Models for default rates in credit portfolios

M. van der Walt

Dissertation submitted in partial fulfilment of the requirements for the degree

Master of Science

in Risk Analysis
at the Potchefstroom campus of the
North-West University

Supervisor: Prof. M.F. Kruger
Co-supervisor: Prof. J.H. Venter

November 2007

Potchefstroom

Acknowledgements

I wish to express my sincere thanks to:

Professor Hennie Venter and Professor Machiel Kruger. Thank you for your patience, guidance, time and input. It was truly a privilege experience to work under your supervision.

Professor Riaan de Jongh – for granting me the opportunity to continue with a Masters degree. Also for the bursary from the National Research Foundation.

Everyone who motivated me throughout the preparation of this dissertation. In particular, my most loyal supporter Wicus.

I thank my parents for the opportunity to study and for their unconditional support.

Soli deo Gloria

Table of Contents

Table of Contents (contd)

Chapter 4: Auto-regressive models with unobserved components

Abstract:

Models for default rates in credit portfolios

The default rate is a measure widely used in credit risk management. This reflects the probability that obligors will default on their credit obligations over a specified time horizon. Our aim is to formulate statistical models that can describe the default rate dynamics and to forecast future default tendencies of credit portfolios. Auto-regressive (AR) models and various extended forms of AR models are used for this purpose. The extended AR models incorporate observed exogenous factors (such as econornic variables) as well as unobserved or latent components. A restricted multivariate vector auto-regressive (VAR) model is also explored in this context.

Monthly default rates data of a mortgage loans portfolio was obtained and used to illustrate the statistical methodology required to fit these models. Since default rates often have very small values and highly skewed distributions, probit and logistic transformations of the rates were necessary before model fitting could be done. For this data, it was found that using only 1 auto-regressive term was sufficient. However, the inclusion of economic variables (e.g. CPIX) and the use of multivariate models were not completely satisfactory and therefore $AR(1)$ models extended with unobserved components were developed and applied to the data. These unobserved components were assumed to have $AR(1)$ dynamics of their own and this made the use of standard software packages impossible when fitting these models by maximum likelihood estimation methods. For this purpose and also for forecasting, methods were developed based on the Kalman filter and the Expectation-Maximization (EM)-algorithm. This formed the main contribution of this work.

OPSOMMING:

Modelle vir nie-nakomingskoerse in krediet portefeuljes

Die nie-nakomingskoers is 'n maatstaf wat algemeen gebruik word vir die bestuur van kredietrisiko. Dit is 'n uitdrukking van die waarskynlikheid dat leners nie hul kredietverpligtinge nakom nie. Ons doelwit is om statistiese modelle te formuleer wat die werking van die nie-nakomingskoerse in krediet portefeuljes kan beskryf en voorspel. Autoregressiewe (AR) modelle en verskeie uitgebreide vorme daarvan word vir hierdie doel gebruik. Die uitgebreide AR-modelle sluit waargenome eksogene faktore (bv. ekonomiese veranderlikes) sowel as nie-waargenome (latente) komponente in. 'n Beperkte meerveranderlike vektor-AR model (VAR) word ook bespreek.

Maandelikse nie-nakomingskoersdata van 'n huisleningsportefeulje is verkry en word gebruik ter illustrasie van die statistiese metodiek benodig vir die passing van hierdie modelle. Nie-nakomingskoerse is geneig om baie klein waardes te hê wat skewe verdelings tot gevolg het. Om hierdie rede word 'n probit- en logistiese transformasie van die koerse benodig voordat modelle gepas kan word. Dit blyk dat een autoregressiewe term in die model voldoende is vir hierdie data. Die insluiting van ekonomiese veranderlikes (bv. CPIX) en die gebruik van meerveranderlike modelle was nie heeltemal bevredigend nie en daarom is $AR(1)$ modelle uitgebrei om nie-waargenome komponente in te sluit. Hierdie modelle is ook op die data toegepas. Die aanname word gemaak dat hierdie nie-waargenome komponente 'n $AR(1)$-dinamika van hul eie besit en dit maak die gebruik van standaard sagteware pakkette vir die passing van hierdie modelle met maksimum aanneemlikheid metodes onmoontlik. Vir hierdie doel en vir voorspelling is metodes gebaseer op die Kalman filter en die "Expectation-Maximization" (EM)-algoritme ontwikkel. Dit vorm die hoofbydrae van hierdie werk.

Chapter 1
Introduction

1.1 Background

Credit activities play a very important role in modern commerce and finance. If potential homeowners could not acquire mortgage loans to finance building of houses, the residential building industry would only be a small fraction of what it actually is. Without vehicle finance far fewer cars will be sold. Without credit card facilities, retailing would be much less convenient and less secure. Banks play a major role in the credit industry. Indeed a large part of the business of banks is to borrow funds from depositors and to lend these funds out on a credit basis at a higher interest rate than that paid to the depositors.

According to the Credit Risk DI500 data of the South African Reserve Bank (2007), the total mortgage loans of all South African banks amounted to R692bn at the start of February 2007. Again their total amount in instalment sales and leases was R210bn and the total outstanding amount on credit cards was R45.7bn at the beginning of February 2007. The sizes of these amounts clearly show how large the South African credit industry has become. The international credit industry is of course larger than this by several orders of magnitude. Underscoring the importance of the credit industry, the governor of the SA Reserve Bank, Mr Tito Mboweni (2007), recently expressed his concerns regarding the inflationary potential of South Africa's increasingly higher credit extension and extremely high household indebtedness. Except for inflation, extreme borrowing also leads to further dangers to the health of the economy when interest rates start to rise while consumers are overcommitted. The current implementation of the SA National Credit Act endeavours to protect credit customers by introducing more control over the marketing of credit, again emphasising the critical role of the credit industry in the soundness of the economy.

Granting credit is a risky business. Some of the obligors (i.e. the people or institutions who were granted credit) may cease repaying outstanding amounts duly, thus defaulting on their credit obligations. This leads to losses for the granter of the credit. For example, the Credit Risk DI500 data of the SA Reserve Bank (2007) shows that the total loss on mortgage loans at the beginning of February 2007 on the books of all South African banks was R4674m, while their loss on instalment sales and leases was R2301m and R327m on credit cards. Such losses have a large impact on the profitability of banks and other credit granting institutions and it is clearly important for them to manage their credit risk very carefully.

The first step in the process of granting credit is to evaluate and analyse credit applications and their corresponding default risk. This is done through assessing whether the obligor has the ability and willingness to honour his debt obligation, a process called credit scoring. According to Koch and MacDonald (2003:590), the following are examples of loan request features that should be considered in the credit scoring process:

1. The character of the obligor (e.g. his commitment and ability to repay debts as stipulated in the loan agreement) and the quality of information that he provided.
2. The use of the loan proceeds.
3. The amount requested for borrowing and time at which the loan will be repaid.
4. The primary repayment source; as well as the secondary source (i.e. the collateral or guarantees) available.

These items assist in assessing the creditworthiness of the obligor. Of course, the amount and form of risk that the lender is prepared to take, are determined by the lender's credit risk appetite and strategy. Risk diversification is an important principle in this regard. Very often small banks grant too many loans in one industry, e.g. in the agriculture industry. This may be due to the specific economic conditions of the bank's trade area. If conditions adversely affect the agriculture industry, the value of such a loan portfolio will deteriorate since the loan portfolio is not sufficiently diversified. Credit scoring is only the first step in managing credit risk. Ongoing monitoring and prediction of the credit portfolio's performance are required to be able to take timely corrective actions to keep the credit portfolio on a sound basis.

1.2 Aims of the dissertation

The Basel Accord plays an increasingly important role in bank management in general and credit risk management in particular. Basel II (Basel Committee on Banking Supervision, 2006:52) requires that banks that follow the advanced internal rating-based approach (IRB) use the following three factors for credit risk calculations on their credit portfolio:

1. *Probability of Default (PD)*: This probability determines how likely the obligor is to default on the credit obligation within a given time horison (e.g. a year).
2. *Exposure At Default (EAD)*: This is the amount the obligor owes to the credit granting institution at the moment of default.
3. *Loss Given Default (LGD):* This measures the amount that the bank might actually lose when an obligor defaults, taking into account possible recoveries after default.

Of the abovementioned three factors, this dissertation will focus on the modelling of the *probability of default*. The other two factors are also important, but we will consider only homogeneous portfolios - in which case those factors may be handled as given constants (fixed, at least to first approximation).

By a *homogeneous credit portfolio* we mean that the portfolio consists of classes of obligors that are similar to each other within a specific class. The similarity is in terms of their default probabilities, exposures and losses given default.
Empirical data may be available in the form of regular periodic (monthly, quarterly or annual) fractions or percentages of obligors that default over that time period. These percentages are often referred to as *empirical default frequencies*, but in this study the shorter term *default rates* will be used.

It is the aim of this dissertation to discuss and compare several different models that can be used to estimate and predict these default rates, possibly also taking into account exogenous variables relating to the economic environment (e.g. interest rates). Possible relations between these model types will be discussed, as well as the procedures and methods that were used in fitting the models to empirical data. The results will be compared after the model fitting. By doing so, the adequacy of the models can be assessed and suggestions on handling possible inadequate models wil be made.

1.3 Overview of the dissertation

We obtained empirical data on historical home loan default rates from a well-known bank. This data will serve as an illustration of the models and methodologies investigated and proposed in the dissertation.

In Chapter 2, this default data is first discussed and analysed to obtain an overview of its statistical properties. Default rates often have very small values and therefore have highly skewed distributions. It is thus important to transform the default rates before analysis and in order to meet this requirement, we will formulate models in terms of the probit and logistic transformed default rates throughout the dissertation. The statistical properties of the home loans data suggest that auto-regressive (AR) models can be used to describe this data.

In Chapter 3, AR default rate models with and without exogenous factors (such as economic variables) are fitted to the transformed default rates. We find that inclusion of the economic factors in the AR model is not particularly helpful in explaining the behaviour of the home loans default rates.

In an effort to improve the AR models, we study the alternative possibility of including unobserved components in the AR models in Chapter 4. Two maximum likelihood estimation methods are presented for fitting the models to our data, namely the Kalman filter (Harvey, 1989) and the expectation-maximisation (EM)-algorithm (Dempster *et al.*, 1977). The results are compared after the models were fitted to the home loans data and the proposed models are also applied to forecasting the default rates. Each chapter contains a brief summary of its contents and Chapter 5 concludes the dissertation with final remarks.

Chapter 2
Historical credit default rate data

2.1 Introduction

A credit default rate data set will be used to motivate and illustrate the types of models that will be developed and studied in this dissertation. Section 2.2 gives the details of the data set and Sections 2.3 to 2.5 present a preliminary exploratory analysis of the data, providing an overview of its main statistical properties. Section 2.6 g ives a summary of this chapter.

2.2 The home loans default rate data set

A local bank provided us with a data set giving the historical default rates of their home loan portfolio. It covers the default rates for 56 months, rang ing from 1 September 2000 to 1 April 2005. Below we will refer to this as the *home loans* data set. The home loans data set is divided into 9 risk classes. The obligors were grouped according to a calculated risk score, where a risk class consists of all obligors with scores that are in the same score bracket. Risk class 1 represents the obligors with highest default risk, while risk class 9 has the obligors that have a low default frequency. The *default rate (dr)* is calculated by dividing the number of obligors that default over the month by the number of obligors present in that risk class at the beginning of the month, i.e. it represents the fraction of obligors in each risk class that defaulted over that month.

Table 2.2.1 on the next page shows the home loans default rate data. Figure 2.2.1 below graphs the time series of default rates over months for each risk class. Clearly, there is substantial variability over time and the default rates decrease from class 1 to class 9. Indeed from class 4 and onwards the default rates are so small that they are not clearly visible in Figure 2.2.1, but inspection of Table 2.2.1 confirms that they continue to decrease, becoming very small in class 9. One way to deal with the issue of handling such small rates is by transformation. Two common possibilities often used in the area of generalised linear models (McCullagh & Nelder, 1989:108) are the probit and logistic transformations to be discussed and motivated further in Chapter 3 below. The probit transformation transforms the default rate r to $\Phi^{-1}(r)$ (the inverse standard normal distribution function) and the logistic transformation transforms it to $\ln(r/(1-r))$. Both are monotone transforms. Tables 2.2.2 and 2.2.3 show the probit and logistic transformed default rates respectively (abbreviated as *pdr* and *ldr*).

Table 2.2.1 Home loans default rates of 9 risk classes over months (1=Sept 2000, 56=April 2005)

Month	dr1	dr2	dr3	dr4	dr5	dr6	dr7	dr8	dr9
1	0.21921	0.04610	0.00640	0.00268	0.00179	0.00081	0.00026	0.00041	0.00008
2	0.22115	0.05517	0.00729	0.00177	0.00126	0.00062	0.00035	0.00033	0.00011
3	0.24763	0.04988	0.00535	0.00218	0.00072	0.00036	0.00027	0.00021	0.00009
4	0.28368	0.05797	0.00829	0.00328	0.00159	0.00036	0.00042	0.00038	0.00010
5	0.25987	0.06053	0.00934	0.00323	0.00142	0.00047	0.00031	0.00043	0.00007
6	0.23094	0.05786	0.00893	0.00178	0.00145	0.00058	0.00066	0.00029	0.00010
7	0.25734	0.05362	0.00703	0.00282	0.00179	0.00029	0.00020	0.00027	0.00007
8	0.20443	0.04426	0.00944	0.00351	0.00178	0.00082	0.00066	0.00032	0.00010
9	0.21314	0.05559	0.00998	0.00199	0.00120	0.00042	0.00056	0.00036	0.00017
10	0.21955	0.05722	0.00865	0.00275	0.00146	0.00042	0.00056	0.00055	0.00011
11	0.30517	0.09800	0.02923	0.00592	0.00148	0.00064	0.00060	0.00039	0.00019
12	0.25941	0.05441	0.01988	0.00325	0.00099	0.00036	0.00035	0.00033	0.00008
13	0.24379	0.06356	0.01816	0.00401	0.00164	0.00043	0.00055	0.00047	0.00012
14	0.25854	0.06545	0.02010	0.00530	0.00147	0.00061	0.00027	0.00038	0.00010
15	0.25802	0.06546	0.01871	0.00545	0.00194	0.00036	0.00013	0.00006	0.00003
16	0.30685	0.08230	0.02409	0.00470	0.00193	0.00038	0.00037	0.00009	0.00006
17	0.17393	0.03817	0.01289	0.00252	0.00189	0.00039	0.00032	0.00027	0.00003
18	0.22396	0.05272	0.01379	0.00350	0.00177	0.00032	0.00022	0.00010	0.00003
19	0.27464	0.08602	0.04213	0.00991	0.00307	0.00125	0.00041	0.00018	0.00006
20	0.25232	0.05312	0.01987	0.00663	0.00260	0.00053	0.00041	0.00027	0.00016
21	0.24647	0.04645	0.01211	0.00482	0.00164	0.00080	0.00053	0.00037	0.00010
22	0.18443	0.03748	0.01046	0.00237	0.00185	0.00052	0.00041	0.00030	0.00007
23	0.20891	0.03569	0.01340	0.00301	0.00218	0.00122	0.00051	0.00059	0.00019
24	0.21200	0.03504	0.01421	0.00295	0.00120	0.00066	0.00041	0.00056	0.00016
25	0.16985	0.04086	0.02512	0.00391	0.00237	0.00042	0.00044	0.00059	0.00012
26	0.17214	0.02329	0.01764	0.00195	0.00261	0.00051	0.00025	0.00027	0.00008
27	0.26150	0.03818	0.03093	0.00323	0.00190	0.00041	0.00056	0.00024	0.00012
28	0.26416	0.04229	0.03082	0.00352	0.00264	0.00066	0.00034	0.00029	0.00020
29	0.21630	0.03022	0.03156	0.00321	0.00215	0.00072	0.00053	0.00026	0.00021
30	0.19699	0.02667	0.01567	0.00370	0.00216	0.00068	0.00036	0.00039	0.00018
31	0.21593	0.03365	0.01521	0.00072	0.00151	0.00052	0.00024	0.00036	0.00018
32	0.26141	0.03873	0.01684	0.00154	0.00212	0.00055	0.00052	0.00030	0.00022
33	0.22785	0.03551	0.02038	0.00480	0.00178	0.00054	0.00037	0.00049	0.00041
34	0.16927	0.03177	0.02280	0.00241	0.00181	0.00053	0.00042	0.00038	0.00018
35	0.22999	0.05088	0.02325	0.00178	0.00219	0.00051	0.00047	0.00024	0.00035
36	0.20702	0.04501	0.02731	0.00222	0.00113	0.00038	0.00022	0.00022	0.00027
37	0.14081	0.05596	0.03092	0.00301	0.00076	0.00019	0.00017	0.00012	0.00017
38	0.14103	0.05605	0.03092	0.00309	0.00078	0.00018	0.00018	0.00011	0.00017
39	0.17010	0.05551	0.04092	0.00338	0.00110	0.00016	0.00027	0.00027	0.00034
40	0.25887	0.07059	0.04104	0.00205	0.00091	0.00017	0.00009	0.00014	0.00020
41	0.20113	0.04446	0.02539	0.00240	0.00054	0.00020	0.00022	0.00029	0.00017
42	0.15548	0.03158	0.02524	0.00340	0.00071	0.00035	0.00033	0.00024	0.00024
43	0.18262	0.04304	0.03673	0.00374	0.00052	0.00028	0.00032	0.00016	0.00009
44	0.24973	0.04741	0.02446	0.00228	0.00041	0.00020	0.00015	0.00009	0.00015
45	0.21341	0.05179	0.07043	0.00742	0.00213	0.00176	0.00207	0.00185	0.00089
46	0.18108	0.03498	0.01449	0.00189	0.00052	0.00050	0.00039	0.00027	0.00013
47	0.23101	0.04490	0.01723	0.00358	0.00049	0.00050	0.00043	0.00045	0.00033
48	0.16377	0.03308	0.01089	0.00262	0.00052	0.00053	0.00043	0.00023	0.00020
49	0.17301	0.02944	0.01264	0.00278	0.00071	0.00034	0.00023	0.00027	0.00013
50	0.20575	0.03347	0.02209	0.00128	0.00058	0.00026	0.00029	0.00018	0.00014
51	0.17026	0.03675	0.02708	0.00187	0.00058	0.00035	0.00017	0.00016	0.00011
52	0.21377	0.04327	0.02840	0.00157	0.00019	0.00034	0.00033	0.00028	0.00013
53	0.19019	0.03689	0.01343	0.00137	0.00033	0.00020	0.00020	0.00019	0.00017
54	0.20306	0.03732	0.01109	0.00136	0.00042	0.00031	0.00020	0.00010	0.00016
55	0.21502	0.03592	0.01834	0.00093	0.00021	0.00022	0.00018	0.00012	0.00009
56	0.23023	0.03777	0.01618	0.00155	0.00038	0.00035	0.00022	0.00018	0.00019

Table 2.2.2 Probits of home loans default rates over months (1=Sept 2000, 56=April 2005)

Month	Pdr1	Pdr2	Pdr3	Pdr4	Pdr5	Pdr6	Pdr7	Pdr8	Pdr9
1	-0.77485	-1.68386	-2.48903	-2.78463	-2.91220	-3.15072	-3.47521	-3.34350	-3.76730
2	-0.76830	-1.59667	-2.44275	-2.91573	-3.02174	-3.22833	-3.38578	-3.40520	-3.70510
3	-0.68196	-1.64606	-2.55210	-2.85133	-3.18744	-3.38469	-3.45561	-3.53266	-3.73499
4	-0.57195	-1.57208	-2.39576	-2.71825	-2.95036	-3.38266	-3.34128	-3.37040	-3.70982
5	-0.64376	-1.55036	-2.35178	-2.72356	-2.98542	-3.30986	-3.42271	-3.33093	-3.81145
6	-0.73575	-1.57296	-2.36833	-2.91467	-2.97849	-3.24674	-3.20982	-3.44481	-3.70958
7	-0.65155	-1.61069	-2.45567	-2.76825	-2.91283	-3.43860	-3.53474	-3.45891	-3.81209
8	-0.82590	-1.70327	-2.34780	-2.69589	-2.91539	-3.14759	-3.20953	-3.41834	-3.71217
9	-0.79557	-1.59288	-2.32700	-2.87948	-3.03558	-3.33675	-3.25831	-3.38519	-3.58412
10	-0.77370	-1.57854	-2.38025	-2.77641	-2.97678	-3.33751	-3.25800	-3.26616	-3.68800
11	-0.50957	-1.29304	-1.89220	-2.51710	-2.97254	-3.22174	-3.23743	-3.36203	-3.55283
12	-0.64516	-1.60351	-2.05616	-2.72104	-3.09233	-3.38497	-3.38764	-3.40713	-3.78184
13	-0.69416	-1.52559	-2.09323	-2.65109	-2.94038	-3.33397	-3.26485	-3.30662	-3.66946
14	-0.64784	-1.51058	-2.05161	-2.55565	-2.97455	-3.23397	-3.45757	-3.36434	-3.72236
15	-0.64947	-1.51045	-2.08118	-2.54603	-2.88808	-3.38299	-3.65792	-3.86131	-3.99655
16	-0.50481	-1.38976	-1.97580	-2.59699	-2.88986	-3.36369	-3.37311	-3.75944	-3.82849
17	-0.93875	-1.77230	-2.22950	-2.80477	-2.89605	-3.36190	-3.41439	-3.46078	-3.99660
18	-0.75889	-1.61907	-2.20308	-2.69656	-2.91579	-3.41025	-3.51490	-3.72303	-3.99597
19	-0.59885	-1.36569	-1.72648	-2.32986	-2.74054	-3.02310	-3.34320	-3.56195	-3.82966
20	-0.66722	-1.61531	-2.05649	-2.47682	-2.79487	-3.27213	-3.34560	-3.46017	-3.59959
21	-0.68566	-1.68033	-2.25368	-2.58857	-2.93952	-3.15691	-3.27534	-3.37458	-3.73190
22	-0.89861	-1.78065	-2.30940	-2.82481	-2.90288	-3.27790	-3.34382	-3.43096	-3.79680
23	-0.81021	-1.80301	-2.21444	-2.74678	-2.85123	-3.03025	-3.28469	-3.24154	-3.55686
24	-0.79949	-1.81145	-2.19134	-2.75340	-3.03489	-3.21025	-3.34635	-3.25920	-3.60570
25	-0.95474	-1.74074	-1.95794	-2.65953	-2.82467	-3.33672	-3.32826	-3.24326	-3.66302
26	-0.94575	-1.99004	-2.10505	-2.88625	-2.79314	-3.28451	-3.48512	-3.45905	-3.76824
27	-0.63873	-1.77223	-1.86724	-2.72348	-2.89512	-3.34398	-3.25977	-3.49163	-3.66491
28	-0.63058	-1.72472	-1.86887	-2.69493	-2.78925	-3.21100	-3.39840	-3.44296	-3.53501
29	-0.78475	-1.87755	-1.85833	-2.72593	-2.85525	-3.18591	-3.27447	-3.47304	-3.52101
30	-0.85240	-1.93221	-2.15264	-2.67823	-2.85409	-3.20365	-3.38464	-3.36111	-3.56109
31	-0.78603	-1.82972	-2.16456	-3.18563	-2.96647	-3.28067	-3.49512	-3.38462	-3.56271
32	-0.63900	-1.76557	-2.12388	-2.96026	-2.85906	-3.26534	-3.28027	-3.43379	-3.51162
33	-0.74595	-1.80530	-2.04600	-2.58975	-2.91543	-3.26804	-3.37123	-3.29852	-3.34368
34	-0.95705	-1.85534	-1.99916	-2.81846	-2.90911	-3.27589	-3.34234	-3.36702	-3.56682
35	-0.73887	-1.63640	-1.99087	-2.91532	-2.84933	-3.28579	-3.30710	-3.49230	-3.38810
36	-0.81681	-1.69527	-1.92188	-2.84516	-3.05346	-3.36384	-3.50975	-3.50949	-3.45580
37	-1.07669	-1.58963	-1.86747	-2.74666	-3.17243	-3.55257	-3.57808	-3.66940	-3.58115
38	-1.07570	-1.58885	-1.86748	-2.73849	-3.16167	-3.57419	-3.56471	-3.68520	-3.58194
39	-0.95377	-1.59362	-1.74011	-2.70820	-3.06296	-3.59425	-3.46200	-3.45988	-3.39783
40	-0.64685	-1.47142	-1.73878	-2.87101	-3.11736	-3.57553	-3.73970	-3.62447	-3.54580
41	-0.83759	-1.70112	-1.95333	-2.81975	-3.26765	-3.54372	-3.51684	-3.43652	-3.57987
42	-1.01319	-1.85804	-1.95587	-2.70646	-3.19043	-3.39016	-3.40558	-3.48768	-3.49412
43	-0.90542	-1.71640	-1.79002	-2.67499	-3.27893	-3.45055	-3.41768	-3.59876	-3.73359
44	-0.67532	-1.67053	-1.96937	-2.83687	-3.34535	-3.54060	-3.62004	-3.74699	-3.60853
45	-0.79464	-1.62776	-1.47258	-2.43611	-2.85791	-2.91753	-2.86686	-2.90254	-3.12316
46	-0.91127	-1.81220	-2.18382	-2.89649	-3.27971	-3.28868	-3.36317	-3.46304	-3.66115
47	-0.73552	-1.69651	-2.11460	-2.68953	-3.29346	-3.29135	-3.33516	-3.32194	-3.40219
48	-0.97910	-1.83735	-2.29415	-2.79204	-3.28118	-3.27473	-3.33099	-3.50741	-3.54509
49	-0.94232	-1.88903	-2.23722	-2.77242	-3.19070	-3.39362	-3.50135	-3.45954	-3.64829
50	-0.82127	-1.83203	-2.01245	-3.01710	-3.24982	-3.46894	-3.44390	-3.56617	-3.63599
51	-0.95314	-1.78973	-1.92555	-2.89953	-3.24700	-3.38990	-3.58813	-3.60190	-3.70295
52	-0.79341	-1.71395	-1.90483	-2.95408	-3.54986	-3.39911	-3.40469	-3.45182	-3.65289
53	-0.87719	-1.78801	-2.21361	-2.99575	-3.40513	-3.54344	-3.53651	-3.55803	-3.58055
54	-0.83074	-1.78267	-2.28730	-2.99788	-3.33862	-3.42026	-3.53611	-3.71235	-3.59693
55	-0.78911	-1.80012	-2.08931	-3.11024	-3.52898	-3.51127	-3.56846	-3.66853	-3.75361
56	-0.73808	-1.77722	-2.13989	-2.95815	-3.36757	-3.39189	-3.50997	-3.56335	-3.55264

Table 2.2.3 Logits of home loans default rates over months (1=Sept 2000, 56=April 2005)

Month	Ldr1	Ldr2	Ldr3	Ldr4	Ldr5	Ldr6	Ldr7	Ldr8	Ldr9
1	-1.27026	-3.02965	-5.04431	-5.91947	-6.32127	-7.11233	-8.27311	-7.79008	-9.40248
2	-1.25895	-2.84059	-4.91421	-6.33257	-6.67805	-7.38103	-7.94338	-8.01435	-9.15523
3	-1.11128	-2.94705	-5.22459	-6.12771	-7.23877	-7.93941	-8.20019	-8.48890	-9.27360
4	-0.92629	-2.78819	-4.78403	-5.71620	-6.44431	-7.93200	-7.78210	-7.88743	-9.17387
5	-1.04666	-2.74223	-4.66386	-5.73232	-6.55855	-7.66932	-8.07861	-7.74486	-9.58022
6	-1.20300	-2.79005	-4.70889	-6.32919	-6.53588	-7.44560	-7.31642	-8.16018	-9.17293
7	-1.05982	-2.87065	-4.95036	-5.86894	-6.32328	-8.13721	-8.49674	-8.21244	-9.58280
8	-1.35883	-3.07243	-4.65307	-5.64860	-6.33150	-7.10160	-7.31542	-8.06256	-9.18316
9	-1.30609	-2.83249	-4.59688	-6.21679	-6.72392	-7.76576	-7.48632	-7.94124	-8.68481
10	-1.26826	-2.80192	-4.74145	-5.89408	-6.53028	-7.76850	-7.48521	-7.51404	-9.08789
11	-0.82278	-2.21967	-3.50281	-5.12413	-6.51644	-7.35798	-7.41289	-7.85705	-8.56537
12	-1.04903	-2.85525	-3.89778	-5.72467	-6.91382	-7.94041	-7.95015	-8.02140	-9.46082
13	-1.13200	-2.69018	-3.98997	-5.51453	-6.41198	-7.75577	-7.50941	-7.65777	-9.01521
14	-1.05355	-2.65882	-3.88655	-5.23484	-6.52299	-7.40076	-8.20745	-7.86543	-9.22348
15	-1.05631	-2.65855	-3.95987	-5.20709	-6.24417	-7.93322	-8.97013	-9.78317	-10.34551
16	-0.81490	-2.41150	-3.70162	-5.35501	-6.24982	-7.86309	-7.89725	-9.37106	-9.64933
17	-1.55802	-3.22671	-4.33831	-5.98192	-6.26957	-7.85660	-8.04803	-8.21940	-10.34571
18	-1.24274	-2.88867	-4.26960	-5.65064	-6.33277	-8.03287	-8.42186	-9.22614	-10.34303
19	-0.97122	-2.36324	-3.12394	-4.60457	-5.78400	-6.68257	-7.78902	-8.60010	-9.65404
20	-1.08630	-2.88058	-3.89860	-5.00980	-5.95116	-7.53515	-7.79768	-8.21711	-8.74418
21	-1.11755	-3.02189	-4.40170	-5.33041	-6.40922	-7.13354	-7.54652	-7.90262	-9.26134
22	-1.48661	-3.24561	-4.54961	-6.04441	-6.29140	-7.55558	-7.79126	-8.10902	-9.52102
23	-1.33151	-3.29644	-4.29907	-5.80307	-6.12740	-6.70624	-7.57965	-7.42732	-8.58072
24	-1.31288	-3.31573	-4.23925	-5.82334	-6.72164	-7.31794	-7.80038	-7.48947	-8.76771
25	-1.58666	-3.15577	-3.65870	-5.53966	-6.04399	-7.76568	-7.73526	-7.43337	-8.99003
26	-1.57056	-3.73603	-4.01960	-6.23832	-5.94582	-7.57903	-8.31012	-8.21298	-9.40626
27	-1.03818	-3.22657	-3.44447	-5.73207	-6.26661	-7.79181	-7.49145	-8.33444	-8.99740
28	-1.02447	-3.12003	-3.44826	-5.64573	-5.93376	-7.32053	-7.98944	-8.15335	-8.49779
29	-1.28735	-3.46851	-3.42377	-5.73951	-6.14007	-7.23346	-7.54342	-8.26500	-8.44488
30	-1.40518	-3.59731	-4.13999	-5.59554	-6.13642	-7.29496	-7.93921	-7.85372	-8.59679
31	-1.28957	-3.35764	-4.17043	-7.23249	-6.49665	-7.56540	-8.34752	-7.93913	-8.60299
32	-1.03864	-3.21153	-4.06701	-6.47646	-6.15211	-7.51114	-7.56398	-8.11946	-8.40951
33	-1.22050	-3.30168	-3.87270	-5.33386	-6.33163	-7.52068	-7.89044	-7.62884	-7.79075
34	-1.59080	-3.41682	-3.75813	-6.02459	-6.31136	-7.54846	-7.78591	-7.87515	-8.61865
35	-1.20835	-2.92611	-3.73804	-6.33126	-6.12140	-7.58357	-7.65945	-8.33698	-7.95182
36	-1.34300	-3.05477	-3.57279	-6.10826	-6.78344	-7.86363	-8.40246	-8.40147	-8.20092
37	-1.80859	-2.82555	-3.44501	-5.80272	-7.18693	-8.56437	-8.66166	-9.01499	-8.67343
38	-1.80676	-2.82388	-3.44503	-5.77777	-7.14988	-8.64682	-8.61062	-9.07692	-8.67647
39	-1.58492	-2.83407	-3.15436	-5.68577	-6.81517	-8.72367	-8.22392	-8.21605	-7.98737
40	-1.05187	-2.57769	-3.15139	-6.18991	-6.99854	-8.65192	-9.29234	-8.84017	-8.53866
41	-1.37925	-3.06768	-3.64767	-6.02861	-7.51930	-8.53079	-8.42915	-8.12955	-8.66851
42	-1.69222	-3.42310	-3.65375	-5.68049	-7.24910	-7.95936	-8.01575	-8.31967	-8.34378
43	-1.49869	-3.10153	-3.26688	-5.58586	-7.55921	-8.18144	-8.06014	-8.74101	-9.26803
44	-1.10003	-3.00042	-3.68615	-6.08222	-7.79677	-8.51893	-8.82304	-9.32136	-8.77860
45	-1.30449	-2.90742	-2.58007	-4.89571	-6.14847	-6.33836	-6.17676	-6.29032	-7.01823
46	-1.50906	-3.31743	-4.21985	-6.27099	-7.56201	-7.59385	-7.86119	-8.22778	-8.98274
47	-1.20260	-3.05749	-4.04361	-5.62947	-7.61085	-7.60334	-7.76004	-7.71259	-8.00332
48	-1.63047	-3.37522	-4.50888	-5.94240	-7.56721	-7.54436	-7.74505	-8.39367	-8.53599
49	-1.56442	-3.49539	-4.35849	-5.88178	-7.25003	-7.97199	-8.37088	-8.21479	-8.93262
50	-1.35076	-3.36295	-3.79047	-6.66274	-7.45643	-8.24973	-8.15680	-8.61619	-8.88482
51	-1.58380	-3.26621	-3.58150	-6.28069	-7.44648	-7.95840	-8.70016	-8.75307	-9.14674
52	-1.30236	-3.09608	-3.53252	-6.45638	-8.55408	-7.99203	-8.01247	-8.18613	-8.95054
53	-1.44876	-3.26231	-4.29691	-6.59241	-8.01408	-8.52971	-8.50346	-8.58517	-8.67114
54	-1.36728	-3.25018	-4.49063	-6.59942	-7.77250	-8.06963	-8.50194	-9.18388	-8.73397
55	-1.29491	-3.28985	-3.98015	-6.97436	-8.47496	-8.40819	-8.62493	-9.01159	-9.34777
56	-1.20700	-3.23786	-4.10755	-6.46961	-7.87715	-7.96567	-8.40328	-8.60544	-8.56466

Figures 2.2.2 and 2.2.3 graph the time series of these transformed default rates for each risk class. Comparison of the default rates over classes is much easier in the transformed forms. These graphs show that the defaults decrease with increasing risk class number, staying roughly at fixed levels but with substantial variation around these levels over time. It also appears that there is some correlation present between the default rates of the various risk classes since they often move together. For example, all the series show an upward fluctuation over months 19 and 45 and similarly many show a downward fluctuation over months 37-39. Similar co-movements are visible elsewhere in the graphs of the probit and logistic transformed data. This suggests that the default rates are correlated over risk classes, and we investigated this issue further below. We shall also look into possible time dependence of each default rate time series.

