Assessing the shape of the distribution of interest rates : lessons from French individual data

Renaud Lacroix*

Banque de France

This version : January 2007

Abstract :

This paper considers the problem of fitting a finite Gaussian mixture with an unknown number of components to interest rates provided by a sample of credit institutions. Indeed, mixture models are able to represent arbitrarily complex probability density functions distributions, and hence outperform the usual non parametric kernel density estimates in several ways. Therefore, this tool enables us to improve our understanding of the credit market. The data are extracted from individual contracts and gathered on a quarterly basis. They provide a loan-by-loan disaggregation of average rates at bank level available in the monthly MIR reports. Loans to non-financial firms and to households are splitted up into eleven categories defined by instrument and maturity of loans. Each category is analyzed separately. Two major econometric issues are dealt with : how to estimate the number of components, and how to estimate the parameters defining the model. Several techniques are compared, which allows us to show how the results may be sensitive to the methodology used, especially when sample survey issues are properly taken into account : for instance, the model is enhanced in order to include discrete components which arise from clusters in the distribution. In a second step, this set-up allows us to estimate precisely the modes of the distribution. Then, these modes are interpreted as resulting from the segmentation of the credit market. We illustrate this topic through a kind of scoring of credit institutions based on ex-post probabilities for loans to be in the neighborhood of a particular mode. Special attention is also paid to the upper side of the distribution, where the usury law may affect the level of interest rates. The effect of the usury threshold on the distribution is first analyzed from a theoretical perspective, through simulations using loans which are not subject to the usury law. Then, for categories defined by the usury law, we provide a rough estimate of the proportion of loans which are not granted because of the threshold effect. Finally, introducing the time series dimension shows that the the shape of the distributions (number and amplitude of modes) is generally not constant over time. This instability raises another interesting issue, especially for short term analysis purposes. This is left for future research.

*Monetary Studies and Statistics Directorate, Statistical Engineering and Coordination Division. I am grateful to Olivier Cousseran, Daniel Gabrielli, Emmanuel Gervais and Hervé Le Bihan for their helpful comments on previous drafts of this paper, and Frédérique Edmond and Bertille Mader for making all the data easy available to me. The views expressed in this paper are those of the author and do not necessarily reflect those of the Banque de France.

1 Introduction

Since January 2003, MFI interest rate (MIR) statistics provide a comprehensive picture of main deposit and loan rates vis-à-vis households and non-financial corporations (NFC) in the euro area. Each month, the BdF collects aggregated data relative to rates and amounts of new transactions from a sample of credit institutions[1]. After aggregation at the country level, one gets useful macroeconomic informations, especially for monetary policy purpose. Indeed, these statistics allow to analyse the extent and the speed of the pass-through of market rates to lending rates faced by households and NFC[2]. Yet, these macro indicators mask the heterogeneity of the banking system, especially concerning the credit market. A direct consequence of this heterogeneity is already observed in the definition and the update of the samples of banks used for the data collection. In order to meet the accuracy criterion defined by the regulation, it proved necessary to include no less than 330 banks from a population of about 900. More specifically, the volatility of interest rates appears to be quite high for loans granted to firms by specialized institutions.

In order to improve our understanding of this heterogeneity, we use another set of individual data obtained from the same sample of respondents, defined on a loan-by-loan basis, and perfectly consistent with the aggregated data used for MIR reports. Indeed, these data provide a disaggregation of monthly aggregates per bank on a quarterly basis. More precisely, the credit institutions included in the sample are required to transmit individual informations on loans every first month of a given quarter, including the amount, the maturity, the type of rate (fixed or floating), the narrowly defined interest rate (designed, in French, by the acronym TESE) and the overall effective rate (TEG) adding compulsory charges to the interest rate component : administration, guarantees, credit insurances. These statistics are primarily used by the BdF to calculate usury rates on loans to households and firms every quarter. All in all, our database is made up of roughly 900000 loans for each quarter. Since this database is not a panel in the strict sense, the time series dimension is difficult to undertake in our analysis : clearly, the longitudinal approach seems more fruitful.

In this work, we seek to investigate two problems:
Firstly, we aim to give the most precise description of the distributions of loan rates, particularly the number of modes which can be identified. The distributions are not weighted by flows of new business, because we want to disentangle interest rate effects and structural effects as much as possible in our analysis. The presence of multimodality can be suggestive of more than one underlying unimodal interest rate distribution. Indeed, these modes may be very informative if they can be associated to micro markets through the identifications of banks and/or instruments which explain their occurrence. Secondly, we try to quantify the various effects of the usury rate in the top of the distribution. In this respect, the recent period appears to be very informative since the usury legislation has been modified in 2003 and 2005 in order to reduce the

[1] For the largest banks, included automaticaly in the sample, subsamples defined at the level of banking desks are used. This facilitates the checking of data (creation of an audit trail).

[2] The impact of monetary policy on income flows is also analysed, through the collection of interest flows (debit and credit) and associated average outstanding amounts. The corresponding interest rates are not analysed in this paper.

perimeter of loans granted to firms subject to usury rates. To assess whether these decisions had a significant effect on the distribution of interest rates is a main objective of the paper. Lastly, for loans to households, the results are part of a set of background studies prepared by Bdf as an input for the reflections on potential improvements of the usury rate mechanism fixation.

The rest of the paper proceeds as follows. The data set and the determination of the usury rates are presented in section 2. The statistical framework is introduced in section 3, with a detailed account of the methodology in section 4. The technical, but essential, question regarding the way the dataset used for the estimation step is selected is dscussed section 5. Section 6 reports the main empirical results. Finally, some technical developments and detailed results are given in the appendix. The estimations have been made with the SAS® V 8.2 software, module IML.

Notations :

- $N\left(m, \sigma^2 | r_1, r_2\right)$ stands for the normal law with parameters $\left(m, \sigma^2\right)$, left and right truncated by r_1 and r_2 respectively; this is the law of $Y = X \times 1_{\{r_1 < X < r_2\}}$ with $X \rightsquigarrow N\left(m, \sigma^2\right)$. When the truncation affects only the right side of the distribution, we write $N\left(m, \sigma^2 | -\infty, r_2\right)$.
- $\#(A)$ is the cardinal of A.
- $\Longrightarrow$ means convergence in distribution when the size n of the sample $(X_1, \ldots, X_n)$ goes to $+\infty$.
- $U(a, b)$ is a random variable following the uniform law with support on $[a, b]$

2 Impact of the usury law on the credit market

For each category of loan, usury rates are defined in the following way : given the average effective rate $\overline{r}_{t-1}$ calculated for the previous quarter, the usury rate for the current quarter is :

$$r_t^u = \frac{4}{3}\overline{r}_{t-1} \tag{1}$$

The average effective rate is a simple mean of the annualized percentage rates observed during the first month of quarter $t-1$. At this point, we emphasize that the calculations are in fact much more involved in order to ensure the reliability of this estimate. Thus, the usury rate is in fact a highly non-linear function of individual observations. Firstly, outliers are dealt with through the use of asymmetric trimmed means which evolve over time, depending on the volatility of individual interest rates. Secondly, a kind of post-stratification is applied to the estimator through a weighting system applied at the network level, and based on flows of new contracts observed for the current quarter and outstanding amounts averaged over the past three years.

Choosing a simple mean for the average interest rate over-estimates the actual cost of loans for costumers as measured by MIR statistics. Indeed, if the amount of the loan is negatively correlated with the interest

rates, a fact often reported in empirical studies, the usury rate is higher than it would be if it was indexed on an average interest rate weighted by flows, the latter being more representative of economic activity.

Relaxing the usury regulation has not modified this definition nor the instruments included in the categories of loans as initially defined by law in 1989. In fact, the perimeter of the categories pertaining to firms has been reduced in two steps. Firstly, the law n^0 2003-721 of 2003 which has taken effect from April 2004 suppressed the usury rate for non financial corporations, with the exception of bank overdrafts. Secondly, the law n^0 2005-882 of 2005 suppressed the usury rate for individual enterprises, the bank overdrafts being still excluded. This law has taken effect from October 2005.

For the sake of clarity, we summarize in table *1a* below the content of each category, and the chronology of the population targeted by the usury rate. We precise that the non-financial corporations include the individual enterprises (IE), for loans granted for professional use, and the Non Profit Institutions Serving Households (NPISH).

N.B : The numbering of the categories will be used throughout the paper in order to shorten the labelling of the instruments included in each category. For instance, cat. 3 will always refer to "Personal loans and other loans over 1524 euros".

<table>
<tr><th>Cat.</th><th>Description</th><th>$\prec$ 2004/04</th><th>$\succ$ 2004/04
$\prec$ 2005/07</th><th>$\succ$ 2005/10</th></tr>
<tr><td></td><td>Consuming loans:</td><td colspan="3" rowspan="9">Individuals
IE (personal use)</td></tr>
<tr><td>1</td><td>- Loans up to 1524 euros</td></tr>
<tr><td>2</td><td>- Bank overdrafts, loan account, instalment credits,</td></tr>
<tr><td></td><td>revolving credit over 1524 euros</td></tr>
<tr><td>3</td><td>- Personal loans over 1524 euros</td></tr>
<tr><td colspan="2"></td></tr>
<tr><td></td><td>Housing loans:</td></tr>
<tr><td>4</td><td>- Loans at fixed rate</td></tr>
<tr><td>5</td><td>- Loans at floating rate</td></tr>
<tr><td>6</td><td>- Bridging loans</td></tr>
<tr><td></td><td>Loans to NFC:</td><td rowspan="4">NFC, IE
NPISH</td><td rowspan="4">IE
NPISH</td><td rowspan="4">NPISH</td></tr>
<tr><td>7</td><td>- Instalment credits</td></tr>
<tr><td>8</td><td>- Loans at floating rate with agreed maturity over 2 years</td></tr>
<tr><td>9</td><td>- Loans at fixed rate with agreed maturity over 2 years</td></tr>
<tr><td>10</td><td>- Bank overdrafts</td><td colspan="3">NFC, IE
NPISH</td></tr>
<tr><td>11</td><td>- Other loans with agreed maturity up to 2 years</td><td>NFC, IE
NPISH</td><td>IE
NPISH</td><td>NPISH</td></tr>
</table>

Table 1a

It should be noticed that bank overdrafts, revolving credit and other kinds of loans without agreed maturiy are included in interest rates on outstanding amounts in MIR reports. Nevertheless, they are considered as

new business according to the usury regulation, and the collected data refer to contractual interest rates which apply within the authorized limits agreed between the credit institution and the customer. The corresponding maximal amounts are then reported as amounts of loans in our database.

