Essays on Finite Mixture Models

ISBN 978 90 3610 134 9

Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **458** of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

This thesis was prepared with LaTeX2e, most computations were done using Ox 4.00 (Doornik, 2007).

Essays on Finite Mixture Models

Essays over finite mixture modellen

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. S.W.J. Lamberts

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 2 juli 2009 om 13:30 uur

door

Abram van Dijk
geboren te Zwolle.

Promotiecommissie

Promotoren:	Prof.dr. R. Paap
	Prof.dr. Ph.H.B.F. Franses
Overige leden:	dr. D. Fok
	Prof.dr. P.J.F. Groenen
	Prof.dr. J.K. Vermunt

Acknowledgements

I spent eight years at the Erasmus University, first as a student and after that I stayed as a PhD candidate. It's been a good time in which I learned a lot, not just about econometrics. The result of this period is the book you now hold in your hands. Although it carries only my name on the cover, other people have contributed in one way or another.

First of all I thank my promotors Richard Paap and Philip Hans Franses. This book literally would not have been here if it was not for Richard. He pulled me into the academia, taught me how to do research, and made significant contributions to each chapter. During these years he knew the answer to every question I had. Sometimes I think I could have finished sooner if I had not waited as long as I did with asking Richard whenever I got stuck on something. Although Philip Hans was very busy, being director of the Econometric Institute and later dean of the whole faculty, he coauthored Chapter 4 and kept coming with new research ideas. If I could have done the research as fast as Philip Hans came up with new ideas, this book would have been twice as big.

I also collaborated with Dennis Fok, Dick van Dijk and Joost van Rosmalen. They coauthored chapters 3, 4 and 5 respectively. I would like to thank them for their contributions and the valuable lessons I learned from them about doing research and writing down the findings.

I would also like to thank my fellow PhD candidates for making the time that I worked at the university so enjoyable. They may not have contributed directly to my research (though Joost did), but the pleasant working days have been important for keeping me enthusiastic about my research. I enjoyed the discussions on all topics during the daily coffee breaks and extended lunches. I have especially good memories of driving along highway 61, or rather, being driven by Remco, watching *A History of Violence* and drinking whisky in Leuven, cycling through Vienna, Joost's Valentines Bash and the swiming pool of our hotel in Singapore. Thank you all for a wonderful time.

Last, but certainly not least, I want to thank my family for encouraging me throughout the years. Thank you Karin for your interest, even though many days I could not explain what exactly I had been doing. One last special 'thank you' goes to Quinten, just for being.

Bram van Dijk
April 2009

Contents

Chapter 1

Introduction and Summary

1.1 General introduction and motivation

Finite mixture distributions are a weighted average of a finite number of distributions. The latter are usually called the mixture components. The weights are usually described by a multinomial distribution and are sometimes called mixing proportions. The mixture components may be the same type of distributions with different parameter values but they may also be completely different distributions (Everitt and Hand, 1981; Titterington *et al.*, 1985). Therefore, finite mixture distributions are very flexible for modeling data. They are frequently used as a building block within many modern econometric models. The specification of the mixture distribution depends on the modeling problem at hand.

In this thesis, we introduce new applications of finite mixtures to deal with several different modeling issues. Each chapter of the thesis focusses on a specific modeling issue. The parameters of some of the resulting models can be estimated using standard techniques but for some of the chapters we need to develop new estimation and inference methods. To illustrate how the methods can be applied, we analyze at least one empirical data set for each approach. These data sets cover a wide range of research fields, such as macroeconomics, marketing, and political science. We show the usefulness of the methods and, in some cases, the improvement over previous methods in the literature.

An often applied finite mixture distribution is the mixture of normals. It is well known that this mixture is very useful to approximate many unbounded continuous distribution. Therefore, it can, for example, be used to describe fat-tailed distributions (see Kon, 1984). Moreover, this approximation property makes finite mixtures very useful for describing distributions of an entirely unknown form. Hence, there are very useful to capture the

often unknown distribution of individual-specific effects in a population (see, for example, Wedel *et al.*, 1999).

One of the best-known applications of finite mixtures within a model is the random effects model (Heckman, 1982). The random effects approach is aimed at capturing unobserved heterogeneity in panel data models due to missing explanatory variables. In the traditional situation, only the intercept is allowed to vary over individuals. The intercepts are assumed to follow from a (usually continuous) population distribution. Most random effects models use a continuous mixture with a normal mixing distribution. As the population distribution is not directly observed, one cannot be sure that the normal distribution is indeed the correct distribution of the unobserved heterogeneity. Heckman and Singer (1984) therefore argue that finite mixtures are to be preferred because of their flexibility to approximate many distributions, see Van Dijk (2006) for a recent exposition in duration models.

In some cases explanatory variables are missing at the individual level but are observed at some aggregate level. In Chapter 2 we show that a finite mixture model can be used to estimate the effect of the explanatory variables on the dependent variable although they are missing at the individual level. The proposed model for the dependent variable is a finite mixture model with a multivariate mixing distribution where the marginal mixing proportions are known.

Another use of finite mixture models is in describing difference in effects of covariates on the dependent variable. For example, individuals may differ with respect to the way they respond to price changes (see Fok *et al.*, 2008, for a survey of the relevant literature). The effect of the explanatory variables on the dependent variable varies across observations, but the modeler does not know the exact specification of this variation. We can describe this phenomenon by allowing for different parameter values across observations using a finite mixture approach. A more extreme case occurs when the differences in the relation between the dependent variable and the explanatory variables across observations cannot be captured by parameter variation alone. In this case we may even need two or more completely different statistical models to describe the differences in the relation between dependent and explanatory variables. This is often referred to as structural heterogeneity. This can be modeled by a finite mixture model, where each mixture component contains a different statistical model, see, for example, Yang and Allenby (2000). In Chapter 3 of this thesis we apply this approach to model differences in ranking capabilities of individuals.

One of the attractive features of finite mixtures is that we can also use them as a tool to segment observations. These segments can usually be given a meaningful interpretation (see, for example, Wedel and Kamakura, 2000). The mixing proportions are the estimates of the sizes of the segments. This feature of mixture models is used in Chapters 4 and 5. Chapter 4 deals with modeling regional house prices in the Netherlands. The mixture model provides a segmentation of the regions in the Netherlands with common house price dynamics. In general, segmentation using mixture models is done in only one dimension, for example segmentation of individuals or segmentation of regions. In Chapter 5 we show that mixture models can also be used for clustering in two dimensions.

In the next section we provide a more detailed outline of the chapters in the thesis.

1.2 Outline and summary

This thesis consists of four chapters. In each chapter we use a finite mixture distribution to deal with a specific modeling problem at hand. Each chapter is self-contained and can be read independently. Therefore, we finish each chapter with a separate conclusion. Below, we give a brief outline of the four chapters with a discussion of their contributions.

Chapter 2 is based on Van Dijk and Paap (2008). In this chapter, we consider the problem that empirical analysis of individual response behavior may be limited due to the lack of explanatory variables at the individual level. We put forward a new approach to estimate the effects of explanatory variables on individual response, where the explanatory variables are unknown at the individual level but observed at some aggregated level. This situation may, for example, occur when the response variable is available at the household level but explanatory variables only at the zip-code level.

We describe the missing individual explanatory variables by a latent variable model which matches the sample information at the aggregate level. The model for individual response is a finite mixture model with a multivariate mixing distribution where the marginal mixing proportions are known. Parameter estimates can be obtained using maximum likelihood or a Bayesian analysis. We illustrate the approach estimating the effects of household characteristics on donating behavior to a Dutch charity. Donating behavior is observed at the household level, while the explanatory variables are only observed at the zip-code level.

Chapter 3 is based on Van Dijk *et al.* (2007a). We consider the situation where one wants to estimate preferences of individuals over a discrete choice set through a survey.

In a traditional setup respondents are asked to select their most preferred option out of a (selected) set of alternatives. It is well known that, in theory, more information can be obtained if respondents are asked to rank the set of alternatives instead. In statistical terms, the preferences can then be estimated more efficiently. However, when individuals are unable to perform (part of) this ranking task, using the complete ranking may lead to a substantial bias in the parameter estimates and hence in the estimated preferences. Therefore, in practical situations one usually opts to use only a part of the reported ranking. The question is then of course which part of the ranking one should use?

To solve this question, we introduce in this chapter a latent-class rank-ordered logit model in which we use latent segments to endogenously identify the ranking capabilities of individuals. Each segment corresponds to a different assumption on the ranking capability. In fact, we have here a finite mixture model, where each component corresponds to a different model specification. Using a Monte Carlo experiment and an empirical application, we show that the new latent-class rank-ordered logit model provides more efficient estimates than a regular multinomial logit model in case at least some individuals are capable to rank more than one item. At the same time it does not suffer from biases due to ranking inabilities of some of the respondents.

Chapter 4 is based on Van Dijk *et al.* (2007b). We develop a panel model for regional house prices, for which both the cross-section and the time series dimension is large. The model allows for stochastic trends, cointegration, cross-equation correlations, and, most importantly, latent-class clustering of regions. Class membership is fully data-driven and based on the average growth rates of house prices, and the relationship of house prices with economic growth.

We apply the model to quarterly data for the Netherlands. The results suggest that there is convincing evidence for the existence of two distinct clusters of regions, with pronounced differences in house price dynamics.

Chapter 5 is based on Van Dijk *et al.* (2009). In this chapter we consider two-mode data. As in the previous chapters, we use finite mixtures to segment observations. The novelty of this chapter is that we jointly cluster two modes. We develop a new Bayesian approach to estimate the parameters of a latent-class model for the joint clustering of both modes of two-mode data matrices. That is, we cluster each mode into its own set of latent classes. Posterior results are obtained using a Gibbs sampler with data augmentation. Our Bayesian approach has three advantages over existing methods. First, in contrast to the frequentist estimation procedures we are able to do statistical inference on the model

parameters. Second, we are able to provide statistical criteria for determining the optimal numbers of clusters. Finally, our Gibbs sampler has fewer problems with local optima in the likelihood function and empty classes than the EM algorithm used in a frequentist approach.

We apply the Bayesian estimation method of the latent-class two-mode clustering model to two empirical data sets. The first data set is the Supreme Court voting data set of Doreian *et al.* (2004). The second data set comprises the roll call votes of the United States House of Representatives in 2007. For both data sets, we show how the results can provide useful insight into the data.

Chapter 2

Explaining Individual Response using Aggregated Data

2.1 Introduction

Empirical analysis of individual behavior is sometimes limited due to the lack of explanatory variables at the individual level. There may be various reasons why individual-level explanatory variables are not available. When using individual revealed preference data, information about explanatory variables may simply not be available as databases cannot be properly linked. For survey data, there may be a missing explanatory variable due to a missing question in the survey or a question which is interpreted the wrong way by the respondents.

In some cases it is possible to obtain information on explanatory variables at some aggregated level. For example, if the zip code of households is known, it may be possible to obtain aggregated information on household characteristics, like income and family size, at the zip-code level. This zip-code level information is usually obtained through surveys. The aggregated information of the variables is summarized in marginal probabilities which reflect the probability that the explanatory variable lies in some interval (income, age) or category (gender, religion) for a household in that zip-code region.

The goal of the current chapter is to estimate the effects of covariates on individual response when the covariates are unobserved at the individual level but observed at some aggregated level. There are several studies in economics which try to link individual and aggregated data, see, for example, Imbens and Lancaster (1994) and Van den Berg and Van der Klaauw (2001). In contrast to our situation, these studies assume that both

individual-level data and aggregated data is available. The aggregated data is assumed to be more reliable and is used to put restrictions on the individual-level data. The situation of missing individual covariates is related to ecological inference, see Wakefield (2004) for an overview. The main difference with regular ecological inference problems is that we observe individual responses, whereas ecological inference also relies on aggregated information on the response variable. The extra information on individual responses may help us overcome certain identification issues in ecological inference.

As far as we know, the only paper that comes close to our situation is Steenburgh *et al.* (2003). The motivation for the use of aggregated data in this paper is, however, different from ours. The authors use zip-code information to describe unobserved heterogeneity in the individual behavior of households instead of estimating the effects of covariates on behavior. Our problem also bears similarities with symbolic data analysis, see Billard and Diday (2003) for an overview. Symbolic data analysis also deals with aggregated explanatory variables and dependent variables at an individual level. The motivation for the use of aggregated data is however different. Aggregation is used to summarize large data sets. Therefore the form of the aggregated information is different and represents, for example, intervals instead of marginal probabilities.

In this chapter we develop a new approach to estimate the effects of covariates on individual response when the covariates are unknown at the individual level but observed at some aggregated level in the form of marginal probabilities. We extend the standard individual response model with a latent variable model describing the missing explanatory variables. This latent variable model describes the missing explanatory variables in such a way that it matches the sample information at the aggregated level. In case of one explanatory variable, the model simplifies to a standard mixture regression with known mixing proportions. A simple simulation experiment shows that this new approach outperforms in efficiency the standard method, where we replace the missing explanatory variables by the observed marginal probabilities at the aggregated level.

Parameter estimates of the individual response model can be obtained using Simulated Maximum Likelihood [SML] or a Bayesian approach. Given the computational burden of SML, the latter approach may be more convenient. To obtain posterior results, we use a Gibbs sampler with data augmentation. The unobserved explanatory variables are sampled alongside the model parameters. Conditional on the sampled explanatory variables, a standard Markov Chain Monte Carlo [MCMC] sampler can be used for the model describing individual response.

The outline of the chapter is as follows. In Section 2.2 we provide a simple introduction into the problem and perform a small simulation experiment to illustrates the merits of our approach. In Section 2.3 we generalize the discussion to a more general setting. Parameter estimation is discussed in Section 2.4. In Section 2.5 we illustrate our approach estimating the effects of household characteristics on donating behavior to a Dutch charity. We use aggregated information on household characteristics at the zip-code level to explain the individual response of households to a direct mailing by the charity. Finally, Section 2.6 concludes.

2.2 Preliminaries

To illustrate the benefits of our new approach, we start the discussion with a simple example. We consider a linear regression model with only one explanatory variable. The explanatory variable x_i can only take the value 0 or 1, for example, a gender dummy. Let the observed response of individual i, y_i, be described by

$$y_i = \alpha + \beta x_i + \varepsilon_i, \tag{2.1}$$

where α is an intercept parameter and where β describes the effect of the 0/1 dummy variables x_i on y_i for $i = 1, \ldots, N$. The error term ε_i is assumed to be normally distributed with mean 0 and variance σ^2. We assume that x_i is unobserved at the individual level but that we have aggregated information on x_i, for example, at the zip-code level. This aggregated information is summarized by $p_i = \Pr[x_i = 1]$ for $i = 1, \ldots, N$.

A simple approach to estimate β is to regress y_i on p_i instead of x_i. The error term of this regression equals

$$\eta_i = (x_i - p_i)\beta + \varepsilon_i. \tag{2.2}$$

The OLS estimator is consistent if $\mathrm{E}[p_i\eta_i|p_i] = 0$. As

$$\begin{aligned} \mathrm{E}[p_i\eta_i|p_i] &= \mathrm{E}\left[p_i \times ((x_i - p_i)\beta + \varepsilon_i)|p_i\right] = \mathrm{E}\left[p_i(x_i - p_i)\beta|p_i\right] + \mathrm{E}[p_i\varepsilon_i|p_i] \\ &= \mathrm{E}\left[p_i x_i \beta|p_i\right] - \mathrm{E}\left[p_i^2\beta|p_i\right] + \mathrm{E}[p_i\varepsilon_i|p_i] \end{aligned} \tag{2.3}$$

this condition is fulfilled if $\mathrm{E}[p_i\varepsilon_i|p_i] = 0$ and $\mathrm{E}[x_i|p_i] = p_i$. Although this OLS estimator is in general consistent, it is clear from (2.2) that the error term is heteroskedastic, and hence the OLS estimator is not efficient. Hence, a GLS estimator may improve upon OLS estimates.

An alternative approach to using the aggregated information to estimate β is to consider a mixture regression, see Quandt and Ramsey (1978), Everitt and Hand (1981) and Titterington *et al.* (1985). To describe the response variable y_i we consider a mixture of two linear regression models where in the first component the x_i variable is 1 and in the second component x_i equals 0. The mixing proportion is p_i which is known but may be different across individuals. Hence, the distribution of y_i is given by

$$y_i \sim \begin{cases} \mathrm{N}(\alpha+\beta, \sigma^2) & \text{with probability } p_i \\ \mathrm{N}(\alpha, \sigma^2) & \text{with probability } (1-p_i). \end{cases} \tag{2.4}$$

The parameters α and β can be estimated using maximum likelihood [ML]. ML estimates can easily be obtained using the EM algorithm of Dempster *et al.* (1977).

To illustrate the efficiency gain of the mixture approach we perform a simulation study. For $N = 1,000$ individuals we simulate 0/1 x_i values according to $\Pr[x_i = 1] = p_i$. We use different simulation schemes for p_i. We either allow the value of p_i to be different across individuals, or we impose that groups of individuals have the same value for p_i corresponding to the idea that these individuals live in the same zip-code region. Furthermore, we allow the range of possible values for p_i to be different. We sample p_i from $U(0.2, 0.4)$ or $U(0.01, 0.99)$. The values of y_i are generated according to $y_i = 1 + 2x_i + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, 1)$.

We estimate the β parameter using four approaches. In the first approach we estimate β using a linear regression model where we include the true x_i as explanatory variables. In practice this solution is of course not feasible but it allows us to compute the efficiency loss due to using explanatory variables at an aggregated level. In the second approach we consider an OLS estimator in a linear regression model with p_i as explanatory variable. The third approach uses a GLS estimator in the same linear regression model. The GLS weights are based on (2.3) and are computed using the true value of β and σ^2. In practice these parameters are of course unknown but the simulation results already show that accounting for heteroskedasticity using the true values does not compensate the efficient loss of the OLS estimator. In the last approach we consider the mixture solution as in (2.4).

Table 2.1 displays the efficiency loss in the estimator for β for the last three estimation approaches compared to using full information. Simulation results are based on 1,000 replications. The efficiency loss is computed using the root mean squared error [RMSE] of the estimates as all estimators are consistent. Several conclusions can be

Table 2.1: Efficiency loss of using aggregated data with respect to using full information for the three estimators

Distribution of p_i	Number of p_i[a]	Efficiency Loss OLS	GLS	Mixture
$U(0.20, 0.40)$	1,000	90.5%	90.5%	33.3%
$U(0.01, 0.99)$	1,000	50.4%	49.8%	24.1%
$U(0.20, 0.40)$	100	90.4%	90.4%	32.4%
$U(0.01, 0.99)$	100	52.5%	52.2%	23.0%
$U(0.20, 0.40)$	10	92.4%	92.3%	31.5%
$U(0.01, 0.99)$	10	62.4%	62.1%	23.0%
$U(0.20, 0.40)$	2	96.6%	96.6%	33.5%
$U(0.01, 0.99)$	2	73.9%	73.9%	31.3%

[a] Number of different p_i values drawn from the uniform distribution. Total number of individuals is 1,000.

drawn from the table. First of all, the mixture approach outperforms the other two estimators. Secondly, the GLS estimator hardly improves upon the OLS estimator, indicating that heteroskedasticity is not the main cause of the efficiency loss of the OLS estimator. Thirdly, all estimators perform better when the range in possible values of p_i is larger, which is not a surprise as a large variation in p_i provides more information about the slope parameter. Finally, the estimators perform better when there are less individuals with the same p_i value. The mixture approach however seems hardly affected by the number of individuals with the same value for p_i.

To illustrate the effects of the efficiency loss, we display in Figure 2.1 the density of $\hat{\beta}$ for the full information estimator, the OLS estimator and the mixture approach, where we use the simulation settings as in the second row of Table 2.1. The graph clearly illustrates the superiority of the mixture approach.

As already indicated by our simulation results, a GLS estimator does not compensate the efficiency loss due to aggregation of the explanatory variables. A second reason why the GLS estimator is not useful, is that constructing a feasible GLS estimator is often not possible if we have more than one explanatory variable. Consider, for example, the case

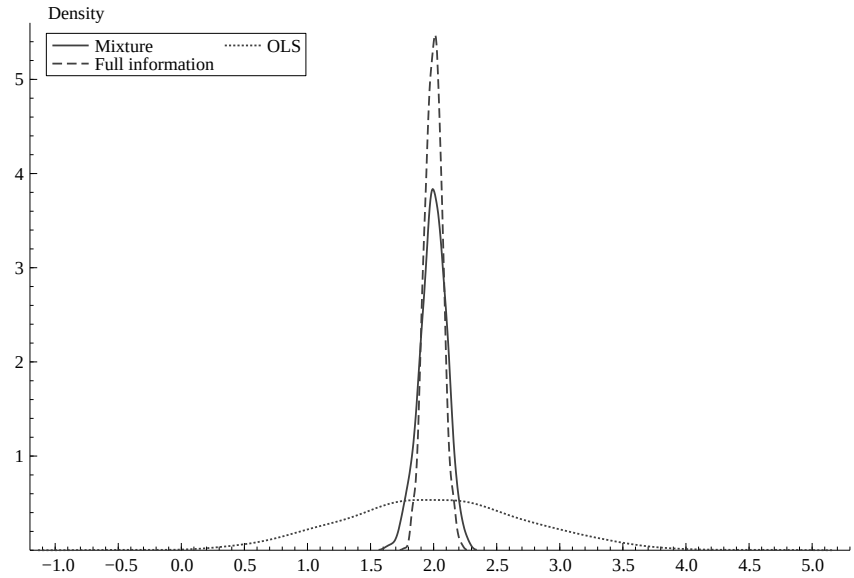


Figure 2.1: Density plots of the three estimators for β

with k explanatory variables which are unobserved at the individual level

$$y_i = \alpha + \sum_{j=1}^{k} \beta_j x_{i,j} + \varepsilon_i, \tag{2.5}$$

where $x_{i,j}$ are unobserved 0/1 dummy variables. Assume that we have aggregated information summarized in k marginal probabilities $\Pr[x_{i,j} = 1] = p_{i,j}$. It is straightforward to extend the proof above and show that the OLS estimator for β_j is consistent when the $x_{i,j}$ are replaced by $p_{i,j}$. In this case, the error term becomes

$$\eta_i = \sum_{j=1}^{k} (x_{i,j} - p_{i,j})\beta_j + \varepsilon_i. \tag{2.6}$$

Although the OLS estimator is consistent, it is impossible to estimate the variance of η_i, because the covariance matrix of x_i is unknown. As in practice we often only observe the marginal probabilities $\Pr[x_{i,j} = 1] = p_{i,j}$ and not the joint probabilities it is not feasible to estimate these covariances.

Before we turn to our solution to this problem, we first consider forecasting. Forecasting individual response when only aggregated explanatory variables are available is hampered for two reasons. First, the effects of the explanatory variables can be estimated less precisely compared to the case where individual data is available. The second reason

Table 2.2: Average RMSE of forecasting y_i using the parameter estimates of 4 methods and using either individual, aggregated or a Bayesian update of the individual data in the forecasts[a]

	using x_i	using p_i	using $\widehat{x_i \mid y_i}$
Full information	1.00	1.35	1.18
OLS	1.06	1.35	1.22
GLS	1.06	1.35	1.22
Mixture	1.00	1.35	1.18

[a] 1,000 out-of-sample observations.

is that the lack of out-of-sample explanatory variables at the individual level introduces more uncertainty in our forecast. To assess the out-of-sample forecasting performance of the four estimation methods, we simulate another set of 1,000 y_i values for each replication. We predict the value of y_i using the estimates of α and β obtained in the first part of the simulation for each of the four estimation procedures.

Table 2.2 displays the average RMSE for each of the four estimation procedures. We only show the results where we simulate p_i from $U(0.2, 0.4)$ and where we draw a distinct value for each individual. The other cases show similar results. We make a distinction between three situations. The second column displays the results when we assume that the out-of-sample x_i are known. In this case the full information approach and the mixture approach have similar average RMSE while the OLS and GLS approach perform worse. In case we only use aggregated out-of-sample information, all approaches perform the same, see third column of Table 2.2. Hence, the loss in forecast precision due to having aggregated out-of-sample information outweighs the efficiency loss in parameter estimation. The final column shows the results in case we only simulate new y_i values using the same x_i values of the original sample. This allows us to estimate the value of x_i given the in-sample information via Bayesian updating. Note that this is only possible in the case of a panel data set and time-invariant x_i variables. Again, the full information approach and the mixture approach have similar average RMSE while the OLS and GLS approach perform worse.

We can conclude from the simulation experiments in this section that the mixture approach is preferred when we want to estimate the effects of explanatory variables which

are only observed at the aggregated level on individual response. In the next section we extend the mixture approach to the situation where there are more than one explanatory variable. The information in the individual responses helps to estimate the unobserved correlations between the covariates.

2.3 Model specification

In this section, we generalize the discussion in the previous section in several ways. First, we relax the assumption that the model for y_i is a linear regression model. Secondly, we allow for m explanatory variables summarized in the m-dimensional vector X_i. Finally, we allow for other types of explanatory variables like ordered and unordered categorical variables. The vector of explanatory variables is written as $X_i = (X_i^{(1)\prime}, X_i^{(2)\prime}, X_i^{(3)\prime})'$, where $X_i^{(1)}$ contains the binary explanatory variables, $X_i^{(2)}$ the ordered categorical explanatory variables and $X_i^{(3)}$ the unordered explanatory variables.

We will use the general model specification

$$y_i = g(X_i\beta, \varepsilon_i), \tag{2.7}$$

where y_i is the observed dependent variable, β is an m-dimensional vector with the parameters of interest, ε_i is a random term, and g is some (non)linear function. The distribution of ε_i is known and depends on the unknown parameter vector θ. We assume that ε_i is independent of X_i.

This general model can be a linear regression model, but also a limited dependent variable model or any other nonlinear model. If the X_i variables are observed, parameter estimation is usually standard. In our case, the X_i variables are unobserved at the individual level but we know the marginal distribution of each X_i, which may or may not vary across individuals. To estimate the model parameters β and θ we extend (2.7) with a latent variable model describing the joint distribution of the X_i variables. Some of the parameters of this latent variable model are fixed to match the available sample information at the aggregated level. In the following subsections we describe the latent variable model for the three different types of explanatory variables. Note that we only discuss them separately to facilitate the exposition. The different types of variables can easily be combined in one multivariate model.

2.3.1 Binary explanatory variables

Assume that $X_i^{(1)}$ consists of k binary variables. The joint distribution of $X_i^{(1)}$ is discrete with 2^k mass points of which the associated probabilities sum up to 1. If we observe these 2^k mass points at some aggregated level, we can follow the mixture approach of Section 2.2 to estimate the β parameters. In practice, however, we typically observe the k marginal probabilities denoted by $P_i^{(1)} = (p_{i,1}^{(1)}, \ldots, p_{i,k}^{(1)})'$. Romeo (2005) proposes a method to estimate the joint discrete distribution from the marginal probabilities. He assumes that the joint distribution is known at an aggregated level. As we do not have this joint distribution at an aggregated level, his method is not suitable for our problem.

The k marginal probabilities plus the fact that probabilities sum up to 1 leave us with $2^k - (k+1)$ degrees of freedom on the 2^k mass points, unless we assume that the explanatory variables are independent. To facilitate modeling the joint distribution of $X_i^{(1)}$, we introduce a latent continuous random vector $X_i^{(1)*} = (x_{i,1}^{(1)*}, \ldots, x_{i,k}^{(1)*})'$ with

$$\begin{array}{lll} x_{i,j}^{(1)} = 1 & \text{if} & x_{i,j}^{(1)*} > 0 \\ x_{i,j}^{(1)} = 0 & \text{if} & x_{i,j}^{(1)*} \leq 0 \end{array} \tag{2.8}$$

for $i = 1, \ldots, N$ and $j = 1, \ldots, k$, see also Joe (1997) for a similar approach. A convenient distribution for $X_i^{(1)*}$ is a multivariate normal. The variance of $x_{i,j}^{(1)*}$ is set equal to 1 for identification. We impose that the mean of $x_{i,j}^{(1)*}$ equals $\Phi^{-1}(p_{i,j}^{(1)})$ for $j = 1, \ldots, k$ and $i = 1, \ldots, N$, where Φ denotes the distribution function of the standard normal distribution. It holds that $\Pr[x_{i,j}^{(1)} = 1] = \Pr[x_{i,j}^{(1)*} > 0] = \Phi(\Phi^{-1}(p_{i,j}^{(1)})) = p_{i,j}^{(1)}$, and hence these restrictions match the marginal distribution of the $X_i^{(1)}$ variables. In sum, we assume that

$$X_i^{(1)*} \sim \mathrm{N}\left(\Phi^{-1}(P_i^{(1)}), \Omega_{1,1}\right), \tag{2.9}$$

where $\Omega_{1,1}$ is a $k \times k$ positive definite symmetric matrix with ones on the diagonal. This leaves us with $\frac{1}{2}k(k-1)$ free parameters, that is, the sub-diagonal elements of $\Omega_{1,1}$. Although we lose some flexibility by assuming this structure, the correlation parameters do get an intuitive interpretation as they are related to correlations between the explanatory variables. The model for $X_i^{(1)}$ is in fact a multivariate probit [MVP] model, see Ashford and Sowden (1970), Amemiya (1974) and Chib and Greenberg (1998). The aggregated data provides the values of the intercepts such that only the sub-diagonal elements of $\Omega_{1,1}$ have to be estimated.

2.3.2 Ordered categorical explanatory variables

The setup for the binary variables can easily be extended to ordered categorical variables. If we have one ordered categorical variable with r categories, the $X_i^{(2)}$ vector in (2.7) contains $r-1$ 0/1 dummies, leaving one category, say the last one, as a reference category. Denote the $r-1$ dummies by $x_{i,1}^{(2)}, \ldots, x_{i,r-1}^{(2)}$. We typically observe the marginal distribution of the categories at some aggregated level which we denote by the r probabilities $P_i^{(2)} = (p_{i,1}^{(2)}, \ldots, p_{i,r}^{(2)})'$.

If we only have one ordered categorical explanatory variable in our model, we can use the simple mixture approach in Section 2.2 to estimate the effects of the r categories. In practice, we usually have a combination of several binary and ordered categorical variables and hence we need to deal with correlation between these variables. To describe correlations between several categorical variables, it is convenient to introduce a normal distributed random variable $x_i^{(2)*}$ and describe the distribution of the categorical variable in the following way

$$\begin{array}{rlll} x_{i,1}^{(2)} = 1 & \text{if} \quad x_i^{(2)*} \leq q_{i,1} & \text{and} \quad x_{i,1}^{(2)} = 0 & \text{otherwise} \\ x_{i,2}^{(2)} = 1 & \text{if} \quad q_{i,1} < x_i^{(2)*} \leq q_{i,2} & \text{and} \quad x_{i,2}^{(2)} = 0 & \text{otherwise} \\ \vdots & & & \\ x_{i,r-1}^{(2)} = 1 & \text{if} \quad q_{i,r-2} < x_i^{(2)*} \leq q_{i,r-1} & \text{and} \quad x_{i,r-1}^{(2)} = 0 & \text{otherwise.} \end{array} \tag{2.10}$$

For identification we impose that the variance of $x_i^{(2)*}$ is 1 such that

$$x_i^{(2)*} \sim \mathrm{N}(0,1)\,. \tag{2.11}$$

To match sample probabilities $P_i^{(2)}$, the limit points $q_{i,1} \ldots q_{i,r-1}$ are set equal to

$$q_{i,j} = \Phi^{-1}\left(\sum_{l=1}^{j} p_{i,l}^{(2)}\right),\ i = 1, \ldots, N,\ j = 1, \ldots, r-1. \tag{2.12}$$

The proposed model for $X_i^{(2)}$ is in fact the ordered probit model of Aitchison and Silvey (1957), see also Cowles (1996) for a Bayesian estimation procedure.