2.3 Summary statistics of the home loans default rates

Here we report some statistical features of the default rate data. Tables 2.2.4, 2.2.5 and 2.2.6 show the means, standard deviations, skewnesses, kurtoses and p-values of the Shapiro-Wilk test for normality (abbreviated to SW p-value) for the default rates and transformed values for each risk class. A few remarks on these statistics follow.

Figure 2.2.1 Home loans default rates over months for all risk classes

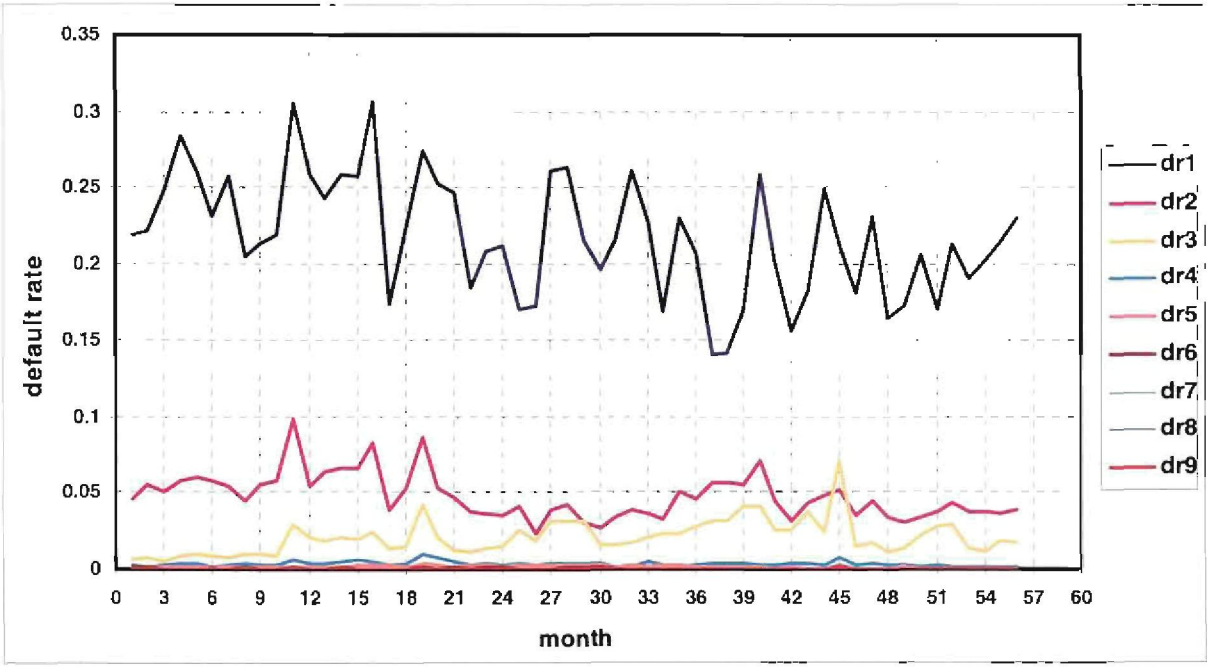


Figure 2.2.2 Probit transformed default rates over months for all risk classes

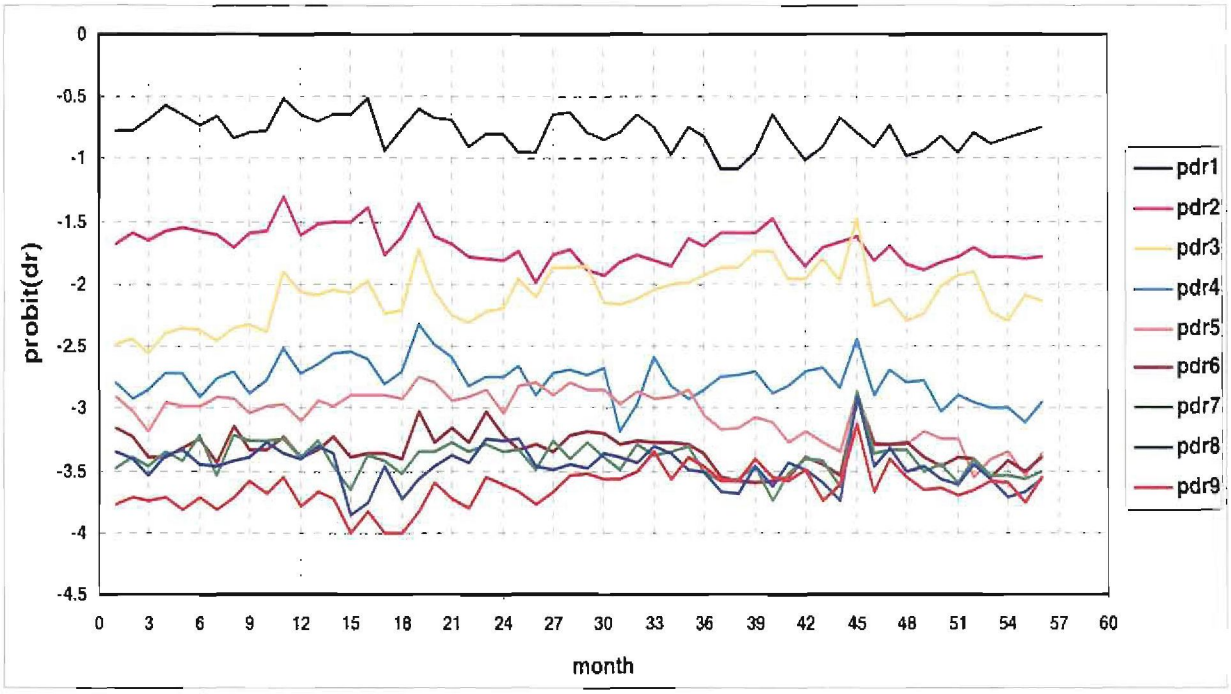
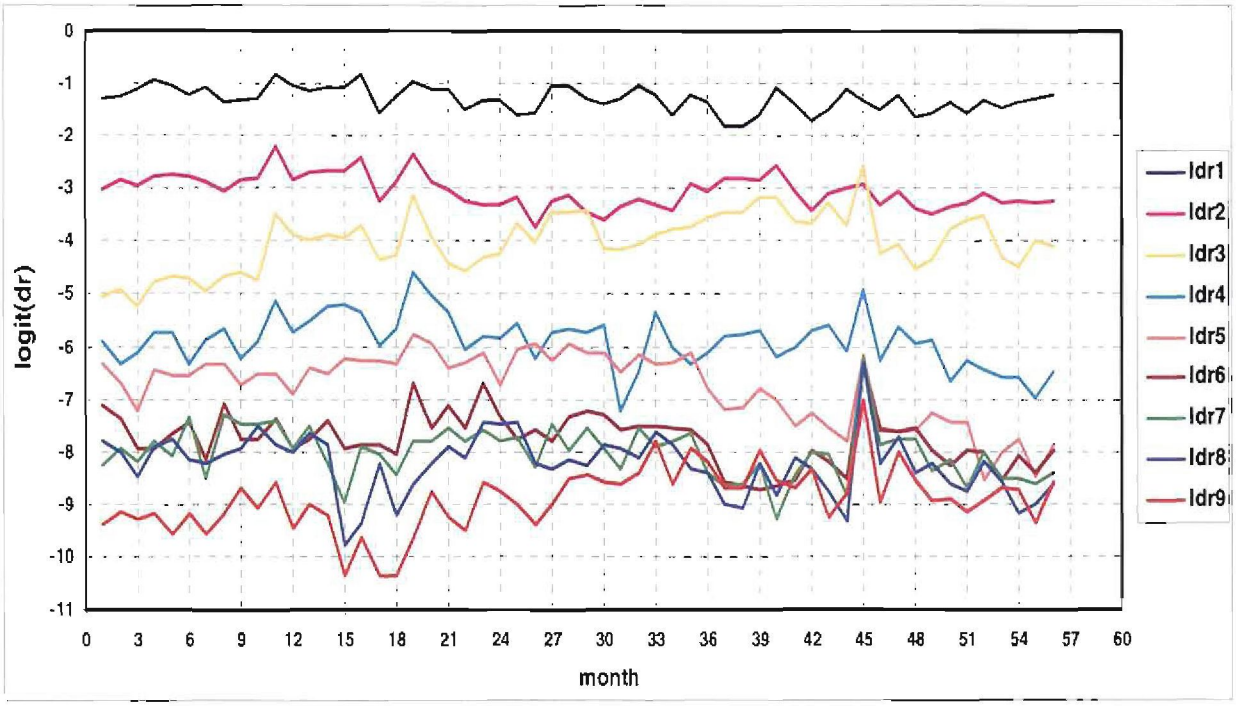


Figure 2.2.3 Logit transformed default rates over months for all risk classes



- The mean default rates decrease monotonically from 21,87% for risk class 1, to 0.016% for risk class 9, again confirming that the risk classes are ordered from highest to lowest default proneness as indicated above. This holds for both the untransformed and the transformed default rates, but the successive decreases are larger for the two transformed series.
- The standard deviations of the untransformed default rates decrease monotonically from 0.03916 for risk class 1, to 0.00013 for risk class 9. This is to be expected since all default rates become smaller as we move from risk class 1 to risk class 9 and therefore their spreads also become smaller. By contrast, the probits' standard deviations do not differ much between the various risk classes, whilst those of the logits show an upward trend towards risk class 9.
- The skewnesses of the untransformed default rates have a small positive value for risk class 1, but tend to increase towards the higher risk classes, especially so for the last three risk classes. By contrast, the skewnesses of the transformed rates do not show a definite tendency but varies around zero, indicating that the transformed rates are more symmetrically distributed.
- The (excess) kurtosis values of the untransformed default rates are quite variable over the risk classes, achieving very high values over the last three risk classes. Again by contrast, the kurtosis values of the transformed default rates are much more constant over the risk classes and also closer to zero.
- For the untransformed default rates, normality is accepted by the Shapiro-Wilk test for only two of the nine risk classes. Again by contrast, normality is accepted for six of the probit transformed risk classes and for eight of the nine logit transformed risk classes. This confirms that transformation of the default rates tends to normalise their distributions. This is in line with what was noted above for the skewnesses and kurtoses.

2.4 Correlation between the home loans default rates

The Pearson correlation coefficients of the default rate data are shown in Table 2.2.7 and for the transformed default rates in Tables 2.2.8 and 2.2.9. From Table 2.2.7 it is evident that adjacent risk classes tend to have highly correlated default rates over time, since the entries close to the diagonal tend to be high (e.g. the highest value of 0.87533 occurs between risk classes 7 and 8). Entries further from the diagonal tend to be lower (or even negative), suggesting that the comovement of default rates of classes far apart in terms of risk proneness is less common. Some exceptions to this remark occur, e.g. there is high

correlation between classes 3 and 7-9. This may be due to the rather prominent upward spike over month 45 visible in Figure 2.2.1. Similar patterns are visible for the transformed default rates in Tables 2.2.8 and 2.2.9.

2.5 Autocorrelations for the home loans default rates

The calculations of the summary statistics and correlations above were aimed at giving an overall feel for the features of the default rates data, not taking into account the time series aspect of the data. Turning to possible time dependence features, Table 2.2.10 below shows the lag 1 autocorrelations of the default rates and their transformed rates. These autocorrelations are noticeably larger for the transformed rates than for the untransformed rates. Overall, the autocorrelations tend to be quite large, except possibly for risk class 7 (in the case of the default rates and also for the probits and logits), as well as class 8 and 9 of the untransformed rates. It is evident that time dependency of the default rates is present, and should be taken into account when doing further analysis on the data.

Table 2.2.4 Summary statistics for the home loans default rates

	dr1	dr2	dr3	dr4	dr5	dr6	dr7	dr8	dr9
Mean	0.21872	0.04731	0.02045	0.00312	0.00138	0.00049	0.00038	0.00031	0.00016
Std Dev	0.03916	0.01481	0.01155	0.00166	0.00073	0.00028	0.00027	0.00025	0.00013
Skewness	0.07687	1.18541	1.70378	1.74984	0.12045	2.27158	4.59272	4.53781	3.71862
Kurtosis	-0.45759	1.94042	5.12548	4.55631	-0.89862	7.40380	28.27431	27.68301	19.36203
SW p-value	0.44860	0.00112	0.00004	0.00002	0.05505	0.00000	0.00000	0.00000	0.00000

Table 2.2.5 Summary statistics for the probit transformed default rates

	pdr1	pdr2	pdr3	pdr4	pdr5	pdr6	pdr7	pdr8	pdr9
Mean	-0.78345	-1.68838	-2.09392	-2.77040	-3.04394	-3.32938	-3.40222	-3.46378	-3.63963
Std Dev	0.13458	0.14205	0.22117	0.16243	0.19945	0.14053	0.13947	0.15845	0.15804
Skewness	-0.15706	0.47127	0.17029	0.10022	-0.74179	0.35478	0.64139	0.24287	0.23945
Kurtosis	-0.45224	0.25506	-0.05837	0.56438	-0.33994	0.62643	2.90464	2.14859	1.61176
SW p-value	0.41536	0.36446	0.89300	0.93537	0.00163	0.18203	0.02172	0.03495	0.09072

Table 2.2.6 Summary statistics for the logit transformed default rates

	ldr1	ldr2	ldr3	ldr4	ldr5	ldr6	ldr7	ldr8	ldr9
Mean	-1.28815	-3.04631	-4.01063	-5.88717	-6.76972	-7.74831	-8.01235	-8.24210	-8.91053
Std Dev	0.23364	0.31009	0.55192	0.49957	0.67727	0.50194	0.50535	0.58694	0.61153
Skewness	-0.21627	0.34472	-0.01585	-0.07286	-0.83096	0.22497	0.41913	0.01658	0.04710
Kurtosis	-0.42751	0.09287	-0.24773	0.54885	-0.13773	0.43162	2.24268	1.68619	1.38660
SW p-value	0.36460	0.60345	0.94371	0.94477	0.00071	0.25417	0.06083	0.05473	0.11412

Table 2.2.7 Pearson correlation coefficients for the home loans default rates

	dr1	dr2	dr3	dr4	dr5	dr6	dr7	dr8	dr9
dr1	1	0.62099	-0.03499	0.36137	0.31761	0.17293	0.10117	0.00765	-0.10862
dr2	0.62099	1	0.21963	0.54373	0.17688	0.07386	0.09299	-0.01775	-0.09161
dr3	-0.03499	0.21963	1	0.45619	0.10417	0.27344	0.42326	0.36451	0.61557
dr4	0.36137	0.54373	0.45619	1	0.53397	0.55761	0.41334	0.35089	0.21644
dr5	0.31761	0.17688	0.10417	0.53397	1	0.57011	0.33162	0.28602	0.04774
dr6	0.17293	0.07386	0.27344	0.55761	0.57011	1	0.75180	0.71901	0.45792
dr7	0.10117	0.09299	0.42326	0.41334	0.33162	0.75180	1	0.87533	0.69375
dr8	0.00765	-0.01775	0.36451	0.35089	0.28602	0.71901	0.87533	1	0.72213
dr9	-0.10862	-0.09161	0.61557	0.21644	0.04774	0.45792	0.69375	0.72213	1

Table 2.2.8 Pearson correlation coefficients for the probit transformed default rates

	pdr1	pdr2	pdr3	pdr4	pdr5	pdr6	pdr7	pdr8	pdr9
pdr1	1	0.58402	-0.07767	0.28517	0.29900	0.24829	0.15411	0.02121	-0.16597
pdr2	0.58402	1	0.13286	0.47667	0.19523	-0.01868	0.04911	-0.11138	-0.19321
pdr3	-0.07767	0.13286	1	0.32949	0.00000	-0.06628	0.04130	-0.05297	0.43939
pdr4	0.28517	0.47667	0.32949	1	0.52429	0.42031	0.37080	0.23251	-0.01711
pdr5	0.29900	0.19523	0.00000	0.52429	1	0.59883	0.42985	0.36292	-0.09817
pdr6	0.24829	-0.01868	-0.06628	0.42031	0.59883	1	0.72713	0.62855	0.14051
pdr7	0.15411	0.04911	0.04130	0.37080	0.42985	0.72713	1	0.73020	0.33860
pdr8	0.02121	-0.11138	-0.05297	0.23251	0.36292	0.62855	0.73020	1	0.43459
pdr9	-0.16597	-0.19321	0.43939	-0.01711	-0.09817	0.14051	0.33860	0.43459	1

Table 2.2.9 Pearson correlation coefficients for the logit transformed default rates

	ldr1	ldr2	ldr3	ldr4	ldr5	ldr6	ldr7	ldr8	ldr9
ldr1	1	0.57438	-0.08661	0.27052	0.29339	0.25585	0.15589	0.02358	-0.16678
ldr2	0.57438	1	0.11327	0.46480	0.19285	-0.03045	0.04168	-0.11828	-0.19717
ldr3	-0.08661	0.11327	1	0.31020	-0.01205	-0.09210	0.01220	-0.08125	0.42372
ldr4	0.27052	0.46480	0.31020	1	0.52361	0.40520	0.36418	0.22122	-0.03220
ldr5	0.29339	0.19285	-0.01205	0.52361	1	0.59507	0.42950	0.36195	-0.10629
ldr6	0.25585	-0.03045	-0.09210	0.40520	0.59507	1	0.72594	0.62093	0.12213
ldr7	0.15589	0.04168	0.01220	0.36418	0.42950	0.72594	1	0.72320	0.32004
ldr8	0.02358	-0.11828	-0.08125	0.22122	0.36195	0.62093	0.72320	1	0.42474
ldr9	-0.16678	-0.19717	0.42372	-0.03220	-0.10629	0.12213	0.32004	0.42474	1

Table 2.2.10 Autocorrelations for the home loan default rates and transformed default rates

Risk class	Default rates (dr)	Probit transformed default rates (pdr)	Logit transformed default rates (ldr)
1	0.34605	0.35506	0.35717
2	0.49029	0.55069	0.55715
3	0.37260	0.59608	0.62541
4	0.33155	0.37690	0.37963
5	0.68527	0.75455	0.75629
6	0.12639	0.36974	0.38896
7	-0.02151	0.10530	0.11548
8	0.02557	0.24891	0.26304
9	0.14326	0.45661	0.47348

2.6 Summary

The analyses of the home loans default rate data above point out that default rates fluctuate at different levels for the different classes, that they tend to be auto-correlated over time within each class and that there is correlation between the rates of different classes. Models to describe the default rates should be able to accommodate these features. Auto-regressive (AR) models are suitable for this purpose and will be discussed in the next chapter.

Chapter 3

Auto-regressive default rate models

3.1. Introduction

As pointed out in Chapter 2, auto-regressive (AR) models can cater for the main features of the home loans default rates data. Before fitting AR models to the data, the default rates need to be transformed appropriately. In Section 3.2 we assume that default takes place via a threshold mechanism and show that this suggests using an inverse distribution function transformation of the default rates. In Section 3.3 we formulate AR models for the transformed default rates and discuss the results of fitting them to our data sets. It turns out that simple $AR(1)$ models fit the data fairly well but it is found that there is some variability in the default rates that cannot be adequately catered for by these models. In an attempt to remedy this situation we extend the $AR(1)$ models by including macro-economic factors in Section 3.4. Again we discuss the results of fitting these extended AR models to our data, but this approach also fails to provide us with satisfactory results. The results in Sections 3.3 and 3.4 are based on univariate fitting methods. As they do not lead to satisfactory conclusions, a multivariate model fitting approach is used in Section 3.5. Here a restricted vector auto-regressive model of order 1 ($VAR(1)$) was fitted to the transformed default rates. Although the $VAR(1)$ model did fit somewhat better than the univariate $AR(1)$ models, the results are still not fully satisfactory. Lastly, Section 3.6 summarises the contents of this chapter.

3.2. Transforming the default rates

Consider a general credit portfolio with obligors categorised into one of K distinct risk classes. Assume that we have data over T time periods on which to base a default rate model. The number of obligors at the start of period t in risk class k is denoted by N_{tk} and the number of defaulters among them over the period by D_{tk} so that the corresponding default rate is $R_{tk} = D_{tk}/N_{tk}$ with k taking values $1,2,\ldots,K$ and t taking values $1,2,\ldots,T$. The R_{tk}'s therefore form a multivariate time series data set, and an appropriate model is required to enable description and prediction of the behaviour of these series over time.

In this dissertation, we restrict attention to large homogeneous portfolios. This means that there are many obligors in each of the risk classes, and that all obligors within each risk class are similar to each other in terms of default probabilities. The "large portfolio" assumption is fair for our data set, especially for the higher numbered risk classes. The number of obligors in risk class 1 varies around the level of 3000 and this number increases with increase in risk

class number, reaching the level of about 100 000 in risk class 9. Whether "homogeneous portfolios" is an appropriate assumption for our data set can only be judged by the extent to which the models fit the data well.

Default models of "threshold" type are widely used in credit risk analysis (see e.g. Schönbucher, 2005:305). Such models assume the existence of a so-called "asset variable" for each obligor which is an unobserved or latent variable and is to be interpreted in a wide sense as the obligor's "creditworthiness" or "ability to pay". It is further assumed that default of the obligor is triggered by the event that the asset variable falls below a certain threshold or critical level. More specifically, let A_{tkn} denote the asset variable of the n-th obligor in risk class k over the period t. To be consistent with the assumption that the obligors are similar, we assume that the A_{tkn}'s are independent and identically distributed for all n's, with a common distribution function of the form $G((a-\nu_{tk})/\tau_{tk})$. Here ν_{tk} and τ_{tk} are location and scale parameters respectively which may vary over different risk classes and at different time periods, and G is some distribution function. If we assume that c_{tk} is the threshold value for risk class k over period t (the same for all obligors, i.e. not dependent on n, in view of the homogeneity assumption), then the n-th obligor in risk class k defaults over period t if $A_{tkn} \le c_{tk}$. Hence this obligor's default probability is $P(A_{tkn} \le c_{tk}) = G((c_{tk}-\nu_{tk})/\tau_{tk})$ which does not depend on n, again consistent with the assumed homogeneity of obligors within risk classes. The total number of defaults D_{tk} in risk class k over period t is then binomially distributed, with parameters N_{tk} and $G((c_{tk}-\nu_{tk})/\tau_{tk})$. By the Law of large numbers, $R_{tk} = D_{tk}/N_{tk} \to G((c_{tk}-\nu_{tk})/\tau_{tk})$ with probability one, as $N_{tk} \to \infty$. Under our assumption that the portfolio consists of large numbers of obligors in each risk class, we can take it that this limit is an equality, i.e. that $R_{tk} = G((c_{tk}-\nu_{tk})/\tau_{tk})$. This implies that

$$G^{-1}(R_{tk}) = (c_{tk}-\nu_{tk})/\tau_{tk} = Y_{tk} \qquad (3.2.1)$$

where Y_{tk} is a variable that combines the effects of the threshold and the location and scale parameters of the asset variable distribution in risk class k over period t. Since the default rate R_{tk} is only affected by the combined variable Y_{tk}, rather than being affected separately by the effects of the threshold and the location and scale parameters, it seems reasonable to base modelling of R_{tk} only on the variable Y_{tk}.

This is also reasonable from another perspective: R_{tk} is a fraction so that $0 \le R_{tk} \le 1$. Traditional statistical modelling is strongly based on normality assumptions, which implicitly assume that the relevant variables vary over $(-\infty, +\infty)$. Thus, working with R_{tk} directly makes

the use of traditional statistical modelling difficult. If G is a distribution function supported on $(-\infty,+\infty)$, then $Y_{tk}=G^{-1}(R_{tk})$ can vary over $(-\infty,+\infty)$ and the transformation alleviates this problem. Once we have modelled Y_{tk} and obtained a fitted or predicted value $\hat{Y}_{tk}$, we can transform back to get a corresponding fitted or predicted value $\hat{R}_{tk}=G(\hat{Y}_{tk})$ for R_{tk}.

The arguments above do not indicate what should be chosen for the distribution function G. We will use both the standard normal and the logistic distribution functions, denoted by $G(x)=\Phi(x)$ and $G(x)=1/(1+e^{-x})$ respectively. The choice to use the standard normal distribution $G(x)=\Phi(x)$ is often made for the so-called *factor models* in credit risk (Schönbucher, 2005:305) and both this choice and other choices for the function G are extensively dealt with in the theory of generalised linear models, where they are known as the "link" function of the model (McCullagh & Nelder, 1989:108).

3.3 AR models for the transformed default rates

Following transformation of R_{tk} to Y_{tk}, the simplest auto-regressive model of lag length p (abbreviated $AR(p)$) is of the form

$$Y_{tk}=\alpha_k+\sum_{i=1}^{p}\beta_{ki}(Y_{t-i,k}-\alpha_k)+e_{tk}=\alpha_k^*+\sum_{i=1}^{p}\beta_{ki}Y_{t-i,k}+e_{tk} \qquad (3.3.1)$$

where α_k represents an intercept, β_{ki} is the lag i AR coefficient for risk class k and $\alpha_k^*=\alpha_k(1-\sum_{i=1}^{p}\beta_{ki})$ is an alternative form in which the intercept is often represented. We prefer the first form of the model in (3.3.1) since the intercept in that form represents a level around which the Y_{tk}'s fluctuate and therefore has more intuitive interpretability. Also, e_{tk} is the error component for risk class k over time period t and the standard distributional assumption is that the e_{tk}'s are $N(0,\sigma_k^2)$ distributed, independently over t but possibly correlated over risk classes k. We write $Cov(e_{tk},e_{tl})=\sigma_{kl}$.

3.3.1 Fitting of the AR models to the home loans data

Auto-regressive models of the form (3.3.1) were fitted to the probit and logit transformed home loans default rates. To cater for possible seasonal effects on the lag terms of the model, lag lengths of up to 13 months were included in the AR models considered. Thus, for each risk class $k=1,2,\ldots,9$ we sought the best fitting AR model among choices of $AR(1)$ to $AR(13)$ models. Several fitting methods and criteria were used in the process and will be discussed in more detail below. The residuals that were obtained from the model fits were further investigated to determine whether model assumptions were met. These include the residuals' normality and homogeneity of their variances.

3.3.2 Methods used in the fitting process

Two SAS tools were used in the fitting process, namely the Time Series Forecasting System (TSFS) as well as PROC AUTOREG. The TSFS provides a user-friendly interface for fitting and forecasting many model types, including AR models. The Autoreg procedure also fits AR models on time series data, then provides some output to enable model selection and is thus useful in the modelling of our default rates. Both TSFS and Autoreg use the first form in (3.3.1). They handle only univariate time series and therefore, in the application to the home loans data, the default rates of the individual risk classes were treated separately.

To allow for possible yearly (twelve-monthly) lag effects, we consider lags up to length 13. The TSFS allows the user to specify a list of models to be fitted to the data after which the best model can be selected according to a specified criterion. Our list consisted of the $AR(1)$ to $AR(13)$ models. The Schwarz's Bayesian information criterion (denoted as SBC, see for e.g. SAS Institute Inc., 2004:544) was used as the criterion of best fit since this is a widely used and standard criterion that penalises a model that includes too many parameters (lag coefficients in this case). It is calculated as $SBC = -2\ln(L) + \ln(T)q$, where L is the likelihood function evaluated at the parameter estimates, T is the number of observations and $q = p + 2$ is the number of estimated parameters (p lag coefficients, the intercept and error variance). Thus, the $AR(p)$ model with smallest SBC value will be chosen as the best model for each risk class. Sometimes the SBC value of the best model does not differ much from that of the second best or even that of the third or lower best models. The results for the best models (up to the third best) are reported below.

PROC AUTOREG was also applied for model selection as follows: Starting with an $AR(13)$ model, stepwise selection was done through using PROC AUTOREG's backstep option in order to decide on the significant lag lengths to include in the model. The stepwise selection sequentially removes the lag length with the highest P-value from the model until only the significant lag lengths at level $\alpha = 5\%$ remain. From these we can select the one lag length that is most significant if we desire a model with only one lag length. In the same way, we can select the two most significant lag lengths if we want a model with two lag lengths, etc. For some classes, the stepwise selection showed no lag lengths to be significant at the significance level $\alpha = 5\%$, so another fitting process was initiated in the same way, but with a higher value of α in order to end with at least one lag length. The results of this procedure may differ from those of TSFS in that only significant individual lag lengths are chosen and not necessarily all lag lengths from 1 to the highest lag length p, i.e. it is not necessarily an $AR(p)$ model that is produced in this way.

3.3.3 Results from the fits

The results from the model selection obtained from TSFS are shown in Table 3.3.1 below, and the results from PROC AUTOREG in Table 3.3.2.

Consider first Table 3.3.1. The first column shows the type of transformed default rate and the risk class number which is dealt with in the corresponding row. Columns 2 to 4 identify the best, second best and third best $AR(p)$ models and columns 5 to 7 show their corresponding SBC values. We note that there is very little difference between the models chosen for the probit and logit transformed default rates. The only difference occurred in risk class 1, where the second and third best models are interchanged. Further, the $AR(1)$ model is chosen as best for five out of the nine risk classes. It is second best for another two classes and for these cases, the SBC values of the best and second best models are virtually the same. The $AR(1)$ model is third best for one further class. It is among the top three best models for all risk classes except for class 5, for which it happens to be the fifth best model (not shown in the table) with an SBC of 214.76. Although there may not be a compelling reason to use the same model for all risk classes it would certainly simplify matters to use one model for all cases. An $AR(1)$ would clearly be the preferred model since it is already the best or close to the best for eight of the nine cases; additionally it has the desirable property of being parsimonious in that only one lag length and thus very few parameters are involved.

Next consider Table 3.3.2. Its first column is similar to the first column of Table 3.3.1. Columns 2 to 4 identify the best, second best and third best lag length to include in the model using as criterion the most significant lag length according to low P-values. Columns 5 to 7 show the corresponding P-values. We note that there is no difference between the models chosen for the probit and logit transformed default rates in this case. The lag length 1 model (equivalent to an $AR(1)$ model) was identified as the best model for six of the nine risk classes. For risk class 9, the lag length 1 was chosen as second best and was also seen as highly significant. Only for risk classes 6 and 7 were the inclusion of lag length 1 not considered important. Although the backward selection process is quite different from the SBC criterion selection approach discussed above, it largely supports the impression that an $AR(1)$ model would be a simple, parsimonious model that would be a reasonable choice for a single model to fit to all the risk classes.

Table 3.3.1 TSFS results: Three best AR models and SBC values for each risk class

	Model			SBC value		
PROBITS	**Best model**	**2nd best**	**3rd best**	**Best model**	**2nd best**	**3rd best**
pdr1	AR(1)	AR(2)	AR(3)	-220.28	-216.91	-216.85
pdr2	AR(1)	AR(2)	AR(3)	-227.21	-225.23	-224.14
pdr3	AR(1)	AR(2)	AR(3)	-184.45	-182.80	-178.78
pdr4	AR(1)	AR(2)	AR(3)	-200.44	-196.92	-195.22
pdr5	AR(3)	AR(4)	AR(2)	-222.74	-218.90	-218.14
pdr6	AR(2)	AR(1)	AR(3)	-217.77	-217.57	-214.83
pdr7	AR(2)	AR(1)	AR(3)	-209.79	-209.57	-207.42
pdr8	AR(1)	AR(2)	AR(3)	-198.81	-195.44	-191.62
pdr9	AR(3)	AR(2)	AR(1)	-211.35	-210.20	-208.71
LOGITS	**Best model**	**2nd best**	**3rd best**	**Best model**	**2nd best**	**3rd best**
ldr1	AR(1)	AR(3)	AR(2)	-159.70	-156.39	-156.37
ldr2	AR(1)	AR(2)	AR(3)	-141.93	-139.85	-138.76
ldr3	AR(1)	AR(2)	AR(3)	-86.97	-85.14	-81.12
ldr4	AR(1)	AR(2)	AR(3)	-76.98	-73.45	-71.87
ldr5	AR(3)	AR(4)	AR(2)	-88.97	-85.11	-84.12
ldr6	AR(2)	AR(1)	AR(3)	-78.77	-78.46	-75.83
ldr7	AR(2)	AR(1)	AR(3)	-68.35	-68.08	-66.04
ldr8	AR(1)	AR(2)	AR(3)	-55.21	-51.77	-48.01
ldr9	AR(3)	AR(2)	AR(1)	-63.28	-62.30	-60.98

Table 3.3.2 PROC AUTOREG results: Lag lengths for the 3 best models for each risk class

	Model			P-values		
PROBITS	**Best model**	**2nd best**	**3rd best**	**Best model**	**2nd best**	**3rd best**
pdr1	lag1	lag11	lag5	0.0006	0.0027	0.0072
pdr2	lag1	lag5	lag10	0.0005	0.0218	0.0245
pdr3	lag1	lag5	lag2	<0.0001	0.1660	0.2709
pdr4	lag1	lag4	lag7	0.0046	0.0604	0.1872
pdr5	lag1	lag3	lag6	0.0003	0.0004	0.3149
pdr6	lag2	lag11	lag8	0.0060	0.0893	0.1300
pdr7	lag2	lag3	lag5	0.0436	0.0660	0.0754
pdr8	lag1	lag5	lag8	0.0643	0.0792	0.3163
pdr9	lag3	lag1	lag2	0.0071	0.0241	0.1136
LOGITS	**Best model**	**2nd best**	**3rd best**	**Best model**	**2nd best**	**3rd best**
ldr1	lag1	lag11	lag5	0.0005	0.0027	0.0079
ldr2	lag1	lag5	lag10	0.0004	0.0296	0.0252
ldr3	lag1	lag5	lag2	<0.0001	0.1825	0.3037
ldr4	lag1	lag4	lag7	0.0043	0.0605	0.1884
ldr5	lag1	lag3	lag6	0.0004	0.0003	03352
ldr6	lag2	lag11	lag8	0.0042	0.0775	0.1240
ldr7	lag2	lag3	lag5	0.0406	0.0613	0.0733
ldr8	lag1	lag5	lag8	0.0501	0.0787	0.3290
ldr9	lag3	lag1	lag2	0.0077	0.0172	0.1268

3.3.4 Parameter estimates

Restricting our attention to $AR(1)$ models we can write (3.3.1) in the form

$$Y_{tk} = \alpha_k + \beta_k (Y_{t-1,k} - \alpha_k) + e_{tk} \qquad (3.3.2)$$

where α_k represents the intercept level and β_k is the lag 1 AR coefficient for risk class k. Table 3.3.3 shows the estimates of the parameters of the $AR(1)$ models with the standard errors in brackets, for each risk class and for both sets of transforms. The intercept level estimates clearly show how the default rates tend to fluctuate at systematically decreasing levels with increasing risk class numbers. These level parameters differ substantially between the two transform types, which is an inherent property of the scale differences of the two transforms. The AR coefficients of the probit and the logit transforms are quite similar. These coefficients are all positive, suggesting a carry-over effect from month to month of the factors that cause default rate changes. The estimates of the AR coefficients vary substantially between the risk classes with class 7 having the smallest and class 5 having the largest values in both cases. However, the standard errors of the estimates are quite large so that the differences between classes may be at least to some extent due to sampling effects.