Our study is restricted to loans subject to the usury regulation, taking for reference the law in force before 2002. The representativeness is quite satisfying for loans to households (around 90% of the total of loans), and still correct for firms (70%) despite the fact that the loans with very high amounts, or associated to specific instruments (leasing) are excluded.

The data used in this study concern the same month, October, over three successive years (2003, 2004 and 2005), and permit us to consider the three states of the usury regulation. The choice of a same month allows us to ignore the impact of seasonality in the credit market which may induce additional variability in the results. More importantly, we expect to interpret the differences in the distributions in the light of the relaxing of the usury regulation.

We provide now some descriptive statistics about our dataset. The relative market share of each category in the total of loans to households or NFC subject to the usury regulation is given in table 1b below. The proportions are calculated without any weighting, or by weighting with the flows. As expected, housing loans is the most important category, according to the weighting flow, of loans to households, although these contracts make up only 6,4% of the total of credit lines. For firms, short and long term loans at fixed rate appear to be prominent. For bank overdrafts, the high values of the percentages weighted by flows result from the fact that the reported flows are measured by the maximum amount allowed by the credit institution, and not by the effective amount that has been effectively drawn.

Cat	*Flows*	*Nb of loans*
1	2	32,3
2	30,4	42,5
3	15,6	18,8
4	28,7	4,1
5	18,9	1,9
6	4,4	0,4
Total	*100*	*100*
7	1,2	1,8
8	7,5	3
9	26,1	15,3
10	38,5	10,9
11	26,7	69
Total	*100*	*100*

Table 1b : structure per category (in %)

We come back to some general time series considerations displayed in fig. 1c. During the period under investigation (October 2003/October 2005), the average interest rates[3] decreased, more for housing loans than for consumer loans. In the meantime, loans to NFC did not exhibit any trend.

[3] Overall effective rate (TEG) for loans to households, and (narrowly defined) interest rates (TESE) for loans to NFC. These data are flow weighted average of individual data, according to the MIR regulation.

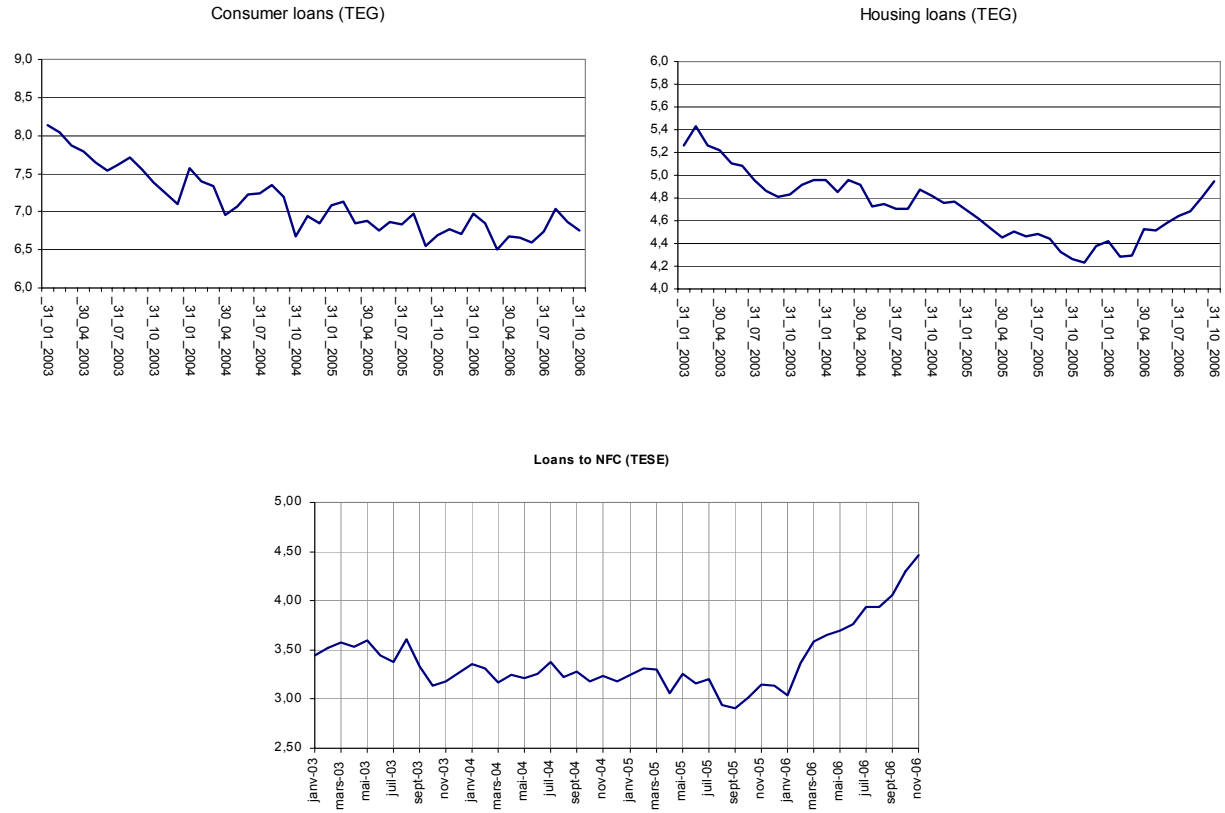


Fig. 1c

We proceed now with two basic indicators representing the distributions of loan effective rates (tables *1d* and *1e*), mean and standard error weighted by sampling weights. The subscript "1" refers to the reference month October 2003, whereas the index "2" refers to October 2004. We observe a lower dispersion for housing loans across categories, and a higher dispersion for loans vis-à-vis NFC, especially for loans whose duration up to two years. It can be noticed that both levels and volatility decreased between 2003 and 2004 for loans to households, whereas most categories of loans to NFC behaved differently.

Category	mean_1	std_1	mean_2	std_2
1	14,1	4,3	13,8	4,7
2	10,4	4,6	9,9	4,3
3	6,9	1,9	6,3	1,7
4	4,9	0,8	4,9	0,7
5	4,4	0,8	4,3	0,6
6	4,8	0,8	4,5	0,8

Table 1d : individuals

Category	$mean_1$	std_1	$mean_2$	std_2
7	6,3	2,7	6,7	2,4
8	4,4	1,0	4,3	0,9
9	4,9	1,1	5,0	1,1
10	8,3	3,1	9,2	2,6
11	4,2	2,1	4,7	2,3

Table 1e : NFC

One important issue for our work is the homogeneity of the categories defined by the usury regulation. Implicitly, an average interest rate makes sense if it really represents the whole distribution, for instance when the dispersion of the individual interest rates is low. A preliminary examination of this question is possible through the traditional approach based on variance analysis. Indeed, we calculate the (residual) within variance (in % of the total variance) in a classical Anova analysis. We first consider a one-way analysis, where we use the credit institution which issued the loan as an explanatory variable : the results are given between brackets in table 1f below. Then, we consider a two-way analysis of variance with the inclusion of the category of loan. We treat separately loans to households and to NFC.

	oct. 2003	*oct. 2004*
Households	(51)23	(52)21
NFC	(75)43	(65)29

Table 1f : Within variance (in % of total variance)

The results indicate an important variability of interest rates within the categories of loan, and more importantly, that this variability is still significant even when the credit institution is added as an explanatory variable : the residual variance is stil around 20 to 30%. This is precisely the objective of this paper to provide insights into the form of the underlying distribution of interest rates, in order to understand this volatility.

The effects of the usury rate on the distribution of the overall effective rates will be deduced them from a careful estimation of the probability density function of the distributions. Two situations may arise when a distribution is right-censored.

- **H1** : the usury rate has only a censoring effect: loans which bear interest rates beyond the legal threshold are not granted, and the distribution shows a truncation effect measured by the proportion p of rejected loans. This proportion is given by the shaded area in fig. 1g below.
- **H2** : the usury rate induces a mode just under the threshold. In this case, there is an accumulation of loans in the neighborhood of the usury rate, and the distribution is distorted (see fig. 1h). The

distortion is not easy to interpret. Indeed, It may indicate that banks adjust their loans rates (for instance, by cutting some of the charges included in the overall effective rates) in order to comply with the legal requirements : some evidence support this hypothesis, like the negative correlation between charges and interest rates observed for some categories of loans[4]. But some credit institutions could also systematically adjust their interest rates (corresponding to a specific instrument or particular costumers) in the immediate vicinity of the usury rate.

