception* of *variety*. ***Factor** 7* mixes the concepts of *assurance* and *reliability* in *customer service* with the perception of *performance risk*. ***Factor 8*** is also concerned with *consumer risk* but from the point of view of the consumers' *image*, and could be labeled as *elitism.* Finally, ***factor 9*** addresses another aspect of the *shopping experience: Effort* associated to the vendors' performance; it is, thus, related to *responsiveness* and *empathy.* These 9 factors are summarized in *table 2.2*.

The application of factor analysis to the data described in section 2.3 resulted in a set of factors which form the basis to assemble characteristic profiles of Internet users. Each factor was interpreted using the factor loadings onto the questionnaire items and the concepts proposed in previous qualitative studies.

FACTOR	DESCRIPTION	ATTRIBUTES
1	*Shopping experience: Compatibility*	*Control and convenience*
2	*Consumer risk perception /* *Environmental control*	*Trust and security*
3	*Customer service*	*Responsiveness and empathy/* *Information richness*
4	*Affordability*	--
5	*Shopping experience: Effort*	*Ease of use*
6	*Product perception*	*Variety: Information richness*
7	*Customer service / Consumer risk*	*Assurance and reliability /* *Performance risk*
8	*Consumer risk: Image risk*	*Elitism*
9	*Shopping experience /* *Customer service*	*Effort /* *Responsiveness and empathy*

Table 2.2: *Descriptive summary of the factor structure from table 2.1.*

2.5. Conclusion

The factor extraction experiments have quantitatively demonstrated the existence of a parsimonious latent structure underlying consumers' opinions of online shopping, explainable in meaningful, operative terms which are consistent with the findings of previous qualitative studies. Therefore, the research question outlined in the introduction of the chapter is answered affirmatively. This simple description of online customers' opinions alleviates the intrinsic drawbacks of high-dimensional data sets, and highlights the practical managerial value of the quantitative analysis approach to e-commerce market analysis.

2.6. Publications related to this chapter

1. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Quantitative characterization and prediction of on-line purchasing behaviour: a latent variable approach. *International Journal of Electronic Commerce,* 4(4), 83-104.

2. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Characterizing and segmenting the business-to-consumer e-commerce market using neural networks. In Lisboa, P.J.G., Vellido, A. and Edisbury, B. (Eds.): *Business Applications of Neural Networks.* Singapore: World Scientific, 29-54.

2.7. Appendix

Description of the questionnaire items, from the GVU's *"Internet Shopping (Part 1) Questionnaire"* (Kehoe and Pitkow, 1998) utilized in the study. The label shown in the factor structure matrix (*table 2.1*) precedes each of the items:

Data set "A"

Question: **We are interested in knowing your opinion of Web-based vendors compared to vendors in other forms of shopping. For each of the following items, please indicate your level of agreement:**

easyfindve: It is easier to find a Web-based vendor that sells the item I wish to purchase.

Easypay: Paying for an item purchased is easier with Web-based vendors.

Infoupdat: Web-based vendors are better at providing information about updates on products I've purchased.

Eascomp: It is easier to compare similar items between different Web-based vendors.

Accessop: Web-based vendors are better at providing me easy access to the opinions of other consumers about products I wish to purchase.

Riskypay: It is more risky to make payment to Web-based vendors when purchasing an item.

Gatherinfo: I would prefer to gather purchase related information through Web-based vendors.

Receitime: It takes longer to receive the item purchased from Web-based vendors.

Quickinfo: I can quickly gather information about products and services I wish to purchase from Web-based vendors.

Timedelivr: Web-based vendors deliver orders/services in a more timely manner.

Easplaceo: It is easier to place orders with Web-based vendors.

Cuserv&as: Web-based vendors provide better customer service and after-sales support.

Quicplacor: Placing an order for an item takes less time with Web-based vendors.

Moreinfo: I can gather more information from Web-based vendors about an item I want to purchase.

Easret&ref: Returns and refunds are easier with Web-based vendors.

Expertopin: Web-based vendors are better at providing me easy access to the opinions of experts about products I wish to purchase.

Question: **We are interested in knowing your general opinion of using the World Wide Web for shopping as compared to other means of shopping (...) For each of the following items, please indicate your level of agreement:**

effenhanc: Shopping over the WWW would enhance my effectivity at shopping.

Trustsafe: I would trust online vendors enough to feel safe shopping over the WWW.

Easylearn: Learning to shop over the WWW would be easy for me.

Payaccexp: Shopping over the WWW would be expensive since it would require me to pay for access to the Internet.

Shoprisky: Shopping over the WWW would be very risky.

Itemselect: Shopping over the WWW would allow me to have better item selection in my shopping.

Afforequip: I could afford to buy the equipment needed to shop over the WWW.

Improimag: Shopping over the WWW would improve my image with those around me.

Trustinfo: I would trust an Internet service provider with transmitting personal information necessary for me to shop over the WWW.

Shopvisibl: Shopping over the WWW is not very visible.

Compsitua: Shopping over the WWW would be completely compatible with my current situation.

Shopcontrl: Shopping over the WWW would give me greater control over my shopping.

Shopsafe: Shopping over the WWW would be a safe way to shop.

Mentaleffrt: Shopping over the WWW would require a lot of mental effort.

Affrpayfee: I could afford to pay a monthly fee to an Internet provider in order to shop over the WWW.

Quickshop: Shopping over the WWW would allow me to do my shopping more quickly.

Prestige: People who shop over the WWW have greater prestige than those who do not.

Shopcomp: Shopping over the WWW would be compatible with all aspects of the way I shop.

Afforequip: Shopping over the WWW would require me to purchase equipment which would be beyond my financial means.

Shopability: Shopping over the WWW would improve my shopping abilities.

Easygener: Overall, I believe that Shopping over the WWW would be easy to do.

Betterprice: Shopping over the WWW would allow me to get better prices when shopping.

Shopclear: Shopping over the WWW would be clear and understandable.

Tryshop: I've had a great deal of opportunity to try shopping over the WWW.

Fitstylesho: Shopping over the WWW fits into my shopping style.

Highprofile: People who shop over the WWW have a high profile.

Shoproduc: Shopping over the WWW would increase my shopping productivity.

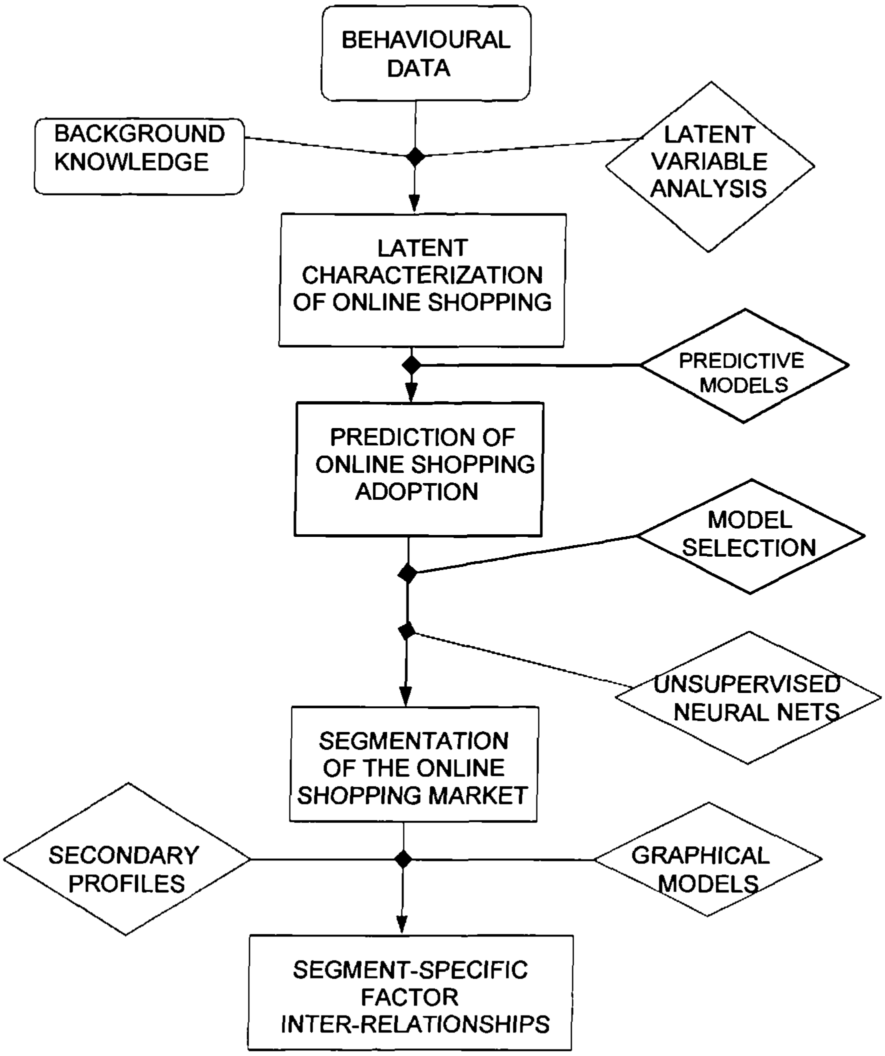
Experiment: I am able to experiment with shopping over the WWW as necessary.

PART 3

Prediction of the propensity to buy online

As long as God knows the truth, it doesn't matter what you tell your customers.

"God is my Broker". Brother Ty


BEHAVIOURAL DATA
BACKGROUND KNOWLEDGE
LATENT VARIABLE ANALYSIS
LATENT CHARACTERIZATION OF ONLINE SHOPPING
PREDICTIVE MODELS
PREDICTION OF ONLINE SHOPPING ADOPTION
MODEL SELECTION
UNSUPERVISED NEURAL NETS
SEGMENTATION OF THE ONLINE SHOPPING MARKET
SECONDARY PROFILES
GRAPHICAL MODELS
SEGMENT-SPECIFIC FACTOR INTER-RELATIONSHIPS

Preface

A parsimonious underlying latent factor description of the Internet users' opinion of online shopping has been obtained in part 2 of the thesis. In part 3, those latent factors are to be used to predict the propensity to buy online.

The aim of **Chapter 3** is two-fold. Firstly, model selection is deployed to unveil which of these factors are the best predictors of the propensity to buy online. Secondly, several inference models are implemented to deal with the quantification of that propensity as a classification problem. Focusing on the Bayesian neural networks as classification models, **Chapter 4** explores the effect of the process of marginalization of the network predictions and the technique of Automatic Relevance Determination on the classification problem results. **Chapter 5** provides a principled method to train the Bayesian neural network in order to deal with the situation of unequal prior class probabilities in the presence of marginalization.

Chapter 3

Quantitative prediction of online purchasing behaviour: A latent variable approach

3.1. Introduction

The thesis now goes one step further in the development of a quantitative framework for the analysis of online consumer purchasing behaviour. As mentioned in the previous chapter, little formal research has been devoted to a quantitative investigation of the underlying factors influencing the propensity to buy online. Likewise, the problem of quantitative prediction of the propensity to buy on-line has barely been tackled, and never resorting to latent variables as predictors.

Online shopping adoption, from a consumer-centered point of view, is now explored in two different ways, using the latent variable description obtained in the previous chapter as a starting point: first, by selecting the latent factors that best predict the propensity to buy online. Secondly, by quantifying the accuracy of the inferences made about that propensity. The latent variable selection will be carried out by logistic discrimination and Bayesian neural network-based methods. The current research reveals that the latter has never been applied to a real-world marketing problem. Its interest rests on its inherently non-linear nature, suited to the non-linearity of the neural network mapping which, in turn, fits the likely non-linearity of the data which are the object of this thesis. The same models plus linear discriminant analysis will be applied to the on-line purchasing prediction/ classification problem.

The success of the model implementation would entail the provision of a selection of factors proven to be predictive of the propensity to buy on-line and, therefore, would prove useful in the design and implementation of marketing strategies.

The questions that this study aims to investigate can be summarized as follows:

> ***Question 1:*** *Which of the latent factors identified in the previous chapter contain the necessary information to best predict Web users' propensity to buy on-line?*
>
> ***Question 2:*** *Are variables, such as age, household income, and Web usage patterns, commonly used in market research, good predictors of that propensity?*
>
> ***Question 3:*** *To what extent can the propensity to buy online be inferred, using quantitative methods, from a selection of the latent factors shown to be the best predictors of such a propensity and how does it compare with the prediction obtained from the complete set of demographic, socio-economic, web usage, and latent variables? Also, is there a significant difference in the predictive performance of the logistic discrimination and neural network models?*

This chapter will continue with a review of the previous literature on quantitative variable selection and prediction in the context of online shopping. This is followed by further information in the data used for the experiments and an outline of the methods adopted for the analyses. Then, the data-based predictive models are deployed and the results reported. The chapter is closed by a section of conclusions.

3.2. Literature on quantitative variable selection and prediction in the context of on-line shopping behaviour

There is a scarcity of quantitative research addressing the selection of variables influencing online shopping behaviour and the prediction of the propensity to buy online. To the best of the authors' knowledge, this has only been accomplished in three main studies.

In the first of them (Jarvenpaa and Todd, 1996/97), four main salient factors constitutive of online shopping behaviour were related, using linear regression, to two dependent variables: *attitudes toward shopping* and *intention toward shopping*. It was concluded that *attitudes* were significantly influenced by *product perception*, *shopping experience* and *consumer risk*, but not by *customer service*. On the other hand, *intention* was influenced by *product perception*, *shopping experience* and *customer service* but not by *consumer risk*. Demographic and socio-economic variables, such as age, education level, household income, years of employment and sex, were shown not to be related to the dependent variable of *attitudes toward shopping*.

In the second study (Novak and Hoffman, 1997), data from the GVU's 7th WWW users survey were used to investigate the relationship between the concept of *flow construct* and consumer attitudes on-line, as well as in other traditional marketing channels. The *flow construct*, in the Web computer-mediated environment, describes the state of mind induced by the Web navigation. The consequences of this flow include "increased learning, increased exploratory and participatory behaviors, and more positive subjective experiences" (Hoffman and Novak, 1997). Web navigation behaviour can be either *goal-directed* or *exploratory*. The first is a "directed search mode (...) in which the consumer is extrinsically motivated to find a particular site or piece of information in a site". It is argued that only experience will lead users to achieve *flow* experiences in a *goal-directed* task such as purchasing online. The variable selection in Novak and Hoffman (1997) was accomplished by significance analysis of the interaction parameters in logistic regression, and the dependent variables were *search* and *purchase* behaviour for different product categories. It was found that *skill* and *challenge*, two antecedents of *flow*, predict online purchasing behaviour in a range of product categories.

Bellman *et al.* (1999) explicitly address the issue of the identification of the factors influencing online buying behaviour. Their study is based on a survey from the Wharton Forum on Electronic Commerce (ecom.wharton.upenn.edu) Stepwise variable selection in a logistic regression model was used in order to find the variables, from the raw data, that best predicted actual purchases. The following variables, in decreasing order of relevance, were found to be the best predictors: "Looking for product information on the Internet", "Number of months online", "Number of e-mails received per day", "Work on the Internet in their offices every week", "Read news online at home every week", "Total household working hours", "Click on banners" and "Agree that Internet improves productivity". The logistic regression model predicted with a 66% of accuracy whether the survey respondents bought anything online or not. The authors also found that demographic data do not seem to influence "whether or not people buy online". It is concluded that "The most important information for predicting shopping habits –online and offline- are measures of past behavior, not demographics".

Where possible, the results of these studies will be compared with those of our own research, which are summarized in section 3.5 of this chapter.

3.3. Description of the data used in this study

The individual factor scores from *data set A*, obtained in the previous chapter, are the starting point of the prediction experiments. For some of these experiments, the factor scores have been complemented with information on *age, household income, years of Internet experience,* and *average of hours a week of Internet usage* (from this point, these four variables to be identified as *data set B*). These variables are likely to be more frequently available to the online marketer. The dependent variable used for the variable selection and classification experiments is binary in nature and responds to the question of whether the respondent has or has not ever purchased online. The size of the data sample utilized is again 2180 individuals, and it is not balanced in terms of class membership: different class-balanced data sets of around 800 individuals were created as part of the n-fold cross validations described in the next sections.

3.4. Selection of the most predictive latent variables

We now try to answer the first and second research questions posed in the introduction: which are the best predictors of the propensity to buy online? The models that will perform that feature selection are first introduced, and this is followed by a summary of the results.

3.4.1. Predictive models

The variable selection, in the context of the prediction of online shopping purchasing behaviour, is carried out by means of Logistic Discrimination and supervised neural network models. Individual factor scores are to be used as the new, non-observable, latent variables, which will undergo a variable selection process within both model frameworks, as described next.

3.4.1.1. Logistic discrimination

Logistic discrimination is a natural extension of linear discriminant analysis for dichotomous dependent variables. Different techniques can be used to assess the inferential significance of individual variables; this study resorts to a stepwise forward likelihood ratio selection method, in which the variable entry is based on the significance score statistic, and the removal on the maximum partial likelihood estimates (Norusis, 1990). The entry probability is set to 0.05, whereas the removal probability is set to 0.1.

3.4.1.2. Neural networks

Neural networks are frequently considered as *black boxes* due to their supposed incapacity to identify the relevance of independent variables in non-linear terms. Nevertheless, in recent years this potential drawback has been addressed: Genetic algorithms and decision trees have been used to assist neural networks in the process of variable selection. Techniques that do not resort to secondary models have also been developed. This study makes use of a statistically sound and non-linear in nature method: Automatic Relevance Determination (ARD), for supervised Bayesian neural networks (MacKay, 1995). A benchmark of different variable selection methods for neural networks (Lisboa *et al.*, 1997) showed ARD to be the most consistent of the methods used. The Bayesian approach to the training of a multi-layer perceptron (MLP) neural network does not refer to a single set of weights, as in the maximum-likelihood approach. Instead, it identifies a probability distribution over the weight-space that reflects the uncertainty in the value of the weights, given that we are limited to the use of finite data sets.

The complexity of neural network models can be controlled by using a regularization term, typically of the form $\frac{1}{2}\alpha\sum_{i=1}^{W} w_i^2$, where $\{w_i\}_{,i=1...W}$ is the vector of network weights. Overly complex models have to be avoided, because they over-fit the data and, as a result, generalize poorly. Within the Bayesian approach, this regularization term suppresses unnecessary weights. Regulation of the regularization coefficient α can be automated using the *evidence approximation* (MacKay, 1992a, 1992b). Therefore, the problem of over-fitting is tackled explicitly: over-complex models are penalized, as they yield a low probability of the target data given the regularization coefficients.

The ARD model extends the basic Bayesian approach by setting a different regularization coefficient for each input, so that the complete regularization term becomes $\frac{1}{2}\sum_{g=1}^{G}\left(\alpha_g \sum_{i=1}^{W_g} w_{gi}^2\right)$, where we have $g=1,\ldots,G$ groups of weights, one for each input. In practice, this works in a similar fashion to a pruning algorithm: high values of α_g will squash down the weights connected to the corresponding input, and its contribution to the overall model will be diminished. Direct inspection of the final values of α_g indicates the relative relevance of each variable. The Bayesian approach to the training of the Multi-Layer Perceptron (MLP) and the ARD technique for variable selection will be described in more detail as part of chapter 4.

3.4.2. Comparative results from the two models

Logistic discrimination

The stepwise forward likelihood-ratio variable selection method, applied to the logistic model using the 9-factor reduction of *data set A*, includes only 5 out of the 9 factors. These are, in order of entry, *2* (odds ratio *0.279*), *1* (odds ratio *0.3825*), *4* (odds ratio

0.5165), *9* (odds ratio *0.5991*) and *5* (odds ratio *0.6802*). See *table 2.2* in chapter 2 for the interpretation of the factors. The significance of the Wald statistic is lower than 0.001 for all of them. *Consumer risk perception/Environmental control* turns out to be the best predictor of purchasing behaviour, followed by *Shopping experience: Compatibility; Affordability; Shopping experience / Customer service* and finally *Shopping experience: Effort.*

Neural networks

The scores of the 9-factor reduction of *Data set A* were applied to a Bayesian-ARD neural network with a single hidden layer consisting of six nodes. This network was trained, within the Bayesian framework, using the *evidence approximation*, and the *scaled conjugate gradients* algorithm (Møller, 1993) to minimize the error function. It has to be borne in mind that the choice of the number of hidden nodes in the Bayesian approach is less restrictive than in the usual maximum-likelihood approach, due to the complexity-optimization process, which is designed to suppress redundant degrees of freedom.

The network performance was evaluated by a rotation (10-fold) cross-validation. Moreover, in order to avoid any bias introduced by the random initialization of the neural network weights, each of the validation data subsets was fed three times into the Bayesian-ARD neural network. As a result, the 9-factor model underwent latent variable selection 30 times. Ranking cumulative results are shown in *figure 3.1*: The outcome is consistent with that of the logistic discrimination; the *Consumer risk* *perception/Environmental control* factor appears neatly as the main predictor, whereas *Consumer risk: Image risk* appears to exert no influence whatsoever in the purchasing behaviour.

The 9-factor scores set plus the variables from *Data set B* (including *age, household income,* and *web usage patterns*), were subjected to the same neural network analysis, this time with a MLP with a single hidden layer consisting of eight nodes. The ranking of cumulative results from the ARD procedure is shown in *figure 3.2*: none of the added variables from *Data set B* turned out to be a good predictor: *household income,* in particular, revealed no discriminatory power. This is consistent with the results in Jarvenpaa and Todd (1996/97) and Bellman *et al.* (1999), reported in section 3.2.

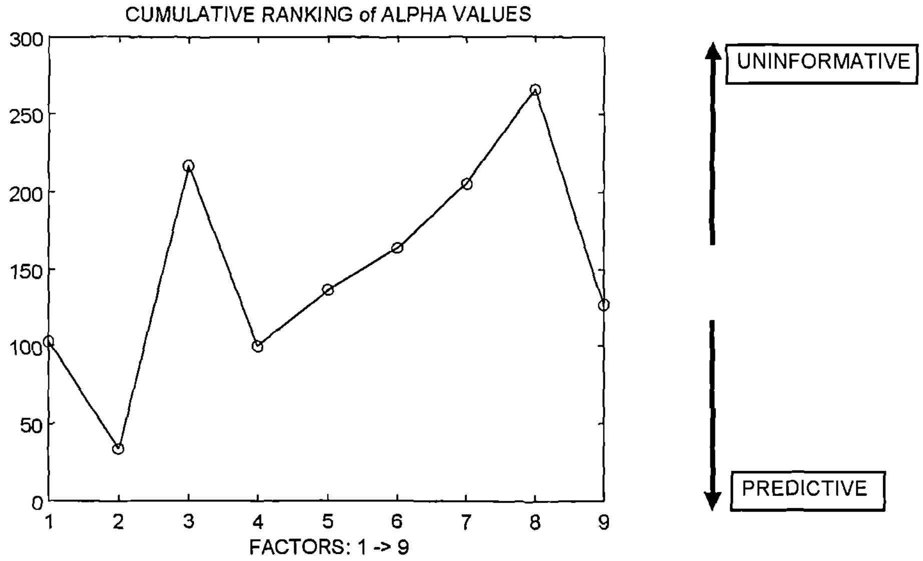


Figure 3.1: *Cumulative relevance calculated according to the ranking of latent variables on each of the 30 Bayesian neural network runs (A value of 30 would indicate that the variable was ranked 1st in all cases, whereas a value of 270 would rank the variable the last in all cases).*

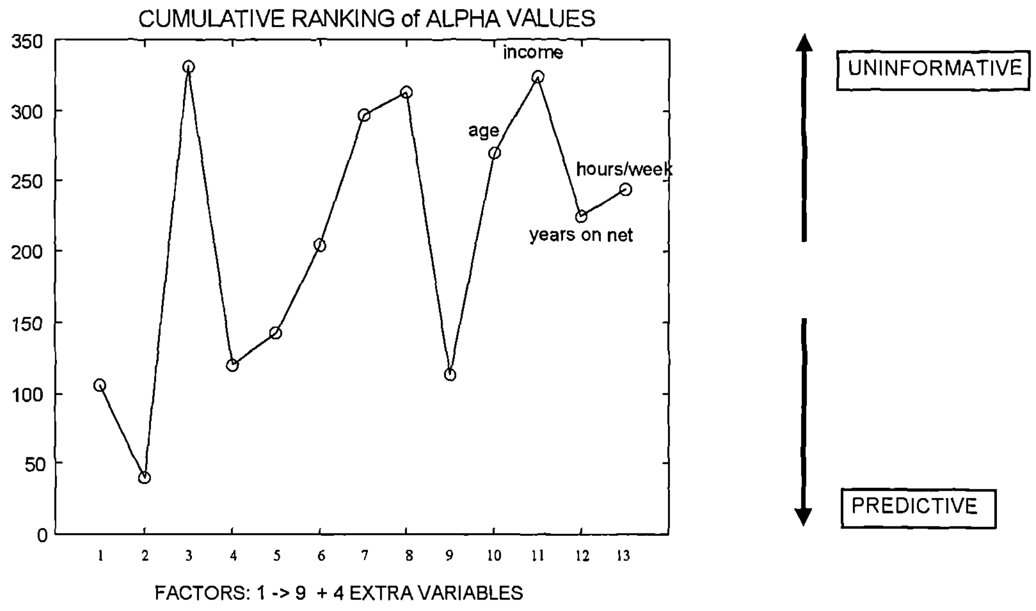


Figure 3.2: *Cumulative relevance calculated according to the ranking of latent and extra variables on each of the 30 Bayesian neural network runs (A value of 30 would indicate that the variable was ranked 1st in all cases, whereas a value of 390 would rank the variable the last in all cases).*

Comparison

The outcomes of the logistic discrimination and neural network-based feature selection methods consistently single out a group of five factors, namely 1, 2, 4, 5 and 9 (see table 2, chapter 2), as the most relevant for the prediction of the propensity to buy online. This 5-factor selection will be used in the experiments of the next section. The only major discrepancy between these results and those reported in Jarvenpaa and Todd (1996/97), summarized in section 3.2, seems to be the different relevance attached to the *consumer risk perception* factor. The relevance of *shopping experience: ease of use* (factor 5 in table 2, chapter 2) indirectly supports the importance conferred in Novak and Hoffman (1997), as reported in section 3.2, to the dimension of *skill* as a predictor of purchasing behaviour. This is because *ease of use* has been identified as an intermediate variable between *skill* and *flow* (Trevino and Webster, 1992).

3.5. Quantitative prediction of online purchasing behaviour

This section intends to provide an answer to the third of the research questions posed in the introduction: we want to find out to what extent can the propensity to buy online be quantitatively predicted. Central to this question is the comparison of the performance of three different inference methods: linear discriminant analysis, logistic discrimination and neural networks. The problem is first described and then the results of the experiments are reported.

3.5.1. Description of the problem

The same neural network and logistic discrimination models employed in the previous section plus linear discriminant analysis are now utilized for the estimation of the online purchasing predictive capabilities of different data sets. The goal is the discrimination of two classes, made up by those who have ever purchased online (*purchasers*) and those who never have (*non-purchasers*).

- We first aim to compare the predictive capabilities of two sets of data, using only the neural network model. The first set consists of the 9-factor reduction of *data set A*, plus the variables in *data set B*. The second consists of the 5 factors selected in the previous section by both the logistic discrimination and neural network models.

- Secondly, we want to compare the performance of the linear discriminant analysis, logistic discrimination and the Bayesian neural network as inference tools, using only the 5-factor selection data set.

The classification results are reported using the Receiver Operating Characteristic (ROC) plots of *sensitivity* against *specificity* (Bradley, 1997), plus overall *accuracy* (correct classification rate). Sensitivity (true positive fraction) and specificity (true negative fraction) are defined as:

SENSITIVITY = TP / (TP + FN)

SPECIFICITY = TN / (TN + FP),

where *TP* stands for true positives (individuals predicted to be purchasers who actually are purchasers), *TN* for true negatives (correctly predicted to be non-purchasers), *FP* for false positives (incorrectly predicted to be purchasers), and *FN* for false negatives (incorrectly predicted to be non-purchasers). Accuracy is defined as the percentage of correctly classified over the total number of samples. These plots offer not just overall classification results (the accuracy plot does) but also information about the partial results for each class. Moreover, they inform of the effect of different classification thresholds in those results. The area under the ROC plot (AURP) is a measure of the overall effectiveness of the model in separating the two classes: a measure of 0.5 indicates no separation capabilities, whereas a measure of 1 indicates perfect separation.

The discrimination between *purchasers* and *non-purchasers* must be approached with caution: following the argument in Anand *et al.* (1998), if those who have ever purchased a product / service online are considered as a class, the rest of the customers are not necessarily the complementary set. The latter might include those who, given the opportunity, would reject to purchase and those who would accept it: these might well share characteristics with those who already purchased in the past. Therefore, a straight dichotomy should not be assumed. Nevertheless, the characteristics of the analyzed sample (Web users more experienced than the average of the Internet population (Rogers, 1998) and, therefore, more exposed to shopping opportunities) alleviate this drawback. Furthermore, even though assuming that both classes inherently overlap, the results should provide valuable information on the level of separability that can be achieved.

3.5.2. Results

3.5.2.1. Performance: 9-factor solution plus data set B, versus 5-factor selection

We report now the compared neural network performance for the 9-factor solution plus data set B and the 5-factor selection data sets. The ROC and accuracy plots are shown in *figures 3.3* and *3.4*. These results correspond to the class-balanced sample mentioned in the data description section. For the 9-factor solution plus data set B, the maximum test accuracy is 79.85%, achieved at threshold 0.37. The best ROC measures (minimum distance to the top left corner, i.e. optimum combined sensitivity and specificity) correspond to a threshold of 0.52 for the test set, indicating a remarkable balance between the results for both the classes separately (simultaneous high specificity and sensitivity for central thresholds).

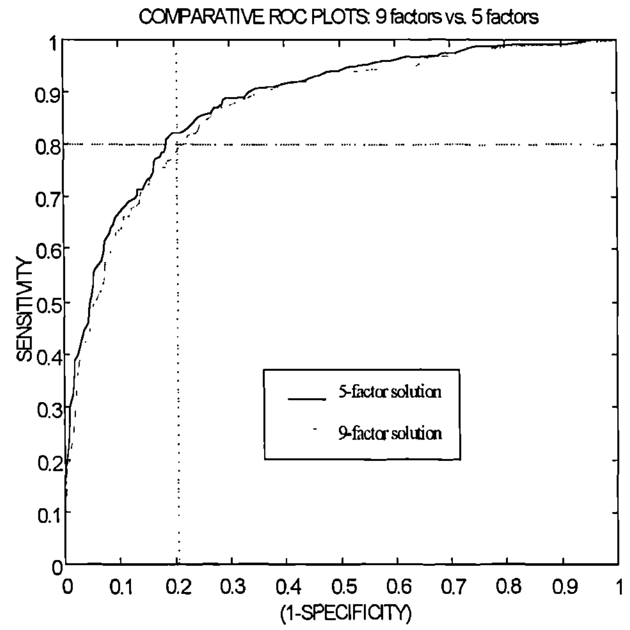


Figure 3.3: *ROC plots with the comparative performance of the Bayesian neural network (ARD) operating with 9-factor and 5-factor reductions of the observable data. Each of the 101 points shaping the curves corresponds, left to right, to a different classification threshold from 0 to 1, in 0.01 intervals. Each of those points is the average result of 30 neural network performances, stemming from a 10-fold cross-validation. Each training-test partition is used 3 times to reduce the bias that might occur due to the random initialization of the network weights. Only test results are shown. The closest point to the left-top corner represents the best performance. In this case, the 5-factor model outperforms the 9-factor one.*

This has to be compared with a peak test accuracy of 81.26% for the 5-factor selection, which is also well class-balanced. All the peak performance results, as well as the AURP measures, are shown in *table 3.1*. The AURP value of 0.8670 for the 9-factor solution plus data set B is equivalent to an 86.70% probability of correctly ranking two individuals, one from each of the classes, and has to be compared to the AURP value of 0.8807 for the 5-factor selection data set.