Table 3.3.3 Parameter estimates with the standard errors in brackets for $AR(1)$ model fits to each risk class of transformed data

	Parameter estimates	
PROBITS	**Level α_k**	**AR-coeff β_k**
pdr1	-0.7829 (0.0260)	0.3551 (0.1270)
pdr2	-1.6902 (0.0346)	0.5507 (0.1133)
pdr3	-2.1049 (0.0582)	0.5961 (0.1083)
pdr4	-2.7725 (0.0322)	0.3769 (0.1275)
pdr5	-3.0534 (0.0696)	0.7546 (0.0925)
pdr6	-3.3282 (0.0276)	0.3677 (0.1264)
pdr7	-3.4026 (0.0209)	0.1053 (0.1361)
pdr8	-3.4637 (0.0274)	0.2489 (0.1322)
pdr9	-3.6402 (0.0343)	0.4566 (0.1209)
LOGITS	**Level α_k**	**AR-coeff β_k**
ldr1	-1.2872 (0.0452)	0.3572 (0.1269)
ldr2	-3.0501 (0.0762)	0.5572 (0.1127)
ldr3	-4.0424 (0.1520)	0.6254 (0.1050)
ldr4	-5.8937 (0.0994)	0.3796 (0.1274)
ldr5	-6.8026 (0.2373)	0.7563 (0.0923)
ldr6	-7.7437 (0.1099)	0.3890 (0.1252)
ldr7	-8.0139 (0.0764)	0.1155 (0.1359)
ldr8	-8.2415 (0.1029)	0.2630 (0.1316)
ldr9	-8.9128 (0.1355)	0.4735 (0.1197)

Table 3.3.4 provides the estimates of the error standard deviations σ_k as well as the estimates of the error correlations between the risk classes, with the latter computed as the correlation coefficients of the $AR(1)$ model residuals. The error standard deviations differ substantially between the two transform types. This is again a reflection of the scale differences between the two transforms. The error variances tend to become larger towards the higher numbered risk classes for the logits but this trend is absent from the probits. The correlations of the probits and the logits are nearly the same for all pairs of risk classes. They are also generally positive, assume the highest values for adjacent risk classes and tend to decrease as the risk class numbers move further apart. This behaviour seems intuitively reasonable.

Table 3.3.4 $AR(1)$ model error standard deviations and correlations

	StDev	Correlations								
PROBITS		pdr1	pdr2	pdr3	pdr4	pdr5	pdr6	pdr7	pdr8	pdr9
pdr1	0.1267	1.0000	0.6484	0.2005	0.2261	0.0929	0.1283	0.0520	-0.0618	0.1168
pdr2	0.1189	0.6484	1.0000	0.5147	0.3848	0.1973	0.1008	0.0887	-0.0791	0.1685
pdr3	0.1750	0.2005	0.5147	1.0000	0.4909	0.3522	0.3372	0.2551	0.2085	0.4502
pdr4	0.1517	0.2261	0.3848	0.4909	1.0000	0.5277	0.4762	0.3737	0.3271	0.2787
pdr5	0.1331	0.0929	0.1973	0.3522	0.5277	1.0000	0.5321	0.4209	0.4048	0.3440
pdr6	0.1297	0.1283	0.1008	0.3372	0.4762	0.5321	1.0000	0.7225	0.6063	0.3968
pdr7	0.1396	0.0520	0.0887	0.2551	0.3737	0.4209	0.7225	1.0000	0.7333	0.4741
pdr8	0.1539	-0.0618	-0.0791	0.2085	0.3271	0.4048	0.6063	0.7333	1.0000	0.5270
pdr9	0.1407	0.1168	0.1685	0.4502	0.2787	0.3440	0.3968	0.4741	0.5270	1.0000
LOGITS		ldr1	ldr2	ldr3	ldr4	ldr5	ldr6	ldr7	ldr8	ldr9
ldr1	0.2197	1.0000	0.6436	0.2085	0.2127	0.0841	0.1297	0.0519	-0.0605	0.1283
ldr2	0.2582	0.6436	1.0000	0.5190	0.3643	0.1830	0.0877	0.0844	-0.0848	0.1786
ldr3	0.4242	0.2085	0.5190	1.0000	0.4678	0.3275	0.3257	0.2390	0.1878	0.4331
ldr4	0.4660	0.2127	0.3643	0.4678	1.0000	0.5183	0.4603	0.3651	0.3142	0.2663
ldr5	0.4508	0.0841	0.1830	0.3275	0.5183	1.0000	0.5255	0.4089	0.3920	0.3376
ldr6	0.4595	0.1297	0.0877	0.3257	0.4603	0.5255	1.0000	0.7151	0.5954	0.3900
ldr7	0.5053	0.0519	0.0844	0.2390	0.3651	0.4089	0.7151	1.0000	0.7246	0.4651
ldr8	0.5679	-0.0605	-0.0848	0.1878	0.3142	0.3920	0.5954	0.7246	1.0000	0.5140
ldr9	0.5388	0.1283	0.1786	0.4331	0.2663	0.3376	0.3900	0.4651	0.5140	1.0000

3.3.5 Testing normality of residuals

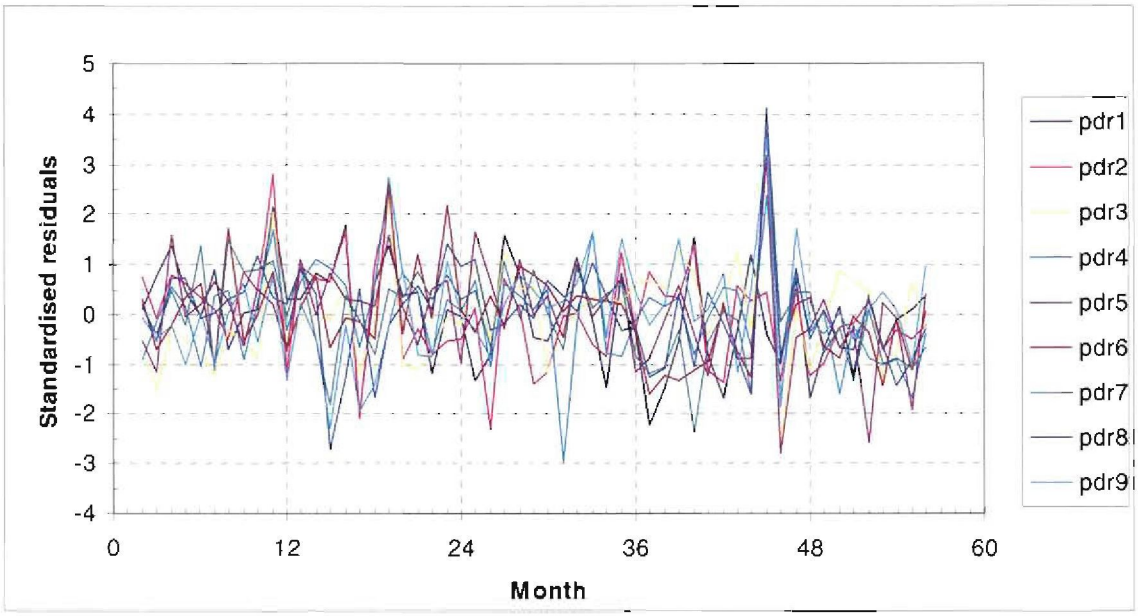
Table 3.3.5 below provides information regarding testing normality of the residuals of the $AR(1)$ models fitted above. Again the first column shows which default rate transformation and risk class is dealt with in the corresponding row. The second column shows the P-values of the Jarque-Bera normality test (SAS Institute Inc., 2004:549) on the $AR(1)$ model residuals. It is clear that the normality assumption is questionable for six of the nine risk classes for the probit transforms and for five of the nine risk classes for the logit transforms. To gain some

further insights into this finding, the residuals were plotted after first standardising them by dividing each residual by its standard deviation. Figure 3.3.1 shows the results for the probit transforms. The equivalent plots for the logit transforms are quite similar and are thus not shown. Clearly, for many of the risk classes the residuals at month 45 are extremely outlying. Leaving out the residuals of month 45, we recalculated the P-values of the Jarque-Bera normality test. These results are shown in the third column of Table 3.3.5. Only two small P-values remain for the probit case and three for the logit case, taking 5% as cut-off. That the default rates of month 45 appear outlying for many risk classes when compared to the other months is also apparent from Figures 2.2.2 and 2.2.3. There are other residuals in Figure 3.3.1 that also appear somewhat outlying and the P-values of the normality tests will become larger if they are also left out. However, we must also recognise that the apparent outlying residuals may be due to model inadequacies rather than data anomalies, and it may be the case that the simple $AR(1)$ models fitted here do not take sufficient factors into account to produce adequately fitting models. Rather than blaming the data for the lack of fit of the models used this far, we shall pursue below the extension of the models to more sophisticated alternatives in order to obtain better descriptions of the processes underlying our data.

Table 3.3.5 P-values of Jarque-Bera normality tests on residuals of $AR(1)$ models

PROBITS	All	Without month 45 residual
pdr1	0.7096	0.6740
pdr2	0.6253	0.6190
pdr3	0.0581	0.6387
pdr4	0.2734	0.1717
pdr5	0.0226	0.0237
pdr6	<0.0001	0.0204
pdr7	<0.0001	0.4539
pdr8	<0.0001	0.0709
pdr9	0.0050	0.8180
LOGITS	**All**	**Without month 45 residual**
ldr1	0.6599	0.6234
ldr2	0.8829	0.8820
ldr3	0.2111	0.6943
ldr4	0.2058	0.0955
ldr5	0.0086	0.0040
ldr6	<0.0001	0.0492
ldr7	<0.0001	0.3678
ldr8	<0.0001	0.0224
ldr9	0.0482	0.6562

Figure 3.3.1 Standardised residuals of $AR(1)$ models of probit transformed default rates



3.3.6 Testing homoscedasticity of residuals

Both the Portmanteau Q-test and LM-test (SAS Institute Inc., 2004:548) were used to test for the assumption of homogeneous error variances. To save space Tables 3.3.6 and 3.3.7 show the P-values for the tests only at some selected orders. A few smallish P-values are found for risk class 5 only; thus the results suggest that the homoscedasticity assumption is acceptable overall.

Table 3.3.6 P-values of the Q-test for homoscedasticity of residuals of $AR(1)$ models

PROBITS	Q(1)	Q(4)	Q(8) c	Q(12)
pdr1	0.5808	0.6999	0.8739	0.8278
pdr2	0.7387	0.8971	0.8142	0.9228
pdr3	0.1162	0.4228	0.5917	0.7131
pdr4	0.6122	0.8097	0.8513	0.3409
pdr5	0.0683	0.3586	0.3286	0.5536
pdr6	0.9925	0.8466	0.9640	0.9845
pdr7	0.8965	0.9181	0.8760	0.9744
pdr8	0.4869	0.8584	0.9885	0.9978
pdr9	0.4116	0.4548	0.5578	0.6203
LOGITS	**Q(1)**	**Q(4)**	**Q(8)**	**Q(12)**
ldr1	0.5947	0.7223	0.8900	0.8305
ldr2	0.8217	0.9232	0.8658	0.9474
ldr3	0.2050	0.5660	0.5181	0.6600
ldr4	0.5456	0.8051	0.8472	0.4085
ldr5	0.0951	0.4502	0.2970	0.5019
ldr6	0.9845	0.8430	0.9490	0.9780
ldr7	0.9020	0.9000	0.7772	0.9339
ldr8	0.4600	0.8375	0.9843	0.9959
ldr9	0.4694	0.4034	0.5007	0.5643

Table 3.3.7 P-values of the LM-test for homoscedasticity of residuals of $AR(1)$ models

PROBITS	LM(1)	LM(4)	LM(8)	LM(12)
pdr1	0.6527	0.8230	0.9538	0.9767
pdr2	0.7283	0.9167	0.7807	0.9072
pdr3	0.1219	0.4862	0.6820	0.7586
pdr4	0.6266	0.8693	0.9286	0.4256
pdr5	0.0722	0.2168	0.2289	0.0653
pdr6	0.9878	0.8812	0.9800	0.9892
pdr7	0.8797	0.9524	0.9150	0.8432
pdr8	0.4857	0.9005	0.9949	0.9997
pdr9	0.4054	0.4398	0.6501	0.6654
LOGITS	**LM(1)**	**LM(4)**	**LM(8)**	**LM(12)**
ldr1	0.6664	0.8355	0.9608	0.9736
ldr2	0.8081	0.9465	0.8054	0.9311
ldr3	0.2093	0.6374	0.6328	0.6891
ldr4	0.5605	0.8608	0.9250	0.5017
ldr5	0.0994	0.2663	0.1812	0.0368
ldr6	0.9643	0.8801	0.9709	0.9835
ldr7	0.8825	0.9418	0.8369	0.7047
ldr8	0.4584	0.8797	0.9924	0.9994
ldr9	0.4586	0.3651	0.5715	0.6288

3.3.7 Testing independence of residuals

Table 3.3.8 shows the P-values of the Durbin-Watson test (SAS Institute Inc., 2004:546) of zero auto-correlation against the alternative of positive auto-correlation. No P-values close to zero were found, suggesting that positive auto-correlation is absent. Since the P-values for testing for zero auto-correlation against the alternative of negative auto-correlation is just one minus the values shown in the table, the large P-values shown for risk class 5 suggest that there may be negative auto-correlation in that specific case.

Table 3.3.8 P-values of the Durbin-Watson test for independence of residuals of $AR(1)$ models

PROBITS	DW (Probits)	LOGITS	DW (Logits)
pdr1	0.3830	**ldr1**	0.3787
pdr2	0.7743	**ldr2**	0.7719
pdr3	0.8520	**ldr3**	0.8557
pdr4	0.5783	**ldr4**	0.5766
pdr5	0.9699	**ldr5**	0.9710
pdr6	0.7852	**ldr6**	0.7986
pdr7	0.5671	**ldr7**	0.5756
pdr8	0.5838	**ldr8**	0.5843
pdr9	0.8407	**ldr9**	0.8443

3.3.8 Summary of Section 3.3

We have shown that the simple $AR(1)$ models go a long way towards describing the home loans transformed default rate data. However, some of the residuals are too large to satisfy the normality assumption and this suggests that there may be additional economic factors influencing the errors over and above what can be accounted for by the $AR(1)$ models. Change of interest rate is a factor that immediately comes to mind: an increase in interest rate will make it more difficult for obligors to meet their payment obligations and thus lead to increases in default rates. Other factors may also play a role. In Section 3.4 we extend the AR models considered so far by incorporating such economic factors.

3.4 Extended AR models for the transformed default rates

The inclusion of macro-economic variables in default rate modelling has been considered by a number of authors. Among these are Nickell, Perraudin and Varotto (2000) as well as Bangia, Diebold, Kronimus, Schagen and Schuermann (2002) who found that the business cycle has a definite influence on default rates (and more generally transition rates between risk classes). In a US context, Bucay and Rosen (2001) studied models that incorporate variables such as industrial production, stock index, consumer price index, retail sales, unemployment level, three-month treasury bill, and short-term, medium-term and long-term government bond yield. In our context we collected data on the CPIX inflation rate (denoted by "INF" below), the three-month bank acceptance rate (denoted by "BAR") and the monthly number of insolvencies (denoted by "INS"). Here we study AR models that were extended to include these variables. Bucay and Rosen (2001) take the yearly returns on the economic variables as the actual factors to incorporate in the models. Presumably the argument is that it is the changes in the variables rather than the actual levels of the variables that act as drivers of default rates. We follow them in this regard and therefore define the "factor" value corresponding to the "variable" for month t by the "return" formula

$$factor_t = 100 \times ln(variable_t / variable_{t-12}).$$

We will thus refer to the changes in inflation as the "inflation factor" and similarly for the bank rate and insolvency.

3.4.1 Extended AR models

The general form of the $AR(1)$ model extended to include observed macro-economic factors is

$$Y_{tk} = \alpha_k + \beta_k (Y_{t-1,k} - \alpha_k) + \sum_{j=1}^{J} \sum_{l=L'_{kj}}^{L''_{kj}} \gamma_{kjl} x_{t-l,j} + e_{tk} \qquad (3.4.1)$$

where α_k and β_k are as before, $x_{t,j}$ is value of the j-th economic factor at time t and γ_{kjl} is the coefficient for this j-th economic factor with lag l in risk class k. Also L'_{kj} and L''_{kj} are

two integers with $L'_{kj} \le L''_{kj}$ and only lag lengths l with $L'_{kj} \le l \le L''_{kj}$ are included in the model. The reason for this set-up is that it may take some months before a factor begins to take effect; the effect may then persist for a while after which it ceases to be active. The interval $L'_{kj} \le l \le L''_{kj}$ therefore indicates the active lag interval and this may be different for different risk classes and economic factors, hence depending on k and j. Finally, e_{tk} is the error component for risk class k over time period t, and the standard distributional assumption is as before - that these e_{tk}'s are $N(0,\sigma_k^2)$ distributed, independently over t but possibly correlated over risk classes with $Cov(e_{tk}, e_{tl}) = \sigma_{kl}$.

3.4.2 Fitting the extended AR models

The following strategy was used to fit extended AR models of the form (3.4.1) to the home loans default rates data. The appropriate lag interval is not known beforehand and must be determined as part of the model selection process. We first need to decide on a maximum value for L''_{kj} and we fixed this at 18, i.e. we assume that all economic factor shocks will have petered out after 18 months. Ideally we would have preferred to make this bound very large but this would mean that an impractically large number of models need to be considered. Even for this choice there are 171 pairs of choices of L'_{kj} and L''_{kj} with $1 \le L'_{kj} \le L''_{kj} \le 18$.

We next draw up three lists of models. The INS list consists firstly of the model with only an $AR(1)$ term and then another 171 models having the $AR(1)$ term combined with each of the possible lag interval terms of the INS factor. We refer to these as the $AR(1)$-INS extended models. Similarly, the INF and BAR lists consist of 172 models based on the INF and BAR factors respectively and these are referred to as the $AR(1)$-INF and $AR(1)$-BAR extended models. Each of these lists was then submitted to TSFS to fit and rank the models according to the SBC criterion.

Table 3.4.1 shows the best three models for each extension factor. Column 2 identifies the best model among the $AR(1)$-INS extended list. In all these cases, the interval consisted of only a single lag value (i.e. $L'_{kj} = L''_{kj}$). The lag value is followed by a "+" or "–" sign, indicating whether the relevant regression coefficient turned out to be positive or negative. The third column identifies the best models among the $AR(1)$-INF extended list. Some of these entries indicate that a genuine lag interval was involved, i.e. the pdr1 entry "5+,6-" means that the INF factor was included at lags 5 and 6 and the coefficient of the lag 5 term was positive, while that of the lag 6 term was negative. The blank entry for ldr3 means that only an AR(1) term was involved, i.e. no actual INF extension. The fourth column has a similar interpretation, where BAR replaces the INF factor. Columns 5 to 7 and 8 to 10 refer to the second and third best

model choices in each category and are interpreted in the same way. We note that the entries for the logits are broadly similar to those of the probits, especially if one allows switches among the model rankings.

The results are disappointing in several respects:

- No convincing patterns of the lag intervals seem to be present. For example, for the INS factor, the intervals consist of single lag values but these seem to vary randomly from risk class to risk class with no visible trend. One would expect that there will be some trend in the lag interval since the obligors become less risky with increasing risk class numbers. Similar remarks hold for the lag intervals of the INF and BAR factors.
- For the best model, five of the lag coefficients of the INS factor have positive signs, and the other four have negative signs. This carries conflicting impressions of the direction in which this factor affects the default rates which is hard to explain. Again, for the BAR factor, four of the signs of the lag coefficients are positive, and six are negative which is intuitively unreasonable since one would expect increasing interest rates to imply an increase in default rates. For the INF factor, most of the signs are positive which seems more reasonable.
- Looking also at the second and third best models does not seem to help much to identify economically meaningful patterns in the results.

We further computed the residuals corresponding to each of the *best* models listed in Table 3.4.1 and applied the Jarque-Bera normality test to the residuals. Table 3.4.2 shows these results. Comparing this table to Table 3.3.5, it is evident that these P-values are quite similar and in particular normality is still rejected for risk classes 5 to 9 for all three (INS, INF and BAR) factors. The P-value increased substantially only in the case of risk class 3 when using the best $AR(1)$-INF model. Thus if we extend the $AR(1)$ model with any of the three economic factors individually as we did above, little benefit above the simple $AR(1)$ was obtained in terms of improved fit.

Further extensions of the $AR(1)$ model were done by including two, or even all three economic factors simultaneously and even longer lists of candidate models were submitted to TSFS for selection. It turned out that combinations of factors were not selected above models with single factor extensions. Therefore, this effort was not fruitful.

It is known that the SBC criterion tend to produce more parsimonious models than other criteria such as the AIC (Akaike Information Criterion). This criterion is defined by $AIC = -2\ln(L) + 2q$. Its penalty factor 2 on the number of parameters (q) is less than half that of $SBC = -2\ln(L) + \ln(T)q = -2\ln(L) + \ln(56)q = -2\ln(L) + 4.03q$ in the home loans data. It may be thought that the use of the SBC criterion is the cause of the disappointing results

found above. To see whether this could be the case, we repeated the model selection process using the AIC in stead of the SBC criterion. Table 3.4.3 and 3.4.4 show the results. Table 3.4.3 clearly shows that much less parsimonious models were selected and that the lag intervals tended to be substantially longer. Because of this, the results are presented by indicating the lag intervals by their begin and end points only. There still does not seem to be a systematic trend or economically meaningful pattern in the choices of the lag intervals. Also the signs of the coefficients of the lagged factors still vary rather randomly between negative and positive (not shown in Table 3.4.3). Finally comparing the P-values of Table 3.4.4 with those of Tables 3.4.2 and 3.3.5 there does not seem to be much difference - so that little improvement in fit is evident.

To summarise, extending the $AR(1)$ models with the economic factors considered here and using the model selection approach discussed above do not yield fruitful results in terms of describing the home loans default rate data better than just the $AR(1)$ models alone.

Table 3.4.1 TSFS results with SBC criterion: Three best models consisting of an $AR(1)$ model extended with a lag interval for each factor. For example, the entry 7+ for the pdr1_INS cell represents the $AR(1)$ model extended with the INS factor at lag 7 and the entry 16+,17+,18- in the ldr5_BAR cell represents the $AR(1)$ model extended with the BAR factor at lags 16 to 18. A + (-) indicates that the relevant coefficient is found to be positive (negative)

	Best model			2nd Best model			3rd Best model		
PROBITS	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**
pdr1	7+	5+,6-	17-	6-	17-	16-	16+	1+	17-,18+
pdr2	13-	12+	18-	2+	14+	17-	10-	13+	16-
pdr3	13-	14+	14+	13-,14-		13+	12-,13-	12+	14+,15-
pdr4	13-	16+	5-,6+	2+	11-,12+	18-	12-,13-	14+	17+,18-
pdr5	8+	3+	18-	11+	3+,4-	17+,18-	8+,9-	2-,3+	3+
pdr6	14+	1+	1+	8+	2+	2+	13-	3+,4-	2+,3-
pdr7	5+	1+	18-	13-	3+,4-	15-	8+	2+	16-
pdr8	5+	3+	12-	8+	1+	14-	12-	2+	18-
pdr9	6-	14+	7+	2+	12+	10+	5+	9+	13+
	Best model			**2nd Best model**			**3rd Best model**		
LOGITS	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**
ldr1	7+	17-	16-	6-	17-	16-	16+	1+	17-,18+
ldr2	13-	14+	17-	2+	14+	17-	10-	13+	16-
ldr3	13-		13+	13-,14-	14+	13+	12-,13-	12+	14+,15-
ldr4	13-	11-,12+	18-	2+	11-,12+	18-	12-,13-	17+	
ldr5	8+	3+,4-	17+,18-	11+	2-,3+	17+,18-	8+,9-	3+,4-	16+,17+,18-
ldr6	14+	2+	2+		2+		8+	3+	2+
ldr7	5+	3+,4-	15-	13-	2+	15-	8+	3+,4-	16-
ldr8	5+	1+	14-		1+		8+	2+	14-
ldr9	6-	12+	10+		14+	13+	5+	12+	10+

Table 3.4.2 P-values of Jarque-Bera normality tests on residuals of the SBC best extended $AR(1)$ models listed in Table 3.4.1

PROBITS	INS	INF	BAR
pdr1	0.5068	0.5145	0.7578
pdr2	0.9776	0.5060	0.6764
pdr3	0.0177	0.1502	0.0020
pdr4	0.2736	0.5171	0.6222
pdr5	0.0255	0.0253	0.0091
pdr6	<0.0001	<0.0001	<0.0001
pdr7	0.0092	<0.0001	<0.0001
pdr8	0.0001	<0.0001	<0.0001
pdr9	<0.0001	0.0234	0.0003
LOGITS	**INS**	**INF**	**BAR**
ldr1	0.4666	0.5388	0.7545
ldr2	0.9552	0.7990	0.9015
ldr3	0.1297	0.2111	0.0235
ldr4	0.1926	0.5168	0.7421
ldr5	0.0088	0.0181	0.0034
ldr6	<0.0001	<0.0001	<0.0001
ldr7	0.0609	<0.0001	<0.0001
ldr8	0.0010	<0.0001	<0.0001
ldr9	0.0021	0.0482	0.0055

Table 3.4.3 TSFS results with AIC criterion: three best models consisting of an $AR(1)$ model extended with a lag interval for each factor. For example, the entry [6,7] for the pdr1_INS cell refers to the $AR(1)$ model extended with the INS factor at lags 6 and 7 and the entry [15,18] in the ldr5_BAR cell refers to the $AR(1)$ model extended with the BAR factor at lags 15 to 18.

	Best model			2nd Best model			3rd Best model		
PROBITS	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**
pdr1	[6,7]	[5,6]	17	7	[5,7]	[17,18]	6	[4,6]	16
pdr2	13	12	18	[13,14]	[14]	17	[12,13]	13	16
pdr3	13	[10,11,12]	14	[13,14]	[11,12]	[14,15]	[12,13]	14	[13,14]
pdr4	[12,13]	[11,12]	[5,6]	[10,13]	[10,12]	[17,18]	[7,13]	[14,16]	[5,7]
pdr5	8	3	[15,18]	[8,9]	[3,4]	[16,18]	[7,8]	[2,3]	[17,18]
pdr6	[13,14]	[4,14]	[2,3]	[13,15]	[3,7]	1	14	[3,12]	[1,2]
pdr7	5	[3,4]	18	13	1	15	[5,6]	[3,5]	16
pdr8	5	[2,7]	12	[4-,5]	[1,7]	14	8	[2,8]	18
pdr9	6	[14,15]	7	[5,6]	[1,3]	[7,8]	2	[1,2]	10
LOGITS	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**	**INS**	**INF**	**BAR**
ldr1	[6,7]	[5,6]	17	7	[5,7]	[10,18]	6	[4,6]	[17,18]
ldr2	13	12	18	[13,14]	14	17	[12,13]	13	16
ldr3	13		14	[13,14]	14	[14,15]	[13,17]	[11,12]	[13,14]
ldr4	[12,13]	[11,12]	[5,6]	[10,13]	[10,12]	[17,18]	[7,13]	16	[5,7]
ldr5	8	3	[15,18]	[8,9]	[2,3]	[17,18]	[7,8]	[3,4]	[16,18]
ldr6	[13,14]	[4,14]	2,3	[13,15]	[3,7]	1	14	[3,12]	[1,2]
ldr7	5	[3,4]	18	13	1	15	[5,6]	[3,7]	[13,14]
ldr8	5	[2,7]	12	8	[1,7]	14	[4,5]	[2,8]	18
ldr9	6	[14,15]	7	[5,6]	[1,3]	7,8	5	[1,2]	13

Table 3.4.4 P-values of Jarque-Bera normality tests on residuals of the AIC best extended $AR(1)$ models listed in Table 3.4.3

PROBITS	INS	INF	BAR
pdr1	0.6543	0.5745	0.7578
pdr2	0.9776	0.5060	0.6764
pdr3	0.0177	0.5996	0.0020
pdr4	0.0443	0.9810	0.6222
pdr5	0.0255	0.0253	0.0010
pdr6	<0.0001	0.0509	<0.0001
pdr7	0.0092	<0.0001	<0.0001
pdr8	0.0001	<0.0001	<0.0001
pdr9	<0.0001	0.0058	0.0003
LOGITS	**INS**	**INF**	**BAR**
ldr1	0.6452	0.5388	0.7545
ldr2	0.9552	0.7990	0.9015
ldr3	0.1297	0.0235	0.0160
ldr4	0.8494	0.0160	0.7421
ldr5	0.0088	0.0181	0.0002
ldr6	<0.0001	0.0822	<0.0001
ldr7	0.0609	0.0023	<0.0001
ldr8	0.0010	<0.0001	<0.0001
ldr9	0.0021	0.0465	0.0055

3.5 Multivariate AR models for the transformed default rates

All the models in Sections 3.3 and 3.4 were formulated and fitted in a univariate manner. The $AR(1)$ model considered allowed only lags of default rates in the same risk class to affect the current default rate and fitting was done separately for each class. It is conceivable that better modelling could be done by including lagged default rates of all classes in the AR model for any particular class. This would mean that we are considering so-called vector auto-regressive (VAR) models. The $VAR(1)$ model has the form

$$Y_{tk} = \alpha_k + \sum_{m=1}^{K} \beta_{km}(Y_{t-1,m} - \alpha_m) + e_{tk} \qquad (3.5.1)$$

where α_k is an intercept as before while β_{km} represents the coefficient expressing the effect of the class m lag 1 default rate on the current class k default rate. The e_{tk}'s are the error components as before. For the home loans data with $K = 9$ there would be 81 β_{km}'s if we allow each risk class to impact on all others. This seems practically excessive and unnecessary since it is unlikely that risk classes that are far from each other would have mutual impacts. Perhaps it is sufficient to allow only adjacent risk class effects. The form of the correlation matrix in Table 3.3.4 also suggests that this could be reasonable. Therefore we shall fit the model (3.5.1) under the restriction that $\beta_{km} = 0$ if $m < k-1$ or $m > k+1$.

The resulting restricted $VAR(1)$ model can be written in the form

$$Y_{tk} = \alpha_k + \sum_{m=k-1}^{k+1} \beta_{km}(Y_{t-1,m} - \alpha_m) + e_{tk} \ . \qquad (3.5.2)$$

This model involves 9 intercept level parameters and $8+9+8=25$ regression coefficients. PROC MODEL of SAS (SAS Institute Inc., 2004:999) handles a simultaneous or multivariate fit of this model. Tables 3.5.1 to 3.5.3 show the output obtained.