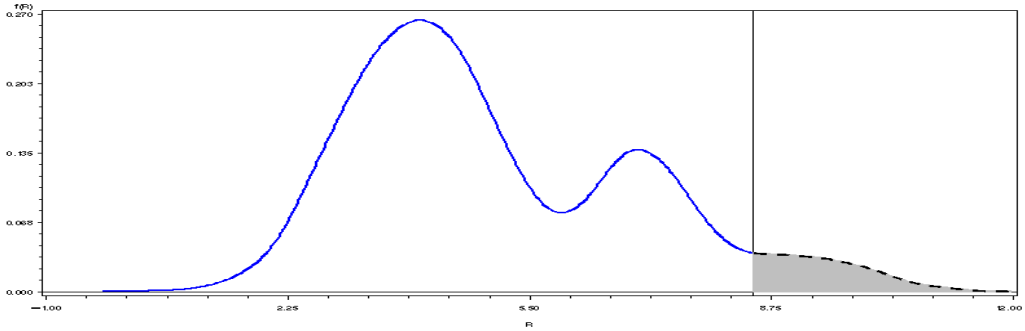


Fig. 1g : distribution f(x) under H1

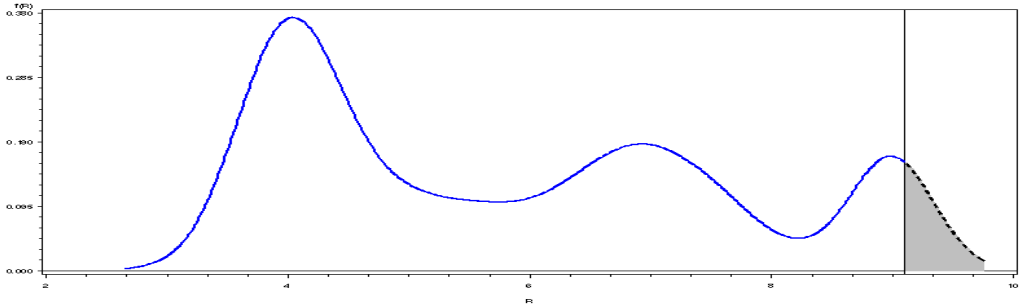


Fig. 1h : distribution f(x) under H2

Of course, both hypotheses are not mutually exclusive : we could observe simultaneously an important truncation of the distribution (H1) and an significant mode near the usury rate (H2) if the demand for loans is highly constrained on the supply side.

Once the distribution has been estimated, it is possible to estimate the eviction rate, i.e, the proportion of rejected loans (corresponding to interest rates larger than the usury rate) in the total population of loans.

At this point, it is worth emphasizing an important side effect of formula 1, namely the use of truncated mean for the definition of the new threshold. In particular, even if the distribution of interest rates is

[4]I thank E. Gervais for pointing out this fact to me.

constant over time, the eviction rate p is time dependent, $p \equiv p_t$ because in this particular case, the usury rate is a decreasing sequence; one could imagine a extreme situation where the usury rate $r_t^u \to 0$ so that $p_t \to 1$. With time dependent distributions, complex (and unexpected) dynamics could arise. In order to illustrate this point, we apply the usury methodology to a category of instruments not covered by the current regulation : leasing for NFC, over the period october 2003-january 2006. In doing so, we build two fictitious sequences of usury rates and the associated eviction rates[5]. In the first sequence (broken line in fig. 1i), the usury rate is computed with all the available observations; in the second sequence (solid line), we use the truncated distribution as according to 1. The gap between the two curves measures the impact of truncation on the dynamic of the usury rate.

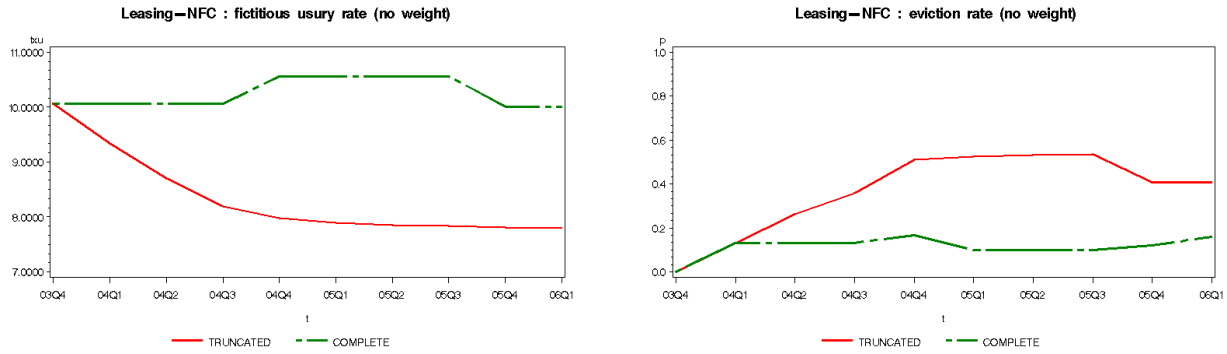


Fig. 1i

The two series of eviction rates diverge apart, with a 'natural' eviction rate fluctuating around 15%, whereas the 'truncated' eviction rate ranges from 40 to 60% at the end of the period. The spread between the two series of usury rates reflects this gap, with a maximum of 300 bp.

One may conclude that uncontrolled dynamics in the trajectory of usury rates can be avoided by cancelling the truncation effects, which means taking into account the unobserved part of the distributions in the calculation of the usury rates. Precisely, the methodology developped in the paper provides such opportunity, although publishing legal rates partially based on econometric estimations of interest rates larger than the usury rate does not seem conceivable.

3 Statistical framework

A basic filtering of the data has been conducted in order to remove credit lines for which either the interest rate or the amount of the loan was unmistakably an outlier. For the moment[6], we suppose that, for any

[5] p_t is estimated with the usual non-parametric estimator of the c.d.f (see 2).
[6] We postpone the discussion of this hypothesis to section 5.

category of loans considered in the study, the overall effective rates can be considered as drawn independently from some unknown law X, hence forming a sequence of *i.i.d* variables denoted by $(X_1, \ldots, X_N)$. The *c.d.f* and *p.d.f* of X are denoted by $F(x)$ and $f(x)$ respectively, the argument x taking its values in the range of values of the interest rate. Finally, a mode of the distribution is simply a local maximum of f.

3.1 Non parametric estimation

The basic tool is the non parametric estimation of both the c.d.f and the p.d.f. For the c.d.f, $F(x)$, the estimator is simply the empirical distribution function:

$$\widehat{F}_n(x) = \frac{\#\text{ observations } \leq x}{n} \tag{2}$$

It is well known that $\widehat{F}_n(x)$ is (uniformly) consistent for $F(x)$.

Unfortunately, the situation is more complicated for the p.d.f $f(x)$. The standard estimator is of the kernel type (see e.g Silverman (1986))[7]:

$$\widehat{f}_n(x) = \frac{1}{nh}\sum_{k=1}^{n} K\left(\frac{x - X_k}{h}\right) \tag{3}$$

It is commonly recognized that the choice of the kernel can be considered as a secondary issue (Silverman, (1986)) since it does not affect much the properties of $\widehat{f}_N(x)$. Then we focus on the Gaussian kernel:

$$K(x) = \varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \tag{4}$$

The bandwidth parameter h drives the smoothness of the curve $\widehat{f}_N$ and depends on N. Indeed, it can be shown that if $h + \frac{1}{nh} \to 0$ when $n \to +\infty$, $\widehat{f}_n(x) \to f(x)$ in probability for all x. Moreover, minimizing the asymptotic variance of $\widehat{f}_n(x)$ allows to define an optimal bandwidth:

$$h_n = C(K, f)\, n^{-\frac{1}{5}} \tag{5}$$

with C constant depending only on K and f which must be estimated, for instance according to the Sheather and Jones (1991) plug-in method. In practice, the value of h used in the estimator is crucial, because even the shape of the curve may be significantly altered when moving from one value of h to another one. In particular, because the parameter C depends in a complicated way of the unknown distribution f, its estimation could suffer from a lack of accuracy.

We can reformulate this point with a different perspective : when K is the Gaussian kernel, the number of modes found in the distribution $\widehat{f}_n$ is a decreasing function of h (Silverman (1981)). Choosing a high value of h is then equivalent to estimate a very smooth distribution with very few modes (only one in an extreme case). On the contrary, as h approaches 0, the curve become more and more irregular.

[7] The kernel K is an even, positive function which attains its maximum at zero, $\int K = 1$, and K decreases rapidly to zero as $x \to \pm\infty$.

We illustrate these considerations with estimations pertaining to category 6 for the reference month October 2003 (fig. 2a). The three estimators $\widehat{f}_n(x)$ differ only by the value of h, with the optimal h given by (5) associated to the solid line. We observe that the main characteristics of the distribution are identical within the three estimates, but local analysis clearly depends on the value of h.

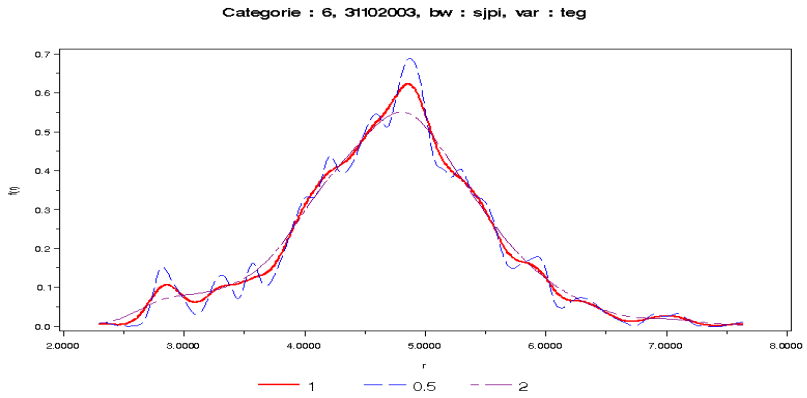


Fig. 2a.

Besides, the choice of the method used to estimate the parameter C in (5) is also important. For the category 1, we show fig. 2b the curves obtained from three classical approaches : Sheather & Jones, Silverman and "Simple Normal Reference". In this case too, we observe that the informations delivered by the curves beyond the overall shape of the distribution may be quite different.

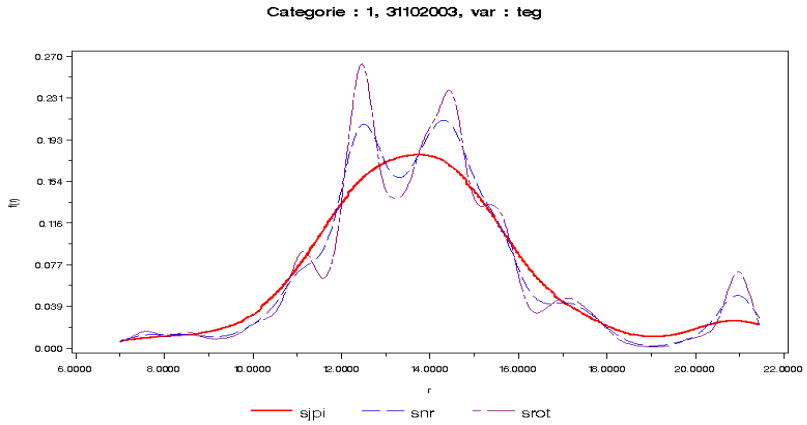


Fig. 2b.