It has been shown that the 5-factor description of the original data can be used to discriminate between the classes of *purchasers* and *non-purchasers* with a higher degree of accuracy than the more complex description consisting of 9 factors plus *data set B*.

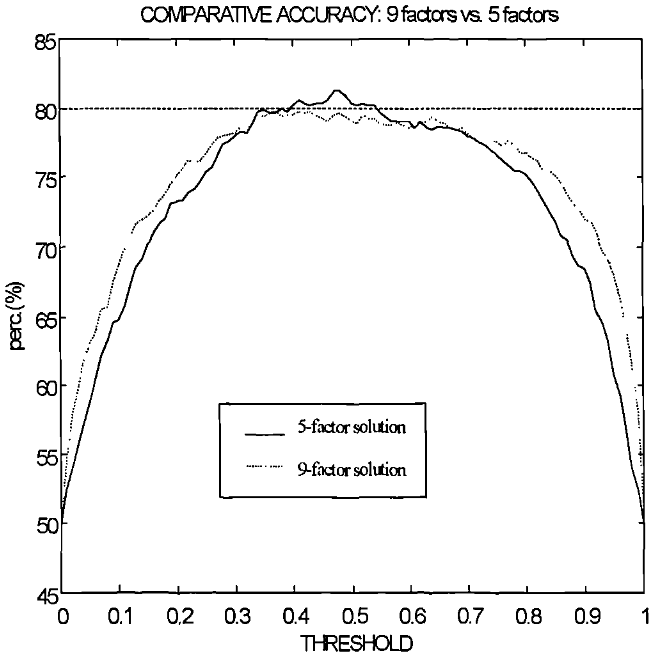


Figure 3.4: *Accuracy (correct classification rate) as a function of the decision threshold: comparative performance of the Bayesian neural network (ARD) operating with 9-factor and selected 5-factor reductions of the observable data. Each of the points in these plots has been obtained following the methodology outlined in fig.3.3. Only test results are displayed. The 5-factor model is shown to predict propensity to buy on-line with higher peak-accuracy.*

3.5.2.2. Five-factor selection: linear discriminant analysis, logistic discrimination and neural network predictions

The outcome of the latent variable selection procedures (for both logistic discrimination and neural network models) consistently pointed out the relevance of a subset of factors: namely, numbers 1, 2, 4, 5 and 9 (See *table 2, chapter 2*). The goal of the current experiment is to compare the predictive capabilities of the linear discriminant analysis, logistic discrimination and neural network models, for the 5-factor selection data set. Two variants of Bayesian neural network models are used. The first, a basic model with a single regularization term (SRT) and the second, the more complex model (ARD) with multiple regularization terms.

Figure 3.5 corresponds to the ROC and accuracy plots describing the neural networks, logistic discrimination and linear discriminant analysis performances. For both neural network models (SRT and ARD), a maximum test accuracy of 81.26% is reached. For SRT, the corresponding threshold is 0.4 and for ARD, the threshold is 0.48. The best ROC measures (minimum distance to the top left corner) are summarized in *table 3.1*.

For the logistic discrimination, a peak accuracy of 80.36% is reached for a threshold of 0.49, with the best ROC measures corresponding to the same threshold. Peak performances and, once again, AURP measures are shown in *table 3.1*.

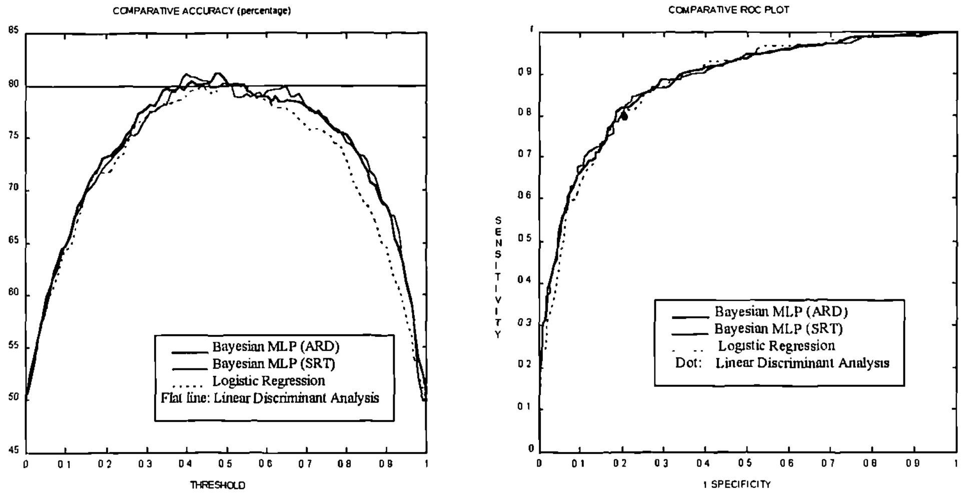


***Figure 3.5:** Accuracy plot (left) and ROC plot (right) summarizing the comparative performance of two neural networks (ARD and SRT), logistic regression and linear discriminant analysis. Each of the 101 points shaping the curves corresponds (left to right for the accuracy, right to left for the ROC) to a different classification threshold from 0 to 1, in 0.01 intervals.*

For the linear discriminant analysis, a test accuracy of 80% is achieved in a leave-one-out procedure. This value is not a function of the decision threshold as in the previous cases. For that reason there is no ROC plot associated.

The results of the application of supervised neural networks to this classification problem are comparable to those obtained applying logistic discrimination. This is true in terms of the overall classification. However, some further interesting results can be obtained from a detailed inspection, across the input range, of the predictions of the different models. On close inspection, the neural network models succeed in correctly classifying outliers that the logistic and linear models are not able to identify.

ACCURACY (Correct classification rate)	NN: 9 factors plus *data set B*	**79.85 %**
	NN (SRT and ARD): 5 factors	**81.26 %**
	LinDA: 5 factors	**80.00 %**
	LogD: 5 factors	**80.36 %**
AURP (Area under ROC plot)	NN: 9 factors plus *data set B*	**0.8670**
	NN: 5 factors (ARD)	**0.8807**
	NN: 5 factors (SRT)	**0.8802**
	LinDA: 5 factors	**NA**
	LogD. 5 factors (balanced)	**0.8746**

Table 3.1: *Best test performances for all the models used in the analyses in section 3.5.2. They show a comparable performance, fitting the data equally well.*

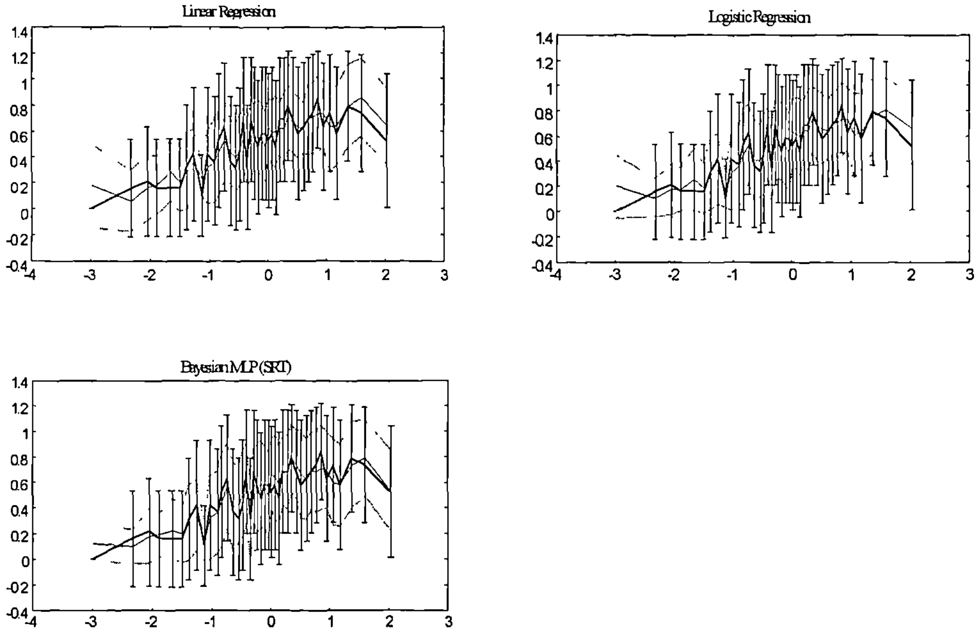


Figure 3.6: *Prediction of the three inference models against the spread of values of the Shopping Experience: Convenience factor. The thick solid line and the solid vertical bars correspond, in turn, to the average of the observed class-membership (0 or 1) across the range of the factor values and its standard deviation. The thin solid line and the dashed line correspond, in turn, to the average model prediction and its standard deviation. The closer the two solid lines and the closer the thin solid and the dashed lines, the better the model prediction.*

The example in *figure 3.6*, concerning the factor/variable *Shopping Experience: Convenience*, shows how the main difference between the predictions of the linear discriminant analysis and logistic regression, in one hand, and the Bayesian neural networks, in the other, is the way in which the latter manage to predict more accurately those cases at the extremes of the input range. The neural network also shows more balanced class-specific results.

3.6. Conclusion

The three research questions outlined in the introduction have been addressed throughout the chapter.

> ***Question 1:*** *Which of the latent factors identified in the previous chapter contain the necessary information to best predict Web users' propensity to buy on-line?*

Using the latent factor scores as the starting point for variable selection procedures, it was found that logistic discrimination and Bayesian neural network models selected the same factors, confirming the importance of *Consumer risk perception Environmental control* as the main discriminator between the classes of online purchasers and non-purchasers. This result entirely supports the statements in Hoffman *et al.* (1999), pointing to the lack of consumer trust –materialised in “concerns over information privacy in electronic, networked environments”- as the main restraining factor for online shopping adoption. The importance of *Affordability* and *Shopping experience: Effort: Ease of use* can be explained as a result of the mainstreaming of the online shopping population, parallel to that of the Internet population described in Rogers (1998), according to which the Internet, as a communication and shopping channel, is no longer an exclusive of high-income technophiles. This conclusion is reinforced by the lack of relevance of the *Consumer risk: Image* factor, and concurs with Chaum (1997) that *"*Ease of use*"* and *"*Access to the hardware needed*"* are two of the three main *"*barriers to widespread adoption of electronic commerce*"*. The combined factor *Shopping experience / Customer service*, also shown to be predictive, raises concern over the recent reports that the worsening of online shopping customer service threatens customer loyalty to this shopping channel (Glasner, 1999). Developing lasting relationships between the vendor and the customer, built upon customer service, should be one of the long-term objectives of commercial Web sites (Zwass, 1999) and on-line marketers. On the other hand, the relevance of *Shopping experience: Compatibility* should encourage online marketers, by pointing to consumers' appreciation of precisely those factors which make online shopping different to traditional retail channels. *Question 1* is thus answered.

> ***Question 2:*** *Are variables, such as age, household income, and Web usage patterns, commonly used in market research, good predictors of that propensity?*

Further experiments have indicated that age, household income, and web usage patterns do not add to the predictive power for online purchasing behaviour. This outcome reinforces some of the findings presented in Jarvenpaa and Todd (1996/97) and Bellman (1999), and answers *Question 2*. Overall, and from a business perspective, these results suggest that management decision making may be focused on factors under in-house control (*Consumer risk perception, Shopping experience, Customer service, Environmental control*), since their ability to influence prospective customers outweighs the effects of demographic, socio-economic or web usage variables.

> ***Question 3:*** *To what extent can the propensity to buy online be inferred, using quantitative methods, from a selection of the latent factors shown to be the best predictors of such a propensity and how does it compare with the prediction obtained from the complete set of demographic, socio-economic, web usage, and latent variables? Also, is there a significant difference in the predictive performance of the logistic discrimination and neural network models?*

It has been demonstrated, in answer to *Question 3*, that the neural network and logistic discrimination models are capable of predicting propensity to buy on-line to a surprising degree of accuracy, the neural network prediction being only marginally more accurate than that of the logistic discrimination. Furthermore, the use of a dimension-reduced data set, made up of the five most relevant factors, maintained the accuracy of larger models in the discrimination of purchasers from non-purchasers. It also results in a much more manageable model than the one including the original 48 observable variables.

3.7. Publications related to this chapter

1. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Quantitative characterization and prediction of on-line purchasing behaviour: a latent variable approach. *International Journal of Electronic Commerce*, 4(4), 83-104

2. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Characterizing and segmenting the business-to-consumer e-commerce market using neural networks. In Lisboa, P.J.G., Vellido, A. and Edisbury, B. (Eds.): *Business Applications of Neural Networks.* Singapore: World Scientific, 29-54.

3. Lisboa, P.J.G., El-Deredy, W., Vellido, A,, Etchells, T. and Pountney, D.C. (1997): Automatic variable selection and rule extraction using neural networks. In *Proceedings of the 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics*, Berlin, 461-466.

Chapter 4

The Bayesian framework for Multi-Layer Perceptrons: the effect of marginalization and ARD in the prediction of online purchasing behaviour.

4.1. Introduction

Within the Bayesian approach to the training of multi-layer perceptrons for classification problems, the interpretation of the outputs as posterior probabilities of class-membership requires to integrate out (marginalize) the network function over the distribution of network weights. MacKay (1992b) suggests an approximation of such analytically intractable integral, in which the integration is over the network output pre-activations. The network predictions can be overoptimistic if this process of marginalization is ignored. This chapter attempts to assess the effect of marginalization, with the approximation aforementioned, on two Bayesian neural network models: one with a single regularization term and another, giving way to a soft feature selection process known as *Automatic Relevance Determination* (ARD), with multiple regularization terms. The classification problem posed in the previous chapter is the test bed for this assessment.

The identification of the factors influencing the consumers' propensity to buy online is of paramount importance to Internet retailers. So is the management of uncertainty in the prediction of that propensity. The process of marginalization tackles the problem of decision uncertainty explicitly, by preventing overoptimistic predictions.

The Bayesian approach to the training of a multi-layer perceptron does not simply attempt to find a single set of weights, as in the maximum-likelihood approach. Instead, it predicts a probability distribution over the weights, which reflects the uncertainty resulting from the use of finite data sets. In that way, the outputs of the neural network in a classification problem can be interpreted as posterior probabilities of class membership given the data and the weights,

$$P(C_i \mid \mathbf{x}, \mathbf{w}) = y(\mathbf{x} \,;\, \mathbf{w}) \qquad (1)$$

where y is the network function, $\mathbf{x}$ is a vector of input data, $\mathbf{w}$ is the vector of the trained network parameters (weights and biases), and C_i is class i. The probability of membership of a specific class for an input vector with which the network has not been trained (test vector), is obtained by marginalizing (1) using the conditional distribution for the weights, to give

$$P(C_i|\mathbf{x}, D) = \int y(\mathbf{x}; \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w} \qquad (2)$$

where D are the target data for the training set. In regression problems, and given a Gaussian approximation for the posterior distribution of the weights $p(\mathbf{w}|D)$, sufficiently narrow to justify a linear expansion of the output around the most probable weight values, an integral analogous to that in (2) can be calculated and the variance of the resulting posterior distribution can be used to calculate error bars on the predicted test outputs.

An alternative to the Laplace approximation is the use of Markov Chain Monte Carlo (MCMC) techniques for random sampling in weight-space directly (Bioch *et al.*, 1995; Neal, 1996), which results in a direct numerical approximation of the integral in (2). A potential drawback associated with MCMC methods is that they are computationally very demanding and convergence is difficult to assess (Neal, 1998). Further limitations are described in Attias (1999) and Ghahramani and Beal (2000). An alternative framework for the analytical calculation of posterior distributions, Variational Bayes (Attias, 1999) has only recently been proposed.

In classification problems, the network output involves a sigmoidal activation function, which can not be approximated by a linear function of the weights, since

$$y = f(a) = \frac{1}{1 + \exp(-a)}, \tag{3}$$

where the activation a is the weighted linear sum of the hidden node responses which is fed into the output nodes. The expression (1) for the marginalized network output becomes

$$P(C_i|\mathrm{x}, D) = \int f(a) p(a|\mathrm{x}, D) da \tag{4}$$

MacKay (1992b) suggested a linear approximation to a,

$$p(a|\mathrm{x}, D) = \int p(a|\mathrm{x}, \mathrm{w}) p(\mathrm{w}|D) d\mathrm{w} = \int \delta(a - a_{MP} - \mathrm{g}^{\mathrm{T}} \Delta\mathrm{w}) p(\mathrm{w}|D) d\mathrm{w}, \tag{5}$$

where $\delta(.)$ is the Dirac delta function, a_{MP} the most probable value of a, and $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}_{\mathrm{MP}}$. The integral in (5) is now tractable, resulting in a normal distribution for a (MacKay, 1992b; Bishop, 1995).

$$p(a|\mathbf{x}, D) = \frac{1}{\left(2\pi s^2\right)^{1/2}} \exp\left(-\frac{\left(a - a_{MP}\right)^2}{2s^2}\right) \tag{6}$$

Having expanded a about its most probable value, allows an approximation to the integral in (4) to be obtained (MacKay, 1992b), resulting in

$$P(C_i|\mathbf{x}, D) \cong f(\kappa(s)a_{MP}); \quad \kappa(s) = (1 + \frac{\pi s^2}{8})^{-\frac{1}{2}} \tag{7}$$

where s^2 is the variance of the posterior distribution (6). The effect of the term κ(s) on the marginalized output (7) for a new (test) input vector, is the *moderation* of the most probable output $y_{MP} = f(a_{MP})$, as its value is pushed towards 0.5. Therefore, should the classification threshold be set to that value, there would be no change in the classification decision. Nevertheless, that decision will be altered if an acceptance/rejection threshold is set.

Neural Networks are frequently considered as *black boxes* due, amongst other things, to their supposed incapacity to identify the relevance of independent variables in non-linear terms. This makes it difficult to automatically carry out input variable selection. Automatic Relevance Determination (ARD), for supervised Bayesian Neural Networks (MacKay, 1995), is a model that tackles that shortcoming: Within the basic Bayesian approach the weight decay or regularization term is interpreted as a Gaussian prior distribution over the network parameters (weights and biases), of the form

$$p(\mathbf{w}) = A\exp\left(-\alpha\sum_{i}^{N_w} w_i^2/2\right) \tag{8}$$

where $\mathbf{w}=\{w_i\}$ is the vector of network parameters, N_w is the total number of network parameters, and A is a normalization factor and is constant. In ARD, individual regularization terms are associated with each group of network parameters. The fan-out weights from each input to the hidden layer are grouped separately, and the remaining weights form two additional groups, namely bias to the hidden nodes and weights plus bias to the output. The prior distribution of the weights now becomes

$$p(\mathbf{w}) = A\exp\left(-\sum_{c}^{C}\alpha_c\sum_{i}^{n_{w(c)}} w_i^2/2\right) \tag{9}$$

where C = (number of inputs + 2) is the number of weight groups, and $n_{w(c)}$ is the number of parameters in group c, so that $\sum_{c}^{C} n_{w(c)} = N_w$. As a result of the network training, the hyperparameters α_c associated with irrelevant inputs will be inferred to be large, and the corresponding weights will be set to small values. Therefore, ARD is performing soft feature selection, and a direct inspection of the final $\{\alpha_c\}$ values indicates the relative relevance of each variable. ARD has shown itself to be a useful feature selection method for classification problems (Penny and Roberts, 1999).

This chapter illustrates the effect of the marginalization of the outputs in the outcome of a Bayesian neural network provided with the real-world data for the classification problem described in the previous chapter. Some comments on these data are made in the next section. This is followed by the presentation of the results and some final comments.

4.2. Description of the data: Unequal class prior probabilities

The five most relevant factors underlying Internet users' opinions of online shooping, labelled according to previously published qualitative studies, that were obtained in the previous chapter, are summarized in *table 4.1*, to ease the interpretation of the results in the next section.

FACTOR	DESCRIPTION	ATTRIBUTES
1	*Shopping experience:* *Compatibility*	*Control and convenience*
2	*Consumer risk perception /* *Environmental control*	*Trust and security*
3	*Affordability*	--
4	*Shopping experience: Effort*	*Ease of use*
5	*Shopping experience /* *Customer service*	*Effort /* *Responsiveness and empathy*

Table 4.1: *Descriptive summary of the factors selected by the ARD model as predictive of online purchasing behaviour.*

The data sample utilised contains records from 2180 Internet users, but it is not balanced in terms of class membership; therefore, a new class-balanced data set with 778 individuals was created and later used in a n-fold cross validation. This enables the neural network to fit both classes properly. Nevertheless, it entails a skewed choice of prior distributions. If we refer to the class of *purchasers* as C_P and to the class of *non-purchasers* as C_{NP}, the balanced priors will obviously satisfy

$P'(C_P) = P'(C_{NP}) = 0.5$. However, the posterior probability of class-membership can be modified to reflect the unequal real distribution of priors. Following Tarassenko (1998), the posterior distribution, taking account of the empirical priors $P(C_P)$ and $P(C_{NP}) = 1 - P(C_P)$, is

$$P(C_P|\mathbf{x}) = \frac{y\frac{P(C_P)}{P'(C_P)}}{y\frac{P(C_P)}{P'(C_P)} + (1-y)\frac{1-P(C_P)}{1-P'(C_P)}}, \qquad (10)$$

where $\mathbf{x}$ is a data vector and y is the network activation. The effect of this modification is that the marginalization now drives $P(C_P|\mathbf{x})$ towards $P(C_P)$ instead of $P'(C_P) = 0.5$. Note that this relies on having a good enough estimate of the prior probabilities of each class, which is the case given the number of samples in our study. The next chapter provides a more principled procedure to deal with unequal prior probabilities in the context of the training of Bayesian neural networks.

4.3. Quantitative prediction of online purchasing behaviour

The classification results are again presented using the Receiver Operating Characteristic (ROC) plots of *sensitivity* against *specificity*, and an overall *accuracy* plot. Let us remark that the Area Under the ROC Plot (AURP) represents the probability that two randomly selected individuals, one from the population of online purchasers and one from the population of non-purchasers, are ranked correctly: in this study, $P(C_P|\mathbf{x}_{\text{purchaser}}, D) > P(C_P|\mathbf{x}_{\text{non-purchaser}}, D)$, for marginalized outputs. The AURP is also the same quantity that is estimated by the nonparametric Wilcoxon (Mann-Whitney) statistic (Hanley and McNeil, 1982).

The results of two Bayesian neural networks will be compared: a model with a single regularization term (SRT), and the ARD model with multiple regularization terms.

4.3.1. Results: Single regularization term (SRT) model

The ROC plot describing the neural network performance is shown in *figure 4.1*. The results correspond to class-balanced samples, as mentioned in the data description section. The best ROC measures (minimum distance to the top left corner, i.e. optimum combined sensitivity and specificity) for the test data are obtained in the threshold interval from 0.4 to 0.5, with an optimum result in 0.4. By effect of the marginalization, the test outputs are pushed towards 0.5. For thresholds over this value, the sensitivity worsens, whereas the specificity improves. For thresholds under that value it is the other way around: specificity worsens and sensitivity improves. This effect, though, is barely noticeable in *figure 4.1*. For a threshold of 0.5 the results remain unchanged. The discrimination capabilities of the model are assessed by the AURP measure, which is shown in *table 4.2*: remarkably, the marginalization of the test outputs does not affect the capability of the model to separate the classes of online purchasers and non-purchasers, as measured by the AURP index. The probability of correctly ranking two randomly selected individuals, one from the population of online purchasers and one from the population of non-purchasers, is 88.02% in the case of non-marginalized outputs, and 88.01% in the case of marginalized outputs.

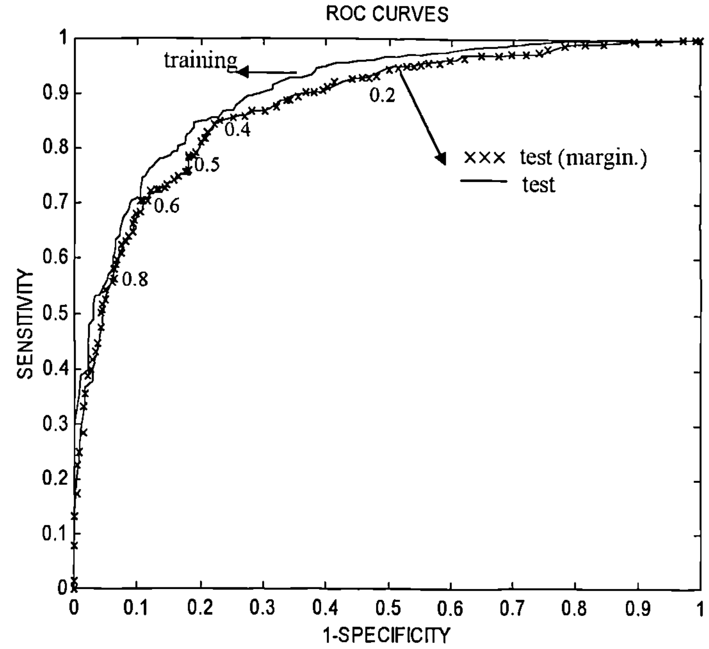


Figure 4.1: *ROC curves corresponding to the SRT model. The test performance for marginalized and non-marginalized outputs was estimated using 10-fold cross-validation. As an illustration, several points of the non-marginalized test curve have been labelled with its corresponding decision threshold.*

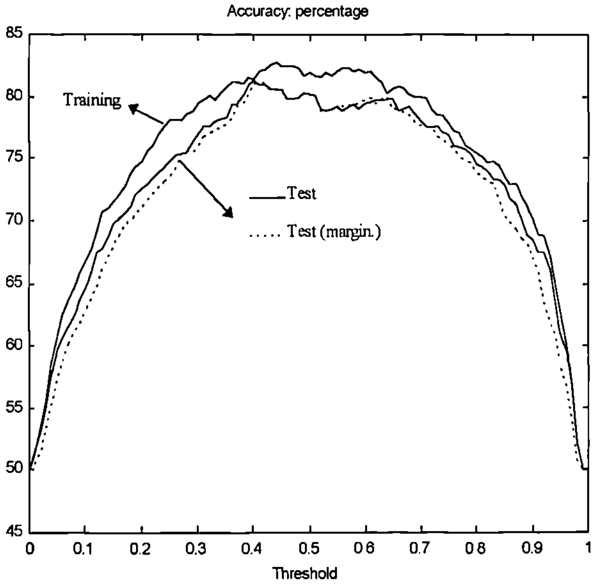


Figure 4.2: Accuracy curves corresponding to the SRT model. They are evaluated over threshold intervals of 0.01 width. Note that the marginalized accuracy is not strictly less that in the absence of marginalization.

AURPs

Training	**0.9057**
Test (non-marginalized)	**0.8802**
Test (marginalized)	**0.8801**

Table 4.2: *Area under the ROC plots (AURP) of figure 1 (SRT model).*

The classification results are summarized in the accuracy plot of *figure 4.2*: A test accuracy of over 80% is achieved by the model with non-marginalized outputs within the threshold interval from 0.38 to 0.51, with a maximum of 81.26% for a threshold of 0.4. The model with marginalized outputs achieve over 80% within the threshold interval from 0.39 to 0.51, with a maximum of 81.13% for a threshold of 0.41. The accuracy for a threshold of 0.5 remains unchanged. The effect of the marginalization of the test outputs in the accuracy results is barely noticeable.

4.3.2. Results: Automatic Relevance Determination (ARD) model

The ROC plot for the ARD model is shown in *figure 4.3*. The best measures for the test data set are obtained in the threshold interval from 0.44 to 0.54, with an optimum result in 0.48. It indicates a remarkable balance between the class-specific performances (simultaneous high specificity and sensitivity for central thresholds). The AURP measures are shown in *table 4.3*: Once again, the marginalization of the test outputs does not affect the discrimination capabilities of the model. The probability of correctly ranking two randomly selected individuals, one from the population of online purchasers and one from the population of non-purchasers, is

88.07% in the case of non-marginalized outputs, and 88.19% in the case of marginalized outputs.

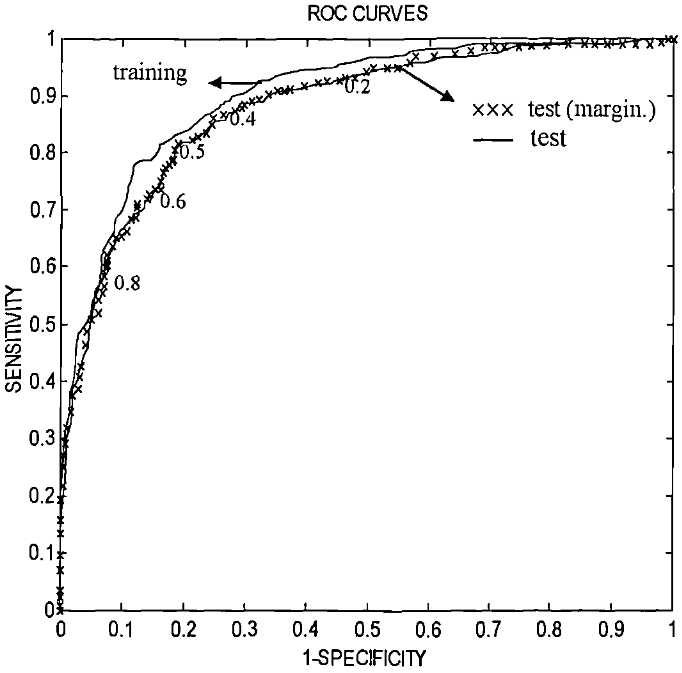


Figure 4.3: ROC curves corresponding to the ARD model. The test curves for marginalized and non-marginalized outputs are the result of a 10-fold cross-validation process. As an illustration, several points of the non-marginalized test curve have been labelled with its corresponding decision threshold.

AURPs

Training	**0.9038**
Test (non-marginalized)	**0.8807**
Test (marginalized)	**0.8819**

Table 4.3: Area under the ROC plots (AURP) of figure 3 (ARD).

The accuracy plot for the ARD model is shown in *figure 4.4*: A test accuracy of over 80% is achieved by the non-marginalized outputs within the threshold interval from 0.4 to 0.54, with a maximum of 81.26% for threshold 0.48 ; the marginalized outputs achieve over 80% within the threshold interval from 0.41 to 0.53, with a maximum of 81.39% for a threshold of 0.48.