Table 3.5.1 Parameter estimates with the standard errors in brackets for the restricted $VAR(1)$ model fitted to the transformed data

	Parameters			
PROBITS	**Intercept α_k**	**AR-coeff $\beta_{k,k-1}$**	**AR-coeff $\beta_{k,k}$**	**AR-coeff $\beta_{k,k+1}$**
pdr1	-0.7829 (0.0481)		0.4881 (0.2326)	-0.0016 (0.2615)
pdr2	-1.7035 (0.0552)	-0.0798 (0.1245)	0.5829 (0.2407)	-0.0841 (0.1178)
pdr3	-2.0313 (0.1558)	-0.3923 (0.2422)	0.7928 (0.1987)	-0.1077 (0.1706)
pdr4	-2.7657 (0.0853)	0.1238 (0.1940)	0.3079 (0.1912)	0.1100 (0.1602)
pdr5	-3.0709 (0.1265)	0.0694 (0.1758)	0.8302 (0.1412)	-0.1929 (0.2117)
pdr6	-3.3418 (0.0525)	0.1532 (0.1392)	0.2699 (0.3205)	0.0981 (0.3115)
pdr7	-3.4054 (0.0396)	0.3036 (0.1989)	0.0557 (0.2534)	0.0331 (0.1635)
pdr8	-3.4638 (0.0482)	0.3732 (0.2466)	0.0867 (0.2091)	0.1556 (0.1762)
pdr9	-3.6140 (0.0846)	-0.3242 (0.2345)	0.8350 (0.2420)	
LOGITS	**Intercept α_k**	**AR- coeff $\beta_{k,k-1}$**	**AR-coeff $\beta_{k,k}$**	**AR-coeff $\beta_{k,k+1}$**
ldr1	-1.2871 (0.0848)		0.4939 (0.2290)	-0.0030 (0.1951)
ldr2	-3.0810 (0.1195)	-0.0849 (0.1550)	0.5837 (0.2254)	-0.0735 (0.1041)
ldr3	-3.8469 (0.4061)	-0.4465 (0.2730)	0.7918 (0.1923)	-0.0693 (0.1366)
ldr4	-5.8715 (0.2658)	0.1537 (0.2397)	0.3045 (0.1926)	0.1017 (0.1496)
ldr5	-6.8614 (0.4369)	0.0745 (0.1984)	0.8267 (0.1456)	-0.1766 (0.2053)
ldr6	-7.7927 (0.1869)	0.1523 (0.1451)	0.2938 (0.3020)	0.0839 (0.3051)
ldr7	-8.0243 (0.1441)	0.3269 (0.1935)	0.0371 (0.2645)	0.0287 (0.1571)
ldr8	-8.2411 (0.1848)	0.3773 (0.2629)	0.0837 (0.2026)	0.1585 (0.1660)
ldr9	-8.8104 (0.3456)	-0.3239 (0.2344)	0.8398 (0.2279)	

Table 3.5.2 Restricted $VAR(1)$ model error standard deviations and correlations

	Stdev	Correlations								
PROBITS		**pdr1**	**pdr2**	**pdr3**	**pdr4**	**pdr5**	**pdr6**	**pdr7**	**pdr8**	**pdr9**
pdr1	0.1302	1.0000	0.6508	0.2748	0.2108	0.0600	0.0627	0.0217	-0.0300	0.2789
pdr2	0.1160	0.6508	1.0000	0.5853	0.4229	0.1515	0.0935	0.1417	-0.0026	0.2558
pdr3	0.1800	0.2748	0.5853	1.0000	0.6109	0.3544	0.4033	0.4447	0.3718	0.3018
pdr4	0.1591	0.2108	0.4229	0.6109	1.0000	0.5355	0.4885	0.4515	0.3955	0.3315
pdr5	0.1334	0.0600	0.1515	0.3544	0.5355	1.0000	0.5357	0.4774	0.4826	0.4551
pdr6	0.1306	0.0627	0.0935	0.4033	0.4885	0.5357	1.0000	0.7492	0.6607	0.5300
pdr7	0.1426	0.0217	0.1417	0.4447	0.4515	0.4774	0.7492	1.0000	0.7806	0.5880
pdr8	0.1612	-0.0300	-0.0026	0.3718	0.3955	0.4826	0.6607	0.7806	1.0000	0.5980
pdr9	0.1419	0.2789	0.2558	0.3018	0.3315	0.4551	0.5300	0.5880	0.5980	1.0000
LOGITS		**ldr1**	**ldr2**	**ldr3**	**ldr4**	**ldr5**	**ldr6**	**ldr7**	**ldr8**	**ldr9**
ldr1	0.2260	1.0000	0.6456	0.2784	0.1973	0.0518	0.0629	0.0178	-0.0280	0.2882
ldr2	0.2524	0.6456	1.0000	0.5859	0.4006	0.1369	0.0746	0.1300	-0.0104	0.2636
ldr3	0.4348	0.2784	0.5859	1.0000	0.5885	0.3264	0.3793	0.4181	0.3437	0.2931
ldr4	0.4873	0.1973	0.4006	0.5885	1.0000	0.5266	0.4707	0.4426	0.3789	0.3181
ldr5	0.4532	0.0518	0.1369	0.3264	0.5266	1.0000	0.5273	0.4652	0.4655	0.4438
ldr6	0.4623	0.0629	0.0746	0.3793	0.4707	0.5273	1.0000	0.7422	0.6496	0.5185
ldr7	0.5121	0.0178	0.1300	0.4181	0.4426	0.4652	0.7422	1.0000	0.7758	0.5770
ldr8	0.5918	-0.0280	-0.0104	0.3437	0.3789	0.4655	0.6496	0.7758	1.0000	0.5829
ldr9	0.5454	0.2882	0.2636	0.2931	0.3181	0.4438	0.5185	0.5770	0.5829	1.0000

Comparing the intercept level parameter estimates of the restricted $VAR(1)$ model in Table 3.5.1 with those of the simple $AR(1)$ model given in Table 3.3.3 we note that they are remarkably similar and they also show the strictly decreasing trend with increasing risk class number that is to be expected in view of the higher creditworthiness corresponding to the higher class numbers.

Comparing the β_{kk}'s in Table 3.5.1 with the β_k's in Table 3.3.3, they seem to correspond in the sense that they move in the same direction as we proceed from class 1 to class 9, but there are notable differences in sizes such as for class 9. We note that the $\beta_{k,k+1}$ estimates tend to have smaller absolute values than the $\beta_{k,k-1}$'s, which suggests that there may be a sort of feed-forward effect but not a feed-backward effect in the interaction pattern of the inter-class default rate dynamics. These interpretations are made difficult by the fairly large standard errors and in fact none of the $\beta_{k,k+1}$ and $\beta_{k,k-1}$ estimates are significantly different from 0.

Comparing the error standard deviations and correlations of the restricted $VAR(1)$ model in Table 3.5.2 with those of the simple $AR(1)$ model given in Table 3.3.4 we note that they are also remarkably similar. This suggests the more elaborate $VAR(1)$ model did not in fact explain much more of the underlying variability in the data than was already done by the simple $AR(1)$. Table 3.5.3 shows the P-values of the normality tests of the $VAR(1)$ model.

Compared to Table 3.3.5 it is clear that classes 3, 6, 7 and 8 are still problematic if the residual of month 45 is included in the tests, while only classes 3 and 6 (and also class 5 for the logits) remain unacceptable if the residual of month 45 is excluded. We also tested homoscedasticity and independence of the residuals of this model and found these acceptable as before. All in all, while the multivariate restricted $VAR(1)$ model did fit somewhat better than the univariate $AR(1)$ model there is still some room for improvement.

Table 3.5.3 P-values of Jarque-Bera normality tests on residuals of the restricted $VAR(1)$ model

PROBITS	All	Without month 45 residual
pdr1	0.7720	0.7347
pdr2	0.4898	0.4534
pdr3	0.0001	<0.0001
pdr4	0.2222	0.1359
pdr5	0.0974	0.1836
pdr6	<0.0001	0.0117
pdr7	<0.0001	0.7803
pdr8	<0.0001	0.1854
pdr9	0.8696	0.7436
LOGITS	**All**	**Without month 45 residual**
ldr1	0.7198	0.6802
ldr2	0.7971	0.7656
ldr3	0.0010	0.0003
ldr4	0.1848	0.0851
ldr5	0.0532	0.0482
ldr6	<0.0001	0.0365
ldr7	<0.0001	0.7170
ldr8	<0.0001	0.1486
ldr9	0.9958	0.6792

3.6 Summary

We fitted three variations of AR models to the home loans default data, namely the simple $AR(1)$ model, the economic factor extended $AR(1)$ model and a restricted multivariate $VAR(1)$ model. Our overall impression is that none of them are completely satisfactory. It is especially disappointing that the effort to include economic factors did not succeed since it is important from a market risk point of view to try to model the impact of changing economic variables on credit risk.

In this regard, McNeil and Wendin (2007:132) states: "Unfortunately, observed variables as proxies for systematic risk are seldom completely satisfactory. The first important issue is the identification of appropriate proxies. Moreover, there may also be a lag between the cycles of a proxy variable and that of the default activity, and this lag may vary stochastically over time. Mastering the lags is of critical importance for regulatory purposes, so that banks do not, for instance, lower capital levels in an apparent upswing of the economy."

For the case of the home loans data, we agree with the sentiments expressed in this statement, since they reflect quite well our experience as reported above. There may be some unobserved factors that also play a role in default rates but for which we cannot find suitable observable proxies and this possibility should be taken into consideration. In their paper McNeil and Wendin (2007:132) pursue credit modelling by incorporating such unobserved factors into their model precisely because of the difficulties experienced with observed variables. In Chapter 4 below we apply this approach in our context.

Chapter 4
Auto-regressive models with unobserved components

4.1 Introduction

The auto-regressive (AR) models of Chapter 3 allowed the inclusion of observable components such as lagged auto-regressive default rates or economic variables. It may happen that there are unobserved or latent components that also influence default rates and in this chapter we consider extensions of the AR models to cater for this possibility.

In Section 4.2 we formulate an $AR(1)$ model with unobserved components. In an economic time series context such as for the default rates that we are considering here, one can not realistically assume that the unobserved components are independent over time. Some form of serial dependence is necessary since economic factors generally have time dependent consequences. Allowing for time dependence in the model assumptions complicates fitting of the model via maximum likelihood methods severely and it does not appear possible to use standard software packages for this purpose.

In Section 4.3 we show how Kalman filter methods can be used to calculate the relevant likelihood function. Numerical optimisation methods can be applied to this function to calculate the maximum likelihood estimates and to do statistical inference. The process is illustrated by means of two simulation test cases which also confirm that our programming is correct. The methodology is then applied to the home loans default rate data. In Section 4.4 we show how a combination of the EM-algorithm and Kalman filtering can be done to approach likelihood estimation from a completely different angle. We apply the models to forecasting in Section 4.5 and again illustrate the process with a forward validation exercise on the home loans data.

Section 4.6 closes with a summary of the contents and contributions of this chapter and a discussion of further research issues. In particular, one of the assumptions used in the methodology is that the errors are normally distributed. When this assumption is tested, normality is rejected for some of the default rate risk classes. This suggests that the normality assumption need to be generalised to make the process fully compatible at least with the home loans data.

4.2 AR models with unobserved components

An $AR(1)$ model with inclusion of unobserved components takes the form

$$Y_{tk} = \alpha_k + \beta_k (Y_{t-1,k} - \alpha_k) + \boldsymbol{\delta}_k' \boldsymbol{U}_t + e_{tk} \ . \tag{4.2.1}$$

where α_k and β_k are as before, the $M \times 1$ vector $\boldsymbol{U}_t$ is the value at time t of M unobserved or latent random components (or factors) that may affect obligor default and the $M \times 1$ vector $\boldsymbol{\delta}_k$ is the corresponding loadings (or coefficients) of these components on the k-th risk class. Note that $\boldsymbol{\delta}_k'$ is the transpose of $\boldsymbol{\delta}_k$. The following distributional assumptions on the $\boldsymbol{U}_t$'s need to be made. We only consider the case where the $\boldsymbol{U}_t$'s are stationary distributed so that $E\boldsymbol{U}_t$ and $Cov(\boldsymbol{U}_t)$ do not depend on time t. Then it involves no loss of generality to take $E\boldsymbol{U}_t = \boldsymbol{0}$ since otherwise we could subtract $E\boldsymbol{U}_t$ from $\boldsymbol{U}_t$ and add $\boldsymbol{\delta}_k' E\boldsymbol{U}_t$ to α_k to get an adjusted form of the model for which $E\boldsymbol{U}_t = \boldsymbol{0}$. Similarly, by adjusting the m-th component δ_{km} of $\boldsymbol{\delta}_k$ we can assume that the m-th component U_{tm} of $\boldsymbol{U}_t$ has variance 1. Also by making a suitable linear transformation of $\boldsymbol{\delta}_k$ if need be, we lose no generality in assuming that the components of $\boldsymbol{U}_t$ are uncorrelated. Hence we are assuming that $Cov(\boldsymbol{U}_t) = \boldsymbol{I}$ with $\boldsymbol{I}$ the $M \times M$ identity matrix. In addition to these moment assumptions we also assume that $\boldsymbol{U}_t$ is multivariate normally distributed, i.e. that $\boldsymbol{U}_t \sim N_M(\boldsymbol{0}, \boldsymbol{I})$. It would not be realistic from an economic point of view to assume that the $\boldsymbol{U}_t$'s are independent over time since economic factors generally have time dependent dynamics. In a related context McNeil and Wendin (2007) assume lag 1 auto-regressive time dependence for the unobserved components in their models. We shall follow them and assume here that each of the components of $\boldsymbol{U}_t$ has this structure, i.e. we assume that

$$U_{tm} = \rho_m U_{t-1,m} + \sqrt{1-\rho_m^2}\,\eta_{tm} \quad \text{for } m = 1,2,...,M \tag{4.2.2}$$

where ρ_m is the lag 1 AR coefficient of the U_{tm} series and η_{tm} are independent and identically $N(0,1)$ distributed, both with respect to t and m. It is easily verified that this assumption is consistent with the former assumption that $\boldsymbol{U}_t \sim N_M(\boldsymbol{0}, \boldsymbol{I})$.

Next we must formulate the distributional assumptions on the error components e_{tk} in (4.2.1). This model is a dynamic factor analysis model (see e.g. Harvey, 2003:449) in the sense that it is similar to classical factor analysis models in multivariate statistics, but the factors encompassed in $\boldsymbol{U}_t$ have their own dynamics rather than being independent over time. In the factor analysis literature the components of $\boldsymbol{U}_t$ are referred to as the "common components" in the sense that they affect all risk classes while the e_{tk}'s are referred to as the "idiosyncratic components" in the sense that they are particular to each risk class. In keeping with this view it is usual to assume that the e_{tk}'s are uncorrelated over k, i.e. that

$Cov(e_{tk}, e_{tl}) = 0$ for $k \neq l$. We continue with the assumption that $e_{tk} \sim N(0, \sigma_k^2)$ as for the models of Chapter 3. Furthermore we assume that the e_{tk}'s are independent of the $\boldsymbol{U}_s$'s over all choices of t, k and s.

Equations (4.2.1) and (4.2.2) can be written in vector-matrix notation. Let $\boldsymbol{Y}_t$, $\boldsymbol{\alpha}$ and $\boldsymbol{e}_t$ denote the $K \times 1$ vectors whose k-th elements are Y_{tk}, α_k and e_{tk} respectively and let $\boldsymbol{\beta}$ denote the $K \times K$ diagonal matrix with k-th diagonal element β_k. Also let $\boldsymbol{\Delta}$ be the $K \times M$ matrix whose k-th row is $\boldsymbol{\delta}_k'$ and let $\boldsymbol{\rho}$ and $\boldsymbol{D}$ be $M \times M$ diagonal matrices whose m-th diagonal elements are ρ_m and $\sqrt{1-\rho_m^2}$ respectively. Then (4.2.1) and (4.2.2) are

$$\boldsymbol{Y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}(\boldsymbol{Y}_{t-1} - \boldsymbol{\alpha}) + \boldsymbol{\Delta}\boldsymbol{U}_t + \boldsymbol{e}_t \quad \text{and} \quad \boldsymbol{U}_t = \boldsymbol{\rho}\boldsymbol{U}_{t-1} + \boldsymbol{D}\boldsymbol{\eta}_t \tag{4.2.3}$$

From this it is clear that we have a so-called state space model with the first of the two equations forming the observational equation and the second the state transition equation (see e.g. Harvey (2003) or Wu *et al.* (1996)). The unobserved components form the states of the system. Statistical inference regarding the model and in particular maximum likelihood estimation can now be done by means of the methodology of Kalman filtering. We explain this in more detail and report the results in the next section. For ease of reference we shall refer to this model as the $AR(1)-U(M)$ model henceforth.

4.3 Maximum likelihood inference via Kalman filtering

4.3.1 Calculation of the log-likelihood function

It is convenient to collect all the parameters in the $AR(1)-U(M)$ model (4.2.3) in the set $\boldsymbol{\Theta} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\Delta}, \boldsymbol{\rho}\}$ where $\boldsymbol{\Sigma}$ is the $K \times K$ diagonal matrix with k-th diagonal σ_k^2 and the other symbols are as defined above. Note that there are in total $K + K + K + KM + M = (3+M)K + M$ parameters in the model. With actually observed data $y_1, y_2, \ldots, y_T$ the log-likelihood function is given by the logarithm of the joint density $f(y_1, y_2, \ldots, y_T)$ of $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T$, i.e. by

$$l(\boldsymbol{\Theta}) = \log f(y_1, y_2, \ldots, y_T) \tag{4.3.1}$$

By the conditional decomposition of the joint density this may be written in the form

$$l(\boldsymbol{\Theta}) = \sum_{t=1}^{T} \log f(y_t \,|\, y_1, y_2, \ldots, y_{t-1}). \tag{4.3.2}$$

where $f(y_t|y_1, y_2, ..., y_{t-1})$ is the conditional density of Y_t given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_{t-1} = y_{t-1}$ (and we take $f(y_t|y_1, y_2, \ldots, y_{t-1}) = f(y_1)$ for $t = 1$). Although not shown explicitly in the notation these densities depend on the parameters $\boldsymbol{\Theta}$. Because of the normality assumptions in the model, $f(y_t|y_1, y_2, \ldots, y_{t-1})$ is also a multivariate normal density and therefore fully characterised by its mean vector and covariance matrix, i.e. if

$$\begin{aligned} \boldsymbol{y}_{t|t-1} &= E[\boldsymbol{Y}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \ldots, \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}] \quad \text{and} \\ \boldsymbol{F}_{t|t-1} &= E[(\boldsymbol{Y}_t - \boldsymbol{y}_{t|t-1})(\boldsymbol{Y}_t - \boldsymbol{y}_{t|t-1})' | \boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \ldots, \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}] \end{aligned} \tag{4.3.3}$$

then

$$\log f(y_t|y_1, y_2, \ldots, y_{t-1}) = -\tfrac{1}{2}(y_t - y_{t|t-1})' F_{t|t-1}^{-1}(y_t - y_{t|t-1}) - \tfrac{1}{2}\log\det(F_{t|t-1}) - \tfrac{K}{2}\log(2\pi) \tag{4.3.4}$$

This means that the log-likelihood function can be computed once we have a method for computing the $y_{t|t-1}$'s and the $F_{t|t-1}$'s. Kalman filtering provides an iterative calculation of these sequences. For this purpose introduce also

$$\begin{aligned} \boldsymbol{u}_{t|t-1} &= E[\boldsymbol{U}_t | \boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \ldots, \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}] \quad \text{and} \\ \boldsymbol{P}_{t|t-1} &= E[(\boldsymbol{U}_t - \boldsymbol{u}_{t|t-1})(\boldsymbol{U}_t - \boldsymbol{u}_{t|t-1})' | \boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \ldots, \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}] \end{aligned} \tag{4.3.5}$$

Then by (4.2.3) we have $y_{t|t-1} = \alpha + \beta(y_{t-1} - \alpha) + \Delta u_{t|t-1}$ and $F_{t|t-1} = \Delta P_{t|t-1}\Delta' + \Sigma$ which enable us to calculate the $y_{t|t-1}$'s and the $F_{t|t-1}$'s easily from the $u_{t|t-1}$'s and the $P_{t|t-1}$'s and we need to focus only the latter. Here $u_{t|t-1}$ is the best mean square error predictor of U_t on the basis of $y_1, y_2, \ldots, y_{t-1}$ and $P_{t|t-1}$ is the covariance matrix of the error of this prediction; hence these quantities are of inherent interest in themselves. Iterative equations to calculate them can be obtained from many sources in the Kalman filtering literature, e.g. Harvey (2003). We shall follow the paper of Wu *et al.* (1996). Adapting the algorithm as given in their Appendix (A.3) to (A.7) to our notation, the iterative equations are given by

$$\begin{aligned} \boldsymbol{H}_t &= \boldsymbol{F}_{t|t-1}^{-1} = (\Delta \boldsymbol{P}_{t|t-1}\Delta' + \Sigma)^{-1} \\ \boldsymbol{K}_t &= \boldsymbol{P}_{t|t-1}\Delta' \boldsymbol{H}_t \\ \boldsymbol{L}_t &= \rho - \rho \boldsymbol{K}_t \Delta \\ \boldsymbol{u}_{t+1|t} &= \rho \boldsymbol{K}_t (y_t - \alpha - \beta(y_{t-1} - \alpha)) + \boldsymbol{L}_t \boldsymbol{u}_{t|t-1} \\ \boldsymbol{P}_{t+1|t} &= \boldsymbol{L}_t \boldsymbol{P}_{t|t-1}\rho' + \boldsymbol{D}\boldsymbol{D}' \end{aligned} \tag{4.3.6}$$

for $t = 1,2,\ldots,T-1$. To initialise the iterations we need $\boldsymbol{u}_{1|0}$, $\boldsymbol{P}_{1|0}$ and $\boldsymbol{y}_0$. Since no conditioning is involved here we take these quantities from the stationary distributions of the $\boldsymbol{U}_t$'s and the $\boldsymbol{Y}_t$'s. Thus $\boldsymbol{u}_{1|0} = E[\boldsymbol{U}_1] = \boldsymbol{0}$ and $\boldsymbol{P}_{1|0} = E\boldsymbol{U}_1\boldsymbol{U}_1' = \boldsymbol{I}$ and these constitute the adaptations of equations (A.1) and (A.2) of Wu *et al.* (1996) to our application. Under the assumption of stationarity $E[\boldsymbol{Y}_0] = \boldsymbol{\alpha}$ and therefore we take $\boldsymbol{y}_0 = \boldsymbol{\alpha}$.

4.3.2 Calculation of maximum likelihood estimates

Once we have programmed the iterative equations we are in a position to calculate the log-likelihood function $l(\boldsymbol{\Theta})$ at any given choice of the parameter set $\boldsymbol{\Theta}$. In principle we also should be able to calculate the maximum likelihood estimates, i.e. that choice of $\hat{\boldsymbol{\Theta}}$ that maximises $l(\boldsymbol{\Theta})$ over all allowed choices of $\boldsymbol{\Theta}$. The relevant parameter constraints are:

$$\begin{aligned} &-1 < \beta_k < 1 \quad \text{for} \quad k = 1,\ldots,K \\ &\sigma_k^2 > 0 \quad \text{for} \quad k = 1,\ldots,K \\ &-1 < \rho_m < 1 \quad \text{for} \quad m = 1,\ldots,M \end{aligned} \tag{4.3.7}$$

Since $l(\boldsymbol{\Theta})$ can only be calculated numerically there is no way to find $\hat{\boldsymbol{\Theta}}$ analytically, i.e. the calculation of $\hat{\boldsymbol{\Theta}}$ must also be done numerically. Numerical optimisation methods also operate iteratively, i.e. we start with an initial estimate $\hat{\boldsymbol{\Theta}}_0$ and then iteratively improve this to $\hat{\boldsymbol{\Theta}}_1, \hat{\boldsymbol{\Theta}}_2, \ldots$. until a suitable convergence criterion is satisfied and the end result is taken as $\hat{\boldsymbol{\Theta}}$. Two things are required for this purpose, namely how to choose the initial estimate $\hat{\boldsymbol{\Theta}}_0$ and what iterative improvement process to use. For the latter we use the Dual Quasi-Newton Method as implemented in the SAS module NLPQN of PROC IML. The choice of $\hat{\boldsymbol{\Theta}}_0$ is difficult in general. A possible strategy is to approximate the $AR(1)-U(M)$ model that we really want to fit by simpler models whose parameters can be estimated more easily and then use these estimates to obtain $\hat{\boldsymbol{\Theta}}_0$. We have found empirically that the following approach works reasonably well for cases with $M \le 2$ which is as high as we shall consider here, due to the rapid increase in number of parameters with increasing M. First ignore the unobserved component term and fit the $AR(1)$ model $\boldsymbol{Y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}(\boldsymbol{Y}_{t-1} - \boldsymbol{\alpha}) + \boldsymbol{e}_t$ using e.g. PROC AUTOREG of SAS. This gives initial estimates $\hat{\alpha}_0$ and $\hat{\boldsymbol{\beta}}_0$ of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Then calculate the residuals $\boldsymbol{r}_t$ of the $\boldsymbol{y}_t$'s following this fit. Intuitively, the model for the $\boldsymbol{r}_t$'s should be about $\boldsymbol{r}_t = \boldsymbol{\Delta}\boldsymbol{u}_t + \boldsymbol{e}_t$ which is like a traditional factor model. Next calculate the first M principal components of the $\boldsymbol{r}_t$'s over time which are rough estimates of the $\boldsymbol{u}_t$'s. Then

the loadings of the r_t's on these principal components provide an initial estimate $\hat{\Delta}_0$ of Δ, while the mean square errors of the next level residuals remaining after fitting the principal component to the r_t's provide initial estimates of the diagonal elements of Σ. The off-diagonal elements are taken as 0 so that we obtain an initial estimate $\hat{\Sigma}_0$ of Σ in this way. We take $\hat{\rho}_0$ as the lag 1 auto-correlation coefficient of the principal component series. Assembling these individual initial estimates leads to $\hat{\Theta}_0 = \{\hat{\alpha}_0, \hat{\beta}_0, \hat{\Sigma}_0, \hat{\Delta}_0, \hat{\rho}_0\}$.

4.3.3 Testing the procedures

The procedures above were programmed in PROC IML of SAS and we next report the results of their application to testing data obtained by simulation with known parameter values. This serves to verify that the programming is correct and will give the reader a better understanding of the methodology. For the first simulation test case we took

$$
\begin{aligned}
&K = 5, M = 1 \\
&\alpha = (6 \;\; 3 \;\; 0 \;\; -3 \;\; -6)' \\
&\beta = diag\{0.4 \;\; 0.2 \;\; 0 \;\; -0.2 \;\; -0.4\} \\
&\Sigma = diag\{1 \;\; 0.5 \;\; 0.25 \;\; 0.5 \;\; 1\} \\
&\Delta = (1 \;\; 0.5 \;\; 0 \;\; -0.5 \;\; -1)' \\
&\rho = \rho_1 = 0.5
\end{aligned}
\qquad (4.3.8)
$$

We then generated 10000 realisations of the vectors Y_t and U_t according to the $AR(1)-U(1)$ model and subsequently applied the program to the first 100, first 1000 and all 10000 of these realisations, i.e. we varied T over the values 100, 1000 and 10000. Table 4.3.1 shows the results. The first column lists the parameters and the second column their true values used to generate the data. The subsequent three blocks of three columns each show the initial estimate at which the optimisation was started, the MLE's and their standard errors for the cases $T = 10000$, $T = 1000$ and $T = 100$ respectively. The standard errors were calculated according to the standard asymptotic formulae of MLE large sample theory using the module NLPFDD of PROC IML of SAS following the optimisation routine.

Consider first the case of the long time series $T = 10000$. Comparing the true parameter values in column 2 of Table 4.3.1 with the initial values in column 3, there seems to be a fair correspondence in the sense that when the true parameter values go up or down or changes sign, this mostly also happens with the initial values. Also in this case we expect the MLE's to be very close to the true parameter values due to the long series used. Comparing the MLE's in the fourth column with the true values this is seen to be the case and, moreover, the differences between the true values and the MLE's are typically smaller than two standard errors. As is to be expected from the shorter series lengths for the other two cases

in the table the resemblances of the initial values to the true values tend to become poorer and also the differences between MLE's and the true values tend to be larger while the standard errors become larger also. Nevertheless the MLE's do not become wildly wrong even for relatively short series and the results are quite encouraging, taking into account that we are dealing with models with many parameters.

Table 4.3.1 Details of maximum likelihood estimates for test case 1

		$T=10000$			$T=1000$			$T=100$		
Parm	True	Init	MLE	SE	Init	MLE	SE	Init	MLE	SE
α_1	6.0000	6.0177	6.0181	0.0327	5.9880	5.9907	0.0976	6.1957	6.2193	0.2587
α_2	3.0000	3.0050	3.0051	0.0136	2.9636	2.9654	0.0442	2.9883	2.9959	0.1512
α_3	0.0000	-0.0042	-0.0042	0.0051	-0.0075	-0.0075	0.0162	-0.0155	-0.0164	0.0607
α_4	-3.0000	-3.0001	-3.0002	0.0091	-2.9779	-2.9789	0.0285	-2.9143	-2.9200	0.0989
α_5	-6.0000	-6.0085	-6.0087	0.0140	-5.9787	-5.9805	0.0431	-5.9729	-5.9806	0.1329
β_1	0.4000	0.5990	0.4028	0.0080	0.5542	0.3630	0.0256	0.4591	0.1443	0.0926
β_2	0.2000	0.3417	0.1808	0.0091	0.3578	0.2000	0.0283	0.4037	0.1585	0.0879
β_3	0.0000	0.0073	0.0073	0.0100	0.0124	0.0124	0.0316	0.1330	0.1301	0.1000
β_4	-0.2000	-0.0395	-0.1876	0.0094	-0.0492	-0.1914	0.0289	0.0352	-0.0677	0.0914
β_5	-0.4000	-0.1845	-0.3958	0.0089	-0.2362	-0.4676	0.0279	-0.1937	-0.3451	0.0892
σ_1^2	1.0000	0.8218	1.0131	0.0210	0.8491	1.0879	0.0677	0.7326	0.7758	0.2185
σ_2^2	0.5000	0.3769	0.4959	0.0084	0.3817	0.5109	0.0271	0.3581	0.4308	0.0839
σ_3^2	0.2500	0.2519	0.2519	0.0036	0.2574	0.2570	0.0115	0.2657	0.2671	0.0379
σ_4^2	0.5000	0.3795	0.4971	0.0083	0.3303	0.4460	0.0240	0.3889	0.4970	0.0784
σ_5^2	1.0000	0.7901	0.9656	0.0206	0.8075	0.8980	0.0641	0.9243	1.1160	0.1909
δ_1	1.0000	0.7109	0.9942	0.0168	-0.7231	-0.9955	0.0529	-0.7604	-1.2015	0.1811
δ_2	0.5000	0.4147	0.5152	0.0101	-0.4151	-0.5130	0.0320	-0.4512	-0.6452	0.1063
δ_3	0.0000	0.0067	0.0060	0.0057	0.0083	0.0001	0.0181	0.0713	0.0654	0.0578
δ_4	-0.5000	-0.4031	-0.4910	0.0098	0.4054	0.5018	0.0296	0.3867	0.4633	0.0916
δ_5	-1.0000	-0.7195	-1.0020	0.0162	0.7112	1.0511	0.0516	0.6409	0.8487	0.1503
ρ_1	0.5000	0.1716	0.4784	0.0125	0.1647	0.4758	0.0389	0.1950	0.4913	0.1136

We note in passing that there is an identifiability issue in the model (4.2.3). Changing Δ to $-\Delta$ and simultaneously U_t to $-U_t$ leaves the model unchanged and $-U_t$ will satisfy all the assumptions made for U_t. Hence Δ can be identified at most up to its sign. The MLE program with the initial values found by our approach chose to converge to a value close to the true value for the case $T=10000$ but to an estimate close to $-\Delta$ for the cases $T=1000$ and $T=100$ as can be seen from the entries in the δ_1 to δ_5 rows in Table 4.3.1.

The second simulation test case used two unobserved factors and we took

$$
\begin{aligned}
&K = 5, M = 2 \\
&\boldsymbol{\alpha} = (6 \;\; 3 \;\; 0 \;\; -3 \;\; -6)' \\
&\boldsymbol{\beta} = diag\{0.4 \;\; 0.2 \;\; 0 \;\; -0.2 \;\; -0.4\} \\
&\boldsymbol{\Sigma} = diag\{1 \;\; 0.5 \;\; 0.25 \;\; 0.5 \;\; 1\} \\
&\boldsymbol{\Delta}_1 = (1 \;\; 0 \;\; 0.5 \;\; 0 \;\; -1)' \\
&\boldsymbol{\Delta}_2 = (0 \;\; 1 \;\; 0.5 \;\; -1 \;\; 0)' \\
&\boldsymbol{\rho} = diag\{0.7 \;\; 0.3\}
\end{aligned}
\tag{4.3.9}
$$

Again we generated 10000 realisations of the vectors $\boldsymbol{Y}_t$ and $\boldsymbol{U}_t$ according to the $AR(1)-U(2)$ model (4.2.3) and then applied the program to the first 100, first 1000 and all 10000 of these realisations. Table 4.3.2 shows the results.

Considering first the $T = 10000$ case, the correspondence between the true parameter values and the initial values used is again fair but also somewhat poorer than for the first test case. This is to be expected since two unobserved components are now involved and this is more complex than the one component model of the first test case. Nevertheless, the MLE's still agree very well with the true parameter values. Naturally for the shorter series in the rest of the table the estimates tend to differ more from the true values and the standard errors become larger but again the results are quite encouraging, taking into account the greater complexity of the model and the even larger number of parameters involved in this case.

The reader should also note that the true values for δ_{11} to δ_{51} given in Table 4.3.2 are actually those for $\Delta_2 = (\delta_{12} \;\; \delta_{22} \;\; \delta_{32} \;\; \delta_{42} \;\; \delta_{52})'$ specified in (4.3.9), while the true values for δ_{12} to δ_{52} are those for $\Delta_1 = (\delta_{11} \;\; \delta_{21} \;\; \delta_{31} \;\; \delta_{41} \;\; \delta_{51})'$ specified in (4.3.9). Also the true values shown for ρ_1 and ρ_2 are actually the values ρ_2 and ρ_1 respectively specified in (4.3.9). In essence we interchanged the roles of the two unobserved components in Table 4.3.2 as compared to the specification in (4.3.9). The reason for this is that the sequencing of the unobserved components is not identifiable and our fitting program chose to deliver the reverse sequence compared to the specification in (4.3.9). Again in order not to confuse the reader we interchanged the true parameter values in Table 4.3.2.