These drawbacks of the non parametric methodology are problematic for our study because we are particularly interested in the details of the distribution, especially for high values of interest rates. Moreover, we want to assess the number of modes for each category of loans. For this reason, we turn now to another approach, based on parametric tools. The price we have to pay is to put more constraints on the common law of the X_k.

Remark : *the statistic $h(\mathcal{M})$ allowing exactly $\mathcal{M}$ modes in the distribution estimated by (3) can be used for a test of the unimodality hypothesis against two modes (Cheng and Hall (1999)). Unfortunately, this test has not yet been extended to the general case of $\mathcal{M}$ modes against the alternative of $\mathcal{M}'$ modes, with $\mathcal{M}' > \mathcal{M}$. It is therefore of little value for our study.*

3.2 Modelling a truncated Gaussian law

Before tackling complicated models, we verify in this section that the distribution of interest rates is neither normal nor log-normal before truncation by the usury rate, whenever it applies. Under the hypothesis of normality, the distributions are obviously unimodal, and the eviction rates could easily be estimated. We then test whether the sample of data, or its logarithm, comes from the distribution $N\left(m, \sigma^2 \left| -\infty, r^u \right.\right)$, where r^u is the usury rate. For a category of loans outside the scope of the usury law, we test the adequation to the standard normal law, which is equivalent to set $r^u = +\infty$ in the developments below. In particular, for the reference months October 2004/2005 and the categories 7,8,9 and 11, we don't take into account the truncation effect which may result from the residual perimeter of loans still subject to the usury law. In other words, we maintain $r^u = +\infty$ in such cases.

Let $\Phi(x)$ be the c.d.f for the standard normal law. Under our null hypothesis, the c.d.f. of the observations is:

$$F(x) = \begin{cases} \Phi\left(\frac{x-m}{\sigma}\right) / \Phi\left(\frac{r^u-m}{\sigma}\right) & \text{if } x \leq r^u \\ 1 \text{ if } x > r^u \end{cases} \tag{6}$$

We use two goodness of fit tests which appear to be well-designed for our purpose:

- *Kolmogorov-Smirnov (K-S)*:

$$KS_n = \sqrt{n} \times \sup_{x \leq r^u} \left| \widehat{F}_N(x) - F(x) \right| \tag{7}$$

- *Anderson-Darling (A-D)*:

$$AD_n = n \int_{-\infty}^{r^u} \frac{\left\{ \widehat{F}_N(x) - F(x) \right\}^2}{F(x)\{1 - F(x)\}} dF(x) \tag{8}$$

These tests reject the hypothesis of normality for high values of the statistic. While K-S is able to detect departures from normality in the middle part of the distribution, A-D does by design a better job in the extreme parts. It is advisable to consider one supplementary test of the Anderson-Darling type, designed for a weaker hypothesis : testing for normality by restricting ourselves to the upper side of the distribution. To achieve this task, we define the modified A-D statistic:

$$\underline{AD}_n = n \int_{F(\theta)}^{1} \frac{\left\{ \widehat{F}_N(x) - F(x) \right\}^2}{1 - F(x)} dF(x) \tag{9}$$

The threshold θ is fixed according to $\theta = 3/4 \times r^u$, which is exactly the average interest rate used for the calculation of the usury rate. It is important to note that these tests are not performed directly on the

whole set of available data, but only on a subsample : details are provided later in section 5. Moreover, the distributions of the test statistics are estimated from bootstrap techniques; a detailed account of the methodology is provided section 9.1.

The tests are performed for both levels and logarithms of interest rates. Results (test statistics and 95% quantiles) are given in tables 3 and 4. An asterisk "*" identifies the categories for which the truncation is not taken into account.

Cat.	**log**						**level**					
	KS	*q95*	**AD**	*q95*	**AD**	*q95*	**KS**	*q95*	**AD**	*q95*	**AD**	*q95*
1	6,05	1,03	56,44	1,21	24,75	0,57	3,7	0,87	16,27	0,71	5	0,24
2	5,82	1,02	46,53	1,21	7,11	0,49	5,92	0,9	66,4	0,75	21,92	0,19
3	4,95	1,06	41,76	1,25	22,97	0,69	3,17	0,85	29,95	0,72	23,48	0,31
4	11,56	1,1	235,96	1,35	127,53	0,77	3,46	1,02	29,78	1,49	11,48	0,63
5	11,77	1,14	253,38	1,56	136,19	0,83	4,47	1,09	45,36	1,84	18,39	0,79
6	8,84	1,08	132,99	1,35	70,83	0,74	3,05	0,98	17,4	1,2	5,9	0,47
7	6,85	1,04	54,53	1,2	11,89	0,63	6,89	0,86	67,82	0,71	23,45	0,27
8	5,49	1,05	57,47	1,25	25,14	0,71	2,33	0,94	9,82	0,99	4,15	0,71
9	5,61	1,05	72,6	1,26	28,05	0,71	2,3	0,88	14,95	0,74	5,64	0,29
10	8,1	1,13	105,83	1,57	8,17	0,64	2,61	0,84	7,3	0,76	3,22	0,37
11	8,57	1,08	93,34	1,39	12,26	0,52	7,82	0,88	106,33	0,71	28,49	0,15

Table 3 : October 2003

Cat.	**log**						**level**					
	KS	*q95*	**AD**	*q95*	**AD**	*q95*	**KS**	*q95*	**AD**	*q95*	**AD**	*q95*
1	5,06	1,05	28,33	1,21	7,35	0,58	3,34	0,84	19,53	0,73	9,08	0,31
2	5,65	1,03	52,34	1,23	7,54	0,47	5,94	0,88	107,27	0,72	36,18	0,21
3	4,93	1,05	48,29	1,22	23,6	0,65	2,36	0,84	24,12	0,72	14,72	0,29
4	11,15	1,1	229,12	1,38	122,7	0,84	4,46	1,04	52,89	1,53	20,16	0,74
5	21,17	1,12	252,15	1,43	138,76	0,84	5,93	1,12	80,91	1,96	37,53	0,97
6	7,83	1,08	119,56	1,27	61,13	0,71	3,3	0,98	18,47	1,19	4,72	0,47
*7**	5,45	0,93	48,12	0,76	7,04	0,26	6,82	0,91	55,75	0,76	4,74	0,31
*8**	1,83	0,92	5,84	0,79	1,11	0,28	2,46	0,93	7,69	0,77	3,14	0,3
*9**	3,69	0,91	29,07	0,76	10,62	0,34	1,53	0,91	5,92	0,74	3,52	0,32
10	3,07	1,05	13,44	1,24	7,27	0,7	2,14	0,84	4,88	0,71	1,47	0,33
*11**	4,17	0,91	16,32	0,75	5,44	0,19	7,03	0,91	71,78	0,74	9,7	0,23

Table 4 : October 2004

For both periods, all the statistics are highly significant. For loans to households, especially housing loans, the specifications in log are massively rejected. For loans to firms, we can notice a slight decrease of the test statistics between 2003 and 2004, perhaps a side effect of the usury reform. Finally, this preliminary study provides clear evidence that the distributions of retail interest rates do not belong to the Gaussian world.

4 Parametric estimation

4.1 The model

The univariate variable X follows a $\mathcal{M}$ component finite Gaussian mixture model if its probability density function can be written as:

$$f(x) = \sum_{k=1}^{\mathcal{M}} \pi_k \varphi_k(x) \tag{10}$$

where $\sum_{k=1}^{\mathcal{M}} \pi_k = 1$, $\pi_k \in]0,1[$, and $\varphi_k(x)$ is the p.d.f of the normal law $\mathbf{N}\left(m_k, \sigma_k^2\right)$:

$$\varphi_k(x) = \frac{1}{\sigma_k} \varphi\left(\frac{x - m_k}{\sigma_k}\right) \tag{11}$$

φ is defined by (4), $\mathcal{M}$ is the smallest integer for which $f(x)$ can be written as in (10), and the couples $\left(m_k, \sigma_k^2\right)$ are all distinct. The model has a simple interpretation : for $k = 1, \ldots, \mathcal{M}$, X is drawn with probability π_k from the normal law with expectation m_k and variance σ_k^2.

We can always suppose that the sequence (m_k) is increasing, which entails that $m_{\mathcal{M}}$ is potentially the value closest to the usury law. We note that the model is invariant to any re-ordering of the regimes. Then, the parameters are defined only up to a permutation of the indexes $\{1, 2, \cdots, \mathcal{M}\}$. However, these conditions are generally not sufficient to ensure the identifiablity of model (10). We do not discuss further this issue, which is largely beyond the scope of this paper, and suppose in the sequel that all the technical requirements needed for the identifiability of the whole set of parameters are satisfied (see Mc Lachlan & Peel (2000) for a detailed discussion of this issue).

The model (10), (11) is very general for our concerns. It is regarded as a flexible tool which is able to approximate any continuous distribution, provided the number of components $\mathcal{M}$ is large enough. Generally, we obtain $\mathcal{M}$ modes in the distribution located at (m_k). The shape of each mode depends of the values of π_k and σ_k^2. However, the number of modes can be reduced, especially when two values of m_k are close. In the case we observe one single mode, the model allows a parsimonious description of departures from normality such as skewness and excess of kurtosis. The distribution in fig.5. illustrates this point, with $\mathcal{M} = 2$, $m_1 = 0, m_2 = 0.5$, $\pi_1 = 0.7$, $\pi_2 = 0.3$, $\sigma_1^2 = 1$ and $\sigma_2^2 = 0.04$:

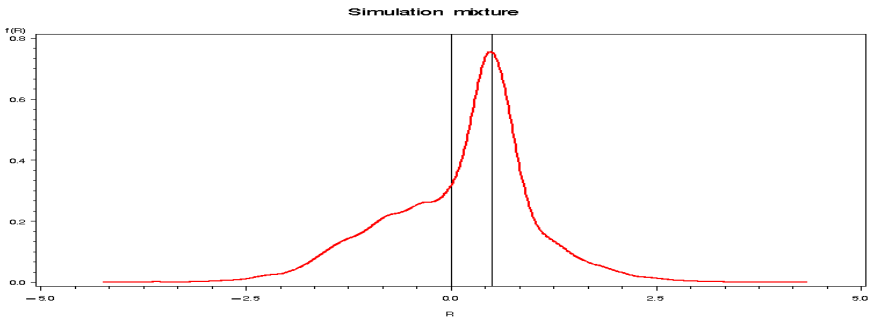


Fig. 5.