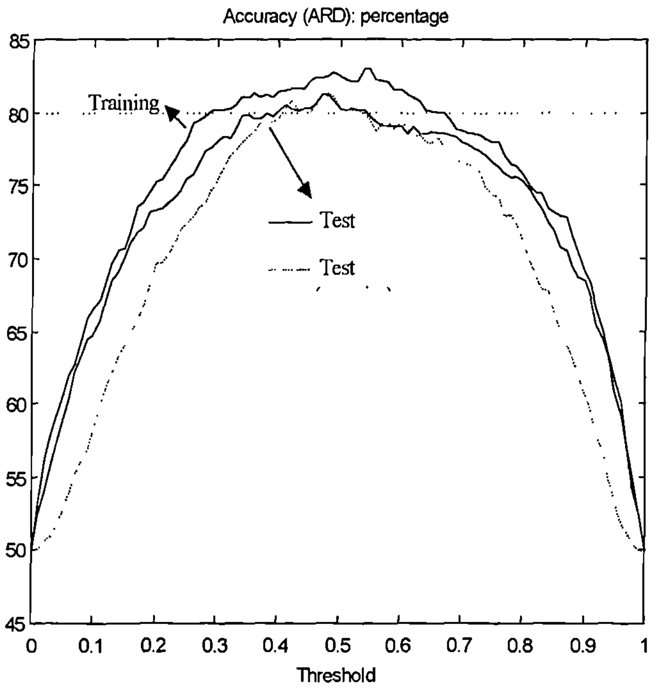


Figure 4.4: Accuracy curves corresponding to the ARD model. They are evaluated over threshold intervals of 0.01 width.

Once again as expected, the accuracy for a threshold of 0.5 remains unchanged. The effect of the moderation of the test outputs, due to the marginalization, in the accuracy results is quite noticeable for thresholds over 0.65 and under 0.35.

Overall, the ARD model yields similar or marginally better results than the model with a single regularization term, although the difference between the AURPs for both models can be shown to be non-significant using the results in Hanley and McNeil (1983). The effect of the use of multiple regularization terms is illustrated by the matrix of input to hidden-layer weights shown in *table 4.4:* large α 's shrank the weights corresponding to irrelevant inputs. ARD works, in fact, as an adaptive pruning procedure with sound statistical grounds.

	hidden node 1	*hidden node 2*	*hidden node 3*	*hidden node 4*	*hidden node 5*	*mean*
input 1 COMPATIBILITY	-0.0237	0.2463	0.0355	-0.1607	0.2683	0.1469
input 2 RISK PERCEPTION	-0.1693	0.1294	-0.1977	0.8925	-1.0342	**0.4846**
input 3 AFFORDABILITY	-0.2543	-0.2674	0.0259	0.2173	0.0088	0.1547
input 4 EASE OF USE	0.0299	0.0482	-0.0339	0.2882	-0.0482	0.0897
input 5 EFFORT	-0.1126	0.1390	-0.0473	0.0014	-0.1008	0.0802

Table 4.4: Input-to-hidden layer matrix of weights for a trained Bayesian neural network with multiple regularization terms. The last column is the mean of the absolute values of the weights coming out of each input. Observe the agreement between these values and the variable selection results in chapter 3. For the interpretation of the inputs, see table 4.1.

4.4. The reject option

In practical applications, when faced with doubtful classification decisions on individual cases, it can be useful to single out those cases and consider them separately. In this study it might be the situation of an individual for which the neural network prediction as potential online purchaser is uncertain: The loss derived of the misclassification has to be balanced with the loss of avoiding to make a decision on that person. This *reject option* entails a compromise between a maximization of the ratio of correctly classified samples, and the minimization of the ratio of rejected samples. For a multiple class problem in which each class is associated to an independent neural network output, the following rejection rule is suggested in Bishop (1995)

$$\text{If } \max_k P(C_k|\mathbf{x}) \begin{cases} \geq \theta, & \textbf{then classify x} \\ \leq \theta, & \textbf{then reject x} \end{cases} \tag{11}$$

for a threshold θ in the range (0,1). For the two-class problem in this study (with classes C_P and C_{NP}), in which a neural network with a single output is used, the posterior probability of the assignment to C_P of an input vector **x** is $P(C_P|\mathbf{x})$, whereas $P(C_{NP}|\mathbf{x}) = 1 - P(C_P|\mathbf{x})$. Therefore, the rejection rule (11) becomes

$$\text{if } P(C_P|\mathbf{x})\begin{cases} \geq \theta \ \ \textbf{or} \ \ \leq 1-\theta, \ \ \textbf{then classify } \mathbf{x} \\ < \theta \ \ \textbf{and} \ \ > 1-\theta, \ \ \textbf{then reject } \mathbf{x} \end{cases} \tag{12}$$

for $0.5 \leq \theta < 1$.

Should the e-marketer be able to quantify the loss associated to the misclassifications from both classes, other risk-minimization strategies could be deployed. For instance the inclusion of a loss matrix (Bishop, 1995), with elements $L_{k,j}$ defined as the penalty associated with assigning to class j an individual that belongs to class k (alternatively, the penalties associated with an individual being a *true positive, true negative, false positive* or *false negative* as defined in section 3). The overall risk would be minimized if an individual were assigned to the class of *purchasers* only if $L_{P,NP}P(C_P|\mathbf{x}) > L_{NP,P}P(C_{NP}|\mathbf{x})$, and, otherwise, to the class of *non-purchasers* (assuming that there is no loss when an individual is correctly classified, i.e. $L_{k,j} = 0$, if $j \quad k$).

Alternatively, this loss can be defined explicitly in terms of the error rates as a function of the predicted class membership probabilities. This more flexible approach, however, requires good accuracy for the model across the complete range of thresholds, known in statistics as the *model calibration*. The calibration accuracy, between thresholds *T*1 and *T*2 is defined as

$$T1 \leq P(C_P|\mathrm{x}) \leq T2 \equiv (TP_{T1} - TP_{T2}) / [(TP_{T1} - TP_{T2}) + (FP_{T1} - FP_{T2})] \quad (13)$$

where *TP* are the true positives and *FP* are the false positives, as defined in section 3. This calibration accuracy is shown in *figure 4.5* for the ARD model.

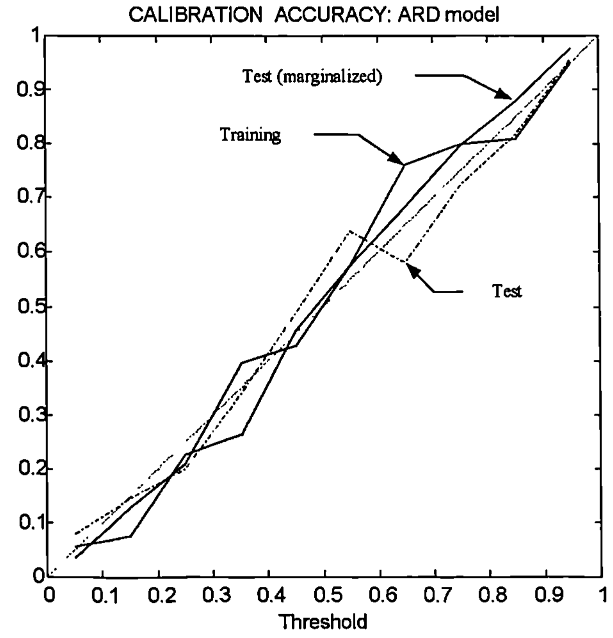


Figure 4.5: *Calibration accuracy for the predictions of the ARD model. The diagonal dashed line in the plot represents the optimum calibration accuracy. The test results are slightly degraded from the training results, as expected, but the effect of marginalization is to level out the calibration accuracy - i.e. stabilize the slope of the above curve- across the range of thresholds. Measurements are made in threshold intervals of 0.1 and they are plotted in the midpoints of these intervals.*

4.4.1. Results: Single regularization term model

The graphic in *figure 4.6a* shows the correct classification and the rejection rates against the threshold for the Bayesian neural network prediction on our data, using the SRT model. For instance, a rejection of 24% of the cases of the test data set is required for a 10% improvement of the correct classification rate; this entails an uncertainty region stretching the interval from 0.31 to 0.69. In the case of the marginalized outputs, the same accuracy increase entails a rejection of 24.65%, but the uncertainty region goes now from 0.32 to 0.68. This is clearly observed in *figure 4.6b,* which shows the accuracy against the rejection rate.

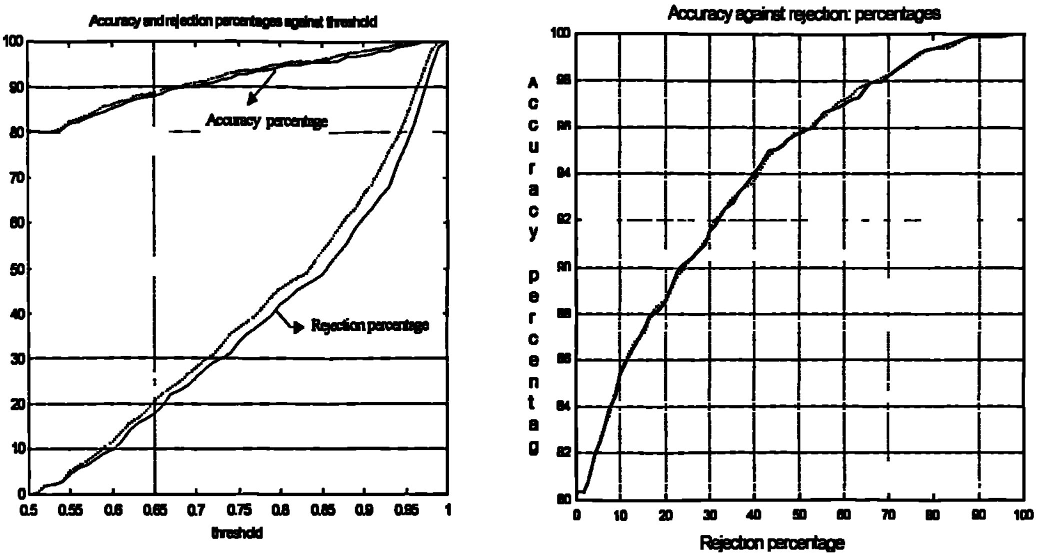


Figure 4.6: a, left) SRT model: Accuracy and rejection percentages, as a function of the uncertainty threshold, for the reject option. The solid lines and the dotted lines correspond, in turn, to the results for non-marginalized and marginalized outputs. b, right) SRT model: Accuracy against rejection percentages for the reject option. The solid lines and the dotted lines correspond, in turn, to the results for non-marginalized and marginalized outputs.

4.4.2. Results: Automatic Relevance Determination (ARD) model

The correct classification and the rejection rates against the threshold for the ARD model are shown in *figure 4.7a.* The marginalization process is more noticeable in this case, as the moderation of the outputs is more intense. It yields slightly worse rejection and correct classification percentages given equal uncertainty regions. For non-marginalized outputs, a 10% improvement of the correct classification rate requires a rejection of 24.13% of the cases of the test data set for an uncertainty region from 0.31

to 0.69, whereas for the marginalized outputs a rejection of 24.39% is required for a narrower uncertainty region from 0.35 to 0.65. *Figure 4.7b* shows the accuracy against the rejection rate.

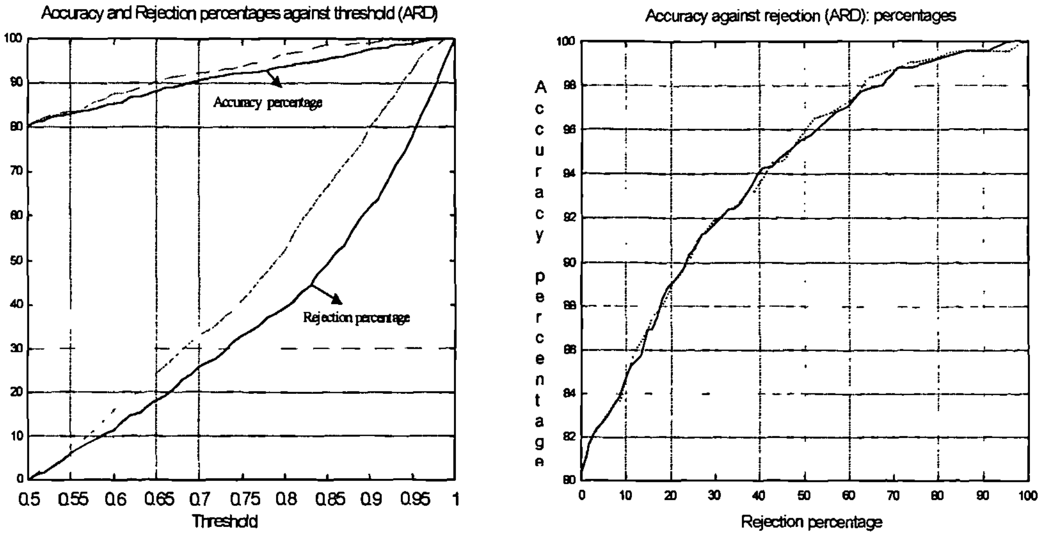


***Figure 4.7:** **a, left)** ARD model: Accuracy and rejection percentages, as a function of the uncertainty threshold, for the reject option. The solid lines and the dotted lines correspond, in turn, to the results for non-marginalized and marginalized outputs. **b, right)** ARD model: Accuracy against rejection percentages for the reject option. The solid lines and the dotted lines correspond, in turn, to the results for non-marginalized and marginalized outputs.*

4.5. Conclusion

Unlike in regression problems, the Bayesian approach for the application of neural networks in classification problems does not permit a straightforward calculation of error bars for the predicted test outputs. An approximation of the marginalization of the test outputs, proposed in MacKay (1992b), can prevent obtaining overoptimistic classification results. It has been shown that, for the application to the classification of Internet users according to their propensity to buy online, the marginalization of the test outputs does not entail any loss of overall discriminatory power for the model. The reject option, to defer decision in circumstances of uncertainty in the neural network predictions, provides guidance to assess the effect of the test output marginalization on the relation between rejection ratio, accuracy and width of the uncertainty region. If the expectations of mass-customising the e-commerce market are ever to be realized, this management of the prediction uncertainty should be at the core of the online retailers' decision making.

All the analyses have been carried out using two different, but related, Bayesian neural network models: one with a single regularization term and another with multiple regularization terms. They have been shown to hold similar predictive capabilities while the latter also provides some explanatory features in the form of a ranking of variables according to their predictive power. This chapter has shown how, beyond the use as a classifier, the probabilistic formulation of the Bayesian MLP can provide systematic feature selection and a framework for the management of decision uncertainty.

4.6. Publications related to this chapter

1. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2001): An electronic commerce application of the Bayesian framework for MLPs: the effect of marginalization and ARD. *Neural Computing and Applications*, forthcoming.
2. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Characterizing and segmenting the business-to-consumer e-commerce market using neural networks. In Lisboa,

P.J.G., Vellido, A. and Edisbury, B. (Eds.): *Business Applications of Neural Networks.* Singapore: World Scientific, 29-54.

Chapter 5

Bias reduction in skewed classification with Bayesian neural networks: an application to the prediction of online purchasing behaviour.

5.1. Introduction

The Bayesian evidence framework has become a standard of good practice for neural network estimation of class conditional probabilities. In this approach, as described in the previous chapter, the conditional probability is marginalised over the distribution of network weights, which is usually approximated by an analytical expression that moderates the network output towards the midrange. In this chapter, it is shown that the network calibration is improved by marginalising to the prior distribution of the classes. Moreover, marginalisation to the midrange can seriously bias the estimates of the conditional probabilities calculated from the evidence framework. In this way, a principled approach to the training of neural networks within the Bayesian approach, for the case of unequal prior class distributions, is provided.

The Bayesian framework for neural networks (MacKay, 1992b) is widely used as a practical methodology for binary classification. Inherent in this framework is the prediction of distributions for the network weights, such that the estimation of the conditional probability of class membership is a marginalisation over the probability density for the weights.

$$y_g(x;\mathbf{w}) = P(Class|x,D) = \int y(x;\mathbf{w})p(\mathbf{w}|D)d\mathbf{w} \qquad (1)$$

where $y(x;\mathbf{w})$ is the actual network output and $y_g(x;\mathbf{w}) \equiv$ *P(Class x,D)* is the conditional. In general, this integral is not analytic due to the non-linearities in the output response

$$y(x) = g(a(x)), g(a) = 1/(1+\exp(-a)), \qquad (2)$$

therefore it is approximated to

$$y_g(x) \approx g\left(\frac{a(x)}{\sqrt{1+\pi s^2(x)/8}}\right), \qquad (3)$$

where *a(x)* represents the output node activation in response to input pattern *x*, and $s^2(x)$ is the associated variance. Whether or not this approximation is employed, the effect of uncertainty in the weights remains the same, namely to moderate the network predictions towards the midrange.

However, the presence of uncertainty should moderate the conditional towards the best unconditional estimate, which is not necessarily the midrange but the class prior. This chapter shows that marginalising to the prior specifies a preferred cost function weighting scheme, which has the effect of substantially reducing the bias in the estimates of the class conditional obtained from Bayesian neural networks.

5.2. Methodology

It is proposed that the estimate of the conditional using the evidence approximation should be obtained in two stages. First, by training the network to equal priors, then, using Bayes' theorem to refer the estimates of the conditional back to the true priors. The prior distribution seen by the model can be equalised by weighting the cost function (Lowe and Webb, 1991). The log-likelihood is expressed:

$$LL = -\sum_{samples}\left(t\log(y).\left(1/_{2P(Class)}\right) + (1-t)\log(1-y).\left(1/_{2(1-P(Class))}\right)\right) \qquad (4)$$

updating the gradient and Hessian calculations in the same way. The network estimates of the conditional are then

$$\tilde{y}_g(x) = \int \tilde{y}(\mathbf{w};x)\tilde{p}(\mathbf{w}|x)\partial\mathbf{w} \cong g\left(\frac{\tilde{a}(x)}{\sqrt{1+\pi\tilde{s}^2(x)\ 8}}\right) \qquad (5)$$

where the tilde denotes equal priors. These estimates are sometimes obtained by sub-sampling the data, or even by augmenting the data using noisy re-samples (Lee, 1999). Our approach utilises all of the actual data, and avoids the need for further artificial data. The correct estimate of the conditional is readily obtained from Bayes' theorem in the usual way,

$$P(Class|x) \cong y_g(x) = \frac{\tilde{y}_g(x)P(Class)}{\tilde{y}_g(x)P(Class) + \left(1-\tilde{y}_g(x)\right)\left(1-P(Class)\right)}, \qquad (6)$$

which has the required property that $y_g(x) \to P(Class)$ when $\tilde{y}_g(x) \to 0.5$. This transformation corresponds to a linear shift of the log-odds ratio of the balanced distribution, which is the network activation, by an amount given by the log-odds of the class priors, since

$$\log\left(\frac{y_g(x)}{\left(1-y_g(x)\right)}\right)=\log\left(\frac{\tilde{y}_g(x)}{\left(1-\tilde{y}_g(x)\right)}\right)+\log\left(\frac{P(Class)}{\left(1-P(Class)\right)}\right). \tag{7}$$

Therefore, the above procedure can be interpreted as estimating the distribution of the network output activations near the region of highest slope for the output response, rather than on the tail of the sigmoid.

5.3. Description of the experiment and results

The effect on the network calibration, of marginalising to an appropriate class prior, is illustrated in this section with the classification problem described in the previous chapters, involving the prediction of propensity to buy online. For the description of the data we refer again to chapters 2 and 3. Figures 5.1 and 5.2 show the calibration curves which result from the standard Bayesian-ARD model, trained *directly on the skewed data*, and then using the proposed marginalisation to the empirical class prior with an identical network structure. Shown are the mean and standard deviation of test results from 20 networks trained on 895 purchasers and 194 non-purchasers, and tested on as many out-of-sample records (i.e. $P(Class_{PUR}) \cong 0.82$ and $P(Class_{NON\text{-}PUR}) \cong 0.18$) The calibration accuracy is quantified by measuring the RMS error between the empirical calibration and the ideal line, digitised into intervals indexed by *i*,

$$Acc_{CAL}=\sqrt{\sum_{i=1}^{noi}\left(cal(i)-y(i)\right)^2 p\left(y(i)\right)}\ , \tag{8}$$

where *noi* is the number of output intervals and $p(y(i))$ is the proportion of observations that belong to interval *i*.

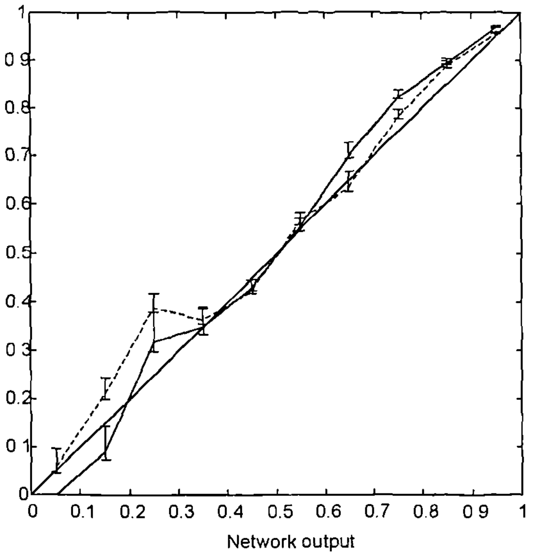


Figure 5.1: *Direct modelling of skewed data: calibration of the original network outputs in the prediction of propensity to purchase online (dashed line) and following marginalisation to the midrange (solid line). A measure of the calibration error is the RMS distance for the calibration line to the diagonal, which is 0.034 before marginalisation, and becomes 0.042 afterwards.*

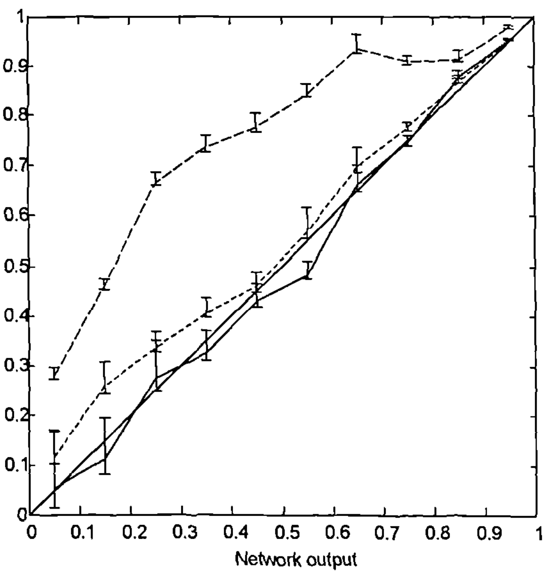


Figure 5.2: *Compensation for uneven priors: calibration of the original network outputs weighting the cost function towards balanced data (long dash), and following referral to the assumed class prior (short-dashed line) and marginalisation (solid line). The calibration accuracies before and after marginalisation are now 0.045 and 0.032.*

It is clear from the numerical results quoted in the figure captions, that the default Bayesian neural network procedure does not help the calibration.

Figures 5.3 and 5.4 show, in turn, the class-specific accuracies that result from the standard Bayesian-ARD model trained *directly on the skewed data*, and then using the proposed marginalization to the empirical class prior with an identical network structure. The undesirable effect of the marginalization to the midrange is clearly observed here, as the crossing point of accuracy shifts the wrong way with the first method (figure 5.3) whereas it remains well placed with the second (figure 5.4).

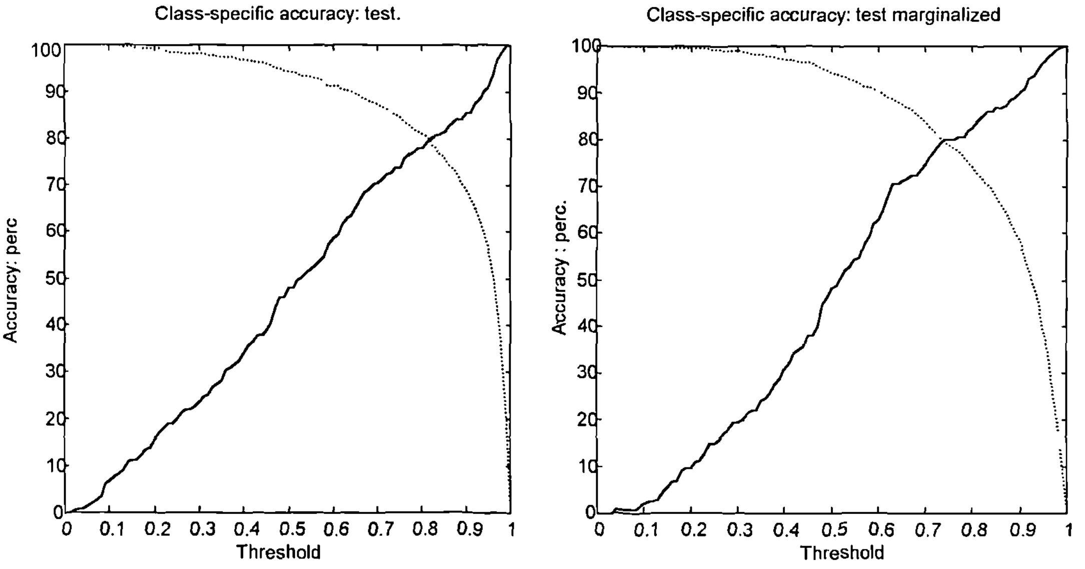


Figure 5.3: *Accuracy test results for each of the classes separately. Solid lines represent the class of non-purchases and dotted lines the class of purchasers. The figure on the left-hand side reproduces results before marginalization and the one on the right after marginalization. The marginalization swifts the crossing point towards the midrange, moving it away from P(Class$_{PUR}$) ≅ 0.82.*

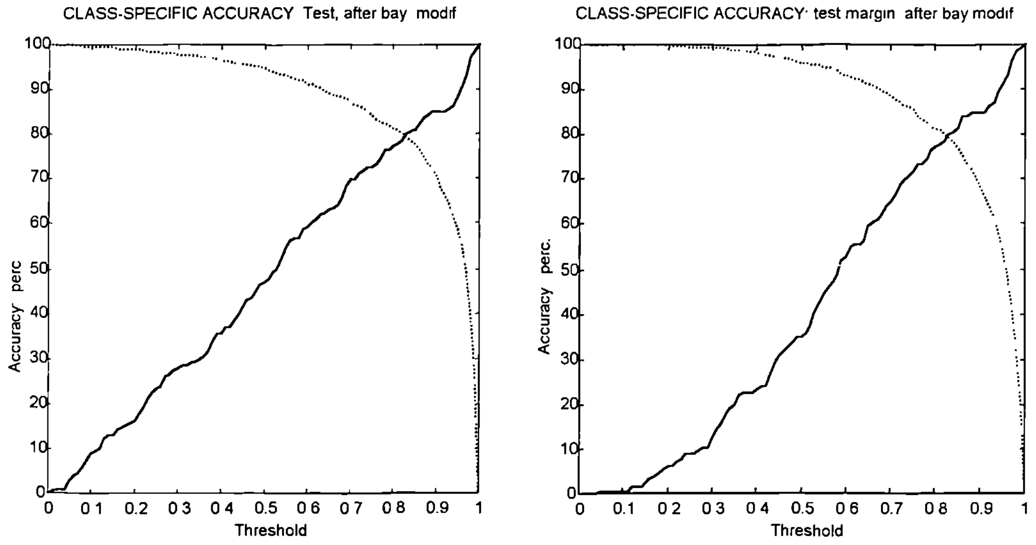


Figure 5.4: Accuracy test results for each of the classes separately. Solid lines represent the class of non-purchases and dotted lines the class of purchasers. The figure on the left-hand side reproduces results before marginalization and the one on the right after marginalization. This time, the marginalization with the method put forward in this chapter does not swift the crossing point towards the midrange but keeps it in the vicinity of P(Class$_{PUR}$) $\cong 0.82$.

5.4. Conclusion

Marginalisation to the class prior in Bayesian neural networks trained with unequal class prevalences, minimises the risk of biasing the calibration, which results from defaulting towards the midrange of the network output.

The procedure introduced in this chapter effectively helps to avoid introducing bias into the model predictions as, in all binary classification problems with skewed data, the default marginalisation introduces a systematic shift away from correct calibration. Moreover, the proposed application of the evidence framework avoids the need to artificially change the sampling procedure in order to equalise the priors, and indicates a preferred weighting scheme for the log-likelihood cost function.

5.5. Publications related to this chapter

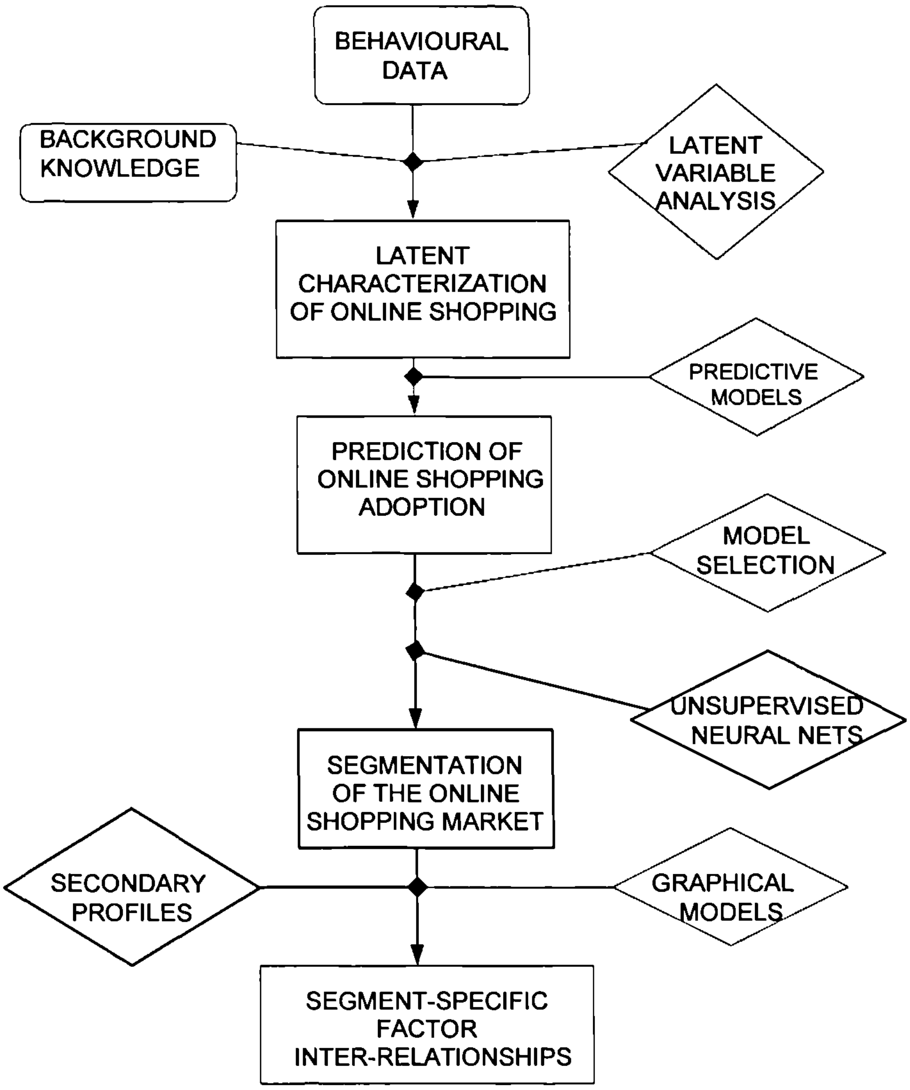
1. Lisboa, P.J.G., Vellido, A., & Wong, H. (2000): Bias reduction in skewed binary classification with Bayesian neural networks. *Neural Networks*, 13, 407-410.

PART 4

Segmentation of the online consumer market

Times on times he divided and measur'd
Space by space in his ninefold darkness,
Unseen, unknown; changes appear'd
Like desolate mountains, rifted furious
By the black winds of perturbation.