The equations (2.10)–(2.12) provide the latent variable model for the case of one ordered categorical explanatory variable. In case we have more categorical variables it is easy to extend the current solution with more latent $x_{i,j}^{(2)*}$ variables and allow them to correlate using a covariance matrix $\Omega_{2,2}$ with ones on the diagonal. It is also possible to correlate the resulting $X_i^{(2)*}$ variables with the latent random variables for the binary variables $X_i^{(1)^*}$ to describe correlations between binary and ordered categorical explanatory variables.

2.3.3 Unordered categorical explanatory variables

We may also encounter an explanatory variable which is categorical with, say, r categories, but without a natural ordering in the categories. We assume here that an individual can only belong to one category. If (s)he can belong to several categories we can apply the approach in Section 2.3.1. To model the effects of such a variable on y_i we include $r-1$ 0/1 dummy variables $x_{i,1}^{(3)},\ldots,x_{i,r-1}^{(3)}$ in $X_i^{(3)}$, leaving the rth category as reference. We observe the marginal probabilities of the r categories at some aggregate level which we denote by $P_i^{(3)}=(p_{i,1}^{(3)},\ldots,p_{i,r}^{(3)})'$.

To deal with the unordered categorical variable we build upon the multinomial probit [MNP] literature, see, for example, Hausman and Wise (1978) and Keane (1992). We introduce $r-1$ normally distributed variables $X_i^{(3)*}=(x_{i,1}^{(3)*},\ldots,x_{i,r-1}^{(3)*})'$ with

$$\begin{aligned} x_{i,1}^{(3)}=1 \quad &\text{if} \quad x_{i,1}^{(3)*}>\max(x_{i,2}^{(3)*},\ldots,x_{i,r-1}^{(3)*},0) \quad \text{and} \quad x_{i,1}^{(3)}=0 \quad \text{otherwise} \\ &\vdots \\ x_{i,r-1}^{(3)}=1 \quad &\text{if} \quad x_{i,r-1}^{(3)*}>\max(x_{i,1}^{(3)*},\ldots,x_{i,r-2}^{(3)*},0) \quad \text{and} \quad x_{i,r-1}^{(3)}=0 \quad \text{otherwise,} \end{aligned} \tag{2.13}$$

which means that $x_{i,1}^{(3)}=\ldots=x_{i,r-1}^{(3)}=0$ if $\max(x_{i,1}^{(3)*},\ldots,x_{i,r-1}^{(3)*})\leq 0$. Hence, the vector $X_i^{(3)*}$ correspond exactly to the utility differences in MNP models. The distribution of $X_i^{(3)*}$ is given by

$$\begin{pmatrix} x_{i,1}^{(3)*} \\ \vdots \\ x_{i,r-1}^{(3)*} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_{i,1}^{(3)*} \\ \vdots \\ \mu_{i,r-1}^{(3)*} \end{pmatrix}, \begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{2} \\ \frac{1}{2} & \cdots & \frac{1}{2} & 1 \end{pmatrix} \right), \tag{2.14}$$

where $\mu_i^{(3)*}=(\mu_{i,1}^{(3)*},\ldots,\mu_{i,r-1}^{(3)*})'$ represents the mean of $X_i^{(3)*}$. As individuals can only belong to one of the categories, we cannot identify the covariance matrix of $X_i^{(3)*}$ and have to fix its elements. For simplicity we take the implied covariance matrix of an MNP model where the individual utilities have a covariance equal to 1/2 times the identity matrix. If we take category r as the base category we end up with same covariance structure as above. The positive correlations are caused by the fact that the value of $x_{i,j}^{(3)*}$ is influenced by both $p_{i,j}^{(3)}$ and the probability of belonging to the reference category $p_{i,r}^{(3)}$. If the reference has a vary small probability, all $x_{i,j}^{(3)*}, j=1,\ldots,r-1$ should have a high value.

The observed probabilities imply $r-1$ restrictions on the distribution parameters of $X_i^{(3)*}$. To match the sample data with the model we have to solve $\mu_i^{(3)*}$ from

$$\begin{aligned} \Pr[x_{i,1}^{(3)*} > x_{i,2}^{(3)*}, \ldots, x_{i,1}^{(3)*} > x_{i,r-1}^{(3)*}, x_{i,1}^{(3)*} > 0] &= p_{i,1}^{(3)} \\ &\vdots \\ \Pr[x_{i,r-1}^{(3)*} > x_{i,1}^{(3)*}, \ldots, x_{i,r-1}^{(3)*} > x_{i,r-2}^{(3)*}, x_{i,r-1}^{(3)*} > 0] &= p_{i,r-1}^{(3)} \\ \Pr[x_{i,1}^{(3)*} \leq 0, \ldots, x_{i,r-1}^{(3)*} \leq 0] &= p_{i,r}^{(3)}. \end{aligned} \tag{2.15}$$

Note that the last restriction is automatically satisfied if the first $r-1$ restrictions hold. Unfortunately, there is no closed form expression for the probabilities from the LHS of (2.15) and hence we have to use numerical methods. If r is small, numerical integration techniques can be used to evaluate the probabilities. For larger values of r the probabilities can be evaluated using the Stern (1992) simulator or the Geweke-Hajivassiliou-Keane [GHK] simulator (Börsch-Supan and Hajivassiliou, 1993; Keane, 1994). The values of $\mu_i^{(3)*}$ can be found using a numerical solver. Notice that the values of $\mu_i^{(3)*}$ have to be determined only once before parameter estimation.

The equations (2.13) and (2.14) provide the latent variable model in case of one unordered categorical explanatory variable. In case there are more categorical variables it is easy to extend the current solution in a similar way as discussed before. It is also possible to correlate the $X_i^{(3)*}$ variables with the $X_i^{(1)*}$ and $X_i^{(2)*}$ variables in a straightforward manner.

2.3.4 Continuous explanatory variables

So far, we only used discrete explanatory variables. Dealing with the case where continuous variables are not observed at the individual level but at some aggregated level is not easy in practice. It is not enough to know the average value of the continuous variable at some aggregate level (e.g. the average value in each zip-code region) unless we make very strong assumptions. To deal with a continuous variable, we need to know the marginal distribution of the variable at the aggregated level. In case of a discrete variable, this distribution is represented by a few probabilities. In case of a continuous variable we need to know the type of distribution and the values of the parameters of the distribution. If the continuous variable is however divided in several intervals and we know the probability distribution over these intervals we can model it like an ordered categorical explanatory variables, see Section 2.3.2 and Section 2.5 for an example.

We summarize this section as follows. The explanatory variables X_i which are missing at the individual level are described by the latent variable $X_i^* = (X_i^{(1)*\prime}, X_i^{(2)*\prime}, X_i^{(3)*\prime})'$. This latent variable has a multivariate normal distribution. The mean of this distribution is determined by the marginal probabilities at the aggregate level. The covariance matrix of X_i^* is denoted by

$$\Omega = \begin{pmatrix} \Omega_{1,1} & \Omega'_{1,2} & \Omega'_{1,3} \\ \Omega_{1,2} & \Omega_{2,2} & \Omega'_{2,3} \\ \Omega_{1,3} & \Omega_{2,3} & \Omega_{3,3} \end{pmatrix}, \tag{2.16}$$

where the matrices $\Omega_{1,1}$ and $\Omega_{2,2}$ contain ones on the diagonal and $\Omega_{3,3}$ is equal to the covariance matrix given in (2.14) in case of just one unordered variable. If there are more unordered variables, $\Omega_{3,3}$ contains as many blocks of the covariance matrix from (2.14) on the diagonal. The remaining elements of Ω are free and describe the correlations between the latent variables X_i^*. We summarize the free elements of Ω in the vector ρ. The models for the X_i^* variables together with (2.7) provide the complete model specification.

2.4 Parameter estimation

To estimate the model parameters of the model proposed in the previous section, we can choose for maximum likelihood or a Bayesian approach. In this section we discuss both approaches and their relative merits.

We first derive the likelihood function. Let the density function of the data y_i for the model in (2.7) conditional on the missing variables X_i be given by

$$f(y_i|X_i;\beta,\theta), \tag{2.17}$$

where β and θ denote the model parameters. To derive the unconditional density of y_i we have to sum over all possible values of X_i, which we will denote by the set χ. Hence, the density of y_i given the observed marginal probabilities P_i is given by

$$f(y_i|P_i;\beta,\theta,\rho) = \sum_{X_i \in \chi} \Pr[X_i|P_i;\rho] f(y_i|X_i;\beta,\theta), \tag{2.18}$$

where $\Pr[X_i|P_i;\rho]$ denotes the probability of observing X_i given the data at the aggregated level which we denote by $P_i = (P_i^{(1)\prime}, P_i^{(2)\prime}, P_i^{(3)\prime})'$. These probabilities depend on the unknown parameter ρ which summarizes the free elements of the covariance matrix Ω as

discussed in the previous section. Hence, the log likelihood function is given by

$$\mathcal{L}(y|P;\beta,\theta,\rho) = \sum_{i=1}^{N} \log f(y_i|P_i;\beta,\theta,\rho), \tag{2.19}$$

where $y = (y_1,\ldots,y_N)'$ and $P = (P_1,\ldots,P_N)'$. The parameters β, θ and ρ have to be estimated from the data.

2.4.1 Maximum likelihood estimation

A maximum likelihood estimator can be obtained by maximizing the log likelihood function (2.19) with respect to (β,θ,ρ). To evaluate the log likelihood function we need to evaluate the probabilities $\Pr[X_i|P_i;\rho]$. Unfortunately, there is no closed form expression to compute these probabilities. For small dimensions it is possible to use numerical integration techniques but in general we have to use simulation methods to evaluate the probabilities. This implies that we end up with a Simulated Maximum Likelihood [SML] estimator, see Lerman and Manski (1981). The probabilities $\Pr[X_i|P_i;\rho]$ can be evaluated using the Stern (1992) simulator or the GHK simulator (Börsch-Supan and Hajivassiliou, 1993; Keane, 1994).

The SML estimator is only consistent if the number of observations and the number of simulations approach infinity. Given the literature on SML in MNP models (see for example, Geweke *et al.*, 1994), we expect that obtaining accurate values of the probabilities $\Pr[X_i|P_i;\rho]$ is computationally intensive, especially when the dimension of the latent X_i^* is large and/or the number of observations N is large. Note that the number of probabilities $\Pr[X_i|P_i;\rho]$ we need to evaluate grows exponentially with the number of variables in X_i.

2.4.2 Bayesian analysis

The model can also be analyzed in a Bayesian framework. To obtain posterior results for the model parameters, we propose a Gibbs sampler (Geman and Geman, 1984) with data augmentation, see Tanner and Wong (1987). The latent X_i^* variables are simulated along side the model parameters (β,θ,ρ). The main advantage of this Bayesian approach is that it does not require the evaluation of the complete likelihood function. If suffices to evaluate the likelihood function conditional on the latent X_i^* which determine X_i.

We focus in this section on the sampling of the latent variable X_i^*. We assume that if we know the X_i^* and hence the X_i variables, an MCMC sampling scheme to simulate

from the posterior distribution of the model parameters β and θ is available. Hence, we do not discuss simulating from the full conditional distribution of β and θ as this is model specific. We do however discuss simulating from the full conditional distribution of ρ as this is part of the model for the latent variable X_i^*.

Sampling of X_i^*

Because X_i is a deterministic function of X_i^* we only need to sample X_i^*. The full conditional density of X_i^* is given by

$$f(X_i^*|y_i, P_i; \beta, \theta, \rho) \propto f(y_i|X_i(X_i^*); \beta, \theta) f(X_i^*|P_i; \rho), \tag{2.20}$$

where $f(y_i|X_i(X_i^*); \beta, \theta)$ is given in (2.17) with $X_i(X_i^*)$ the deterministic mapping of X_i^* to X_i given in (2.8), (2.10) and (2.13). The function $f(X_i^*|P_i; \rho)$ denotes the density of X_i^* implied by the latent variable model for X_i^*. Given the structure of the latent variable model, X_i^* has a multivariate normal distribution with a mean μ_i which is determined by P_i and a covariance matrix Ω of which the free elements are denoted by ρ, that is,

$$f(X_i^*|P_i; \rho) = \phi(X_i^*; \mu_i(P_i), \Omega(\rho)), \tag{2.21}$$

where ϕ denotes the multivariate normal density function. Sampling the complete X_i^* vector at once is very difficult. Therefore, we sample the individual elements of X_i^* separately from their full conditional distribution. Let us consider the jth element of X_i^* denoted by $x_{i,j}^*$. The full conditional density of $x_{i,j}^*$ is given by

$$f(x_{i,j}^*|y_i, P_i, X_{i,-j}^*, \beta, \theta, \rho) \propto f(y_i|x_{i,j}(x_{i,j}^*), X_{i,-j}(X_{i,-j}^*); \beta, \theta) f(x_{i,j}^*|X_{i,-j}^*, P_i; \rho), \tag{2.22}$$

where $X_{i,-j}^*$ and $X_{i,-j}$ denote the vector X_i^* and X_i without $x_{i,j}^*$ and $x_{i,j}$, respectively.

The full conditional density of $x_{i,j}^*$ consists of two parts. We start with the second part, $f(x_{i,j}^*|X_{i,-j}^*, P_i; \rho)$, which is the conditional density of one of the elements of X_i^* which is of course a normal density with known mean, say, $\bar{\mu}_{i,j}$, and variance, say, $\bar{s}_j^2$, which are functions of $\mu_i(P_i)$ and $\Omega(\rho)$. The first part $f(y_i|x_{i,j}(x_{i,j}^*), X_{i,-j}(X_{i,-j}^*); \beta, \theta)$ is a function of $X_i(X_i^*)$ and can take a discrete number of values depending on the value of $x_{i,j}^*$.

In case $x_{i,j}^*$ corresponds to a binary explanatory variable, $x_{i,j}$ can take two values depending on whether $x_{i,j}^*$ is larger or smaller than 0. We can sample $x_{i,j}^*$ from in one step from its full conditional posterior distribution using the inverse CDF method but it is computationally more efficient to sample $x_{i,j}^*$ in two steps. In the first step, we determine

whether $x^*_{i,j}$ is larger or smaller than 0, that is whether $x_{i,j}$ is 1 or 0, respectively. From (2.22) it follows that

$$\Pr[x_{i,j}=1|y_i,P_i,X^*_{i,-j},\beta,\theta,\rho]=\frac{k_{i,j,1}\left(1-\Phi\left[\frac{-\bar{\mu}_{i,j}}{\bar{s}_j}\right]\right)}{c_{i,j,0}\Phi\left[\frac{-\bar{\mu}_{i,j}}{\bar{s}_j}\right]+c_{i,j,1}\left(1-\Phi\left[\frac{-\bar{\mu}_{ij}}{\bar{s}_j}\right]\right)}, \tag{2.23}$$

where $c_{i,j,0}=f(y_i|x_{i,j}=0,X_{i,-j}(X^*_{i,-j});\beta,\theta)$ and $c_{i,j,1}=f(y_i|x_{i,j}=1,X_{i,-j}(X^*_{i,-j});\beta,\theta)$.

In the second step, we sample $x^*_{i,j}|x_{i,j}$ from a truncated normal distribution with mean $\bar{\mu}_{i,j}$ and variance $\bar{s}^2_j$. We sample $x^*_{i,j}$ either positive or negative, depending on the whether $x_{i,j}$ is 1 or 0, respectively. For this truncated sampling we use the efficient accepting algorithm in Geweke (2005, pp. 113), see also Geweke (1991).

The other types of variables can be sampled in a similar manner. Appendix 2.A also provides the sampling schemes in case $x^*_{i,j}$ is associated with an ordered or an unordered categorical variable.

Sampling of ρ

To complete the Gibbs sampler, we need to sample the parameters in ρ from their full conditional posterior distribution. The vector ρ contains the free elements of the covariance matrix of X^*_i which is denoted by Ω, see (2.16). As discussed in Section 2.3, identification requires several restrictions on the covariance matrix Ω. In the first place, all diagonal elements of Ω are equal to 1 and hence Ω is a correlation matrix. Furthermore, the correlations between elements of the same unordered categorical variable are set equal to 1/2. Hence, the full conditional distribution of Ω is not an inverted Wishart distribution.

There exists several algorithms to sample a correlation matrix, see, for example, Chib and Greenberg (1998), Manchanda *et al.* (1999), and Liechty *et al.* (2004). In this chapter we follow Barnard *et al.* (2000). They suggest sampling one correlation at a time from their full conditional posterior distribution using a griddy-Gibbs sampler, see Ritter and Tanner (1992).

Suppose we want to draw the jth correlation in ρ denoted by ρ_j. Denote the vector ρ without ρ_j as ρ_{-j} and let $X^*=(X^*_1,\ldots,X^*_N)'$ and $\mu(P)=(\mu_1(P_1),\ldots,\mu_N(P_N))'$, where $\mu_i(P_i)$ denotes the mean of X^*_i determined by P_i for $i=1,\ldots,N$. The full conditional

posterior density of ρ_j is given by

$$\begin{aligned} f(\rho_j|y,P,X^*,\beta,\theta,\rho_{-j},) &\propto f(X^*|P,\rho_j,\rho_{-j})f(\rho_j|\rho_{-j}) \\ &\propto \prod_{i=1}^{N}\phi(X_i^*;\mu_i(P_i),\Omega(\rho_j,\rho_{-j}))f(\rho_j|\rho_{-j}) \\ &\propto |\Omega(\rho_j,\rho_{-j})|^{-\frac{N}{2}}\exp\left(-\frac{1}{2}(X^*-\mu(P))'\Omega(\rho_j,\rho_{-j})^{-1}(X^*-\mu(P))\right)f(\rho_j|\rho_{-j}), \end{aligned} \tag{2.24}$$

where $f(\rho_j|\rho_{-j})$ denotes the prior density of the jth element of ρ, conditional on all other elements of ρ. Barnard *et al.* (2000) show how to determine the range of values for which ρ_j leads to a positive definite matrix. Within this range we can define a set of grid points to evaluate the kernel (2.24) for the griddy-Gibbs sampler.

As correlations in Ω which are related to the jth explanatory variable are not identified if $\beta_j = 0$, we suggest to impose an informative prior for the parameters in ρ. We use a truncated normal prior with variance ω^2 for the parameters in ρ, that is,

$$f(\rho) \propto I[\Omega(\rho) = \text{PD}]\prod_j \exp(-\rho_j^2/(2\omega^2)), \tag{2.25}$$

where $I[\Omega(\rho) = \text{PD}]$ is an indicator function which is 1 if $\Omega(\rho)$ is positive definite and 0 otherwise. Hence, we concentrate the probability mass around zero.

2.5 Application

To illustrate our approach, we consider in this section an application where we analyze the characteristics of households who donate to a large Dutch charity in the health sector. Households receive a direct mailing from the charity with a request to donate money. The household may not respond and donate nothing or respond and donate a positive amount. We have no information about the characteristics of the households apart from their zip code. At the zip-code level we know aggregated household characteristics.

Our sample contains 10,000 households which are randomly selected from the database. The mailing we consider took place in February 1997. The response rate is 39.0%. The average donation is 3.39 euros and the average donation conditional on response is 8.68 euros. We match these data with aggregated data at the zip-code level (4 digits) from Statistics Netherlands (CBS). Table 2.3 shows the relevant aggregated data at the zip-code level. As can be observed from the table we have aggregated data for different types

Table 2.3: Available explanatory variables at the zip-code level

Variable	Type	Description
Church	Binary	Goes to church every week
Not-active	Binary	Not active in labor force
Reference: Family with kids		
Single	Unordered (3 cat.)	Lives alone
Family no kids	Unordered (3 cat.)	Family without kids
Reference: Average income		
Income low	Ordered (3 cat.)	Income in lowest 40% nationally
Income high	Ordered (3 cat.)	Income in highest 20% nationally
Reference: Age between 25 and 44		
Age 0-24	Ordered (4 cat.)	Age between 0 and 24
Age 45-64	Ordered (4 cat.)	Age between 45 and 64
Age 65+	Ordered (4 cat.)	Age over 65
Urbanization	Observed	Measure for the degree of urbanization

of explanatory variables, that is, for binary, unordered, and ordered categorical variables. Note that we only know *urbanization* at the zip-code level. As it is the same for each individual in the zip-code region, this variable is treated as an observed variable.

To describe donating behavior we consider a censored regression (Tobin, 1958) because the donated amount is censored at 0. We use the log of (1 + amount) as dependent variable which leads to the following model specification

$$\log(1+y_i) = \begin{cases} x_i'\beta + \varepsilon_i & \text{if} \quad x_i'\beta + \varepsilon_i > 0 \\ 0 & \text{if} \quad x_i'\beta + \varepsilon_i \leq 0, \end{cases} \tag{2.26}$$

with $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. As explanatory variables we take the variables displayed in Table 2.3.

To estimate the effects of the covariates on response, we use two approaches. First, we follow the simple regression approach of Section 2.2, which means that we replace the unknown household characteristics by their sample averages at the zip-code level. The parameters of (2.26) are estimated using ML. Although we have only shown in Section 2.2 that OLS in a linear regression model provides consistent estimates, simulations suggest that this result carries over to the ML estimator in a censored regression model. Secondly,

Table 2.4: Posterior means, posterior standard deviations, and HPD regions of the model parameters for the mixture approach together with ML results for the simple approach

	mixture approach			simple approach	
	mean	s.d.	95% HPD	ML	s.e.[a]
Intercept	-1.77	0.05	(-1.87,-1.68)	-2.89	0.81
Urbanization	0.05	0.03	(-0.01, 0.11)	0.63	0.35
Church	0.64	0.02	(0.59, 0.69)	0.24	0.25
Not-active	-0.72	0.01	(-0.75,-0.69)	-1.03	0.74
Single	3.63	0.04	(3.55, 3.71)	0.52	0.55
Family no kids	3.61	0.04	(3.53, 3.70)	3.51	1.20
Income low	-0.01	0.03	(-0.07, 0.05)	0.41	1.17
Income high	0.71	0.02	(0.68, 0.75)	1.60	0.87
Age 0-24	-0.01	0.03	(-0.07, 0.06)	2.96	1.35
Age 45-65	-3.60	0.09	(-3.77,-3.42)	-1.42	1.28
Age 65+	0.72	0.02	(0.68, 0.75)	1.80	1.05
σ	0.17	0.00	(0.17, 0.18)	2.36	0.02

[a] Heteroskedastic-consistent standard errors, see White (1982).

we use the mixture approach to estimate the censored regression parameters, where we opt for a Bayesian approach. For ρ we take the informative prior (2.25) with $\omega^2 = 1/4$. For β we use a normal prior with mean 0 and standard deviation 0.5 and for σ^2 we use an Inverted Gamma-2 prior with parameters 12.5 and 50. These priors help to obtain a smoother convergence of the MCMC sampler. Posterior results turn out not to be sensitive to moderate changes of this prior specification.

We use a total of $120,000$ draws, which took about six hours on an Pentium 4, 2.8 Ghz processor. The first $20,000$ draws were used as burn-in period. Furthermore, we only used every 20th draw to obtain a random sample of $5,000$ draws. Our code is tested using the approach of Geweke (2004).

Table 2.4 displays the estimation results for both approaches. It is clear from the table that the posterior standard deviations of the mixture approach are much smaller than the standard errors of the ML estimator, where the unknown household characteristics are replaced by their sample averages at the zip-code level. Although the number of observations is very large, the estimated standard errors of the simple approach are still

Table 2.5: Posterior means of the correlations between unobserved variables with posterior standard deviations in parentheses

	Church	Not-active	Single	Family no kids	Income	Age
Church	1 (-)					
Not-active	0.19 (0.10)	1 (-)				
Single	-0.34 (0.10)	-0.14 (0.14)	1 (-)			
Family no kids	-0.72 (0.09)	0.19 (0.08)	0.50 (-)	1 (-)		
Income	0.06 (0.09)	-0.54 (0.16)	0.33 (0.10)	-0.27 (0.10)	1 (-)	
Age	0.33 (0.10)	0.27 (0.06)	0.29 (0.08)	-0.19 (0.06)	-0.17 (0.06)	1 (-)

substantial. This illustrates the huge efficiency gain of using our method. This efficiency gain enables us to identify more significant influences from household characteristics. When using the simple approach, only *Family no kids* and *Age 0-24* are identified as having a significant impact on the donating behavior. But, using our mixture approach it becomes clear that many other household characteristics also influence this decision. Being Religious has a positive effect, while not being active in the labor force has a negative effect. Both single households and families without children tend to donate more. Household with higher income tend to donate more, while the effect of age is nonlinear. The highest posterior density [HPD] interval shows that urbanization grade has no influence on donating behavior.

There are two differences in the results of the two methods. First, both find that families without children donate more than families with children, however, the ML results suggest that single households donate about the same as families with children while the mixture approach suggest that their donating behavior is more comparable to families without children. The second difference concerns the effect of age. The main difference is the level of the reference category, as all parameters have a higher value in the ML estimates. Moreover, according to the ML results individuals younger than 25 are the

most lucrative group, while the mixture approach suggests that this group consists of individuals over 65.

Table 2.5 displays the estimated correlation matrix of X_i^*. The diagonal elements are fixed for identification. The correlation between the variables *Single* and *Family no kids* is fixed at 1/2, because they belong to the same unordered categorical variable. Many of the posterior means of the correlations are more than two times larger than their posterior standard deviation, illustrating the importance of our approach. In general, the correlations have the expected sign. For example, there is a negative correlation between being not active in the labor force and income, and a positive correlation between being religious and age.

2.6 Conclusions

In this chapter we have developed a new approach to estimate the effects of explanatory variables on individual response where the response variable is observed at the individual level but the explanatory variables are only observed at some aggregated level. This approach can, for example, be used if information about individual characteristics is only available at the zip-code level. To solve the limited data availability, we extend the model describing individual responses with a latent variable model to describe the missing individual explanatory variables. The latent variable model is of the probit type and matches the sample information of the explanatory variables at the aggregated level. Parameter estimates for the effects of the explanatory variables in the individual response model can be obtained using maximum likelihood or a Bayesian approach.

A simulation study shows that our new approach clearly outperforms a standard approach in efficiency. The efficiency loss which is due to aggregation is about 50% smaller than for the standard method. We illustrated the merits of our approach by estimating the effects of the household characteristics on donating behavior to a Dutch charity. For this application we used data of donating behavior at the household level, while the covariates were only observed at the zip-code level.

There are several areas for future research. It may be interesting to investigate whether the proposed method can be used to deal with nonresponse in survey data. Another topic for future research is to consider the complement case where explanatory variables are observed at the individual level but that the response variable is only observed at some aggregated level.

2.A Derivation of full conditional distributions

In this appendix we provide the simulation schemes for missing ordered categorical variables and unordered categorical variables. As starting point we take the general form of the full conditional density of $x^*_{i,j}$ given in (2.22).

2.A.1 Ordered categorical variable

For an ordered categorical variable with r categories, we have to sample the variable $x^{(2)*}_{i,j}$. If we assume that r is the reference category, the $x^{(2)*}_{i,j}$ variable determines the $r-1$ 0/1 dummy variables $x^{(2)}_{i,1},\ldots,x^{(2)}_{i,r-1}$. Let $P^{(2)}_i = (p^{(2)}_{i,1},\ldots,p^{(2)}_{i,r})'$ denote the observed marginal probabilities that the individual belongs to the r categories. The threshold levels $q_{i,t}$ are set equal to $\Phi^{-1}(\sum_{l=1}^{t} p^{(2)}_{i,l})$ for $i = 1,\ldots,N$ and $t = 1,\ldots,r-1$. Let $\bar{\mu}_{i,j}$ denote the mean of $x^{(2)*}_{i,j}|X^*_{i,-j},\rho$ in the latent model and let $\bar{s}^2_j$ denote the variance of $x^{(2)*}_{i,j}|X^*_{i,-j},\rho$, where $X^*_{i,-j}$ denotes X^*_i without $x^{(2)*}_{i,j}$.

Sampling of $x^{(2)*}_{i,j}$ proceeds in the same way as for the binary variables except for the fact that there are now r possible values for $f(y_i|X_i;\beta,\theta)$ instead of only 2, that is,

$$\begin{aligned}
c_{i,j,1} &= f(y_i|X_{i,-j}(X^*_{i,-j}), x^{(2)}_{i,1}=1, x^{(2)}_{i,2}=0,\ldots,x^{(2)}_{i,r-1}=0;\beta,\theta)\\
&\vdots\\
c_{i,j,r-1} &= f(y_i|X_{i,-j}(X^*_{i,-j}), x^{(2)}_{i,1}=0,\ldots,x^{(2)}_{i,r-2}=0, x^{(2)}_{i,r-1}=1;\beta,\theta)\\
c_{i,j,r} &= f(y_i|X_{i,-j}(X^*_{i,-j}), x^{(2)}_{i,1}=0,\ldots,x^{(2)}_{i,r-1}=0;\beta,\theta),
\end{aligned}$$

where $X_{i,-j}$ denotes X_i without $x^{(2)}_{i,1},\ldots,x^{(2)}_{i,r-1}$.

First we draw $x^{(2)}_{i,1},\ldots,x^{(2)}_{i,r-1}$ using the fact that they are mutually exclusive and

$$\Pr[x^{(2)}_{i,t}=1, x^{(2)}_{i,-t}=0|y_i,P_i,X_{i,-j}(X^*_{i,-j}),\beta,\theta,\rho] = \frac{c_{i,j,t}(\bar{p}_{i,t}-\bar{p}_{i,t-1})}{\sum_{l=1}^{r} c_{i,j,l}(\bar{p}_{i,l}-\bar{p}_{i,l-1})} \tag{2.27}$$

for $t = 1,\ldots,r-1$, where $\bar{p}_{i,t} = \Phi\left(\frac{q_{i,t}-\bar{\mu}_{i,j}}{\bar{s}_j}\right)$ with $\bar{p}_{i,0}=0$ and $\bar{p}_{i,r}=1$, and where $x^{(2)}_{i,-t}$ denotes $x^{(2)}_{i,1},\ldots,x^{(2)}_{i,r-1}$ without $x^{(2)}_{i,t}$.

Next, we sample $x^{(2)*}_{i,j}|x^{(2)}_{i,1},\ldots,x^{(2)}_{i,r-1}$ from a truncated normal distribution with mean $\bar{\mu}_{i,j}$ and variance $\bar{s}^2_j$. If $x^{(2)}_{i,t}=1$ we sample $x^{(2)*}_{i,j}$ between $q_{i,t-1}$ and $q_{i,t}$ for $t=1,\ldots,r-1$, where $q_{i,0}=-\infty$. If $x^{(2)}_{i,1}=\ldots=x^{(2)}_{i,r-1}=0$, we sample $x^{(2)*}_{i,j}$ larger than $q_{i,r-1}$. We again use the acceptance sampling algorithm of Geweke (2005, pp. 113).

2.A.2 Unordered categorical variable

For an unordered categorical variable with r categories, we add $r-1$ 0/1 dummies in $X_i^{(3)}$, say, $x_{i,1}^{(3)},\ldots,x_{i,r-1}^{(3)}$. The $r-1$ normally distributed random variables which belong to this unordered categorical variable are denoted by $(x_{i,1}^{(3)*},\ldots,x_{i,r-1}^{(3)*})'$.