4.3.4 Estimating the unobserved components

We noted above that $\boldsymbol{u}_{t|t-1} = E[\boldsymbol{U}_t \,|\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \ldots, \boldsymbol{Y}_{t-1} = \boldsymbol{y}_{t-1}]$ is the best mean square error predictor of $\boldsymbol{U}_t$ on the basis of the observed data $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{t-1}$ prior to time t and this is automatically produced as part of the Kalman filter application. We can also compute the

Table 4.3.2 Details of maximum likelihood estimates for second test case

		$T = 10000$			$T = 1000$			$T = 100$		
Parm	True	Init	MLE	SE	Init	MLE	SE	Init	MLE	SE
α_1	0.2000	5.9973	5.9976	0.0422	5.9638	5.9647	0.1310	6.1469	6.1475	0.3264
α_2	0.1000	2.9785	2.9784	0.0190	3.0001	2.9992	0.0580	3.1257	3.1228	0.1834
α_3	0.0000	-0.0197	-0.0197	0.0146	-0.0166	-0.0161	0.0402	0.0208	0.0179	0.0870
α_4	-0.1000	-2.9870	-2.9870	0.0128	-2.9969	-2.9960	0.0383	-3.0727	-3.0687	0.1260
α_5	-0.2000	-5.9873	-5.9871	0.0181	-5.9958	-5.9967	0.0539	-6.0713	-6.0655	0.1494
β_1	0.4000	0.6880	0.4019	0.0092	0.7041	0.4141	0.0287	0.6223	0.3152	0.0997
β_2	0.2000	0.3783	0.1970	0.0085	0.3513	0.1835	0.0280	0.3358	0.2087	0.0912
β_3	0.0000	0.3375	0.0030	0.0092	0.3040	0.0222	0.0292	0.3829	0.2090	0.0941
β_4	-0.2000	0.0030	-0.1986	0.0090	-0.0153	-0.1877	0.0286	0.0361	-0.1379	0.1020
β_5	-0.4000	-0.1027	-0.3937	0.0096	-0.1452	-0.4123	0.0295	-0.2162	-0.5131	0.0888
σ_1^2	1.0000	0.6424	1.0356	0.0209	0.6141	0.9217	0.0641	0.5852	0.7739	0.1899
σ_2^2	0.5000	0.3221	0.4973	0.0138	0.3256	0.4720	0.0463	0.3439	0.5918	0.1384
σ_3^2	0.2500	0.1909	0.2399	0.0059	0.1745	0.2362	0.0187	0.1357	0.2209	0.0408
σ_4^2	0.5000	0.3277	0.5036	0.0140	0.3433	0.5960	0.0440	0.2878	0.3623	0.1491
σ_5^2	1.0000	0.6066	0.9724	0.0199	0.6104	0.9763	0.0645	0.5256	0.7177	0.1779
δ_{11}	0.0000	0.2857	0.0113	0.0205	0.2863	0.0804	0.0677	0.0202	0.2705	0.1900
δ_{21}	1.0000	0.6453	0.9918	0.0125	0.6269	0.9538	0.0403	0.6999	0.7117	0.1324
δ_{31}	0.5000	0.4637	0.5098	0.0111	0.4546	0.5113	0.0330	0.2661	0.3364	0.0702
δ_{41}	-1.0000	-0.6451	-0.9972	0.0126	-0.6155	-0.8928	0.0397	-0.6950	-0.8283	0.1360
δ_{51}	0.0000	-0.3018	-0.0044	0.0205	-0.3196	-0.0545	0.0677	0.0503	-0.1314	0.1954
δ_{12}	1.0000	0.8038	0.9864	0.0210	0.7666	1.0119	0.0652	0.7866	0.9700	0.2034
δ_{22}	0.0000	-0.3914	-0.0025	0.0244	-0.4150	-0.1960	0.0768	-0.1601	-0.4248	0.1868
δ_{32}	0.5000	0.1201	0.5035	0.0155	0.1408	0.4183	0.0473	0.2294	0.1453	0.0919
δ_{42}	0.0000	0.3986	0.0055	0.0245	0.4176	0.1254	0.0729	0.1234	0.3991	0.1971
δ_{52}	-1.0000	-0.8371	-0.9911	0.0185	-0.8195	-0.9959	0.0573	-0.8164	-0.9881	0.1680
ρ_1	0.3000	0.0688	0.2981	0.0136	0.0360	0.2723	0.0433	0.0163	0.1784	0.1570
ρ_2	0.7000	0.1325	0.6924	0.0100	0.1793	0.6602	0.0332	0.2020	0.6486	0.1035

best mean square error estimate $u_{t|T} = E[U_t | Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T]$ on the basis of all the data $y_1, y_2, \ldots, y_T$, which is known as the "smoothing" estimate of the unobserved component U_t. This can be done by means of a backward iteration through the data after doing the forward iteration (4.3.6). Adapting the algorithm as given in Wu *et al.* (1996) Appendix (A.8) to (A.11) to our notation, the iterative steps are as follows.

Initialise a $M \times 1$ vector sequence $\boldsymbol{v}_t$ and a $M \times M$ matrix sequence $\boldsymbol{V}_t$ at $t = T$ by $\boldsymbol{v}_T = \boldsymbol{0}$ and $\boldsymbol{V}_T = \boldsymbol{0}$ and then calculate successively for $t = T, T-1, \ldots, 1$

$$\begin{aligned} \boldsymbol{v}_{t-1} &= \boldsymbol{\Delta}' \boldsymbol{H}_t (y_t - \alpha - \boldsymbol{\beta} y_{t-1} - \boldsymbol{\Delta} \boldsymbol{u}_{t|t-1}) + \boldsymbol{L}_t' \boldsymbol{v}_t \\ \boldsymbol{V}_{t-1} &= \boldsymbol{\Delta}' \boldsymbol{H}_t \boldsymbol{\Delta} + \boldsymbol{L}_t' \boldsymbol{V}_t \boldsymbol{L}_t \\ \boldsymbol{u}_{t|T} &= \boldsymbol{u}_{t|t-1} + \boldsymbol{P}_{t|t-1} \boldsymbol{v}_{t-1} \end{aligned} \qquad (4.3.10)$$

Finally for $t = 0$ calculate $\boldsymbol{u}_{0|T} = \boldsymbol{\rho}' \boldsymbol{v}_0$. Here $\boldsymbol{H}_t$, $\boldsymbol{L}_t$, $\boldsymbol{u}_{t|t-1}$ and $\boldsymbol{P}_{t|t-1}$ are the same as calculated in (4.3.6) above so that these quantities need not be computed again. We programmed this calculation also and applied it to our two test cases discussed above. Since the actual values of the unobserved components are available from the simulation generation of the data in these cases we can compare the estimates with the actual values. By way of illustration, Figure 4.3.1 provides a scatter plot of the actual values against the estimated (smoothed) values of the unobserved component for the first test case with $T = 1000$ observations. Note that they are highly negatively correlated. This is due to the identifiability issue referred to above: the estimate chose to estimate the negative of the actual unobserved component. Figures 4.3.2a and 4.3.2b provide similar scatter plots again for $T = 1000$ observations for the second test case, but here we plot respectively the second estimated component against the first actual and the first estimated component against the second actual, again due to the sequencing identifiability issue noted above. With this arrangement the correlations are quite high. These high correlations between actual and estimated values are consistent with the programming being correct.

4.3.5 Fitted values and testing fit

Having estimated the parameters and the unobserved components we can also define fitted values, calculate residuals and develop model diagnostics and tests of fit. The fitted value and the residual at time t are defined by

$$y_{t|T} = \hat{\alpha} + \hat{\boldsymbol{\beta}} y_{t-1} + \hat{\boldsymbol{\Delta}} \boldsymbol{u}_{t|T} \quad \text{and} \quad e_{t|T} = y_t - y_{t|T} = y_t - (\hat{\alpha} + \hat{\boldsymbol{\beta}} y_{t-1} + \hat{\boldsymbol{\Delta}} \boldsymbol{u}_{t|T}) \ . \qquad (4.3.11)$$

These are often referred to as the smoothed fitted values and residuals since they are based on the smoothed estimates of the unobserved components. We can also define "predictive" and "filtered" fitted values and residuals based on the predictive and filtered estimates $\boldsymbol{u}_{t|t-1}$ and $\boldsymbol{u}_{t|t}$ respectively. As these are not based on all the available data we shall not discuss them further here. Figure 4.3.3 provides an illustration of the fitted values based on our first test case using the first 100 observations.

Figure 4.3.1 Comparison of actual and estimated unobserved component of the first test case

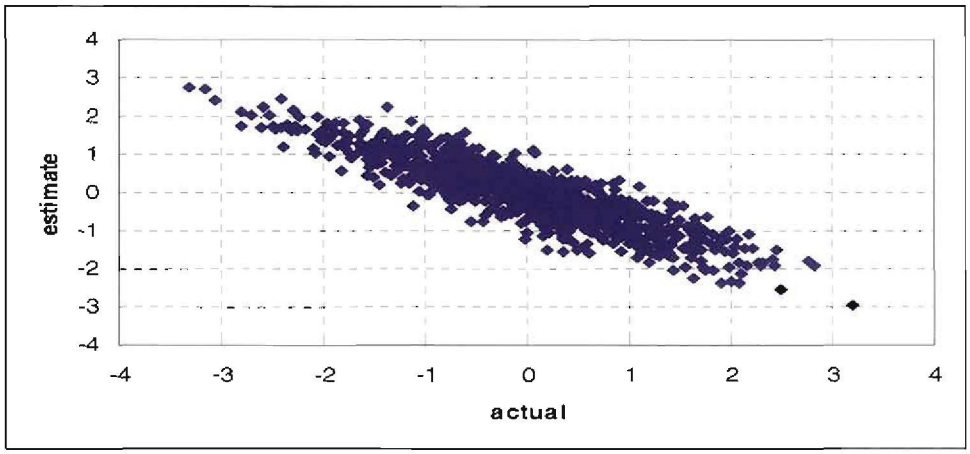


Figure 4.3.2a Comparison of actual first and estimated second unobserved component of the second test case

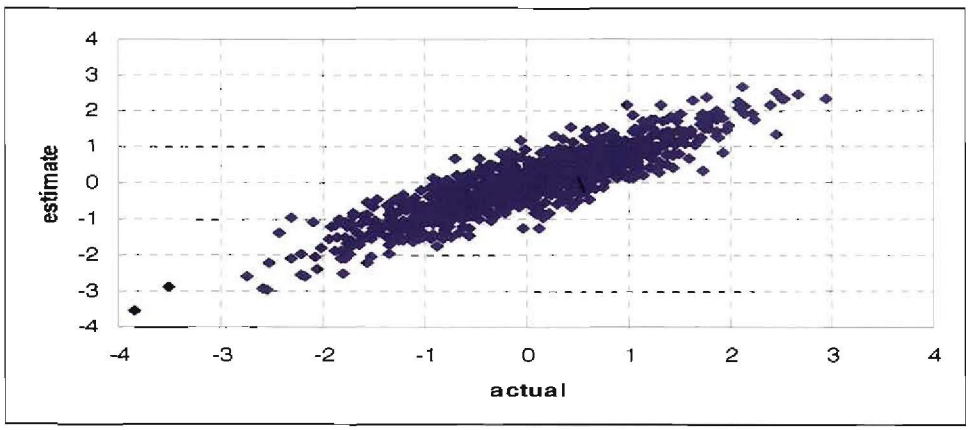


Figure 4.3.2b Comparison of the actual second and estimated first unobserved component of the second test case

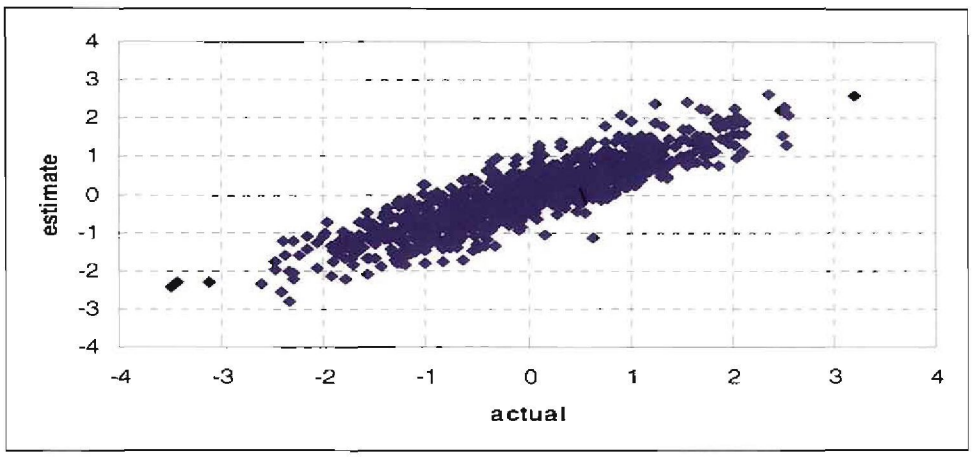
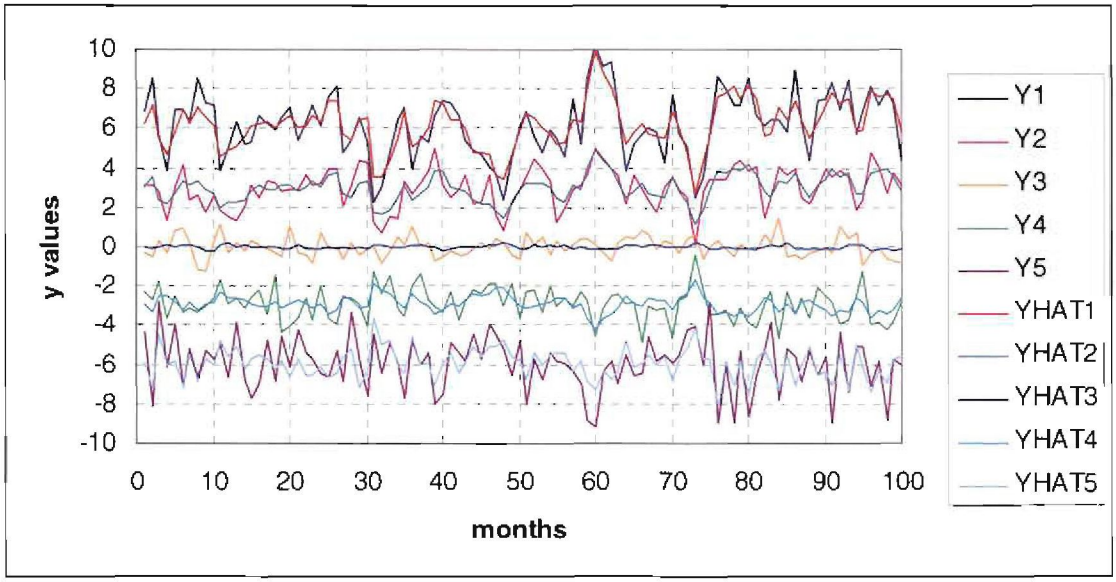


Figure 4.3.3 Comparison of the actual and fitted series for the first test case



In as much as the $e_{t|T}$'s are estimates of the e_t's, one could expect them to be approximately independently and multivariate normally distributed. However, this could be the case at best *approximately* since the estimation process may introduce both dependencies and non-normalities in the distributions of the $e_{t|T}$'s. Developing the required distribution theory on which to base fully justified tests of fit is beyond the scope of this dissertation. Here we merely report the outcome of standard univariate normality and independence tests on each of the components of the residuals when applied to our two test cases above and briefly discuss the results.

Tables 4.3.3 and 4.3.4 show the results for test cases 1 and 2 respectively. Each table shows the P-values of the Jarque-Bera normality test. They also show the P-values of the Durbin-Watson test of zero auto-correlation against the alternative of positive auto-correlation. Considering first the P-values of the normality tests, there are only two smallish entries, namely for the second test case at 10000 observations. With thirty tests being done one should expect the occasional small P-value even when the normality null hypothesis is true. Hence the results are consistent with acceptance of the normality hypothesis. Of course normality was used when the data was generated so that one expects to accept the normality hypothesis here. The point of actually doing the testing here was to illustrate the reasonableness of the process of subjecting the residuals to normality tests. Regarding the P-values for testing independence, a number of values close to 1 were obtained especially at the large sample size cases. Since the P-values for testing for zero auto-correlation against the alternative of negative auto-correlation is just one minus the values shown in the table, the P-values close to 1 in the tables suggest that there may be negative auto-

correlation in the smoothed residuals. Since independent errors were in fact used in the generation of the data, it appears that the null-distribution of the Durbin-Watson test statistic in the present context needs to be studied anew and that the P-values produced by standard packages are not reliable in this instance.

Table 4.3.3 P-values for testing normality and independence for test case one

Hypothesis	Class	$T=100$	$T=1000$	$T=10000$
Normality	1	0.4222	0.4879	0.8138
	2	0.8183	0.9367	0.5447
	3	0.4215	0.1893	0.4369
	4	0.7335	0.4220	0.8473
	5	0.7968	0.9577	0.4328
Independence	1	0.5509	0.8641	0.9985
	2	0.2111	0.5475	0.9886
	3	0.4074	0.4963	0.4997
	4	0.6288	0.6998	0.8250
	5	0.5393	0.9748	1.0000

Table 4.3.4 P-values for testing normality and independence for test case two

Hypothesis	Class	$T=100$	$T=1000$	$T=10000$
Normality	1	0.7056	0.6550	0.9286
	2	0.5165	0.6655	0.0611
	3	0.7199	0.2605	0.0584
	4	0.6444	0.1282	0.8827
	5	0.5880	0.4278	0.4194
Independence	1	0.9309	0.9667	1.0000
	2	0.6829	0.9512	0.9997
	3	0.4526	0.9582	1.0000
	4	0.8513	0.8256	0.9977
	5	0.7821	0.7432	1.0000

4.3.6 Application of an $AR(1)-U(1)$ model to the home loans transformed default rates

Here we apply an auto-regressive model with unobserved components of the form (4.3.2) to the transformed default rates and we begin with the simplest case of one unobserved component, i.e. we fit the $AR(1)-U(1)$ model and treat the probit case first. Table 4.3.5 shows the results regarding parameter estimates. The first column in this table lists the parameters. The second column shows the initial estimates at which likelihood optimisation was started. The third column shows the final MLE's obtained at convergence of the likelihood optimisation. The fourth column shows the standard errors of the MLE's obtained from the standard asymptotic (large sample) maximum likelihood methodology. Since our sample size of 56 is not really large we were concerned that these standard errors may not be reliable. Therefore we also calculated parametric bootstrap standard errors (Davison & Hinkley, 1997:15) and list these in the fifth column of Table 4.3.5. This calculation was done as follows:

At the MLE's listed in Table 4.3.5, generate a data set of length 56 months according to the $AR(1)-U(1)$ model; then calculate starting estimates and MLE's from this generated data set. Do many repetitions of these steps (we did 3000) thus building up a set of repeated MLE's each calculated on independently generated data sets proceeding from the MLE's of the fit on the probit data set. The standard deviations of these repeated MLE's constitute the parametric bootstrap standard error estimates. Comparing the fourth and fifth columns of Table 4.3.5 we see that there is good agreement between the two sets of standard errors and conclude that they are reliable. Below we shall use the entries in column 4. Column 6 shows the t-statistics corresponding to the entries in columns 3 and 4. For comparison purposes columns 7 to 9 of Table 4.3.5 also shows the MLE's, standard errors and t-statistics of the $AR(1)$ model fits to each of the nine default series of the individual risk classes. The estimates of the intercepts and AR coefficients of these $AR(1)$ models are slightly different from those shown in Table 3.3.1. The entries in Table 3.3.1 were produced by PROC AUTOREG as described before. The AR coefficients in Table 4.3.5 were calculated by maximising the $AR(1)$ likelihood functions using our own program which was specially written to ensure that exactly the same method and assumptions were used as in the $AR(1)-U(1)$ program. Column 10 of Table 4.3.5 will be used to discuss variance relations between the two models below.

Consider first the intercept level results in the rows labelled α_1 to α_9. The initial estimates for the $AR(1)-U(1)$ model are the same as the MLE's of the $AR(1)$ models which is due to our initialising strategy as discussed above. The MLE's of the $AR(1)-U(1)$ model are seen to be quite close to these initial estimates suggesting that the initialising strategy works well for the intercept parameters. The standard errors of the $AR(1)-U(1)$ and $AR(1)$ estimates are similar with the exception of α_5 and α_9 and they vary considerably between the risk classes.

Next consider the AR coefficient results in the rows labelled β_1 to β_9. Again, due to our initialising strategy the initial $AR(1)-U(1)$ estimates are the same as the MLE's of the $AR(1)$ model. However, the MLE's of the $AR(1)-U(1)$ model are generally not close to the MLE's of the $AR(1)$ models and tend to be larger especially for β_6 to β_9. The t-statistics of the $AR(1)-U(1)$ estimates show that they are all significantly different from 0 and this is generally more emphatically the case than for the $AR(1)$ estimates.

Table 4.3.5 MLE's of $AR(1)-U(1)$ and $AR(1)$ models fitted to probit transformed rates

	$AR(1)-U(1)$ model					$AR(1)$ model			Variance relation
Parm	Init	MLE	SE	BSE	t	MLE	SE	t	
α_1	-0.7829	-0.7832	0.0246	0.0251	-31.8704	-0.7829	0.0257	-30.4663	-
α_2	-1.6901	-1.6908	0.0345	0.0353	-49.0784	-1.6901	0.0343	-49.3312	-
α_3	-2.1162	-2.1502	0.0859	0.0735	-25.0221	-2.1162	0.0568	-37.2593	-
α_4	-2.7727	-2.7749	0.0337	0.0338	-82.4306	-2.7727	0.0319	-86.9998	-
α_5	-3.0334	-3.0351	0.0612	0.0654	-49.5730	-3.0334	0.0670	-45.2743	-
α_6	-3.3272	-3.3239	0.0310	0.0300	-107.0906	-3.3272	0.0270	-123.1502	-
α_7	-3.4026	-3.4054	0.0235	0.0230	-144.6484	-3.4026	0.0205	-166.0358	-
α_8	-3.4635	-3.4631	0.0324	0.0325	-106.7843	-3.4635	0.0269	-128.8793	-
α_9	-3.6418	-3.6504	0.0453	0.0454	-80.5836	-3.6418	0.0338	-107.8793	-
β_1	0.3614	0.3367	0.1247	0.1257	2.6992	0.3614	0.1248	2.8965	-
β_2	0.5605	0.5747	0.1075	0.1179	5.3481	0.5605	0.1112	5.0385	-
β_3	0.5791	0.7138	0.1164	0.1036	6.1325	0.5791	0.1129	5.1276	-
β_4	0.3838	0.4852	0.1075	0.1139	4.5142	0.3838	0.1252	3.0645	-
β_5	0.7648	0.7768	0.0774	0.0993	10.0390	0.7648	0.0933	8.1943	-
β_6	0.3658	0.5524	0.0980	0.1037	5.6377	0.3658	0.1249	2.9289	-
β_7	0.1053	0.3685	0.1058	0.1062	3.4828	0.1053	0.1337	0.7881	-
β_8	0.2490	0.4795	0.1065	0.1066	4.5007	0.2490	0.1299	1.9172	-
β_9	0.4597	0.6570	0.1024	0.1037	6.4177	0.4597	0.1193	3.8544	-
σ_1^2	0.0148	0.0150	0.0029	0.0028	5.2635	0.0155	0.0029	5.2914	0.0155
σ_2^2	0.0119	0.0125	0.0024	0.0023	5.1806	0.0136	0.0026	5.2915	0.0136
σ_3^2	0.0208	0.0221	0.0046	0.0045	4.7560	0.0323	0.0061	5.2913	0.0330
σ_4^2	0.0116	0.0141	0.0030	0.0028	4.7743	0.0222	0.0042	5.2908	0.0224
σ_5^2	0.0092	0.0105	0.0022	0.0021	4.7479	0.0174	0.0033	5.2915	0.0174
σ_6^2	0.0065	0.0068	0.0017	0.0016	3.9852	0.0168	0.0032	5.2921	0.0174
σ_7^2	0.0077	0.0073	0.0020	0.0018	3.7336	0.0189	0.0036	5.2909	0.0201
σ_8^2	0.0119	0.0112	0.0026	0.0025	4.2222	0.0231	0.0044	5.2911	0.0243
σ_9^2	0.0114	0.0106	0.0023	0.0022	4.4987	0.0194	0.0037	5.2919	0.0203
δ_1	0.0183	0.0213	0.0181	0.0177	1.1758	-	-	-	-
δ_2	0.0241	0.0335	0.0170	0.0164	1.9661	-	-	-	-
δ_3	0.0555	0.1043	0.0262	0.0247	3.9864	-	-	-	-
δ_4	0.0547	0.0911	0.0205	0.0202	4.4484	-	-	-	-
δ_5	0.0480	0.0832	0.0176	0.0177	4.7350	-	-	-	-
δ_6	0.0533	0.1031	0.0173	0.0171	5.9642	-	-	-	-
δ_7	0.0557	0.1131	0.0189	0.0187	5.9953	-	-	-	-
δ_8	0.0562	0.1147	0.0211	0.0209	5.4320	-	-	-	-
δ_9	0.0476	0.0985	0.0197	0.0188	5.0090	-	-	-	-
ρ_1	-0.2766	-0.4385	0.1408	0.1522	-3.1138	-	-	-	-

Next, consider the error variance results in the rows labelled σ_1^2 to σ_9^2. Here the initial $AR(1)-U(1)$ estimates are smaller than the MLE's of the $AR(1)$ model since an initial estimate of the unobserved component was used to reduce the error variance below that of the $AR(1)$ model. This worked well since the final MLE's of the $AR(1)-U(1)$ model are quite close to the initial estimates of $AR(1)$ model. The error variance estimates of the $AR(1)-U(1)$ model are all smaller than those of the $AR(1)$ model. This can best be understood by taking the loadings of the unobserved component also into consideration. A relationship exists between the error variances of the $AR(1)-U(1)$ and $AR(1)$ models. If we write $\tilde{e}_{tk} = \boldsymbol{\delta}_k' \boldsymbol{U}_t + e_{tk}$ in (4.2.1) then $\tilde{e}_{tk}$ would be the error term in the $AR(1)$ model. From the model assumptions we can easily derive that

$$\tilde{\sigma}_k^2 = Var(\tilde{e}_{tk}) = Var(\boldsymbol{\delta}_k' \boldsymbol{U}_t) + Var(e_{tk}) = \boldsymbol{\delta}_k' \boldsymbol{\delta}_k + \sigma_k^2 . \qquad (4.3.12)$$

With only one unobserved component this relation reduces to $\tilde{\sigma}_k^2 = \delta_k^2 + \sigma_k^2$. We can test this relation in our results as follows. The MLE of $\tilde{\sigma}_k^2$ is shown in column 6 of the σ_k^2 labelled row and the MLE's of δ_k and σ_k^2 are shown in column 3 of the δ_k and σ_k^2 labelled rows respectively. Adding the square of the δ_k estimate to the σ_k^2 estimate we get the entries in the last column of the table and according to the relation (4.3.12) these should be about the same as the corresponding entries in column 6. The agreement is quite close. It is clear that δ_k^2 represents the reduction in error variance for the k-th risk class due to the inclusion of the unobserved component in the $AR(1)$ model. This reduction will be larger where the estimate of δ_k is more different from zero and this is especially notable for risk classes 3 to 9 where the estimated reductions in error variances ranged from 30% to 60%. Looking at the t-values in column 5 it is clear that the estimates of δ_k are very significantly different for risk classes 3 to 9, while that of class 2 is borderline and that of class 1 is not significantly different from 0. Thus, the unobserved factor has strong impacts on the default rates of classes 3 to 9 and less so on classes 1 and 2.

Finally, the MLE of the unobserved factor auto-correlation coefficient ρ_1 is significantly negative. A possible reason for this effect might be alternating month lengths. Since a longer month leaves more room for defaulting than a shorter month and longer and shorter months tend to alternate, it may be expected that there is a negative auto-correlated small driving force coming from this effect. We looked into this possibility by extending the $AR(1)-U(1)$ model by including month length as an observable factor but this did not have

a significant effect. The auto-correlation of the remaining unobserved factor stayed negative and was minimally affected. It is an open challenge to try to find a more convincing interpretation of this negative auto-correlation effect in the unobserved component.

Figures 4.3.4a and 4.3.4b compare the fitted with the actual probit default rates. Broadly speaking the $AR(1)-U(1)$ model seems to fit the data well. It is especially notable that the fit is good at month 45 where a prominent shock appears in most of the series. The good fit is due to the estimated unobserved component having a prominent jump at that month. This feature is evident in Figure 4.3.5 which plots the estimated component series.

We also fitted the $AR(1)-U(1)$ model to the logit transformed default rates and repeated all the calculations above for this case. Table 4.3.7 shows the equivalent of Table 4.3.5 for the logit transforms. Although there are some differences in the values of the estimates, the overall findings detailed above for the probits are also valid for the logits. The equivalents of Figures 4.3.4a and 4.3.4b for the logit case are not exhibited here since they show no new features. Figure 4.3.5 shows the estimated component series for both the probit and the logit fits and it is remarkable that they are so similar that their plots are indistinguishable. This again indicates that the actual underlying properties of the default rates are not affected much by the type of transformation used in the modelling process.

Table 4.3.8 gives P-values of the Jarque-Bera normality test on the residuals of both the $AR(1)-U(1)$ and the $AR(1)$ models fitted to the probit and logit transformed rates. The P-values in columns 2 and 4 were obtained from PROC AUTOREG while those in column 3 are parametric bootstrap based P-values (see e.g. Davison and Hinkley, 1997:140). Only two of the risk classes have problematically small P-values for the $AR(1)-U(1)$ model while this is the case for at least five classes for the $AR(1)$ model. This shows substantial progress in modelling the underlying error variability as being normally distributed but raises the question of whether further progress would be possible by including more unobserved components in the $AR(1)-U(1)$ model, i.e. fitting an $AR(1)-U(M)$ model with $M \geq 2$.

Other evidence motivating that the $AR(1)-U(1)$ model may not be adequate is contained in the correlations of the $AR(1)-U(1)$ model residuals. Table 4.3.9a shows these correlations. Apart from normality we also assumed that the error components e_{tk} are independent between risk classes under the $AR(1)-U(1)$ model, i.e. that $Cov(e_{tk}, e_{tl}) = 0$ for $k \neq l$. Seeing that the residuals $e_{t|T}$ are estimates of the e_t's one would therefore expect that the empirical correlation coefficients between the components of the $e_{t|T}$'s given by the off-diagonal elements in Table 4.3.9a should be close to zero. In many instances this seems to

Figure 4.3.4a Actual probit and fitted values of classes 1 to 5 for $AR(1)-U(1)$ model

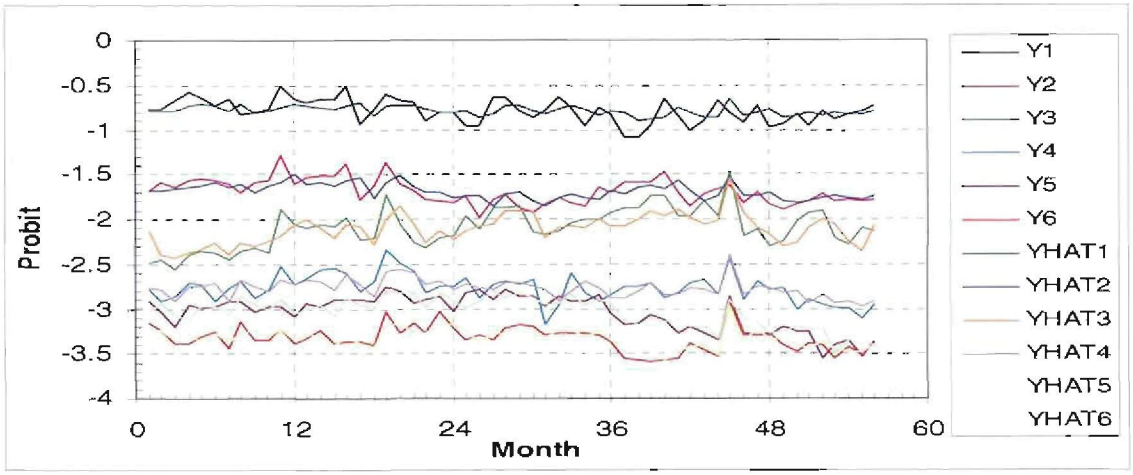


Figure 4.3.4b Actual probit and fitted values of classes 6 to 9 for $AR(1)-U(1)$ model

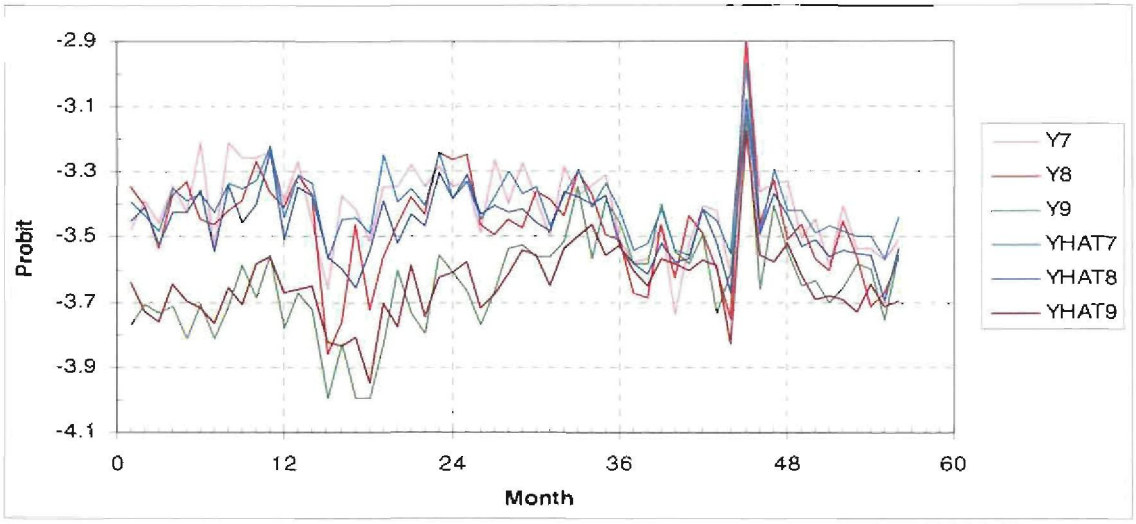


Figure 4.3.5 Estimated unobserved component $u_{t|T}$ of $AR(1)-U(1)$ model fitted to the probit and logit transformed default rates