"The Book of Urizen". William Blake


BEHAVIOURAL DATA
BACKGROUND KNOWLEDGE
LATENT VARIABLE ANALYSIS
LATENT CHARACTERIZATION OF ONLINE SHOPPING
PREDICTIVE MODELS
PREDICTION OF ONLINE SHOPPING ADOPTION
MODEL SELECTION
UNSUPERVISED NEURAL NETS
SEGMENTATION OF THE ONLINE SHOPPING MARKET
SECONDARY PROFILES
GRAPHICAL MODELS
SEGMENT-SPECIFIC FACTOR INTER-RELATIONSHIPS

Preface

Two stages of a general, systematic methodology for the quantitative analysis of Internet users opinions of online shopping have already been outlined and developed: latent variable characterization and prediction of the propensity to buy online. This part of the thesis deals with another aspect of the methodology deemed to be of paramount importance: the segmentation of the business-to-consumer e-commerce market.

This part of the thesis is structured in three chapters. The first, **chapter 6,** introduces the problem of market segmentation, and the unsupervised neural network-based model proposed to address it: the Generative Topographic Mapping (GTM), a statistically principled alternative to the Self-Organizing Map (SOM). **Chapter 7** is a general case study of e-commerce market segmentation, using the model described in the previous chapter. **Chapter 8** presents a development of the GTM: Selective Mapping Smoothing (SMS), a result of the use of the Bayesian approach with multiple regularization for model training.

Chapter 6

The Generative Topographic Mapping as a principled model for e-commerce data visualization and market segmentation.

6.1. Introduction

The process of extracting knowledge from our data involves the discovery of those patterns of interest that might be present in it. One type of such patterns is the clusters in which the data points are grouped (Chen *et al.*, 1996). In the context of Internet retailing, the identification of clusters of consumer types has been stated as "the most important use of data mining, as this type of information is useful in a myriad of other planning and development tasks" (Slater *et al.*, 1999).

The interactive nature of the Internet requires a readjustment of the assumed relationship between marketers and customers. They become bound to establish both an economic and a social contract, as the transactions in this medium cannot be discrete but necessarily relationship exchanges (Hoffman *et al.*, 1999). Consumers are empowered by the information richness of the medium, which enables unprecedented availability of access to most aspects of the shopping process. Comparison of products, services and their prices are within everybody's keystroke reach, whereas opening times or geographical limitations are hardly an issue anymore. In this new commercial framework, those issues entail deep transformations of the marketing mix (Gordon and De Lima-Turner, 1997; Brannback, 1997) and on-line vendors have to make an extra effort to differentiate their offer and attract the on-line browsers towards their electronic outlets. Market segmentation techniques, grounded on the benefits sought by the customers, can give the marketer a leading edge: the identification of such segments can be the basis for effective targeting, enabling the redirection of made-to-measure content towards the customer.

The Internet is a paramount medium for the gathering and dissemination of information on consumers' habits of usage. Furthermore, the ever-increasing computer capabilities for electronic data storage and processing, coupled with the newest data mining and warehousing techniques, pose an unmatched opportunity for extracting knowledge out of the on-line shopping process. Market segmentation is just one of the ways in which such knowledge can be represented. The uniqueness of the on-line shopping channel has generated an interesting controversy about the way market segmentation should be conducted, if used as a tool for targeting. It confronts two opposite views: advocates of personalized and one-to-one marketing argue that it is possible, in the Internet medium, to take to its extreme the segmentation rationale, reaching and targeting "segments of one". This should be placed in the context of a post-modern model of markets in increasing fragmentation (Firat and Shultz, 1997) and ultimately justified by the expected benefits of "reaching individual consumers in order to satisfy their unique needs and wants in the best way" (Kara and Kaynak, 1997). Against this view, several of its shortcomings have been highlighted, as for instance the lack of economic viability that entails the creation of massive amounts of customized content, and the costs of maintaining, over time, the service for personalized interaction. Also, the potentially counterproductive breach of the privacy sphere involved in the pursuit of detailed information about individuals. The neural network-based models utilized in this and the next chapters can be helpful tools to bridge the gap between these controversial points of view, as they can accommodate cluster-based segmentation strategies of different levels of segment detail.

Market segmentation techniques frequently combine quantitative and qualitative methods, but it has recently been recognized that their design and deployment can be more clearly grounded in a sound statistical framework (Wedel and Kamakura, 1998). This chapter introduces the Generative Topographic Mapping (GTM), a non-linear latent variable model devised and developed by Bishop, Svensén and Williams (1998a; 1998b), as a model for data clustering and visualization that responds to these demands. This is combined with a non-linear predictive model, to form a principled quantitative methodology for market segmentation.

A key issue in the design of statistical frameworks for data analysis is the reproducibility and robustness of the results obtained. Non-linear models, in particular, do not replicate well when applied repeatedly even to the same data, unless they are implemented within a statistically constrained framework. The probabilistic formulation of the GTM is consistent with the methodology of mixture models proposed by Wedel and Kamakura (1998). It also makes possible to extend its basic description in a principled manner, making all the modelling assumptions explicit.

The chapter is organized in the following manner: A review of the exiting literature on segmentation of the online market is presented in the second section. The GTM model is then described in the third section, where its main advantages over the SOM are summarised. Section 6.4 outlines the rationale for proposing the GTM as a principled model for data visualization and market segmentation. The chapter is then closed by a section of general conclusions.

6.2. Segmentation of the on-line market: Review of the literature

Market segmentation has been highlighted as one of the important and necessary avenues of research needed in the field of electronic commerce (Chang, 1998) and as an issue that Internet vendors cannot ignore (O'Connor and O'Keefe, 1997). Despite this, only a few studies have addressed the subject: Gordon and De Lima-Turner (1997) examined how Internet users make a trade-off among attributes associated with Internet advertising policy, within the framework of a theory of *social contract* previously applied to direct mail marketing. Segmentation was carried out according to the differences in the way the on-line customers balance their privacy and interests. Several segments were found, including: *Complacent Customers, Efficiency Experts and Agitated Activists*, concluding that the majority of the Internet users appear to take a passive approach towards privacy issues on Internet.

The inherently world-wide structure of the Internet matches naturally with an international market segmentation perspective. McDonald (1996) explored the Internet usage motivations of consumers of several countries, aiming to find segments that transcended national boundaries. The study reveals the existence of a small *Social Shoppers* group, stable across geographical location; this segment, though, is outsized by others such as *Entertainment Seekers*, *Fact Collectors,* and *Avid Adventurers*.

6.3. The Generative Topographic Mapping model

In this section, the main principles and properties of the Generative Topographic Mapping model are briefly described, followed by a summary of its advantages over the model to which it intends to be a principled alternative: the Self Organizing Map (SOM; Kohonen 1982, 1995).

6.3.1. Principles of the Generative Topographic Mapping

The Generative Topographic Mapping (GTM) (Bishop *et al.*, 1998a, 1998b) is a non-linear latent variable model that generates a probability density in the multi-dimensional data space, using a set of latent variables of smaller dimension. This non-linear mapping is described by the generalized linear regression model

$$\mathbf{y} = \mathbf{W}\phi(\mathbf{u}) \, , \qquad (1)$$

where **u** is an L-dimensional vector of latent variables, **W** is the matrix that generates the explicit mapping from latent space to an L-dimensional manifold embedded in data space, and ϕ is a set of R basis functions which, in this study, are chosen to be Gaussians. For the non-linear mapping to remain analytically and computationally tractable, and also to elaborate a principled alternative to the SOM, the prior distribution of **u** in latent space is defined as a discrete grid, similar in spirit to the grid of the SOM

$$p(\mathbf{u}) = \frac{1}{M}\sum_{i=1}^{M}\delta(\mathbf{u} - \mathbf{u}_i), \tag{2}$$

where M is the number of its nodes. Since the data do not necessarily lie in an L-dimensional space, it is necessary to make use of a noise model for the distribution of the data points **x**. The integration of this data distribution over the latent space distribution, gives

$$p(\mathbf{x}|\mathbf{W},\beta) = \int p(\mathbf{x}|\mathbf{u},\mathbf{W},\beta)p(\mathbf{u})d\mathbf{u} = \frac{1}{M}\sum_{i=1}^{M}\left(\frac{\beta}{2\pi}\right)^{\frac{D}{2}} \exp\left\{-\frac{\beta}{2}\|\mathbf{m}_i - \mathbf{x}\|^2\right\}, \tag{3}$$

where D is the dimensionality of the input space, and $\mathbf{m}_i = \mathbf{W}\phi(\mathbf{u}_i)$ for the discrete node representation (2), according to expression (1). Using the SOM terminology, $\mathbf{m}_i$ can be considered as *reference vectors*, each of them the centre of an isotropic Gaussian distribution in data space (Bishop *et al.* 1998b). A log-likelihood can now be defined as

$$L(\mathbf{W},\beta) = \sum_{n\ 1}^{N}\ln p(\mathbf{x}^n|\mathbf{W},\beta) \tag{4}$$

for the whole input data set $\{\mathbf{x}^n\}$.

The distribution (3) corresponds to a constrained Gaussian mixture model (Hinton *et al.*, 1992), hence its parameters, **W** and β, can be determined using the Expectation-Maximization (EM) algorithm (Dempster *et al.*, 1977), details of which can be found in Bishop *et al.* (1998a). As part of the Expectation step, the mapping from latent space to data space, defined by (1), can be inverted using Bayes' theorem so that the posterior probability of a GTM node *i*, given a data-space point, is defined as

$$R_i^n \equiv p\left(\mathbf{u}_i \middle| \mathbf{x}^n\right) = \frac{\exp\left[-\frac{\beta}{2}\left\|\mathbf{m}_i - \mathbf{x}^n\right\|^2\right]}{\sum_{i'}^{M} \exp\left[-\frac{\beta}{2}\left\|\mathbf{m}_{i'} - \mathbf{x}^n\right\|^2\right]} . \qquad (5)$$

This is known as the *responsibility* taken by each node *i* for each point *n* in the data space. It will prove itself extremely useful, given a 2-dimensional latent space, for data visualization and also for cluster analysis in the context of market segmentation.

The complexity of the mapping generated by the GTM model is mainly controlled by the number and form of the basis functions. Further control of this effective complexity can be achieved with the addition of a regularization term to the error function (4), in such a way that the training of the GTM would consist of the maximization of a *penalized* log-likelihood

$$L_{PEN}(\mathrm{W},\beta) = \sum_{n\;1}^{N} \ln p(\mathbf{x}_n | \mathrm{W},\beta) + \frac{1}{2}\alpha\left\|\mathbf{w}\right\|^2, \qquad (6)$$

where $\mathbf{w}$ is a vector shaped by concatenation of the different column vectors of the weight matrix $\mathbf{W}$. This regularization term is effectively preventing the GTM to fit the noise in the data and is used under the assumption that there exist an underlying data generator which is a combination of the density functions for each of the segments.

The optimum values for all these complexity-controlling parameters should ideally be evaluated in a continuous space of solutions. Given that the GTM is formulated within a probabilistic framework, this can be accomplished using the Bayesian formalism and, more specifically, the evidence approximation (MacKay, 1992a). The application of this methodology (Bishop *et al.*, 1998b) produces update formulae for the regularization coefficient α and for the inverse variance of the noise model β.

Once the parameters α and β have been adaptively optimized, the best GTM model (in the sense that it reaches the best compromise between fitting the data and representing the underlying distribution from which the data were generated) can be obtained by experimenting with different combinations of the number of Gaussian basis functions and its width, σ. The number of basis functions can also be dealt with in a continuous space of solutions as part of the Bayesian approach. This development is considered in chapter 8.

6.3.2. Advantages of the GTM over the SOM model

The main advantage of the GTM over the SOM model is that the former generates a density distribution in data space so that the model can be described and developed within a principled probabilistic framework in which all the modelling assumptions are made explicit. The GTM also provides the well-defined objective function described in (4); its maximization using either standard techniques for non-linear optimization or the EM-algorithm has been proved to converge. As part of this process, the GTM *learning parameters* calculation is grounded in a sound theoretical basis. One of the main limitations of the SOM model is the impossibility to define such an objective function. Finally, the *magnification factor* for the GTM (described in section 6.4.2) can be calculated as a continuous function of the latent variables, avoiding the discrete approximation to which the SOM is limited.

6.4. The GTM as a principled model for data visualization and market segmentation

In this section we intend to put the Generative Topographic Mapping (GTM) forward as a principled model for data visualization and market segmentation. We address the fundamental questions of statistical validation of the clustering model and its suitability to the requirements of the segmentation techniques.

6.4.1. Justification of the GTM as a principled model for data visualization and market segmentation.

The GTM, as a non-linear latent variable model based on a constrained mixture of Gaussians, whose parameters are estimated with the EM algorithm, is defined within a sound statistical framework. Referring generally to mixture of distribution models, Wedel and Kamakura (1998: p.33) argue that "the statistical approach clearly is a major step forward in segmentation research." That is in contrast with traditional non-overlapping clustering algorithms used for segmentation, grouped in two main types: *hierarchical* (e.g. Ward's method) and *nonhierarchical* (e.g. K-means). These algorithms, as well as a neural network-based model such as the SOM, are limited by its heuristic nature and their results can not be justified by standard statistical theory.

One of the consequences of the definition of the GTM within a statistical framework is that the posterior probability (5) of each component of the mixture of Gaussians (or node in the latent space), given the data, can be calculated. Each of the nodes in latent space can be interpreted as a cluster, as with the SOM model (Ripley, 1996). In the context of market segmentation, these clusters correspond to segments of the market. Therefore, the GTM model defines a posterior probability of cluster membership for each cluster and each data point. Assuming that each data point / individual belongs only to one cluster (non-overlapping clustering) but the information it conveys is insufficient to uniquely assess its cluster membership (McLachlan and Basford, 1988; Wedel and Kamakura, 1998), the GTM can then be used as a *fuzzy-clustering* tool.

The GTM defines a mapping from latent to data space. This mapping can be inverted to give rise to the posterior distribution (5), providing the model with the data visualization capabilities that other models that project high-dimensional data into a visualization space possess. The mapping defined by (1) ensures the insightful property of *topographic ordering*: any two points that are close in latent space will be mapped to points that are necessarily close in data space. Consequently, neighbouring clusters in latent space correspond to groups of data that are also close in data space. Should the data sets contain information for the same entities at different moments in time, the continuity of this mapping could be used for visualizing their evolution through the map of clusters and even for hypothesizing about their most likely future *segment trajectories*.

This preservation of the topographic order, naturally brings about another feature of the GTM which is most relevant for its implementation as a market segmentation tool: besides the definition of each node in latent space as a cluster / segment, these nodes can be aggregated in a principled way to form macro-clusters. Therefore, the GTM provides a way to address the always contentious issue of the level of detail or *granularity* with which the segmentation should be carried out. The way the GTM enables the aggregated segmentation is described next.

6.4.2. Magnification factors for the GTM

The GTM maps points from a regular grid in latent space into a multi-dimensional data space. In doing so, a region in latent space will undergo a distortion. The regularity of the latent grid and hence of the visualization map will not necessarily reflect the separation of natural groupings as it is in the data space. The same problem affects the SOM algorithm and, in order to tackle it for this model, Ultsch (1993) and Kraaijveld *et al.* (1995) suggested visualization strategies based on maps of pseudo-colour levels representing distances between *code-book* vectors (similar to the GTM *reference vectors*). Definite macro-clusters are expected to be indicated by separate groupings of close *code-book* vectors, which can be visually assessed. In the context of the GTM, Bishop *et al.* (1997a) have addressed this problem making use of the concept of the *magnification factors*. It is at this point that the GTM definition of a probability density, in the form of a lower-dimensional manifold embedded in data space, reveals its importance: the local *magnification factor* can now be calculated, resorting to differential geometry, as a continuous function of the latent space variables.

For the sake of brevity, only the main results from Bishop *et al.* (1997a) will be presented here. Given a two-dimensional latent space, the GTM will map an infinitesimal rectangle with area $dA = \prod_i dx^i$ from this space into another infinitesimal rectangle in the manifold embedded in data space and defined by (1), with area dA'. The relation between those areas and, therefore, the *magnification factor* is proved to be

$$\frac{dA'}{dA} = \det{}^{\frac{1}{2}}\left(g_{ij}\right) , \tag{7}$$

where g_{ij} is the metric tensor defined as

$$g_{ij} = \delta_{kl} \frac{\partial y^k}{\partial x^i} \frac{\partial y^l}{\partial x^j} \tag{8}$$

In matrix form, and making use of (1), the expression (6) can thus be written as

$$\frac{dA'}{dA} = \det{}^{\frac{1}{2}}\left(\psi^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}\psi\right), \tag{9}$$

where ψ is a matrix of partial derivatives of the basis functions ϕ with respect to the latent variables. Now the magnification factor $\frac{dA'}{dA}$ can be plotted in the latent space visualization map, using a pseudo-colour representation. The map areas indicate varying mapping distortions, with extreme shades in a grey-scale representation representing very large or very small distortions. These can be cautiously associated to inter-cluster or within-cluster regions. As a result, the macro-clusters that are a pre-requisite for aggregate segmentation can be defined.

6.5. Conclusions

A quantitative segmentation methodology based on the Generative Topographic Mapping (GTM) has been proposed. It addresses the need for models that reach a compromise between the definition of a finite set of homogeneous segments and the continuous distribution of heterogeneity in populations, identifying segments but also allowing for a certain degree of heterogeneity within each of them (Wedel and Kamakura, 1998: p.338). This problem of "continuous distribution of heterogeneity versus market segments" (Wedel and Kamakura, 1998: p.331) confronts two different views. The first being where the heterogeneity amongst consumers is not so pronounced that the possibility of grouping them into segments is rejected. The second would consider the partition of the consumer representation continuum into segments as an artefact (Allenby and Lenk, 1994), thus assuming that markets are perfectly heterogeneous. The latter is the basis for any attempt to define models of *one-to-one* marketing or *finer* segmentation (Firat and Shultz, 1997; Kara and Kaynak, 1997). The GTM grid of nodes / clusters in latent space can be modified at will: either augmented so that the distribution of cases in the visualization map becomes sparser (although the differences between nodes / clusters will be smaller), or reduced so that all cases agglomerate in a few clusters (more heterogeneous).

The probabilistic formulation of this non-linear latent variable model provides estimates for the posterior probability of cluster/segment-membership for each individual, enabling it to be used for micro-segmentation and as a fuzzy clustering tool. Moreover, the GTM is naturally integrated into the family of mixture models proposed by Wedel and Kamakura (1998), providing it with an extended context in segmentation research.

6.6. Publications related to this chapter

1. Vellido, A., Lisboa, P.J.G. & Meehan, K. (1999): Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications,* 17(4), 303-314.

2. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Characterizing and segmenting the business-to-consumer e-commerce market using neural networks. In Lisboa, P.J.G., Vellido, A. and Edisbury, B. (Eds.): *Business Applications of Neural Networks.* Singapore: World Scientific, 29-54.

Chapter 7

Segmenting the consumer e-commerce market using the Generative Topographic Mapping: a case study

7.1. Introduction

In the chapters of part 3 of the thesis, a variable selection model, ARD, associated to Bayesian neural networks, was applied to the factor scores resulting from the factor analysis of the GVU data in order to find which amongst them bore maximum predictive power as to predicting the propensity to buy online. The five most relevant factors were labeled according to previously published qualitative studies, and they are again summarized in *table 7.1,* to ease the interpretation of the results. The segmentation properties of the Generative Topographic Mapping (GTM) model, described in the previous chapter, will be illustrated in this one with the very same data.

The application of variable selection methods (chapter 3) to these factors yielded a ranking, according to which factor 2 (from now on referred to in this study as *Risk Perception*) in *table 7.1* bears most of the predictive power of the model, followed by factors 1 and 3 (referred to as *Compatibility* and *Affordability*) and finally, by the not-so-relevant factors 4 and 5 (referred to as *Ease of use* and *effort responsiveness*).

The chapter is organized in the following manner: section 7.2 provides a description and defense of our *tandem approach* to market segmentation. This is followed by a general section of model implementation and presentation of the results. The chapter is closed by a section of conclusions.

FACTOR	DESCRIPTION	*ATTRIBUTES*
1	*Shopping experience:* *Compatibility*	*Control and convenience*
2	*Consumer risk perception* *Environmental control*	*Trust and security*
3	*Affordability*	--
4	*Shopping experience: Effort*	*Ease of use*
5	*Shopping experience /* *Customer service*	*Effort /* *Responsiveness and empathy*

Table 7.1: Descriptive summary of the factors selected by the ARD model as predictive of online purchasing behaviour.

7.2. The tandem approach to market segmentation

The practice of factor analysing the observable data, followed by the clustering of the obtained factor scores is known, in the marketing literature, as the *tandem approach* towards segmentation. Factor analysis helps to overcome the limitations associated with survey-based segmentation studies, i.e. noisy data, poorly measured variables based on subjective ratings, and unbalanced item selection across the domain of the surveyed constructs (Green and Krieger, 1995). Some authors have argued against the tandem approach on the basis that it might discard relevant information and distort the true cluster structure of the data (Arabie and Hubert, 1994). Schaffer and Green (1998), alternatively, suggest either the clustering of the observable variables or, compromising to preserve the dimensionality reduction, the clustering of non-standardized, non-rotated principal component scores. Nevertheless, even this "compromise" alternative does not provide an interpretation (labelling) of the factor structure.

Green and Krieger (1995) and Schaffer and Green (1998) acknowledge that a good validation of the segmentation results might stem from the relation of alternative clusterings to some profit-based measure. This provides a justification for our use of the *tandem approach*: in this study, the segmentation is carried out according to factors which have been shown to discriminate, quite accurately, the classes of *purchasers* and *non-purchasers*, each of which offers a profit/non-profit opportunity to the marketer. Therefore, although the GTM is a descriptive clustering tool, the whole model is closer to the *criterion-based* or *predictive* types of segmentation models such as CHAID (Magidson, 1994) or clusterwise regression (Wedel and Kistemaker, 1989). The GTM is expected to generate clusters that discriminate between consumers that belong to each of the classes defined by the dependent variable. This property has been observed in the SOM model (Serrano-Cinca, 1996, Vellido *et al.*, 1999b). The resulting cluster solution has an explicitly profit-based interpretation, as each cluster can be described and targeted in a marketing campaign with the combined knowledge of the latent factors that shape it and the propensity to buy which they predict.

7.3. Segmentation of the e-commerce market: Implementation and results

This section gives a brief account of implementation of the GTM, followed by the presentation and discussion of the segmentation results.

The model by which we illustrate the use of the GTM consists of a grid of 5x5 basis functions with a common width $\sigma = 1$. The GTM was trained with a class-balanced data set of 778 individuals, and the values for the regularization coefficient and the inverse variance of the noise model after training were, in turn, $\alpha = 1.12$ and $\beta = 1.57$. A fixed grid in latent space of 15x15 nodes was selected as a compromise on the level of detail or *granularity* of the cluster / segment solution.

The expression (5) in chapter 6 provides a posterior probability of cluster/segment membership for each individual. In order to visualize that probability for a complete set of data, the information has to be somehow summarized. This can be done (Bishop *et al.* 1998a) by calculating the *mean* of the distribution for each point $\mathbf{x}^n$ in data space

$$\langle \mathbf{u} | \mathbf{x}^n, \mathbf{W}^*, \beta^* \rangle = \sum_{i=1}^{M} R_{in} \mathbf{u}_i \,, \qquad (1)$$

where $\mathbf{W}^*$ and β^* are the values of $\mathbf{W}$ and β for the trained GTM. The *mode* of the distribution, given by

$$i^{\max} = \arg\max_{\{i\}} R_{in} \,, \qquad i = 1, \ldots, M \qquad (2)$$

can also be used. In this case, the information is visualized in the way that is usual with the SOM model.

7.3.1. Class discrimination

Figure 7.1 represents the whole data set by the *mode* (2), as mapped onto the nodes of the trained GTM. Each individual has been assigned a colour, black or white, depending on its *true* class membership: online *purchasers* or *non-purchasers*. Grey nodes indicate that individuals from different classes have been mapped onto them. Figure 7.1(a) shows that the GTM, *without any prior information on class membership* has managed to separate, quite clearly, the classes of *purchasers* and *non-purchasers*. Such a result partially justifies our semi-predictive approach (as described in section 7.2) of using factors shown to predict the propensity to buy online as the bases for segmentation.

This approach would be fully justified if it were shown that the original, observable variables, can not separate both classes with more accuracy than the factor description of the data. A criterion has to be defined to quantify that capability. We propose the following entropy-based measure, which represents the information that the GTM map contains about class membership (Kullback, 1968):

$$S1(C_1,C_2) = -\sum_{clusters} P(clusters) \sum_{classes} P(class|clusters) \ln P(class|clusters) = \\ -\sum_{i=1}^{M} \frac{N_i}{N}\left(p_i^1 \ln p_i^1 + p_i^2 \ln p_i^2\right) \quad , \qquad (3)$$

where C_1 and C_2 are, in turn, the classes of *purchasers* and *non-purchasers*, and p_i^1 and p_i^2 are the ratios of *purchasers* n_i^1/N_i and *non-purchasers* n_i^2/N_i mapped into the latent-space node i. N_i is the total number of individuals mapped into node i, so that

$N = \sum_{i=1}^{M} N_i$ is the number of individuals in the complete data set. M is the total number of latent-space nodes. The minimum entropy value is 0, which corresponds to the ideal case in which every node/cluster has only members of one of the classes mapped into it. The maximum entropy is $\ln(2) \cong 0.69$.

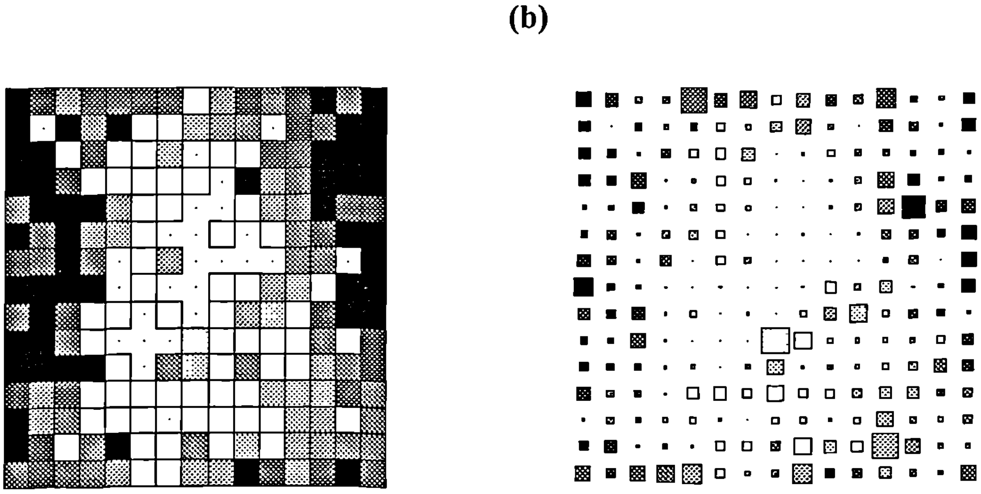


Figure 7.1: GTM class-membership maps. All the individuals in the training sample are mapped onto the nodes in the latent visualization space that correspond to their modes as defined by (2): a) Each individual has been shaded according to its class-membership: white for purchasers, black for non-purchasers. Those GTM nodes with more than one pattern mapped onto them are depicted in shades of grey, corresponding to the class-membership proportion. The nodes with no individuals mapped onto them are depicted as dots. b) The same pseudo-colour representation has been used, but now the relative size of the squares/nodes corresponds to the number of patterns mapped onto them. Regions of high pattern occurrence in latent space can thus be visualized.

The expression (3) will be evaluated for the three different GTM maps resulting from the use of: a) the 44 original observable variables (see chapter 2); b) the complete 9-factor solution prior to factor selection (see chapter 2); c) the 5-factor selection obtained in chapter 3.

Even more important than the entropy measure for the trained GTM itself, is the quantification of the generalization capabilities of this model, i.e. to what extent is it able to discriminate the classes of *purchasers* and *non-purchasers* from a sample of consumers with which the model has not been trained (a hold-out sample). Although the GTM can be a very powerful exploratory and descriptive tool, it would be of little help unless it was able to generalize, giving the marketer the possibility to assess the propensity to buy on-line of any future potential customers. Therefore, the entropy expression (3) for a hold-out sample will be calculated for the three GTM models described above. The results for both the training and the generalization models are summarized in table 7.2.

Entropy measures

	MODELS		
	5-factor	**9-factor**	**44-variables**
Training	0.3302	0.3504	***0.3187***
Generalization	***0.2317***	0.2882	0.2512

Table 7.2: *Values of the entropy measure (12) for the trained GTM (first row) and for the hold-out sample (second row). The latter are lower because the measure is sensitive to the total number of samples in the map and the size of the hold-out sample was, in this case, half the size of the training sample. The best results have been highlighted.*

The use of the original 44 variables yields slightly better training entropy results than any of the factor models. This can be put down to the fact that those 44 variables convey more information than any of the factor models. Nevertheless, the 5-factor selection model is shown to produce a lower generalization entropy, which can be justified on the basis that all the information that is redundant in terms of class-discrimination has been removed from this model. Therefore, the use of the *tandem approach* as described in section 7.2 can now be argued to be an appropriate segmentation methodology for market profit optimization.

7.3.2. Characterization of macro-clusters

Figure 7.1(b) conveys information about the size of each of the nodes / micro-clusters (number of individuals whose *mode* corresponds to each of them) showing agglomerations of data in specific areas. This might provide a starting point for the definition of the macro-clusters of an aggregate segmentation strategy. Lewis *et al.* (1997) propose, for the SOM model, to designate as macro-cluster centroids those nodes with the largest number of patterns mapped onto it. Each of the remaining nodes will be added to the macro-cluster with the closest centroid. The nodes with fewer patterns mapped into them could also give an indication of the boundaries between clusters (Zhang and Li, 1993).

These methods can be combined, whereas the *magnification factor*, described in section 6.4.2 of chapter 6 and shown in figure 7.2, can be used as a complementary source of information. In this thesis, we are considering a different methodology to define macro-clusters from the trained GTM. The strategy described in the previous paragraph relied on the assumption that a frequently *hit* map unit corresponds to a densely populated area in multi-dimensional space. Therefore we are only making use of the information contained in the *mode*, as defined by (2). Alternatively, all the information conveyed by the posterior probability of cluster membership defined in expression (5), from chapter 6, can be used. We now define the cumulative posterior probability of cluster membership or *cumulative responsibility* for each node i as

$$R_{CUM,i} = \sum_n R_i^n \equiv \sum_n p\left(\mathbf{u}_i \middle| \mathbf{x}^n\right) = \sum_n \frac{\exp\left[-\frac{\beta}{2}\left\|\mathbf{m}_i - \mathbf{x}^n\right\|^2\right]}{\sum_{i'}^{M} \exp\left[-\frac{\beta}{2}\left\|\mathbf{m}_{i'} - \mathbf{x}^n\right\|^2\right]} \tag{4}$$

Regions of the map with high values of $R_{CUM,i}$ will roughly correspond with regions of high *mode* occurrence, unless the distribution is strongly multi-modal. A further advantage of this criterion is that it also indicates the approximate extension of the macro-clusters. The reason is that sparse clusters of points in the multidimensional space are compacted in small areas of the GTM map. Consequently, the posterior probability of a point in such type of cluster is expected to be narrow and concentrated in only a few map units. On the other hand, the mapping of points from compact clusters in multidimensional space will not entail large distortions. These points will occupy larger regions of the map and their posterior probabilities will be broad.