Suppose that we want to sample $x_{i,j}^{(3)*}$ from its full conditional posterior distribution. The full conditional posterior density is given by (2.22). The second part is given by $f(x_{i,j}^{(3)*}|X_{i,-j}^*,P_i;\rho)$, where $X_{i,-j}^*$ denotes X_i^* without $x_{i,j}^{(3)*}$. This is a normal density with known mean, say $\bar{\mu}_{i,j}$, and variance, say $\bar{s}_j^2$. The first part is given by $f(y_i|x_{i,j},X_{i,-j}(X_{i,-j}^*);\beta,\theta)$. This can only take two values, conditional on $X_{i,-j}^*$. Define $x_{i,l}^{(3)*}=\max(x_{i,1}^{(3)*},\ldots,x_{i,j-1}^{(3)*},x_{i,j+1}^{(3)*},\ldots,x_{i,r-1}^{(3)*})$. The two possible values are given by

$$\begin{aligned}c_{i,j,0} &= f(y_i|x_{i,1}^{(3)}=0,\ldots,x_{i,l-1}^{(3)}=0,x_{i,l}^{(3)}=1,x_{i,l+1}^{(3)}=0,\ldots x_{i,r-1}^{(3)}=0,X_{i,-j}(X_{i,-j}^*);\beta,\theta)\times\\ &\qquad I[x_{i,l}^{(3)*}>0]+f(y_i|x_{i,1}=0,\ldots,x_{i,r-1}=0,X_{i,-j}(X_{i,-j}^*);\beta,\theta)I[x_{i,l}^{(3)*}\leq 0]\\ c_{i,j,1} &= f(y_i|x_{i,1}^{(3)}=0,\ldots,x_{i,j-1}^{(3)}=0,x_{i,j}^{(3)}=1,x_{i,j+1}^{(3)}=0,\ldots x_{i,r-1}^{(3)}=0,X_{i,-j}(X_{i,-j}^*);\beta,\theta),\end{aligned}$$

where $X_{i,-j}$ denotes X_i without $x_{i,1}^{(3)},\ldots,x_{i,r-1}^{(3)}$.

In the cases of a binary or ordered categorical variable the distribution of $x_{i,j}^*|X_i$ is a univariate truncated normal. However, for an unordered variable this is not the case anymore, as its distribution also depends on the value of $x_{i,l}^{(3)*}$. The sampling of $x_{i,1}^{(3)*},\ldots,x_{i,r-1}^{(3)*}$ given $x_{i,1}^{(3)},\ldots,x_{i,r-1}^{(3)}$ becomes non-standard. Therefore, it is more efficient to use the inverse CDF method to draw from $x_{i,j}^{(3)*}$ and $x_{i,j}^{(3)}$ simultaneously.

The full conditional posterior density of $x_{i,j}^{(3)*}$ is given by

$$\begin{aligned}d_{i,j}(c_{i,j,0}\phi(x_{i,j}^{(3)*};\bar{\mu}_{i,j},\bar{s}_j)I[x_{i,j}^{(3)*}\leq\max(x_{i,l}^{(3)*},0)]+\\ k_{i,j,1}\phi(x_{i,j}^{(3)*};\bar{\mu}_{i,j},\bar{s}_j)I[x_{i,j}^{(3)*}>\max(x_{i,l}^{(3)*},0)]),\end{aligned}\quad(2.28)$$

where the integrating constant $d_{i,j}$ equals

$$d_{i,j}=\left((c_{i,j,0}-c_{i,j,1})\Phi\left(\frac{\max(x_{i,l}^{(3)*},0)-\bar{\mu}_{i,j}}{\bar{s}_j}\right)+c_{i,j,1}\right)^{-1}.\quad(2.29)$$

Straightforward derivation leads to the following inverse CDF

$$x_{i,j}^{(3)*}(u)=\begin{cases}\Phi^{-1}\left(\frac{u}{d_{i,j}c_{i,j,0}}\right)\bar{s}_j+\bar{\mu}_{i,j} & \text{if } u\leq\bar{u}\\ \Phi^{-1}\left(\frac{u}{d_{i,j}c_{i,j,1}}+\frac{c_{i,j,1}-c_{i,j,0}}{c_{i,j,1}}\Phi\left(\frac{\max(x_{i,l}^{(3)*},0)-\bar{\mu}_{i,j}}{\bar{s}_j}\right)\right)\bar{s}_j+\bar{\mu}_{i,j} & \text{if } u>\bar{u},\end{cases}\quad(2.30)$$

where $\bar{u}=d_{i,j}c_{i,j,0}\Phi\left(\frac{\max(x_{i,l}^{(3)*},0)-\bar{\mu}_{i,j}}{\bar{s}_j}\right)$.

Chapter 3

A Rank-Ordered Logit Model with Unobserved Heterogeneity in Ranking Capabilities

3.1 Introduction

To determine preferences of individuals, researchers often rely on surveys. In the traditional setup, a survey is created in which each respondent is only asked to select the most preferred option out of a set of presented alternatives. To estimate the preferences based on such a survey one can use a standard discrete choice model, like the multinomial logit [MNL] model. It is well known that more information can be obtained from a respondent if (s)he is asked to give a complete ranking of *all* presented alternatives. The rank-ordered logit [ROL] model has become the standard tool to analyze the preferences in case rank data is available.

The ROL model was introduced in the literature by Beggs *et al.* (1981). The model can be used to analyze the preferences of individuals over a set of alternatives, where the preferences are partially observed through surveys or conjoint studies. Empirical applications describing preferences using the ROL model can be found in many fields such as voter preferences (Koop and Poirier, 1994), aging studies (Hsieh, 2005), marketing (Ahn *et al.*, 2006; Dagsvik and Liu, 2006), empirical labor economics (Van Beek *et al.*, 1997), school choice (Mark *et al.*, 2004; Drewes and Micheal, 2006), demand for classical music (Van Ophem *et al.*, 1999) and transportation studies (Kockelman *et al.*, 2008; Calfee *et al.*, 2001).

In theory, when individuals are asked to rank the alternatives instead of only choosing the most preferred option, the parameters of the choice model and hence the preferences can be estimated more efficiently. However, in practice respondents may be unable to perform (part of) the ranking task. This may be due to several reasons. First of all, respondents may actually not be able to perform the task itself. In some cases there may be too many alternatives to rank. Secondly, the respondent may not be able to distinguish between his/her less-preferred alternatives. In any case, straightforwardly using reported rankings may lead to a substantial bias in the parameter estimates in the ROL model, as Chapman and Staelin (1982) have argued, and Hausman and Ruud (1987) have shown in practice. To solve this issue, Chapman and Staelin (1982) suggest to only use the first few ranks in the estimation. They consider several rules to determine the appropriate number of ranks to use, in their words "the explosion depth". One of these rules is based on a pooling test for the equality of parameter estimates based on different ranks. Hausman and Ruud (1987) provide a detailed analysis of the ROL model and some specification tests. Ultimately, they still only use the first few ranks in the estimation, and provide an alternative method to test for the number of ranks to use in the estimation. Note that in both approaches this number is assumed to be the same for all respondents. If ranking capabilities differ across individuals, this may lead to an efficiency loss.

Hausman and Ruud (1987) also consider a model where each rank in the estimation receives a weight. This makes it possible that the most preferred rank contain more information than lower ranks. These weights are estimated alongside the model parameters. But again, these weights are the same for each individual, and, moreover, if some respondents enter the lower ranks in a random way, estimates using this method will still be biased.

To solve the ranking inability issue one may considering extending the choice model with a consideration set stage, see, for example, Chiang *et al.* (1999) and Bronnenberg and Vanhonacker (1996). The idea behind this approach is that a choice decision involves two steps. In the consideration stage the individual first limits the number of choice alternatives by deleting alternatives he or she does not consider. In the second stage the individual chooses from (or in our case ranks) the alternatives considered. This approach however does not solve the ranking inability issue. Introduction of a consideration stage implies that one assumes that individuals cannot rank alternatives outside the consideration set which may not be true. Furthermore, in surveys individuals are usually forced to

consider all alternatives presented to them. Hence, the consideration set approach does not seem to be appropriate.

In this chapter we propose a model in which we use all observed rankings while taking into account the fact that the rankings may not completely reflect the true preferences. To this end we introduce a ROL model which uses latent segments to endogenously identify the unobserved ranking capabilities of respondents. Each segment corresponds to a particular assumption on the unobserved ranking capability. Hence, in contrast to the existing literature, we allow for individual-specific ranking capabilities. We show that in case at least some individuals are able to rank, our model results in a clear efficiency gain relative to a standard MNL model. At the same time it does not suffer from biases due to ranking inabilities of some of the individuals. We refer to our model as the latent-class rank-ordered logit [LCROL] model.

Apart from the efficiency gain, the new model also allows us to learn about the unobserved ranking capabilities of respondents. On itself, this information can also be very valuable. For example, one can use this information to construct more efficient ranking tasks in surveys and for deciding the number of respondents necessary to reach a preferred precision. To investigate the presence of respondents with a particular type of ranking capability, we propose to use a likelihood ratio test. We can, for example, test whether some individuals are able to rank all alternatives. Furthermore, the statistical test can be helpful to remove redundant segments from our model, which may lead to efficiency loss if present.

This study is not the first attempt to make maximal use of the information content on preferences from individual ranked ordered data. Van Ophem *et al.* (1999) designed a multichoice logit [MCL] model to use more information, and hence to be more efficient than the MNL model, while avoiding the bias that a ROL model can have due to ranking inability of some respondents. In their setup, respondents need to sequentially divide the items into three preference groups where an item in a higher group is always preferred to an item in a lower group. Finally, the items in the top-ranked group were completely ranked according to preferences. The final ranking can be seen as an extreme case of dividing the items across groups, where each group just contains one item. The MCL model describes the preference ordering of these groups. By not considering the preferences within all groups, they hope to avoid biases. However, it is still assumed that ranking capabilities are the same across individuals.

Our solution to deal with differences in ranking capabilities also bears some similarities with the misreporting literature. If an individual is not able to rank properly, his/her answer may not be according to his/her preferences. In other words in this case there may be misreporting, or a classification error. In a non-linear model, measurement errors in the dependent variables in general result in biased estimates, see Hausman (2001) for an overview. Many studies have considered specifications to account for different kinds of misclassification, see, for example, Chua and Fuller (1987), Poterba and Summers (1995), Hausman *et al.* (1998), and Abrevaya and Hausman (1999). In our model we account for a more general form of misclassifications, some ranks are correct whereas other ranks are "misclassified".

The remainder of this chapter is set up as follows. In Section 3.2 we introduce the LCROL model which incorporates the unobserved ranking capabilities of individuals. To illustrate the merits of our model specification we perform a simulation study in Section 3.3. For the case that not all respondents are able to provide a full ranking, simulations show an estimation bias in the standard ROL model. For the general case that some individuals are not able to give a complete ranking and some individuals are able to give a partial ranking, estimates resulting from our LCROL model show no bias and are more efficient than MNL estimates. Section 3.4 provides an application of the LCROL model. We apply the model to data obtained from a small survey among 91 students, who are asked to rank six different platforms for computer games. The estimation results show that the LCROL is a useful tool for analyzing preference data. The parameter estimates from the LCROL model have the smallest standard errors, even in this small sample situation. Furthermore, the resulting segmentation of the students is plausible. Finally, in Section 3.5 we conclude and give topics for further research.

3.2 Model specification

In this section we develop our approach to capture unobserved heterogeneity in the ranking ability of respondents. In Section 3.2.1 we start the discussion with the standard rank-ordered logit model and in the following sections we will extend this model. Throughout this chapter, we consider the simple case that each respondent is confronted with a fixed set of alternatives. The respondent either chooses the most preferred item from this set or is asked to provide a complete preference ordering. Of course more complicated surveys

are also possible, for example, in a conjoint setting, see Van Ophem *et al.* (1999). However, in order to focus the discussion we keep the setup of the survey simple.

3.2.1 Rank-ordered logit model

Suppose that we want to learn the (determinants of) preferences of individuals over a discrete set of items. Classic examples of this problem include preferences over different modes of transportation or different brands in a product category. Preferences can be recovered from historical data but if some items are not available yet, a survey is usually the only option.

Denote the number of alternatives by J. The utilities for individual i are given by a set of variables $U_{i,1}, \ldots, U_{i,J}$, where $i = 1, \ldots, N$ indexes individuals and $j = 1, \ldots, J$ indexes the items. In the traditional setup, respondents are asked to choose their most preferred option out of the complete set of J alternatives. Let $y_{i,j} = 1$ denote that respondent i prefers alternative j most. The information $y_{i,j} = 1$ implies that for this respondent the utility of alternative j is larger than all other alternatives, that is,

$$U_{i,j} \geq \max\{U_{i,1}, \ldots, U_{i,J}\}. \tag{3.1}$$

It is generally assumed that the respondent actually knows all $U_{i,j}$, $j = 1, \ldots, J$, so from the respondents point of view a deterministic decision is made.[1] The econometrician does not observe $U_{i,j}$, we therefore have to make a (stochastic) model for the utilities. From the point of view of the econometrician the respondent then makes a probabilistic decision. We use the random utility framework, see Manski (1977), to represent the preferences of individuals. The random utilities for individual i are defined as

$$U_{i,j} = V_{i,j} + \varepsilon_{i,j}, \tag{3.2}$$

The utilities consist of two parts: $V_{i,j}$ is the location component of the utility, determined by observed individual characteristics and $\varepsilon_{i,j}$ is the random component of the utility of alternative j for individual i. In general the deterministic part of the utility is modeled as

$$V_{i,j} = x_i' \beta_j, \tag{3.3}$$

[1] One can relax this assumption by assuming that the individual only knows that the utility of the chosen alternative is at least as large as the utility of the other alternatives. This relaxation is without consequences for our econometric models. For simplicity of discussion we however maintain the assumption that the respondents know the utility levels.

where x_i is an m-dimensional vector with characteristics of individual i and β_j is an m-dimensional parameter vector specific to alternative j.

The probability of observing individual i choosing for alternative j now depends on the distribution of $\varepsilon_{i,j}$. If we assume that the population distribution of $\varepsilon_{i,j}$ is an type-I extreme value distribution and that the $\varepsilon_{i,j}$'s are independent of each other, we have the setup of a multinomial logit [MNL] model, see McFadden (1973, 1974). This leads to the well-known expression for the probability that item j is most preferred by individual i

$$\begin{aligned} \Pr[y_{i,j} = 1; \beta] &= \Pr[U_{i,j} \geq \max\{U_{i,1}, \ldots, U_{i,J}\}] \\ &= \frac{\exp(V_{i,j})}{\sum_{l=1}^{J} \exp(V_{i,l})}, \end{aligned} \tag{3.4}$$

where $\beta = \{\beta_1, \ldots, \beta_J\}$ and β_J is put equal to zero for identification.

The information on the most preferred item is enough to be able to estimate the model parameters. However, as discussed before, more information per respondent can be obtained if we ask for a ranking of alternatives and this will in general result in an efficiency gain. We will denote the response of respondent i by the vector $y_i = (y_{i,1}, \ldots, y_{i,J})'$, where $y_{i,j}$ now denotes the rank that individual i gives to item j. For example, if $y_{i,j} = 2$ this means that the respondent considers alternative j the second most preferred option. For notational convenience we will also use the equivalent notation $r_i = (r_{i,1}, \ldots, r_{i,J})'$, where $r_{i,j}$ denotes the item number that received rank j by individual i. The relation between r_i and y_i is given by

$$y_{i,k} = j \iff r_{i,j} = k \tag{3.5}$$

for $j, k = 1, \ldots, J$.

An observed ranking for a respondent implies a complete ordering of the underlying utilities. An individual will prefer an item with a higher utility over an item with a lower utility. As it is assumed that the individual knows all utility values, the respondent can easily provide a full ranking. From his/her point of view the ranking is deterministic. If we observe a full ranking r_i, we know that

$$U_{i,r_{i,1}} > U_{i,r_{i,2}} > \ldots > U_{i,r_{i,J}}. \tag{3.6}$$

It is obvious that (3.6) provides more information compared to (3.1). Under the utility assumption in (3.2) and the assumption of the extreme value distribution, we obtain the rank-ordered logit [ROL] model, see Beggs *et al.* (1981) and Chapman and Staelin (1982).

To the econometrician the probability of observing a particular ranking r_i equals

$$\begin{aligned} \Pr[r_i;\beta] &= \Pr[U_{i,r_{i,1}} > U_{i,r_{i,2}} > \ldots > U_{i,r_{i,J}}] \\ &= \prod_{j=1}^{J-1} \frac{\exp(V_{i,r_{i,j}})}{\sum_{l=j}^{J} \exp(V_{i,r_{i,l}})}. \end{aligned} \tag{3.7}$$

The ROL model can be seen as a series of MNL models: an MNL for the most preferred item; another MNL for the second-ranked item to be preferred over all items except the one with rank 1, and so on. Finally, the probability of a complete ranking is made up of the product of these separate MNL probabilities. The product contains only $J-1$ probabilities, because ranking the least preferred item is done with probability 1. Note that this result holds due to the IIA property of the MNL model, see Beggs *et al.* (1981) for a derivation. Furthermore, we do not have to assume that the respondent actually makes the decisions in this order. As discussed above, the respondent knows all utility values. Equation (3.7) only implies that we can look at the ranking *as if* consecutive choices are made.

3.2.2 Ranking ability

In the standard ROL model we implicitly assume that respondents are able to rank each item according to the random utility model. However, it has already been noted by Chapman and Staelin (1982) that for the less preferred items, this assumption does not always hold. One of the possible reasons for this is that the respondent perhaps has no experience with some of the items, and hence is not able to indicate the proper ranking order. It is also possible that respondents tend to find the least preferred items less important and rank those randomly. In practice this means that the observed rank order of the least preferred items may not be according to the model.

It is important to exactly identify the utility information that the respondent has in this case. Before we assumed that the respondent knows *all* utility values. Obviously, we can no longer maintain this assumption. Suppose that the respondent is only able to rank the top-k items. In utility terms this can be seen as that the respondent knows the utilities of these k items and (s)he knows that the utility of the other items is below the minimum of the utilities of the top-k items. In other words the respondent now also has only partial information on the utilities.

If the least preferred items are not ranked according to the underlying utility model, the use of those ranks in the estimation will bias the parameter estimates towards zero,

see Chapman and Staelin (1982, p. 292). Indeed, Hausman and Ruud (1987, p. 89) notice in an application on mobile phones that including more ranks in the estimation procedure leads to a decline in the absolute magnitude of the parameters.

The common solution to this problem is rather simple. If the ranks beyond k are biased, then do not use these lower rankings in the estimation procedure. This can be done very easily by letting the product in (3.7) only go up to k instead of $J-1$. The probability of observing a particular ranking by individual i if (s)he only ranks the k most preferred items according to utility values and ranks the remaining items $(J-k)$ randomly is given by

$$\begin{aligned} \Pr[y_i|k;\beta] &= \Pr[U_{i,r_{i,1}} > U_{i,r_{i,2}} > \ldots > U_{i,r_{i,k}} > \max\{U_{i,r_{i,k+1}},\ldots,U_{i,r_{i,J}}\}] \\ &= \prod_{j=1}^{k} \frac{\exp(V_{i,r_{i,j}})}{\sum_{l=j}^{J}\exp(V_{i,r_{i,l}})} \frac{1}{(J-k)!}, \end{aligned} \tag{3.8}$$

where we condition on the ranking capabilities k, that is, being able to rank the k most preferred items correctly. The second term in (3.8), $1/(J-k)!$, is usually not present in studies. The reason for this is that one usually considers the probability of observing only the first k rankings, that is, $\Pr[r_{i,1}, r_{i,2}, \ldots, r_{i,k}|k;\beta]$ instead of considering the full ranking y_i. For reasons that will become clear below, we also have to take into account the probability of observing a particular ranking of the least preferred items. We assume that the least preferred $J-k$ items are ordered randomly, therefore all $(J-k)!$ possible orderings are equally likely. Hence, the last term in (3.8) contains the probability of observing one particular ordering of the last $J-k$ items.

The assumption that the last $J-k$ ranks are given randomly follows from the assumption that individuals are unable to determine their preference over these items. More formally, a respondent may not know the relative values of $U_{i,r_{i,k+1}},\ldots,U_{i,r_{i,J}}$. Therefore, the individual does not have the information to make an ordering of the items. The ordering provided by the individual may therefore not reflect the underlying ordering of the utilities. Thus, if a respondent does not know if (s)he prefers one item over another, the provided ranking must be random. If it is not random, the respondent does have a preference for one of the items, and hence is capable to provide this ranking. Note that we do not assume that an individual knows nothing about $U_{i,r_{i,k+1}},\ldots,U_{i,r_{i,J}}$. We assume that (s)he does know that the maximum of these is smaller than $U_{i,r_{i,k}}$.

If one assumes k to be the same for the whole sample, the last term in (3.8) can be ignored in the estimation as it becomes a constant in the log likelihood, see Chapman

and Staelin (1982) and Hausman and Ruud (1987). These two papers propose estimating different ROL models, each using a different number of ranks, that is, a different value for k. The papers propose different methods of choosing a model from the resulting set. One can, for example, use a Hausman (1978) test to test for differences in the β parameter for different values of k. The tradeoff here is that using more ranks gives more efficient parameter estimates, but may introduce a bias in the results.

In the present chapter we relax the assumption that k is equal for all individuals. In the next subsection we will allow k to vary over the individuals using a latent-class approach.

3.2.3 Latent-class rank-ordered logit model

In the applications of the ROL model in the literature one either assumes that k is known a-priori, where $k = J$ is also possible, or one determines k by comparing the results for several values of k. The underlying assumption here is that there is no heterogeneity in the population concerning the capability to order the alternatives. In this section we will introduce such heterogeneity. Hence, determining the value of k becomes part of the model.

The introduction of heterogeneity of ranking abilities solves many practical issues. Firstly, one usually does not know beforehand how many ranks should be used for estimating a ROL model to balance the efficiency against a possible bias. Secondly, by allowing for the heterogeneity we make efficient use of the available data. For example, assume that 10% of the respondents can only give the most preferred item, and give a random ordering for the other items, and the rest (90%) is able to provide a complete ordering. In the standard model we would then be forced to only consider the first rank in the estimation for all individuals to avoid a bias in the estimated parameters. The additional information available in the responses of the 90% will not be used to make the estimates more efficient. Using the additional information is only possible, when heterogeneity in the ranking ability is incorporated in the model.

To allow for such heterogeneity, we divide our individuals into J latent classes, see, for example, Wedel and Kamakura (2000). For class $k = 0, 1, \ldots, J-1$ we impose that the individual can rank k most preferred items, more formally that they are based on the underlying utilities. The individual ranks the remaining $(J-k)$ items randomly. The

probability of observing a particular ranking for individual i, now becomes

$$\Pr[y_i;\beta,\pi] = \sum_{k=0}^{J-1} \pi_k \Pr[y_i|k;\beta], \tag{3.9}$$

where π_k is the probability that individual i belongs to segment k with $0 \leq \pi_k \leq 1$ and $\sum_{k=0}^{J-1} \pi_k = 1$, and where the probability of observing ranking y_i when only the k most preferred items are ranked according to the random utility model $\Pr[y_i|k;\beta]$ is given in (3.8). Equation (3.9) explains the reason for including the second term in (3.8). In order to be able to compare the segments, they must all contain the probability of observing the full ranking, not just the probability of observing the k most-preferred ranks.

The resulting model is a finite mixture model, see, for example, Titterington *et al.* (1985) and Everitt and Hand (1981). To estimate the model parameters β and the mixing proportions $\pi = (\pi_0, \ldots, \pi_{J-1})$ we rely on Maximum Likelihood. The likelihood function is given by,

$$\begin{aligned} l(\beta,\pi) &= \prod_{i=1}^{N} \Pr[y_i;\beta,\pi] \\ &= \prod_{i=1}^{N} \sum_{k=0}^{J-1} \pi_k \Pr[y_i|k;\beta] \\ &= \prod_{i=1}^{N} \sum_{k=0}^{J-1} \frac{\pi_k}{(J-k)!} \left[\prod_{l=1}^{k} \frac{\exp(x_i'\beta_{r_{i,l}})}{\sum_{m=l}^{J} \exp(x_i'\beta_{r_{i,m}})} \right]. \end{aligned} \tag{3.10}$$

The log likelihood is given by

$$\mathcal{L}(\beta,\pi) = \sum_{i=1}^{N} \log \left\{ \sum_{k=0}^{J-1} \pi_k \exp \left[-\log\left((J-k)!\right) + \sum_{l=1}^{k} \left(x_i'\beta_{r_{i,l}} - \log\left(\sum_{m=l}^{J} \exp(x_i'\beta_{r_{i,m}}) \right) \right) \right] \right\}, \tag{3.11}$$

where we have rewritten the product of probabilities as the exponent of a sum for numerical stability. The likelihood function can be maximized using numerical optimization algorithms. In our simulations and in the empirical section, we use a constrained optimization procedure to enforce the restrictions on the π parameters. Standard errors for the parameters can straightforwardly be obtained using the second-order derivative of the log likelihood.

Testing for empty classes

The total number of potential latent classes is equal to the number of items. If some of these classes are not present in the data, this may lead to an efficiency loss in the estimation of the β parameters. Therefore it is useful to test whether some of the classes can be removed from the model. It is especially interesting to test is whether $\pi_0 = 0$, that is whether each individual can at least provide his/her most preferred item. If this is not the case, even the standard MNL model will provide parameter estimates that are biased towards zero, see Section 3.4 for an illustration.

To test for the redundancy of a segment, we propose to use a standard likelihood ratio [LR] test. Because the β parameters are the same in each latent class, we do not suffer from the Davies (1977) problem, which is usually the case when testing for the number of latent classes. To test for the restriction of the absence of class j we consider the hypothesis $\pi_j = 0$. As the alternative is $\pi_j > 0$, we have a test for a parameter on the boundary of the parameter space. The asymptotic distribution of the LR test statistic is then a mixture of χ^2 distributions, for the case of testing one parameter the appropriate distribution is $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$, see Wolak (1989a,b). Hence, if we want to test at a 5% level of significance we have to use the 90% percentile of a χ^2_1 distribution, which is equal to 2.705. In case we want to perform a joint test for the absence of two or more segments, the asymptotic distribution will be a weighted average of χ^2 distributions where the weights follow from the covariance matrix of the estimated π parameters, see Wolak (1989a,b) for details.

3.2.4 Extension

So far we have assumed that respondents are only able to rank the first few items correctly. However, it may be possible that in some particular applications respondents are also able to indicate which alternatives they least prefer[2]. For example, it could be that they have tried something and were dissatisfied with it. This allows us to take advantage of the information in the lowest ranks. The efficiency in the parameter estimates will then increase even more relative to a standard MNL model.

In this case we denote the latent segments by two indices (k, l), where k denotes the number of most preferred items and l the number of least preferred ranks that can be ranked correctly. The case $l = 0$ corresponds to the previous discussion. We focus

[2]Note that this situation cannot be captured in a consideration set model.

the discussion below on the situation $l = 1$, where the lowest ranked alternative is also consistent with the utility model. We exclude the combinations $(J-2, 1)$ and $(J-1, 1)$. These cases actually correspond to an individual who is able to rank all alternatives. These individuals are classified in segment $(J-1, 0)$. Hence, we have to add $J-2$ additional latent classes to the model, with $l = 1$ and k ranging from 0 to $J-3$. The rank probabilities in the new segments $(k, 1)$ are given by

$$\Pr[y_i|(k,1);\beta] = \Pr[y_i|k;\beta]\Pr[U_{i,r_{i,J}} \leq U_{i,r_{i,m}} \; \forall m > k;\beta](J-k) \tag{3.12}$$

for $k = 0, \ldots, J-3$. The probability that alternative $r_{i,J}$ is preferred least in the set of items $(k+1, k+2, \ldots, J)$, $\Pr[U_{i,r_{i,J}} \leq U_{i,r_{i,m}} \; \forall m > k;\beta]$, is derived in Appendix 3.A. Due to the IIA property of the logit specification this probability can be seen as independent of the k top ranked alternatives. The factor $(J-k)$ is added in (3.12) because the number of possible random combinations of the $(J-k)$ least preferred items is reduced by the factor $(J-k)$ as we now know the least preferred item.

A similar exercise can be performed for $l > 1$. The number of classes will increases quickly, and one has to consider the tradeoff between the number of new latent classes and the potential efficiency gain.

3.3 Monte Carlo simulation

To illustrate the merits of our latent-class rank-ordered logit [LCROL] model, we perform a small simulation study. In this simulation we assume the presence of four alternatives $(J = 4)$. To facilitate matters, we assume that respondents are able to rank the most preferred items in a correct way and have problems ranking the remaining items. However, we do allow for all four potential latent segments, where in each segment the individuals are able to rank 0, 1, 2 or all most-preferred items correctly.

In the data generating process we put the probability of the first segment (ranking 0 items correctly) to 0 for two reasons. First, this restriction implies that the parameters of a multinomial logit [MNL] model can be estimated consistently, which allows us to make a fair comparison with our latent-segment model. The second reason is that it allows us to analyze the size of our proposed LR test for the redundancy of this segment. The size of the other segments are $\pi_1 = \pi_2 = 0.30$ and $\pi_3 = 0.40$.

The latent utilities are generated according to $U_{i,j} = \beta_{0,j} + x_{i,1}\beta_{1,j} + x_{i,2}\beta_{2,j} + \varepsilon_{i,j}$, for $j = 1, \ldots, 4$ and $i = 1, \ldots, N$. The variable $x_{i,1}$ is generated from a standard normal

distribution, $x_{i,2}$ is a 0/1 dummy variable with probability 0.5 of being 1. The fourth category is considered to be a base category and hence $\beta_{m,4} = 0$ for $m = 0, 1, 2$. The disturbances $\varepsilon_{i,j}$ are independently drawn from a type-I extreme value distribution. The values of the β parameters are displayed in the second column of Table 3.1.

In each replication of the simulation experiment we simulate utilities for $N = 1,000$ individuals who provide their ranking of the four alternatives. The 1,000 individuals are divided in one of the three segments according to the mixing proportions. The ranking of each individual is adjusted according to the imposed abilities in each segment, that is, a random ranking is imposed if individuals are assumed not being able to rank. For example, if an individual belongs to segment 2, we keep the preference order of the first two most preferred items but replace the preference order of the final two items by a random order.

We estimate four different models, that is, a standard MNL model, a rank-ordered logit [ROL] model, a LCROL model with 4 segments (including segment 0) and a LCROL model with 3 segments (without segment 0). Table 3.1 provides the means and the root mean squared errors [RMSE] of the parameter estimates over 10,000 replications.

The third column of Table 3.1 displays the results for the MNL model. The mean of the parameters is almost the same as the DGP parameters and hence one can estimate the β parameters in the utilities without a bias as expected. The fourth column shows the results for the usual ROL model. Note that this model is misspecified as about 60% of the individuals are not able to rank all four alternatives properly. This results in a clear bias towards zero in almost all of the β parameters, ranging from 4% to about 56%, which is clearly substantial. This confirms the findings of Chapman and Staelin (1982) and Hausman and Ruud (1987).

The fifth and sixth column of Table 3.1 display the results for our LCROL model. If we consider the case where we do not restrict π_0 to be zero, we see that the difference between the mean of the estimated parameters and the true parameters is at most 0.04. The RMSE of the estimator is however smaller than for the estimator of the same parameters in the MNL model. If we impose that $\pi_0 = 0$ the results even improve more. The mean of the estimated parameters and the true parameters differs at most 0.01. The RMSE of the estimator is even smaller.

The simulation results clearly show the advantage of our approach. With our LCROL model we use as much information as possible, while still obtaining unbiased results. An efficiency gain can be obtained if we remove redundant segments from our LCROL model.