Our interpretation of the results will focus on modes rather than regimes. Indeed, we believe that modes are easier to interpret than regimes because they can be related in more straightforwardly to market segmentation issues. The statistical model is just a black-box which allows us to obtain the more accurate description of the p.d.f. of the data.

4.2 Estimation

The parameters to be estimated are the number of regimes, $\mathcal{M}$, then $\boldsymbol{\theta}_{\mathcal{M}} = \left(\boldsymbol{\pi}, \mathbf{m}, \boldsymbol{\sigma}^2\right)'$ with:

$$\begin{cases} \boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\mathcal{M}-1}) \\ \mathbf{m} = (m_1, \ldots, m_{\mathcal{M}}) \\ \boldsymbol{\sigma}^2 = \left(\sigma_1^2, \ldots, \sigma_{\mathcal{M}}^2\right) \end{cases}$$

We suppose that $\mathcal{M}$ is bounded from above, $\mathcal{M} \leq \mathcal{M}_{\max}$. We note the density $f(x) : f(x; \boldsymbol{\theta}_{\mathcal{M}})$. For a given number of regimes $\mathcal{M}$, the model is estimated by maximum likelihood. The log-likelihood is:

$$l_n(\boldsymbol{\theta}_{\mathbf{M}}) = \sum_{j=1}^{n} \log\left(\sum_{k=1}^{\mathcal{M}} \pi_k \varphi_k(X_j)\right) \tag{12}$$

There exists a sequence of local maximizers of the likelihood which is consistent and asymptotically normal (Bickel *and alii*, (1998)). However, as it is well known in this kind of model, a major difficulty arises from the fact that the likelihood diverges to $+\infty$ when at least one variance σ_j^2 approaches zero, i.e for values of the parameters converging to the bounds of the parameter space (Mc Lachlan & Peel (2000)). The estimation becomes challenging : instead of searching for a global maximum, we must identify an efficient local maximum. Besides, estimates near to zero for a subset of the variance parameters may occur in some particular situations, for instance when clusters of interest rate exist in the data, that is, a significant number of observations for which $X_j = c^{ste}$. To prevent the estimation procedure from converging to pathological values, we combine two additional procedures: the first one will be discussed later on section 5.2, and consists in a preliminary analysis of the data. The second one concerns the optimization of the likelihood, which is completed by the following constraints, derived from Hathaway (1985):

$$\left[\begin{array}{l} \text{(A)} : \forall j : \sigma_j > 10^{-4} \text{ and } \sigma_j < 10 \times \overline{\sigma} \\ \text{(B)} : \forall j : \pi_j \geq 5 \times 10^{-3} \end{array}\right. \tag{13}$$

$\overline{\sigma}$ is the standard error of the observations (X_i) ; the constraints prevent the algorithms from converging to the bounds of the parameter space. Through the global maximization of $l_n(\boldsymbol{\theta}_{\mathcal{M}})$ under the constraints (13), we get a consistent estimate [8] of $\boldsymbol{\theta}_{\mathcal{M}}$.

Another route to deal with these constraints is the use of the penalized likelihood advocated by Hamilton (1991). Unfortunately, in our framework, results (not reported in this paper) indicate that this procedure does not preclude the possibility to get bad estimates of σ_j and π_j.

[8] Obviously, the true value of the parameter is supposed to satisfy these conditions.

The model is not complete without the constraint of right censoring associated to the usury rate[9]. With this constraint, the p.d.f $\underline{f}$ of X is:

$$\underline{f}(x) = \frac{f(x)}{\int_{-\infty}^{r^u} f(t)\,dt} \text{ for } x \leq r^u$$

The log-likelihood of the observations is then:

$$\underline{l_n}(\boldsymbol{\theta}_{\mathcal{M}}) = l_n(\boldsymbol{\theta}_{\mathcal{M}}) - n \log\left\{\left(\sum_{k=1}^{\mathcal{M}} \pi_k \varphi_k(r^u)\right)\right\} \tag{14}$$

Instead of a direct numerical optimization of (14) which appears to be a quite formidable task, we proceed as follows.

- **Maximization of the complete likelihood (without censoring)** $l_n(\boldsymbol{\theta}_{\mathcal{M}})$.

  We use here the E-M algorithm (see the appendix for a synthetic description, and Mc Lachlan & Peel (2000) for a comprehensive survey). Starting from some initial guess for the parameter, $\boldsymbol{\theta}_{\mathcal{M}}^{(0)}$, the algorithm builds up a sequence of estimations $\boldsymbol{\theta}_{\mathcal{M}}^{(j)}$ such that, after any iteration, $l_n\left(\boldsymbol{\theta}_{\mathcal{M}}^{(j+1)}\right) \geq l_n\left(\boldsymbol{\theta}_{\mathcal{M}}^{(j)}\right)$. A fixed point of the algorithm (noted $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}\left[\boldsymbol{\theta}_{\mathcal{M}}^{(0)}\right]$ to keep in mind the dependency on the initial condition) is a local maximum of the likelihood. Because we seek to obtain a global maximum, we then iterate this algorithm for a set of initial values designed to spread the space of admissible values of the parameters. Let Θ_M be this set. The final estimator is then:

  $$\widehat{\boldsymbol{\theta}}_{\mathcal{M}}[\Theta_M] = \underset{\boldsymbol{\theta}_{\mathcal{M}}^{(0)} \in \Theta_M}{\operatorname{argmax}} l_N\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}\left[\boldsymbol{\theta}_{\mathcal{M}}^{(0)}\right]\right) \tag{15}$$

  Let $\mathcal{L}_n(\mathcal{M})$ be the log-likelihood at the optimum $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}[\Theta_M]$. Each run of E-M is stopped when at least one of the following conditions is true : *a)* one of the constraints (13) is not satisfied, *b)* the number of iterations exceeds 600, *c)* the convergence criterion is met:

  $$\max\left\{\frac{\left|\mathbf{m}^{(j+1)} - \mathbf{m}^{(j)}\right|}{\left|\mathbf{m}^{(j)}\right|}, \frac{\left|\boldsymbol{\sigma}^2_{(j+1)} - \boldsymbol{\sigma}^2_{(j)}\right|}{\left|\boldsymbol{\sigma}^2_{(j)}\right|}, \left|\boldsymbol{\pi}^{(j+1)} - \boldsymbol{\pi}^{(j)}\right|\right\} \leq \varepsilon = 10^{-4} \tag{16}$$

- **Maximization of the likelihood with censoring** $\underline{l_n}(\boldsymbol{\theta}_{\mathcal{M}})$

  This step uses a numerical algorithms for the maximization, with the optimal value (15) resulting from the previous step used input for initialization. Unfortunately, this procedure provided deceptive results. Indeed, it didn't improve the likelihood in a significant way, unless parameters were allowed to vary freely, without the constraints imposed to ensure interpretable values. The huge amount of parameters explains certainly the problems we encountered. We will work thereafter with the estimation obtained from the E-M methodology, $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}[\Theta_M]$.

[9] We neglect left-censoring issues arising from the positivity of interest rates.

- **Choice of initial values.**

  The determination of Θ_M being an important ingredient of our strategy, we provide a detailed account of the way it is built. Let $x_0 = \inf X_i$ and $x_{\mathcal{M}+1} = \sup X_i$. Three strategies are used:

1. Partial random initialization :

   Calculate $h(\mathcal{M})$, smallest bandwidth for which the non-parametric estimator (3) has exactly $\mathcal{M}$ modes, $x_1, \ldots, x_{\mathcal{M}}$. Then, there exists for $\mathcal{M} > 1$, $\mathcal{M} - 1$ local minima (the antimodes) $\in]x_0, x_{\mathcal{M}+1}[$ noted $z_1, \ldots, z_{\mathcal{M}-1}$, with:

   $$z_k \in]x_k; x_{k+1}[$$

   We set the end-points $z_0 = x_0$ and $z_{\mathcal{M}} = x_{\mathcal{M}+1}$. Now, define for $k = 1, \ldots, \mathcal{M}$, the initial values:

   $$m_k^{(0)} = x_k$$
   $$\pi_k^{(0)} = \frac{\#\text{ observations } X_i \in [x_{k-1}; x_k]}{n}$$
   $$\sigma_k^{(0)} = \overline{\sigma} \times \sqrt{U\left(\frac{1}{10}, 2\right)}$$

   For the first initialization, we also use:

   $$\sigma_k^{(0)} = \sqrt{\mathcal{V}_{\text{emp}}\left(X_i \in [x_{k-1}; x_k]\right)}$$

   A limited simulation experiment indicates that this procedure provides plausible starting values for π and m.