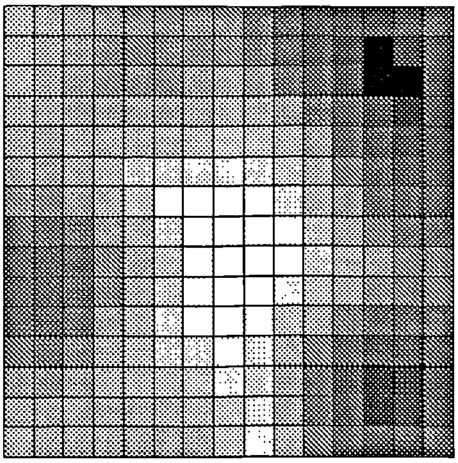


Figure 7.2: *Magnification factor map for the GTM, according to section 6.4.2 of chapter 6, in pseudo-colour representation. Light areas correspond to regions of low distortion in the mapping from latent into data space. Dark areas, consequently, correspond to regions of high distortion in that mapping.*

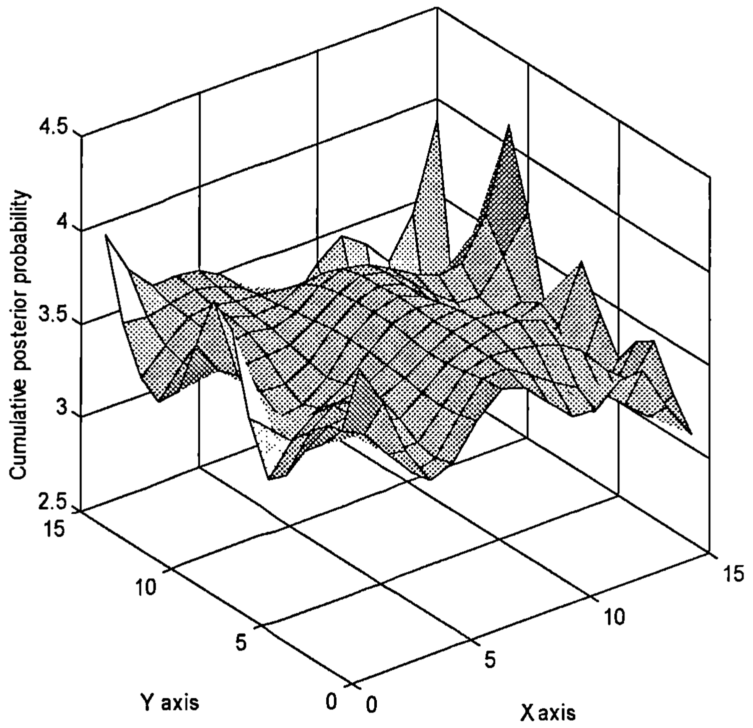


Figure 7.3: 3-dimensional plot of the cumulative posterior probability of cluster membership (cumulative responsibility) for the trained GTM. Peaks correspond to areas of data concentration and valleys correspond to regions of more sparse data presence. This probability distribution can be used as a benchmark for the results of the contiguity-constrained clustering algorithm. The X and Y axes correspond, in turn, to the horizontal and vertical axes in the rest of the GTM maps in this chapter.

Overall, a landscape representation of (4) will be characterized by broad plateaux, corresponding to compact clusters, and narrow high peaks, corresponding to sparse groupings in the data. Such representation for our data can be found in figure 7.3. Notice the resemblance of these contours with the magnification factor in figure 7.2. In fact, the cumulative responsibility implicitly contains the information provided by the magnification map.

7.3.2.1. Contiguity-constrained agglomerative clustering

We already have a reasonable idea of where the macro-clusters lie in the map. Now we intend to confirm that idea using a clustering algorithm on the *reference vectors* $\mathbf{m}_i = \mathbf{W}\phi(\mathbf{u}_i)$. This time there will be no *a priori* specification of the centroids. Following Murtagh (1995), a contiguity-constrained agglomerative algorithm is proposed. As described in section 6.4.1 (chapter 6), points which are close in the multidimensional space will be close in their latent space representation, hence the imposition of a contiguity condition. The algorithm will proceed as follows:

1. Each unit i of the map is initialized as a macro-cluster with its center at $\mathbf{m}_i$, and uniquely labelled.
2. At every step of the algorithm, the two closest neighbouring macro-clusters are merged. This merger entails substituting the previous centers with their mean.
3. Repeat step 2 until the macro-cluster partition approximates the representation in figure 7.3.

The distances in the algorithm are taken to be Euclidean. The contiguity or neighbouring condition in step 2 consists of only considering, as candidates to be merged, those macro-clusters that contain neighbour units in the map. For units neither in the edges nor in the corners of the map, the eight surrounding units are considered as its neighbours.

An example of that procedure is the 3-segment solution displayed in figure 7.4. It strongly resembles the map in figure 7.1(a), with two segments at the sides roughly corresponding to the class of *non-purchasers*, and a segment occupying the center of the map mainly corresponding to the class of purchasers.

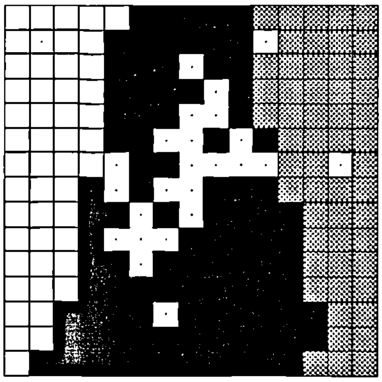


Figure 7.4: A 3-segment solution as an example of the clustering procedure described in the text.

7.3.2.2. Cluster description by means of reference vectors interpretation.

Those segments have to be interpreted in terms of the segmentation bases. For that purpose and, as defined in section 6.3.1 of chapter 6, the *reference vectors* $\mathbf{m}_i = \mathbf{W}\phi(\mathbf{u}_i)$ in data space, for each node *i* in the latent visualization space, will be used. Each of the variables shaping these vectors can be visualized using a *reference map*, which is a pseudo-colour representation of its numerical values. The reference maps for our trained GTM are shown in figure 7.5. The meaning of the numerical values (white for high, black for low in a grey-shaded palette) is described in table 7.3.

The segment on the left-hand side of figure 7.4 is dominated by low values of *risk perception* and *effort/responsiveness*, and high values of *affordability* (See table 7.1). It might be characterised as a segment of consumers who, seeing shopping on-line as affordable, are deterred by perceptions of security and economic risks. For the segment on the right-hand side of the map the *risk perception* seems to be attenuated, whereas the *affordability* becomes very low; *ease of use* and *effort/responsiveness* are also locally very low. The main characteristic of this segment seems to be that the consumers in it do not find shopping on-line economically or otherwise compatible with their current situation. Finally, the central segment, very compact according to the *magnification factor* in figure 7.2 and the *cumulative responsibility* in figure 7.3, has medium to high values for all the factors, indicating that, in general, it corresponds to consumers convinced of the advantages of shopping on-line. That is consistent with the fact that most of those individuals belong to the class of*purchasers.*

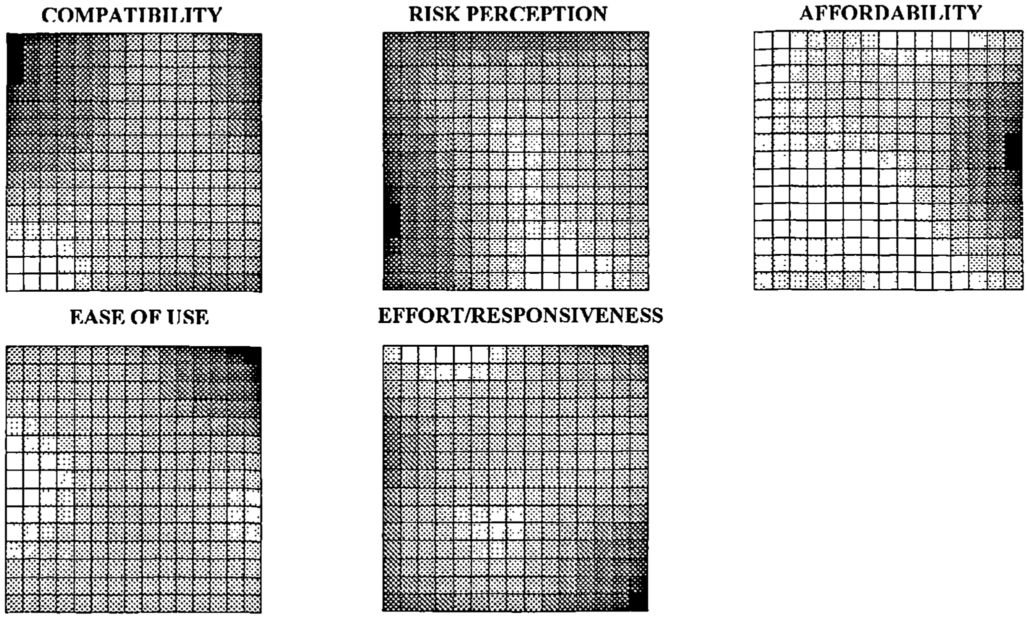


Figure 7.5: Reference maps of the trained GTM, associated with each of the factors in the 5-factor selection, in a pseudo-colour representation. Light shades of grey correspond to high values of the elements of the reference vectors, whereas dark colours correspond to low values. The interpretation of those high and low values is reported in table 7.3.

	POSITIVE VALUES	NEGATIVE VALUES
Factor 1 *Compatibility:* *Control / Convenience*	People who perceive shopping on the WWW as compatible and convenient, feeling they are in control of the shopping process.	People who do not perceive shopping on the WWW as either compatible or convenient, or do not feel in control of the shopping process.
Factor 2 *Risk perception:* *Trust / Security*	People who do not perceive that shopping online entails a differential risk. They find online vendors trustworthy.	People who perceive security and economic risks as major deterrents to shop online. They find online vendors untrustworthy.
Factor 3 *Affordability*	People who find the whole prerequisites for online shopping as affordable.	People who find the whole prerequisites for online shopping as unaffordable.
Factor 4 *Shopping Experience:* *Effort Ease of use*	People who find shopping online easy and unproblematic.	People who find shopping online as a complicated undertaking, difficult to learn, that requires a lot of mental effort.
Factor 5 *Shop.Exp.;Customer Service* *Effort Responsiveness*	People who consider that online vendors provide a responsive customer service, reducing the effort involved in the online shopping process.	People who do not consider that online vendors provide a responsive customer service.

Table 7.3: *Interpretation of the low and high (negative / positive) values of the factor scores (inputs to the model) and reference vectors of the trained GTM.*

7.3.2.3. Contribution of the factors to the segment structure

The influence of each of the factors on the overall cluster / segment structure can be further assessed (Kaski *et al.*, 1998) by measuring the contribution of each reference vector component to the total distance between map units, defined as

$$\left|m_{ik} - m_{jk}\right| / \left\|\mathbf{m}_i - \mathbf{m}_j\right\|, \qquad (5)$$

where i and j are two neighbouring units in latent space, and $k = 1,\ldots,5$ for our 5-factor selection. For each unit, these contributions can be averaged over all its neighbours and visualized using a pseudo-colour representation. The resulting *contribution maps*, in figure 7.6, complement the information conveyed by the previously displayed *reference maps*.

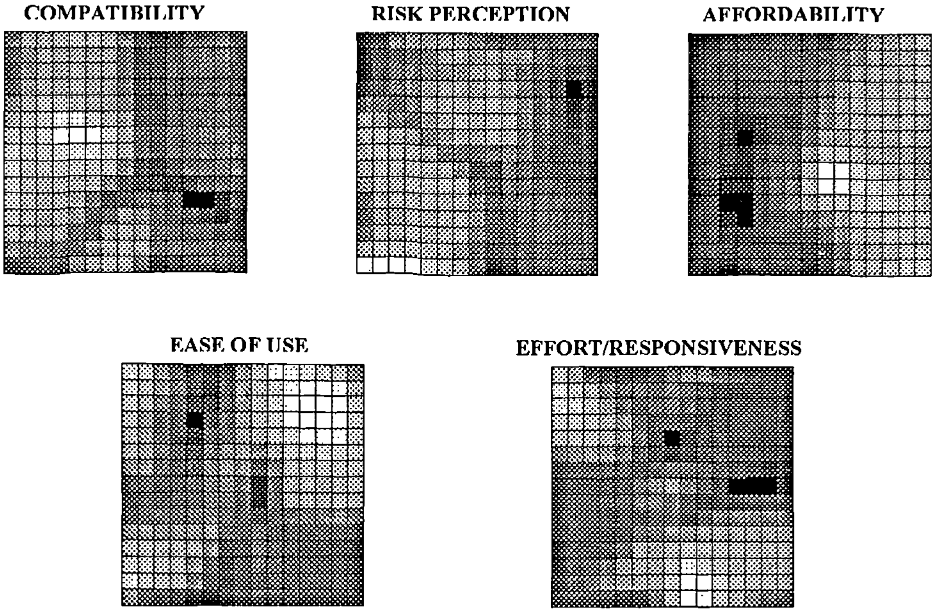


Figure 7.6: *Contribution maps for the trained GTM, describing the influence of each of the factors in the 5-factor selection on the cluster structure of each region of the map. The numerical values resulting from the expression (5) are visualized in a pseudo-colour representation. (Light-dark) shades of grey correspond to (high-low) values.*

7.3.2.4. A finer segmentation solution

A finer segmentation that fits the *cumulative responsibility* distribution shown in figure 7.3 reasonably well is exemplified by the 7-segment solution in figure 7.7. The interpretation of the segments according to the *reference* and the *contribution* maps, their labeling and sizes are summarized in table 7.4. Although, for the sake of brevity, not all the results are reported here, different segment solutions from 5 to 10 segments were investigated. A 10-segment solution, for instance, would split the lower part of segment 4 to produce an extra segment that, according to the reference maps, could be labeled as *Security and Cost Confident*. Furthermore, segment 5 would be split in 3 small groups, two of them in an intermediate position with respect to segments 3 and 6. This possible split of segment 5 will be further analyzed in the next section.

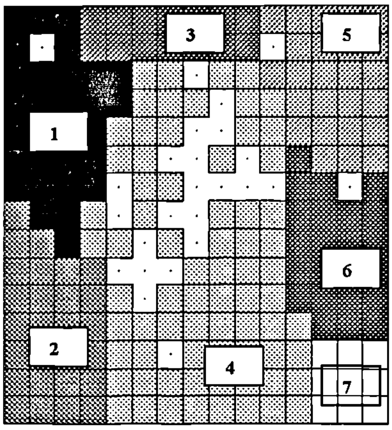


***Figure** 7.7: A 7-segment solution obtained with the clustering procedure described in the text. It is interpreted in table 7.4 with the help of the reference maps in figure 7.5.*

	SEGMENT DESCRIPTION	SEGMENT LABEL and SIZE (%)
1	Very low on *Compatibility* and rather low on *Perception of risk*, although rather high on *Affordability*. *(20.2% purchasers – 79.8% non-purchasers)*	*Unconvinced* (14.0%)
2	High *Compatibility* compounded with low values of *Perception of risk*. *(29.4% purchasers – 70.6% non-purchasers)*	*Security conscious* (13.1%)
3	Similar to the *Unconvinced* segment but scoring higher in the factor of *Compatibility*. *(51.5% purchasers – 48.5% non-purchasers)*	*Undecided* (8.7%)
4	All factors present medium-to-high values. Values of *Perception of risk* and *Affordability* are specially high. *(84.3% purchasers – 15.7% non-purchasers)*	*Convinced* (35.0%)
5	Low *Compatibility* and very low *Ease of use*. Medium to high *Affordability*. *(17.1% purchasers – 82.9% non-purchasers)*	*Complexity avoiders* (14.3%)
6	Most factors present medium values except *Affordability*, which is very low. *(45.2% purchasers – 54.8% non-purchasers)*	*Cost conscious* (10.8%)
7	This small group scores very low in *Effort/responsiveness*, but medium-to-high in the remaining factors. *(50% purchasers – 50% non-purchasers)*	*Customer service wary* (4.1%)

***Table** 7.4: Description of the 7-segment solution proposed in the text. Segments are numbered, according to figure 7.7, in the first column of the table and described in the second column according to table 7.3 (This column also includes the percentages of purchasers and non-purchasers present in each segment). The relative size and proposed label of each segment are included in the third column.*

7.3.2.5. Segment profiling

There are several criteria that define the feasibility of a market segmentation strategy. Amongst others, *substantiality* and *actionability*, which are met by the segment solution provided in this study. Another two criteria, *identifiability* and *accesibility* might sometimes depend on the availability of secondary information, such as demographic and socio-economic data. This type of information has been shown to be neither the most effective in developing segments (Wedel and Kamakura, 1998) nor a good predictor of the propensity to buy on-line (see chapter 3). Nevertheless, it might sometimes be the only kind of data available to the marketer.

We profile now the 7-segment solution provided in the previous section, making use of the following variables: *age, household income, years of Internet experience, average of hours a week of Internet usage* and *gender*. This selection of bases is by no means exhaustive. Bearing in mind that this is just an illustration of a procedure that ultimately depends on the availability of secondary information, and for the sake of brevity, the profiling, summarized in table 7.5, will be limited to the aforementioned variables. Similar segmentation variables have also been used in other studies (Gordon and De Lima-Turner, 1997).

Many conclusions can be drawn from these results. Table 7.5, though, is quite self-explanatory, so that only some general ideas will be sketched here. Firstly, all the background bases show significant differences over the clusters, as measured by a χ^2 test. Segment 4 (*Convinced*), mainly composed of*purchasers*, can be characterised as mostly male in their late twenties to early forties, with high-band income and Internet-savvy. Most of the segments with a majority of *non-purchasers* tend to be dominated by females, an interesting feature given the level of significance of the *gender* basis reported in the table. This is especially relevant in the case of the core non-purchaser segment 1, the *Unconvinced*, because it reveals women as the type of customers that find buying online incompatible with their life and shopping styles (See *reference maps*, figure 7.5)

The division of segment 5, the *Complexity Avoiders*, into 3 sub-segments seems strongly justified by their very different profiles. Sub-segment 5a is the top-left part of segment 5 and is close to segment 3, the *Undecided*, which makes it a somehow "softer" target for potential marketing campaigns. Its profile is most interesting: quite older than the rest of segment 5, more Internet-savvy and with very high average income level. It is also more male-dominated. Sub-segment 5c, in the top-right corner of the map is the hard core of the segment, with a profile that is extremely young, barely Internet experienced and with the lowest average income in the map. Sub-segment 5b is half-way between the other two sub-segments and presents a strongly female profile. Segment 6, the *Cost Conscious*, add rather lower incomes (which is consistent with the low values of the *affordability* factor that the individuals in this segment present) to lack of Internet experience. Segment 3 (*Undecided*) are similar to the *Cost Conscious*, with a young profile and average-to-low Internet usage and experience. Their main difference lies in the higher income distribution of the *Undecided*, which makes this segment highly attractive to marketers. Segment 7

(*Customer Service Wary*) is almost a variant on the *Unconvinced*, only younger and with lower Internet usage.

The levels of Internet usage and experience seem to be related to the membership of segments dominated by *purchasers*, which is reinforced by the results of the χ^2 test for the corresponding background variables reported in table 7.5. This can be understood in the light of the *flow construct*, developed for the Web computer-mediated environment by Hoffman and Novak (1996) to describe the state of mind induced by the Web navigation.

		Whole sample	**Seg. 1**	**Seg. 2**	**Seg. 3**	**Seg. 4**	**Seg. 5a**	**Seg. 5b**	**Seg. 5c**	**Seg. 6**	**Seg. 7**
Age (A) *	A ≤ 25	26.1	27.5	25.5	35.3	19.8	16.0	24.6	52.4	32.1	34.4
	25 <A ≤35	25.2	22.0	26.5	30.9	29.3	12.0	16.9	4.8	26.2	25.0
	35 <A ≤45	23.4	17.4	25.5	19.1	24.2	48.0	23.1	28.6	19.1	28.1
	A > 45	25.3	33.1	22.5	14.7	26.7	24.0	35.4	14.2	22.6	12.5
Hours (H) **	H ≤ 10	34.1	31.2	44.1	38.2	22.7	48.0	49.2	52.4	35.7	43.7
	10 <H ≤20	31.9	35.8	30.4	35.3	28.9	40.0	29.2	23.8	38.1	28.1
	H > 20	34.0	33.0	25.5	26.5	48.4	12.0	21.6	23.8	26.2	28.2
Years (Y) **	Y ≤ 1	18.2	23.8	22.5	13.2	6.6	20.0	36.9	42.9	23.8	25.0
	1 <Y ≤3	38.2	35.8	37.2	42.6	35.5	32.0	43.1	38.1	46.4	34.4
	3 <Y ≤5	22.4	21.1	24.5	26.5	26.0	28.0	13.8	4.8	16.7	21.9
	Y > 5	21.2	19.3	15.8	17.7	31.9	20.0	6.2	14.2	13.1	18.7
Income (I) *	I ≤ 30K	26.1	30.3	24.5	27.9	20.1	4.0	32.3	52.4	33.3	31.2
	30K<I ≤50K	28.1	23.8	36.3	25.0	23.1	28.0	29.2	23.8	40.5	34.4
	50K<I ≤70K	18.8	22.0	16.7	16.2	23.8	28.0	12.3	9.5	10.7	12.5
	I > 70K	27.0	23.9	22.5	30.9	33.0	40.0	26.2	14.3	15.5	21.9
Gender (G) **	Male	51.6	42.2	47.1	47.1	64.5	60.0	40.0	47.6	42.9	40.6
	Female	48.4	57.8	52.9	52.9	35.5	40.0	60.0	52.4	57.1	59.4
Segment	**Size (abs.)**	778	109	102	68	272	25	65	21	84	32
%		*100*	*14.0*	*13.1*	*8.7*	*35.0*	*3.2*	*8.4*	*2.7*	*10.8*	*4.1*

*Table 7.5: The bases age, average of hours a week of Internet usage, years of Internet experience, household income and gender are represented, in the left-side of the table, as Age, Hours, Years, Income (in US $) and Gender. *$\chi^2$ test, significant at $p<0.05$; ** significant at $p<0.001$. All the figures in the table, corresponding to those 5 bases, are percentages of the segment size (or the sample size in the case of the first column). All the segment sizes and the percentage of the sample size they represent are shown in the last row of the table.*

The consequences of this flow include "increased learning, increased exploratory and participatory behaviors, and more positive subjective experiences" (Hoffman and Novak, 1997). The authors characterize consumers' Web navigation behaviour as either goal-directed or exploratory. The first is a "directed search mode (...) in which the consumer is extrinsically motivated to find a particular site or piece of information in a site", whereas the second "corresponds to a nondirected, exploratory search mode" (Hoffman and Novak, 1997). It is argued that early Web users' flow experiences are generally of the exploratory kind, whereas only experience will lead users to achieve flow experiences in a goal-directed task such as purchasing online.

7.3.3. Fuzzy clustering and segments of one

The posterior probability of node / segment-membership defined in chapter 6 (expression 5) provides a wealth of information for each consumer that is lost in the representations of the complete set of data. This information would be the basis for the use of the GTM as a fuzzy clustering tool. Figure 7.8 illustrates some specific cases.

We can explore in more detail cases 2 and 3 in figure 7.8. Making use of the reference maps in figure 7.5, it is clear that *case 2* would be much more likely to become a purchaser if the *perception of risk* (factor 2 in table 7.1) was alleviated. *Case 3*, in the *Cost Conscious* segment, is close enough to segment 4, the *Convinced*, to consider that information targeted to tackle his/her perception of lack of *affordability* (factor 3 in table 7.1) would increase the probability of this person becoming a purchaser. As remarked in Block *et al.* (1996): "...by understanding the segment of customers which are most interesting, specific marketing efforts can be targeted to similar individuals, currently non customers."

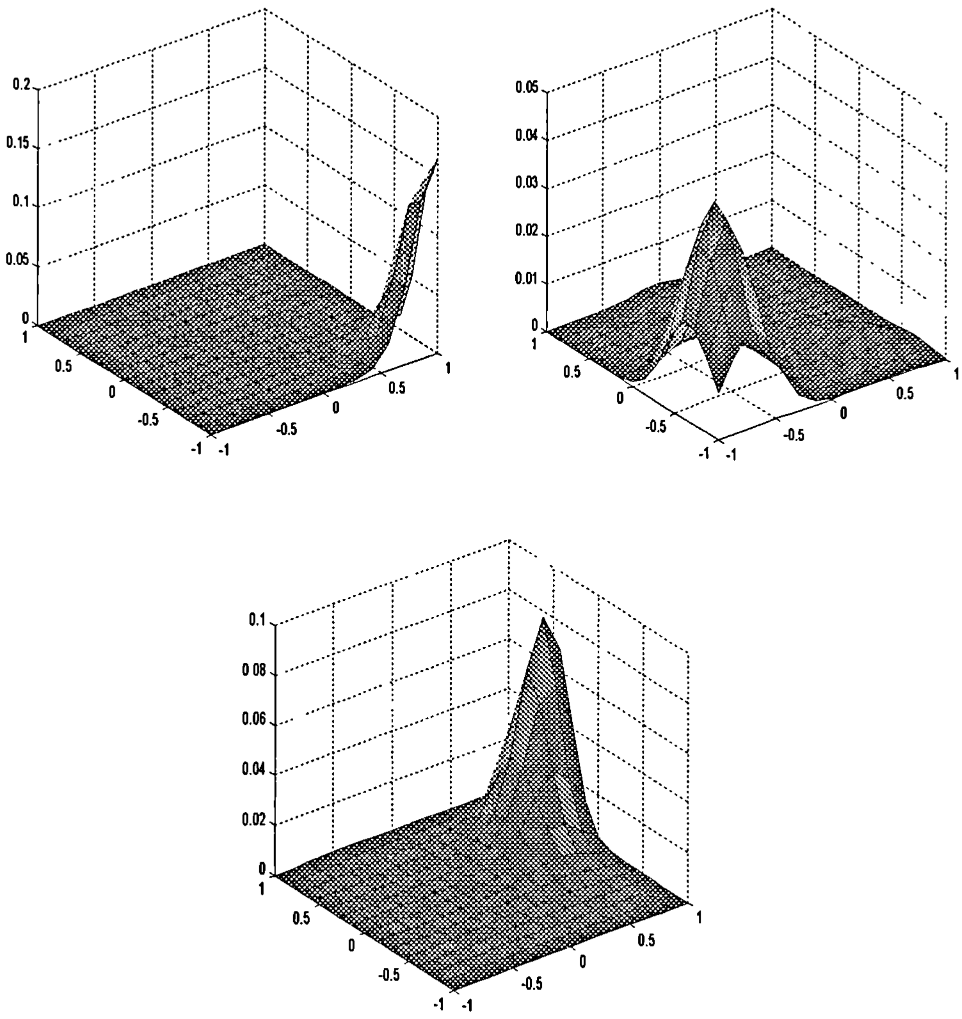


***Figure 7.8:** Posterior probability of segment-membership for three specific consumers. This probability adds up to 1 over the 225 nodes of the 15x15 grid. Case 1 (top-left) clearly corresponds to segment 7 (Customer service wary) of the 9-segment solution described in table 4. Case 2 (top-right) belongs to segment 2 (Security Conscious) but it is very close to segment 4 (Convinced). Case 3 (bottom) belongs to segment 5 (Cost conscious)*

The GTM can also be used to discriminate *segments of one*, otherwise referred to as micro-segmentation, one-to-one-segmentation or finer-segmentation (Kara and Kaynak, 1997), by augmenting the latent-space grid to a sufficient resolution. The resulting segments can be characterized using the *reference maps* as described above.

Nevertheless, it has to be borne in mind that segments do not naturally arise from the data description of the market, and there is always a degree of subjectivity regarding its definition, motivated by the need to obtain a market description of managerial relevance. The characteristics of the segments have to be adapted to accommodate a trade off between the costs of the segmentation and the efficiency of consumer response (Wind, 1978; Wedel and Kamakura, 1998: p.329). The online medium, associated with consumer electronic commerce, has been cited as enabling the necessary interactivity to reach the individual as a *one-person segment*. The heterogeneity of the market, though, can only do without aggregate discretization by using a large amount of information on each individual and this information may not always be available to the marketer. This drawback can be especially poignant in the online medium, where the "primary barrier to the successful commercial development (...) is the current lack of consumer trust" due to the lack of satisfactory response to "mounting consumer concerns over information privacy in electronic, networked environments" (Hoffman *et al.*, 99). Furthermore, it still remains to be seen whether the benefits derived from augmented consumer response can outstrip the costs of content personalization. This conflict between personalization and aggregate segmentation is nicely worded in Kiang *et al.* (2000) within a product-based perspective:

> "...high product customization requires extensive profiling and customization tools to identify and target individual customers ... When product customization is low, one would still need tools that can broadly cluster customers for target marketing."

Either way, and given that "for different strategic goals, different segments may need to be identified in the same population" (Wedel and Kamakura, 1998: p.328), the GTM has proved to be a flexible model, suited to market segmentation tasks.

7.4. Conclusions

The application of a quantitative segmentation methodology to our data, based on the Generative Topographic Mapping (GTM) model, has revealed several useful features of this approach. The GTM shows considerable discrimination of the two main groups of on-line purchasers and non-purchasers without the need of *a priori* information about class membership of the individuals in the sample.

An entropy measure has been proposed to quantify the information content of equal-size unsupervised maps about an externally imposed class label. According to this measure, it has been shown that the parsimonious description of the original data, represented by a 5-factor selection model, results in a trained GTM that optimizes the separation of *purchasers* and *non-purchasers* in a hold-out sample. This comes to justify the *tandem* approach to market segmentation, utilized in this study, on the basis of a profit-based measure.

The use of the GTM for aggregate segmentation has been described. The *topographic ordering* of the latent space representation, together with the *magnification factor* and the *cumulative responsibility*, make the GTM a well-principled aggregate segmentation model. Some patterns of managerial interest can be found in the macro-cluster segmentation results. From the *magnification factor* map in figure 7.2 and the *cumulative responsibility* in figure 7.3, the group of *purchasers* seems to be compact and homogeneous, corresponding to the *Convinced* in figure 7.7, whereas the *non-purchasers* seem to be much more heterogeneous in nature. Segments shaped mainly by those *non-purchasers* might be targeted with modified marketing strategies that address the specific concerns affecting the individuals within them. For instance, and following the definitions in tables 7.3 and 7.4, the *Complexity Avoiders* may become engaged in the on-line purchasing process through the simplification of its procedures and the utilization, at all steps of the transaction, of user-friendly interfaces. The *Undecided* might become engaged if the marketer made every effort to develop exchange relationships based upon trust, which, as remarked by Hoffman *et al.* (1999), are more likely to succeed. Strategies of this kind could also be designed to target the rest of those segments with strong *non-purchasers* presence.

7.5. Publications related to this chapter

1. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): Segmenting the e-commerce market using the Generative Topographic Mapping. In *Proceedings of the MICAI-2000, Acapulco, Mexico*, 470-481.
2. Vellido, A., Lisboa, P.J.G. & Meehan, K. (2000): The Generative Topographic Mapping as a principled model for data visualization and market segmentation: an electronic commerce case study. *International Journal of Computers, Systems and Signals,* forthcoming.