Table 3.1: Mean and RMSE of the distribution of the parameters for N=1000 based on 10,000 Monte Carlo replications

Parameter	True	MNL[a]	ROL[b]	LCROL[c]	LCROL[d]
$\beta_{0,1}$	1.00	1.00	0.80	1.03	1.00
		(0.14)	(0.22)	(0.12)	(0.11)
$\beta_{1,1}$	0.75	0.76	0.54	0.78	0.76
		(0.12)	(0.22)	(0.11)	(0.09)
$\beta_{2,1}$	-0.30	-0.30	-0.32	-0.30	-0.30
		(0.21)	(0.12)	(0.17)	(0.16)
$\beta_{0,2}$	0.25	0.25	0.24	0.26	0.25
		(0.16)	(0.08)	(0.12)	(0.12)
$\beta_{1,2}$	-0.50	-0.50	-0.42	-0.51	-0.50
		(0.12)	(0.11)	(0.10)	(0.09)
$\beta_{2,2}$	0.45	0.45	0.31	0.47	0.45
		(0.21)	(0.18)	(0.17)	(0.17)
$\beta_{0,3}$	-0.25	-0.25	-0.11	-0.26	-0.25
		(0.18)	(0.17)	(0.14)	(0.13)
$\beta_{1,3}$	1.00	1.01	0.62	1.04	1.00
		(0.14)	(0.39)	(0.13)	(0.11)
$\beta_{2,3}$	0.80	0.80	0.50	0.83	0.80
		(0.23)	(0.32)	(0.19)	(0.18)
π_0	0.00	-	-	0.02	-
				(0.04)	
π_1	0.30	1	-	0.29	0.30
				(0.06)	(0.06)
π_2	0.30	-	-	0.29	0.30
				(0.10)	(0.10)
π_3	0.40	-	1	0.40	0.40
				(0.09)	(0.08)

[a] Standard multinomial logit model

[b] Rank-ordered logit model

[c] Latent-class rank-ordered logit model.

[d] Latent-class rank-ordered logit model with $\pi_0 = 0$

Table 3.2: Theoretical and empirical size of the LR test for $\pi_0 = 0$ (sample size is 1000)

Theoretical size	Empirical size	Critical value
1%	1.1%	5.41
5%	5.2%	2.71
10%	9.5%	1.64
20%	19.2%	0.71

To illustrate the applicability of the LR test for redundant segments, we compute in each replication of the simulation exercise the LR test for $p_0 = 0$ in the LCROL model. In Table 3.2 we report the empirical size, based on the asymptotic critical values (quantiles of the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ distribution). It can be seen that the empirical size of the LR test is close to the theoretical size. Also the mean (0.49) and the standard deviation (1.15) of the test statistic are very close to the theoretical mean and standard deviation (0.5 and 1.12 respectively, see 3.B) of the asymptotic distribution of the LR test.

In sum, we can conclude that the LCROL model provides consistent estimates in situations where not all individuals can fulfil a rank task properly. Furthermore, the RMSE of the estimator of the LCROL is smaller than for the MNL model where we only use the most preferred rank. An LR test for the absence of a segment is correctly sized, and hence it can be used to determine which latent segments are present in the data.

3.4 Application

To illustrate the practical usefulness of our latent-class rank-ordered logit model, we consider the results of a survey among 91 Dutch students. The students were asked to consider buying a computer to play games either for the first time or to replace their current platform. They had to rank 6 different platforms suitable to play computer games. The 6 platforms are the X-box (360), the PlayStation (2 or 3), the Gamecube (or Wii), the PlayStation Portable, the Gameboy (color/advance/DS/...) or just a regular PC. Note that the survey did not distinguish between different generations of the same platform. At the time of the survey, the Nintendo Wii and the Playstation 3 were not yet available. In addition we know which of the 6 platforms the student owns at the moment (if any) and the average number of hours that each student spends on gaming each week.

Model specification

First we estimate a multinomial logit [MNL] model for the most preferred platform. As explanatory variables we include platform intercepts, time (hours gaming) and a 0/1 dummy to indicate whether the student owns the platform, where 1 corresponds to ownership. The base alternative is a personal computer. The second column of Table 3.3 displays the parameter estimates. We notice that current platform ownership has a positive effect on preference for purchasing the same platform again and that individuals who spend more time gaming seem to prefer a personal computer over a real game computer. However, this effect is not significant for any platform.

The third column of Table 3.3 displays the parameter estimates for a standard rank-ordered logit [ROL] model. Hence, we implicitly assume that each student is capable of performing the complete ordering task. The parameter values differ substantially from the MNL estimates which suggests that this assumption is not valid. The Hausman (1978) test statistic for equal parameters equals 34.3, which is significant at the 5% level of significance.

The estimation results suggests that it seems necessary to include latent ordering abilities in the model. We include 6 classes indicating that the individuals cannot rank at all (segment 0), rank only the most preferred item (segment 1), the first 2, 3, 4 most preferred items (segment 2, 3, and 4) and all items (segment 5) with corresponding mixing proportions π_k for $k = 0, \ldots, 5$. The fourth column in Table 3.3 displays the parameter estimates of the latent-class rank-ordered logit [LCROL] model. The parameters seem to be different from the MNL estimates. This difference can be explained by the fact that about 23% (π_0) of the students are not able to rank the platforms at all. Indeed, the likelihood ratio test for $\pi_0 = 0$ equals 14.42 and hence this segment cannot be neglected. As additional check we compute the LR statistic for the restriction $\pi_1 = 1$ which leads to the MNL model. The value of the statistic is 110.20 and hence this restriction is clearly rejected. The ROL model is obtained when we impose the restriction $\pi_0 = \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0$. Again, this restriction is clearly rejected.

The estimates of π_2, π_3 and π_4 are relatively small. The LR statistic for the restriction $\pi_2 = \pi_3 = \pi_4 = 0$ equals 1.85 and hence we cannot reject the redundancy of these 3 segments. After imposing this restriction we obtain the model presented in the final column of Table 3.3. Using these results we find that the median individual prefers the

Table 3.3: Parameters estimates results for the MNL and ROL models with standard errors in parenthesis.

Variable	MNL	ROL	LCROL[a]	LCROL[b]	LCROL[c]
			intercept		
XBox	0.92	1.41	1.53	1.47	1.51
	(0.49)	(0.29)	(0.51)	(0.42)	(0.53)
Playstation	0.58	0.93	1.11	1.05	1.04
	(0.45)	(0.27)	(0.47)	(0.40)	(0.47)
Playstation portable	-0.03	0.80	0.44	0.79	0.53
	(0.59)	(0.28)	(0.52)	(0.51)	(0.56)
GameCube	0.49	-0.00	-3.50	-0.65	-2.18
	(0.59)	(0.30)	(1.61)	(0.55)	(1.16)
GameBoy	-1.47	0.08	-2.71	-0.51	-1.59
	(0.99)	(0.29)	(1.41)	(0.70)	(0.99)
			hours spent on gaming		
XBox	-0.10	-0.17	-0.14	-0.14	-0.14
	(0.06)	(0.05)	(0.06)	(0.06)	(0.06)
Playstation	-0.11	-0.13	-0.11	-0.11	-0.11
	(0.07)	(0.04)	(0.06)	(0.05)	(0.06)
Playstation portable	-0.10	-0.23	-0.36	-0.35	-0.40
	(0.11)	(0.05)	(0.12)	(0.12)	(0.13)
GameCube	-0.39	-0.18	-0.01	-0.22	-0.15
	(0.24)	(0.05)	(0.15)	(0.11)	(0.16)
GameBoy	-0.05	-0.23	-0.23	-0.33	-0.33
	(0.18)	(0.05)	(0.15)	(0.11)	(0.15)
Platform ownership	1.78	0.97	1.72	1.45	1.71
	(0.38)	(0.19)	(0.37)	(0.29)	(0.35)
π_0	-	-	0.23	-	0.21
			(0.07)		(0.07)
π_1	1	-	0.20	0.34	0.27
			(0.09)	(0.11)	(0.08)
π_2	-	-	0.07	0.01	-
			(0.08)	(0.09)	
π_3	-	-	0.07	0.08	-
			(0.09)	(0.12)	
π_4	-	-	0.00	0.00	-
			(0.16)	(0.62)	
π_5	-	1	0.43	0.57	0.52
LR statistic[d]	110.20	32.73	-	14.42	1.85

[a] Latent-class rank ordered logit model

[b] Latent-class rank ordered logit model with $\pi_0 = 0$

[c] Latent-class rank ordered logit model with $\pi_2 = \pi_3 = \pi_4 = 0$

[d] LR statistic to test against the LCROL model in fourth column.

PC for playing games.[3] The PC has a probability of 0.45 of being the most preferred platform. The Xbox and the Playstation come second and third with 0.28 and 0.19, respectively.

Students who spend more time playing games, have more preference for the personal computer. Owning a platform has a positive effect on the preference for that platform. This can be due to two effects. The first is because of reverse causality, a students owns the platform, because he/she likes it. But since we consider multiple generations of the same platform this can also be interpreted as a backwards compatibility effect. Someone who owns a lot of PS2 games would rather have a PS3 than a Nintendo Wii, because the PS3 can, for example, still play the PS2 games.

Interpretation of the segments

In the final model, the mixing proportions divide the students in three segments. About 52% of the students know enough about the different platforms to give a complete ranking. The remaining 48% of the students can either not provide a clear ranking at all (21%), or only know which of the platforms they prefer most (27%).

We expect that the first segment of students consists of heavy gamers, who spend a lot of time on gaming. The other 2 segments contain students who game less. To segment the individual students given their responses we can compute the conditional segment membership probabilities defined by

$$\hat{\pi}_{i,k} = \frac{\pi_k \Pr[y_i|k;\beta]}{\sum_{l\in\mathcal{K}} \pi_l \Pr[y_i|l;\beta]}, \tag{3.13}$$

where $\hat{\pi}_{i,k}$ is the conditional probability that individual i belongs in latent class k given the observed ranking, π_k are the estimated mixing proportions of the segments and $\Pr[y_i|k;\beta]$ is given in (3.8). The set $\mathcal{K}$ contains the segments included in the model. In our final model we had the segments 0, 1 and 5.

The segment membership probabilities (3.13) allow us to assign the students to the three groups. The average value of the largest conditional segment membership probability over the students is 0.80, which indicates that the model is capable to make a clear distinction between the segments of students. In Table 3.4 we display the average value of the explanatory variables in the 3 segments. If we consider the number of hours spent on gaming, the results are as we expect. Students who spend more time on gaming are

[3]The median individual spends 2 hours per week on gaming, and only owns a PC.

Table 3.4: Platform ownership and average time spend per week on gaming in hours for each latent segment

	Segment 0	Segment 1	Segment 5
Hours spent on gaming	2.54	3.47	4.64
Ownership XBox	11%	17%	12%
Ownership PlayStation	30%	27%	35%
Ownership PSP	7%	13%	10%
Ownership Gamecube	11%	14%	5%
Ownership Gameboy	21%	12%	11%
Ownership PC	83%	85%	91%

better able to rank the platforms than students who game less. The differences are less clear with ownership, except for the PC where we find that respondents who are not able to rank are less likely to own a PC, although differences are small.

3.5 Conclusions

The respondents inability to accurately provide a full ranking of all presented alternatives in a survey leads to a bias in the application of the well-known rank-ordered logit model. To remove this bias, while still taking maximum advantage of the information in the ranked data, we propose to augment the rank-ordered logit model with latent segments. Each latent segment is associated with a particular ranking ability. More specifically, given J alternatives we define J segments. In the k-th segment we assume that the respondent is only able to rank the k most preferred items correctly. We also allow k to be zero, in this case the respondent does not even report the most preferred item correctly. Under this situation even the multinomial logit model would provide biased estimates.

Using simulation and an empirical application, we show that our new model is indeed robust against inabilities of individuals to give proper ranks. Moreover, it is more efficient than a standard multinomial logit approach. All this taken together, our model makes it very attractive to ask respondents in a survey to rank all options instead of asking them to select their preferred option. The application also shows that it is not unlikely that one may encounter respondents who cannot rank the alternatives at all. The direct application of the multinomial logit model is therefore not always appropriate.

Our analysis is based on the logit framework. Therefore we have to assume that the IIA property holds. If one wants to relax this assumption, our proposed strategy can be incorporated in rank-ordered probit models (Hajivassiliou and Ruud, 1994) or mixed rank-ordered logit model of Calfee *et al.* (2001), which do not suffer from the IIA assumption.

3.A Derivation of probability of least preferred item

In this appendix we derive the probability that a particular item is preferred least in a set of J items, see Van Ophem *et al.* (1999) for an alternative derivation. The probability of interest is given by

$$\Pr[U_{i,r_{i,J}} \leq U_{i,r_{i,m}} \ \forall m > k]. \tag{3.14}$$

The IIA property of the logit models implies that this probability does not depend on the utilities of the k most preferred items. In fact, this probability is the same as the probability that item $r_{i,J}$ has rank $J-k$ in the set of items $(r_{i,k+1}, \ldots, r_{i,J})$. To simplify the discussion, we derive without loss of generality the probability that item 1 is ranked last in a full set of alternatives, which can easily be generalized to the probability in (3.14).

$$\begin{aligned}
\Pr[y_1 = J] &= \Pr[U_1 \leq U_2, U_1 \leq U_3, \ldots, U_1 \leq U_J] \\
&= \Pr[V_1 + \varepsilon_1 \leq V_2 + \varepsilon_2, \ldots, V_1 + \varepsilon_1 \leq V_J + \varepsilon_J] \\
&= \Pr[\varepsilon_2 > V_1 - V_2 + \varepsilon_1, \ldots, \varepsilon_J > V_1 - V_J + \varepsilon_1] \\
&= \int_{-\infty}^{\infty} f(\varepsilon_1) \int_{V_1-V_2+\varepsilon_1}^{\infty} f(\varepsilon_2) \cdots \int_{V_1-V_J+\varepsilon_1}^{\infty} f(\varepsilon_J) d\varepsilon_J \ldots d\varepsilon_2 d\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f(\varepsilon_1) \int_{V_1-V_2+\varepsilon_1}^{\infty} f(\varepsilon_2) \cdots \int_{V_1-V_{J-1}+\varepsilon_1}^{\infty} f(\varepsilon_{J-1}) \times \\
&\qquad\qquad [\exp(-e^{-\varepsilon_J})]_{V_1-V_J+\varepsilon_1}^{\infty} d\varepsilon_{J-1} \ldots d\varepsilon_2 d\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f(\varepsilon_1) \int_{V_1-V_2+\varepsilon_1}^{\infty} f(\varepsilon_2) \cdots \int_{V_1-V_{J-1}+\varepsilon_1}^{\infty} f(\varepsilon_{J-1}) \times \\
&\qquad\qquad [1 - \exp(-e^{V_J-V_1-\varepsilon_1})] d\varepsilon_{J-1} \ldots d\varepsilon_2 d\varepsilon_1 \\
&= \int_{-\infty}^{\infty} f(\varepsilon_1)[1 - \exp(-e^{V_2-V_1-\varepsilon_1})] \cdots [1 - \exp(-e^{V_J-V_1-\varepsilon_1})] d\varepsilon_1,
\end{aligned} \tag{3.15}$$

where we suppress the subscript i for notational convenience. Expanding the terms in brackets leads to

$$\begin{aligned}
\Pr[y_1 = J] = &\int_{-\infty}^{\infty} f(\varepsilon_1) d\varepsilon_1 - \sum_{i=2}^{J} \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{V_i - V_1 - \varepsilon_1}) d\varepsilon_1 \\
&+ \sum_{i=2}^{J-1} \sum_{j=i+1}^{J} \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{V_i - V_1 - \varepsilon_1}) \exp(-e^{V_j - V_1 - \varepsilon_1}) d\varepsilon_1 \\
&+ \ldots + (-1)^{J-1} \int_{-\infty}^{\infty} f(\varepsilon_1) \exp(-e^{V_2 - V_1 - \varepsilon_1}) \cdots \exp(-e^{V_J - V_1 - \varepsilon_1}) d\varepsilon_1 \\
= 1 &- \sum_{i=2}^{J} \frac{\exp(V_1)}{\exp(V_1) + \exp(V_i)} + \sum_{i=2}^{J-1} \sum_{j=i+1}^{J} \frac{\exp(V_1)}{\exp(V_1) + \exp(V_i) + \exp(V_j)} \\
&+ \ldots + (-1)^{J-1} \frac{\exp(V_1)}{\exp(V_1) + \exp(V_2) + \cdots + \exp(V_J)},
\end{aligned} \tag{3.16}$$

where we use the standard logit characteristic. To simplify notation, we define the set T_i to contain the sums of all possible combinations of i distinct elements from the set $\{\exp(V_2), \ldots, \exp(V_J)\}$, that is, T_i contains $\binom{J-1}{i}$ elements. We will denote one specific element from the set T_i as $T_{i,j}$, $j = 1, \ldots, \binom{J-1}{i}$. The probability (3.16) can now be written as

$$\Pr[y_1 = J] = 1 + \sum_{i=1}^{J-1} (-1)^i \sum_{j=1}^{\binom{J-1}{i}} \frac{\exp(V_1)}{\exp(V_1) + T_{i,j}}. \tag{3.17}$$

3.B Derivation of the mean and variance of the LR test under the null hypothesis

In this appendix we give the mean and standard deviation of a mixture of a χ^2_0 and a χ^2_1 distribution, where both distributions have an equal weight. The density function is given by

$$f(L) = \frac{1}{2} I(L = 0) + \frac{1}{2} \frac{L^{-\frac{1}{2}} \exp(-\frac{L}{2})}{\sqrt{2}\Gamma(\frac{1}{2})}. \tag{3.18}$$

The expectation of L is given by

$$\begin{aligned} E(L) &= \frac{1}{2} \times 0 + \frac{1}{2} \int_0^\infty L \frac{L^{-\frac{1}{2}} \exp(-\frac{L}{2})}{\sqrt{2}\Gamma(\frac{1}{2})} dL \\ &= \frac{1}{2\sqrt{2}\Gamma(\frac{1}{2})} \int_0^\infty L^{\frac{1}{2}} \exp(-\frac{L}{2}) dL. \end{aligned} \tag{3.19}$$

Notice that the integral in (3.19) is a kernel of a χ^2_3 distribution, therefore we also know what the value of the integral is.

$$\begin{aligned} E(L) &= \frac{1}{2\sqrt{2}\Gamma(\frac{1}{2})} \int_0^\infty L^{\frac{1}{2}} \exp(-\frac{L}{2}) dL \\ &= \frac{2\sqrt{2}\frac{1}{2}\Gamma(\frac{1}{2})}{2\sqrt{2}\Gamma(\frac{1}{2})} \\ &= \frac{1}{2}. \end{aligned} \tag{3.20}$$

To compute the variance (or standard deviation), we only need to compute $E(L^2)$, as $V(L) = E(L^2) - E(L)^2$.

$$\begin{aligned} E(L^2) &= \frac{1}{2} \times 0^2 + \frac{1}{2} \int_0^\infty L^2 \frac{L^{-\frac{1}{2}} \exp(-\frac{L}{2})}{\sqrt{2}\Gamma(\frac{1}{2})} dL \\ &= \frac{1}{2\sqrt{2}\Gamma(\frac{1}{2})} \int_0^\infty L^{1\frac{1}{2}} \exp(-\frac{L}{2}) dL \\ &= \frac{4\sqrt{2}\frac{3}{2}\frac{1}{2}\Gamma(\frac{1}{2})}{2\sqrt{2}\Gamma(\frac{1}{2})} \\ &= \frac{3}{2}, \end{aligned} \tag{3.21}$$

where we again use the trick with the χ^2 kernel. The variance is now equal to

$$V(L) = \frac{3}{2} - \frac{1}{2}^2 = 1\frac{1}{4}. \tag{3.22}$$

Thus, the standard deviation is equal to

$$\sigma(L) = \sqrt{1\frac{1}{4}} \approx 1.118. \tag{3.23}$$

Chapter 4

Modeling Regional House Prices

4.1 Introduction

Real estate prices in many countries have experienced a dramatic boom in recent years (IMF, 2004). At the same time, the extent of the price increase appears to vary substantially across different regions within a given country. In the Netherlands, for example, it is commonly believed that house prices in *Amsterdam* and the densely populated western part of the country have increased far more than prices in the smaller cities and rural areas in the east. As house prices are typically available per region or city, we may analyze these data at such a disaggregate level, to examine whether indeed regions or cities behave differently, perhaps in terms of trends, but also in terms of response to outside economic shocks. In this chapter we develop a time series model that suits this purpose.

Most regional house prices have the following properties. First, they tend to display a trend, and historical price patterns suggest that this trend probably is not deterministic but stochastic. In particular, house prices show 'bubble'-type behavior, where prolonged periods of steady increases of the price level suddenly end with a sharp drop followed by a period of low price levels, suggesting that trends are unlikely to be deterministic. Second, for different regions within a country these stochastic trends should somehow be linked. It is not plausible that prices in different regions would diverge indefinitely or that certain regions would not respond to common macroeconomic shocks. So, a model for regional house prices should allow for some form of common trends. Third, it can be expected that adjacent regions show similar price patterns, although this may also be the case for regions far apart geographically but with similar economic or demographic characteristics. Hence, a suitable model should allow for similarities in the dynamic behavior of house

prices across regions. An intuitively appealing possibility is to consider a model that allows for groups or clusters of regions, where house price dynamics in regions within a given cluster are the same, while they are different across clusters. Preferably, such a model should not require ex-ante or exogenous assignment of regions to specific clusters. In fact it would be best if the data themselves were allowed to indicate if clusters exist and if so, which regions belong to which cluster.

In this chapter we extend the latent-class panel time series model introduced by Paap *et al.* (2005) to capture these different properties of regional house prices. The key feature of this model is that the clustering of regions is purely data-driven, where cluster membership is based on characteristics corresponding to two specific research questions we want to address. The first question is whether prices in all regions have the same average growth rate. Note that a common trend specification across the regions entails that their growth rates must be somehow compatible, but it still leaves open the possibility that house prices in some regions grow faster than in others. The second question we consider is the way the house prices in each region react to changes in the overall economic situation, which we measure by GDP. We examine both the size of the effect from GDP on the house prices and the speed at which regions react to changes in GDP.

We apply our model to house price data for the Netherlands, comprising 76 regions for which we have quarterly data for the period 1985Q1-2005Q4. We find that the 76 regions can be grouped into two clusters. The first cluster consists mainly of regions in the east of the country. These are mainly rural areas that are close to the larger cities, especially close to the *Randstad* (consisting of *Utrecht*, *Amsterdam*, *Den Haag*, *Rotterdam* and other cities in the area). This cluster reacts both stronger and faster to changes in GDP. The average growth rate does not vary over the regions.

There are not many studies that describe regional house prices. Cameron *et al.* (2006) build a model from inverse demand equations. They have, however, only a limited number (9) of regions, and their model would not work in our situation where we have many more (76) regions, as we will describe below. Malpezzi (1999) constructs an error correction model for regional house prices. The parameters of this model are however not allowed to vary across regions. Holly *et al.* (2008) model US house prices at the state level. Their model is 'fully heterogenous' in the sense that it has different parameters for each region. In this chapter we cover the middle ground, that is, the model parameters are allowed to vary across *groups* of regions but not across each region individually.

Before we propose our latent-class model, we first provide some details on the house price data in Section 4.2. We consider two decades of quarterly house prices on 76 regions in the Netherlands. We discuss their trending behavior by performing panel unit root tests and we also show that the growth rates in different regions show strong cross-correlations. Using multidimensional scaling techniques we get a first impression if and how these 76 regions could get clustered. Then, in Section 4.3, we put forward our model specification, highlighting the underlying data-driven clustering mechanism. In addition, we describe the method used for parameter estimation. In Section 4.4 we first present our estimation results, and give interpretation to the various outcomes. Next, we take a look at impulse response functions of the house prices with respect to a shock in GDP and in the interest rate. In Section 4.5 we conclude with some limitations and we outline topics for further research.

4.2 Data

The Dutch real estate agent association [NVM] publishes quarterly data on house prices for $N = 76$ regions in the Netherlands. Our dataset covers the sample period 1985Q1-2005Q4 ($T = 84$ quarters). Hence, we have a panel database where both the cross-section dimension N and the time dimension T are fairly large.

The way the country is divided into 76 regions is determined by the NVM. Macroeconomic data, such as output and inflation, are not available for this particular specification of regions. Other (macro) variables that we use in our model are therefore measured at the country level. In particular, this concerns the interest rate (obtained from the Dutch Central Bank) and quarterly real GDP (from Statistics Netherlands). The GDP series is available until 2005Q2. We obtain real house prices by deflating with the consumer price index [CPI] (from Statistics Netherlands). In addition, we seasonally adjust the real GDP series using the Census X-12 algorithm (available in EViews 5.1). We denote the real house price in region i at time t as $p_{i,t}$, and real GDP as y_t.

Figure 4.1 shows time series of $\log(p_{i,t})$ for three specific regions: *Noordwest-Friesland*, which usually is the least expensive region, *Bunnik/Zeist*, which usually is the most expensive region, and *Amsterdam*, which is in between. On top we also plot $\log(y_t)$ (scaled to limit the size of the vertical axis in the graph). Comparing the graphs in Figure 4.1 suggests that real house prices increase slightly faster than real GDP. Prices in *Bunnik/Zeist* and *Amsterdam* show substantial variations in the trend growth rate over time,

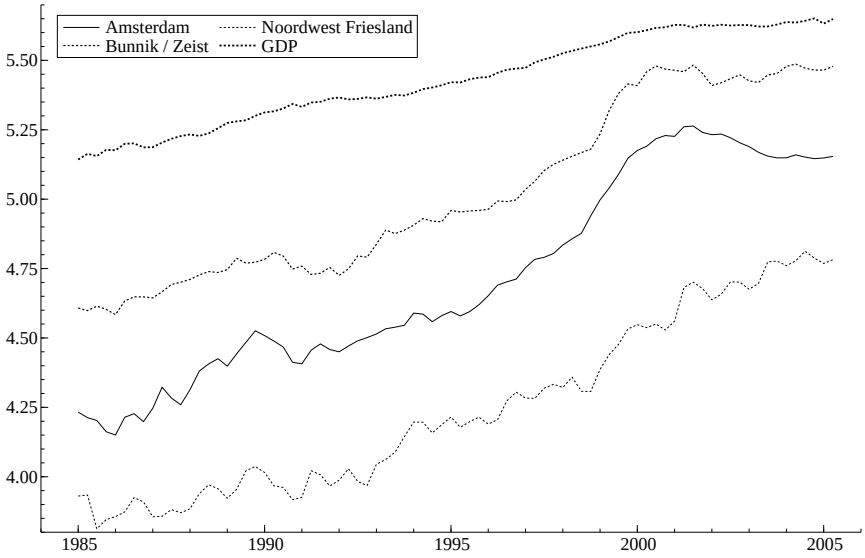


Figure 4.1: Log house prices for 3 distinct regions, and log GDP.

with alternating periods of steep price increases and of stable or falling prices. Especially the 'hump' in the prices around 2000 stands out clearly. This suggests that the trend in the house prices is stochastic rather than deterministic. Furthermore, as the trending behavior of the different price series seems quite similar regional house prices may well be cointegrated.

4.2.1 Unit roots and cointegration

To test whether these visual impressions from Figure 4.1 can be given more formal statistical support, we perform panel unit root tests on the regional house prices. Two of the most popular tests in the literature are those from Levin *et al.* (2002) [LLC] and Im *et al.* (2003) [IPS], see Breitung and Pesaran (2008). These tests have as null hypothesis the presence of a unit root in all the series in the panel. The alternative hypotheses are different however. Levin *et al.* (2002) assume that the house price dynamics are the same for each region, and therefore the alternative hypothesis is that all regional house prices are stationary. Im *et al.* (2003), however, have as alternative hypothesis that at least one regional house price is stationary. Both these tests assume that there is no cross-correlation between different series in the panel. In fact, they are not consistent if such a dependency is present, which is quite likely in our case. Alternative tests that do allow

for cross-section dependence are available, like the one in Moon and Perron (2004), but these usually rely on asymptotics that require T to be much larger then N, while in our case they are about equal.

To meet our data characteristics, we therefore employ the cross-sectionally augmented IPS [CIPS] test, recently developed in Pesaran (2007). This allows for cross-sectional dependence, and is also valid when N is larger than T. The idea of the CIPS test is to add the cross-section averages of the lagged levels and first differences to the familiar augmented Dickey-Fuller [ADF] regression equation. If it can be assumed that the cross-correlations are caused by a common factor, then this common factor must also be present in the cross-section averages. Adding these to the ADF equations should then get rid of the common factor in the residuals and thus correct for the presence of cross-correlations.

As the CIPS test is known to have reduced power relative to the IPS and LLC tests in case cross-correlation is not present, we test whether we really should use the CIPS test instead of these simpler tests. For this purpose we use the cross-section dependence [CD] test of Pesaran (2004) and the adjusted LM [LM_{adj}] test of Pesaran *et al.* (2008). These tests both use the cross-correlations between the residuals of the individual ADF regressions for the different regions. The CD test takes a simple sum which is scaled such that it has a standard normal distribution under the null hypothesis of no cross-sectional dependence. Therefore, the CD test has little power in case there are both positive and negative correlations such that the average is close to zero. The LM_{adj} test, however, is also valid in this case as it employs the squares of the cross correlations in the construction of the test statistic. However, the LM_{adj} test is less robust against non-normally distributed error terms and exhibits size distortions, especially when N is much larger than T.

Table 4.1 gives the result of these tests for the panel of quarterly growth rates in house prices $\Delta \log(p_{i,t})$, where Δ denotes the first-difference filter, and of $\log(p_{i,t}) - \log(p_{34,t})$, which is the difference of each series with the log house prices in *Amsterdam* (region 34, see Appendix 4.A). The number of lagged (first) differences is allowed to vary across each (C)ADF equation and is determined by minimizing BIC. Adding a lagged variable means losing one observation, therefore we actually minimize BIC$/T$, see Cameron and Trivedi (2005, pp. 279) or the definition of BIC given in Franses and Paap (2001). Each (C)ADF regression equation contains an intercept and a trend.

From the second column of Table 4.1 we see that for the first difference of the log house prices there is substantial cross-sectional dependence, according to both the CD and LM_{adj} tests. Next, we see that all three unit root tests reject the presence of a

Table 4.1: Results of the CD test, the LM_{adj} test and three different tests for a unit root for two series (boldface numbers indicate rejection of the null hypothesis).

Test	Series	$\Delta[\log(p_{i,t})]$	$\log(p_{i,t}) - \log(p_{34,t})$
CD[a]		**92.0**	**144.2**
LM_{adj}[a]		**60.4**	**175.1**
LLC[a]		**-61.2**	2.0
IPS[a]		**-55.9**	1.9
CIPS[b]		**-8.9**	**-3.5**

[a] Test statistic is asymptotically distributed as normal

[b] Tables with critical values for various values for N and T are given by Pesaran (2007), in the presence of an intercept and a trend in the CADF equations and for $N = T = 70$ the critical value at the 95%-level is -2.58, for $N = T = 100$ it is -2.56.

unit root in these growth rate series. Results for the difference between the log price in a region and the log price in *Amsterdam* (region 34) appear in the third column of Table 4.1. The reason for examining the log price differences with respect to *Amsterdam* is that finding these to be stationary, we can conclude that the house prices in each region are cointegrated. Again, the CD and LM_{adj} tests indicate that there is substantial cross-sectional dependence. Next, the LLC and IPS unit root tests do not reject the presence of a unit root, but the CIPS test does. Since the LLC and IPS tests are not valid in case of cross-sectional dependence, we rely on the CIPS test and conclude that the log house prices in each region are cointegrated. Note that the $(1, -1)$ cointegration relationships suggested by the results in Table 4.1 are quite plausible. It means that the difference between the log of house prices, or, equivalently the ratio of house prices, in each region is a stationary process. This constrains the long-term growth of house prices in each region to be about the same.