2. Full random initialization:

   $$\begin{cases} m_k^{(0)} = U(x_0, x_{\mathcal{M}+1}) \\ \pi_k^{(0)} = \frac{\widetilde{\pi}_k^{(0)}}{\sum_{k=1}^{M} \widetilde{\pi}_k^{(0)}} \text{ with } \widetilde{\pi}_k^{(0)} = U(0.05, 0.95) \\ \sigma_k^{(0)} = \overline{\sigma} \times \sqrt{U\left(\frac{1}{10}, 2\right)} \end{cases}$$

3. Use of $\widehat{\boldsymbol{\theta}}_{\mathcal{M}+\mathbf{1}}\left[\Theta_{M+1}\right]$

   When estimators for the mixture with $M+1$ components are available, we can identify among these components the pair of nearest neighbours (k_1, k_2) in the sense of Kullback contrast (Figueiredo *et alii* (1999)). Then, it is possible to calculate the values of the parameters $\left(m_{\underline{k}}, \sigma_{\underline{k}}^2\right)$ for the new component resulting from the merging of k_1 and k_2, that is, such that:

   $$\frac{\pi_{k_1}}{\pi_{k_1} + \pi_{k_2}} \varphi_{k_1}(x) + \frac{\pi_{k_2}}{\pi_{k_1} + \pi_{k_2}} \varphi_{k_2}(x) \cong \varphi_{\underline{k}}(x) \tag{17}$$

We then get a set of initial values for the model with $\mathcal{M}$ components:

$$\begin{cases} \boldsymbol{\pi} = \{(\pi_1, \ldots, \pi_{\mathcal{M}+1}) \setminus \{\pi_{k_1}, \pi_{k_2},\}\} \cup \{\pi_{k_1} + \pi_{k_2}\} \\ \mathbf{m} = \{(m_1, \ldots, m_{\mathcal{M}+1}) \setminus \{m_{k_1}, m_{k_2},\}\} \cup \{m_{\underline{k}}\} \\ \boldsymbol{\sigma}^2 = \left\{\left(\sigma_1^2, \ldots, \sigma_{\mathcal{M}+1}^2\right) \setminus \left\{\sigma_{k_1}^2, \sigma_{k_2,}^2\right\}\right\} \cup \left\{\sigma_{\underline{k}}^2\right\} \end{cases}$$

Finally, we use 100 set of initial values obtained from strategy 1 and 2 when $\mathcal{M} = \mathcal{M}_{\max}$, and 50 when $\mathcal{M} < \mathcal{M}_{\max}$. For these values of $\mathcal{M}$, strategy 3 is used as a complementary source, with one initialization[10].

4.3 Choosing $\mathcal{M}$

4.3.1 Formal test

We define for $\mathcal{M} = 1, 2, \ldots, \mathcal{M}_{\max} - 1$, the sequential tests:

$$\begin{array}{l} \mathbf{H}_0^{\mathcal{M}} : \ f \text{ has } \mathcal{M} \text{ mixtures} \\ \mathbf{H}_a^{\mathcal{M}} : \ f \text{ has } \mathcal{M}+1 \text{ mixtures} \end{array}$$

The sequence stops when $\mathbf{H}_0^{\mathcal{M}}$ can not be rejected for for $\mathcal{M} = \widetilde{\mathcal{M}}$. Then one can concludes that $\widehat{\mathcal{M}} = \widetilde{\mathcal{M}}$.

The test statistic for testing $\mathbf{H}_0^{\mathcal{M}}$ against $\mathbf{H}_a^{\mathcal{M}}$ is the likelihood-ratio:

$$\mathcal{LR}_n\left(\mathcal{M} \,|\mathcal{M}+1\right) = -2\left\{\mathcal{L}_n\left(\mathcal{M}\right) - \mathcal{L}_n\left(\mathcal{M}+1\right)\right\} \tag{18}$$

Remark:

The test procedure is ascending, starting from low values of $\mathcal{M}$. Alternative procedures may be used, such as, descending procedure, testing $\mathcal{M}$ against $\mathcal{M}_{\max}$ mixtures. In these two methods, we need to estimate models with large values of $\mathcal{M}$, which is a quite problematic task (see the discussion below). For this reason, we maintain the ascending scheme for our tests.

The statistic does not follow the usual Chi-2 asymptotic because under $\mathbf{H}_0^{\mathcal{M}}$, the parameters of one regime are not identified, which implies that the likelihood and its derivatives do not depend on their values[11] (Hansen (1996)). Despite this drawback, the asymptotic law of $\mathcal{LR}_n$ has been obtained (Dacunha-Castelle and Gassiat (1999)), but unfortunately this limit distribution is defined in a very implicit way and depends in a quite intricate way of nuisance parameters related to the law of (X_i).

The results of Andrews (1999, 2001) could be used, but their practical implementation appear quite demanding. For this reason, we will use subsampling techniques for the estimation of the quantiles of this limit law (Politis and Romano (1994), Politis *et alii* (1999), see the appendix for a brief summary). This approach relies upon estimations on B subsamples $\widetilde{s}_b, b = 1, \ldots, B$ drawn from $\widetilde{s}$ (here, $B = 500$). Closely related to the bootstrap, but valid in a much more general context, this technique should be used whenever the theoretical validity of the bootstrap has not yet been established, or when it is known that bootstrap analysis doses not work. The main drawback of subsampling is that the subsamples must have a reduced size $n_b < n$. More precisely, $n_b / n \to 0$ when $n \to +\infty$. As a consequence, in empirical studies, these techniques are relevant only when the amount of data is very important, which is precisely our case.

For the estimation step, the initial values $\boldsymbol{\theta}_{\mathcal{M}}^{(0)}$ are obtained in the same way as before, completed by $\boldsymbol{\theta}_{\mathcal{M}}^{(0)} = \widehat{\boldsymbol{\theta}}_{\mathcal{M}}\left[\Theta_M\right]$, the estimated value of the parameter for the the "master" sample : we work with a total of 10 initial values.

[10] For $M = 1$, the estimation is trivial, and we need only one initial value.

[11] However, the likelihood is always increasing with the number of regimes.

Once the quantiles $q_n(1-\alpha)$ of the limit law of $\mathcal{LR}_n(\mathcal{M}\,|\mathcal{M}+1)$ are estimated from the estimations performed on each subsample, we are able to reject $\mathbf{H}_0^{\mathcal{M}}$ at level α when:

$$\mathcal{LR}_n(\mathcal{M}\,|\mathcal{M}+1) > q_n(1-\alpha)$$

4.3.2 Information criteria

We introduce in this section an alternative approach based on the use of Information Criteria. In our context, these criteria belong to the "Minimum Message Length" family. They are used in the field of pattern and speech recognition:

$$\begin{aligned}\mathbf{MML}_1(\mathcal{M},n) &= \frac{\mathcal{L}_n(\mathcal{M})}{n} - \frac{(3\mathcal{M}-1)}{2}\times\frac{\log n}{n} \\ \mathbf{MML}_2(\mathcal{M},n) &= \tfrac{\mathcal{L}_n(\mathcal{M})}{n} - \tfrac{\mathcal{M}-1}{2}\times\tfrac{\log n}{n} - \tfrac{1}{n}\textstyle\sum_{k=1}^{\mathcal{M}}\log n\pi_k \\ &= \tfrac{\mathcal{L}_n(\mathcal{M})}{n} - \tfrac{(3\mathcal{M}-1)}{2}\times\tfrac{\log n}{n} - \tfrac{1}{n}\textstyle\sum_{k=1}^{\mathcal{M}}\log \pi_k\end{aligned} \tag{19}$$

$\mathbf{MML}_1$ is similar to the usual BIC criterion; the add value of $\mathbf{MML}_2$ stems from the factor $n\pi_k$ (instead of n), which is the number of observations really useful for the theoretical estimation of the component k of the mixture. on the other hand, the $\mathcal{M}-1$ parameters π_k are estimated with the complete data. This criterion is asymptotically equivalent to MML_1 and uses a smaller penalty designed to correct the tendency for MML_1 to underestimate $\mathcal{M}$ in finite sample, as reported by for instance by Figueiredo and *alii* (1999).

An ultimate version of this criterion, still equivalent to MML_1 has been proposed by Figueiredo and Jain (2002): this is the one used in our empirical work.

$$\mathbf{MML}_3(\mathcal{M},n) = \frac{\mathcal{L}_n(\mathcal{M})}{n} - \frac{1}{n}\sum_{k=1}^{\mathcal{M}}\log\left(\frac{n\pi_k}{12}\right) - \frac{\mathcal{M}}{2n}\left\{\log\frac{n}{12}+3\right\} \tag{20}$$

$\mathcal{M}$ is finally estimated from the values taken by the criterion according to the rule:

$$\widehat{\mathcal{M}}_3 = \underset{1\leq\mathcal{M}\leq\mathcal{M}_{\max}}{\operatorname{argmax}}\ \mathbf{MML}_3(\mathcal{M},n) \tag{21}$$

When n goes to $+\infty$, $\widehat{\mathcal{M}}_3 \to \mathcal{M}$ in probability (Keribin (2000), Gassiat (2002)). In empirical applications, this method appears to deliver reliable results (see for instance the simulation experiments reported by Psaradakis and Spagniolo (2002) in the related context of Markov-Switching models).

However, these two approaches suffer from the same drawback: estimation of models with high values of $\mathcal{M}$ can potentially generate trivial components with $\sigma_k \approx 0$ and/or $\pi_k \approx 0$. We know already that when $\sigma_k \to 0$, the likelihood diverges and it is easily seen that $\mathbf{MML}_3$ also diverges towards infinity when $\pi_k \to 0$: in both cases, an over parametrized model might ermerge from the selection process, even though the true value of $\mathcal{M}$ is smaller. The constraints (13) don't really solve the problem because it is likely that despite using numerous initial values, it won't be possible to get at least one final estimate fulfilling these constraints. Thus, if $\mathcal{M} > 5$, or when the variable is taken in log, it is quite common to observe numerical

problems during the iterations of the EM algorithm (more specifically, we can't prevent σ_k to converge to 0). In addition, these difficulties are amplified in the subsampling experiment, because the size of the sample is dramatically reduced. Intuitively, the number of régimes we are really able to identify decreases, and the risk of obtaining spurious estimates when $\mathcal{M}$ is large increases. We can even obtain non-increasing sequences of log-likelihood when the number of regimes is increasing !

At last, we could avoid the problem by reducing the value of $\mathcal{M}_{\max}$. However, the diversity of the lending market in France is a well established stylized fact : the population of credit institutions is quite heterogeneous and numerous instruments coexist in the categories of loans under investigation. As a consequence, we should maintain a high value for $\mathcal{M}_{\max}$, typically $\mathcal{M}_{\max}$=8. The counterpart of this choice is that the procedure (4.3.1) is extremely time consuming from an IT perspective.