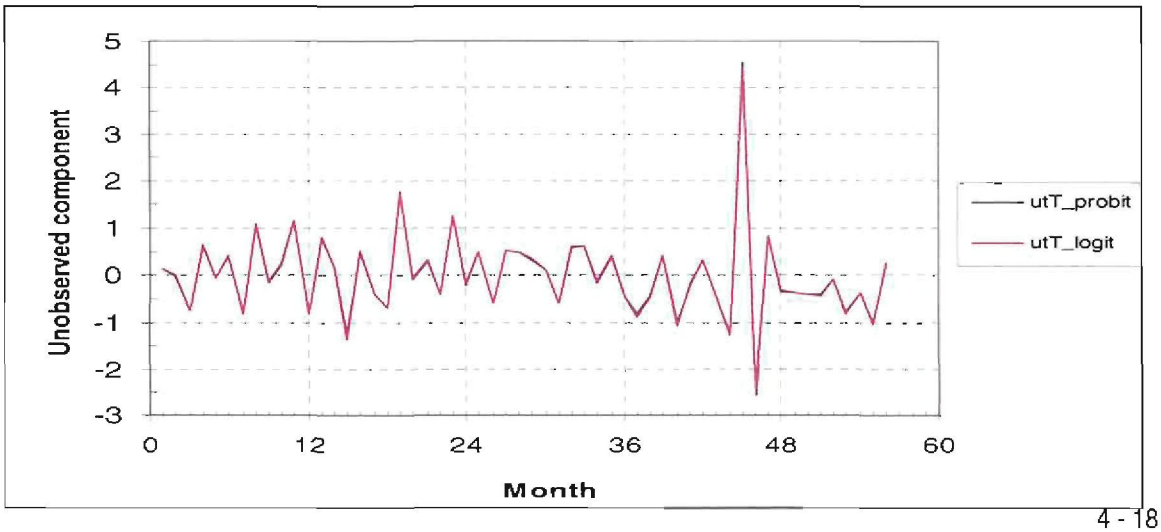


Table 4.3.7 MLE's of $AR(1)-U(1)$ and $AR(1)$ models fitted to logit transformed rates

	$AR(1)-U(1)$ model					$AR(1)$ model			Variance relation
Parm	Init	MLE	SE	BSE	t	MLE	SE	t	
α_1	-1.2870	-1.2876	0.0428	0.0429	-30.1012	-1.2870	0.0447	-28.7814	-
α_2	-3.0496	-3.0513	0.0758	0.0778	-40.2626	-3.0496	0.0754	-40.4527	-
α_3	-4.0789	-4.1796	0.2463	0.1927	-16.9691	-4.0789	0.1495	-27.2856	-
α_4	-5.8941	-5.9016	0.1030	0.0999	-57.2887	-5.8941	0.0983	-59.9554	-
α_5	-6.7333	-6.7439	0.2070	0.2164	-32.5813	-6.7333	0.2284	-29.4769	-
α_6	-7.7394	-7.7307	0.1111	0.1109	-69.5563	-7.7394	0.0986	-78.4735	-
α_7	-8.0139	-8.0252	0.0850	0.0846	-94.3887	-8.0139	0.0751	-106.6608	-
α_8	-8.2406	-8.2410	0.1202	0.1203	-68.5513	-8.2406	0.1013	-81.3797	-
α_9	-8.9197	-8.9565	0.1792	0.1752	-49.9834	-8.9197	0.1333	-66.9145	-
β_1	0.3635	0.3389	0.1248	0.1282	2.7145	0.3635	0.1247	2.9161	-
β_2	0.5670	0.5796	0.1071	0.1164	5.4095	0.5670	0.1106	5.1249	-
β_3	0.6063	0.7339	0.1248	0.1018	5.8782	0.6063	0.1118	5.4214	-
β_4	0.3865	0.4770	0.1087	0.1133	4.3891	0.3865	0.1251	3.0895	-
β_5	0.7666	0.7735	0.0781	0.0988	9.8993	0.7666	0.0932	8.2241	-
β_6	0.3852	0.5540	0.0980	0.1037	5.6509	0.3852	0.1239	3.1090	-
β_7	0.1171	0.3620	0.1065	0.1087	3.3982	0.1171	0.1335	0.8774	-
β_8	0.2651	0.4770	0.1068	0.1066	4.4662	0.2651	0.1294	2.0491	-
β_9	0.4767	0.6645	0.1027	0.1045	6.4714	0.4767	0.1181	4.0356	-
σ_1^2	0.0443	0.0452	0.0086	0.0084	5.2618	0.0465	0.0088	5.2915	0.0466
σ_2^2	0.0564	0.0594	0.0115	0.0112	5.1811	0.0643	0.0121	5.2920	0.0643
σ_3^2	0.1241	0.1377	0.0287	0.0272	4.7895	0.1921	0.0363	5.2915	0.1960
σ_4^2	0.1130	0.1374	0.0286	0.0278	4.7972	0.2094	0.0396	5.2912	0.2109
σ_5^2	0.1076	0.1232	0.0260	0.0250	4.7456	0.2000	0.0378	5.2917	0.1996
σ_6^2	0.0829	0.0859	0.0218	0.0206	3.9459	0.2110	0.0399	5.2916	0.2171
σ_7^2	0.1029	0.0964	0.0262	0.0242	3.6793	0.2474	0.0468	5.2912	0.2612
σ_8^2	0.1656	0.1554	0.0369	0.0343	4.2137	0.3148	0.0595	5.2916	0.3287
σ_9^2	0.1675	0.1588	0.0354	0.0338	4.4888	0.2845	0.0538	5.2912	0.2964
δ_1	0.0325	0.0367	0.0316	0.0305	1.1643	-	-	-	-
δ_2	0.0523	0.0701	0.0373	0.0367	1.8778	-	-	-	-
δ_3	0.1346	0.2416	0.0650	0.0617	3.7190	-	-	-	-
δ_4	0.1666	0.2712	0.0630	0.0640	4.3011	-	-	-	-
δ_5	0.1613	0.2764	0.0598	0.0597	4.6176	-	-	-	-
δ_6	0.1894	0.3623	0.0609	0.0608	5.9496	-	-	-	-
δ_7	0.2023	0.4060	0.0678	0.0659	5.9839	-	-	-	-
δ_8	0.2068	0.4163	0.0776	0.0755	5.3662	-	-	-	-
δ_9	0.1836	0.3709	0.0756	0.0720	4.9076	-	-	-	-
ρ_1	-0.2831	-0.4261	0.1450	0.1557	-2.9381	-	-	-	-

Table 4.3.8 P-values of Jarque-Bera normality tests on residuals of $AR(1)-U(1)$ and $AR(1)$ models

PROBITS	$AR(1)-U(1)$ model Autoreg	$AR(1)-U(1)$ model Bootstrap	$AR(1)$ model Autoreg
pdr1	0.5936	0.5059	0.7096
pdr2	0.93990	0.9365	0.6253
pdr3	0.8719	0.5481	0.0581
pdr4	0.0175	0.0268	0.2734
pdr5	0.0741	0.0279	0.0226
pdr6	0.6743	0.8062	<0.0001
pdr7	0.6089	0.5508	<0.0001
pdr8	0.3214	0.1322	<0.0001
pdr9	0.5975	0.3878	0.0050
LOGITS			
ldr1	0.5603	0.4607	0.6599
ldr2	0.9813	0.9775	0.8829
ldr3	0.9394	0.5645	0.2111
ldr4	0.0017	0.0125	0.2058
ldr5	0.0051	0.0075	0.0086
ldr6	0.7307	0.8196	<0.0001
ldr7	0.6216	0.5940	<0.0001
ldr8	0.2231	0.0836	<0.0001
ldr9	0.6857	0.5053	0.0482

Table 4.3.9a Correlations of the residuals of the $AR(1)-U(1)$ model fitted to the probit default rates

	Correlations for $AR(1)-U(1)$ model								
PROBITS	**pdr1**	**pdr2**	**pdr3**	**pdr4**	**pdr5**	**pdr6**	**pdr7**	**pdr8**	**pdr9**
pdr1	1.0000	0.6272	0.1491	0.1598	-0.0355	-0.0347	-0.1706	-0.2839	0.1006
pdr2	0.6272	1.0000	0.4739	0.2684	-0.0103	-0.2079	-0.2381	-0.4632	0.0342
pdr3	0.1491	0.4739	1.0000	0.1856	-0.1192	-0.2743	-0.2582	-0.3948	0.0540
pdr4	0.1598	0.2684	0.1856	1.0000	0.1616	-0.1005	-0.3680	-0.2486	-0.1923
pdr5	-0.0355	-0.0103	-0.1192	0.1616	1.0000	0.0453	-0.3274	-0.1560	-0.1808
pdr6	-0.0347	-0.2079	-0.2743	-0.1005	0.0453	1.0000	-0.0790	-0.1048	-0.3884
pdr7	-0.1706	-0.2381	-0.2582	-0.3680	-0.3274	-0.0790	1.0000	0.1023	-0.1186
pdr8	-0.2839	-0.4632	-0.3948	-0.2486	-0.1560	-0.1048	0.1023	1.0000	-0.0197
pdr9	0.1006	0.0342	0.0540	-0.1923	-0.1808	-0.3884	-0.1186	-0.0197	1.0000

Table 4.3.9b Bootstrap P-values for testing for zero correlation against positive alternatives of the residuals of the $AR(1)-U(1)$ model fitted to the probit default rates

	Correlations for $AR(1)-U(1)$ model								
PROBITS	**pdr1**	**pdr2**	**pdr3**	**pdr4**	**pdr5**	**pdr6**	**pdr7**	**pdr8**	**pdr9**
pdr1	-	0.0000	0.1043	0.1035	0.5505	0.5109	0.8476	0.9724	0.1901
pdr2	0.0000	-	0.0006	0.0124	0.4468	0.8786	0.9179	0.9996	0.3043
pdr3	0.1043	0.0006	-	0.0313	0.6340	0.8885	0.8435	0.9899	0.1584
pdr4	0.1035	0.0124	0.0313	-	0.0414	0.4336	0.9637	0.8581	0.7643
pdr5	0.5505	0.4468	0.6340	0.0414	-	0.0957	0.9202	0.6262	0.7497
pdr6	0.5109	0.8786	0.8885	0.4336	0.0957	-	0.1210	0.2899	0.9643
pdr7	0.8476	0.9179	0.8435	0.9637	0.9202	0.1210	-	0.0173	0.3344
pdr8	0.9724	0.9996	0.9899	0.8581	0.6262	0.2899	0.0173	-	0.1967
pdr9	0.1901	0.3043	0.1584	0.7643	0.7497	0.9643	0.3344	0.1967	-

be the case, but there are also large values such as between risk classes 1 and 2 having the value 0.6272. Of course we need a way to judge what is "large" in this context, e.g. by establishing critical values to enable formal testing of the zero correlation hypothesis here. This would require deriving the distribution of the empirical correlations of the residuals under the independence hypothesis which is a problem beyond the scope of this dissertation. As an alternative we can compute parametric bootstrap P-values for testing zero correlation. Table 4.3.9b shows these P-values. This indicates that a correlation of 0.6272 between risk classes 1 and 2 is significantly different from 0 as is also true for a number of other pairs of classes. The extremely high P-value for classes 2 and 8 also suggests significant negative correlation between these classes as is also the case for a few other pairs of classes. If we compare the entries in Table 4.3.9a with those in Table 3.3.4 it is clear that the $AR(1)-U(1)$ model did little in terms of explaining the correlation between risk classes 1 and 3. This is different for the other classes: comparing the correlations just above the diagonal (adjacent classes) in Table 4.3.9a with those of Table 3.3.4 we see that the $AR(1)-U(1)$ model did quite a lot in terms of explaining the correlations of risk classes 3 to 9. Here again classes 1 and 2 seem to be different from 3 to 9 in terms of what can be achieved by adding one unobserved component to the $AR(1)$ model.

4.3.7 Application of an $AR(1)-U(2)$ model to the home loans transformed default rates

Here we apply the model $AR(1)-U(2)$ to the transformed default rates and we begin with the probits. Table 4.3.10 shows the parameter estimates for the $AR(1)-U(2)$ model in columns 2 to 5. For ease of reference, the estimates for the $AR(1)-U(1)$ model as given in Table 4.3.5 are repeated in columns 6 to 9 of Table 4.3.10. These can now be compared to the corresponding estimates under the $AR(1)-U(2)$ model. The estimates of the level parameters are virtually the same for both models. The MLE's of the AR coefficients of the $AR(1)-U(2)$ model are again notably larger for risk classes 1 to 3, but the others are quite similar. The MLE's of the error variances of the $AR(1)-U(2)$ model are smaller for risk classes 1 to 3 and also for risk class 8. Note especially that the error variance estimate for risk class 2 is very small. Indeed, the value 0.0001 given in Table 4.3.10 is the lower bound specified by our fitting program for such error variances. It appears that the fitted model chose the second unobserved component such that the fit to the data of the second risk class was almost exact. This is a somewhat puzzling aspect of this model fit. The factor loadings for the first unobserved component (i.e. the δ_{k1}'s) are comparable in magnitude for the $AR(1)-U(1)$ and $AR(1)-U(2)$ models, indicating that the first unobserved component

Table 4.3.10 MLE's of $AR(1)-U(2)$ and $AR(1)-U(1)$ models fitted to the probit transformed rates

	$AR(1)-U(2)$ model				$AR(1)-U(1)$ model			
Parm	**Init**	**MLE**	**SE**	**t**	**Init**	**MLE**	**SE**	**t**
α_1	-0.7829	-0.7887	0.0354	-22.2896	-0.7829	-0.7832	0.0246	-31.8704
α_2	-1.6902	-1.7076	0.0419	-40.7181	-1.6901	-1.6908	0.0345	-49.0784
α_3	-2.1039	-2.5200	0.1405	-17.9330	-2.1162	-2.1502	0.0859	-25.0221
α_4	-2.7725	-2.7818	0.0350	-79.4363	-2.7727	-2.7749	0.0337	-82.4306
α_5	-3.0522	-3.0467	0.0588	-51.7914	-3.0334	-3.0351	0.0612	-49.5730
α_6	-3.3282	-3.3301	0.0284	-117.2022	-3.3272	-3.3239	0.0310	-107.0906
α_7	-3.4026	-3.4085	0.0223	-153.0874	-3.4026	-3.4054	0.0235	-144.6484
α_8	-3.4637	-3.4666	0.0308	-112.4669	-3.4635	-3.4631	0.0324	-106.7843
α_9	-3.6402	-3.6604	0.0474	-77.2361	-3.6418	-3.6504	0.0453	-80.5836
β_1	0.3606	0.5790	0.1085	5.3349	0.3614	0.3367	0.1247	2.6992
β_2	0.5564	0.7416	0.0756	9.8118	0.5605	0.5747	0.1075	5.3481
β_3	0.5711	0.9681	0.0431	22.4533	0.5791	0.7138	0.1164	6.1325
β_4	0.3742	0.5142	0.1062	4.8416	0.3838	0.4852	0.1075	4.5142
β_5	0.7261	0.7694	0.0761	10.1046	0.7648	0.7768	0.0774	10.0390
β_6	0.3641	0.5137	0.0964	5.3283	0.3658	0.5524	0.0980	5.6377
β_7	0.1055	0.3327	0.1038	3.2046	0.1053	0.3685	0.1058	3.4828
β_8	0.2490	0.4570	0.0993	4.6019	0.2490	0.4795	0.1065	4.5007
β_9	0.4568	0.6668	0.1060	6.2927	0.4597	0.6570	0.1024	6.4177
σ_1^2	0.0065	0.0086	0.0018	4.8499	0.0148	0.0150	0.0029	5.2635
σ_2^2	0.0024	0.0001	0.0016	0.0618	0.0119	0.0125	0.0024	5.1806
σ_3^2	0.0159	0.0141	0.0031	4.5261	0.0208	0.0221	0.0046	4.7560
σ_4^2	0.0105	0.0138	0.0028	4.9450	0.0116	0.0141	0.0030	4.7743
σ_5^2	0.0093	0.0107	0.0022	4.8451	0.0092	0.0105	0.0022	4.7479
σ_6^2	0.0055	0.0065	0.0016	4.0849	0.0065	0.0068	0.0017	3.9852
σ_7^2	0.0054	0.0066	0.0017	3.7884	0.0077	0.0073	0.0020	3.7336
σ_8^2	0.0058	0.0077	0.0021	3.6870	0.0119	0.0112	0.0026	4.2222
σ_9^2	0.0114	0.0109	0.0023	4.7245	0.0114	0.0106	0.0023	4.4987
δ_{11}	0.0183	0.0428	0.0315	1.3598	0.0183	0.0213	0.0181	1.1758
δ_{21}	0.0241	0.0720	0.0357	2.0146	0.0241	0.0335	0.0170	1.9661
δ_{31}	0.0555	0.1457	0.0286	5.0916	0.0555	0.1043	0.0262	3.9864
δ_{41}	0.0547	0.0934	0.0203	4.5998	0.0547	0.0911	0.0205	4.4484
δ_{51}	0.0480	0.0772	0.0200	3.8499	0.0480	0.0832	0.0176	4.7350
δ_{61}	0.0533	0.0919	0.0237	3.8785	0.0533	0.1031	0.0173	5.9642
δ_{71}	0.0557	0.1011	0.0265	3.8217	0.0557	0.1131	0.0189	5.9953
δ_{81}	0.0562	0.0970	0.0356	2.7222	0.0562	0.1147	0.0211	5.4320
δ_{91}	0.0476	0.0951	0.0211	4.5003	0.0476	0.0985	0.0197	5.0090

Table 4.3.10(contd) MLE's of $AR(1)-U(2)$ and $AR(1)-U(1)$ models fitted to the probit transformed rates

	$AR(1)-U(2)$ model				$AR(1)-U(1)$ model			
Parm	**Init**	**MLE**	**SE**	**t**	**Init**	**MLE**	**SE**	**t**
δ_{12}	0.0662	0.0764	0.0209	3.6610	-	-	-	-
δ_{22}	0.0708	0.0947	0.0258	3.6791	-	-	-	-
δ_{32}	0.0523	0.0445	0.0474	0.9392	-	-	-	-
δ_{42}	0.0255	-0.0042	0.0338	-0.1249	-	-	-	-
δ_{52}	-0.0052	-0.0279	0.0281	-0.9922	-	-	-	-
δ_{62}	-0.0246	-0.0468	0.0315	-1.4828	-	-	-	-
δ_{72}	-0.0359	-0.0543	0.0349	-1.5547	-	-	-	-
δ_{82}	-0.0570	-0.0834	0.0343	-2.4281	-	-	-	-
δ_{92}	-0.0087	-0.0215	0.0326	-0.6572	-	-	-	-
ρ_1	-0.2766	-0.4634	0.1306	-3.5473	-0.2766	-0.4385	0.1408	-3.1138
ρ_2	0.0223	-0.1653	0.1447	-1.1422	-	-	-	-

of the $AR(1)-U(2)$ model is roughly the same as the unobserved component of the $AR(1)-U(1)$ model, loading mostly on risk classes 3 to 9. Further evidence to this effect is given in Figure 4.3.6 below, which plots the estimated unobserved components of both models.

By contrast, the second unobserved component loads mostly on risk classes 1 and 2 (and to some extent, also on class 8) since its loadings (i.e. the δ_{k2}'s) are not significantly different from zero for the other risk classes. Lastly, the MLE of ρ_1 of the $AR(1)-U(2)$ model is virtually the same as that of the $AR(1)-U(1)$ model, again suggesting that the first unobserved component of the $AR(1)-U(2)$ is the same as the unobserved component of the $AR(1)-U(1)$ model. The MLE of ρ_2 of the $AR(1)-U(2)$ model is not significantly different from zero and this is consistent with the second unobserved component contributing a new feature in the data over and above what can be catered for by the $AR(1)-U(1)$ model.

Figure 4.3.7 compares the actual and fitted probit default series of risk classes 1 and 2 for the two models. In the case of risk class 1, there is a clear improvement in the fit of the $AR(1)-U(2)$ model over and above that of the $AR(1)-U(1)$ model. In the case of risk class 2, the plot of the $AR(1)-U(2)$ model is so close to the actual data that they are indistinguishable on the graph. This is consistent with the puzzling feature noted above that the $AR(1)-U(2)$ model seems to overfit on this risk class. To save space we do not show

Figure 4.3.6 Estimated unobserved components $u_{t|T}$ of the $AR(1)-U(1)$ model, and $u_{t|T1}$ and $u_{t|T2}$ of the $AR(1)-U(2)$ model fitted to the probit transformed default rates

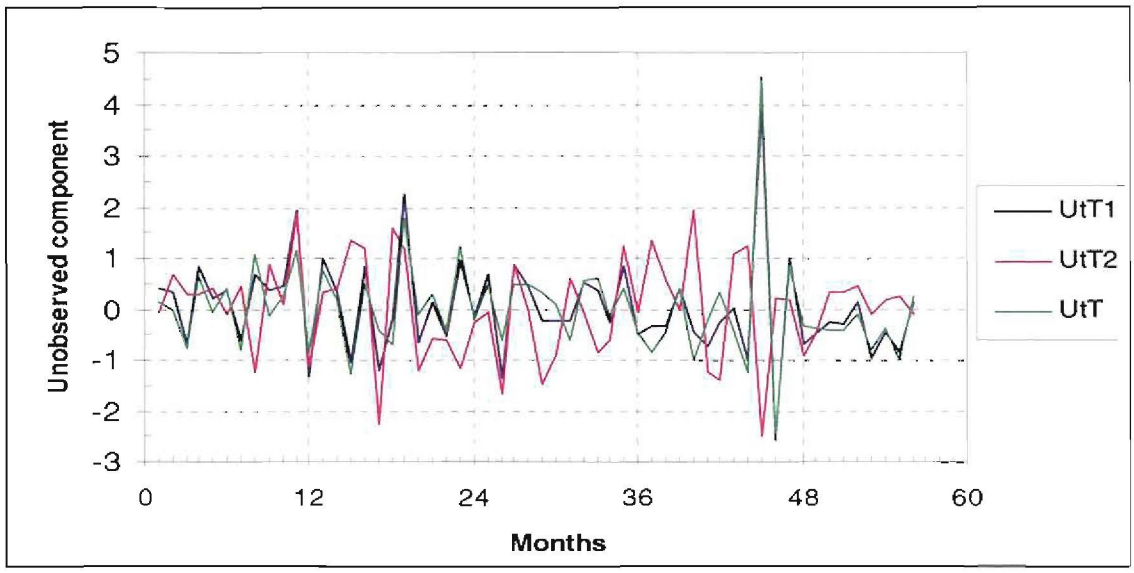
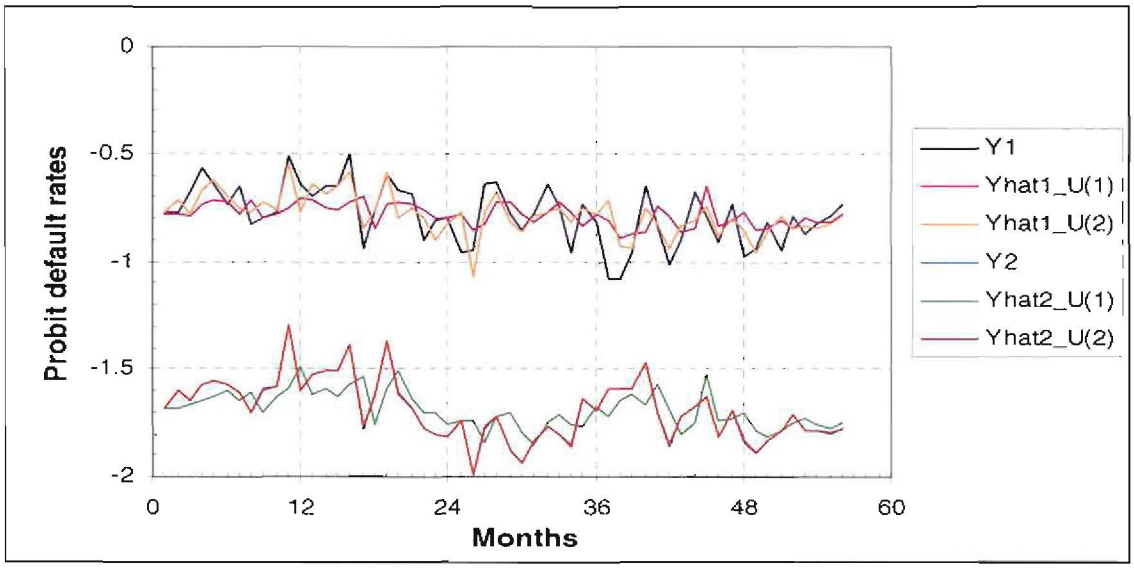


Figure 4.3.7 Actual probit and fitted values of classes 1 and 2 for $AR(1)-U(1)$ and $AR(1)-U(2)$ models



the plots of fitted and actual series for the risk classes 3 to 9 since the $AR(1)-U(2)$ model does not improve these relative to what is achieved by the $AR(1)-U(1)$ as plotted in Figures 4.3.4a and 4.3.4b. The analysis above for the probits was also done for the logits. The results were largely similar and in particular the overfitting on risk class 2 also occurs. We thus do not show the details of this analysis for the logits also.

It might be thought that the puzzling overfit of the $AR(1)-U(2)$ model on risk class 2 could be due to the likelihood optimisation converging to an incorrect local rather than global

maximum. To look into this possibility we investigated starting the optimisation at different initial parameter estimates. One possibility is to use the MLE's of the $AR(1)-U(1)$ model together with the choice of 0 as initial value for the δ_{k2}'s and for ρ_2. It turned out that the optimisation iteration started there, again converged to the MLE's shown in Table 4.3.10. Another possibility is to start at randomly chosen initial estimates. We ran the optimisation repeatedly each time starting from a set of estimates obtained by adding randomly chosen perturbations to each of the 47 parameter values selected according to our initialisation strategy. We recorded the parameter values towards which convergence took place as well as the value of the log-likelihood function there. We also varied the scale of the perturbations to see what difference this would make. It turned out that when small scale perturbations were made convergence always occurred to the MLE's listed in Table 4.3.10 above. As the scale of the perturbations increased, convergence occasionally occurred to values different from the MLE's in Table 4.3.10. However, invariably the value of the log-likelihood function in such cases was lower than that at the MLE's of Table 4.3.10. This investigation indicates that the log-likelihood function may have local maxima. Apparently the one given as the MLE in Table 4.3.10 is hard to improve on and it may be the global maximum. However, a 47-dimensional space is very large and we cannot be 100% certain that the MLE's of Table 4.3.10 are truly globally maximising. This is a general issue in any multi-parameter numerical maximum likelihood estimation problem.

Practically speaking the issue remains: what can be done about the overfit on risk class 2? Inspecting Table 4.3.10 again the MLE's of the error variances of all the risk classes other than class 2 are quite similar taking their standard errors into account. The issue can be viewed as an unbalanced treatment of risk classes and a possible solution is to insist on more balance in this regard, e.g. by requiring that the σ_k^2's are the same for all risk classes. This has the additional advantage of reducing the number of parameters by 8. Table 4.3.11 compares the MLE's under this restriction with those in Table 4.3.10 where the error variances were not forced to be equal. The MLE of the common error variance was found to be 0.00892 which happens to be close to the average 0.00878 of the 9 separate estimates under the unrestricted model. The estimates of the other parameters are quite similar under the two models and most of the discussion above continues to apply with some changes. As in Figure 4.3.6, the estimate of the first unobserved component remains close to that of the $AR(1)-U(1)$ model while the second one differs from that shown in Figure 4.3.6. The overfit on risk class 2 in Figure 4.3.7 is no longer present when we graph the fits under the restricted model.

Table 4.3.11 MLE's of $AR(1)-U(2)$ models with and without equal variances fitted to the probit transformed rates

	$AR(1)-U(2)$ equal var model				$AR(1)-U(2)$ model			
Parm	**Init**	**MLE**	**SE**	**t**	**Init**	**MLE**	**SE**	**t**
α_1	-0.7829	-0.7984	0.0333	-23.9954	-0.7829	-0.7887	0.0354	-22.2896
α_2	-1.6902	-1.7156	0.0468	-36.6187	-1.6902	-1.7076	0.0419	-40.7181
α_3	-2.1039	-2.4588	0.1555	-15.8163	-2.1039	-2.5200	0.1405	-17.9330
α_4	-2.7725	-2.7900	0.0336	-83.0592	-2.7725	-2.7818	0.0350	-79.4363
α_5	-3.0522	-3.0438	0.0583	-52.1880	-3.0522	-3.0467	0.0588	-51.7914
α_6	-3.3282	-3.3277	0.0319	-104.2431	-3.3282	-3.3301	0.0284	-117.2022
α_7	-3.4026	-3.4050	0.0238	-142.9101	-3.4026	-3.4085	0.0223	-153.0874
α_8	-3.4637	-3.4600	0.0323	-107.0207	-3.4637	-3.4666	0.0308	-112.4669
α_9	-3.6402	-3.6604	0.0443	-82.6365	-3.6402	-3.6604	0.0474	-77.2361
β_1	0.3606	0.5385	0.1168	4.6119	0.3606	0.5790	0.1085	5.3349
β_2	0.5564	0.6769	0.1049	6.4538	0.5564	0.7416	0.0756	9.8118
β_3	0.5711	0.9435	0.0546	17.2785	0.5711	0.9681	0.0431	22.4533
β_4	0.3742	0.5641	0.0899	6.2717	0.3742	0.5142	0.1062	4.8416
β_5	0.7261	0.7835	0.0708	11.0692	0.7261	0.7694	0.0761	10.1046
β_6	0.3641	0.5259	0.1087	4.8363	0.3641	0.5137	0.0964	5.3283
β_7	0.1055	0.3106	0.1177	2.6395	0.1055	0.3327	0.1038	3.2046
β_8	0.2490	0.4392	0.1093	4.0195	0.2490	0.4570	0.0993	4.6019
β_9	0.4568	0.6650	0.0990	6.7158	0.4568	0.6668	0.1060	6.2927
σ_1^2	0.0081	0.0089	0.0006	13.9731	0.0065	0.0086	0.0018	4.8499
σ_2^2	0.0081	0.0089	0.0006	13.9731	0.0024	0.0001	0.0016	0.0618
σ_3^2	0.0081	0.0089	0.0006	13.9731	0.0159	0.0141	0.0031	4.5261
σ_4^2	0.0081	0.0089	0.0006	13.9731	0.0105	0.0138	0.0028	4.9450
σ_5^2	0.0081	0.0089	0.0006	13.9731	0.0093	0.0107	0.0022	4.8451
σ_6^2	0.0081	0.0089	0.0006	13.9731	0.0055	0.0065	0.0016	4.0849
σ_7^2	0.0081	0.0089	0.0006	13.9731	0.0054	0.0066	0.0017	3.7884
σ_8^2	0.0081	0.0089	0.0006	13.9731	0.0058	0.0077	0.0021	3.6870
σ_9^2	0.0081	0.0089	0.0006	13.9731	0.0114	0.0109	0.0023	4.7245
δ_{11}	0.0183	0.0350	0.0236	1.4812	0.0183	0.0428	0.0315	1.3598
δ_{21}	0.0241	0.0566	0.0246	2.2966	0.0241	0.0720	0.0357	2.0146
δ_{31}	0.0555	0.1484	0.0284	5.2277	0.0555	0.1457	0.0286	5.0916
δ_{41}	0.0547	0.1039	0.0190	5.4741	0.0547	0.0934	0.0203	4.5998
δ_{51}	0.0480	0.0819	0.0176	4.6470	0.0480	0.0772	0.0200	3.8499
δ_{61}	0.0533	0.0923	0.0206	4.4785	0.0533	0.0919	0.0237	3.8785
δ_{71}	0.0557	0.0969	0.0234	4.1374	0.0557	0.1011	0.0265	3.8217
δ_{81}	0.0562	0.0984	0.0287	3.4353	0.0562	0.0970	0.0356	2.7222
δ_{91}	0.0476	0.0974	0.0195	4.9960	0.0476	0.0951	0.0211	4.5003

Table 4.3.11 (contd) MLE's of $AR(1)-U(2)$ models with and without equal variances fitted to the probit transformed rates

	$AR(1)-U(2)$ equal var model				$AR(1)-U(2)$ model			
Parm	**Init**	**MLE**	**SE**	**t**	**Init**	**MLE**	**SE**	**t**
δ_{12}	0.0662	0.0719	0.0203	3.5406	0.0662	0.0764	0.0209	3.6610
δ_{22}	0.0708	0.0746	0.0201	3.7100	0.0708	0.0947	0.0258	3.6791
δ_{32}	0.0523	0.0672	0.0359	1.8699	0.0523	0.0445	0.0474	0.9392
δ_{42}	0.0255	0.0119	0.0293	0.4071	0.0255	-0.0042	0.0338	-0.1249
δ_{52}	-0.0052	-0.0216	0.0244	-0.8862	-0.0052	-0.0279	0.0281	-0.9922
δ_{62}	-0.0246	-0.0377	0.0264	-1.4312	-0.0246	-0.0468	0.0315	-1.4828
δ_{72}	-0.0359	-0.0524	0.0273	-1.9220	-0.0359	-0.0543	0.0349	-1.5547
δ_{82}	-0.0570	-0.0800	0.0275	-2.9101	-0.0570	-0.0834	0.0343	-2.4281
δ_{92}	-0.0087	-0.0183	0.0265	-0.6915	-0.0087	-0.0215	0.0326	-0.6572
ρ_1	-0.2766	-0.4910	0.1269	-3.8703	-0.2766	-0.4634	0.1306	-3.5473
ρ_2	0.0223	-0.0358	0.1749	-0.2047	0.0223	-0.1653	0.1447	-1.1422

4.4 Maximum likelihood inference via the EM algorithm combined with Kalman filtering

4.4.1 The EM algorithm

The EM ("expectation-maximisation") algorithm of Dempster *et al.* (1977) provides an approach to maximum likelihood inference that is particularly well suited to models involving unobserved components. We saw above that the likelihood function of such models is too complex to handle analytically which makes it impossible to calculate the MLE's analytically and forces one to use numerical optimisation methods. As mentioned above, one cannot be absolutely sure that the numerical optimisation actually yields a global maximum rather than a local maximum especially when we work with short series and large numbers of parameters as is the case with the home loans data.

The EM algorithm follows an entirely different approach which enables at least part of the optimisation to be done analytically. According to some comparative studies reported in the literature this leads to greater stability in the results (see e.g. Metaxoglou and Smith (2007)). In this section we apply the EM-approach to maximum likelihood estimation for our AR model with unobserved components.