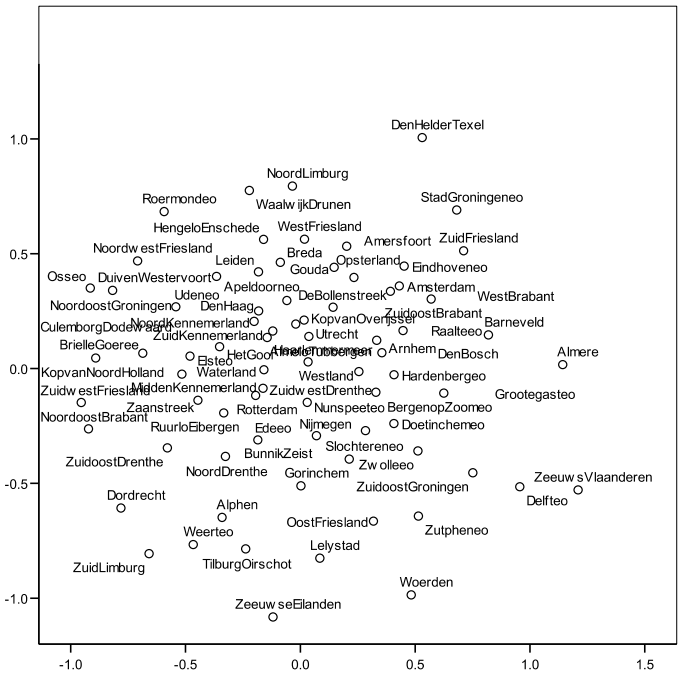


Figure 4.2: Multidimensional scaling plot of the regions, based on the correlations of the first differences of the log house prices over the period 1985Q1-2005Q4.

4.2.2 Clusters

Before we turn to our conditional clustering analysis using latent class techniques we consider unconditional clustering based on the correlations of the house price growth rates or of the residuals of the ADF regressions used above. For this purpose, we use multi-dimensional scaling [MDS], which results in the graphs shown in Figure 4.2 and 4.3.

Although the graphs in these figures are rather different, they basically lead to the same conclusion that there are no apparent clusters. Hence, dividing the regions into different groups based only on the cross-correlations of the regional house prices is not

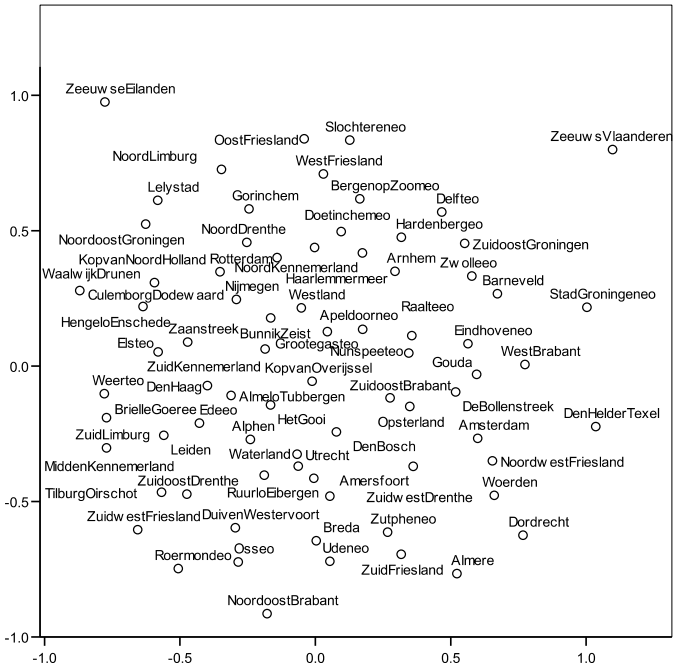


Figure 4.3: Multidimensional scaling plot of the regions, based on the correlations of the residuals of the ADF regressions for the log house prices over the period 1985Q1-2005Q4.

a meaningful possibility. Apparently, we need a more sophisticated clustering method, perhaps based on latent classes, as we will propose in the next section.

4.3 The model

In this section we put forward the specification of the latent-class panel time series model for describing the regional house prices. We first discuss the characteristics of the model, and then we outline the parameter estimation procedure.

4.3.1 Representation

Our starting point is the latent-class panel time series model developed by Paap *et al.* (2005). The crucial idea behind this model is that the individual time series may be grouped into a limited number of clusters. Within each cluster, a linear model is assumed to describe the dynamic behavior of the time series. The clusters are defined such that the model parameters are the same for all time series within a cluster, but they are different across clusters. Hence, this model covers the middle ground between a pooled regression model, where the model parameters are constrained to be the same for all regions, and a 'fully heterogenous' model, where the parameters are allowed to be different for each individual region. Whereas a pooled regression model may be too restrictive, a fully heterogenous model may be too flexible and ignores the possible similarities between regions. Finally, the key feature of the model of Paap *et al.* (2005) is that the number of clusters in the model as well as the allocation of the individual time series to different clusters is purely data-based. This avoids *ex ante*, and necessarily subjective, grouping of regions according to geographical location or economic or demographic characteristics, for example.

In our model for quarterly growth rates of house prices we allow for more flexibility than was done in Paap *et al.* (2005). As mentioned, there are two research questions we want to answer with our model and each question corresponds to different parameters that can vary across the latent classes. The first is whether the mean growth rates of house prices are the same across all regions. We therefore allow the clusters to have a different average growth rate by allowing for a class-specific intercept. To facilitate interpretation, we demean all other variables in the model such that the intercept is equal to the average growth rate of the house prices in the regions in a cluster.

The second question we wish to answer with our model is whether the house prices in regions follow the trend in real GDP. We add an error correction variable linking regional real house prices and real GDP, where the long-run parameter should be estimated. This long-run parameter determines the size of the effect of GDP on the house prices. The adjustment parameter indicates how fast the house prices in a region react to changes in GDP.

Based on the above discussion, we propose the following latent-class panel time series model for regional house prices in the Netherlands

$$\Delta \log(p_{i,t}) = \beta_{0,k_i} + \beta_{1,k_i}[\log(p_{i,t-1}) + \gamma_{k_i} \log(y_{t-1})] + \eta_{i,t}. \tag{4.1}$$

The β and γ parameters are class-specific parameters, where the subscript $k_i = 1, \ldots, K$ denotes the latent class which region i belongs to with K being the number of latent classes. We denote the probability that a region belongs to latent class k, or the mixing proportions, as π_k. Naturally it must hold that, $0 < \pi_k < 1$ and that $\sum_{k=1}^{K} \pi_k = 1$.

As the house prices of each regions are cointegrated with GDP, they are also cointegrated amongst themselves. This can easily be seen in the following way. Both $p_{i,t} - \gamma_{k_i} y_t$ and $p_{j,t} - \gamma_{k_j} y_t$ are stationary series. Now, consider the following expression,

$$(p_{i,t} - \gamma_{k_i} y_t) - \delta(p_{j,t} - \gamma_{k_j} y_t) = (p_{i,t} - \delta p_{j,t}) - (\gamma_{k_i} - \delta\gamma_{k_j}) y_t. \tag{4.2}$$

The LHS of (4.2) is stationary, therefore the RHS is also a stationary series. For $\delta = \gamma_{k_i}/\gamma_{k_j}$ the second term on the RHS of (4.2) will disappear, therefore regions i and j must have a $(1, -\delta)$ cointegration relationship. Two regions in the same cluster will therefore have a $(1, -1)$ cointegration relationship, because they share the same γ parameter. As we have seen in Section 4.2.1, there is support for exactly this relationship.

Even though model (4.1) includes $\log(y_{t-1})$, which is the same for all regions, there may still be some cross-section correlation among the house prices that is not captured. Therefore, following Holly *et al.* (2008), we allow the error term $\eta_{i,t}$ in (4.1) to be correlated across regions, but assume that this correlation is due to dependence on certain common factors. To be precise, we consider the specification

$$\eta_{i,t} = \alpha_{1,i} \Delta \log(y_{t-1}) + \alpha_{2,i} I_{t-1} + \alpha_{3,i} \Delta \log(p_{t-1}) + \varepsilon_{i,t}, \tag{4.3}$$

where I_{t-1} denotes the interest rate at time $t-1$, p_{t-1} denotes the average house price in the Netherlands at time $t-1$ and where $\alpha_{l,i}$ for $l = 1, 2, 3$ are region-specific parameters. The residuals $\varepsilon_{i,t}$ are now assumed to be independently normally distributed with a region-specific variance σ_i^2.

In the application below, we demean all variables in (4.1) and (4.3) and hence the intercepts β_{0,k_i} in (4.1) are equal to the average growth rates of the house prices in the latent classes k_i for $k_i = 1, \ldots, K$.

4.3.2 Estimation

The parameters in our model (4.1) with (4.3) can be estimated as outlined in Paap *et al.* (2005), using the EM algorithm of Dempster *et al.* (1977). This makes use of the full data log-likelihood function, that is, the joint density of the house prices and the latent classes

k_i, which we specify in detail below. The EM algorithm is an iterative maximization algorithm, which alternates between two steps until convergence occurs. In the first step (E-step) we compute the expected value of the full data log-likelihood function with respect to the latent classes k_i, $i = 1, \ldots, N$, given the house prices and the current values of the model parameters. In the second step (M-step) we maximize the expected value of the full data log-likelihood function with respect to the model parameters. As the model given the class memberships can be written as a standard linear regression, the M-step amounts to a series of (weighted) regressions. As the EM algorithm maximizes the log-likelihood function, the resulting estimates of the model parameters are equal to the maximum likelihood [ML] estimates. We can therefore compute standard errors of the estimates using the second derivative of the log-likelihood function.

Note that due to the presence of the term $\beta_{1,k_i}[\log(p_{i,t-1}) + \gamma_{k_i} \log(y_{t-1})]$ the model in (4.1) is actually nonlinear in the parameters. To deal with this issue, we follow Boswijk (1994) and rewrite the model as

$$\Delta \log(p_{i,t}) = \beta_{0,k_i} + \beta_{1,k_i} \log(p_{i,t-1}) + \beta_{2,k_i} \log(y_{t-1}) + \eta_{i,t}, \tag{4.4}$$

where $\beta_{2,k_i} = \beta_{1,k_i}\gamma_{k_i}$. Note that (4.4) is linear in the parameters, which facilitates estimation. The ML estimate $\hat{\gamma}_{k_i}$ can be obtained from the ML estimates of β_{1,k_i} and β_{2,k_i} as $\hat{\beta}_{2,k_i}/\hat{\beta}_{1,k_i}$.

The full data likelihood function, that is, the joint density of $\mathcal{P} = \{\{\Delta \log p_{i,t}\}_{t=1}^{T}\}_{i=1}^{N}$ and $\mathcal{K} = \{k_i\}_{i=1}^{N}$ is given by

$$l(\mathcal{P}, \mathcal{K}; \theta) = \prod_{i=1}^{N} \left(\prod_{k=1}^{K} \left(\pi_k \prod_{t=1}^{T} \frac{1}{\sigma_i} \phi(\varepsilon_{i,t}^{k}/\sigma_i) \right)^{I[k_i=k]} \right), \tag{4.5}$$

where $\phi(\cdot)$ denotes the probability density function of a standard normal random variable and θ is a vector containing all model parameters. The error term at time t for region i belonging to cluster k is defined as

$$\varepsilon_{i,t}^{k} = \Delta \log p_{i,t} - x_{i,t}'\beta_k - w_t'\alpha_i, \tag{4.6}$$

where $x_{i,t}$ is the (3×1) vector with the regressors appearing in (4.4) and β_k contains the corresponding parameters for cluster k. Similarly, w_t is the (3×1) vector with common factors in the specification for $\eta_{i,t}$ in (4.3), and $\alpha_i = (\alpha_{1,i}, \alpha_{2,i}, \alpha_{3,i})'$ containing the parameters for region i.

The expectation of the full data log-likelihood function with respect to $\mathcal{K}|\mathcal{P}, \theta$ [E-step] is given by

$$\mathcal{L}(\mathcal{P};\theta) = \sum_{i=1}^{N}\left(\sum_{k=1}^{K}\hat{\pi}_{i,k}\left(\ln\pi_k + \sum_{t=1}^{T} -\frac{1}{2}\ln\sigma_i^2 - \frac{1}{2}\ln 2\pi - \frac{(\varepsilon_{i,t}^k)^2}{2\sigma_i^2}\right)\right), \tag{4.7}$$

where $\hat{\pi}_{i,k}$ denotes the conditional probability that region i belongs to class k. This is equal to

$$\hat{\pi}_{i,k} = \frac{\pi_k \prod_{t=1}^{T}\frac{1}{\sigma_i}\phi\left(\varepsilon_{i,t}^k/\sigma_i\right)}{\sum_{l=1}^{K}\pi_l\prod_{t=1}^{T}\frac{1}{\sigma_i}\phi\left(\varepsilon_{i,t}^l/\sigma_i\right)}. \tag{4.8}$$

In the M-step, we need to maximize (4.7) with respect to the parameters β_k, π_k, $k = 1, \ldots, K$ and α_i, σ_i^2 for $i = 1, \ldots, N$. We perform this maximization step sequentially. First, we optimize over β_k keeping the other parameters fixed. This can be done by a simple weighted regression of $\Delta\log(p_{i,t}) - w_t'\alpha_i$ on $x_{i,t}$ with weights given by $\sqrt{\hat{\pi}_{i,k}}/\sigma_i$. Clearly, we want regions with a larger probability of belonging to class k to have a larger weight in estimating β_k. At the same time, regions with a larger standard deviation of the error term σ_i should get a smaller weight, as their house prices contain relatively more noise and less information about β_k. Each β_k, $k = 1, \ldots, K$ is estimated in a separate weighted regression.

Second, we optimize the log-likelihood function over α_i for $i = 1, \ldots, N$. We do this by regressing $\sum_{k=1}^{K}\hat{\pi}_{i,k}\left[\Delta\log(p_{i,t}) - x_{i,t}\beta_k\right]$ on w_t. The dependent variable in this regression is the conditional expectation of $\eta_{i,t}$. We perform these regressions for each region separately.

Next, the new estimate of σ_i^2 is given by

$$\sigma_i^2 = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}\hat{\pi}_{i,k}\left(\varepsilon_{i,t}^k\right)^2 \tag{4.9}$$

for $i = 1, \ldots, N$. Finally, the mixing proportions are updated by averaging the conditional class membership probabilities, that is,

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N}\hat{\pi}_{i,k} \tag{4.10}$$

for $k = 1, \ldots, K$.

As we maximize over the parameters sequentially in the M-step, we do not reach the optimum of the expected full data log-likelihood function (4.7) in each iteration of the

EM-algorithm. We can repeat the individual update steps until convergence, but this is not necessary. Indeed, Meng and Rubin (1993) have shown that an increase in the full-data log-likelihood function in the M-step is sufficient for the EM algorithm to converge to the maximum of the log-likelihood function.

Determining the appropriate number of latent classes is not straightforward. We cannot use a standard statistical test, due to the Davies (1977) problem of unidentified nuisance parameters under the null hypothesis. The usual approach is using a criterion function balancing the fit and the complexity of the model, where the model fit is measured by the value of the log-likelihood function while the number of model parameters provides a measure of complexity. The most well-known criteria are the Akaike information criterion [AIC] and the Bayesian information criterion [BIC]. Bozdogan (1994) suggests that the AIC should have a penalty factor of 3 instead of 2 in the case of mixture models. Indeed, Andrews and Currim (2003) show that this AIC-3 criterion outperforms other criteria. Bozdogan (1987) modifies the AIC into the so-called consistent Akaike information criterion [CAIC], which is almost equal to BIC. He shows that when the sample size is large the CAIC and BIC criteria perform better than AIC. We will consider all four criteria below.

4.4 Empirical results

In this section we discuss the results of applying our model to the regional house price data for the Netherlands described in Section 4.2. The effective sample period ranges from 1985Q3 (because we have $\Delta \log(p_{t-1}) = \log(p_{t-1}) - \log(p_{t-2})$ in our model) to 2005Q2 (because we only have real GDP data until 2005Q2), giving $T = 80$ data points in the time series dimension. To obtain a first impression of the extent of similarities across regions, we start by estimating a fully heterogenous model allowing for different parameters for each region. Next, we provide estimation results for the model with a limited number of latent classes. Finally, we consider impulse-response functions for three interesting scenarios to provide further interpretation of the model.

4.4.1 A fully heterogenous model

We first estimate the parameters in a fully heterogenous model, that is, we estimate the model in (4.1) with (4.3) allowing for different parameters for each individual region. This

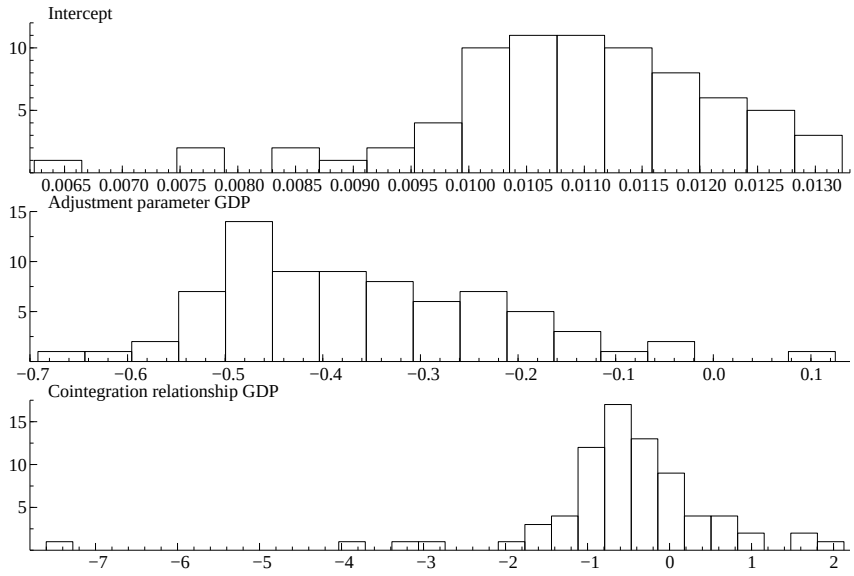


Figure 4.4: Histograms of the estimated values of the parameters β_j, $j = 0, 1$, and γ in (4.1) in the fully heterogenous model with 76 classes.

essentially is a model with K=76 latent classes, in which case each region forms a separate class.

The mean, minimum and maximum of each parameter of the 76 regions can be found in Table 4.2. Figure 4.4 displays the histograms for the 76 estimated values for each of the parameters β_j, $j = 0, 1$, and γ in (4.1). The top panel shows the intercepts, β_0, which equal the quarterly growth rates. These are all positive, reflecting the upward trend in the house prices, and range between 0.6% and 1.3% per quarter. The middle panel of

Table 4.2: Results for the fully heterogenous model.

Parameter	Mean	Minimum	Maximum
β_0	0.011	0.006	0.013
β_1	-0.363	-0.692	0.125
γ	-0.591	-7.601	2.127
α_1	-0.271	-1.851	0.703
α_2	-0.004	-0.013	0.004
α_3	0.277	-0.340	0.739

Figure 4.4 shows the results for the adjustment parameter for the cointegration term with GDP. We find some positive values, which is not as expected, as these imply divergence between GDP and the house prices in that region. Finally, the histogram in the bottom panel shows the parameter γ in the cointegration relationship with GDP, which we expect to be negative as we expect the house prices and GDP to move in the same direction. Table 4.2 also shows the results for the α parameters from (4.3). Again we find that they show some counterintuitive signs and a relatively large spread.

We can see from these results that some form of aggregation may be useful, as we now get a wide variety of parameter estimates, with sometimes quite implausible results. At the same time, this variety also suggests that we should perhaps better not restrict the parameters to be the same across all regions. Hence, it may be optimal to allow for a limited number of different clusters.

4.4.2 A model with latent classes

A major issue for successful application of the latent-class panel time series model is of course determining the appropriate number of latent classes. As discussed in Section 4.3.2, we consider four different information criteria for this purpose. Table 4.3 shows the values of these criteria for models with one to four and 76 classes. For all criteria, we see that going from a homogenous model (with a single class) to two classes amounts to a relatively large improvement in the balance of model fit and complexity. After this, adding more classes does not improve any of the criteria. We therefore focus on the model with two latent classes.

The estimation results for the model with two latent classes are given in Table 4.4. Additionally, Table 4.5 gives the results for a series of Wald tests which we use to examine whether the parameters for the different classes are significantly different from each other.

Table 4.3: Criteria values for different numbers of latent classes (boldface numbers indicate the optimum).

Criterion \ K	1	2	3	4	76
AIC	-3.937	**-4.059**	-4.058	-4.057	-3.962
AIC-3	-3.887	**-4.008**	-4.006	-4.004	-3.862
BIC	-3.598	**-3.716**	-3.710	-3.704	-3.292
CAIC	-3.584	**-3.665**	-3.658	-3.652	-3.192

Table 4.4: Estimation results for $K = 2$ latent classes.

Class	Estimate	Standard error
intercept $\beta_{0,k}$		
1	0.012	0.001
2	0.011	0.000
adjustment parameter GDP $\beta_{1,k}$		
1	-0.178	0.019
2	-0.131	0.007
cointegration relationship GDP γ_k		
1	-1.888	0.083
2	-1.684	0.035
mixing proportions π_k		
1	0.202	0.159
2	0.798	.

The estimation results show that the regions in the two latent classes do indeed differ from each other in several important respects. First, the estimated intercepts show that the average growth rate in class 1 is slightly higher than in class 2[1]. This difference is not significant though, as can be seen from the second row of Table 4.5. The average growth rate in class 1 is equal to 1.2% per quarter, or 4.8% annually, while the house prices in class 2 grow with 1.1% per quarter, or 4.4% annually.

Second, examining the cointegration relationship with GDP, we find that class 1 has a significantly larger adjustment parameter. Thus, the house prices in regions belonging to cluster 1 react faster to changes in GDP than the house prices in class 2.

Finally, The cointegration relationship between house prices and GDP itself, is also significantly different across the classes. For class 1, it is $(1, -1.89)$, meaning that in the long run the house prices in the regions in this cluster grow almost twice as fast as GDP. In class 2 the cointegration relationship is $(1, -1.68)$. These long term relationships may not be very plausible, however, as we could already see from Figure 4.1, they are a good description of the development of house prices and GDP in the sample period.

[1]Recall that we demeaned all other variables the model, so the intercepts represent the average growth rates.

Table 4.5: Wald tests for equality of the parameters across the two classes in (4.1).

Restriction	Wald statistic	p-value
$\beta_{0,1} = \beta_{0,2}$	0.61	0.41
$\beta_{1,1} = \beta_{1,2}$	7.10	0.01
$\gamma_1 = \gamma_2$	6.90	0.01

As we showed in Section 4.3, the cointegration relationship of each region with GDP entails that the regions are also cointegrated among themselves. The long term parameter is only influenced by the γ parameters of the two regions involved, and thus only depends on the class membership of the two regions. Table 4.6 shows these cointegration relationships between the house prices of regions from any of the two clusters. First, we see that two regions that belong to the same cluster are $(1, -1)$ cointegrated. This is actually very intuitive, as they have follow the same trend relative to the trend of GDP, they must follow the exact same trend themselves. Next we find that a region from cluster 1 is $(1, -1.12)$ cointegrated with a region from cluster 2. This corresponds with the slightly higher growth rate in class 1.

The parameters in (4.3) are region-specific, and full estimation results are not reported to save space. Only 11% of the $\alpha_{1,i}$ parameters is significant, suggesting that the impact of GDP on the house prices is mostly captured by the cointegration term. Moreover, only 22% has the expected positive sign. The $\alpha_{2,i}$ parameters are mostly negative, and only one region has an (insignificant) positive value. Furthermore, for 63% of the regions the $\alpha_{2,i}$ parameter is significant at the 5% level, indicating that the interest rate indeed influences the house prices in the expected direction. The $\alpha_{3,i}$ parameters, relating the growth of

Table 4.6: Cointegration relationships between the regions from clusters i and j.

$i \setminus j$	1	2
1	(1, -1)	(1, -1.12)
2	(1, -0.89)	(1, -1))

the house price in a region to growth of the average house price in the Netherlands in the previous quarter, is positive for 88% of the regions, but only significant for 42% of these regions.

The latent classes

The parameter estimation results obviously become more interesting if we know which regions belong to each of the two classes. Therefore, we compute the conditional class membership probabilities using (4.8). The resulting classification of the regions is shown in Figure 4.5. Regions are colored based on $\hat{\pi}_{i,1}$, the probability of belonging to class 1. Regions are colored in four shades of grey. For the regions that are colored in the lightest shade it holds that $\hat{\pi}_{i,1} \leq 0.2$. For regions colored in subsequently darker shades of grey it holds that $0.2 < \hat{\pi}_{i,1} \leq 0.4$, $0.4 < \hat{\pi}_{i,1} \leq 0.6$, or $0.6 < \hat{\pi}_{i,1} \leq 0.8$. There were no regions with $\hat{\pi}_{i,1} > 0.8$. It can be seen that most regions are either very dark or very light, suggesting that the classification is very clear for most regions. In fact, the average value of $\max(\hat{\pi}_{i,1}, \hat{\pi}_{i,2})$ is equal to 0.83.

We find that class 1 contains mainly rural regions surrounding the big cities in the Netherlands. The regions in this class mainly cover parts of *Noord-Brabant* and the *Veluwe*. Even though the East belongs almost completely to class 1, the larger cities of the East, like *Zwolle*, *Almelo*, *Hengelo*, *Enschede*, and *Arnhem* are part of class 2.

Class 2 contains different types of regions. First, it contains many large cities in different parts of the country, like *Breda* and *Groningen*, as well as almost all of the regions in the *Randstad*, the densely populated western part of the country. At the same time some rural regions, like *Zeeland*, *Zuid-Limburg* and regions in the North belong to this class with high probability. Note that these rural regions are not as close to the *Randstad* as most of those in class 1.

A possible explanation for our results is the increased number of commuters that live in the regions belonging to class 1 and who work in the large western cities. If the number of commuters increases, it is likely that they move to regions in cluster 1, as these are still at traveling distance from the *Randstad*. This development has two consequences for the regions in class 1. First, the average income in these regions is likely to increase, as the individuals who move away from the cities are relatively wealthy. The second consequence is an increase of housing quality in these regions, as wealthier people leaving the cities will increase the demand for more luxurious houses.

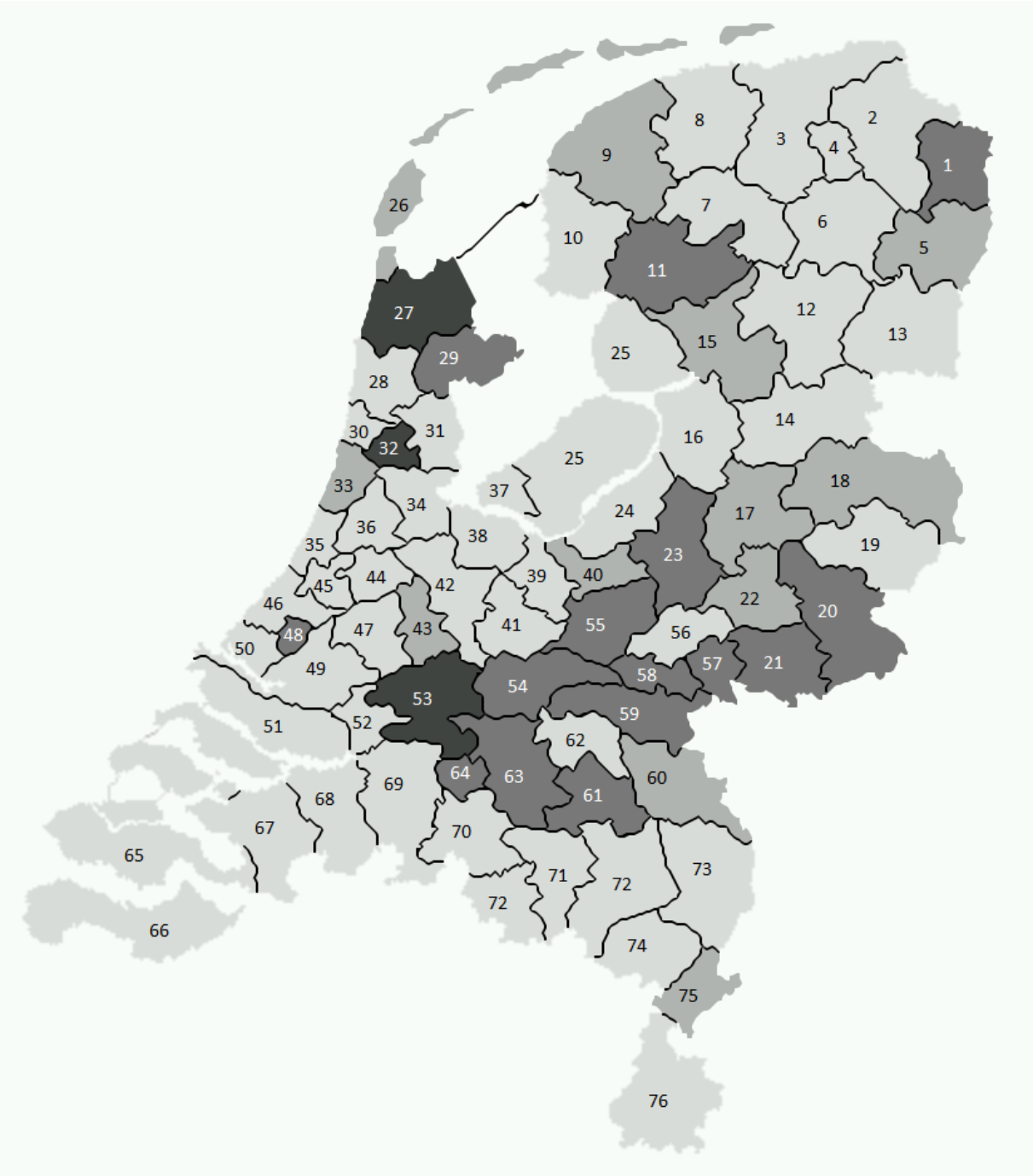


Figure 4.5: Clustering of regions. Regions with a high probability of belonging to class 1 are colored dark, regions with a low probability of belonging to class 1 are colored lighter. The numbers inside the regions correspond to the ones in Appendix 4.A.

These potential structural changes within the regions of cluster 1 are consistent with all of our findings. First, the increase in housing quality will result in a larger increase in the average house prices in class 1 as compared to class 2. Our second finding is that house prices in these regions react faster to changes in GDP. This may be caused by tha fact that the increase of their income may influence the decision of these individuals to move and start commuting. Our last and most striking finding is that the house prices in class 1 increase almost twice as fast as GDP. Note however that the increase is not corrected for higher housing quality.

4.4.3 Impulse-response functions

To give further interpretation to our estimation results we compute impulse-response functions for two interesting scenarios, each occurring in the second quarter of 2005. In the first scenario real GDP receives a shock of 1%. In the second scenario real GDP stays the same, but the interest rate receives a shock of 1%-point. We forecast the house prices for each of the scenarios and compare with a no-change scenario, for the subsequent three-year period from 2005Q3 until 2008Q2.

In order to compute the impulse responses up to 12 quarters ahead, we also need forecasts for GDP and the interest rate, as these variables also affect house prices, see (4.1). Here we assume that the interest rate stays the same during the forecast period. In scenario 3, the interest rate is higher, but still assumed to be constant over the whole forecast period. To obtain forecasts for GDP we construct a simple AR(q) model with intercept for $\Delta \log y_t$. We choose q based on out-of-sample forecasting performance, where we use the last 3 years as a hold-out sample. It turns out that $q = 8$ gives the best performance.

Figure 4.6 shows the impulse-response functions of the log house prices with respect to the log of GDP. The y-axis gives the relative change in house prices between the two scenarios, that is, a value of 0.01 means that the house price is 1% higher than the reference forecast. We calculate the impulse response functions for each of the 76 regions. We then aggregate these to average responses in the two clusters.

We find that the effect of an increase in GDP is initially negative in both clusters, which is caused by the many negative $\alpha_{1,i}$ parameter in both clusters. However, this negative effect lasts only one quarter, and after that the house prices are higher compared to the reference forecasts. As expected, we find that the house prices in cluster 1 react both faster and more on the change in GDP.