Chapter 8

Selective smoothing of the Generative Topographic Mapping

8.1. Introduction

In the previous chapter, we have applied the Generative Topographic Mapping (Bishop *et al.*, 1998a) to the segmentation of the business to consumer e-commerce market. The probabilistic formulation of the GTM makes it possible to develop this model in a principled manner. In particular, the training of the GTM can be integrated into the Bayesian framework for the estimation of its controlling parameters (Bishop *et al.*, 1998b) In a similar way, this was done in chapter 4 with the Multi-Layer Perceptron.

In this chapter we propose the use of the principle of *Automatic Relevance Determination* (ARD: MacKay, 1995) for the inclusion of the number of basis functions, controlling the complexity of the mapping, within this Bayesian framework. This process will result in the selective smoothing of the mapping from the latent to the multi-dimensional data space, through a soft pruning of irrelevant basis functions. Therefore, the complexity of the mapping is optimized and only the width of the basis functions remains as a parameter to be selected "by hand" outside the Bayesian framework.

We also propose a new information measure to represent the class-based information contained by the clustering. This measure provides a new criterion to select the number of clusters/segments in a partition.

The chapter is organized in the following manner: The basics of the Bayesian approach to the training of the GTM with a single regularization term are summarized in the first part of the next section. The second part describes the *Selective Mapping Smoothing* procedure. Section 8.3 introduces the information measure that will be used to compare the performance of the Selective GTM Smoothing model with those of the Bayesian GTM with single regularization term and the unregularized GTM model.

8.2. Bayesian inference of the GTM hyperparameters and selective mapping smoothing

The complexity of the mapping generated by the GTM model is mainly controlled by the number and form of the basis functions. Further control of this effective complexity can be achieved with the addition of a regularization term to the error function. This entails a modification of expression (4), in section 6.3.1 of chapter 6, so that the training of the GTM consists now of the maximization of the *penalized* log-likelihood

$$L_{PEN}(\mathbf{W},\beta) = \sum_{n=1}^{N} \ln p\left(\mathbf{x}^n \mid \mathbf{W},\beta\right) + \frac{1}{2}\alpha\|\mathbf{w}\|^2, \tag{1}$$

where **w** is a vector shaped by concatenation of the different column vectors of the weight matrix **W**, and α is the regularization coefficient. This regularization term is effectively preventing the GTM to fit the noise in the data.

The optimum values for all these parameters should ideally be evaluated in a continuous space of solutions. Given that the GTM is formulated within a probabilistic framework, this can be accomplished using the Bayesian formalism and, more specifically, the *evidence approximation* (MacKay, 1992). The application of this methodology to the regularization coefficient (Bishop *et al*, 1998b) is summarized next. Given that α and β control other parameter distributions, they will be referred to as *hyperparameters*. We intend to find those values of the hyperparameters that maximize the posterior distribution

$$p\left(\alpha,\beta|\left\{\mathbf{x}^{n}\right\}\right)=\frac{p\left(\left\{\mathbf{x}^{n}\right\}|\alpha,\beta\right)p(\alpha,\beta)}{p\left(\left\{\mathbf{x}^{n}\right\}\right)} \tag{2}$$

Assuming uninformative priors for the hyperparameters, that is equivalent to the maximization of their *evidence* or marginal likelihood

$$p\left(\left\{\mathbf{x}^{n}\right\}|\alpha,\beta\right)=\int p\left(\left\{\mathbf{x}^{n}\right\}|\mathbf{w},\beta\right)p(\mathbf{w}|\alpha)d\mathbf{w} \tag{3}$$

The prior distribution over the weights is defined as

$$p(\mathbf{w}|\alpha)=\left(\frac{\alpha}{2\pi}\right)^{\frac{W}{2}}\exp\left(-\frac{1}{2}\alpha\|\mathbf{w}\|^{2}\right), \tag{4}$$

where *W* is the number of weights in **W** (*D*x*R*, following the definitions in section 6.3.1, chapter 6). Let us now define

$$S(\mathbf{w},\alpha,\beta) = -\ln\left[p\left(\left\{\mathbf{x}^n\right\}\mid\mathbf{w},\beta\right)p(\mathbf{w}\mid\alpha)\right]. \tag{5}$$

Using a second order Taylor expansion of (5), we find that the *evidence* can now be expressed as

$$p\left(\left\{\mathbf{x}^n\right\}\mid\alpha,\beta\right) = \exp\left[-S(\mathbf{w}_*,\alpha,\beta)\right](2\pi)^{\frac{W}{2}}\left|\mathbf{A}_*\right|^{-\frac{1}{2}} , \tag{6}$$

where $\mathbf{w}_*$ is the weight vector that generates a maximum of the posterior distribution of the data, and $\mathbf{A}_*$ is the Hessian of S evaluated at $\mathbf{w}_*$. The log-evidence for α and β is thus given by

$$\ln p\left(\left\{\mathbf{x}^n\right\}\mid\alpha,\beta\right) = L(\mathbf{w}_*,\beta) - \frac{1}{2}\alpha\left\|\mathbf{w}_*\right\|^2 - \frac{1}{2}\ln\left|\mathbf{A}_*\right| + \frac{W}{2}\ln\alpha , \tag{7}$$

where the constant terms have been dropped. The maximization of this expression for α and β leads to the following updating formulae: For the regularization coefficient α,

$$\alpha_{NEW} = \frac{\gamma_{OLD}}{\left\|\mathbf{w}_*\right\|^2}, \tag{8}$$

where

$$\gamma_{OLD} = \sum_{i\ 1}^{W}\frac{\lambda_i - \alpha_{OLD}}{\lambda_i}, \tag{9}$$

and λ_i are the eigenvalues of $\mathbf{A}_*$.

For the inverse variance of the noise model β,

$$\beta_{NEW} = \frac{ND - \gamma_{OLD}}{\sum_{n=1}^{N}\sum_{i=1}^{M}R_i^n\left\|\mathbf{x}^n - \mathbf{m}_i\right\|^2}, \tag{10}$$

where *N, D, M,* the responsibility R_i^n and the reference vector $\mathbf{m}_i$ have all been defined in the previous chapter. In practice, the updating of the hyperparameters is intertwined with the calculations of the E-M algorithm. The advantage of the evaluation of the complexity parameters within this framework is that, instead of having to resort to the calculation of the determinant of Hessians, only the derivative of the *evidence* with respect to the hyperparameters is required. The latter makes use of the trace of the inverse of the Hessian, which is numerically better conditioned (MacKay, 1995).

8.2.1. Selective smoothing of the Generative Topographic Mapping

The calculation of the optimum number of effective basis functions in a continuous space of solutions can also be accomplished within the Bayesian framework. We propose the principle of *Automatic Relevance Determination* (ARD) (MacKay, 1994) as the framework to integrate the number of basis functions in a continuous model space. Instead of using a single regularization coefficient α for the whole mapping, a different regularization coefficient α_r will be defined for each basis function. In this way, the prior distribution for the weights is now defined as

$$p(\mathbf{w}|\alpha) = p(\mathbf{w}|\{\alpha_r\}) = \prod_{r-1}^{R}\left(\frac{\alpha_r}{2\pi}\right)^{\frac{D}{2}} \exp\left(-\sum_{r=1}^{R}\frac{1}{2}\alpha_r\|\mathbf{w}_r\|^2\right), \tag{11}$$

where $\mathbf{w}_r$ is the vector of weights in **W** associated to the hyperparameter *r*. This leads to a re-formulation of the *log-evidence* or marginal log-likelihood in the form

$$\ln p\left(\{\mathbf{x}^n\}|\{\alpha_r\},\beta\right) = L(\mathbf{w}_*,\beta) - \sum_{r=1}^{R}\frac{1}{2}\alpha_r\|\mathbf{w}_{*r}\|^2 - \frac{1}{2}\ln\left|\mathbf{A}_{*\{\alpha_r\}}\right| + \sum_{r=1}^{R}\frac{D}{2}\ln\alpha_r\,, \tag{12}$$

Its maximization with respect to β produces the same results as the single regularization term case, so that β will follow the updating expression (10). The maximization with respect to each α_r results in the set of r equations

$$-\frac{1}{2}\left\|\mathbf{w}_{*r}\right\|^2 - \frac{1}{2}\frac{d}{d\alpha_r}\left(\sum_i^W \ln\left(\lambda_i + \alpha_r\right)\right) + \frac{D}{2\alpha_r} = 0 \quad , \tag{13}$$

They have solutions for

$$\alpha_r\left\|\mathbf{w}_{*r}\right\|^2 = D - \alpha_r Trace_r\left(\mathbf{A}^{-1}\right) = \gamma_r\,, \tag{14}$$

where the trace is over those $D = WR$ weights associated to α_r and γ_r can be interpreted as the number of *effective* weights associated to each regularization coefficient (i.e. to each basis function), laying between 0 and D. This leads to the following updating formula for each of the regularization parameters

$$\alpha_{r,NEW} = \frac{\gamma_{r,OLD}}{\left\|\mathbf{w}_{*r}\right\|^2} \tag{15}$$

High values of α_r will push the weights associated to their corresponding basis function towards zero, effectively minimizing their effect in the mapping. The smoothness of the GTM mapping is largely determined by the properties of the basis functions: the more or the narrower, the more flexible the mapping will be, which could lead to over-fitting. The less or the wider, the smoother the mapping. Therefore, the selective smoothing of the basis functions accomplished by ARD will provide mappings of optimum complexity, for a given width of the basis functions, without limiting the number of basis functions utilized. The only parameter still depending on a heuristic procedure for its determination is now the width of the basis functions but, having already adaptively optimized the rest of the hyperparameters, the best GTM model can be obtained by experimenting with different values of these widths. The use of multiple regularization terms brings about some changes in the M-step of the EM algorithm for updating **W**, which has now to be calculated as

$$\left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{G}\mathbf{\Phi}+\frac{1}{\beta}\Lambda\right)\mathbf{W}_{new}^{\mathrm{T}}=\mathbf{\Phi}^{\mathrm{T}}\mathbf{R}\mathbf{X} \quad , \qquad (16)$$

where Λ is a square matrix with $\{\alpha_r\}$ in the diagonal and zeros elsewhere.

8.3. The information measure: experiments with Selective Mapping Smoothing

We now return to the data set made up of the selection of five factors defining Internet users' opinions of online shopping that was utilized in the previous chapters (table 7.1, chapter 7). The performance of the *selective smoothing* method will be compared with those of the non-regularized GTM, and the Bayesian GTM with a single regularization term. This performance will be assessed according to an extension of the information measure defined in section 7.3.1 of chapter 7. That entropy was only valid for the comparison of partitions of the data set with equal number of clusters. The new measure extends its validity to the comparison of partitions with any number of clusters.

Let us define the entropy, or average information provided by the class discrimination, given a cluster partition, as

$$S1=-\sum_{clusters}P(clusters)\sum_{classes}P(class|cluster)\ln P(class|cluster) \qquad (17)$$

For two classes *-purchasers (C1)* and *non-purchasers (C2)-* and M clusters -either GTM nodes or aggregate macro-clusters- (17) becomes

$$S1 = -\sum_{i=1}^{M} \frac{N_i}{N}\left(p_i^{C1} \ln p_i^{C1} + p_i^{C2} \ln p_i^{C2}\right) \tag{18}$$

which is the entropy defined in section 7.3.1, chapter 7. p_i^{C1} and p_i^{C2} are the ratios of *purchasers* n_i^{C1}/N_i and *non-purchasers* n_i^{C2}/N_i, mapped into cluster i and N_i is the total number of observations mapped into cluster i, so that $N = \sum_{i=1}^{M} N_i$ is the number of individuals in the complete data set.

Let us define next the entropy due to the own distribution of observations across the cluster partition as

$$S2 = -\sum_{clusters} P(clusters) \ln P(clusters) = -\sum_{i=1}^{M} \frac{N_i}{N} \ln \frac{N_i}{N} \tag{19}$$

and the entropy associated to the prior class distributions as

$$SI = -\sum_{classes} P(class) \ln\left(P(class)\right). \tag{20}$$

We propose a new information measure to represent the class-based information contained by the clustering. Given the joint entropy

$$J = -\sum_{classes} \sum_{clusters} P(class, cluster) \ln P(class, cluster), \tag{21}$$

the optimal partition is defined to be that which minimizes the increase in information due to adding the class labels, given a fixed cluster partition, together with the increase in information due to the cluster partition itself, given the class prevalences prior to clustering

$$S = \frac{J - S2}{S2} + \frac{J - SI}{SI} = \frac{S1}{S2} + \frac{(S1 + S2) - SI}{SI}. \tag{22}$$

The behavior of this entropy metric is illustrated with several extreme cases in table 8.1. It is clear that a one-to-one correspondence between clusters and classes reduces S to its absolute minimum of zero, given that the numerators are information divergences, hence positive semi-definite. Therefore, a minimum of S will correspond to a parsimonious cluster partition.

*1. **All the data contained in a single cluster*** $S1 = SI$ $S2 = 0$ $\frac{S1 + S2 - SI}{SI} = 0$ $S = \infty$	*2. **Each cluster is a separate class*** $S1 = 0$ $S2 = SI$ $\frac{S1 + S2 - SI}{SI} = 0$ $S = 0$
*3. **Each data observation is in a separate cluster*** $S1 = 0$ $S2 = \ln(N)$ $\frac{S1 + S2 - SI}{SI} = \frac{\ln(N) - SI}{SI}$ $S = \frac{\ln(N) - SI}{SI}$	*4. **Clusters and classes are completely independent*** $S1 = SI$ $S2 = -\sum_{i=1}^{M} \frac{N_i}{N} \ln\left(\frac{N_i}{N}\right)$ $S = \frac{SI}{S2} + \frac{S2}{SI}$

***Table 8.1**: Extreme cases of the information measure (22)*

Figure 8.1 shows the values of the entropy for the three models mentioned above across the range of cluster solutions. Table 8.2 summarizes the best results. The two models trained within the Bayesian approach show a very similar performance across the range of cluster solutions, and very different to the non-regularized model. The latter is heavily penalized for parsimonious cluster partitions. The reason is that this model has a very uneven distribution of the observations across clusters, with one of them considerably outsizing the rest. Unlike the non-regularized GTM, both regularized models have limited the effect of over-fitting, as reflected by the fact that their training entropies are higher than the test ones. The ARD model generates the overall minimum entropy.

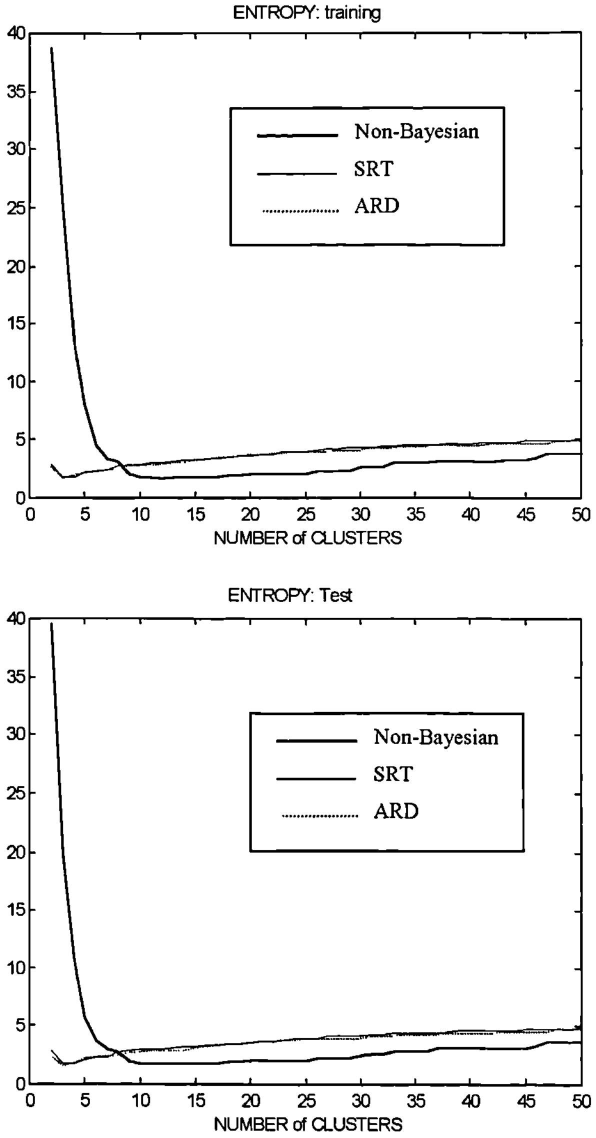


***Figure 8.1:** Entropy for the different models (training above, test below) described in the text, as a function of the number of clusters in the partition. This follows the agglomerative procedure described in section 7.3.2.1 of chapter 7.*

GTM model		**Minimun entropy**	**No. of clusters**
Non-regularized	*training*	1.8118	12
	test	1.8357	17
Bayesian-single regularization term	*training*	1.8717	3
	test	1.8664	4
Bayesian Selective Smoothing	*training*	**1.7468**	3
	test	**1.6977**	3

Table 8.2: *Best entropy for each of the GTM models and its corresponding number of clusters.*

The 3-segment solution, shown to be the best in table 8.2, was described in chapter 7, section 7.3.2.1, and displayed in figure 7.4. Its comparison with figure 8.1 reveals that the entropy metric is picking up almost a natural class-discrimination by the GTM.

8.4. Conclusions

The number of basis functions utilized in the GTM can be optimized using the Automatic Relevance Determination principle, within the Bayesian approach to the network training. Such procedure, named *Selective Mapping Smoothing (SMS)*, capable of producing maps of optimal complexity, has been described and implemented in the present chapter using the data concerning Internet users' opinions of online shopping as a test bed for assessment.

The ARD principle has shown to be a useful feature selection method for MLPs applied to classification (Penny and Roberts, 1999) and a principled environment for the support of its empirical assessment has been developed (Neal, 1998). It has also recently been applied with success to a Bayesian treatment of the Support Vector Machine (SVM) model, called Relevance Vector Machine (RVM: Tipping, 2000). In this model, the number of *kernel functions* –used for the mapping from inputs to outputs in regression and classification problems- is optimized using ARD.

The performance of the *SMS* model has been compared to those of the Bayesian GTM with a single regularization term and the non-regularized GTM. To measure that performance, a new class-discrimination entropy-based information measure has been defined. This measure is the sum of the appropriately normalized additional information from adding the external label to the cluster structure, and from imposing the clusters onto the prior class labels. It achieves a global minimum when the two structures match identically. The Bayesian approach- based regularized models have been shown to limit the occurrence of over-fitting, whereas the *SMS* model has generated an overall entropy minimum.

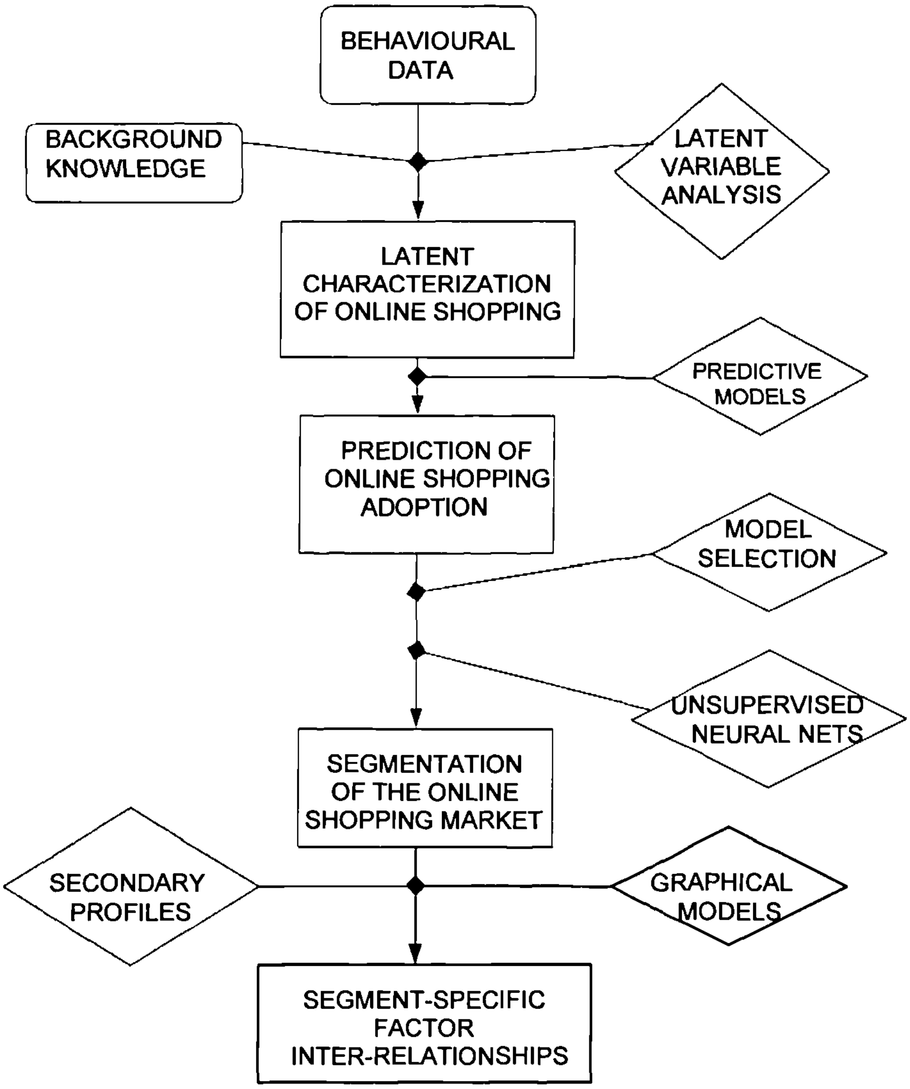
There is no contradiction between the best segment solution found in section 8.3 of this chapter and the segment solutions proposed in chapter 7. The latter did not make explicit use of the class-membership information, so that the segments can be considered as *natural* data clusters. On the other hand, the use of the entropy metric in this chapter provides a segment solution optimized in terms of class-discrimination. Depending on the goals of its marketing use, the analyst can opt for either segmentation solution or try to reach a compromise between them.

PART 5

Graphical modelling of online market segments

...but eventually I came to see this chance as a form of readiness, a way of saving myself through the minds of others.

"Moon Palace". Paul Auster


BEHAVIOURAL DATA
BACKGROUND KNOWLEDGE
LATENT VARIABLE ANALYSIS
LATENT CHARACTERIZATION OF ONLINE SHOPPING
PREDICTIVE MODELS
PREDICTION OF ONLINE SHOPPING ADOPTION
MODEL SELECTION
UNSUPERVISED NEURAL NETS
SEGMENTATION OF THE ONLINE SHOPPING MARKET
SECONDARY PROFILES
GRAPHICAL MODELS
SEGMENT-SPECIFIC FACTOR INTER-RELATIONSHIPS

Preface

We reach now a new stage of the quantitative methodology outlined in the introduction. The latent variable description of the data, containing factors shown to be good predictors of the propensity to buy online, was the bases, in the previous part, to segmenting the e-commerce market using an unsupervised neural network model. The resulting segments were characterized and profiled, but this characterization did not provide information on how the segmentation bases are related to each other within the segment and the inter-relationships between the propensity to buy online and those segmentation bases. Graphical models, the subject of this chapter, will be the method to address these problems. Their application should reveal aspects of the structure of specific segments that are not obvious in the visual inspection of individual factors. The results are expected to improve the actionability of the segments in the partition and, therefore, contribute to the segmentation procedure itself.

Chapter 9

Segment-specific graphical modeling of the e-commerce market

9.1. Introduction

In the previous chapters, the business-to-consumer electronic market was segmented using the unsupervised neural network-based GTM model. The resulting segments were characterized, in an exploratory way, using visual aids such as the reference maps and profiling with exogenous variables (demographics, Web usage...). These characterization methods do not provide information on how the bases are related to each other within the segment and this is, precisely, what the graphical models, subject of this chapter, can offer. Graphical models should reveal aspects of the structure and characteristics of specific segments that are not obvious in the visual inspection of individual factors. The resulting new insights are expected to improve the actionability of the segments in the partition and, therefore, contribute to the methodology of segmentation itself.

Furthermore, we want to explore the within-segment inter-relationships between the propensity to buy online and the factors used as segmentation bases. In particular, the graphical models can help to discover which factors are conditionally independent of the propensity to buy online. This information is especially useful from a marketing point of view because it would permit the marketer to focus resources and efforts on factors which have a direct influence on the propensity to buy online.

Some basic principles of graphical modeling theory are gathered in section 9.2. Then, the results of the application of graphical Gaussian models to the GTM 7-segment solution from chapter 7 are described and commented on in some detail in section 9.3. A discrete variable, describing whether individuals in the sample have ever purchased online or not, is introduced in the analysis, and graphical mixed model techniques are applied to obtain independence graphs for the same 7-segment solution. The corresponding results are summarized in section 9.4. The chapter is closed by section 9.5, which offers some overall conclusions as well as comments on the connection of this chapter with the overall methodology considered in the thesis.

9.2. Basics of graphical modeling theory

Graphical modeling is a technique of multivariate analysis that fits graphs to data models. In this framework, multivariate data are assumed to arise from a joint probability density distribution $f_\theta(x_1,\dots,x_m)$, where θ is an unknown parameter and X_1, ..., X_m are m random variables, corresponding to x_1, ..., x_m. A model is defined as a family of possible probability densities f_θ.

Graphs are a powerful way to visualize the inter-relationships between the variables of a data set. A graph $G=(V,E)$ is defined by a set of vertices $V-\{X_1, \dots, X_m\}$, representing the variables of the data, and a set of edges E, connecting those vertices.

Two vertices X_i and X_j are said to be *adjacent,* and represented as $[X_i\ X_j]$, if they are connected by an edge. Models are related to graphs in the following form: a graph only retains those edges that connect two variables which are *not* conditionally independent given the rest of variables of the data set. This can formally be defined (Whittaker, 1990) saying that the conditional independence graph of X is the undirected graph $G=(V,E)$ where $[X_i\ X_j] \notin E$ if and only if $X_i \perp X_j \mid V \setminus \{X_i, X_j\}$. This last expression can be read as "X_i is conditionally independent of X_j, given the rest of the variables of the data set". Two variables X_i and X_j are conditionally independent given a third one X_k, i.e. $X_i \perp X_j \mid X_k$, if the conditional distribution $f_{X_i|X_j,X_k}(x_i|x_j,x_k)$ is not a function of x_k. Alternatively (Dawid, 1979), $X_i \perp X_j \mid X_k$ if the joint density $f_{X_i,X_j,X_k}(x_i,x_j,x_k)=f_1(x_j,x_k)f_2(x_i,x_k)$ can be broken into factors that are, in turn, independent of x_i and x_j.

In this chapter, two different graphical models will be utilized. Those that only include continuous variables, called graphical Gaussian models, and those which include both continuous and discrete variables, called graphical mixed models. They are now described separately in more detail.

9.2.1. Graphical Gaussian models

These are models based on the multivariate normal distribution for continuous variables. Let us assume that the m-dimensional random variable X is described by a multivariate normal distribution with mean μ and a covariance matrix Σ. We can write the distribution X as

$$f_X(x) = 2\pi^{-m/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(x-\mu)'\Omega(x-\mu)\right\} \tag{1}$$

where $\Omega=\Sigma^{-1}$. This expression can be rewritten as a combination of constant , linear and quadratic terms on x

$$f_X(x) = \exp\left(\alpha + \beta' x - \frac{1}{2}x'\Omega x\right) = \exp\left(\alpha + \sum_{i=1}^{m}\beta^j x_j - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\omega^{ij}x_i x_j\right), \tag{2}$$

where ω^{ij} are the elements of the inverse covariance matrix Ω. Now, given the definitions of conditional independence in the previous section, it can be seen that $X_i \perp X_j \mid V \setminus \{X_i, X_j\}$ if and only if $\omega^{ij} = 0$. According to the concepts explained in the introduction to this section, the connection with the graphs is straightforward: setting $\omega^{ij} = 0$ corresponds to deleting the edge $[X_i\ X_j]$ from the graph.

If a sample of N m-dimensional data observations has a sample mean $\bar{x}$ and a sample covariance S, the log-likelihood of the sample will be expressed (Edwards, 1995) as

$$L(\mu,\Omega) = -Nm\ln(2\pi)/2 - N\ln|\Sigma|/2 - N\mathrm{tr}(\Omega S)/2 - N(\bar{x}-\mu)'\Omega(\bar{x}-\mu)/2\ , \tag{3}$$

where $\mathrm{tr}(\Omega S)$ is the trace of the matrix ΩS. The maximum likelihood estimates can be calculated using iterative algorithms, and the maximized log-likelihood difference or *deviance difference* between two models M_0 and M_1 can be calculated as

$$d = 2\left(L_{\max,0} - L_{\max,1}\right) = N\ln\left(|\Sigma_{\max,0}|/|\Sigma_{\max,1}|\right)$$

This deviance difference has an asymptotic χ^2 distribution under M_0, and can be used as a test for the deletion of edges, i.e. for model selection. This will be explored in section 9.2.3.

9.2.2. Graphical mixed models

Models with both continuous and discrete variables are now described. Let us assume that there are r discrete and m continuous variables, represented as (i,x) with corresponding random variables (I,X), where i is a r-tuple of discrete values and x is a m-tuple of continuous values. The probability of the discrete variables taking the value i is p_i , and the conditional Gaussian (CG) distribution for X, given $I = i$ is defined

$$f(i,x) = p_i|2\pi\Sigma_i|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)\right\} \tag{4}$$

where Σ is the covariance matrix and μ the mean vector as defined in the previous section. This can be rewritten as

$$f(i,x) = \exp\left\{\alpha_i + \beta_i'x - \frac{1}{2}x'\Omega_i x\right\},\text{ where }\Omega = \Sigma^{-1}. \tag{5}$$

α, β and Ω, called *canonical parameters*, can be expanded as sums of interaction terms, and the generation of a model entails setting some of the higher-order interaction terms to zero.

The expansion of the canonical parameter α is described by discrete *generators* $d_1,\ldots,$ d_r. A generator is a maximal interaction term, or *clique*, in the model, also defined as a subset of vertices that is maximally complete, i.e. all the vertices are adjacent to each other and if a vertex outside the subset is added this property will be lost. The expansion of β is described by linear generators $l_1,\ldots, l_s$ that contain only one continuous variable and can also contain discrete variables. The expansion of Ω is described by quadratic generators $q_1,\ldots, q_t$, that *must* contain at least one continuous variable. Therefore, the model formula has the form $d_1, \ldots, d_r / l_1, \ldots, l_s \; q_1, \ldots, q_t$.

The connection between a model and its graph can be obtained following a procedure similar to the one used for the continuous models, although some specific syntax rules have to be observed (Edwards, 1995). Similarly, an expression for the log-likelihood can be obtained and maximized, and a *deviance difference*, to be used for model selection, can be calculated.

9.2.3. Model selection

For each of the segments in which the data set has been partitioned using the GTM and the contiguity-constrained agglomerative algorithm, a graphical model, representing the conditional relationships between variables, has to be found. A method to find simple models consistent with the data is stepwise selection. In this procedure, the deletion of *edges* between variables is tested iteratively, starting from an initial model and until a termination criterion is met.

Significance tests are performed for the deletion of edges. More specifically, χ^2-tests based on the *deviance difference* between the model previous to the edge deletion and the model resulting from that deletion (the *deviance difference* is referred to a χ^2

distribution), as defined in the previous section. All possible edge-removals are tested and p-values are obtained. The edge corresponding to the largest non-significant p-value is then removed from the model. This procedure is repeated until no non-significant p-value is obtained. In the case of our experiments, that significance level was set to 0.05.