The first step in the EM-algorithm is to write down an expression for the "complete" log-likelihood function $\log L^*(\boldsymbol{\Theta})$, i.e. the logarithm of the joint density function of the observed as well as unobserved components, $y_1, y_2, \ldots, y_T$ and $\boldsymbol{u}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_T$. The E-step finds an expression for $Q(\boldsymbol{\Theta}, \boldsymbol{\Theta}_0) = E_{\boldsymbol{\Theta}_0}[\log L^*(\boldsymbol{\Theta}) | \boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T]$ where $E_{\boldsymbol{\Theta}_0}$ means that the parameter

value is at Θ_0 when the conditional expectation is calculated. The M-step then maximises $Q(\Theta,\Theta_0)$ over Θ. From an initial estimate Θ_0 this maximisation leads to a better estimate Θ_1. Iterating this process constitutes the EM-algorithm (see e.g. Dempster *et al.* (1977) or Liu *et al.* (1998) or the other papers cited above).

For the first step we use conditional decomposition again to get the complete log-likelihood function. Given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_{t-1} = y_{t-1}$ and also $U_0 = u_0, U_1 = u_1, \ldots, U_t = u_t$ the distribution of Y_t is $N_K(\alpha + \beta y_{t-1} + \Delta u_t, \Sigma)$ so that this conditional distribution has log-density function

$$-\tfrac{1}{2}(y_t - \alpha - \beta y_{t-1} - \Delta u_t)'\Sigma^{-1}(y_t - \alpha - \beta y_{t-1} - \Delta u_t) - \tfrac{1}{2}\log(\det(\Sigma)) - \tfrac{K}{2}\log(2\pi) \quad . \tag{4.4.1}$$

Further given $Y_1 = y_1, Y_2 = y_2, \ldots, Y_{t-1} = y_{t-1}$ and also $U_0 = u_0, U_1 = u_1, \ldots, U_{t-1} = u_{t-1}$ the distribution of U_t is $N_M(\rho u_{t-1}, DD')$ with log-density

$$-\tfrac{1}{2}(u_t - \rho u_{t-1})'(DD')^{-1}(u_t - \rho u_{t-1}) - \tfrac{1}{2}\log(\det(DD')) - \tfrac{M}{2}\log(2\pi) \quad . \tag{4.4.2}$$

We can then add (4.4.1) and (4.4.2) to get the log of the joint density of Y_t and U_t given $U_0 = u_0, U_1 = u_1, \ldots, U_{t-1} = u_{t-1}$ and $Y_1 = y_1, Y_2 = y_2, \ldots, Y_{t-1} = y_{t-1}$. This can then be summed over $t = 1, 2, \ldots, T$ to get the complete log-likelihood function. Dropping constant terms and replacing the fixed values of the unobserved components $u_0, u_1, \ldots, u_T$ by their random equivalents $U_0, U_1, \ldots, U_T$ we get

$$\begin{aligned} \log L^*(\Theta) = \; & -\tfrac{1}{2}(U_0'U_0) \\ & -\tfrac{1}{2}\textstyle\sum_{t=1}^{T}(U_t - \rho U_{t-1})'(DD')^{-1}(U_t - \rho U_{t-1}) - \tfrac{T}{2}\log(\det(DD')) \\ & -\tfrac{1}{2}\textstyle\sum_{t=1}^{T}(y_t - \alpha - \beta y_{t-1} - \Delta U_t)'\Sigma^{-1}(y_t - \alpha - \beta y_{t-1} - \Delta U_t) - \tfrac{T}{2}\log(\det(\Sigma)) \end{aligned} \tag{4.4.3}$$

For the E-step we need an expression for $Q(\Theta,\Theta_0) = E_{\Theta_0}[\log L^*(\Theta) | Y_1 = y_1, \ldots, Y_T = y_T]$ where E_{Θ_0} signifies that the conditional distribution of the U_t's given $Y_1 = y_1, \ldots, Y_T = y_T$ used for this expectation operates under the parameter value Θ_0. For this purpose we use and extend the smoothing estimate notation of Section 4.4, namely

$$\begin{aligned} u_{t|T} &= E_{\Theta_0}[U_t | Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T] \\ P_{t|T} &= E_{\Theta_0}[(U_t - u_{t|T})(U_t - u_{t|T})' | Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T] \\ P_{t,t-1|T} &= E_{\Theta_0}[(U_t - u_{t|T})(U_{t-1} - u_{t-1|T})' | Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_T] \end{aligned} \tag{4.4.4}$$

At any given $\boldsymbol{\Theta}_0$ these quantities can again be calculated by Kalman filter iterations which will be detailed below. From (4.4.4) we have the expressions

$$\begin{aligned} E_{\boldsymbol{\Theta}_0}[\boldsymbol{U}_t\boldsymbol{U}_t'|\boldsymbol{Y}_1=\boldsymbol{y}_1,\boldsymbol{Y}_2=\boldsymbol{y}_2,\ldots,\boldsymbol{Y}_T=\boldsymbol{y}_T]&=\boldsymbol{P}_{t|T}+\boldsymbol{u}_{t|T}\boldsymbol{u}_{t|T}' \\ E_{\boldsymbol{\Theta}_0}[\boldsymbol{U}_t\boldsymbol{U}_{t-1}'|\boldsymbol{Y}_1=\boldsymbol{y}_1,\boldsymbol{Y}_2=\boldsymbol{y}_2,\ldots,\boldsymbol{Y}_T=\boldsymbol{y}_T]&=\boldsymbol{P}_{t,t-1|T}+\boldsymbol{u}_{t|T}\boldsymbol{u}_{t-1|T}' \end{aligned} \tag{4.4.5}$$

In particular, from (4.4.3) the first term in the evaluation of $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}_0)$ is $-\frac{1}{2}E_{\boldsymbol{\Theta}_0}[\boldsymbol{U}_0\boldsymbol{U}_0'|\boldsymbol{Y}_1=\boldsymbol{y}_1,\ldots,\boldsymbol{Y}_T=\boldsymbol{y}_T]=-\frac{1}{2}(\boldsymbol{P}_{0|T}+\boldsymbol{u}_{0|T}\boldsymbol{u}_{0|T}')$ which does not depend on $\boldsymbol{\Theta}$ and will have no effect on the M-step and may thus be ignored . To get the other terms we rewrite the summations in (4.4.3) somewhat. Using standard matrix trace properties the summation in the second line in (4.4.3) is

$$\begin{aligned} &\textstyle\sum_{t=1}^{T} tr\{(\boldsymbol{U}_t-\boldsymbol{\rho}\boldsymbol{U}_{t-1})'(\boldsymbol{D}\boldsymbol{D}')^{-1}(\boldsymbol{U}_t-\boldsymbol{\rho}\boldsymbol{U}_{t-1})\} \\ &=\textstyle\sum_{t=1}^{T} tr\{(\boldsymbol{U}_t-\boldsymbol{\rho}\boldsymbol{U}_{t-1})(\boldsymbol{U}_t-\boldsymbol{\rho}\boldsymbol{U}_{t-1})'(\boldsymbol{D}\boldsymbol{D}')^{-1}\} \\ &=tr\{[\textstyle\sum_{t=1}^{T}\boldsymbol{U}_t\boldsymbol{U}_t'-2\sum_{t=1}^{T}\boldsymbol{U}_t\boldsymbol{U}_{t-1}'\boldsymbol{\rho}'+\boldsymbol{\rho}\sum_{t=1}^{T}\boldsymbol{U}_{t-1}\boldsymbol{U}_{t-1}'\boldsymbol{\rho}'](\boldsymbol{D}\boldsymbol{D}')^{-1}\} \end{aligned} \tag{4.4.6}$$

Applying $E_{\boldsymbol{\Theta}_0}[.|\boldsymbol{Y}_1=\boldsymbol{y}_1,\ldots,\boldsymbol{Y}_T=\boldsymbol{y}_T]$ we get

$$\begin{aligned} &tr\{[\textstyle\sum_{t=1}^{T}(\boldsymbol{P}_{t|T}+\boldsymbol{u}_{t|T}\boldsymbol{u}_{t|T}')-2\sum_{t=1}^{T}(\boldsymbol{P}_{t,t-1|T}+\boldsymbol{u}_{t|T}\boldsymbol{u}_{t-1|T}')\boldsymbol{\rho}'+ \\ &\qquad\boldsymbol{\rho}\textstyle\sum_{t=1}^{T}(\boldsymbol{P}_{t-1|T}+\boldsymbol{u}_{t-1|T}\boldsymbol{u}_{t-1|T}')\boldsymbol{\rho}'](\boldsymbol{D}\boldsymbol{D}')^{-1}\} \\ &=tr\{[\boldsymbol{A}-2\boldsymbol{B}\boldsymbol{\rho}'+\boldsymbol{\rho}\boldsymbol{C}\boldsymbol{\rho}'](\boldsymbol{D}\boldsymbol{D}')^{-1}\} \\ &=\textstyle\sum_{m=1}^{M}\{a_{mm}-2\rho_m b_{mm}+\rho_m^2 c_{mm}\}/(1-\rho_m^2) \end{aligned} \tag{4.4.7}$$

Here a_{mm} is the m-th diagonal element of $\boldsymbol{A}=\sum_{t=1}^{T}(\boldsymbol{P}_{t|T}+\boldsymbol{u}_{t|T}\boldsymbol{u}_{t|T}')$, b_{mm} is the m-th diagonal element of $\boldsymbol{B}=\sum_{t=1}^{T}(\boldsymbol{P}_{t,t-1|T}+\boldsymbol{u}_{t|T}\boldsymbol{u}_{t-1|T}')$ and c_{mm} is the m-th diagonal element of $\boldsymbol{C}=\sum_{t=1}^{T}(\boldsymbol{P}_{t-1|T}+\boldsymbol{u}_{t-1|T}\boldsymbol{u}_{t-1|T}')$ and we used the facts that $\boldsymbol{\rho}$ and $\boldsymbol{D}\boldsymbol{D}'$ are diagonal with m-th diagonal elements ρ_m and $1-\rho_m^2$ respectively. Also $\log(\det(\boldsymbol{D}\boldsymbol{D}'))=\sum_{m=1}^{M}\log(1-\rho_m^2)$ and the terms involving the ρ_m's in $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}_0)$ become

$$-\tfrac{1}{2}\textstyle\sum_{m=1}^{M}\{(a_{mm}-2\rho_m b_{mm}+\rho_m^2 c_{mm})/(1-\rho_m^2)+T\log(1-\rho_m^2)\} \quad . \tag{4.4.8}$$

For the last summation in (4.4.3) introduce the abbreviation $\tilde{\boldsymbol{y}}_t=\boldsymbol{y}_t-\boldsymbol{\alpha}-\boldsymbol{\beta}\boldsymbol{y}_{t-1}$.

Then it becomes

$$\begin{aligned}&\sum_{t=1}^{T} tr\{(\tilde{\boldsymbol{y}}_t - \boldsymbol{\Delta}\boldsymbol{U}_t)'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{y}}_t - \boldsymbol{\Delta}\boldsymbol{U}_t)\}\\&= \sum_{t=1}^{T} tr\{(\tilde{\boldsymbol{y}}_t - \boldsymbol{\Delta}\boldsymbol{U}_t)(\tilde{\boldsymbol{y}}_t - \boldsymbol{\Delta}\boldsymbol{U}_t)'\boldsymbol{\Sigma}^{-1}\}\\&= tr\{[\sum_{t=1}^{T}\tilde{\boldsymbol{y}}_t\tilde{\boldsymbol{y}}_t' - 2\boldsymbol{\Delta}\sum_{t=1}^{T}\boldsymbol{U}_t\tilde{\boldsymbol{y}}_t' + \boldsymbol{\Delta}\sum_{t=1}^{T}\boldsymbol{U}_t\boldsymbol{U}_t'\boldsymbol{\Delta}']\boldsymbol{\Sigma}^{-1}\}\end{aligned} \tag{4.4.9}$$

Applying $E_{\boldsymbol{\Theta}_0}[.|\boldsymbol{Y}_1 = \boldsymbol{y}_1,\ldots,\boldsymbol{Y}_T = \boldsymbol{y}_T]$ and inserting (4.4.5) this becomes

$$\begin{aligned}&tr\{[\sum_{t=1}^{T}\tilde{\boldsymbol{y}}_t\tilde{\boldsymbol{y}}_t' - 2\boldsymbol{\Delta}\sum_{t=1}^{T}\boldsymbol{u}_{t|T}\tilde{\boldsymbol{y}}_t' + \boldsymbol{\Delta}\sum_{t=1}^{T}(\boldsymbol{P}_{t|T} + \boldsymbol{u}_{t|T}\boldsymbol{u}_{t|T}')\boldsymbol{\Delta}']\boldsymbol{\Sigma}^{-1}\}\\&= \sum_{t=1}^{T}(\tilde{\boldsymbol{y}}_t - \boldsymbol{\Delta}\boldsymbol{u}_{t|T})'\boldsymbol{\Sigma}^{-1}(\tilde{\boldsymbol{y}}_t - \boldsymbol{\Delta}\boldsymbol{u}_{t|T}) + tr\{\boldsymbol{\Delta}\sum_{t=1}^{T}\boldsymbol{P}_{t|T}\boldsymbol{\Delta}'\boldsymbol{\Sigma}^{-1}\}\\&= \sum_{t=1}^{T}(\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\boldsymbol{y}_{t-1} - \boldsymbol{\Delta}\boldsymbol{u}_{t|T})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}_t - \boldsymbol{\alpha} - \boldsymbol{\beta}\boldsymbol{y}_{t-1} - \boldsymbol{\Delta}\boldsymbol{u}_{t|T}) + tr\{\boldsymbol{\Delta}\sum_{t=1}^{T}\boldsymbol{P}_{t|T}\boldsymbol{\Delta}'\boldsymbol{\Sigma}^{-1}\}\end{aligned} \tag{4.4.10}$$

Again taking into account that $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are diagonal with k-th diagonal elements β_k and σ_k^2 respectively and writing $\boldsymbol{S} = \sum_{t=1}^{T}\boldsymbol{P}_{t|T}$ then (4.4.10) may be written in the form

$$\sum_{k=1}^{K}\{\sum_{t=1}^{T}(y_{tk} - \alpha_k - \beta_k y_{t,k-1} - \boldsymbol{\delta}_k'\boldsymbol{u}_{t|T})^2 + \boldsymbol{\delta}_k'\boldsymbol{S}\boldsymbol{\delta}_k\}/\sigma_k^2 \quad . \tag{4.4.11}$$

Since $\log(\det(\boldsymbol{\Sigma})) = \sum_{k=1}^{K}\log(\sigma_k^2)$ the terms involving the σ_k^2's in $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}_0)$ become

$$-\tfrac{1}{2}\sum_{k=1}^{K}[\{\sum_{t=1}^{T}(y_{tk} - \alpha_k - \beta_k y_{t,k-1} - \boldsymbol{\delta}_k'\boldsymbol{u}_{t|T})^2 + \boldsymbol{\delta}_k'\boldsymbol{S}\boldsymbol{\delta}_k\}/\sigma_k^2 + T\log(\sigma_k^2)] \quad . \tag{4.4.12}$$

We have now evaluated all the terms in $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}_0)$. Dropping those that do not depend on $\boldsymbol{\Theta}$ we may take the result as the sum of (4.4.8) and (4.4.12).

Now consider the M-step which calls for maximising $Q(\boldsymbol{\Theta},\boldsymbol{\Theta}_0)$ with respect to $\boldsymbol{\Theta}$. We start with ρ_m which is involved only in the m-th term of (4.4.8). We need to minimise the expression $(a_{mm} - 2\rho_m b_{mm} + \rho_m^2 c_{mm})/(1-\rho_m^2) + T\log(1-\rho_m^2)$ over $-1 < \rho_m < 1$. This cannot be done analytically but a (one-dimensional) Newton-Raphson search algorithm easily and reliably achieves this goal. Next α_k, β_k and $\boldsymbol{\delta}_k$ occur only in the k-th term in (4.4.12) and we need to minimise $\sum_{t=1}^{T}(y_{tk} - \alpha_k - \beta_k y_{t,k-1} - \boldsymbol{\delta}_k'\boldsymbol{u}_{t|T})^2 + \boldsymbol{\delta}_k'\boldsymbol{S}\boldsymbol{\delta}_k$ over arbitrary α_k, β_k with $-1 < \beta_k < 1$ and arbitrary $\boldsymbol{\delta}_k$. For given $\boldsymbol{\delta}_k$ the optimal choices of α_k and β_k can be found by ordinary least squares minimisation as follows.

Put

$$
\begin{aligned}
&x_{tk} = y_{t-1,k}\ ,\ \bar{y}_k = \textstyle\sum_{t=1}^{T} y_{tk}/T\ ,\ \bar{x}_k = \textstyle\sum_{t=1}^{T} x_{tk}/T\ ,\bar{\boldsymbol{u}} = \textstyle\sum_{t=1}^{T} \boldsymbol{u}_{t|T}/T \\
&\hat{b}_k(y) = \textstyle\sum_{t=1}^{T}(x_{tk}-\bar{x}_k)y_{tk} \Big/ \sum_{t=1}^{T}(x_{tk}-\bar{x}_k)^2 \\
&\hat{\boldsymbol{b}}_k(u) = \textstyle\sum_{t=1}^{T}(x_{tk}-\bar{x}_k)\boldsymbol{u}_{t|T} \Big/ \sum_{t=1}^{T}(x_{tk}-\bar{x}_k)^2 \\
&\hat{y}_{tk} = y_{tk} - \bar{y}_k - \hat{b}_k(y)(x_{tk}-\bar{x}_k) \\
&\hat{\boldsymbol{u}}_{tk} = \boldsymbol{u}_{t|T} - \bar{\boldsymbol{u}} - \hat{\boldsymbol{b}}_k(u)(x_{tk}-\bar{x}_k)
\end{aligned}
\tag{4.4.13}
$$

Then for given $\boldsymbol{\delta}_k$ the optimal choices of α_k and β_k are

$$
\begin{aligned}
\hat{\alpha}_k &= \bar{y} - \hat{\beta}_k(y)\bar{x} - \boldsymbol{\delta}_k'(\bar{\boldsymbol{u}} - \hat{\beta}_k(u)\bar{x}) \\
\hat{\beta}_k &= \hat{\beta}_k(y) - \boldsymbol{\delta}_k'\hat{\beta}_k(u)
\end{aligned}
\tag{4.4.14}
$$

and when these are substituted back we need to choose $\boldsymbol{\delta}_k$ to minimise

$$
\begin{aligned}
&\textstyle\sum_{t=1}^{T}(\hat{y}_{tk} - \boldsymbol{\delta}_k'\hat{\boldsymbol{u}}_{tk})^2 + \boldsymbol{\delta}_k'\boldsymbol{S}\boldsymbol{\delta}_k \\
&= \textstyle\sum_{t=1}^{T}\hat{y}_{tk}^2 - 2\boldsymbol{\delta}_k'\sum_{t=1}^{T}\hat{\boldsymbol{u}}_{tk}\hat{y}_{tk} + \boldsymbol{\delta}_k'(\sum_{t=1}^{T}\hat{\boldsymbol{u}}_{tk}\hat{\boldsymbol{u}}_{tk}' + \boldsymbol{S})\boldsymbol{\delta}_k
\end{aligned}
\tag{4.4.15}
$$

Again this has the analytic solution

$$
\hat{\boldsymbol{\delta}}_k = (\textstyle\sum_{t=1}^{T}\hat{\boldsymbol{u}}_{tk}\hat{\boldsymbol{u}}_{tk}' + \boldsymbol{S})^{-1}\sum_{t=1}^{T}\hat{\boldsymbol{u}}_{tk}\hat{y}_{tk} \quad . \tag{4.4.16}
$$

Hence we first calculate $\hat{\boldsymbol{\delta}}_k$ from (4.4.16) and substitute these into (4.4.14) to get the final minimising choices of $\boldsymbol{\delta}_k$, α_k and β_k. These can now be substituted back into (4.4.12) and the result minimised with respect to $\sigma_k^2 > 0$. This has the solution

$$
\hat{\sigma}_k^2 = \{\textstyle\sum_{t=1}^{T}(y_{tk} - \hat{\alpha}_k - \hat{\beta}_k y_{t,k-1} - \hat{\boldsymbol{\delta}}_k'\boldsymbol{u}_{t|T})^2 + \hat{\boldsymbol{\delta}}_k'\boldsymbol{S}\hat{\boldsymbol{\delta}}_k\}/T \quad . \tag{4.4.17}
$$

Thus except for the calculation of the optimising ρ_m (which requires only a one-dimensional search) the optimising values of all the other parameters can be found analytically.

The EM-algorithm to calculate the MLE's proceeds as follows. Start with an initial estimate $\hat{\boldsymbol{\Theta}}_0$ and with $\boldsymbol{\Theta}_0$ replaced by $\hat{\boldsymbol{\Theta}}_0$ carry out the M-step spelled out above and call the optimising values $\hat{\boldsymbol{\Theta}}_1$. Then redo the E- and M-step calculations but with $\boldsymbol{\Theta}_0$ replaced by $\hat{\boldsymbol{\Theta}}_1$ obtaining new optimising values $\hat{\boldsymbol{\Theta}}_2$. Carry on with repetitions until convergence is achieved, taking the final optimising values as the MLE's. On each iteration we need to recompute the quantities in (4.4.4) at the current $\boldsymbol{\Theta}_0$.

Calculation of $\boldsymbol{u}_{t|T}$ was already explained as part of the backward data pass in (4.3.10). For $\boldsymbol{P}_{t|T}$ and $\boldsymbol{P}_{t,t-1|T}$ (4.3.10) can be extended as in equations (A.12) and (A.13) of the Appendix of Wu *et al.* (1996). In our notation these equations are

$$\begin{aligned} \boldsymbol{P}_{t|T} &= \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1}\boldsymbol{V}_{t-1}\boldsymbol{P}_{t|t-1} \\ \boldsymbol{P}_{t,t-1|T} &= (\boldsymbol{I} - \boldsymbol{P}_{t|t-1}\boldsymbol{V}_{t-1})\boldsymbol{L}_{t-1}\boldsymbol{P}_{t-1|t-2} \end{aligned} \qquad (4.4.18)$$

We programmed the EM-algorithm in PROC IML of SAS and to verify the programming we applied the program to the two test cases of Section 4.3, starting at the same initial estimates. The values of the MLE's of the parameters produced by the EM algorithm differed from those presented in Tables 4.3.1 and 4.3.2 by less than 0.0005, depending on how strict the convergence criteria were set.

4.4.2 Application to the home loans transformed default rates

Here we fit the $AR(1)-U(1)$ model to the transformed default rates and as before we begin with the probit case. Columns 2 to 5 of Table 4.4.1 show the parameter estimates. For comparison purposes we repeated columns 2, 3 and 4 of Table 4.3.5 in columns 2, 5 and 6 of Table 4.4.1. These represent the initial estimates chosen by our initialisation strategy and the MLE's obtained from direct numerical maximisation as well as the standard errors of the MLE's. Column 3 of Table 4.4.1 shows the MLE's to which the EM-algorithm converged when started from the initial estimates in column 2 of this table. Comparing these EM-MLE's with the entries in column 5 they are seen to be similar (at least to two decimals and often better). We also started our direct numerical maximisation from these EM-MLE's and found that they quickly converged to the values shown in column 4 which are all very close to the entries of column 5. Our conclusion is that using the EM-algorithm on the $AR(1)-U(1)$ model essentially leads to the same results as obtained from the direct numerical maximisation. We also fitted the $AR(1)-U(1)$ model on the logit transforms and the results are similar. We do not include the corresponding table here.

Further we fitted the $AR(1)-U(2)$ model to the probit transformed default rates and Table 4.4.2 shows the results in the same format as above in Table 4.4.1. Again the EM-algorithm converged to MLE's which are close to those obtained by direct numerical maximisation with one exception, namely that the error variance estimate of risk class 2 is not so small. Thus overfitting appears to be less of a problem and this is consistent with the impression that the EM-approach yields more stable results. We also looked into estimation of the unobserved components, the fitted transformed default rates and the quality of fit. Figures 4.4.1a and Figure 4.4.1b compare the estimated unobserved components following from the EM-algorithm MLE estimates with those following from the direct numerical maximisation MLE's.

Table 4.4.1 MLE's of $AR(1)-U(1)$ models fitted to the probit transformed rates using the EM algorithm compared to direct MLE's

Parm	Init	EM(Init)	MLE(EM)	MLE(Init)	SE
α_1	-0.7829	-0.7831	-0.7831	-0.7832	0.0246
α_2	-1.6901	-1.6906	-1.6905	-1.6908	0.0345
α_3	-2.1162	-2.0962	-2.1494	-2.1502	0.0859
α_4	-2.7727	-2.7738	-2.7748	-2.7749	0.0337
α_5	-3.0334	-3.0633	-3.0347	-3.0351	0.0612
α_6	-3.3272	-3.3311	-3.3238	-3.3239	0.0310
α_7	-3.4026	-3.4036	-3.4053	-3.4054	0.0235
α_8	-3.4635	-3.4657	-3.4629	-3.4631	0.0324
α_9	-3.6418	-3.6374	-3.6502	-3.6504	0.0453
β_1	0.3614	0.3385	0.3378	0.3367	0.1247
β_2	0.5605	0.5744	0.5747	0.5747	0.1075
β_3	0.5791	0.6823	0.7129	0.7138	0.1164
β_4	0.3838	0.4827	0.4847	0.4852	0.1075
β_5	0.7648	0.7748	0.7774	0.7768	0.0774
β_6	0.3658	0.5464	0.5528	0.5524	0.0980
β_7	0.1053	0.3595	0.3686	0.3685	0.1058
β_8	0.2490	0.4740	0.4797	0.4795	0.1065
β_9	0.4597	0.6482	0.6570	0.6570	0.1024
σ_1^2	0.0148	0.0150	0.0150	0.0150	0.0029
σ_2^2	0.0119	0.0125	0.0125	0.0125	0.0024
σ_3^2	0.0208	0.0225	0.0221	0.0221	0.0046
σ_4^2	0.0116	0.0142	0.0141	0.0141	0.0030
σ_5^2	0.0092	0.0106	0.0105	0.0105	0.0022
σ_6^2	0.0065	0.0071	0.0068	0.0068	0.0017
σ_7^2	0.0077	0.0077	0.0073	0.0073	0.0020
σ_8^2	0.0119	0.0115	0.0112	0.0112	0.0026
σ_9^2	0.0114	0.0108	0.0106	0.0106	0.0023
δ_1	0.0183	0.0196	0.0212	0.0213	0.0181
δ_2	0.0241	0.0310	0.0335	0.0335	0.0170
δ_3	0.0555	0.0943	0.1043	0.1043	0.0262
δ_4	0.0547	0.0840	0.0911	0.0911	0.0205
δ_5	0.0480	0.0766	0.0832	0.0832	0.0176
δ_6	0.0533	0.0943	0.1031	0.1031	0.0173
δ_7	0.0557	0.1030	0.1131	0.1131	0.0189
δ_8	0.0562	0.1049	0.1146	0.1147	0.0211
δ_9	0.0476	0.0901	0.0985	0.0985	0.0197
ρ_1	-0.2766	-0.4030	-0.4382	-0.4385	0.1408

Table 4.4.2 MLE's of $AR(1)-U(2)$ models fitted to the probit transformed rates using the EM algorithm compared to direct MLE's

Parm	Init	EM(Init)	MLE(EM)	MLE(Init)	SE
α_1	-0.7829	-0.7822	-0.7887	-0.7887	0.0354
α_2	-1.6902	-1.6916	-1.7076	-1.7076	0.0419
α_3	-2.1039	-2.0962	-2.5200	-2.5200	0.1405
α_4	-2.7725	-2.7737	-2.7818	-2.7818	0.0350
α_5	-3.0522	-3.0610	-3.0467	-3.0467	0.0588
α_6	-3.3282	-3.3304	-3.3301	-3.3301	0.0284
α_7	-3.4026	-3.4030	-3.4085	-3.4085	0.0223
α_8	-3.4637	-3.4650	-3.4666	-3.4666	0.0308
α_9	-3.6402	-3.6366	-3.6604	-3.6604	0.0474
β_1	0.3606	0.5686	0.5790	0.5790	0.1085
β_2	0.5564	0.6968	0.7416	0.7416	0.0756
β_3	0.5711	0.7773	0.9681	0.9681	0.0431
β_4	0.3742	0.5077	0.5142	0.5142	0.1062
β_5	0.7261	0.7579	0.7694	0.7694	0.0761
β_6	0.3641	0.4725	0.5137	0.5137	0.0964
β_7	0.1055	0.2834	0.3327	0.3327	0.1038
β_8	0.2490	0.4013	0.4570	0.4570	0.0993
β_9	0.4568	0.6609	0.6668	0.6668	0.1060
σ_1^2	0.0065	0.0085	0.0086	0.0086	0.0018
σ_2^2	0.0024	0.0033	0.0001	0.0001	0.0016
σ_3^2	0.0159	0.0160	0.0141	0.0141	0.0031
σ_4^2	0.0105	0.0136	0.0138	0.0138	0.0028
σ_5^2	0.0093	0.0110	0.0107	0.0107	0.0022
σ_6^2	0.0055	0.0067	0.0065	0.0065	0.0016
σ_7^2	0.0054	0.0060	0.0066	0.0066	0.0017
σ_8^2	0.0058	0.0071	0.0077	0.0077	0.0021
σ_9^2	0.0114	0.0105	0.0109	0.0109	0.0023
δ_{11}	0.0183	0.0541	0.0428	0.0428	0.0315
δ_{21}	0.0241	0.0739	0.0719	0.0720	0.0357
δ_{31}	0.0555	0.1248	0.1457	0.1457	0.0286
δ_{41}	0.0547	0.0867	0.0934	0.0934	0.0203
δ_{51}	0.0480	0.0645	0.0772	0.0772	0.0200
δ_{61}	0.0533	0.0734	0.0919	0.0919	0.0237
δ_{71}	0.0557	0.0806	0.1011	0.1011	0.0265
δ_{81}	0.0562	0.0705	0.0970	0.0970	0.0356
δ_{91}	0.0476	0.0860	0.0951	0.0951	0.0211

Table 4.4.2 (contd) MLE's of $AR(1)-U(2)$ models fitted to the probit transformed rates using the EM algorithm compared to direct MLE's

Parm	Init	EM(Init)	MLE(EM)	MLE(Init)	SE
δ_{12}	0.0662	0.0598	0.0764	0.0764	0.0209
δ_{22}	0.0708	0.0586	0.0947	0.0947	0.0258
δ_{32}	0.0523	0.0104	0.0445	0.0445	0.0474
δ_{42}	0.0255	-0.0224	-0.0042	-0.0042	0.0338
δ_{52}	-0.0052	-0.0434	-0.0279	-0.0279	0.0281
δ_{62}	-0.0246	-0.0667	-0.0468	-0.0468	0.0315
δ_{72}	-0.0359	-0.0798	-0.0542	-0.0543	0.0349
δ_{82}	-0.0570	-0.1047	-0.0834	-0.0834	0.0343
δ_{92}	-0.0087	-0.0407	-0.0215	-0.0215	0.0326
ρ_1	-0.2766	-0.4097	-0.4634	-0.4634	0.1306
ρ_2	0.0223	-0.1014	-0.1653	-0.1653	0.1447

They are clearly very similar suggesting that the fitted default rates and the quality of fit testing based on these EM-MLE's will be quite similar to those reported in Paragraph 4.3.7. We found that this is indeed the case and there is no need to elaborate further on the results.

Although little new inference features are found from this EM exercise, it was still useful in that a quite different approach gave similar results as the direct optimisation MLE's, thus strengthening the reliability of the MLE results.

4.5 Application to default rate forecasting

Fitting statistical models to default rate data helps the user to understand the dynamics of the default processes. This was demonstrated in the results reported in Chapter 3 and the sections of Chapter 4 above. The fitted models can also be used to forecast future default rates. This is important since it enables anticipation and appropriate management of credit risk processes. In this section we discuss forecasting using the $AR(1)-U(M)$ models and apply the methodology to the home loans data.

Recall that the model equations (4.2.3) are

$$\boldsymbol{Y}_t = \boldsymbol{\alpha} + \boldsymbol{\beta}(\boldsymbol{Y}_{t-1} - \boldsymbol{\alpha}) + \boldsymbol{\Delta}\boldsymbol{U}_t + \boldsymbol{e}_t \quad \text{and} \quad \boldsymbol{U}_t = \boldsymbol{\rho}\boldsymbol{U}_{t-1} + \boldsymbol{D}\boldsymbol{\eta}_t \tag{4.5.1}$$

and assume that we have data available at times $t = 1,2,\ldots,T$ which was used to fit the model, yielding parameter estimates $\hat{\boldsymbol{\alpha}}$, $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\Delta}}$, $\hat{\boldsymbol{\rho}}$ and $\hat{\boldsymbol{D}}$ as well as estimated unobserved components $\boldsymbol{u}_{t|T}$. We wish to forecast $\boldsymbol{Y}_{T+h}$ for $h = 1,2,\ldots,H$ and carry out the following simulation approach.