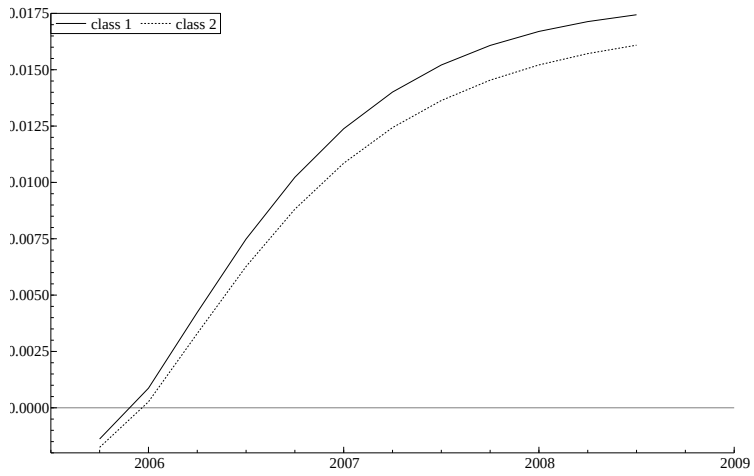


Figure 4.6: Impulse-response function of $\log(p_{i,t})$ with respect to $\log(y_t)$ for 3 regions.

In the second scenario, the interest rate receives a shock, and increases from 2.06% to 3.06%. We find that the house prices are falling. After three years the house prices are about 2% lower in lower in cluster 2 and almost 3% lower in cluster 1, as compared to the reference forecasts.

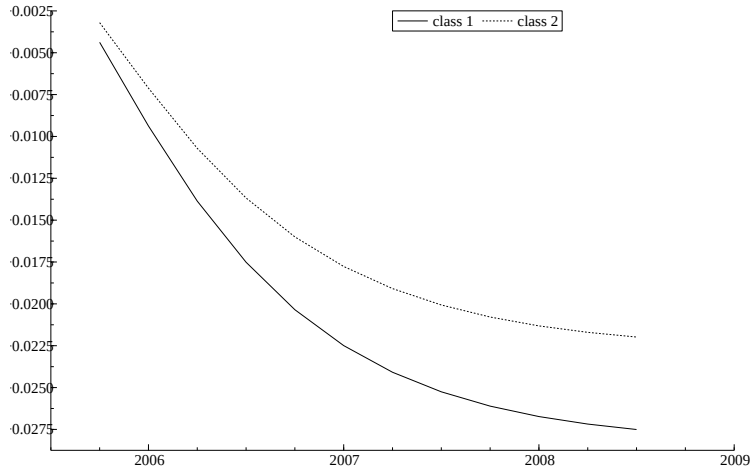


Figure 4.7: Impulse-response function of $\log(p_{i,t})$ with respect to $\log(I_t)$ for 3 regions.

4.5 Conclusions

In this chapter we developed a latent-class panel time series model for describing several key characteristics of regional house prices in the Netherlands between 1985 and 2005. An important feature of the model is that we cluster the regions in separate classes, where the price dynamics of house prices in regions within the same class are similar, while they are different across the classes. For the 76 regions in the Netherlands we find that two classes are sufficient. The first class contains mainly rural regions close to large cities. The second class contains both the larger cities and some more remote rural regions. The house prices in regions in the first class are characterized by slightly higher average growth rates, and stronger and faster reactions to changes in GDP. These findings may be caused by the increased number of commuters. Indeed, the number of people working in the larger cities, but living in the regions of class 1, has increased substantially during our sample period.

Our model allows for the analysis of rather detailed data. To fully exploit its properties one would want to analyze even further disaggregated data. The collection of such more detailed series is left to further research. Another issue for further research is to make the class probabilities dependent on certain explanatory variables.

4.A Regions by number

1	Noordoost-Groningen	27	Kop v. Noord-Holland	52	Dordrecht
2	Slochteren +s	28	Noord-Kennemerland	53	Gorinchem
3	Grootegast +s	29	West-Friesland	54	Culemborg/Dodewaard
4	Stad Groningen +s	30	Midden-Kennemerland	55	Ede +s
5	Zuidoost-Groningen	31	Waterland	56	Arnhem
6	Noord-Drenthe	32	Zaanstreek	57	Duiven/Westervoort
7	Opsterland	33	Zuid-Kennemerland	58	Elst +s
8	Oost-Friesland	34	Amsterdam	59	Nijmegen
9	Noordwest-Friesland	35	De Bollenstreek	60	Noordoost-Brabant
10	Zuidwest-Friesland	36	Haarlemmermeer	61	Uden +s
11	Zuid-Friesland	37	Almere	62	Oss +s
12	Zuidwest-Drenthe	38	Het Gooi	63	Den Bosch
13	Zuidoost-Drenthe	39	Amersfoort	64	Waalwijk/Drunen
14	Hardenberg +s	40	Barneveld	65	Zeeuwse Eilanden
15	Kop van Overijssel	41	Bunnik/Zeist	66	Zeeuws-Vlaanderen
16	Zwolle +s	42	Utrecht	67	Bergen op Zoom +s
17	Raalte +s	43	Woerden	68	West-Brabant
18	Almelo Tubbergen	44	Alphen	69	Breda
19	Hengelo Enschede	45	Leiden	70	Tilburg/Oirschot
20	Ruurlo Eibergen	46	Den Haag	71	Eindhoven +s
21	Doetinchem +s	47	Gouda	72	Zuidoost-Brabant
22	Zutphen +s	48	Delft +s	73	Noord-Limburg
23	Apeldoorn +s	49	Rotterdam	74	Weert +s
24	Nunspeet +s	50	Westland	75	Roermond +s
25	Lelystad	51	Brielle/Goeree	76	Zuid-Limburg
26	Den Helder/Texel				

Note: +s means including surrounding area.

Chapter 5

A Bayesian Approach to Two-Mode Clustering

5.1 Introduction

Clustering algorithms divide a single set of objects into segments based on their similarities and properties or the dissimilarities between them (see, for example, Hartigan, 1975). Such methods typically operate on one mode (dimension) of a data matrix; we refer to these methods as one-mode clustering. Two-mode clustering techniques (Van Mechelen *et al.*, 2004) cluster two sets of objects into segments based on their interactions. In two-mode clustering, both rows and columns of data matrix are clustered simultaneously.

Many clustering methods, such as k-means clustering and Ward's method, lack a method to ascertain the significance of the results and rely on arbitrary methods to determine the number of clusters. To solve these problems, one may consider using model-based techniques for clustering data. For one-mode data, model-based clustering methods have been developed (see, for example, Fraley and Raftery, 1998; Frühwirth-Schnatter, 2006). These model-based clustering methods use statistical tools for inference.

In this chapter, we extend the model-based one-mode clustering approach to two-mode clustering. In two-mode clustering, we cluster both the rows and the columns of a data matrix into groups in such a way that the resulting block structure is homogenous within blocks but differs between blocks. This requires matrix-conditional data, which basically means that each element should be measured on the same scale. Methods for two-mode clustering are in general not model-based, see among others, Candel and Maris (1997), Doreian *et al.* (2004), Brusco and Steinley (2006), and Van Rosmalen *et al.* (2008).

Model-based methods usually rely on latent-class techniques. It is not straightforward to extend these techniques to two-mode data, because, unlike one-mode data, two-mode data cannot be assumed to be independent. Despite this problem, Govaert and Nadif (2003, 2008) have been able to use a latent-class approach to cluster two-mode data. They use a frequentist approach to estimate the parameters, but they are only able to optimize an approximation of their likelihood function using the EM algorithm (Dempster *et al.*, 1977). In this chapter we use the same model as Govaert and Nadif (2003, 2008). We, however, propose a Bayesian estimation procedure. This enables us to estimate the model parameters properly and to do statistical inference on the results.

The contribution of our Bayesian approach is threefold. First, our approach allows for statistical inference on the parameter estimates. Govaert and Nadif (2003, 2008) estimate the model parameters in a frequentist setting, but they are unable to compute standard errors of the estimated parameters. Our method provides posterior standard deviations and posterior distributions of the parameters. Therefore, our approach enables hypothesis testing, which is not feasible in the frequentist setting.

Second, our Bayesian method has fewer computational problems than the maximum likelihood approach. By using proper priors, we avoid the computational issues with empty classes which is a known problem with the EM algorithm. Furthermore, because of the more flexible way Markov Chain Monte Carlo methods search the parameter space, our Bayesian algorithm is less likely to get stuck in a local optimum of the likelihood function. This flexibility may, however, cause *label switching*, see Celeux *et al.* (2000). Geweke (2007) has however shown that this problem can be solved.

Finally, our method can help indicate the optimal number of segments. The Bayesian approach can be used to derive selection criteria such as *Bayes factors*. Methods previously proposed in the literature for selecting the optimal number of clusters (see, for example, Schepers *et al.*, 2008) seem somewhat arbitrary and lack theoretical underpinnings.

We illustrate our Bayesian approach on two data sets. The first is a data set of Supreme Court voting and was also used by Doreian *et al.* (2004) and Brusco and Steinley (2006). Our approach results in a similar solution; however, the optimal numbers of segments are lower than in either of their solutions. Our second application is a large data set concerning roll call voting in the United States House of Representatives. The individual votes data are available at `http://www.GovTrack.us`. We use our model to cluster both the representatives and the bills simultaneously.

The remainder of this chapter is organized as follows. In Section 5.2 we introduce our new Bayesian approach for clustering two-mode data. We compare it with the existing frequentist approaches of Govaert and Nadif (2003, 2008). In Section 5.3 we discuss the posterior simulator for our Bayesian approach together with the selection of the number of segments. The Bayesian approach is illustrated on the Supreme Court voting data in Section 5.4. Section 5.5 deals with our second application, which concerns roll call votes of the United States House of Representatives in 2007. Finally, Section 5.6 concludes.

5.2 The latent-class two-mode clustering model

In this section, we present our Bayesian approach to clustering both modes of two-mode data simultaneously. We first give a derivation of the likelihood function and then discuss Bayesian parameter estimation for the latent-class two-mode clustering model.

5.2.1 The likelihood function

For illustrative purposes, we start this discussion with one-mode data, that is, we have N observations denoted by $y = (y_1, \ldots, y_N)'$. These observations can be discrete or continuous, and one-dimensional or multidimensional. We assume that each observation comes from one of K segments, and that the elements within each segment are independently and identically distributed. As a result, all observations must be independent. Furthermore, we assume that the observations come from a known distribution that is the same for all segments; only the parameters of the distribution vary over the segments. This can be described by a mixture model. Let $k_i \in \{1, \ldots, K\}$ be an indicator for the segment to which observation y_i belongs, and let $k = (k_1, \ldots, k_N)'$. Then, the likelihood of observing y_i belonging to segment q is

$$f(y_i|k_i = q) = g(y_i|\theta_q), \tag{5.1}$$

where $g(y_i|\theta_q)$ is the density function of y_i in segment q with parameter vector θ_q. The segment membership is unknown. We assume that the probability that observation y_i belongs to segment q is given by κ_q for $q = 1, \ldots, K$, with $\kappa_q > 0$ and $\sum_{q=1}^{K} \kappa_q = 1$. We collect the so-called mixing proportions κ_q in the vector $\kappa = (\kappa_1, \ldots, \kappa_K)'$. The likelihood function of this model is given by

$$l(y|\theta, \kappa) = \prod_{i=1}^{N} \left\{ \sum_{q=1}^{K} \kappa_q g(y_i|\theta_q) \right\}, \tag{5.2}$$

where $\theta = (\theta_1, \ldots, \theta_K)'$.

To cluster two-mode data, we would like to extend (5.2) to two-mode data matrices, with a simultaneous clustering of both rows and columns. We aim to construct a model in which the observations that belong to the same row cluster and the same column cluster are independently and identically distributed. In two-mode clustering, unlike in one-mode clustering, this assumption does not ensure that all observations are independent. As a result, a straightforward extension of the one-mode likelihood function to two modes will not adequately model the dependence structure in the data.

Assume that Y is an $(N \times M)$ matrix, and that we want to cluster the rows into K latent classes and the columns into L latent classes. The obvious natural extension of (5.2) to two-mode data yields

$$l(Y|\theta, \kappa, \lambda) = \prod_{i=1}^{N}\prod_{j=1}^{M}\sum_{q=1}^{K}\sum_{r=1}^{L}\kappa_q\lambda_r g(Y_{i,j}|\theta_{q,r}), \tag{5.3}$$

where $\kappa = (\kappa_1, \ldots, \kappa_K)'$ gives the size of each row segment, $\lambda = (\lambda_1, \ldots, \lambda_L)'$ gives the size of each column segment, and $\theta_{q,r}$ contains the parameters of observations in row segment q and column segment r. Model (5.3) fails to impose that all elements in a row belong to the same row cluster and also does not impose that all elements in a column belong to the same column cluster; using this model, the data matrix Y would effectively be modeled as a vector of one-mode data.

To solve this problem, we first rewrite the one-mode likelihood function (5.2) as

$$\begin{aligned}
l(y|\theta, \kappa) &= \prod_{i}^{N}\left\{\sum_{q=1}^{K}\kappa_q g(y_i|\theta_q)\right\} \\
&= \left\{\sum_{q=1}^{K}\kappa_q g(y_1|\theta_q)\right\}\left\{\sum_{q=1}^{K}\kappa_q g(y_2|\theta_q)\right\}\ldots\left\{\sum_{q=1}^{K}\kappa_q g(y_N|\theta_q)\right\} \\
&= \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\cdots\sum_{k_N=1}^{K}\prod_{i=1}^{N}\kappa_{k_i} g(y_i|\theta_{k_i}) \\
&= \sum_{k\in\mathcal{K}}\prod_{q=1}^{K}\kappa_q^{N_k^q}\prod_{i=1}^{N} g(y_i|\theta_{k_i}),
\end{aligned} \tag{5.4}$$

where we introduce some new notation in the last line. First, the set $\mathcal{K}$ contains all possible divisions of items into the segments, and thus has K^N elements if there are N items and K possible segments. Second, N_k^q equals the number of items belonging to

segment q according to segmentation k. Thus, $\sum_{q=1}^{K} N_k^q = N$ for a fixed classification k. The fact that these two representations of the likelihood function of a mixture model are equivalent was already noticed by Symons (1981).

Using this representation, we can extend the mixture model to clustering two modes simultaneously. The resulting likelihood function equals

$$l(Y|\theta,\kappa,\lambda) = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}}\prod_{q=1}^{K}\kappa_q^{N_k^q}\prod_{r=1}^{L}\lambda_r^{M_l^r}\prod_{i=1}^{N}\prod_{j=1}^{M} g(Y_{i,j}|\theta_{k_i,l_j}), \tag{5.5}$$

where $\mathcal{L}$ denotes all possible division of the columns into L segments, M_l^r equals the number of items belonging to segment r according to column segmentation $l = (l_1,\ldots,l_M)'$. Note that it is impossible to rewrite (5.5) as a product of likelihood contributions as in the one-dimensional case (5.2).

5.2.2 Parameter estimation

The likelihood function (5.5) was already proposed by Govaert and Nadif (2003), who estimate the parameters of this model in a frequentist setting. However, this approach has several limitations. First, unlike in the one-mode case, the likelihood function (5.5) cannot be written as a product over marginal/conditional likelihood contributions. As Govaert and Nadif (2003) already notice, we only have a sample of size 1 from the joint distribution of Y, k, and l. Hence, given the sample size of 1, the standard results for the asymptotic properties of the maximum likelihood estimator do not apply here.

Second, standard approaches to maximize the likelihood function (5.5) and estimate the model parameters are almost always computationally infeasible. Enumerating the $K^N L^M$ possible ways to assign the rows and columns to clusters in every iteration of an optimization routine is only possible for extremely small data sets. To solve this problem, Govaert and Nadif (2003) instead consider the so-classed classification likelihood approach, in which k and l are parameters that need to be optimized. Hence one maximizes

$$l(Y|\theta,\kappa,\lambda,k,l) = \prod_{q=1}^{K}\kappa_q^{N_k^q}\prod_{r=1}^{L}\lambda_r^{M_l^r}\prod_{i=1}^{N}\prod_{j=1}^{M} g(Y_{i,j}|\theta_{k_i,l_j}) \tag{5.6}$$

with respect to θ, κ, λ, $k \in \mathcal{K}$, and $l \in \mathcal{L}$. As the parameter space contains discrete parameters k and l, standard asymptotic theory for maximum likelihood does not hold. Govaert and Nadif (2008) also consider the optimization of an approximation to the likelihood function (5.5). This approximation is based on the assumption that the two

classifications (that is, the classification of the rows and the classification of the columns) are independent.

We solve the aforementioned problems by considering a Bayesian approach. This has several advantages. The first advantage is that we do not have to rely on asymptotic theory for inference. We can use the posterior distribution to do inference on the model parameters. Furthermore, it turns out that we do not need to evaluate the the likelihood specification (5.5) to obtain posterior results. Posterior results can easily be obtained using a Markov Chain Monte Carlo [MCMC] sampler with data augmentation (Tanner and Wong, 1987). Data augmentation means that the latent variables (in our case k and l) are simulated alongside the model parameters θ, κ, and λ. This amounts to applying the Gibbs sampler to the complete data likelihood in (5.6). As Tanner and Wong (1987) show, posterior results from the complete data likelihood function are the same as posterior results from the likelihood function. As we can rely on the complete data likelihood (5.6) and do not have to consider (5.5), obtaining posterior results is computationally feasible. Furthermore, unlike previous authors (see, for example, Govaert and Nadif, 2003, 2008), we can provide statistical rules for choosing the numbers of segments, which we will do in Section 5.3.2. Finally, our method does not suffer as much from computational difficulties when searching the global optimum of the likelihood function. The EM algorithm is known to get stuck in local optima of the likelihood function, which often occurs in local optima with one or more empty segments. Because we rely on MCMC methods, our approach has fewer problems with local optima. Furthermore, using proper priors, we can avoid solutions with empty segments, see also Dias and Wedel (2004) for similar arguments.

5.3 Posterior simulator

As discussed previously, we rely on MCMC methods to estimate the posterior distributions of the parameters of the two-mode mixture model. We propose a Gibbs sampler (Geman and Geman, 1984) with data augmentation (Tanner and Wong, 1987), in which we sample the vectors k and l alongside the model parameters. This approach allows us to sample from the posterior distributions of the parameters without evaluating the full likelihood function and therefore requires limited computation time. We assume independent priors for the model parameters with density functions $f(\kappa)$, $f(\lambda)$, and $f(\theta)$. In Section 5.3.1, we derive the Gibbs sampler. The choice of the number of segments is discussed in Section 5.3.2.

5.3.1 The Gibbs sampler

In each iteration of the Gibbs sampler, we need to draw the parameters θ, κ, and λ together with the latent variables k and l from their full conditional distributions. The MCMC simulation scheme is as follows

- Draw $\kappa, \lambda | \theta, k, l, Y$
- Draw $k | \kappa, \lambda, \theta, l, Y$
- Draw $l | \kappa, \lambda, \theta, k, Y$
- Draw $\theta | \kappa, \lambda, k, l, Y$

Below we derive the full conditional posteriors, which are needed for the Gibbs sampler. The Gibbs sampler results in a series of draws from the posterior distributions of the parameters θ, κ, and λ. These draws can be used to compute posterior means, posterior standard deviations, and highest posterior density regions. Because we use data augmentation, we also obtain draws from the posterior distributions of k and l. This enables us to compute the posterior distributions for each row and column of data over the segments. We can store the posterior distributions in matrices Q and R, where Q is of size $(N \times K)$, and R is of size $(M \times L)$. Each row of Q contains the posterior distribution of a row of data over the K possible row segments, and each row of R contains the posterior distribution of a column of data over the L possible column segments.

Sampling of κ and λ

The full conditional density of κ is given by

$$\begin{aligned} f(\kappa | \theta, \lambda, k, l, Y) &\propto l(Y | \theta, \kappa, \lambda, k, l) f(\kappa) \\ &\propto \prod_{q=1}^{K} \kappa_q^{\sum_{i=1}^{N} I(k_i = q)} f(\kappa), \end{aligned} \tag{5.7}$$

where $f(\kappa)$ is the prior density of κ, and $I(.)$ is an indicator function that equals 1 if the argument is true and 0 otherwise. The first part of (5.7) is the kernel of a Dirichlet distribution (see, for example, Frühwirth-Schnatter, 2006). If we specify a Dirichlet$(d_1, d_2, \ldots, d_K)$ prior distribution for κ, the full conditional posterior is also a Dirichlet distribution with parameters $\sum_{i=1}^{N} I(k_i = 1) + d_1$, $\sum_{i=1}^{N} I(k_i = 2) + d_2$, …, $\sum_{i=1}^{N} I(k_i = K) + d_K$.

If we take a Dirichlet$(d_1, d_2, \ldots, d_L)$ prior for λ, the λ parameters can be sampled in exactly the same way. The full conditional posterior density is now given by

$$f(\lambda|\theta, \kappa, k, l, Y) \propto \prod_{r=1}^{L} \lambda_r^{\sum_{j=1}^{M} I(l_j=r)} f(\lambda), \tag{5.8}$$

where $f(\lambda)$ denotes the prior density. Hence, we can sample λ from a Dirichlet distribution with parameters $\sum_{j=1}^{M} I(l_j = 1) + d_1, \sum_{j=1}^{M} I(l_j = 2) + d_2, \ldots, \sum_{j=1}^{M} I(l_j = L) + d_L$.

Sampling of k and l

We sample each element of k and l separately. The full conditional density of k_i is given by

$$\begin{aligned} p(k_i|\theta, \kappa, \lambda, k_{-i}, l, Y) &\propto p(Y_i|k_i, \theta, \kappa, \lambda, l) \\ &\propto \kappa_{k_i} \prod_{j=1}^{M} g(Y_{i,j}|\theta_{k_i,l_j}), \end{aligned} \tag{5.9}$$

for $k_i = 1, \ldots, K$, where k_{-i} denotes k without k_i, and Y_i denotes the ith row of data from Y. Hence, k_i can be sampled from a multinomial distribution for $i = 1, \ldots, N$. In a similar way we can derive the full conditional density of l_j which equals

$$p(l_j|\theta, \kappa, \lambda, k, l_{-j}, Y) \propto \lambda_{l_j} \prod_{i=1}^{N} g(Y_{i,j}|\theta_{k_i,l_j}), \tag{5.10}$$

for $l_j = 1, \ldots, L$, where l_{-j} denotes l without l_j. We can sample l_j from a multinomial distribution.

Sampling of θ if Y has a Bernoulli distribution

The sampling of the θ parameters depends on the specification of $g(Y_{i,j}|\theta_{q,r})$. With our application in mind we now assume that $Y_{i,j}$ is a binary random variable with a Bernoulli distribution with probability $p_{q,r}$ for an element in row segment q and column segment r, that is,

$$g(Y_{i,j}|\theta_{q,r}) = Y_{i,j}^{p_{q,r}} (1 - Y_{i,j})^{1-p_{q,r}}. \tag{5.11}$$

Let P denote the $(K \times L)$ matrix containing these probabilities for each combination of a row segment and a column segment, so that $\theta = P$.

To sample $p_{q,r}$, we need to derive its full conditional density, which is given by

$$\begin{aligned} f(p_{q,r}|P_{-q,r},\kappa,\lambda,k,l,Y) & \\ &\propto f(Y|p_{q,r},P_{-q,r},\kappa,\lambda,k,l)f(p_{q,r}) \\ &\propto \prod_{i\in\mathcal{Q}}\prod_{j\in\mathcal{R}} p_{q,r}^{Y_{i,j}}(1-p_{q,r})^{1-Y_{i,j}} f(p_{q,r}) \\ &\propto p_{q,r}^{\sum_{i=1}^{N}\sum_{j=1}^{M} I(k_i=q)I(l_j=r)Y_{i,j}}(1-p_{q,r})^{\sum_{i=1}^{N}\sum_{j=1}^{M} I(k_i=q)I(l_j=r)(1-Y_{i,j})} f(p_{q,r}), \end{aligned} \quad (5.12)$$

where $\mathcal{Q}$ is the set containing all rows that belong to segment q, $\mathcal{R}$ contains all columns that belong to segment r, $P_{-q,r}$ denotes P without $p_{q,r}$ and $f(p_{q,r})$ denotes the prior density of $p_{q,r}$. The first part of (5.12) is the kernel of a beta distribution. If we specify a Beta(b_1, b_2) prior distribution, the full conditional posterior distribution is also a beta distribution with parameters $\sum_{i=1}^{N}\sum_{j=1}^{M} I(k_i = q)I(l_j = r)Y_{i,j} + b_1$ and $\sum_{i=1}^{N}\sum_{j=1}^{M} I(k_i = q)I(l_j = r)(1 - Y_{i,j}) + b_2$.

Sampling of θ if Y has a Normal distribution

If $Y_{i,j}$ is a normally distributed variable, with mean $\mu_{q,r}$ and variance $\sigma^2_{q,r}$ in row segment q and column segment r, then

$$g(Y_{i,j}|\theta_{q,r}) = \frac{1}{\sqrt{2\pi\sigma^2_{q,r}}} \exp\left\{-\frac{1}{2}\frac{(Y_{i,j}-\mu_{q,r})^2}{\sigma^2_{q,r}}\right\}. \quad (5.13)$$

Let μ and Σ denote the $(K \times L)$ matrices containing the means and variances, respectively, for each combination of a row segment and a column segment; hence $\theta = \{\mu, \Sigma\}$.

To sample $\mu_{q,r}$, we need to derive its full conditional distribution, which is given by

$$\begin{aligned} f(\mu_{q,r}|\mu_{-q,r},\Sigma,\kappa,\lambda,k,l,Y) & \\ &\propto f(Y|\mu_{q,r},\mu_{-q,r},\Sigma,\kappa,\lambda,k,l)f(\mu_{q,r}) \\ &\propto \exp\left[-\frac{\sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{R}}(Y_{i,j}-\mu_{q,r})^2}{2\sigma^2_{q,r}}\right] f(\mu_{q,r}) \\ &\propto \exp\left[-\frac{(\mu_{q,r}-1/N^{q,r}_{k,l}\sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{R}}Y_{i,j})^2}{2\sigma^2_{q,r}/N^{q,r}_{k,l}}\right] f(\mu_{q,r}), \end{aligned} \quad (5.14)$$

where $\mu_{-q,r}$ denotes μ without $\mu_{q,r}$, $f(\mu_{q,r})$ denotes the prior of $\mu_{q,r}$, and $N^{q,r}_{k,l}$ denotes the number of elements that are both in row segment q and column segment r according to k and l and is equal to $\sum_{i=1}^{N}\sum_{j=1}^{M} I(k_i = q)I(l_j = r)$. As some segments may become

empty in one of the iterations, we propose to use a proper prior. In this case, the first part of (5.14) is the kernel of a normal distribution; therefore, using a normal prior is an obvious choice. Let the prior distribution have mean μ_0 and variance σ_0^2. Then, the full conditional posterior distribution is also normal with

$$f(\mu_{q,r}|\mu_{-q,r},\Sigma,\kappa,\lambda,k,l,Y) \sim \mathcal{N}\left(\frac{\sigma_0^{-2}}{\sigma_0^{-2}+s^{-2}}\mu_0 + \frac{s^{-2}}{\sigma_0^{-2}+s^{-2}}\bar{\mu}, (\sigma_0^{-2}+s^{-2})^{-1}\right), \quad (5.15)$$

where $\bar{\mu} = 1/N_{k,l}^{q,r}\sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{R}}Y_{i,j}$, the sample average within the cluster, and $s^2 = \sigma_{q,r}^2/N_{k,l}^{q,r}$.

The full conditional distribution of $\sigma_{q,r}^2$ is given by

$$\begin{aligned} f(\sigma_{q,r}^2|\mu,\Sigma_{-q,r},\kappa,\lambda,k,l,Y) & \\ \propto\ & f(Y|\mu,\sigma_{q,r}^2,\Sigma_{-q,r},\kappa,\lambda,k,l)f(\sigma_{q,r}^2) \\ \propto\ & (\sigma_{q,r}^2)^{N_{k,l}^{q,r}/2}\exp\left[-\frac{\sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{R}}(Y_{i,j}-\mu_{q,r})^2}{2\sigma_{q,r}^2}\right]f(\sigma_{q,r}^2), \end{aligned} \quad (5.16)$$

where $\Sigma_{-q,r}$ denotes Σ without $\sigma_{q,r}^2$, and $f(\sigma_{q,r}^2)$ denotes the prior of $\sigma_{q,r}^2$. The first part of (5.16) is the kernel of an inverted gamma-2 distribution. Again, using a proper prior may be advisable. If we specify an inverted gamma-2 prior with parameters g_1 and g_2, the full conditional posterior has an inverted gamma-2 distribution with parameters $N_{k,l}^{q,r}+g_1$ and $\sum_{i\in\mathcal{Q}}\sum_{j\in\mathcal{R}}(Y_{i,j}-\mu_{q,r})^2+g_2$. The prior essentially adds g_1 observations with an average variance of g_2/g_1 to each combination of a row segment and a column segment.

Note that if $Y_{i,j}$ has neither a Bernoulli nor a Normal distribution, we only have to adjust the term $g(Y_{i,j}|\theta_{k_i,l_j})$ for sampling k and l and the sampling θ in the appropriate manner.

5.3.2 Selecting the numbers of segments

The usual way to determine the numbers of clusters in a finite mixture model in a frequentist framework is to use information criteria such as AIC, AIC-3, BIC, and CAIC (see, for example, Fraley and Raftery, 1998; Andrews and Currim, 2003). The reason for this is that standard tests for determining the optimal number of classes in latent-class models are not valid due to the Davies (1977) problem. Within a Bayesian framework, we can avoid this problem by computing Bayes factors (see, for example, Berger, 1985;

Kass and Raftery, 1995; Han and Carlin, 2001). Unlike hypothesis tests, Bayes factors can be used to compare several possibly nonnested models simultaneously; Bayes factors naturally penalize complex models. The Bayes factor for comparing Model 1 with Model 2 is defined as

$$B_{21} = \frac{f(Y|M_2)}{f(Y|M_1)}, \tag{5.17}$$

where $f(Y|M_i)$ denotes the marginal likelihood of model M_i, $i = 1, 2$.

Computing the value of the marginal likelihood is not an easy task. Theoretically, its value can be estimated by averaging the likelihood function over draws from the prior distribution. If the support of the prior distribution does not completely match with the support of the likelihood function, the resulting estimate will be very poor. Another strategy is to use the harmonic mean estimator of Newton and Raftery (1994). However, this estimator is often quite unstable. In this chapter, we estimate the marginal likelihood using the fourth estimator proposed by Newton and Raftery (1994, p. 22), which is also used by DeSarbo *et al.* (2004) in a similar model. This estimator uses importance sampling to compute the marginal likelihood value. The importance sampling function is a mixture of the prior and the posterior distribution with mixing proportion δ. Using the fact that the marginal likelihood is the expected value of the likelihood function with respect to the prior, one can show that the marginal likelihood $f(Y)$ can be estimated using the iterative formula

$$\widehat{f(Y)} = \frac{\delta m/(1-\delta) + \sum_{i=1}^{m} (f(Y|\vartheta^{(i)})/(\delta \widehat{f(Y)} + (1-\delta) f(Y|\vartheta^{(i)})))}{\delta m/(1-\delta)\widehat{f(Y)} + \sum_{i=1}^{m} (\delta \widehat{f(Y)} + (1-\delta) f(Y|\vartheta^{(i)}))^{-1}}, \tag{5.18}$$

where m denotes the number of draws $\vartheta^{(i)}$ from the posterior distribution and where we drop the model M for notational convenience. To apply the formula, we need to choose the value δ; Newton and Raftery (1994) recommend using a low value of δ, which we set to 0.001 in our application below.