Not all edges are considered for deletion at each step of the stepwise backward selection. In fact, if the test for the removal of an edge results in a significant p-value and, therefore, the edge is not rejected, it will not be tested in the following iterations of the procedure. This is called the principle of *coherence* (Gabriel, 1969). Furthermore, only *decomposable* models (Lauritzen, 1992) are considered at each step, which means that edge-deletions leading to non-decomposable models are not considered. Decomposable models are those with closed-form maximum likelihood estimates. An interesting property of decomposable models is that they are easier to interprete, as it can be shown that they can be reduced to a succession of univariate conditional models that, in turn, can be interpreted as some sort of causal ordering.

The use of decomposable models also allows the use of F-tests for the deletion of edges, instead of χ^2-tests. This is important when we are dealing with small data sets for which the χ^2-tests can be unreliable (Porteous, 1989).

9.3. Segment-specific graphical Gaussian models using the original segmentation bases

We would expect graph models to give us further insights into the structure and characteristics of the segments obtained with unsupervised mapping and clustering models. Eventually, these insights should provide the online vendor with some marketing advantages. As a new source of marketing intelligence, the graphical models should improve the actionability of the segments in the partition. In chapter 7, the segments were described in terms of the segmentation bases and the corresponding profiles. This does not offer any information on how the bases are related to each other within the segment and that is, precisely, what the independence graphs can offer.

The results of the application of the stepwise backward model selection method to the 7-segment partition of the data set reported in figure 7.7 of chapter 7, is summarized next. Only the 5 continuous factors utilized as bases in the GTM segmentation are used here to build the graphical Gaussian models. The names of the factors have been replaced in the calculations according to the convention shown in *table 9.1*. In all cases, the selection procedure is started with a fully connected model, from which individual edges are removed iteratively. For each segment, the deviance and the *p*-values corresponding to either the χ^2-tests or F-tests are provided. The independence graph is also shown. A graphical model for the whole data set is also generated as a benchmark for comparison with the segment-specific models.

Compatibility	V
Perception Of Risk	W
Affordability	X
Ease Of Use	Y
Effort / Responsiveness	Z

Table 9.1: *Notational convention for the model selection procedure*

Segment 1 (Unconvinced)

The selection proceeded as summarized by the statistics in *table 9.2*. The edges are listed in the order they were removed.

Edge removed	G^2	*p-value*
[WZ]	1.1195	0.2900
[VZ]	1.0262	0.3111
[VY]	1.5827	0.2084
[VW]	2.1833	0.1395
[VX]	1.6815	0.1947

Table 9.2: *Summary statistics of the model selection for segment 1. G^2 is the deviance difference between the models with and without the tested edge.*

The formula of the final model selected was //V, WXY, XYZ. It is represented by the independence graph in figure 9.1.

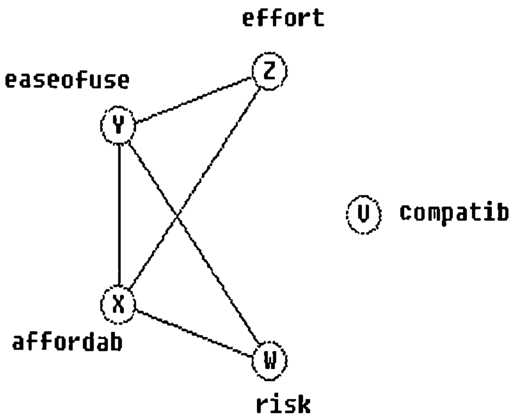


***Figure 9.1:** Independence graph for segment 1 (Unconvinced)*

Comments

The main feature of the resulting graph is the conditional independence of the *Compatibility* factor with respect to all the rest. This is particularly interesting, because the exploration of the reference maps of the GTM (Chapter 7, figure 7.7) reveals that the main descriptor of this segment is precisely *Compatibility*. Given its low value and the results of the ARD variable selection in chapter 3, this factor is clearly behind the fact that the segment is mainly made up of *non-purchasers* (Chapter 7, table 7.4). The fact that *Compatibility* does not seem to be influenced by any other factor reinforces the view of this segment as the hardcore of customers reluctant to shop online.

Segment 2 (Security Conscious)

The selection proceeded as summarized by the statistics in *table 9.3*. The formula of the final model selected was //VWY, VWZ, VXZ. It is represented by the independence graph in figure 9.2.

Edge removed	G^2	*p-value*
[YZ]	0.0077	0.9301
[XY]	1.7250	0.1891
[WX]	0.0288	0.8653

Table 9.3: *Summary statistics of the model selection for segment 2.*

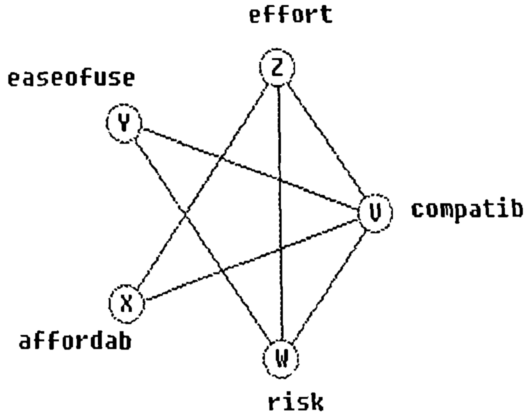


Figure 9.2: *Independence graph for segment 2 (Security Conscious)*

Comments

The *Security Conscious*, who are mainly *non-purchasers*, generate a graph which is distinctively different to that of the *Unconvinced*. In fact, *Convenience* is now linked to all the rest of the factors. There is another feature, though, that links this graph with those of the *Undecided* and the *Complexity Avoiders*: a very strong dependence (resulting in very small *p*-values, not reported in the tables) between *Compatibility, Affordability* and *Effort / Responsiveness*. The highly entangled web of dependencies between factors is yet another aspect of the graph that could be exploited.

Segment 3 (Undecided)

The selection proceeded as summarized by the statistics in *table 9.4*. The formula of the final model selected was //VXZ, WY. It is represented by the independence graph in figure 9.3.

Edge removed	*G^2*	*p-value*
[YZ]	0.0081	0.9285
[VY]	0.1040	0.7568
[VW]	0.1362	0.7121
[XY]	0.0178	0.8939
[WX]	0.5497	0.9902
[WZ]	2.0190	0.1553

***Table 9.4**: Summary statistics of the model selection for segment 3.*

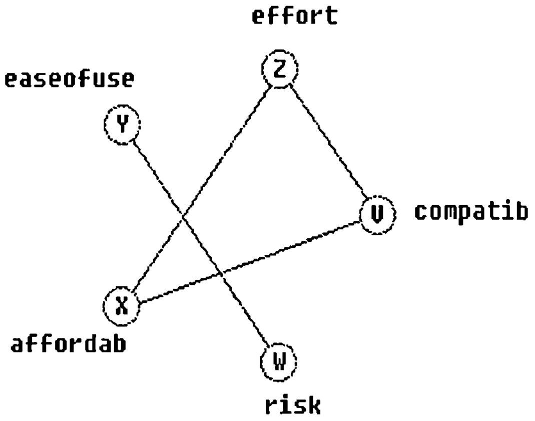


***Figure 9.3**: Independence graph for segment 3 (Undecided)*

Comments

The interpretation of the graph corresponding to this segment can be enhanced by comparing it to that corresponding to the *Complexity Avoiders*. Both are strongly characterized by the relation between *Compatibility, Affordability* and *Effort /*

Responsiveness, together with the lack of conditional independence between *Perception of Risk* and *Ease of Use*. Nevertheless, for the *Undecided*, there is conditional independence between *Affordability* and *Ease of Use*.

Segment 4 (Convinced)

The selection proceeded as summarized by the statistics in *table 9.5*. The formula of the final model selected was //VWYZ, WXY. It is represented by the independence graph in figure 9.4.

Edge removed	G^2	*p-value*
[VX]	0.4060	0.5240
[XZ]	0.8969	0.3436

Table 9.5: *Summary statistics of the model selection for segment 4.*

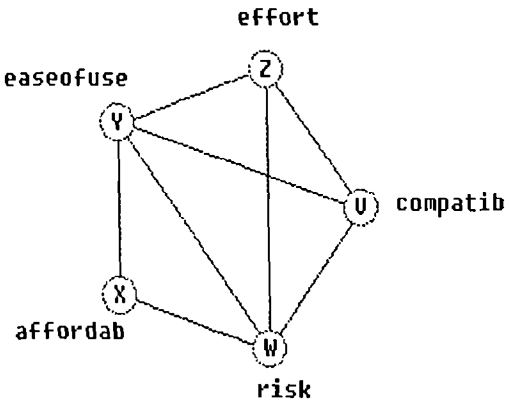


Figure 9.4: *Independence graph for segment 4 (Convinced)*

Comments

The analysis of the *Convinced* segment is, indeed, of special interest, as it is mainly made up by *purchasers*. The corresponding graph model could shed some more light into the mechanisms behind the decision to buy online. There are two main unique features on the graph shown in figure 9.4. First, this is the more densely connected graph of all the segments. It suggests that the experience of buying somehow makes all factors more interdependent, so that their individual effects are balanced and compensated. Secondly, the *Perception of Risk*, one of the best predictors of the propensity to buy online, is now adjacent to all the rest of the factors, which reinforces the previous statement.

Segment 5 (Complexity Avoiders)

The selection proceeded as summarized by the statistics in *table 9.6*. The formula of the final model selected was //VXZ, WY, XY. It is represented by the independence graph in figure 9.5.

Edge removed	G^2	*p-value*
[VY]	0.0446	0.8328
[VW]	0.1459	0.7025
[WZ]	0.2717	0.6022
[WX]	0.6925	0.4053
[YZ]	0.9477	0.3303

***Table 9.6**: Summary statistics of the model selection for segment 5.*

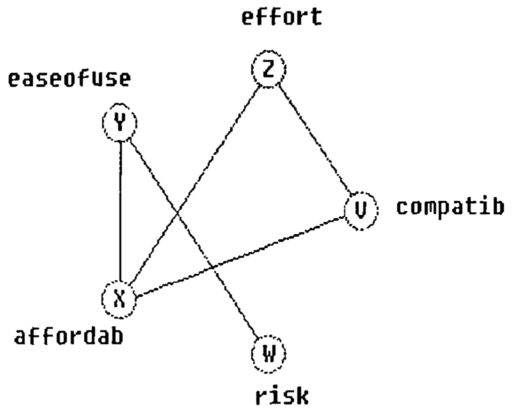


***Figure 9.5**: Independence graph for segment 5 (Complexity Avoiders)*

Comments

Following on the comments for segment 3, the graph for the *Complexity Avoiders* reveals that *Ease of use* is now not conditionally independent of *Affordability*, given the rest of the factors. This is perhaps not surprising: the inspection of the segment profiles (Chapter 7, table 7.4) indicates that this segment corresponds to combined below-the-average web usage/experience and household income.

Segment 6 (Cost Conscious)

The selection proceeded as summarized in *table 9.7*. The formula of the final model selected was //VY, WZ. It is represented by the independence graph in figure 9.6.

Comments

The graph for the *Cost Conscious* segment shows a very generalized conditional independence between factors. The only remaining links are rather obvious, established between *Ease of Use* and *Compatibility*, and then *Perception of Risk* and *Effort Responsiveness*. It is revealing that the only factor conditionally independent of all the rest is precisely *Affordability*, in the same way that *Compatibility* was independent in the *Unconvinced* segment and, not surprisingly, *Effort/Responsiveness* is to be independent in the *Customer Service Wary* segment. *Affordability*, in the GTM reference maps, was the clearest descriptor of the segment.

Edge removed	*G^2*	*p-value*
[VZ]	0.0005	0.9829
[XZ]	0.1287	0.7198
[WX]	0.0660	0.7972
[VX]	0.1365	0.7118
[YZ]	0.1433	0.7050
[VW]	0.7054	0.4010
[WY]	1.0312	0.3099
[XY]	2.4482	0.1177

Table 9.7: Summary statistics of the model selection for segment 6.

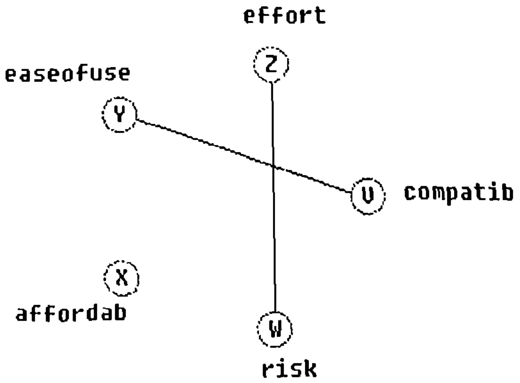


Figure 9.6: Independence graph for segment 6 (Cost Conscious)

Segment 7 (Customer Service Wary)

The selection proceeded as summarized by the statistics in *table 9.8*. The formula of the final model selected was //VW, WX, XY, Z. It is represented by the independence graph in figure 9.7.

Edge removed	G^2	*p-value*
[VZ]	0.3108	0.5772
[XZ]	0.2774	0.5984
[VX]	0.3113	0.5769
[YZ]	1.1511	0.2833
[VY]	1.7487	0.1860
[WY]	0.0253	0.8736
[WZ]	2.8917	0.0890

Table 9.8: *Summary statistics of the model selection for segment 7.*

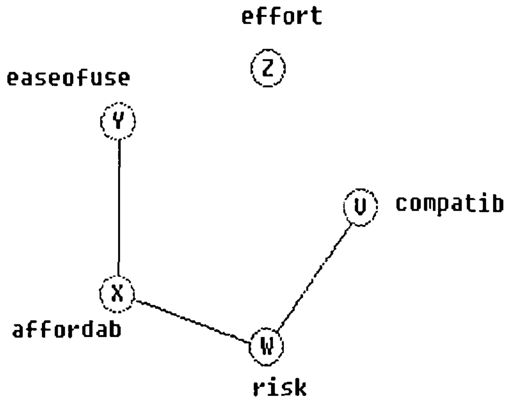


Figure 9.7: *Independence graph for segment 7 (Customer Service Wary)*

Comments

Yet another sparsely connected graph for the *Customer Service Wary*. As stated in the comments to the previous segment, the main feature of this graph is the conditional independence of *Effort Responsiveness* with respect to the remaining factors. *Effort*

Responsiveness, a factor describing perceptions of customer service, was obviously the main descriptor of this segment, according to the GTM reference maps.

All data

The selection proceeded as summarized by the statistics in *table 9.9*. The formula of the final model selected was //YW. It is represented by the independence graph in figure 9.8.

Edge removed	G^2	*p-value*
[WZ]	0.0829	0.7734
[YZ]	0.1351	0.7133
[VY]	0.2270	0.6337
[VZ]	0.3407	0.5594
[VX]	0.1274	0.7211
[XZ]	0.5464	0.4598
[WX]	0.8701	0.3509
[VW]	1.2827	0.2574
[XY]	1.6536	0.1985

Table 9.9: *Summary statistics of the model selection for the complete data set.*

Comments

The only use of the analysis of the independence graph for the whole data set is to find out how it compares with the segment-specific results. Basically, almost all the factors are pairwise conditionally independent given the remaining ones. Even the edge between *Ease of Use* and *Perception of Risk* remains on the basis of a *p*-value (0.0398) that is just within the limits of the cut-off limit, set to 0.05. The fact that the individual segments yield a very diverse, and more or less complex, web of significant linking edges indicates that the segment-specific independence graphs can be seen as a supplementary source of marketing intelligence.

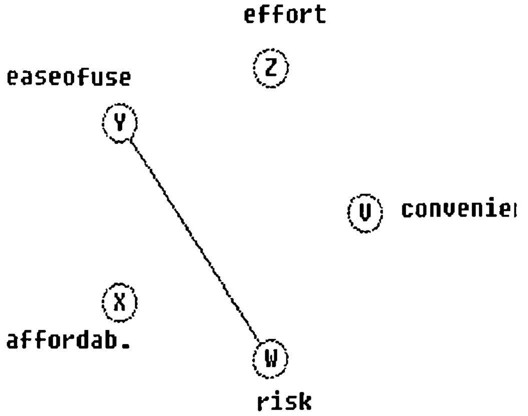


Figure 9.8: Independence graph for the whole data set

9.3.1. Marketing oriented comments

The graphical model for the whole data set indicated the pairwise conditional independence of virtually all factors, given the remaining ones. It means that, from the market analyst point of view, each factor has to be dealt with separately. On the other hand, each segment is graphically represented by a more or less complex web of adjacent factors, therefore departing from the overall model. Let us focus on segments dominated by non-purchasers. In the previous section, it has been shown that there are three segments: *Unconvinced*, *Cost Conscious* and *Customer Service Wary*, for which a single factor: namely, in turn, *Compatibility*, *Affordability* and *Effort Responsiveness*, is conditionally independent of all the rest. It is quite interesting that, in every case, it is the factor that strongly defines the segment. For the marketer, this could mean that those factors should be tackled on their own and that changes to those factors could strongly change the overall features of the segment.

On the other hand, there are several segments for which any marketing action could be focused on the relationships between factors, instead of being based on their conditional independence. Let us now go back to the similarities between the graphs for the *Security Conscious*, the *Undecided* and the *Complexity Avoiders*. For all these segments, there is a strong triangular relationship between *Compatibility*, *Affordability* and *Effort / Responsiveness*. It could be argued that acting on any of those factors might trigger changes in the others, which are not obvious otherwise. For instance, making an effort to improve the vendor's responsiveness could entail improving the perception of affordability and making the customer feel that the process of shopping online is more convenient and/or compatible with his/her shopping and life styles.

9.4. Segment-specific graphical mixed models

The framework for the graphical modeling with continuous variables is now extended to include a discrete variable informing of whether the Internet user has ever purchased online or not. This is the same variable that was described and utilized in parts 3 and 4 of the thesis. Given that this variable is not continuous, mixed models that encompass continuous and discrete variables, described in section 9.2.2, have to be utilized.

The rationale behind the inclusion of this discrete variable as part of the graphical analyses is straightforward: although it was not used as a segmentation basis, it has been shown in previous chapters that different segments show different purchasing behaviour. The definition of the independence graphs will provide segment-specific information on how the different factors relate to the propensity to buy online. More precisely, it would be useful to discover which factors are conditionally independent of the propensity to buy online, because it would permit the marketer to focus marketing efforts on those factors which are not.

The results of the application of the stepwise backward model selection method, using mixed models, to the GTM segment partition is summarized next. For the sake of brevity not all the segment analyses are reported. The names of the factors follow again the convention shown in *table 9.1*, and the discrete variable, that will be referred to as *Purchase*, is represented by letter A. In all cases, the selection procedure is started with a saturated model, from which individual edges are removed iteratively. For each segment, the deviance and the *p*-values corresponding to either the χ^2-tests or F-tests are provided.

Segment 3 (Undecided)

The selection proceeded as summarized by the statistics in *table 9.10*. The formula of the final model selected was A/ Z, X, Y, AW, AV/ VZ, VXY, AW, AV. It is represented by the independence graph in figure 9.9.

Comments

Three factors turn out to be conditionally independent of the propensity to buy online: *Affordability*, *Ease of Use* and *Effort Responsiveness*. Interestingly, only *Compatibility* acts as a mediator for all of them. It might be argued that those three factors only influence the propensity to buy online through the perception of compatibility and not through the *Perception of Risk*, although this factor is also adjacent to *Purchase*. From the point of view of the marketer, in the case of the *Undecided*, it might be worth prioritizing effort and resources on tackling the perceptions of compatibility and risk.

Edge removed	G^2	***p-value***
[WY]	0.3341	0.8462
[WZ]	0.5833	0.7470
[YZ]	0.8575	0.6513
[XZ]	0.8732	0.6462
[AZ]	3.6928	0.2966
[WX]	2.7563	0.2520
[AX]	4.4566	0.3477
[AY]	6.8064	0.0783
[VW]	5.7604	0.0561

Table 9.10: *Summary statistics of the model selection for segment 3.*

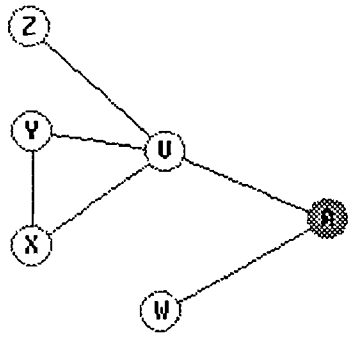


Figure 9.9: *Independence graph for segment 3 (Undecided)*

Segment 4 (Convinced)

The selection proceeded as summarized by the statistics in *table 9.11*. The formula of the final model selected was A/ AV,AW, AX, AY, AZ/ AY, AWZ, AVX, AVW. It is represented by the independence graph in figure 9.10.

Edge removed	G^2	*p-value*
[XY]	0.5228	0.7700
[WX]	1.0203	0.6004
[XZ]	2.1468	0.3418
[WY]	2.5877	0.2742
[VY]	3.3631	0.1861
[VZ]	4.7034	0.0952
[YZ]	4.9062	0.0860

Table 9.11: *Summary statistics of the model selection for segment 4.*

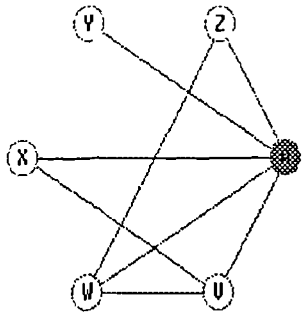


Figure 9.10: *Independence graph for segment 4 (Convinced)*

Comments

The *Convinced* segment is characterized by the fact that none of the factors is conditionally independent of the propensity to buy online. In fact, the *p*-values for the deletion test in the backward selection procedure (not reported in the tables) indicate that the removal of the edges [AV], [AW], [AX] and [AZ] is strongly rejected. The only other segment for which none of the factors is independent of *Purchase* is precisely the *Unconvinced.* These two segments, the *Convinced* and the *Unconvinced,* occupy the extremes of the propensity to buy online spectrum. The results imply that, in the case of the former, the marketer can not leave any of the factors "unattended" and, in the case of the latter, the marketer has to actively pursue to influence the potential customer's decision making through all the factors simultaneously.

Segment 5 (Complexity Avoiders)

The selection proceeded as summarized by the statistics in table 9.12. The formula of the final model selected was A/ X, W, AZ, AY, AV/ WX, VXY, AVZ, AVY. It is represented by the independence graph in figure 9.11.

Edge removed	G^2	*p-value*
[WY]	0.5790	0.7486
[WZ]	1.0190	0.6008
[YZ]	1.7439	0.4181
[XZ]	3.7997	0.1496
[AW]	8.4216	0.0773
[VW]	0.0014	0.9697
[AX]	7.5129	0.1111

Table 9.12: Summary statistics of the model selection for segment 5.

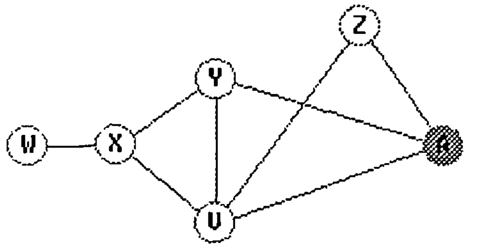


***Figure 9.11**: Independence graph for segment 5 (Complexity Avoiders)*

Comments

The graph for this segment indicates that the *Perception of Risk* and *Affordability* factors are conditionally independent of the propensity to buy online. In this case *Ease of Use* (somehow naturally for the *Complexity Avoiders*) and *Compatibility* seem to be the mediators with respect to *Purchase*. From the point of view of the marketer, this can be a confirmation that special attention should be paid to these mediators as well as to the factor of *Effort Responsiveness*, which is also adjacent to *Purchase.*

Segment 6 (Cost Conscious)

The selection proceeded as summarized by the statistics in *table 9.13*. The formula of the final model selected was A/ Z, Y, AX, AV, AW/ VXZ, VXY, AVX, AVW. It is represented by the independence graph in figure 9.12.

Comments

This time, two factors: *Ease of Use* and *Effort Responsiveness,* turn out to be conditionally independent of the propensity to buy online. This time, the mediators are

Affordability (Let us bear in mind that we are describing the *Cost Conscious*) and, again, *Compatibility*, which seems to assume this mediation role in most of the segments. Marketing action could relegate *Ease of Use* and *Effort / Responsiveness* in favor of the mediators and the *Perception of Risk*.

Edge removed	G^2	*p-value*
[WX]	0.5583	0.7564
[WZ]	0.9021	0.6370
[YZ]	0.9949	0.6081
[WX]	3.4792	0.1756
[AY]	3.9239	0.416
[AZ]	9.3991	0.05

Table 9.13: Summary statistics of the model selection for segment 6.

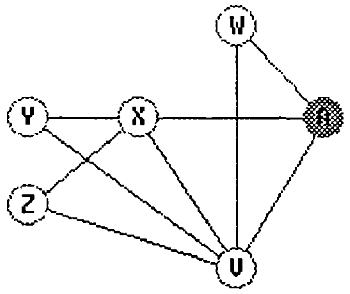


Figure 9.12: Independence graph for segment 6 (Cost Conscious)

9.5. Conclusion

In this chapter we have applied graphical modeling techniques to reveal the segment-specific structure of the inter-relationships between segmentation bases and also between these and the segment-specific propensity to buy online. In the previous chapter, the description of segments using visualization techniques such as the GTM reference maps and their characterization using profiles, built with variables that are exogenous to the segmentation, are both blind to the likely existence of these inter-relationships.

With the application of graphical models, we have mapped the multivariate relationships for each segment. The resulting independence graphs provide extra marketing intelligence that could not have been provided by the segmentation methodologies themselves. These new insights can help to improve the actionability of the individual segments, which confirms the potential of graphical modeling as a tool for market analysis that is able to amplify the power of cluster-based segmentation methodologies.

Some interaction graphs of interest have appeared in the graphical Gaussian models in section 9.3. Several segments contain factors that are conditionally independent of all the rest, and these are precisely the factors that define those segments more strongly. From the graphical point of view, such segments could be labeled as "factor-specific", and a marketer could opt to act on them only, or specially, through those factors. On the other hand, other group of segments reveals an entangled web of relations between factors. In this case, a marketer could choose to target these "spread inter-relation" groups with a mixed strategy that addresses all factors.

The exploration of the inter-relationships between the propensity to buy online and the factors used as segmentation bases has helped to discover which factors are conditionally independent of that propensity. This information can orient the online marketer to focus resources and efforts on factors that have a direct influence on the consumers' buying decision. The results in section 4 indicate that, for some segments, all factors are related to the propensity to buy online. Nevertheless, most of the graphical mixed models fitted to the segments indicate that some factors only influence the propensity to buy online through mediation factors, a valuable result for the reasons stated above.

We have now come a long way into the description of a quantitative methodology for the analysis of data related to Internet users' opinions on e-commerce. The original data were first reduced to a more parsimonious description using latent variable methods. The power of these underlying factors to predict the propensity to buy online was analyzed next and, with this, a first approach to the discovery of causal relationships in the data was attempted. The most predictive factors were then used as bases for cluster-based segmentation. Now, in this chapter, the segment-specific inter-relationships between segmentation bases, and between these and the propensity to buy online, have been mapped. With this, not only a new stream of information has been channeled to the marketing intelligence flow available to the online vendor but, also, the framework of an overall methodology has come full circle as the independence graphs set new bases for a more refined causal exploration of the data.

PART 6

Conclusions and directions for future research

Por primera vez en nuestra historia, somos contemporáneos de todos los hombres.

Octavio Paz

Chapter 10

Conclusions and directions for future research

10.1. Introduction

In this final part, the overall structure and proposals of the thesis are recapitulated. This is followed by a brief summary of the most important results obtained in each of the parts and chapters in which the thesis is organized. Finally, several proposals are made for future research that could build upon the results obtained in this thesis.

10.2. Recapitulation of the thesis

A complete and systematic quantitative methodology for the analysis of consumers' opinions of business-to-consumer electronic commerce has been presented. This methodology is centered on neural network models and comprises four successive stages of implementation, as follows:

- **Latent variable analysis and dimensionality reduction:** A survey data set containing a large number of online behaviour-related variables, from a sample of Internet users' opinions of online shopping and online vendors, is made available for our study. It also includes demographic, socio-economic and Web usage information about the Internet users. Questionnaire data involves a large number of questions, many of which are related. Before carrying out a quantitative analysis, it is necessary to extract the latent traits behind these consumer opinions. Factor analysis is a common statistical method for doing this, since it provides a well founded procedure to determine how many factors are required to represent the data, and is a preprocessing methodology generally accepted by marketers. Therefore, the dimensionality of the data is reduced, whereas its interpretation is considerably eased and undesirable artifacts inherent to survey design are removed.

- **Predictive modelling:** The latent factors, together with the demographic, socio-economic and Web usage information, are used to construct a global predictive model of the propensity to buy online. Different prediction models, including neural networks, are used and their results compared. A Bayesian Neural Network model is utilized for this purpose, and it is implemented using Automatic Relevance Determination (ARD) for model selection. This provides a pruning of the descriptive features identified by factor analysis, keeping only those which are predictive within this non-linear model. Furthermore, a lower limit is set on the predictive power of the segmented model developed in the next stages.
- **Cluster-based segmentation:** Those factors selected as the best predictors of the propensity to buy online are used as bases for the segmentation of the online consumer market, utilizing a neural network-based model for clustering and visualization: the Generative Topographic Mapping (GTM). For this model, the distribution of network coefficients provides a ready labelling for each of the resulting segments. These segments are further characterized and made actionable

through the use of demographic, socio-economic and Web usage profiles. The statistically principled nature of the GTM model makes it possible to provide a posterior probability of cluster membership, given the data, as well as to properly define the cluster boundaries. The visualization provided by the mappings of this model entails topologically ordered maps, which directly reflect the structure of the data in the multivariate sense, and gives an indication of whether sample points lay on inside clusters or near cluster boundaries.

- **Graphical modelling:** Having identified the different consumer segments, it is now possible to apply Graphical Modelling and, more specifically, Conditional Independence Maps (CIM) separately to each segment, to produce graphs representing the inter-relationships between the predictive factors, or segmentation bases, and between these and the propensity to buy online. The resulting qualitative description of each segment, mapping its structure of inter-relationships, goes beyond the simple descriptive characterization of the segments provided in the previous stage of the methodology. This description, as a source of marketing intelligence, provides a better handle to the actionability of the market segments.