The difficulties reported in this section are quite serious, and may obliterate the overall significance of the results. That's the reason why we proceed now to an alternative methodology which is in our view more efficient, although it doesn'y belong to the traditionnal econometric toolbox. The two previous methods will be used for illustration purpose with $\mathcal{M}_{\max}$=5, a value for which the estimation remains feasible.

4.4 A combined approach

We propose to estimate simultaneously $\mathcal{M}$ and the parameters of the mixture, an approach which is commonly used in pattern recognition, as advocated by Figueiredo and Jain (2002). It relies on a learning algorithm which permits to decrase progressively $\mathcal{M}$ during the learning, according to a "top to bottom" scheme. Starting from $\mathcal{M}_{\max}$ regimes and some initial guess $\boldsymbol{\theta}^{(0)}_{\mathcal{M}_{\max}}$, we iterate the EM algorithm as in (13), but with a improvement in the M-step which consists in setting to zero the π_k which appear to be no significant. In other words, the M-step performs component annihilation through an explicit rule for moving from the current value of $\mathcal{M}$ to a smaller one. The significativity of π_k is assessed from the posterior probabilities for a loan j to be produced by a given component. More precisely, we use the criterion below (the superscript "t" refers to the loop in the iteration of the EM algorithm):

$$\left\{ \sum_{j=1}^{n} P^{(t)} \left(j \text{ comes from regime } k \right) < 1 \right\} \Longrightarrow \pi_k^{(t+1)} = 0$$

The components for which $\pi_k = 0$ become irrelevant because they do no longer contribute to the log-likelihood. The number of regimes fluctuates during the iterations of the EM algorithm : let $\mathcal{M}_0^{(t)}$ be this variable. The algorithm is stopped if the following condition is true:

$$\frac{\left| \mathbf{MML}_3 \left(\mathcal{M}_0^{(t+1)}, n \right) - \mathbf{MML}_3 \left(\mathcal{M}_0^{(t)}, n \right) \right|}{\left| \mathbf{MML}_3 \left(\mathcal{M}^{(t)}, n \right) \right|} \leq \varepsilon = 10^{-6} \tag{22}$$

Once the convergence has been achieved, we obtain an estimate for the number of components, say $\mathcal{M}_0$, and for the parameters, $\boldsymbol{\theta}_{\mathcal{M}_0}$: this ends step "0". Turning now to step "1", we repeat exactly the same

operations, starting from $\mathcal{M}_0 - 1$ regimes, and an initial value $\boldsymbol{\theta}^{(0)}_{\mathcal{M}_0-1}$ obtained from the merging of the two closest components in $\boldsymbol{\theta}_{\mathcal{M}_0}$ in the sense of criterion (17). Then, the convergence of the EM algorithm provides the estimates:

$$\mathcal{M}_1 \leq \mathcal{M}_0 - 1 \text{ and } \boldsymbol{\theta}_{\mathcal{M}_1}$$

We iterate until step "$\mathcal{S}$" which provides only one regime as an input of the next loop, that is:

$$\mathcal{M}_S = 1$$

Therefore, we have obtained a set of estimated values for parameters, each set being associated to a specific value of $\mathcal{M} : (\boldsymbol{\theta}_{\mathcal{M}_0}, \ldots, \boldsymbol{\theta}_{\mathcal{M}_S})$. The "best" model $\left(\mathcal{M}, \widehat{\boldsymbol{\theta}}_{\mathcal{M}}\right)$ is then chosen according to the $\mathbf{MML}_3$ criterion, after elimination of models for which the constraints[12] are not fulfilled (13). The final result depends of course of the first initialization used at the beginning of the loop. We recall this dependence in the following compact notation:

$$\left(\mathcal{M}, \widehat{\boldsymbol{\theta}}_{\mathcal{M}}\right) = g\left(\boldsymbol{\theta}^{(0)}_{\mathcal{M}_{\max}}\right)$$

The whole procedure is repeated for different set of initial values, and we obtain the set of "final" estimates:

$$\left\{\left(\mathcal{M}, \widehat{\boldsymbol{\theta}}_{\mathcal{M}}\right)\right\} = g\left(\boldsymbol{\theta}^{(0)}_{\mathcal{M}_{\max}} \middle| \boldsymbol{\theta}^{(0)}_{\mathcal{M}} \in \Theta_M\right)$$

Again, we routinely select the best model according to the MML_3 criterion and the constraints (13). The final model corresponding to the initialization with $\mathcal{M}_{\max}$ regimes is then:

$$\left(\mathcal{M}^{(1)}, \widehat{\boldsymbol{\theta}}^{(1)}_{\mathcal{M}}\right) = g\left(\mathcal{M}_{\max}\right) \tag{23}$$

Now, the procedure is entirely repeated with $\mathcal{M}_{\max} - 1$, $\mathcal{M}_{\max} - 2, \ldots, 1$ initial regimes; it yields:

$$\left(\mathcal{M}^{(p)}, \widehat{\boldsymbol{\theta}}^{(p)}_{\mathcal{M}}\right) = g\left(\mathcal{M}_{\max} - p\right) \text{ pour } 0 \leq p \leq \mathcal{M}_{\max} - 1 \tag{24}$$

The number of initial values depends on the value of p: we consider 21 initial values if $p \geq 3$, 10 for 2 regimes, and 2 for one regime.

4.5 Level and logarithm

The estimation procedure described in the previous sections is performed for the level and the logarithm of the interest rate. We next have to choose between these two specifications. Obviously, both models are encompassed in the Box-Cox specification:

$$\frac{X^\lambda - 1}{\lambda} \rightsquigarrow f\left(x; \boldsymbol{\theta}_{\mathcal{M}}\right)$$

$\lambda \in [0, 1]$. The limit case $\lambda = 0$ corresponds to $\log(X)$, and $\lambda = 1$ corresponds to X because:

$$X - 1 \rightsquigarrow f\left(x; \boldsymbol{\theta}_{\mathcal{M}}\right) \Leftrightarrow X \rightsquigarrow f\left(x; \overline{\boldsymbol{\theta}}_{\mathcal{M}}\right)$$

[12] In the worst case, no model fulfills the constraints, and we retain by default the best model in the sense of MML_3.

with $\overline{\boldsymbol{\theta}}_{\mathcal{M}} \equiv \boldsymbol{\theta}_{\mathcal{M}}$ except for $\overline{\mathbf{m}}=(m_1+1,\ldots,m_{\mathcal{M}}+1)$.

The joined estimation of λ and $\boldsymbol{\theta}_{\mathcal{M}}$ by maximum likelihood appears to be a quite impossible task; indeed, the EM algorithm lacks simplicity because numerical optimizations would be required in each iteration, for both E and M steps. For this reason, we will restrict ourselves to a comparison between the MML criteria of the two models constraints by $\lambda = 1$ and 0 respectively. Let $\mathbf{L}_N(\lambda=1)$ and $\mathbf{L}_N(\lambda=0)$ be the corresponding log-likelihood for each model. The log-likelihood for $\lambda = 0$ is calculated as follows: the p.d.f of the variable in level, X, $f_{\mathbf{n}}(x;\boldsymbol{\theta}_{\mathcal{M}})$ is related to the p.d.f of $\log X$, $f_{\mathbf{l}}(y;\boldsymbol{\theta}_{\mathcal{M}})$ by:

$$f_{\mathbf{n}}(x;\boldsymbol{\theta}_{\mathcal{M}}) = \frac{f_{\mathbf{l}}(\log x;\boldsymbol{\theta}_{\mathcal{M}})}{x} \tag{25}$$

Thus,

$$\begin{array}{ll}\mathcal{L}_N(\lambda=0) & = \sum_{k=1}^{n}\log f_{\mathbf{n}}(X_k;\boldsymbol{\theta}_{\mathcal{M}}) \\ & = \sum_{k=1}^{n}\log f_{\mathbf{l}}(\log X_k;\boldsymbol{\theta}_{\mathcal{M}}) - \sum_{k=1}^{n}\log X_k\end{array}$$

The first factor in the r.h.s. is simply the log-likelihood (12) when the input variable is $\log X$: this is the result of the optimization procedure. Then, we calculate the standard BIC (MML_0 in our notations), because we believe that the refinements introduced in the definition of MML_2 and MML_3 are certainly not pertinent for the estimation of λ.

$$\mathbf{MML}_0(\lambda,n) = \frac{\mathcal{L}_n(\lambda)}{n} - (3\mathcal{M}_\lambda - 1)\times\frac{\log n}{n} \tag{26}$$

This approach is purely empirical: the analysis of the theoretical properties of the implied estimator of λ is largely beyond the scope of this paper.

4.6 Some simple indicators

We introduce in this section several indicators pertaining to the hypothesis we want to investigate. They will allow us to quantify the potential concentration of loans in the upper side of the distribution. Let κ_l for $l=1,\ldots,\mathcal{M}_-$ be the modes obtained from $f_{\mathbf{n}}(x;\boldsymbol{\theta}_{\mathcal{M}})$ for the model in level, and from (25) for the model in log. Generally, $\mathcal{M}_- \leq \mathcal{M}$. If r^u is the usury rate for the category under investigation, it is possible to observe $\kappa_{\mathcal{M}_-} > r^u$, especially when the usury regulation is relaxed : this is the case for categories 7,8,9 and 11 in October 2004 and 2005. In the other cases, an estimated mode may be slightly larger than the usury rate. In order to facilitate the temporal comparisons, we will restrict our attention to modes smaller, or in the immediate vicinity of the usury rate. This condition reads as follows:

$$\kappa_l \leq 1,1\times r^u \tag{27}$$

The largest mode satisfying (27) is then $\kappa_{\mathcal{M}_-}$.

1. **The eviction rate** : it is given for the model in level by:

$$p = 1 - F(r^u;\boldsymbol{\theta}_{\mathcal{M}}) \tag{28}$$

with

$$F\left(x;\boldsymbol{\theta}_{\mathcal{M}}\right)=\sum_{k=1}^{\mathcal{M}}\widehat{\pi}_k\Phi\left(\frac{x-\widehat{m}_k}{\widehat{\sigma}_k}\right) \tag{29}$$

For a model in logarithm, F is estimated from $\log X$ and:

$$p=1-F\left(\log r^u;\boldsymbol{\theta}_{\mathcal{M}}\right) \tag{30}$$

It is worth noting that for loans to NFC in October 2004 and 2005, some observed interest rates are larger than usury rates, thanks to the new regulation. In October 2005, only loans to NPISH are still subject to the usury law; then, p is simply the proportion of loans whose interest rates exceed the usury rate. However, for the sake of simplicity, we will still use the same designation for p.