Obtaining an accurate value of the marginal likelihood for any moderately sophisticated model tends to be hard, as was noted by Han and Carlin (2001). Therefore, we also propose a simpler alternative method to choose the numbers of segments, based on information criteria. Simulations in Andrews and Currim (2003) suggest that the AIC-3 of Bozdogan (1994) performs well as a criterion for selecting numbers of segments. To evaluate the AIC-3, we need the maximum likelihood value and the number of parameters. To compute the maximum likelihood value, we take the highest value of the likelihood function (5.6) across the sampled parameters.

Determining the number of parameters is not as straightforward as usual. The parameters θ, κ, and λ contain wKL, $K-1$, and $L-1$ parameters, respectively, where w denotes the number of parameters in θ per cluster. However, even though k and l contain the same numbers of parameters for all numbers of latent classes, the number of possible values for each parameter increases. We can think of k as an $(N \times K)$ indicator matrix, where each row indicates to which segment an object belongs. This means that k and l contain $N(K-1)$ and $M(L-1)$ free parameters, respectively. Hence, the effective total number of parameters is $wKL + NK + ML + K + L - M - N - 2$.

5.4 Application 1: Supreme Court voting data

We apply the latent-class two-mode clustering model to two empirical data sets. The first data set, which is discussed in this section, is the Supreme Court voting data of Doreian *et al.* (2004). We use this data set to compare the results of our approach with the results of previous authors, and we discuss this data set relatively briefly. The second data set will be analyzed in greater detail in the next section. The Supreme Court voting data set comprises the decisions of the nine Justices of the United States Supreme Court on 26 important issues. The data are displayed in Table 5.1. In this table, a 1 reflects that the Justice voted with the majority, and a 0 means that the Justice voted with the minority.

To describe the votes, we use a Bernoulli distribution with a Beta$(1,1)$ prior for the probability, which is equivalent to a uniform prior on (0,1). Furthermore, we use an uninformative Dirichlet$(1,1,\ldots,1)$ prior for both κ and λ. To determine the optimal numbers of segments, we compute the marginal likelihoods for several values of K and L. Table 5.2 display the values of $\ln f(Y)$ for each combination of $K = 1,\ldots,6$ rows segments and $L = 1,\ldots,6$ column segments. The highest marginal likelihood is achieved with $K = 2$ segments for the issues and $L = 3$ segments for the Justices.

To analyze the results, it is possible to weight the results with different numbers of segments according to the posterior model probabilities that follow from the marginal likelihoods. However, we find it more convenient to consider the results for only one value of K and L. Therefore, we focus on the solution with the highest marginal likelihood value, that is, $K = 2$ segments of issues and $L = 3$ segments of Justices. Note that we find fewer segments than Doreian *et al.* (2004), who applied blockmodeling to this data set and found 7 clusters for the issues and 4 clusters for the Justices. Brusco and Steinley (2006) ended up with 5 clusters for the issues and 3 clusters for the Justices.

Table 5.1: The Supreme Court Voting Data

	Supreme Court Justice										Supreme Court Justice								
Issue	Br	Gi	So	St	OC	Ke	Re	Sc	Th	Issue	Br	Gi	So	St	OC	Ke	Re	Sc	Th
2000 Presidential Election	0	0	0	0	1	1	1	1	1	Title VI Disabilities	0	0	0	0	1	1	1	1	1
Illegal Search 1	1	1	1	1	1	1	0	0	0	PGA vs. Handicapped	1	1	1	1	1	1	1	0	0
Illegal Search 2	1	1	1	1	1	1	0	0	0	Immigration Jurisdiction	1	1	1	1	0	1	0	0	0
Illegal Search 3	1	1	1	0	0	0	0	1	1	Deporting Criminal Aliens	1	1	1	1	1	0	0	0	0
Seat Belts	0	0	1	0	0	1	1	1	1	Detaining Criminal Aliens	1	1	1	1	0	1	0	0	0
Stay of Execution	1	1	1	1	1	1	0	0	0	Citizenship	0	0	0	1	0	1	1	1	1
Federalism	0	0	0	0	1	1	1	1	1	Legal Aid for the Poor	1	1	1	1	0	1	0	0	0
Clean Air Act	1	1	1	1	1	1	1	1	1	Privacy	1	1	1	1	1	1	0	0	0
Clean Water	0	0	0	0	1	1	1	1	1	Free Speech	1	0	0	0	1	1	1	1	1
Cannabis for Health	0	1	1	1	1	1	1	1	1	Campaign Finance	1	1	1	1	1	0	0	0	0
United Foods	0	0	1	1	0	1	1	1	1	Tobacco Ads	0	0	0	0	1	1	1	1	1
New York Times Copyright	0	1	1	0	1	1	1	1	1	Labor Rights	0	0	0	0	1	1	1	1	1
Voting Rights	1	1	1	1	1	0	0	0	0	Property Rights	0	0	0	0	1	1	1	1	1

Table 5.2: Log marginal likelihoods for the Supreme Court Voting Data

	Segments of Justices					
Segments of issues	1	2	3	4	5	6
1	-155.56	-160.96	-164.17	-166.28	-168.13	-169.96
2	-172.98	-117.69	**-109.98**	-111.00	-112.27	-114.02
3	-183.29	-124.17	-112.17	-113.83	-114.63	-116.56
4	-189.60	-129.92	-115.57	-117.72	-119.27	-118.61
5	-194.88	-133.45	-118.78	-120.00	-121.50	-121.51
6	-199.67	-137.27	-122.29	-122.66	-125.84	-126.28

We experience *label switching* in our MCMC sampler. Two of the segments of Justices switched places twice in our chain of 100,000 draws. However, we could easily identify where the switching took place. As suggested by Geweke (2007), we solved the *label switching* problem by sorting the draws in an appropriate way.

The posterior means and standard deviations of P, κ, and λ are shown in Table 5.3. Tables 5.4 and 5.5 show the marginal posterior distributions of the issues and the Justices over the segments. We find that Justices *Ginsburg*, *Stevens*, *Breyer*, and *Souter* constitute the liberal wing (that is, the left wing) of the Supreme Court. The Court's moderate wing comprises Justices *O'Connor* and *Kennedy*, and the conservative wing (that is, the right wing) consists of Justices *Rehnquist*, *Scalia*, and *Thomas*. The segments of the issues consist of issues that resulted in liberal decisions (segment 1) and issues that resulted in conservative decisions (segment 2). We find strong partisan tendencies in the Supreme Court: liberal Justices support liberal decisions in 97% of the cases, and conservative

Table 5.3: Mean posterior results, with posterior standard deviations in parentheses, for $K = 2$ and $L = 3$ in the Supreme Court Data.

	Segment of Justices			
Segment of issues	1	2	3	Posterior segment size
Interpretation	liberal	moderate	conservative	
1 (liberal majority)	0.97 (0.03)	0.68 (0.10)	0.10 (0.07)	0.46 (0.10)
2 (conservative majority)	0.26 (0.07)	0.84 (0.07)	0.97 (0.03)	0.54 (0.10)
Posterior segment size	0.42 (0.14)	0.25 (0.12)	0.33 (0.13)	

Table 5.4: Marginal posterior distribution of the Justices over the segments.

Justice	1	2	3
Breyer	1.00	0.00	0.00
Ginsburg	1.00	0.00	0.00
Stevens	1.00	0.00	0.00
Souter	1.00	0.00	0.00
O'Connor	0.00	1.00	0.00
Kennedy	0.00	0.98	0.02
Rehnquist	0.00	0.00	1.00
Thomas	0.00	0.00	1.00
Scalia	0.00	0.00	1.00
Interpretation	liberal	moderate	conservative

Justices also support conservative decisions with a 97% probability. The liberal Justices sometimes (in 26% of the cases) vote for a conservative decision, whereas conservative Justices seldom support a liberal decision. Because of their centrist position in the court, the moderate Justices usually are in the majority. However, the moderate Justices are slightly more likely to support conservative decisions than liberal decisions. In general, the uncertainty in these classifications is low, especially given the relatively small size of the data set. The Justices and almost all issues can be assigned to one segment with a posterior probability close to 1.

The clustering of the Justices, as displayed in Table 5.4, is equal to the one found in Brusco and Steinley (2006). However, Doreian *et al.* (2004) divide the Justices into four segments. The two moderate Justices, *O'Connor* and *Kennedy*, are both an entire cluster on their own.

The segmentation of the issues deviates more from the solutions of Doreian *et al.* (2004) and Brusco and Steinley (2006), who find 7 and 5 segments, respectively. First, they both divide our liberal majority cluster into 3 clusters, depending on whether *O'Connor*, *Kennedy* or both of them voted with the liberal Justices. Second, they divide our conservative majority cluster into 2 or 4 clusters. They both find a cluster where all liberal justices vote with the minority, except that Doreian *et al.* (2004) add *Free Speech* to this cluster. With the remaining issues, both some or all of the liberal Justices, and some or all of the conservative Justices voted with the majority. The allocation of these issues is

Table 5.5: Marginal posterior distribution of the issues over the segments.

Issue \ Segment	1	2	Issue \ Segment	1	2
2000 Presidential Election	0.00	1.00	Clean Air Act	0.20	0.80
Federalism	0.00	1.00	Illegal Search 3	0.57	0.43
Clean Water	0.00	1.00	PGA vs. Handicapped	1.00	0.00
Title VI Disabilities	0.00	1.00	Illegal Search 1	1.00	0.00
Tobacco Ads	0.00	1.00	Illegal Search 2	1.00	0.00
Labor Rights	0.00	1.00	Stay of Execution	1.00	0.00
Property Rights	0.00	1.00	Privacy	1.00	0.00
Citizenship	0.00	1.00	Immigration Jurisdiction	1.00	0.00
Free Speech	0.00	1.00	Detaining Criminal Aliens	1.00	0.00
Seat Belts	0.00	1.00	Legal Aid for the Poor	1.00	0.00
United Foods	0.00	1.00	Voting Rights	1.00	0.00
New York Times Copyright	0.00	1.00	Deporting Criminal Aliens	1.00	0.00
Cannabis for Health	0.01	0.99	Campaign Finance	1.00	0.00
Majority	liberal	conservative	Majority	liberal	conservative

less marked, especially *Illegal Search 3*, which has an almost equal posterior probability of belonging to either cluster. This is the only issue where both moderate Justices voted with the minority and, according to Doreian *et al.* (2004), it is a cluster of its own. The *Clean Air Act* is the only issue that was decided unanimously and is therefore also a bit harder to classify. Doreian *et al.* (2004) cluster it together with *New York Times Copyright* and *Cannabis for Health* into a segment for which there was a (nearly) unanimous decision.

5.5 Application 2: Roll call voting data

5.5.1 Data

To apply our method to a larger data set, we consider the voting behavior of the entire United States House of Representatives. The details of each roll call vote of the United States congress are published on the website `http://www.GovTrack.us`. We gathered data on all roll call votes from the House of Representatives in 2007. We only use data on votes that are related to a bill. We thus obtain data on 766 roll call votes from 427

Table 5.6: Sample means of the individual characteristics for the whole House, Democrats, and Republicans.

	House	Democrats	Republicans
Size	1.00	0.54	0.46
Democrat	0.54	1.00	0.00
Female	0.16	0.20	0.10
Age	55.73	56.54	54.78
Region			
Pacific	0.16	0.19	0.12
Mountain	0.07	0.05	0.09
West North Central	0.07	0.07	0.08
East North Central	0.15	0.14	0.17
Middle Atlantic	0.14	0.18	0.10
New England	0.05	0.09	0.01
West South Central	0.11	0.08	0.15
East south Central	0.06	0.05	0.08
South Atlantic	0.18	0.15	0.22

members of the House of Representatives in 2007. There are four possible types of votes: *yea*, *nay*, *no vote*, and *present*. A *no vote* means that the representative was absent at the moment of voting; this is the case for 3.5% of the observations. A *present* vote means that the representative is present, but votes neither *yea* nor *nay*, which happens only 143 times (0.00%).

We did not recode our data, such that the majority vote always gets a 1, or that the side which the (majority of) Democrats prefer always gets a 1. These recodings have been done in the past, for example with the Supreme Court voting data from Section 5.4. The argument for doing so is that for each bill it is possible to provide an opposite bill for which every representative will vote the opposite. However, this argument does not take into account that there is a status quo, from which a bill deviates. This means that the hypothetical opposite bill is one that goes from the new position towards the status quo. We think that it is interesting to see in which direction a bill deviates from the status quo; this information would be lost if we recoded the votes.

Table 5.7: The numbers of bills prepared by each House Committee

Administration	14	Intelligence (Permanent Select)	15
Agriculture	18	Judiciary	51
Appropriations	291	Natural Resources	50
Armed Services	43	Oversight and Government Reform	57
Budget	14	Rules	9
Education and Labor	51	Science and Technology	43
Energy and Commerce	44	Small Business	26
Financial Services	98	Transportation and Infrastructure	69
Foreign Affairs	44	Veterans' Affairs	15
Homeland Security	55	Ways and Means	42

We collected some additional information on the representatives from GovTrack.us. We have data on their party membership, gender, age on January 1st 2007, and state from which they were elected. Table 5.6 shows the means for these variables for the entire House of Representatives and for the Democrats and Republicans separately. In 2007, the Democrats had a majority in the House of 53.9%, and there were no third-party or independent representatives. There is a fairly large difference in the share of female representatives between the Democrats (20.4%) and the Republicans (10.1%). The age is about the same for representatives from both parties. We aggregated the representatives' home state into nine regions.

We also collected more information on the bills. Before a bill comes to a vote in the House of Representatives, it is prepared by one or more committees. There are twenty such committees in the House. Table 5.7 shows the committees and how many bills they prepared. The committee that handles the largest number of bills is *Appropriations*, which controls the disbursement of funds. The *Rules* committee influences what is discussed and voted upon; this committee is not primarily concerned with bills and only prepared nine of them. Most other committees deal with specific topics. The committee(s) that prepared a bill is an indication for the subject of the bill. Having this information should help us interpret the clusters of bills. Identifying the segments of bills helps us to understand the segments of representatives in a better way, as we know what types of bills they support and oppose.

Other papers have analyzed roll call voting as well. Poole and Rosenthal (1991), Heckman and Snyder Jr. (1997), and Nelson (2002) try to estimate latent preferences

of representatives, based on their voting behavior. De Leeuw (2006) plots the relative positions of representatives into a two-dimensional space. The paper that most closely resembles ours is Hartigan (2000), who clusters the members of the United States Senate, as well as the bills that they vote on. However, he does not cluster the two dimensions simultaneously, but alternates between clustering one dimension conditional on the segmentation of the other dimension, until convergence.

5.5.2 Results

We apply the latent-class two-mode clustering model to the roll call voting data. We assign a 1 to *yea* votes and a 0 to *nay* votes; we treat the response options *no vote* and *present* as missing observations. Again, we describe the individual votes using a Bernoulli distribution with a Beta$(1, 1)$ prior for the probability. Furthermore, we use an uninformative Dirichlet$(1, 1, \ldots, 1)$ prior for both κ and λ.

To determine the numbers of segments, we now opt for the AIC-3 criterion as described in Section 5.3.2. To prevent the Gibbs sampler from getting stuck in a local optimum of the likelihood function, we sample 10 sets of 10 MCMC chains, and each of the 100 MCMC chains has length 200. For each set, the MCMC chain that attains the highest

Table 5.8: AIC-3 values for $K = 2, \ldots, 12$ segments of bills and $L = 2, \ldots, 10$ segments of representatives.

Segments	Segments of representatives								
of bills	2	3	4	5	6	7	8	9	10
2	251,889	245,490	242,752	242,630	243,333	244,062	245,130	246,267	247,433
3	206,373	197,752	193,995	193,132	193,360	193,966	194,899	195,951	197,020
4	191,907	181,359	177,459	176,081	176,193	176,556	177,376	178,328	179,349
5	186,493	175,547	170,848	169,409	169,370	169,698	170,476	171,396	172,353
6	181,998	170,774	165,959	164,315	164,261	164,480	165,229	166,115	167,061
7	179,889	168,300	163,112	161,363	161,263	161,449	162,178	163,044	163,920
8	179,620	167,790	162,484	160,722	160,516	160,698	161,421	162,244	163,136
9	179,565	167,459	162,080	160,182	160,015	160,056	160,745	161,567	162,449
10	179,531	167,209	161,744	159,715	**159,530**	159,553	160,194	161,011	161,837
11	180,904	168,414	162,822	160,696	160,443	160,346	160,986	161,810	162,621
12	182,500	169,944	164,292	162,167	161,802	161,671	162,248	163,065	163,854

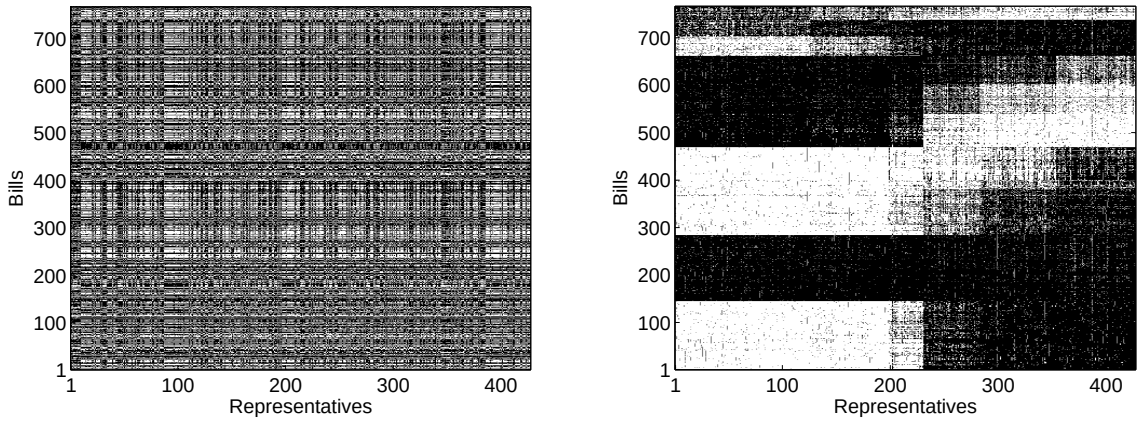
Table 5.9: Posterior means and posterior standard deviations in parentheses of P, κ, and λ.

	Segment of representatives						
Segment of bills	1	2	3	4	5	6	κ
1	0.00	0.01	0.20	0.87	0.98	0.99	0.19
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
2	1.00	1.00	1.00	1.00	1.00	0.98	0.18
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)
3	0.01	0.03	0.17	0.45	0.83	0.93	0.13
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.00)	(0.01)
4	0.01	0.01	0.07	0.13	0.30	0.79	0.12
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
5	0.99	0.97	0.84	0.14	0.02	0.02	0.09
	(0.00)	(0.00)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
6	1.00	0.99	0.95	0.65	0.25	0.10	0.08
	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
7	1.00	1.00	0.99	0.95	0.83	0.48	0.08
	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.01)	(0.01)
8	0.11	0.41	0.78	0.93	0.97	0.98	0.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.01)
9	0.71	0.91	0.95	0.97	0.98	0.97	0.05
	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.00)	(0.01)
10	0.74	0.49	0.34	0.25	0.09	0.15	0.04
	(0.01)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)	(0.01)
λ	0.29	0.17	0.08	0.12	0.17	0.17	
	(0.02)	(0.02)	(0.01)	(0.02)	(0.02)	(0.02)	

likelihood value is chosen, and this MCMC chain is allowed to run for an additional 3,000 iterations. The highest likelihood value that is attained during these 3,000 iterations over all sets of MCMC chains is then used as the final maximum likelihood value. This likelihood value serves as input for the AIC-3 information criterion.

Table 5.8 displays the AIC-3 values. The lowest AIC-3 value is attained with $K = 10$ segments of bills and $L = 6$ segments of representatives; the corresponding log-likelihood value is $-66,108.70$. For these numbers of segments, we sample an additional 100,000

Figure 5.1: Graphical Representation of Voting Data Set Before and After Reordering of Rows and Columns.



iterations from the chain that had the highest likelihood value. Due to the large size of the data set, we have no problems with label switching (Celeux *et al.*, 2000; Geweke, 2007). In the remainder of this section, we present and interpret the results for this model specification.

Table 5.9 shows the posterior means and standard deviations for P, κ, and λ. The first thing to notice is that except for segments (of bills) 9 and 10, the posterior means of the *yea* voting probabilities in all rows are monotonously increasing or decreasing. There are only deviations from monotonicity in segments 9 and 10 in representative class 6. These results imply that the political preferences in the House are one-dimensional. Bills from segment 2 are approved more or less unanimously, and bills from segments 7 and 9 are also widely supported. Bills from other segments seem to be backed by representatives from either segments 1-3 or segments 4-6. In the next subsection, we show that these segments mainly contain Democrats and Republicans, respectively.

To show the effectiveness of our two-mode clustering method, we show graphical representations of the roll call voting data set before and after reordering the rows and columns according to their segment in Figure 5.1. For this reordering, we used the segmentation k and l that yielded the highest likelihood value. Before reordering the rows and columns, it is already apparent that some structure exists in the data; after reordering, the nature of the block structure becomes clear.

Interpretation of segments

For each row (bill) and for each column (representative), we compute the marginal posterior distribution over the segments. This allow us to compute the means of our explanatory variables within each of the latent classes. Table 5.10 shows the posterior means of the individual characteristics of the representatives over the latent classes. The main result is that the first three segments consist of Democrats, and the last three contain Republicans. We know from Table 5.9 that voting behavior is monotonous; therefore, we can interpret segments 1 and 6 as very partisan Democrats and Republicans, respectively. Segments 2 and 5 seem to be typical Democrats and Republicans, respectively, and the ones in segments 3 and 4 are relatively moderate. Note that segments 3 and 4 are not completely

Table 5.10: Posterior means of personal variables for each segment of representatives.

Segment of representatives	1	2	3	4	5	6	Total
Democrat	1.00	1.00	0.97	0.02	0.00	0.00	0.54
Female	0.29	0.08	0.16	0.16	0.08	0.08	0.16
Age	59.30	57.44	50.79	57.68	55.19	55.06	55.73
Region 9							
Pacific	0.27	0.12	0.03	0.10	0.14	0.12	0.16
Mountain	0.03	0.06	0.09	0.08	0.03	0.15	0.07
West North Central	0.03	0.12	0.09	0.08	0.07	0.08	0.07
East North Central	0.13	0.14	0.16	0.26	0.14	0.15	0.15
Middle Atlantic	0.22	0.12	0.13	0.27	0.06	0.03	0.14
New England	0.12	0.07	0.00	0.02	0.00	0.00	0.05
West South Central	0.04	0.12	0.15	0.00	0.21	0.19	0.11
East South Central	0.01	0.06	0.16	0.02	0.15	0.04	0.06
South Atlantic	0.14	0.18	0.19	0.18	0.20	0.25	0.18
Region 4							
West	0.30	0.18	0.13	0.18	0.17	0.27	0.22
Mid West	0.17	0.26	0.25	0.34	0.21	0.23	0.23
North East	0.34	0.20	0.13	0.29	0.06	0.03	0.19
South	0.19	0.36	0.50	0.20	0.57	0.48	0.36
Segment size	0.29	0.17	0.08	0.12	0.17	0.17	1.00

homogenous, which means that there is a little overlap between these moderate Democrats and Republicans.

Further, we can see that there are relatively more females in the left wing. Not only are females more often Democrats than Republicans, also within their parties they seem to be on the left side. The effect of age seems to be non-existent within the Republican party, but within the Democratic party, the younger representatives seem more moderate than the older ones.

There are also some clear regional patterns. Representatives from states in the *West* are more extreme in their voting behavior, as there are few representatives from these states that are in the moderate clusters 3 and 4. Interestingly, it seems that representatives from the *Pacific* are responsible for the left wing, while the right wing representatives seem to come mainly from the *Mountain* states. Representatives from the *Mid West* on the other hand seem to be more moderate than the national average, though the effect is not as strong. In the *North East*, we find that the Democrats are hard liners, and the Republicans are moderate. Lastly, in the *South*, we find that the Democrats are moderate, whereas the Republicans belong more to the hard line segments.

Table 5.11 contains the posterior means of the committees for the segments of bills. The results are less pronounced than for the representatives. For example, the posterior means for segment 1 closely resemble the entire sample (that is, the final column in the table), except that segment 1 contains relatively many bills from the *Financial Services* committee.

Nevertheless, there are some striking results. Bills from the *Veterans affairs* committee all belong to segment 2, which contains bills that receive nearly unanimous support. *Transportation and Infrastructure* is relatively common in segments 5, 6, 7, and 10, which are all primarily favored by Democrats. Bills from the *Judiciary* committee can primarily be found in segments 2, 7, and 9. These segments are the ones for which the voting is almost unanimously *yea*. Segment 4 almost solely contains bills from the *Appropriations* committee. Only the hard-line Republicans from class 6 vote in majority (79%) *yea* for these bills. To a lesser extent, this is also true for bills from segment 3, though there is a little more support for these bills, even from some of the moderate Democrats in segment 3.

Table 5.11: Posterior means of committees for each segment of bills.

Segment of bills	1	2	3	4	5	6	7	8	9	10	Total
Administration	0.01	0.03	0.00	0.00	0.04	0.02	0.00	0.00	0.09	0.03	0.02
Agriculture	0.01	0.01	0.03	0.00	0.03	0.00	0.02	0.02	0.00	0.28	0.02
Appropriations	0.36	0.05	0.66	0.92	0.27	0.24	0.07	0.58	0.37	0.38	0.38
Armed Services	0.06	0.05	0.02	0.00	0.12	0.00	0.05	0.02	0.14	0.27	0.06
Budget	0.03	0.00	0.00	0.00	0.07	0.05	0.02	0.02	0.00	0.00	0.02
Education and Labor	0.07	0.05	0.08	0.00	0.11	0.15	0.09	0.00	0.06	0.10	0.07
Energy and Commerce	0.07	0.11	0.01	0.00	0.07	0.08	0.05	0.00	0.03	0.10	0.06
Financial Services	0.24	0.09	0.15	0.03	0.03	0.24	0.14	0.12	0.03	0.07	0.13
Foreign Affairs	0.05	0.08	0.03	0.00	0.06	0.02	0.09	0.00	0.09	0.35	0.06
Homeland Security	0.08	0.05	0.02	0.02	0.19	0.05	0.05	0.12	0.14	0.13	0.07
Intelligence (Permanent Select)	0.04	0.00	0.00	0.00	0.07	0.02	0.02	0.05	0.03	0.00	0.02
Judiciary	0.04	0.14	0.00	0.00	0.07	0.03	0.16	0.02	0.23	0.07	0.07
Natural Resources	0.08	0.09	0.03	0.00	0.10	0.14	0.05	0.02	0.03	0.10	0.07
Oversight and Government Reform	0.03	0.17	0.02	0.00	0.10	0.07	0.14	0.05	0.09	0.10	0.07
Rules	0.01	0.00	0.00	0.00	0.04	0.03	0.00	0.00	0.06	0.03	0.01
Science and Technology	0.04	0.10	0.04	0.01	0.05	0.02	0.09	0.07	0.09	0.10	0.06
Small Business	0.05	0.05	0.01	0.00	0.01	0.02	0.11	0.00	0.00	0.10	0.03
Transportation and Infrastructure	0.06	0.09	0.05	0.03	0.17	0.18	0.15	0.07	0.06	0.10	0.09
Veterans' Affairs	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
Ways and Means	0.06	0.04	0.00	0.01	0.18	0.11	0.05	0.02	0.11	0.00	0.05
Segment size	0.19	0.18	0.13	0.12	0.09	0.08	0.08	0.05	0.05	0.04	1.00

5.6 Conclusions

We have developed a Bayesian approach to do inference in a latent-class two-mode clustering model. The advantage of the Bayesian approach is that it allows for statistical inference on the model parameters, which is not possible using the maximum likelihood approach. The Bayesian approach also allows us to do statistical inference on the number of segments using marginal likelihoods. An alternative way to select the numbers of segments is to consider information criteria. The third advantage of using Bayesian estimation techniques is that we run into fewer computational problems during estimation.

We have applied our model to the Supreme Court voting data set of Doreian *et al.* (2004) and Brusco and Steinley (2006). The marginal likelihoods used to determine the optimal numbers of segments indicate fewer segments than were found in these previous studies. In the second example, we consider roll call votes from the United States House

of Representatives in 2007. We detect six segments of representatives and ten segments of bills. Three of the individual segments contain Democrats and the other three segments contain Republicans, though there is a little overlap. There are also clear regional effects.

Another advantage of our approach is that it can easily be extended in several directions. First, it can be adapted for use with data matrices with arbitrary distributions, though we have only derived posterior samplers for Bernoulli and normally distributed data. Second, it can be adapted for three-mode data (Schepers *et al.*, 2006). Third, explanatory variables can be added, either with segment-dependent effects or as concomitant variables, that is, variables explaining why a row (or column) belongs to a certain segment, see Dayton and MacReady (1988).

Nederlandse samenvatting (Summary in Dutch)

In dit proefschrift introduceren we nieuwe toepassingen van finite mixture verdelingen in econometrische modellen. Een finite mixture verdeling is een type statistische verdeling. Een statistische verdeling beschrijft de kans op elke mogelijke uitkomst van een variabele, bijvoorbeeld de prijs van een willekeurig huis in Rotterdam of het aantal ogen bij het gooien van een dobbelsteen. Verschillende typen verdelingen kunnen verschillende patronen in de data beschrijven. Als een variabele niet een standaardpatroon heeft of als het patroon van een variabele vooraf onbekend is wordt vaak een finite mixture verdeling toegepast, omdat die heel flexibel is en veel verschillende datapatronen kan beschrijven. Een finite mixture verdeling is namelijk een gewogen gemiddelde van een bepaald aantal andere verdelingen. Deze andere verdelingen worden vaak de mixture componenten genoemd en de gewichten de mixture proporties.

Een econometrisch model dat gebruik maakt van een finite mixture verdeling noemen we een finite mixture model. De mixture componenten worden dan vaak latente klassen genoemd. Elk van de observaties die gebruikt worden om de parameters van het model te schatten behoort nu tot één van de latente klassen. Dit maakt finite mixture modellen heel geschikt als clustermethode. De latente klassen worden nu geïnterpreteerd als clusters of segmenten en de mixture proporties geven de grootte van ieder cluster.

Dit proefschrift bevat vier hoofdstukken. Elk hoofdstuk behandelt een specifiek probleem en kan onafhankelijk van de andere hoofdstukken worden gelezen. In elk hoofdstuk gebruiken we minimaal één dataset om te laten zien hoe de modeluitkomsten inzicht geven in de onderliggende data. Deze datasets komen uit verschillende onderzoeksgebieden, zoals macro-economie, marketing en politicologie. We laten steeds het nut van het model zien en soms ook de verbetering ten opzichte van bestaande methodes in de literatuur. Nu volgt er een kort overzicht van ieder hoofdstuk.

Hoofdstuk 2 behandelt de situatie dat er geen data beschikbaar is voor de verklarende variabelen. Econometrische modellen worden vaak gebruikt om de invloed van deze verklarende variabelen op een responsvariabele te berekenen. Om deze invloed te berekenen hebben we data nodig van zowel de responsvariabele als de verklarende variabelen. Ook als er geen individuele data beschikbaar is voor (een aantal van) de verklarende variabelen, kunnen we in sommige gevallen toch informatie verkrijgen over deze variabelen op een geaggregeerd niveau. Het kan bijvoorbeeld zo zijn dat we de responsvariabele op huishoudniveau hebben, maar de verklarende variabelen alleen maar op postcodeniveau. We nemen aan dat deze geaggregeerde informatie over de verklarende variabelen de vorm heeft van een discrete verdeling over intervallen (inkomen of leeftijd) of categorieën (geslacht of religie) voor een huishouden in een bepaald postcodegebied. We ontwikkelen een model om in deze situatie toch de effecten te meten van de verklarende variabelen op de responsvariabele.