Figure 4.4.1a Comparison of estimated first unobserved component according to $AR(1)-U(2)$ model fitted by the EM algorithm (UtT1_EM) and the direct optimisation MLE (UtT1) to probit default rates

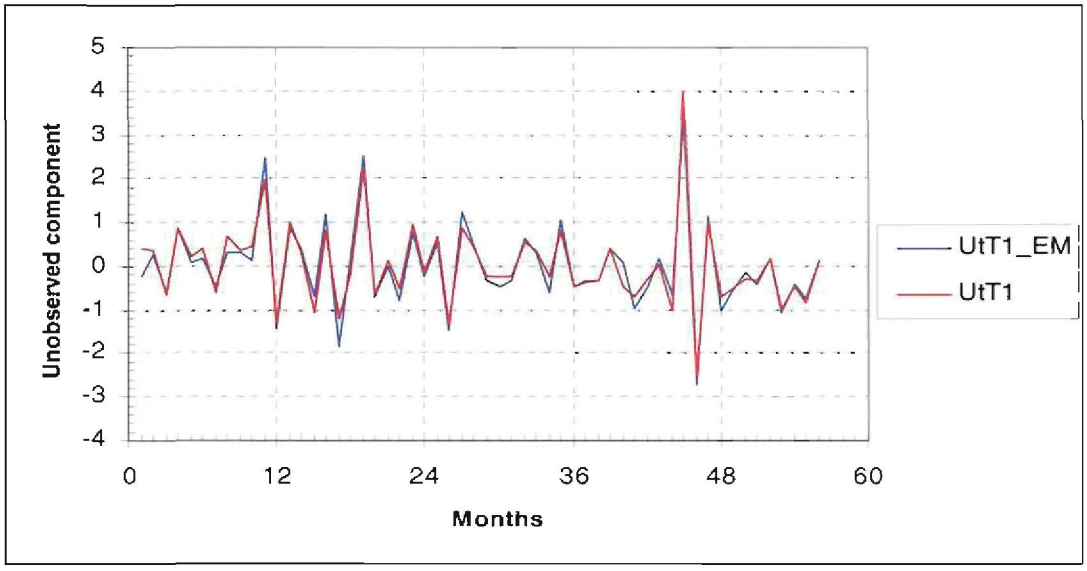
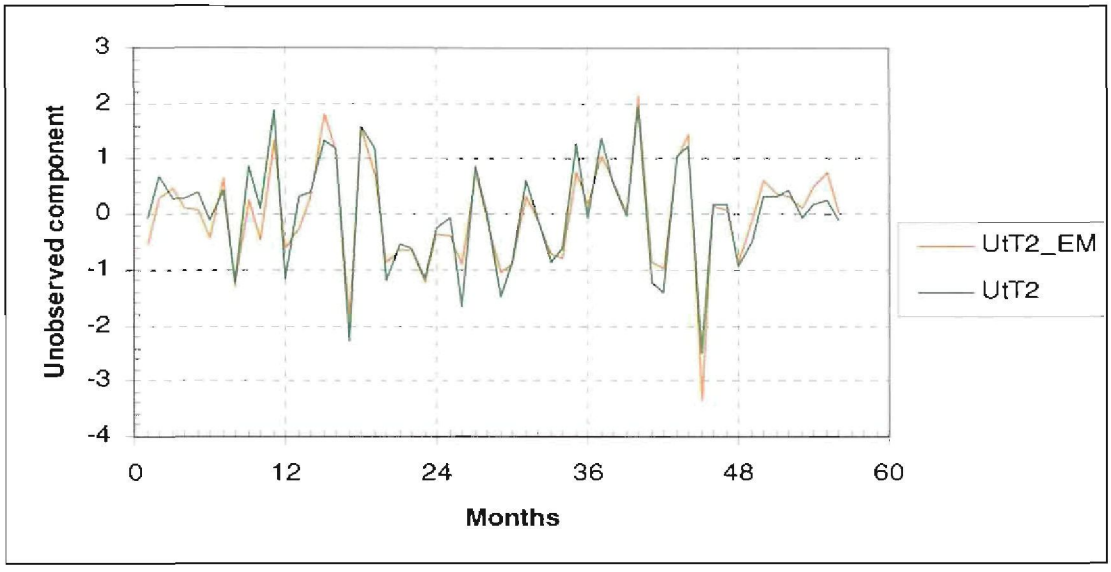


Figure 4.4.1b Comparison of estimated second unobserved component according to $AR(1)-U(2)$ model fitted by the EM algorithm (UtT2_EM) and the direct optimisation MLE (UtT2) to probit default rates



1. From (4.5.1) we have $\boldsymbol{U}_{T+1} = \boldsymbol{\rho}\boldsymbol{U}_T + \boldsymbol{D}\boldsymbol{\eta}_T$ and generate L one step ahead simulation based forecast values of $\boldsymbol{U}_{T+1}$ by generating $\boldsymbol{\eta}_{T+1}(l)$ from the $N_M(\boldsymbol{0},\boldsymbol{I})$ distribution and calculating $\hat{\boldsymbol{U}}_{T+1}(l) = \hat{\boldsymbol{\rho}}\boldsymbol{u}_{T|T} + \hat{\boldsymbol{D}}\boldsymbol{\eta}_{T+1}(l)$ for $l = 1,2,\ldots,L$.

2. Generate $\boldsymbol{Z}_{T+1}(l)$ from the $N_K(\boldsymbol{0},\boldsymbol{I})$ distribution, set $\boldsymbol{e}_{T+1}(l) = \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{Z}_{T+1}(l)$ and $\hat{\boldsymbol{Y}}_{T+1}(l) = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}(\boldsymbol{Y}_T - \hat{\boldsymbol{\alpha}}) + \hat{\boldsymbol{\Delta}}\hat{\boldsymbol{U}}_{T+1}(l) + \hat{\boldsymbol{e}}_{T+1}(l)$ for $l = 1,2,\ldots,L$ to obtain L one step ahead simulation based forecast values of $\boldsymbol{Y}_{T+1}$.

3. Similarly, for $h > 1$ generate $\boldsymbol{\eta}_{T+h}(l)$ from the $N_M(\boldsymbol{0},\boldsymbol{I})$ distribution, $\boldsymbol{Z}_{T+h}(l)$ from the $N_K(\boldsymbol{0},\boldsymbol{I})$ distribution, set $\hat{\boldsymbol{U}}_{T+h}(l) = \hat{\boldsymbol{\rho}}\hat{\boldsymbol{U}}_{T+h-1}(l) + \hat{\boldsymbol{D}}\boldsymbol{\eta}_{T+h}(l)$, $\boldsymbol{e}_{T+h}(l) = \hat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\boldsymbol{Z}_{T+h}(l)$ and $\hat{\boldsymbol{Y}}_{T+h}(l) = \hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{Y}}_{T+h-1}(l) - \hat{\boldsymbol{\alpha}}) + \hat{\boldsymbol{\Delta}}\hat{\boldsymbol{U}}_{T+h}(l) + \hat{\boldsymbol{e}}_{T+h}(l)$ for $l = 1,2,\ldots,L$ to obtain L h steps ahead simulation based forecast values of $\boldsymbol{Y}_{T+h}$.

4. Then the average $\bar{\boldsymbol{Y}}_{T+h} = \sum_{l=1}^{L}\hat{\boldsymbol{Y}}_{T+h}(l)/L$ serves as our point forecast of $\boldsymbol{Y}_{T+h}$. Also the diagonals of $\boldsymbol{C}_{T+h} = \sum_{l=1}^{L}[\hat{\boldsymbol{Y}}_{T+h}(l) - \bar{\boldsymbol{Y}}_{T+h}][\hat{\boldsymbol{Y}}_{T+h}(l) - \bar{\boldsymbol{Y}}_{T+h}]'/L$ estimate the variances of the components of $\boldsymbol{Y}_{T+h}$ and their square roots estimate the standard deviations. Confidence intervals can be formed by adding and subtracting multiples of the corresponding standard deviations from the components of averages $\bar{\boldsymbol{Y}}_{T+h}$.

5. Transform the average $\bar{\boldsymbol{Y}}_{T+h}$ back to default rates by taking the inverses of the probit or logistic transforms of the components of $\bar{\boldsymbol{Y}}_{T+h}$ respectively. The upper and lower bounds of the confidence intervals can also be transformed back to get corresponding confidence intervals for the forecasted (untransformed) default rates.

We programmed these steps and applied them to the home loans default rates. We used the data of the first 50 months (i.e. now $T = 50$) to fit the model and then forecasted 6 months ahead (i.e. $H = 6$). The forecasts can be compared to the actual observed data over months 51 to 56. Using $L = 10000$ repetitions in the simulation process we obtained the results below. To begin with we fitted the $AR(1)-U(1)$ model on the probit transformed default rates.

Figure 4.5.1a shows the following for risk classes 1, 4 and 7: The observed probits over the whole period, the fitted values based on the data of the first 50 months, the forecasted values over the last 6 months and one standard deviation confidence bounds. In preparing these graphs the observed data is represented by connecting the individual monthly points

by straight lines, using one colour (e.g. green is used for class 4 in Fig.4.5.1a). Further, using a second colour (e.g. orange used for class 4 in Fig.4.5.1a), the fitted and forecasted values as well as the confidence bounds are represented by connecting the individual monthly points by straight lines. Note that connecting the monthly points is simply a graphical way of making the interpretation of the graphs easier and in particular, does not mean that we have observations, forecasts or confidence bounds for the times between months.

Figure 4.5.1b shows the same plots for classes 2, 5 and 8 and Figure 4.5.1c shows them for classes 3, 6 and 9. The non-adjacent groupings and one standard deviation confidence bounds were used to separate the plots enough to make interpretation easier. Looking at Figure 4.5.1a we note that the observed probits of class1 stayed within the one standard deviation interval over the forecast period, those of class 4 dropped below the lower bound (although it would be close to the two standard deviation lower bound) and those of class 7 followed the one standard deviation lower bound. Similar remarks apply to Figures 4.5.1b and 4.5.1c. There are in total 9x6=54 forecasts in these 3 figures and 35 of them remain inside the one standard deviation intervals; this is 65% which is fairly consistent with what one would expect from normal one standard deviation intervals (keeping in mind that the inside-outside events are not independent here). Also 49 of them remain inside the two standard deviation intervals, which is 91% and is also fairly consistent with normal two standard deviation intervals.

We transformed the forecasts and the two-standard deviation intervals back to default rates and Table 4.5.1 shows the results for all risk classes and the six time horizons. For each risk class, a block of 4 rows is shown consisting of the actual observed default rate (identified by "dr"), the lower bound of the interval ("lo"), the forecasted default rate ("for") and the upper bound of the interval ("up"). Note that the intervals are not symmetric around the forecasted values. We also did the above analysis for the logits and transformed them back to default rates. We found that the equivalent logit table has entries very close to Table 4.5.1 and therefore did not include that table. Again the conclusion is that it makes little difference whether the probit or logit transformation is used.

We further fitted and forecasted the home loans data using the $AR(1)-U(2)$ model on the probit and logit transformed default rates. Figures 4.5.2a, b and c show the results for the probit transforms in the same form as Figures 4.5.1a, b and c above. The two sets of forecasts appear quite similar and as far as can be judged from these graphs it seems that little benefit was derived from using the more complex $AR(1)-U(2)$ model. We also prepared the equivalent of Table 4.5.1 but using the $AR(1)-U(2)$ model. Again the results

Figure 4.5.1a Comparison of actual, fitted and forecasted probit default rates for risk classes 1,4 and 7 using the $AR(1)-U(1)$ model

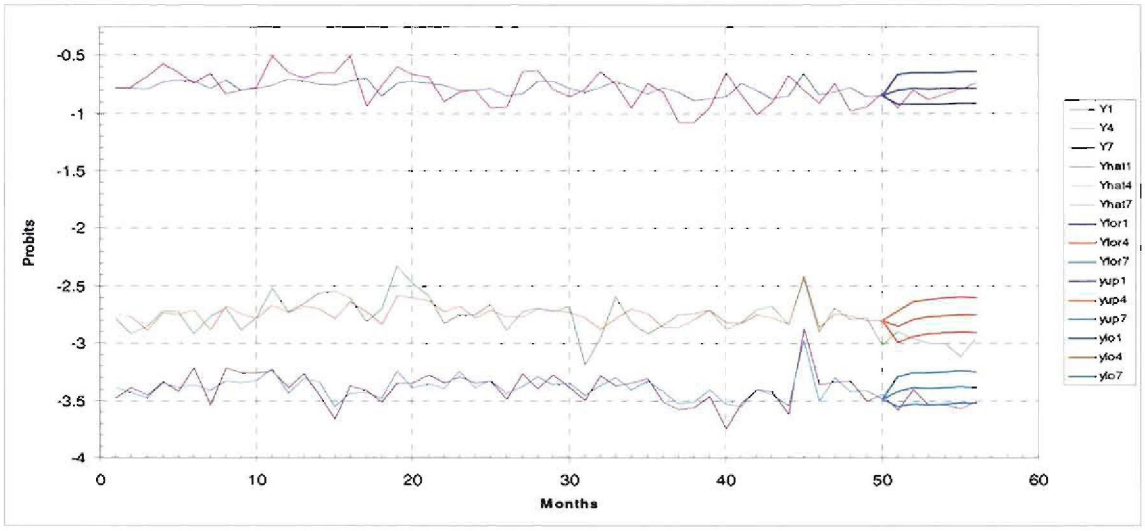


Figure 4.5.1b Comparison of actual, fitted and forecasted probit default rates for risk classes 2,5 and 8 using the $AR(1)-U(1)$ model

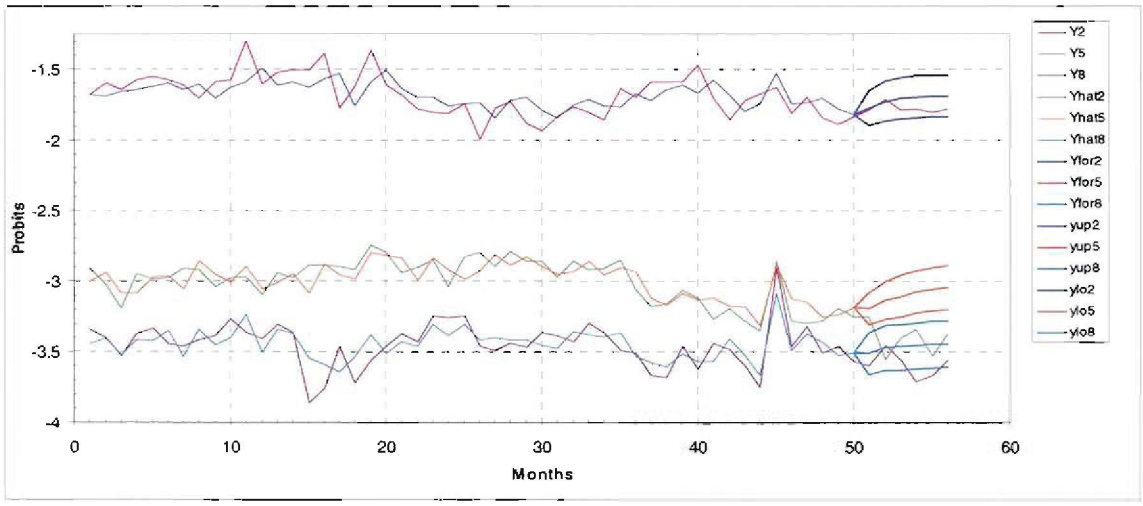


Figure 4.5.1c Comparison of actual, fitted and forecasted probit default rates for risk classes 3,7 and 9 using the $AR(1)-U(1)$ model

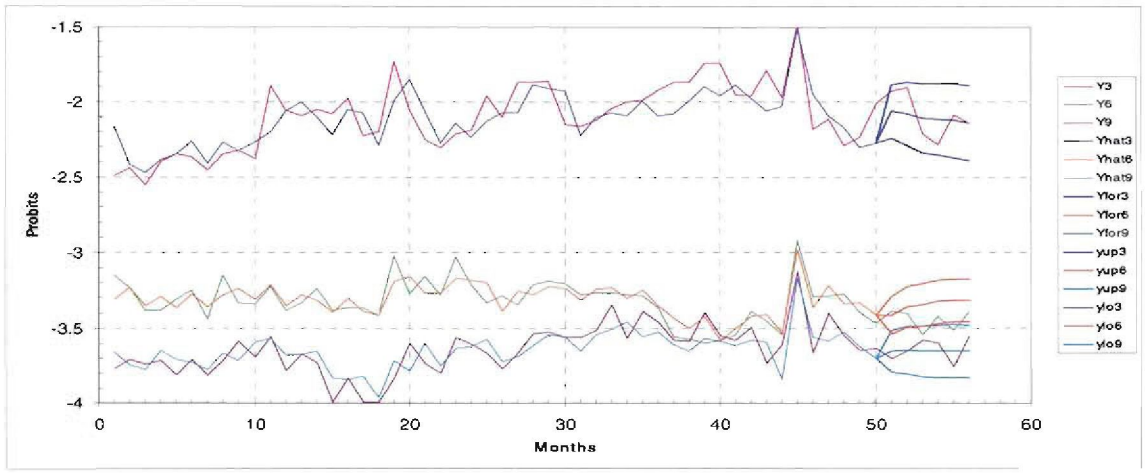


Table 4.5.1 Forecasted default rates and confidence intervals compared to observed default rates based on probit transforms

Class		Forecast horizons					
	h	1	2	3	4	5	6
1	dr	0.170259	0.213768	0.190192	0.203059	0.215023	0.230232
	lo	0.148601	0.146815	0.147570	0.146983	0.147144	0.146816
	for	0.214807	0.216608	0.217362	0.216653	0.216820	0.216667
	up	0.300113	0.307618	0.308226	0.307444	0.307590	0.307764
2	dr	0.036749	0.043269	0.036887	0.037320	0.035921	0.037766
	lo	0.023136	0.023129	0.023817	0.024184	0.024470	0.024508
	for	0.038672	0.041174	0.043331	0.044289	0.044969	0.045414
	up	0.063958	0.072257	0.077564	0.079743	0.081209	0.082645
3	dr	0.027080	0.028401	0.013428	0.011089	0.018340	0.016182
	lo	0.009182	0.006444	0.004942	0.003984	0.003310	0.002804
	for	0.022151	0.020039	0.019300	0.018165	0.017323	0.016459
	up	0.052468	0.060568	0.072334	0.078824	0.085566	0.090567
4	dr	0.001869	0.001568	0.001369	0.001359	0.000935	0.001547
	lo	0.000949	0.001034	0.001122	0.001152	0.001173	0.001162
	for	0.002267	0.002590	0.002844	0.002874	0.002929	0.002923
	up	0.005409	0.006471	0.007192	0.007154	0.007292	0.007336
5	dr	0.000583	0.000193	0.000331	0.000421	0.000209	0.000379
	lo	0.000354	0.000353	0.000368	0.000386	0.000401	0.000413
	for	0.000765	0.000858	0.000972	0.001035	0.001103	0.001152
	up	0.001652	0.002080	0.002561	0.002770	0.003032	0.003204
6	dr	0.000350	0.000338	0.000197	0.000313	0.000223	0.000347
	lo	0.000145	0.000144	0.000149	0.000153	0.000160	0.000159
	for	0.000351	0.000374	0.000410	0.000421	0.000435	0.000438
	up	0.000850	0.000969	0.001129	0.001158	0.001178	0.001205
7	dr	0.000167	0.000331	0.000203	0.000203	0.000180	0.000224
	lo	0.000133	0.000122	0.000123	0.000123	0.000125	0.000124
	for	0.000348	0.000327	0.000341	0.000335	0.000340	0.000340
	up	0.000911	0.000877	0.000945	0.000915	0.000923	0.000934
8	dr	0.000158	0.000278	0.000187	0.000103	0.000122	0.000183
	lo	0.000082	0.000077	0.000079	0.000080	0.000082	0.000082
	for	0.000244	0.000249	0.000267	0.000266	0.000270	0.000273
	up	0.000730	0.000801	0.000897	0.000881	0.000888	0.000906
9	dr	0.000107	0.000130	0.000171	0.000161	0.000087	0.000191
	lo	0.000049	0.000038	0.000034	0.000032	0.000031	0.000030
	for	0.000144	0.000131	0.000129	0.000125	0.000124	0.000122
	up	0.000417	0.000443	0.000487	0.000488	0.000495	0.000491

Figure 4.5.2a Comparison of actual, fitted and forecasted probit default rates for risk classes 1,4 and 7 using the $AR(1)-U(2)$ model

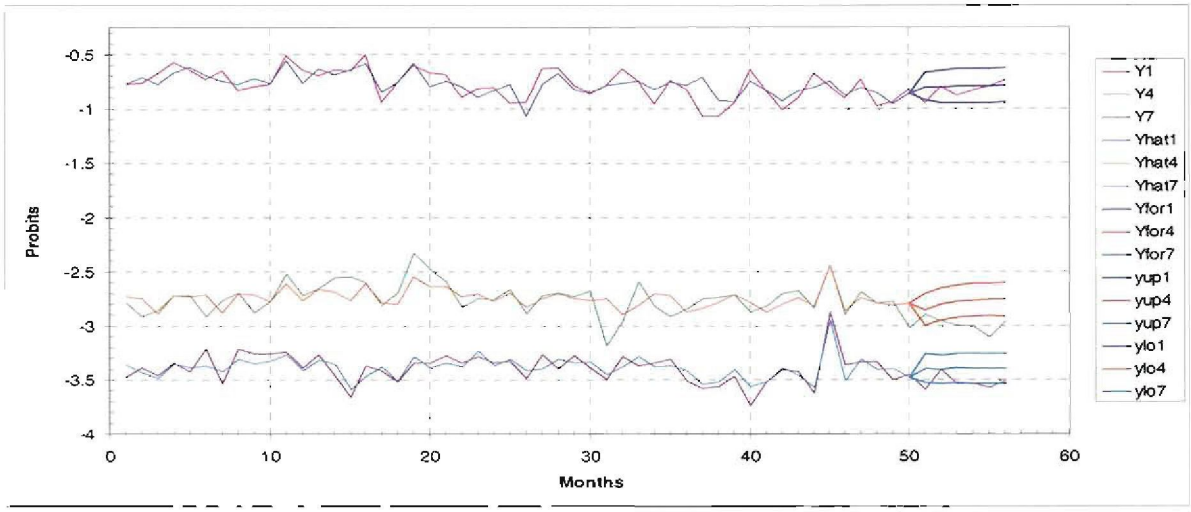


Figure 4.5.2b Comparison of actual, fitted and forecasted probit default rates for risk classes 2,5 and 8 using the $AR(1)-U(2)$ model

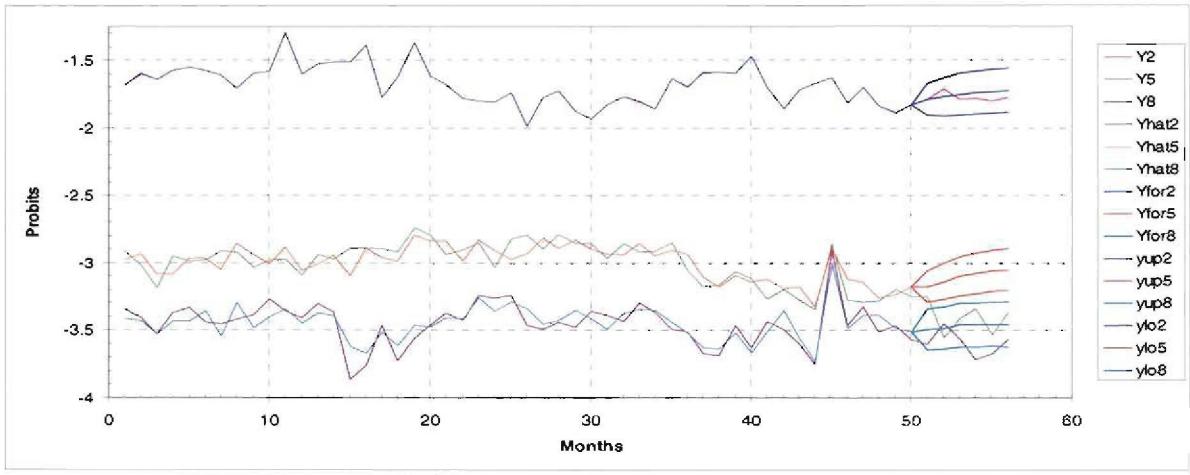
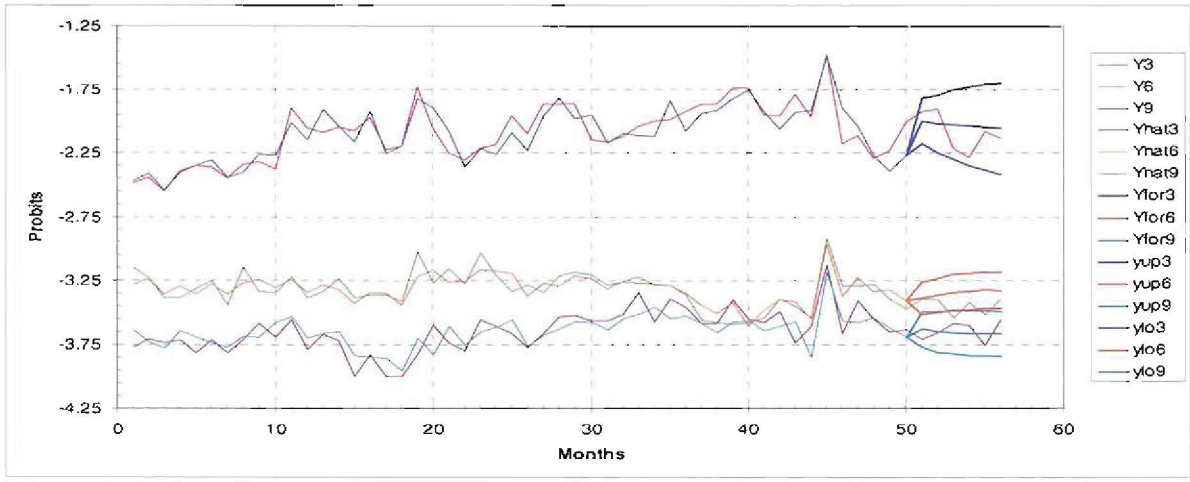


Figure 4.5.2c Comparison of actual, fitted and forecasted probit default rates for risk classes 3,7 and 9 using the $AR(1)-U(2)$ model



are comparable to those in Table 4.5.1. The confidence intervals are slightly shorter for risk classes 4 to 9 when using the $AR(1)-U(2)$ model but somewhat longer for classes 1 to 3 and the empirical coverage rates are the same. Again using the $AR(1)-U(2)$ does not seem to yield much benefit over the $AR(1)-U(1)$ model.

4.6 Summary

This chapter contains the main contributions of this dissertation, namely the formulation of AR multivariate time series models with unobserved components. We also developed the maximum likelihood inference methodology for this type of model to some extent when the assumptions of normally distributed error terms and unobserved components apply. The process was illustrated by application to simulation data as well as the home loans default data.

Some issues remain for further investigation. For the home loans data it turned out that the normality assumption is rejected for some of the risk classes and some of the unobserved components when formally tested. For example Table 4.6.1 shows the P-values of the Jarque-Bera normality test on the estimated probit residuals and unobserved components of the $AR(1)-U(2)$ model fitted by direct maximum likelihood both without and with equal variance restrictions and also by the EM algorithm.

Table 4.6.1 P-values of the Jarque-Bera normality test on probit residuals and unobserved components of the $AR(1)-U(2)$ model

Class	MLE (equal var)	MLE (unrestricted)	MLE (EM)
pdr1	0.0089	<0.0001	<0.0001
pdr2	0.6526	0.9638	0.8906
pdr3	0.9740	0.5766	0.4980
pdr4	<0.0001	<0.0001	<0.0001
pdr5	0.1029	0.0555	0.0622
pdr6	0.8117	0.7503	0.5831
pdr7	0.4673	0.5661	0.5978
pdr8	0.2601	0.2245	0.2241
pdr9	0.5598	0.1526	0.4672
Unob Comps	**MLE (equal var)**	**MLE (unrestricted)**	**MLE (EM)**
1	<0.0001	<0.0001	0.0002
2	0.5748	0.9985	0.0174

In the cases of risk classes 1 and 4, normality is rejected for all three fits and this also happens for the first unobserved component. This is not altogether unexpected since it is often found in an economic modelling context that the normality assumption is not adequate

for the distributions of error terms and random factors. When this is the case, MLE's based on the normality assumption are often described as Quasi-MLE's and may still be valid in the sense of being consistent and asymptotically normally distributed, but not necessarily efficient in the sense of having the smallest possible variance.

Of course it may also be that there are outliers or data errors which cause the rejection of the normality assumption. For example, if the extreme estimated unobserved component of month 45 (see Figure 4.3.6) is left out when testing for normality the P-values improve substantially suggesting that the data of month 45 is suspect. This was the case with the data right from the beginning as an inspection of Figures 2.2.2 and 2.2.3 shows. Although the fitted model did accommodate the extreme movements in the default rates at month 45 this was at the cost of a non-normal fluctuation in the first unobserved component. It is an open issue that requires further research to study the effects of the lack of normality of the error terms and unobserved components on the MLE's in the context of the $AR(1)-U(M)$ models used here. One approach would be to replace the normality assumption by more general distributional assumptions, e.g. the Normal Inverse Gaussian (NIG) distribution assumption (Venter *et al.,* 2006), but this is beyond the scope of this dissertation.

An important question is how many unobserved components to use. We treated this issue rather informally in our exposition above. We stopped at two unobserved components since this already entails estimating 47 parameters. Each additional unobserved component added to the model would require another $K+1$ (=10 for the home loans data) parameters and this becomes rather demanding compared to our relatively short data series. It would be preferable to use model selection criteria such as the AIC or SBC to determine the number of unobserved components. Of course there may also be reason to question restricting attention to only $AR(1)$ choices in the $AR(1)-U(M)$ combination. More generally one may wish to consider $ARMA(p,q)-U(M)$ models also. The model selection issue then becomes even more serious but we leave these matters for future research.

A further issue not discussed here is to explain the unobserved component in economically meaningful terms. For example one could try to establish a relation between observed economic factors (e.g. CPIX, insolvency rates and bank acceptances rate or other variables) and the unobserved components as estimated from the data. This would help the user to understand the underlying economic dynamics of the processes better.

Thus although progress is reported and software enabling maximum likelihood inference has been developed here, further work is required on the issues listed above.

Chapter 5

Concluding remarks

The importance of proper credit risk management is highlighted by the so-called sub-prime crisis that is currently causing world-wide turmoil on the financial markets. The seriousness of this crisis is illustrated by the following news item taken from Reuters: *"Charles Prince resigned on Sunday as chairman and chief executive of Citigroup Inc. as the bank said it may write off $11 billion of subprime mortgage losses, on top of a $6.5 billion write-down last quarter".* (Stempel & Wilchins, 2007). Evidently, granting credit and managing credit portfolios are extremely important aspects of proper bank management and government. It is equally important to model the default rate dynamics of credit portfolios and to forecast future default tendencies in order to take appropriate action timeously.

The aims of this dissertation were to make contributions to the literature on modelling and forecasting credit default rates. Our effort focused on describing simultaneous default tendencies of the obligors in the different risk classes forming a credit portfolio.

Our main contribution centres around the use of auto-regressive models extended to include unobserved or latent components. We developed a methodology to fit such models and discussed its application to the home loans portfolio. We showed how default rates can be forecasted using these models. As discussed at the end of Chapter 4, some open issues regarding the application of these models still require further research. Nevertheless, the contributions in this dissertation should form a useful starting point for managers wishing to model and forecast default rates of their credit portfolios.

Bibliography

BANGIA, A., DIEBOLD, F.X., KRONIMUS, A., SCHAGEN, C. & SCHUERMANN, T. 2002. Ratings migration and the business cycle, with application to credit portfolio stress testing. *Journal of banking & finance*, 26:445-474. Available: Elsevier.

BASEL COMMITTEE ON BANKING SUPERVISION. 2006. International convergence of capital measurement and capital standards: A revised framework (comprehensive version). http://www.bis.org/publ/bcbs128.pdf

BUCAY, N. & ROSEN, D. 2000. Applying portfolio credit risk models to retail portfolios. *Algo research quarterly*, 3(1):45-73, Mar.

DAVISON, A.C. & HINKLEY, D.V. 1997. Bootstrap methods and their application. Cambridge: University of Cambridge Press. 582 p.

DEMPSTER, A.P., LAIRD, N.M. & RUBIN, D.B. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the. Royal Statistical Society*, Series B(39): 1-39.

DURBIN, J. 1970. Testing for serial correlation in least-squares regression when some of the regressors are lagged dependent variables. *Econometrica*, 38:410-421.

HARVEY, A.C. 1989. Forecasting, structural time series models and the Kalman filter. Cambridge: University of Cambridge Press. 554 p.

JARQUE, C.M. & BERA, A.K. 1980. Efficient tests for normality, homoskedasticity and serial independence of regression residuals. *Economics letters*, 6:255-259.

KOCH, T.W. & MACDONALD, S.S. 2003. Bank Management. 5th ed. Mason, Oh: Thomson South-Western. 888p.

LIU, C., RUBIN, D.B. & WU, Y.N. 1998. Parameter expansion to accelerate the EM: The PX-EM Algorithm. *Biometrika*, 85(4):755-770, Dec. Available: JSTOR.

MBOWENI, T.T. 2007. Address by Mr T.T. Mboweni, Governor of the South African Reserve Bank at the National Consumer Forum Conference to celebrate World Consumer Rights Day, Johannesburg, 15 March 2007.
http://www.reservebank.co.za /Addresses by Governors/Addresses by Governors/2007-03-15 Date of access: 16 May 2007.

Bibliography (contd)

McCULLAGH, P. & NELDER, J.A. 1989. Generalized Linear Models. 2nd ed. London: Chapman & Hall. 511 p.

McNEIL, A.J. & WENDIN, J.P. 2007. Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of empirical finance*, 14(2007):131-149. Available: ScienceDirect.

METAXOGLOU, K. & SMITH, A. 2007. Maximum likelihood estimation of VARMA models using a state-space EM algorithm. *Journal of time series analysis*, 28(5):666-685.

NICKELL, P., PERRAUDIN, W. & VAROTTO, S. 2000. Stability of rating transitions. *Journal of banking & finance*, 24:203-227. Available: Elsevier.

SAS INSTITUTE Inc. 2004. SAS/ETS 9.1 User's Guide. Cary, NC: SAS Publishing. http://support.sas.com/documentation/onlinedoc/91pdf/index_91.html#ets Date of access: 14 Nov. 2007.

SAS INSTITUTE Inc. SAS 9.1.3 Help and documentation.

STEMPEL, J. & WILCHINS, D. 2007. Citigroup may face $11 billion writeoff. *Reuters*: 5 Nov. http://www.reuters.com/article/newsOne/idUSWEN234820071105 Date of access: 7 Nov. 2007

SCHÖNBUCHER, P.J. 2005. Credit derivatives pricing models: Models, pricing and implementation. Chichester: Wiley. 375 p.

SOUTH AFRICAN RESERVE BANK. 2007. DI500 data (February 2007). www.reservebank.co.za/SARB activities/Bank Supervision/Total Banks data/DI500/Febr 2007 Date of access: 15 May 2007.

VENTER, J.H., DE JONGH, P.J. & GRIEBENOW, G. 2006. GARCH-type volatility models based Brownian inverse Gaussian intra-day return processes. *The Journal of Risk*, 8(4): 97-116.

WU, L.S., PAI, J.S. & HOSKING, J.R.M. 1996. An algorithm for estimating parameters of state-space models. *Statistics & Probability letters*, 28(1996): 99-106. Available: Elsevier.