This last stage closes the methodological loop. The market representation embedded in the qualitative structural maps for each segment is a new stream of information channeled to the online vendor. Also, the framework of an overall methodology has come full circle as the CIMs set new bases for a more refined causal exploration of the data. This can also lead to a refinement of the original questionnaire design, given the qualitative models derived from the e-commerce market segmentation, either by introducing more detailed questions on key predictive features or, where the predictive model is felt to need further improvement, by considering segment-specific additional questions, on the back of the qualitative knowledge already derived. The updated closed-loop methodology is represented by figure 10.1:

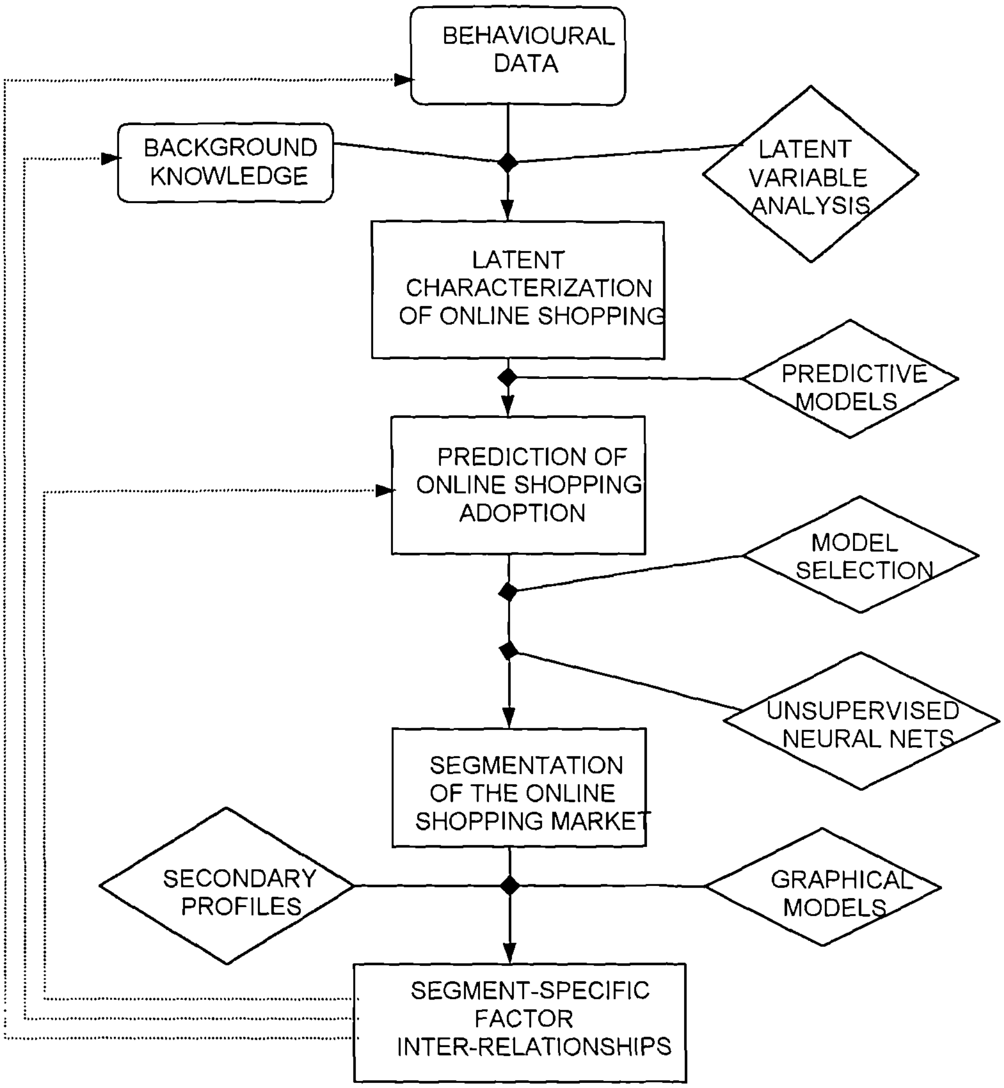

BEHAVIOURAL DATA
BACKGROUND KNOWLEDGE
LATENT VARIABLE ANALYSIS
LATENT CHARACTERIZATION OF ONLINE SHOPPING
PREDICTIVE MODELS
PREDICTION OF ONLINE SHOPPING ADOPTION
MODEL SELECTION
UNSUPERVISED NEURAL NETS
SEGMENTATION OF THE ONLINE SHOPPING MARKET
SECONDARY PROFILES
GRAPHICAL MODELS
SEGMENT-SPECIFIC FACTOR INTER-RELATIONSHIPS

***Figure 10.1**: An augmented closed-loop scheme of the methodology described in the thesis.*

10.3. Main results of the thesis

What follows is a brief account of the main results related to each of the parts in which the thesis has been organized, as listed in the previous section.

- **Latent variable analysis and dimensionality reduction:**

  In chapter 2, the factor extraction experiments have quantitatively demonstrated the existence of a parsimonious latent structure underlying the Internet users' opinions of online shopping, explainable in meaningful, operative terms which are consistent with the findings of existing mainly qualitative studies (Jarvenpaa and Todd, 1996/97; Chaum, 1997; Hoffman and Novak, 1997). This simple description of online customers' opinions alleviates the intrinsic drawbacks of high-dimensional data sets and survey-based data. It is the basis for the rest of the experiments carried out in this thesis.

- **Predictive modelling:**

  In chapter 3, using the factor scores obtained in the previous part for variable selection, the following factors, in order of relevance, are found to be the best predictors of the propensity to purchase online. *Consumer risk perception Environmental control*, *Shopping experience: Compatibility*, *Affordability*, *Shopping experience: Effort: ease of use* and *Shopping experience Customer service: responsiveness and empathy*. It is also found that age, household income, and web usage patterns do not add to the predictive power for online purchasing behaviour. This outcome agrees with findings in previous studies.

Finally, it is shown that the propensity to buy on-line can be predicted to a surprising degree of accuracy using neural network and logistic discrimination models. Furthermore, the use of the five most relevant factors retains the accuracy of larger models in the discrimination of purchasers from non-purchasers, resulting in a much more manageable model than the one including the original 48 observable variables from the raw data.

In chapter 4, the Bayesian approach for the training of neural networks is explored. It is shown that, for the classification problem described in the previous chapter, the marginalization of the test outputs did not entail any loss of overall discriminatory power for the model. The reject option, to defer decision in circumstances of uncertainty in the neural network predictions, is described as providing guidance to assess the effect of the test output marginalization on the relation between rejection ratio, accuracy and width of the uncertainty region. This management of the prediction uncertainty should be at the core of the online retailers' decision making. This chapter shows how, beyond the use as a classifier, the probabilistic formulation of the Bayesian MLP can provide systematic feature selection and a framework for the management of decision uncertainty.

Finally, in chapter 5, a principled procedure for the training of Bayesian neural networks for the case of class-unbalanced data sets is provided. It is illustrated with the classification problem described in chapter 3. This procedure helps to avoid introducing bias into the model predictions as, in all binary classification problems with skewed data, the default marginalisation introduces a systematic shift away from correct calibration. Moreover, it avoids the need to change the sampling procedure in order to equalise the priors, having to train the network with artificially class-balanced data sets.

- **Cluster-based segmentation:**

In chapter 6, the GTM model is presented as a statistically principled alternative to the Self-Organizing Map (SOM) and its suitability for market segmentation is argued and defended. In chapter 7, the GTM is used to carry out cluster-based market segmentation, within the *tandem approach* of segmenting using the factors (obtained in chapter 2 and selected in chapter 3) as bases. The GTM is shown to be a powerful data visualization tool, capable of assisting the data analysis carried out by supervised models in part 3 of the thesis, by providing these with further explanatory capabilities. It is also successfully applied, in a completely unsupervised mode, to discover the clusters or segments in which the data are naturally organized. Its use for micro- and macro-segmentation is illustrated and, for the latter, two different strategies are described. The segmentation carried out with the most parsimonious set of bases (5 factors) is shown to provide the best class-discrimination according to an entropy measure. The resulting segments are profiled using secondary information in order to make the segment solution more actionable from the marketing point of view.

In chapter 8, we present a novel method, the *Selective Mapping Smoothing (SMS)*, for the automatic optimization of one of the GTM complexity-controlling parameters. A general entropy metric is defined to evaluate the class-discriminatory capabilities of different cluster solutions produced by the *SMS* model and other existing GTM models. The *SMS* model is shown to outperform the rest according to this entropy metric.

- **Graphical modelling:**

In chapter 9, graphical modelling techniques are applied to reveal the segment-specific structure of the inter-relationships between segmentation bases and also between these and the segment-specific propensity to buy online. The segment partition obtained using the GTM model in chapter 7 is the test bed for this application.

The experiments reveal that several segments contain factors that are conditionally independent of all the rest, and these are precisely the factors that define those segments more strongly (chapter 7). From the graphical point of view, such segments could be labeled as "factor-specific", and a marketer could opt to act on them only, or specially, through those factors. On the other hand, other group of segments reveals an entangled web of relations between factors. In this case, a marketer could choose to target these "spread inter-relation" groups with a mixed strategy that addresses all factors.

The exploration of the inter-relationships between the propensity to buy online and the factors used as segmentation bases has helped to discover which factors are conditionally independent of that propensity. The experiments indicate that, for some segments, all factors are related to the propensity to buy online. Nevertheless, most of the graphical mixed models fitted to the segments indicate that some factors only influence the propensity to buy online through mediation factors.

10.4. Directions for future research

The methodology presented in this thesis can provide the online market analyst with a very detailed snapshot of the state of affairs in the consumer e-commerce market. In most real-world applications of the methodology, though, it would be much more informative, and probably necessary, to deal with time dependent data. The GTM model can handle the analysis of time-dependent data (Bishop *et al.*, 1997a). Further research is encouraged in this area, given that the most accessible information to online marketers is precisely time-dependent Web log data (Chang, 1998), which can be transformed into formats analyzable by modern data mining techniques (Slater *et al.*, 1999). Invaluable information on the dynamics of consumers across the map of segments would be generated, and the calculation of their *most likely trajectories* would enable the online marketer to develop dynamic segmentation forecasting techniques and assess the stability of existing segments. It has recently been stated that "highly sophisticated tracking tools to monitor changing customer preferences are necessary to maintain the flexibility of the online marketing channel" (Kiang *et al.*, 2000).

In a practical implementation of the proposals of this thesis, the methodology that has been put forward might be conceptualized as a *user modelling* toolbox. From this point of view, such a toolbox could be integrated into a more complex framework oriented to the *personalization* of the contents directed towards segments of consumers or even individual consumers. One such framework, based on *third-generation Web pages* and *intelligent agent*-based, assistant-like, technologies, has been developed and deployed in real-world situations by Wahlster (2000). It responds to the "unique identification of the temporal and logical structure of interaction sequences of single users" by generating personalized Web pages created on the fly. For this, the user modelling techniques provide user "stereotypes" by "statistical evaluation of the interaction behavior when accessing Web pages", so that the "anticipation of further interaction behaviour of a user by assigning him to a stereotype" becomes possible.

Acknowledgements

The author gratefully acknowledges the following persons for their contribution to the completion of this labour of love:

Paulo, for trusting me so much, for being right most of the time, and for he is such an awesome example to follow.

Karon, for her advice has always been sound.

Wael, for no guardian angel ever knew so much about Cuban music.

Helen and Barbara, for fun and gossips (and the odd serious discussion)

References

Allenby, G.M. and Lenk, P.J. (1994): Modelling household purchase behaviour with logistic normal regression. *Journal of the American Statistical Association,* 89(December), 1218-1231.

Altman, E.I., Marco, G. & Varetto, F. (1994): Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks. *Journal of Banking and Finance,* 18, 505-529.

Anand, S.S., Patrick, A.R., Hughes J.G., and Bell, D.A. (1998): A data mining methodology for cross-sales. *Knowledge-Based Systems,* 10(7), 449-461.

Andrews, R. and Diederich, J. (eds.) (1996): *Rules and Networks. Proceedings of the Rule Extraction From Trained Artificial Neural Networks Workshop, AISB'96.* Queensland University of Technology, UK.

Arabie, P., and Hubert, L. (1994): Cluster analysis in market research. In Bagozzi, R.P. (Ed.), *Advanced methods in marketing research.* Oxford: Blackwell & Company, 160-189.

Attias, H. (1999): Inferring parameters and structure of latent variable models by Variational Bayes. To appear in: *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*.

Back, B., Laitinen, T., and Sere, K. (1996): Neural networks and genetic algorithms for bankruptcy predictions. *Expert Systems with Applications,* 11(4), 407-413.

Back, B., Irjala, M., Sere, K, and Vanharanta, H. (1997): Competitive financial benchmarking using self-organizing maps. *Paperi Ja Puu - Paper and Timber*, 79(1), 42-49.

Balakrishnan, P.V.S., Cooper, M.C., Jacob, V.S.. and Lewis, P.A. (1996): Comparative performance of the FSCL neural net and K-means algorithm for market segmentation. *European Journal of Operational Research,* 93, 346-357.

Bellman, S., Lohse, G.L., and Johnson, E.J. (1999): Predictors of online buying behaviour, *Communications of the ACM,* 42(12), 32-38.

Bioch JC, van der Meer O, Potharst R. (1995): Classification using Bayesian neural nets. In: Decaestecker C (ed.). *Proceedings Benelearn '95. Universiteit Brussel, Brussel,* , pp. 79-90.

Bishop, C.M. (1995): *Neural Networks for Pattern Recognition.* Oxford: Oxford University Press.

Bishop, C.M., Hinton, G.E., and Strachan, I.G.D. (1997a): GTM through time, in *Proceedings IEE Fifth International Conference on Artificial Neural Networks,* Cambridge, U.K., 111-116.

Bishop, C.M., Svensén, M. and Williams, C.K.I. (1997b): Magnification factors for the GTM algorithm, in *Proceedings IEE Fifth International Conference on Artificial Neural Networks,* Cambridge, U.K., 64-69.

Bishop, C.M., Svensén, M., and Williams, C.K.I. (1998a): GTM: the Generative Topographic Mapping. *Neural Computation,* 10(1), 215-234.

Bishop, C.M., Svensén, M., and Williams, C.K.I. (1998b): Developments of the Generative Topographic Mapping. *Neurocomputing* 21(1-3), 203-224

Bloch, M., Pigneur, Y., and Segev, A. (1996): On the road to electronic commerce – a business value framework, gaining competitive advantage and some research issues. *Proceedings of the Ninth International EDI-IOS Conference,* Bled: Slovenia.

de Bodt, E., Henrion, E-F., Cottrell, M., and van Wymeersch, C. (1998). Self-organizing maps for data analysis: An application to the Belgian leasing market. *Journal of Computational Intelligence in Finance,* 6(6), 5-24.

Borowsky, M. (1995): Scoring puts up higher numbers. *US Banker,* 105(1), 44-46, 63.

Bradley, P. (1997): The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition,* 30(7), 1145-1159.

Brannback, M. (1997): Is the Internet changing the dominant logic of marketing? *European Management Journal,* 15(6), 698-707.

Brockett, P.L., Xia, X.H., and Derrig, R.A. (1998): Using Kohonen's self-organizing feature map to uncover automobile bodily injury claims fraud. *The Journal of Risk and Insurance,* 65(2), 245-274.

Chang, S. (1998): Internet Segmentation: State-of-the-art marketing applications. *Journal of Segmentation in Marketing,* 2(1), 19-34.

Chatfield, C. (1993): Neural networks: Forecasting breakthrough or passing fad?. *International Journal of Forecasting,* 9, 1-3.

Chatfield, C. (1995): Positive or negative?. *International Journal of Forecasting*, 11, 501-502.

Chaum, D. (1997): How much do you trust Big Brother? *IEEE Internet Computing*, 1(6), 8-16.

Chen, S.K., Mangiameli, P., and West, D. (1995). The comparative ability of self-organizing neural networks to define cluster structure. *Omega: International Journal of Management Science*, 23(3), 271-279.

Chen, M.S., Han, J. and Yu, P.S. (1996): Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering*, 8(6), 866-884.

Curry, B. and Morgan, P. (1996): Neural networks: A need for caution. *Omega: International Journal of Management Science*, 25(1), 123-133.

Dasgupta, C.G., Dispensa, G.S. and Ghose. S. (1994): Comparing the predictive performance of a neural network model with some traditional market response models. *International Journal of Forecasting*, 10, 235-244.

Davies, F., Moutinho, L. and Curry, B. (1996): ATM user attitudes: a neural network analysis. *Marketing Intelligence & Planning*, 14(2), 26-32.

Dawid, A.P. (1979): Conditional Independence in statistical theory (with discussion). *Journal of the Royal Statistical Society (B)*, 41, 1-31.

Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1), 1-38.

Deng, P-S. (1993): Automatic knowledge acquisition and refinement for decision support: A connectionist inductive inference model. *Decision Sciences*, 24(2), 371-393.

Desay, V.S., Crook, J.N. and Overstreet Jr., G.A. (1996): A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, 95, 24-37.

Edwards, D. (1995): *Introduction to Graphical Modelling*. New York: Springer-Verlag.

Farquhar, B.J. and Balfour, A. (1998): Consumer needs in global electronic commerce: The role of standards in addressing consumer concerns. *EM: Electronic Markets Newsletter*, 8 (2), 1-5.

Firat, A.F., and Shultz II, C.J. (1997): From Segmentation to fragmentation: Markets and marketing strategy in the postmodern era. *European Journal of Marketing,* 31(3-4), 183-207.

Fish, K.E., Barnes, J.H. and Aiken, M.W. (1995): Artificial neural networks - A new methodology for industrial market segmentation. *Industrial Marketing Management,* 24, 431-438.

Gabriel, K.R. (1969): Simultaneous test procedures: Some theory of multiple comparisons. *Annals of Mathematics and Statistics,* 40, 224-250.

Ghahramani, Z. and Beal, M.J. (2000): Variational inference for Bayesian mixture of factor analysers. To appear in: *Advances in Neural Information Processing Systems,* 12.

Glasner, J. (1999): A cranky e-commerce Christmas, *Wired News*, (19 January),
URL: http://www.wired.com/news

Glorfeld, L.W. and Hardgrave, B.C. (1996): An improved method for developing neural networks: the case of evaluating commercial loan creditworthiness. *Computers & Operations Research,* 23(10), 933-944.

Gordon, M.E., and De Lima-Turner, K. (1997): Consumer attitudes towards Internet advertising: A social contract perspective. *International Marketing Review,* 14(5), 362-375.

Green, P.E. and Krieger, A.M. (1995): Alternative approaches to cluster-based market segmentation. *Journal of the Market Reseach Society,* 37(3), 221-239.

Grover, R., and Srinivasan, V. (1989): An approach for tracking within-segment shifts in market shares. *Journal ofMarketing Research,* 26, 230-236.

Ha, S.H. and Park, S.C. (1998): Application of data mining tools to hotel data mart on the Intranet for database marketing. *Expert Systems With Applications,* **15**, 1-31.

Hand, D.J. (1997): *Construction and Assessment of Classification Rules*, Chichester: John Wiley & Sons.

Hanley, J.A., and McNeil, B.J. (1982): The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, 143, 29-36.

Hanley, J.A., and McNeil, B.J. (1983): A method for comparing the areas under Receiver Operating Characteristic curves derived from the same cases. *Radiology,* 148, 839-843.

Hinton, G.E., Williams, C.K.I., and Revow, M.D. (1992): Adaptive elastic models for hand-printed character recognition. In Moody, J.E., Hanson, S.J., and Lippmann, R.P. (Eds.) *Advances in Neural Information Processing Systems,* 4, 512-519. Morgan Kauffmann.

Hoffman, D.L., Kalsbeek, W.D. and Novak, T.P. (1996): Internet and Web use in the United States: baselines for commercial development. *Communications of the ACM,* 39 (Dec.), 36-46.

Hoffman, D.L. and Novak, T.P. (1996): Marketing in hypermedia computer-mediated environments: Conceptual foundations. *Journal of Marketing,* 60(July): 50-68.

Hoffman, D.L., and Novak, T.P. (1997): A new marketing paradigm for electronic commerce. *The Information Society,* 13(1), 43-54.

Hoffman, D.L., Novak, T.P., and Peralta, M.A. (1999): Information privacy in the marketspace: implications for the commercial uses of anonymity on the Web. *The Information Society,* 15(2), 129-139.

Iconocast (2000): *Internet at a glance.* URL: www.iconocast.com

Jarvenpaa, S.L., and Todd, P.A. (1996/1997): Consumer reactions to electronic shopping on the WWW. *International Journal of Electronic Commerce,* 1(2), 59-88.

Jo, H., Han, I. and Lee, H. (1997): Bankruptcy prediction using case-based reasoning, neural network, and discriminant analysis. *Expert Systems with Applications,* 13(2), 97-108.

Jung, D. and Burns, J.R. (1993): Connectionist approaches to inexact reasoning and learning systems for executive and decision support - Conceptual design. *Decision Support Systems,* 10, 37-66

Kaiser, H.F. (1974): An Index of Factorial Simplicity, *Psychometrica*, 39, 31-36.

Kara, A., and Kaynak, E. (1997): Markets of a single customer: exploiting conceptual developments in market segmentation. *European Journal of Marketing,* 31(11-12), 873-895.

Kaski, S., Nikkilä, J., and Kohonen, T. (1998): Methods for interpreting a self-organized map in data analysis. In: *Proceedings of European Symposium on Artificial Neural Networks (ESANN),* Bruges, Belgium.

Kehoe, C., Pitkow, J., and Rogers, J.D. (1998): 9th GVU's WWW user survey.
URL: http://www.gvu.gatech.edu/user_surveys/survey-1998-04/

Kiang, M.Y., Raghu, T.S., and Shang, K.H-M. (2000): Marketing on the Internet-who can benefit from an online marketing approach. *Decision Support Systems,*27, 383-393.

Kiviluoto, K., and Bergius, P. (1998a): Two level self-organizing maps for analysis of financial statements. *Proceedings of the International Joint Conference on Neural Networks (IJCNN'98), Anchorage, Alaska,* 189-192.

Kiviluoto, K. (1998b): Predicting bankruptcies with the self-organizing map. *Neurocomputing,* 21(1-3), 203-224.

Kohonen, T. (1982): Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* 43(1), 59-69.

Kohonen, T. (1995): *Self-organizing Maps.* Berlin: Springer-Verlag.

Kraaijveld, M.A., Mao, J.C., and Jain, A.K. (1995): A nonlinear projection method based on Kohonen's topology preserving-maps. *IEEE Transactions on Neural Networks,* 6(3), 548-559.

Krzanowski, W.J. (1996): *Principles of Multivariate Analysis.* Oxford: Clarendon Press.

Kullback, S. (1968): *Information Theory and Statistics*. New York: Dover.

Lauritzen, S.L. (1992): Graphical association models (draft version). Research report, Institute of Electronic Systems. University of Aalborg.

Lee, S.S. (1999): Regularization in skewed binary classification. *Computational Statistics*, 14(2), 277-292.

Lee, K.C., Han, I. and Kwon, Y. (1996): Hybrid neural network models for bankruptcy predictions. *Decision Support Systems*, 18, 63-72.

Lenard, M.J., Alam, P. and Madey, G.R. (1995): The application of neural networks and a qualitative response model to the auditor's going concern uncertainty decision. *Decision Sciences*, 26(2), 209-227.

Lewis, O.M., Ware, J.A., and Jenkins, D. (1997): A novel neural network technique for the valuation of residential property. *Neural Computing & Applications*, 5(4), 224-229.

Lisboa, P.J.G., El-Deredy, W., Vellido, A,, Etchells, T. and Pountney, D.C. (1997): Automatic variable selection and rule extraction using neural networks. In *Proceedings of the 15th IMACS World Congress on Scientific Computation, Modelling and Applied Mathematics*, Berlin, 461-466.

Lowe, D., and Webb, A.R. (1991): Optimized feature extraction and the Bayes decision in feed-forward classifier networks. *IEEE-PAMI*, 13(4), 355-364.

Mackay, D.J.C. (1992a): A practical Bayesian framework for back-propagation networks. *Neural Computation*, 4(3), 448-472.

Mackay, D.J.C. (1992b): The evidence framework applied to classification networks. *Neural Computation*, 4(5), 720-736.

Mackay, D.J.C. (1995): Probable networks and plausible predictions - a review of practical Bayesian methods for supervised neural networks. *Network: Computation in Neural Systems,* 6, 469-505.

Magidson, J. (1994): The CHAID Approach to Segmentation Modelling: CHI-squared Interaction Detection, in *Advanced Methods in Marketing Research*, (Bagozzi, R.P., ed). Oxford: Blackwell & Company.

Martin-del-Brio, B. and Serrano-Cinca, C. (1993): Self-organizing neural networks for the analysis and representation of data: Some financial cases. *Neural Computing & Applications,* 1, 193-206.

Mazanec, J.A. (1992): Classifying tourists into market segments: A neural network approach. *Journal of Travel & Tourism Marketing,* 1(1), 39-59.

McDonald, W.J. (1996). Internet customer segments: an international perspective. In Droge, C., & Calantone, R. (Eds.), *Enhancing knowledge development in marketing.* Chicago, IL: American Marketing Association, 338-344.

McLachlan, G.J. and Basford, K.E. (1988): *Mixture Models: Inference and Applications to Clustering.* New York: Marcel Dekker.

Møller, M. (1993): A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks,* 6(4), 525-533.

Murtagh, F. (1995): Interpreting the Kohonen self-organizing feature map using contiguity-constrained clustering. *Pattern Recognition Letters,* 16(4), 399-408.

Neal, R.M. (1996): Bayesian Learning from Neural Networks. New York: Springer-Verlag.

Neal, R.M. (1998): Assessing relevance determination methods using DELVE. In: Bishop CM (ed.). *Generalization in Neural Networks and Machine Learning.* New York: Springer-Verlag, 97-129

Norusis, M.J. (1990): *SPSS Advanced Statistics User's Guide.* Chicago IL: SPSS Inc.

Novak, T.P. and Hoffman D.L. (1997): Measuring the flow experience among web users. Presented at Interval Research Corporation, draft 1.0, (July 31)
URL: http://www2000.ogsm.vanderbilt.edu/papers/pdf/flow.pdf

O'Brien, S., and Ford, R. (1988): Can we at last say goodbye to social class? *Journal of the Market Research Society,* 30(3), 289-332.

O'Connor, G.C., and O'Keefe, B. (1997): Viewing the Web as a marketplace: the case of small companies. *Decision Support Systems,* 21(3), 171-183.

Penny, W.D, and Roberts, S.J. (1999): Bayesian neural networks for classification: how useful is the evidence framework? *Neural Networks,* 12, 877-892.

Pepermans, R., Verleye, G., and Van Cappellen, S. (1996): 'Wallbanking', innovativeness and computer attitudes: 25-40-year-old ATM-users on the spot. *Journal of Economic Psychology,* 17(6), 731-748.

Petersohn, H. (1998): Assessment of cluster analysis and self-organizing maps. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* 6(2), 139-149.

Piramuthu, S., Shaw, M.J. and Gentry, J.A. (1994): A Classification approach using multi-layered neural networks. *Decision Support Systems,* 11, 509-525.

Porteus, B.T. (1989): Stochastic inequalities relating to a class of log likelihood ratio statistics to their asymptotic χ^2 distribution. *Annals of Statistics,* 17, 1723-1734.

Refenes, A-P.N. (1994): Comments on 'Neural networks: Forecasting breakthrough or passing fad' by C. Chatfield. *International Journal of Forecasting,* 10, 43-46.

Rice, R.E., Grant, A.E., Schmitz, J. and Torobin, J. (1990): Individual and network influences on the adoption and perceived outcomes of electronic messaging. *Social Networks,* 12, 27-55.

Ripley, B. (1996): *Pattern recognition and neural networks.* Cambridge: Cambridge University Press.

Robins, G. (1993a): Neural networks - Automatic data analysis finds retail applications. *Stores,* January, 39-42.

Robins, G. (1993b): Credit scoring: Can retailers benefit from neural networks?. *Stores,* April, 34-35.

Rogers, J.D. (1998): GVU's 9th WWW user survey: Executive summary

URL: http://www.gvu.gatech.edu/user_surveys/survey-1998-04

Schaffer, C.M., and Green, P.E. (1998): Cluster-based market segmentation: some further comparisons of alternative approaches. *Journal of the Market Reseach Society,* 40(2), 155-163.

Scharl A. and Brandtweiner, R. (1998): A conceptual research framework for analyzing the evolution of electronic markets. *Electronic Markets Newsletter,* 8(2): 1-6.

Schocken, S. and Ariav, G. (1994): Neural networks for decision support: Problems and opportunities. *Decision Support Systems,* 11, 393-414.

Schwartz, E.I. (1997): What sells on the Web? *Microsoft Internet Magazine*, (April). URL: http://home.microsoft.com/reading/archives/

Serrano-Cinca, C. (1996): Self organizing neural networks for financial diagnosis. *Decision Support Systems.* 17, 227-238.

Setiono, R., Thong, J.Y.L. and Yap, C.S. (1998): Symbolic rule extraction from neural networks- an application to identifying organizations adopting IT. *Information & Management,* 34(2), 91-101.

Slater, D., Mulvenna, M., Büchner, A., and Moussy, L. (1999): Mining marketing intelligence from Internet retailing data: user requirements & business process description. In *Proceedings of the European Multimedia, Microprocessor Systems and Electronic Commerce (EMMSEC'99) Annual Conference*. Stockholm, Sweden.

Tam, K.Y. and Kiang, M.Y. (1992): Managerial applications of the neural networks: The case of bank failure predictions. *Management Science,* 38(7), 926-947.

Tarassenko, L. (1998): *A Guide to Neural Computing Applications*. London: Arnold.

Tipping, M. E. (2000): The relevance vector machine. In *Advances in Neural Information Processing Systems*, San Mateo, CA: Morgan Kaufmann.

Trevino, L.K., and Webster, J. (1992): Flow in computer-mediated communication. *Communication Research,* 19(5), 539-573.

Ultsch, A. (1993): Self-organizing neural networks for visualization and classification. In Opitz, O., Lausen, B., & Klar, R. (Eds.), *Information and classification. Concepts, methods and applications.* Berlin: Springer-Verlag, 307-313.

Vellido, A., Lisboa, P.J.G., and Vaughan, J. (1999a): Neural networks in business: a survey of applications (1992-1998). *Expert Systems with Applications* 17(1): 51-70.

Vellido, A., Lisboa, P.J.G. and Meehan, K. (1999b): Segmentation of the on-line shopping market using neural networks. *Expert Systems with Applications,* 17(4), 303-314.

Vesanto, J. (1999): SOM-based data visualization. *Intelligent Data Analysis,* 3(2), 111-126.

Wahlster, W. (2000): *Intelligent multimedia intelligent agents.* Tutorial, *MICAI-2000, Acapulco, Mexico.*

Wallin, E.O. (1999): Consumer personalization technologies for e-commerce on the Internet: a taxonomy. In: Roger, J-Y, Standford-Smith B, Kidd, PT (eds.). *Proceedings of the European Multimedia, Microprocessor Systems and Electronic Commerce (EMMSEC'99) Annual Conference.* IOS Press, Amsterdam.

Wedel, M. and Kamakura, W.A. (1998): *Market Segmentation. Conceptual and Methodological Foundations.* Massachusetts: Kluwer Academic Publishers.

Wedel, M. and Kistemaker, C. (1989): Consumer Benefit Segmentation Using Clusterwise Linear Regression. *International Journal of Research in Marketing*, 6, 45-49.

Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley & Sons.

Willems, T.M. and Brandts, L.E.M.W. (1997): Sub-symbolic management decision support system: an application of neural networks in management. *Production Planning & Control,* 8(2), 123-132.

Wind, Y. (1978): Issues and Advances in Segmentation Research. *Journal of Marketing Research*, 15(August), 317-337.

Zeleznikov, J., Strainieri, A. and Gawler, M. (1996): Project report: Split-Up - A legal expert system which determines property division upon divorce. *Artificial Intelligence and Law,* 3(4), 267-275.

Zhang, X. and Li, Y. (1993): Self-organizing map as a new method for clustering and data analysis. In *Proceedings of the International Joint Conference on Neural Networks* (IJCNN'93), Nagoya: Japan, 2448 - 2451.

Zwass, V. (1999): Structure and macro-level impacts of electronic commerce: From technological infrastructure to electronic marketplaces. Kendall, K.E. (ed.), *Emerging Information Technologies*. Thousand Oaks, CA: Sage Publications, 289-315.

World without end, remember me.

"Bright Red". Laurie Anderson