Om de effecten van deze geaggregeerde variabelen te meten breiden we het standaard responsmodel uit met een latentevariabelenmodel dat de ontbrekende verklarende variabelen beschrijft op individueel niveau. Hierbij wordt er rekening gehouden met de informatie die we op het geaggregeerde niveau hebben. Als er maar één verklarende variabele is, is dit een standaard finite mixture model waarbij de mixing proporties bekend zijn. In een simpel simulatie-experiment laten we zien dat deze aanpak leidt tot efficiëntere parameterschattingen dan de standaardmethode, waar we de ontbrekende verklarende variabele vervangen door zijn geobserveerde marginale kans op het geaggregeerde niveau.

De parameters van het individuele responsmodel kunnen zowel klassiek als Bayesiaans worden geschat. Computationeel gezien is de Bayesiaanse methode te prefereren. De posteriorresultaten kunnen worden verkregen door het gebruik van een Gibbs simulator met data-augmentatie. Hierbij simuleren we de ontbrekende verklarende variabelen naast de modelparameters. Conditioneel op deze gesimuleerde verklarende variabelen kunnen we de parameters van het individuele responsmodel schatten met standaard Makov Chain Monte Carlo methoden.

We illustreren deze methode met een dataset over donaties aan een goed doel. In deze dataset weten we de respons van huishoudens op een direct mailing van het goede doel. Verder is alleen de postcode van deze huishoudens bekend. Op postcodeniveau hebben we informatie over sommige huishoudkarakteristieken zoals leeftijd en inkomen. We tonen aan dat ouderen met een hoog inkomen het meeste doneren.

In Hoofdstuk 3 willen we de voorkeuren van individuen bepalen over een aantal opties, bijvoorbeeld verschillende spelcomputers, vakantielanden of merken van een bepaald product. Hierbij maken we gebruik van enquêteresultaten. Traditioneel worden respondenten alleen gevraagd om aan te geven welke optie zij het beste vinden. Het is bekend dat we meer informatie kunnen krijgen van een respondent als we hem vragen om alle opties te rangschikken. Hierdoor kunnen de parameters van het keuzemodel efficiënter geschat worden. In de praktijk kan het echter zo zijn dat sommige respondenten niet in staat zijn om hun voorkeuren weer te geven voor sommige opties in de keuzeset, bijvoorbeeld omdat ze geen ervaring hebben met die opties. Het is bekend in de literatuur dat de parameterschattingen hierdoor een onzuiverheid kunnen krijgen richting nul. De standaardoplossing is om niet de complete rangschikking te gebruiken, maar bijvoorbeeld alleen de top drie. De vraag is nu welk gedeelte van de rankschikkingen we veilig kunnen gebruiken.

Om dit op te lossen introduceren we een nieuw rank-ordered logit model met latente klassen. Dit model houdt expliciet rekening met het feit dat de gegeven rangschikking niet altijd overeenkomt met de voorkeuren van de respondent. De latente klassen identificeren endogeen hoeveel opties de respondenten goed kunnen rangschikken. Hierdoor hoeft dit aantal niet meer vooraf te worden gekozen. Bovendien kan dit aantal nu per respondent verschillen, wat met eerdere methodes ook niet mogelijk was. Dit resulteert in een finite mixture model waarbij elk mixture component correspondeert met een specifieke aanname over hoe goed een respondent de opties kan rangschikken.

Naast de toename in efficiëntie kunnen we door het toepassen van dit model ook leren hoe goed respondenten in staat zijn om de opties te rangschikken. Deze informatie kan ook heel waardevol zijn. Bijvoorbeeld om vast te stellen hoeveel respondenten er nodig zijn om een gewenste precisie te bereiken. We stellen een likelihood ratio toets voor om te bepalen of een bepaald segment aanwezig is. Deze toets kan bijvoorbeeld gebruikt worden om overbodige segmenten uit het model te verwijderen. We illustreren het model met een simulatie-experiment en een empirische toepassing. Hieruit blijkt dat ons model inderdaad leidt tot efficiëntere parameterschattingen dan standaardmethodes. Bovendien blijven de parameterschattingen ook zuiver als sommige respondenten niet in staat zijn om de complete rangschikking uit te voeren.

In Hoofdstuk 4 beschrijven we de ontwikkeling van regionale huizenprijzen in Nederland tussen 1985 en 2005 met behulp van een panelmodel. In dit model zijn de volgende eigenschappen van de reeksen mogelijk: stochastische trends, co-integratie, correlaties

tussen de regio's en een clustering van de regio's op basis van latente klassen. De clustering is volledig door de data bepaald en gebaseerd op de gemiddelde groeivoet van de huizenprijzen en de relatie van de huizenprijzen met het BBP. Hierbij wordt er onderscheid gemaakt tussen hoe sterk en hoe snel de huizenprijzen van een regio reageren op veranderingen in het BBP.

We passen het model toe op kwartaaldata van Nederlandse huizenprijzen binnen 76 regio's. De resultaten laten zien dat er twee verschillende clusters zijn. De kleinste omvat zo'n 20% van de regio's, voornamelijk in het oosten van het land. Opvallend is dat de grotere steden in het oosten zoals Zwolle en Arnhem juist weer in het grote cluster vallen. Er is geen significant verschil in de groeivoet van de huizenprijzen tussen de twee clusters. Wel reageren de huizenprijzen in het kleine cluster sterker en sneller op veranderingen in het BBP. Deze verschillen kunnen zijn veroorzaakt door het toegenomen aantal forenzen die in regio's van het kleine cluster wonen.

Met dit model kunnen we ook scenarioanalyses uitvoeren. We bekijken twee scenario's, in de eerste stijgt het BBP met één procent extra. We zien dat de huizenprijzen in het kleine cluster iets sneller en sterker reageren op deze verandering, maar uiteindelijk zijn de verschillen klein en stijgen de huizenprijzen tussen de 1,5 en 1,7% extra. In het tweede scenario stijgt de rente met een procentpunt. Hier zijn de verschillen groter. De huizenprijzen in het kleine cluster zakken ruim $2\frac{1}{2}$% terwijl die in het grote cluster maar 2% zakken.

Ook in Hoofdstuk 5 houden we ons bezig met clusteren met behulp van finite mixtures. Klassieke clustermethodes verdelen een enkele reeks objecten over segmenten. In dit hoofdstuk clusteren we zowel de rijen als de kolommen van een datamatrix simultaan in segmenten. De rijen en kolommen hebben beiden hun eigen verzameling van segmenten. We clusteren de rijen en kolommen zodanig dat de blokstructuur die dan ontstaat zo veel mogelijk hetzelfde gedrag laat zien binnen de blokken, maar verschillend gedrag tussen de blokken. We ontwikkelen een nieuwe Bayesiaanse methode om de parameters te schatten van een latenteklassenmodel voor dit simultaan clusteren. De posteriorresultaten kunnen we verkrijgen met behulp van een Gibbs simulator met data-augmentatie.

Onze Bayesiaanse methode heeft drie voordelen over bestaande methodes in de literatuur. Ten eerste kunnen we statistische inferentie toepassen op de modelparameters. Als we bestaande frequentistische schattingsprocedures gebruiken, is het niet mogelijk om standaardfouten te berekenen en is ook toetsen en bepalen of effecten significant zijn niet meer mogelijk. Verder kunnen we statistische criteria geven voor het bepalen van de

optimale aantallen segmenten. Ten slotte heeft onze Gibbs simulator minder problemen met lokale optima in de likelihoodfunctie en lege klassen dan het EM algoritme dat wordt gebruikt in een frequentistische benadering.

We passen ons model toe op twee datasets. De eerste dataset is heel klein en bestaat uit het stemgedrag van de 9 rechters van het Amerikaanse hooggerechtshof over 26 onderwerpen. Deze dataset is vaker gebruikt in de literatuur om clusteralgoritmes te illustreren. Wij laten zien dat eerdere oplossingen uit de literatuur te veel segmenten hebben. De tweede dataset is juist groot en bevat alle stemmen van het Amerikaanse huis van afgevaardigden over wetsvoorstellen in 2007. We laten zien dat de afgevaardigden met 6 klassen kunnen worden beschreven en de wetsvoorstellen met 10 klassen. Het stemgedrag voor een klasse wetsvoorstellen is over de 6 klassen van afgevaardigden altijd monotoon dalend of stijgend. Hieruit kunnen we concluderen dat er maar één dimensie is te ontwaren in de politieke voorkeuren van de afgevaardigden.

De vier hoofdstukken in dit proefschrift ontwikkelen nieuwe modellen of breiden bestaande modellen uit en ontwikkelen nieuwe schattingsmethoden voor de parameters van deze modellen. Ook komen we tot een aantal opvallende empirische resultaten. We hopen hiermee een waardevolle bijdrage te leveren aan de econometrische literatuur.

Bibliography

Abrevaya, J. and J. A. Hausman (1999), Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells, *Annales D'Economie et de Statistique*, **55-56**, 243–275.

Ahn, J., J. Lee, J.-D. Lee, and T.-Y. Kim (2006), An Analysis of Consumer Preferences among Wireless LAN and Mobile Internet Services, *ETRI Journal*, **28**, 205–215.

Aitchison, J. and S. D. Silvey (1957), The Generalization of Probit Analysis to the Case of Multiple Responses, *Biometrika*, **44**, 131–140.

Amemiya, T. (1974), Bivariate Probit Analysis: Minimum Chi-Square Methods, *Journal of the American Statistical Association*, **69**, 940–944.

Andrews, R. L. and I. S. Currim (2003), A Comparison of Segment Retention Criteria for Finite Mixture Logit Models, *Journal of Marketing Research*, **40**, 235–243.

Ashford, J. R. and R. R. Sowden (1970), Multi-Variate Probit Analysis, *Biometrics*, **26**, 536–546.

Barnard, J., R. McCulloch, and X. L. Meng (2000), Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage, *Statistica Sinica*, **10**, 1281–1311.

Beggs, S., S. Cardell, and J. A. Hausman (1981), Assessing the Potential Demand for Electric Cars, *Journal of Econometrics*, **16**, 1–19.

Berger, J. (1985), *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag, New York.

Billard, L. and E. Diday (2003), From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis, *Journal of the American Statistical Association*, **98**, 470–487.

Börsch-Supan, A. and V. A. Hajivassiliou (1993), Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models, *Journal of Econometrics*, **58**, 347–368.

Boswijk, H. P. (1994), Testing for an Unstable Root in Conditional and Structural Error Correction Models, *Journal of Econometrics*, **63**, 37–60.

Bozdogan, H. (1987), Model Selection and Akaike's Information Criterion (AIC): The General Theory and its Analytical Extensions, *Psychometrika*, **52**, 345–370.

Bozdogan, H. (1994), Mixture-Model Cluster Analysis Using Model Selection Criteria and a New Information Measure of Complexity, in H. Bozdogan (ed.), *Proceedings of the First US/Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach*, vol. 2, Kluwer, Boston.

Breitung, J. and M. H. Pesaran (2008), Unit Roots and Cointegration in Panels, in L. Matyas and P. Sevestre (eds.), *The Econometrics of Panel Data (Third Edition)*, Kluwer, Boston.

Bronnenberg, B. J. and W. R. Vanhonacker (1996), Limited Choice Sets, Local Price Response, and Implied Measures of Price Comptetition, *Journal of Marketing Research*, **33**, 163–174.

Brusco, M. and D. Steinley (2006), Inducing a Blockmodel Structure of Two-Mode Binary Data using Seriation Procedures, *Journal of Mathematical Psychology*, **50**, 468–477.

Calfee, J., C. Winston, and R. Stempski (2001), Econometric Issues in Estimating Consumer Preferences From Stated Preference Data: A Case Study of the Value of Automobile Travel Time, *The Review of Economics and Statistics*, **83**, 699–707.

Cameron, A. C. and P. K. Trivedi (2005), *Microeconometrics*, Cambridge University Press, New York.

Cameron, G., J. Muellbauer, and A. Murphy (2006), Was there a British House Price Bubble? Evidence from a Regional Panel, Economics Series Working Papers 276, University of Oxford, Department of Economics.

Candel, M. J. J. M. and E. Maris (1997), Perceptual Analysis of Two-Way Two-Mode Frequency Data: Probability Matrix Decomposistion and Two Alternatives, *International Journal of Research in Marketing*, **14**, 321–339.

Celeux, G., M. Hurn, and C. P. Robert (2000), Computational and Inferential Difficulties with Mixture Posterior Distributions, *Journal of the American Statistical Association*, **95**, 957–970.

Chapman, R. G. and R. Staelin (1982), Exploiting Rank Ordered Choice Set Data Within the Stochastic Utility Model, *Journal of Marketing Research*, **19**, 288–301.

Chiang, J., S. Chib, and C. Narasimhan (1999), Markov Chain Monte Carlo and Models of Consideration Set and Parameter Heterogeneity, *Journal of Econometrics*, **9**, 223–248.

Chib, S. and E. Greenberg (1998), Analysis of Multivariate Probit Models, *Biometrika*, **85**, 347–361.

Chua, T. C. and W. A. Fuller (1987), A Model for Multinomial Response Error Applied to Labor Flows, *Journal of the American Statistical Association*, **82**, 46–51.

Cowles, M. K. (1996), Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models, *Statistics and Computing*, **6**, 101–111.

Dagsvik, J. K. and G. Liu (2006), A Framework for Analyzing Rank Ordered Panel Data with Application to Automobile Demand, Discussion Papers 480, Research Department of Statistics Norway.

Davies, R. B. (1977), Hypothesis Testing when a Nuisance Parameter is Present Only Under the Alternative, *Biometrika*, **64**, 247–254.

Dayton, C. M. and G. B. MacReady (1988), Concomitant Variable Latent Class Models, *Journal of the American Statistical Association*, **83**, 173–178.

De Leeuw, J. (2006), Principal Component Analysis of Binary Data by Iterated Singular Value Decomposition, *Computational Statistics & Data Analysis*, **50**, 21–39.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977), Maximum Likelihood Estimation from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

DeSarbo, W. S., D. K. H. Fong, J. C. Liechty, and M. K. Saxton (2004), A Hierarchical Bayesian Procedure for Two-Mode Cluster Analysis, *Psychometrika*, **69**, 547–572.

Dias, J. G. and M. Wedel (2004), An Empirical Comparison of EM, SEM and MCMC Performance for Problematic Gaussian Mixture Likelihoods, *Statistics and Computing*, **14**, 323–332.

Doornik, J. A. (2007), *Object-Oriented Matrix Programming using Ox*, 3rd edn., Timberlake Consultants Press, Londen.

Doreian, P., V. Batagelj, and A. Ferligoj (2004), Generalized Blockmodeling of Two-Mode Network Data, *Social Networks*, **26**, 29–53.

Drewes, T. and C. Micheal (2006), How do Students Choose a University?: An Analysis of Applications to Universities in Ontario, Canada, *Research in Higher Education*, **47**, 781–800.

Everitt, B. S. and D. J. Hand (1981), *Finite Mixture Distributions*, Chapman and Hall, London.

Fok, D., R. Paap, and P. H. Franses (2008), Incorporating Responsiveness to Marketing Efforts in Brand Choice Modeling, Econometric Institute report 2008-15, http://hdl.handle.net/1765/13051.

Fraley, C. and A. E. Raftery (1998), How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis, *The Computer Journal*, **41**, 578–588.

Franses, P. H. and R. Paap (2001), *Quantitative Models in Marketing Research*, Cambridge University Press, Cambridge.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, Springer, New York.

Geman, S. and D. Geman (1984), Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Geweke, J. (1991), Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints, in E. M. Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, Interface Foundation of North America, Fairfax, VA.

Geweke, J. (2004), Getting it Right: Joint Distribution Tests of Posterior Simulators, *Journal of the American Statistical Association*, **99**, 799–804.

Geweke, J. (2005), *Contemporary Bayesian Econometrics and Statistics*, Wiley, New York.

Geweke, J. (2007), Interpretation and Inference in Mixture Models: Simple MCMC Works, *Computational Statistics & Data Analysis*, **51**, 3529–3550.

Geweke, J., M. Keane, and D. Runkle (1994), Alternative Computational Approaches to Inference in the Multinomial Probit Model, *The Review of Economics and Statistics*, **76**, 609–632.

Govaert, G. and M. Nadif (2003), Clustering with Block Mixture Models, *Pattern Recognition*, **36**, 463–473.

Govaert, G. and M. Nadif (2008), Block Clustering with Bernouilly Mixture Models: Comparison of Different Approaches, *Computational Statistics & Data Analysis*, **52**, 3233–3245.

Hajivassiliou, V. A. and P. A. Ruud (1994), Classical Estimation Methods for LDV Models using Simulation, in R. Engle and D. McFadden (eds.), *Handbook of Econometrics*, vol. 4, Elsevier, Amsterdam.

Han, C. and B. P. Carlin (2001), Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review, *Journal of the American Statistical Association*, **96**, 1122–1133.

Hartigan, J. A. (1975), *Clustering Algorithms*, Wiley, New York.

Hartigan, J. A. (2000), Bloc Voting in the United States Senate, *Journal of Classification*, **17**, 29–49.

Hausman, J. A. (1978), Specification Tests in Econometrics, *Econometrica*, **46**, 1251–1271.

Hausman, J. A. (2001), Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left, *Journal of Economic Perspectives*, **15**, 57–67.

Hausman, J. A. and P. A. Ruud (1987), Specifying and Testing Econometric Models for Rank-Ordered Data, *Journal of Econometrics*, **34**, 83–104.

Hausman, J. A. and D. A. Wise (1978), A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Hetrogeneous Preferences, *Econometrica*, **46**, 403–426.

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998), Misclassification of the Dependent Variable in a Discrete-Response Setting, *Journal of Econometrics*, **87**, 239–269.

Heckman, J. J. (1982), Statistical Models for Analysis of Discrete Panel Data, in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data*, MIT Press, Cambridge.

Heckman, J. J. and B. Singer (1984), A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data, *Econometrica*, **52**, 271–320.

Heckman, J. J. and J. M. Snyder Jr. (1997), Linear Probability Models of the Demand for Attributes with an Empirical Application to Estimating the Preferences of Legislators, *RAND Journal of Economics*, **28**, S142–S189.

Holly, S., M. H. Pesaran, and T. Yamagata (2008), A Spatio-Temporal Model of House Prices in the US, *Journal of Econometrics*, forthcoming, Cambridge Working Papers in Economics No. 654, University of Camebridge, and CESInfo Working Paper Series No. 1826.

Hsieh, C. (2005), Age and Relative Importance of Major Life Domains, *Journal of Aging Studies*, **19**, 503–512.

Im, K. S., M. H. Pesaran, and Y. Shin (2003), Testing for Unit Roots in Heterogenous Panels, *Journal of Econometrics*, **115**, 53–74.

Imbens, G. W. and T. Lancaster (1994), Combining Micro and Macro Data in Microeconometric Models, *Review of Economic Studies*, **61**, 655–680.

International Monetary Fund (2004), The Global House Price Boom, *World Economic Outlook*, 1491–1517.

Joe, H. (1997), *Multivariate Models and Dependence Concepts*, Chapman and Hall, London.

Kass, R. E. and A. E. Raftery (1995), Bayes Factors, *Journal of the American Statistical Association*, **90**, 773–795.

Keane, M. P. (1992), A Note on Identification in the Multinomial Probit Model, *Journal of Business & Economic Statistics*, **10**, 193–200.

Keane, M. P. (1994), A Computationally Practical Simulation Estimator for Panel Data, *Econometrica*, **62**, 95–116.

Kockelman, K. M., K. V. Podgorski, M. Bina, and S. Gadda (2008), Public Perceptions of Pricing Existing Roads and Other Transportation Policies: The Texas Perspective, *Journal of the Transportation Research Forum*, forthcoming.

Kon, S. J. (1984), Models of Stock Returns – A Comparison, *Journal of Finance*, **39**, 147–165.

Koop, G. and D. J. Poirier (1994), Rank-Ordered Logit Models: An Empirical Analysis of Ontario Voter Preferences, *Journal of Applied Econometrics*, **9**, 369–388.

Lerman, S. R. and C. F. Manski (1981), On the Use of Simulated Frequencies to Approximate Choice Probabilities, in C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT press, Cambridge.

Levin, A., C. F. Lin, and C. J. Chu (2002), Unit Root Tests in Panel Data: Asymptotic and Finite-Sample Asymptotics, *Journal of Econometrics*, **108**, 1–24.

Liechty, J. C., M. W. Liechty, and P. Müller (2004), Bayesian Correlation Estimation, *Biometrika*, **91**, 1–14.

Malpezzi, S. (1999), A Simple Error Correction Model of House Prices, *Journal of Housing Economics*, **8**, 27–62.

Manchanda, P., A. Ansari, and S. Gupta (1999), The "Shopping Basket": A Model for Multicategory Purchase Incedence Decisions, *Marketing Science*, **18**, 95–114.

Manski, C. F. (1977), The Structure of Random Utility Models, *Theory and Decision*, **8**, 229–254.

Mark, D. R., J. L. Lusk, and M. Scott Daniel (2004), Recruiting Agricultural Economics Graduate Students: Student Demand for Program Attributes, *American Journal of Agricultural Economics*, **86**, 175–184.

McFadden, D. (1973), Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York.

McFadden, D. (1974), The Measurement of Urban Travel Demand, *Journal of Public Economics*, **3**, 303–328.

Meng, X.-L. and D. B. Rubin (1993), Maximum Likelihood Estimation via the ECM Algorithm: a General Framework, *Biometrika*, **80**, 267–278.

Moon, H. R. and B. Perron (2004), Testing for a Unit Root in Panels with Dynamic Factors, *Journal of Econometrics*, **122**, 81–126.

Nelson, J. P. (2002), "Green" Voting and Ideology: LCV Scores and Roll-Call Voting in the U.S. Senate, 1988-1998, *The Review of Economics and Statistics*, **84**, 518–529.

Newton, M. A. and A. E. Raftery (1994), Approximate Bayesian Inference by the Weighted Likelihood Bootstrap, *Journal of the Royal Statistical Society, Series B*, **56**, 3–48.

Paap, R., P. H. Franses, and D. J. C. Van Dijk (2005), Does Africa Grow Slower Than Asia, Latin America and the Middle East? Evidence from a New Data-Based Classification Method, *Journal of Development Economics*, **77**, 553–570.

Pesaran, M. H. (2004), General Diagnostic Tests for Cross Section Dependence in Panels, Cambridge Working Papers in Economics No. 435, University of Cambridge, and CESInfo Working Paper Series No. 1229.

Pesaran, M. H. (2007), A Simple Panel Unit Root Test in the Presence of Cross-Section Dependence, *Journal of Applied Econometrics*, **22**, 265–312.

Pesaran, M. H., A. Ullah, and T. Yamagata (2008), A Bias-Adjusted LM Test of Error Cross Section Independence, *The Econometrics Journal*, **11**, 105–127.

Poole, K. T. and H. Rosenthal (1991), Patterns of Congressional Voting, *American Journal of Political Science*, **35**, 228–278.

Poterba, J. M. and L. H. Summers (1995), Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification, *The Review of Economics and Statistics*, **77**, 207–216.

Quandt, R. E. and J. B. Ramsey (1978), Estimating Mixtures of Normal Distributions and Switching Regressions, *Journal of the American Statistical Association*, **73**, 730–738.

Ritter, C. and M. A. Tanner (1992), Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler, *Journal of the American Statistical Association*, **87**, 861–868.

Romeo, C. J. (2005), Estimating Discrete Joint Probability Distributions for Demographic Characteristics at the Store Level Given Store Level Marginal Distributions and a City-Wide Joint Distribution, *Quantitative Marketing and Economics*, **3**, 71–93.

Schepers, J., I. Van Mechelen, and E. Ceulemans (2006), Three-Mode Partitioning, *Computational Statistics & Data Analysis*, **51**, 1623–1642.

Schepers, J., E. Ceulemans, and I. Van Mechelen (2008), Selecting among Multi-Mode Partitioning Models of Different Complexities: A Comparison of Four Model Selection Criteria, *Journal of Classification*, **25**, 67–85.

Steenburgh, T. J., A. Ainslie, and P. H. Engebretson (2003), Massively Categorical Variables: Revealing the Information in Zip Codes, *Marketing Science*, **22**, 40–57.

Stern, S. (1992), A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models, *Econometrica*, **60**, 943–952.

Symons, M. J. (1981), Clustering Criteria and Multivariate Normal Mixtures, *Biometrics*, **37**, 35–43.

Tanner, M. A. and W. H. Wong (1987), The Calculation of Posterior Distributions by Data Augmentation, *Journal of the American Statistical Association*, **82**, 528–540.

Titterington, D. M. T., A. F. M. Smith, and U. E. Makov (1985), *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

Tobin, J. (1958), Estimation of Relationships for Limited Dependent Variables, *Econometrica*, **26**, 24–36.

Van Beek, K. W. H., C. C. Koopmans, and B. M. S. Van Praag (1997), Shopping at the Labour Market: A Real Tale of Fiction, *European Economic Review*, **41**, 295–317.

Van den Berg, G. J. and B. Van der Klaauw (2001), Combining Micro and Macro Unemployment Duration Data, *Journal of Econometrics*, **102**, 271–309.

Van Dijk, B. (2006), Treatment Effect of Job-Training Programmes on Unemployment Duration in Slovakia, *Statistica Neerlandica*, **60**, 57–72.

Van Dijk, B. and R. Paap (2008), Explaining Individual Response using Aggregated Data, *Journal of Econometrics*, **146**, 1–9.

Van Dijk, B., D. Fok, and R. Paap (2007a), A Rank-Ordered Logit Model with Unobserved Heterogeneity in Ranking Capabilities, Econometric Institute Report 2007-07, http://hdl.handle.net/1765/8533.

Van Dijk, B., P. H. Franses, R. Paap, and D. J. C. Van Dijk (2007b), Modeling Regional House Prices, Econometric Institute Report 2007-55, http://hdl.handle.net/1765/11723.

Van Dijk, B., J. Van Rosmalen, and R. Paap (2009), A Bayesian Approach to Two-Mode Clustering, Econometric Institute Report 2009-06, http://hdl.handle.net/1765/15112.

Van Mechelen, I., H.-H. Bock, and P. De Boeck (2004), Two-Mode Clustering Methods: A Structured Overview, *Statistical Methods in Medical Research*, **13**, 363–394.

Van Ophem, H., P. Stam, and B. M. S. Van Praag (1999), Multichoice Logit: Modeling Incomplete Preference Rankings of Classical Concerts, *Journal of Business & Economic Statistics*, **17**, 117–128.

Van Rosmalen, J., P. J. F. Groenen, J. Trejos, and W. Castillo (2008), Global Optimization Strategies for Two-Mode Clustering, *Journal of Classification*, forthcoming.

Wakefield, J. (2004), Ecological Inference for 2×2 Tables, *Journal of the Royal Statistical Society, Series A*, **167**, 385–445.

Wedel, M. and W. A. Kamakura (2000), *Market Segmentation: Conceptual and Methodological Foundations*, second edn., Springer, Boston.

Wedel, M., W. A. Kamakura, N. Arora, A. Bemmaor, J. Chiang, T. Elrod, R. Johnson, P. Lenk, S. Neslin, and C. S. Poulsen (1999), Discrete and Continuous Representations of Unobserved Heterogeneity in Choice Modeling, *Marketing Letters*, **10**, 219–232.

White, H. (1982), Maximum Likelihood Estimation of Misspecified Models, *Econometrica*, **50**, 1–25.

Wolak, F. A. (1989a), Local and Global Testing of Linear and Nonlinear Inequality Constraints in Nonlinear Econometric Models, *Econometric Theory*, **5**, 1–35.

Wolak, F. A. (1989b), Testing Inequality Constraints in Linear Econometric Models, *Journal of Econometrics*, **41**, 205–235.

Yang, S. and G. Allenby (2000), A Model for Observation, Structural, and Household Heterogeneity in Panel Data, *Marketing Letters*, **11**, 137–149.

Author index

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus University Rotterdam, University of Amsterdam and VU University Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

415. F. RAVAZZOLO *Forecasting financial time series using model averaging.*
416. M.J.E. KABKI *Transnationalism, local development and social security: the functioning of support networks in rural Ghana.*
417. M. POPLAWSKI RIBEIRO *Fiscal policy under rules and restrictions.*
418. S.W. BISSESSUR *Earnings, quality and earnings management: the role of accounting accruals.*
419. L. RATNOVSKI *A Random Walk Down the Lombard Street: Essays on Banking.*
420. R.P. NICOLAI *Maintenance models for systems subject to measurable deterioration.*
421. R.K. ANDADARI *Local clusters in global value chains, a case study of wood furniture clusters in Central Java (Indonesia).*
422. V.KARTSEVA *Designing Controls for Network Organizations: A Value-Based Approach.*
423. J. ARTS *Essays on New Product Adoption and Diffusion.*
424. A. BABUS *Essays on Networks: Theory and Applications.*
425. M. VAN DER VOORT *Modelling Credit Derivatives.*
426. G. GARITA *Financial Market Liberalization and Economic Growth.*
427. E.BEKKERS *Essays on Firm Heterogeneity and Quality in International Trade.*
428. H.LEAHU *Measure-Valued Differentiation for Finite Products of Measures: Theory and Applications.*
429. G. BALTUSSEN *New Insights into Behavioral Finance.*
430. W. VERMEULEN *Essays on Housing Supply, Land Use Regulation and Regional Labour Markets.*
431. I.S. BUHAI *Essays on Labour Markets: Worker-Firm Dynamics, Occupational Segregation and Workplace Conditions.*
432. C. ZHOU *On Extreme Value Statistics.*
433. M. VAN DER WEL *Riskfree Rate Dynamics: Information, Trading, and State Space Modeling.*

434. S.M.W. PHLIPPEN *Come Close and Co-Create: Proximities in pharmaceutical innovation networks.*
435. A.V.P.B. MONTEIRO *The Dynamics of Corporate Credit Risk: An Intensity-based Econometric Analysis.*
436. S.T. TRAUTMANN *Uncertainty in Individual and Social Decisions: Theory and Experiments.*
437. R. LORD *Efficient pricing algorithms for exotic derivatives.*
438. R.P. WOLTHOFF *Essays on Simultaneous Search Equilibrium.*
439. Y.-Y. TSENG *Valuation of travel time reliability in passenger transport.*
440. M.C. NON *Essays on Consumer Search and Interlocking Directorates.*
441. M. DE HAAN *Family Background and Children's Schooling Outcomes.*
442. T. ZAVADIL *Dynamic Econometric Analysis of Insurance Markets with Imperfect Information.*
443. I.A. MAZZA *Essays on endogenous economic policy.*
444. R. HAIJEMA *Solving large structured Markov Decision Problems for perishable-inventory management and traffic control.*
445. A.S.K. WONG *Derivatives in Dynamic Markets.*
446. R. SEGERS *Advances in Monitoring the Economy.*
447. F.M. VIEIDER *Social Influences on Individual Decision Making Processes.*
448. L. PAN *Poverty, Risk and Insurance: Evidence from Ethiopia and Yemen.*
449. B. TIEBEN *The concept of equilibrium in different economic traditions: A Historical Investigation.*
450. P. HEEMEIJER *Expectation Formation in Dynamic Market Experiments.*
451. A.S. BOOIJ *Essays on the Measurement Sensitivity of Risk Aversion and Causal Effects in Education.*
452. M.I. LÓPEZ YURDA *Four Essays on Applied Microeconometrics.*
453. S. MEENTS *The Influence of Sellers and the Intermediary on Buyers' Trust in C2C Electronic Marketplaces.*
454. S. VUJIĆ *Econometric Studies to the Economic and Social Factors of Crime.*
455. F. HEUKELOM *Kahneman and Tversky and the Making of Behavioral Economics.*
456. G. BUDAI-BALKE *Operations Research Models for Scheduling Railway Infrastructure Maintenance.*
457. T.R. DANIELS *Rationalised Panics: The Consequences of Strategic Uncertainty during Financial Crises.*