2. **The posterior probabilities** that any given interest rate x_i (not necessarily included in the sample used in the estimation step) comes from the jth normal component of the mixture can be achieved as follows. Let $I_{ij}=1$ if x_i comes from component "j", 0 otherwise:

$$\mathbf{P}\left(j\left|x_i\right.\right)=\mathbf{P}\left(I_{ij}=1\left|x_i\right.\right)=\frac{\pi_j\varphi_j\left(x_i\right)}{\sum_{l=1}^{\mathcal{M}}\pi_l\varphi_l\left(x_i\right)} \tag{31}$$

We set $I_{ij}=1$ when $\mathbf{P}\left(I_{ij}\left|x_i\right.\right)>0,5$. The "market share" of each component is defined through the proportion of loans such that $I_{ij}=1$:

$$\theta_j=\frac{\sum_{i,I_{ij}=1}P_im_i}{\sum_i P_im_i} \tag{32}$$

m_i is the amount of loan "i" and P_i its sampling weight. Thus $\theta_{_{\mathcal{M}}}$ measures the economic weight of the regime close to the usury rate. In the same spirit, this indicator can be calculated at the bank level h, such as:

$$\theta_h\left(j\right)=\frac{\sum_{i\in h}P_i\times\mathbf{P}\left(j\left|x_i\right.\right)}{\sum_{i\in h}P_i} \tag{33}$$

$\sum_{j=1}^{\mathcal{M}}\theta_h\left(j\right)=1$ and "$i\in h$" means that the loan "i" was granted by bank "h". However, the regimes have above all a statistical interpretation. When the distributions are clearly multimodal, we favour interpretations based on modes rather than mean values across regimes. That's the reason why we associate to each mode κ_l for $l=1,\ldots,\mathcal{M}_-$ a set $K\left(l\right)$ of components k in accordance with the following criterion[13]:

$$\begin{array}{ll}\lambda=1: & \left|m_k-\kappa_l\right|\leq 1\\ \lambda=0: & \left|\exp\left(m_k-\sigma_k^2\right)-\kappa_l\right|\leq 1\end{array} \tag{34}$$

By doing so, we define a mapping between regimes and modes. When a regime k is associated to several modes l, we select the mode which minimize the distance $\left|m_k-\kappa_l\right|$. Regimes which can't be associated to at least one mode are gathered in $K\left(0\right)$. Finally, we obtain $K\left(l\right)$ for $l=0,\ldots,\mathcal{M}_{--}$,

[13] $\exp\left(m_k-\sigma_k^2\right)$ is the mode of a log-normal distribution with parameters $\left(m_k,\sigma_k^2\right)$

with $\mathcal{M}_{--}$ the effective number of modes connected to at least one regime. The modes are supposed to be sort in ascending order:

$$\kappa_1 < \kappa_2 < \dots$$

The pseudo posterior probabilities that an interest rate x_i comes from mode $l = 0, \dots, \mathcal{M}_{--}$ is then defined by:

$$\mathbf{P}\left(l \left| x_i \right.\right) = \sum_{j \in K(l)} \frac{\pi_j \varphi_j \left(x_i\right)}{\sum_{l=1}^{\mathcal{M}} \pi_l \varphi_l \left(x_i\right)} \tag{35}$$

The aggregation of these $\mathcal{M}_{--} + 1$ indicators for each bank h provides an new indicator similar to (33):

$$\theta_h \left(l\right) = \frac{\sum_{i \in h} P_i \times \mathbf{P}\left(l \left| x_i \right.\right)}{\sum_{i \in h} P_i} \tag{36}$$

$\sum_{l=0}^{\mathcal{M}_{--}} \theta_h \left(l\right) = 1$: this distribution provides a summary of the situation of the bank on the credit market.

We may also define an indicator similar to (32) :

$$\theta_{--} = \frac{\sum_{i, I_{i,\mathcal{M}_{--}}=1} P_i m_i}{\sum_i P_i m_i} \tag{37}$$

with

$$I_{i,\mathcal{M}_{--}} = 1 \text{ if } \mathbf{P}\left(\mathcal{M}_{--} \left| x_i \right.\right) > 0,5, \text{ zero otherwise} \tag{38}$$

The relative importance of mode $\kappa_{\mathcal{M}_{--}}$ is given from the theoretical mixing parameters π_j by:

$$\pi_{--} = \sum_{j \in K(\mathcal{M}_{--})} \pi_j$$

This quantity may be estimated with a larger dataset with:

$$\pi_{--} = \frac{\sum_{i, I_{i,\mathcal{M}_{--}}=1} P_i}{\sum_i P_i} \tag{39}$$

Remark : *The indicators π_{--}, θ_{--} become spurious when the distribution is unimodal, and may be very difficult to analyse when the distributions are unstable over time. Indeed, in the latter case, lack of persistence in the structure and amplitude of modes would ruin the potential use of such indicators for short-term analysis purposes.*

1. **Test of constancy of the eviction rate between 2003 and 2004, or 2004 and 2005** against the alternative of an increase (or a decrease) of p:

$$\begin{array}{l} \mathbf{H}_0 : p_2\left(\boldsymbol{\theta}_{\mathcal{M}_2}\right) = p_1\left(\boldsymbol{\theta}_{\mathcal{M}_1}\right) \\ \mathbf{H}_a : p_2\left(\boldsymbol{\theta}_{\mathcal{M}_1}\right) \lesseqgtr p_1\left(\boldsymbol{\theta}_{\mathcal{M}_1}\right) \end{array}$$

For loans to NFC, we expect, all things being equal and under hypothesis H2, an increase of p between 2003 and 2004:

$$p_2 > p_1$$

By contrast, if H1 is true, but H2 is not, then we should observe:

$$p_2 \simeq p_1$$

We use a Wald test whose implementation is easy in our context since $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are independent. Let us introduce:

$$\underline{p}\left(\theta_1, \theta_2\right) = p_2 - p_1$$

The test statistic is:

$$\mathcal{W}_n = n\left\{\underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)\right\}^2 \times \left(\sum_{k=1}^{2} \frac{\partial}{\partial \theta_k^{'}} \underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right) \widehat{V}_{\text{asym}}\left(\widehat{\theta}_k\right) \frac{\partial}{\partial \theta_k} \underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)\right)^{-1} \tag{40}$$

under $\mathbf{H}_0$, $\mathcal{W}_n$ has the usual $\chi_2\left(1\right)$ asymptotic.

Remark : *n is defined without any ambiguity since the size of our samples is fixed for each quarter (see section 5 below).*

This test can be easily cast in the subsampling framework, since the subsamples provide an alternative way to estimate the variance of $\underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)$: all we need is $\underline{p}\left(\theta_{b1}^*, \theta_{b2}^*\right)$ for each couple of sub-samples $(\widetilde{s}_{b1}, \widetilde{s}_{b2})$, for $b = 1, \ldots, B$. Next:

$$\widehat{V}_{\text{asym}}\left[\underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)\right] = V_{\text{emp}}\left\{\sqrt{n_b}\left[\underline{p}\left(\theta_{b1}^*, \theta_{b2}^*\right) - \underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)\right]\right\} \tag{41}$$

It remains to replace the expression between brackets in (40) by (41) to obtain the modified test statistic:

$$\mathcal{W}_N^* = n \times \underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)^2 \times \left(\widehat{V}_{\text{asym}}\left[\underline{p}\left(\widehat{\theta}_1, \widehat{\theta}_2\right)\right]\right)^{-1} \tag{42}$$

5 Selection of the data

We discuss in this section some theoretical and practical problems closely connected to the huge amount of information at our disposal for estimation purposes. Paradoxically, a potentially severe problem comes from the large number of observations which appears to exceed the capacities of the econometric routines available in our IT system : for each category, 2000 observations appear to be an upward limit for the quantity of information we are able to process within reasonable time execution constraints. It means that we have to work with a subsample denoted by $\widetilde{s}$ extracted from our database $s = (x_i)_{1 \leq i \leq n}$. It is important to note that this subsample should be representative of the whole population of loans granted during the reference period. Of course, although the whole population is not observed, it can be extrapolated from s through the

sampling weights P_i. Consequently, it is easily seen that our subsample $\widetilde{s}$ of size $\widetilde{n}$ should be drawn from s with replacement according to a proportional to size scheme given by the probabilities of inclusion p_i defined as:

$$p_i = \frac{P_i}{\sum_{i \in s} P_i}$$

We fix $\widetilde{n} = 2000$ for each quarter. It is easy to verify that $\widetilde{s} = (X_1, \dots X_{\widetilde{n}})$ is representative of the global distribution of interest rates. Indeed, let $\overline{X}$ and σ_X^2 be the mean and variance calculated from $\widetilde{s}$. We still denote[14] by x_i the interest rate for the loan i in s. Conditionally on s, the X_i are *i.i.d*, and:

$$\mathbb{E}\left(X_i \,|s\right) = \sum_{i \in s} p_i x_i$$

It yields:

$$\mathbb{E}\left(\overline{X} \,|s\right) = \frac{\sum_{i \in s} P_i x_i}{\sum_{i \in s} P_i} = \widehat{r}$$

$\widehat{r}$ is the standard estimator of the average non-weighted interest rate which is obtained from the 'master' sample s. We get similarly:

$$\sigma_X^2 = \frac{1}{\widetilde{n} - 1} \sum_{j=1}^{\widetilde{n}} \left(X_j - \overline{X}\right)^2$$

Standard manipulations yield finally:

$$\mathbb{E}\left(\sigma_X^2 \,|s\right) = \sum_{i \in s} p_i \left(x_i - \widehat{r}\right)^2 = \frac{\sum_{i \in s} P_i \left(x_i - \widehat{r}\right)^2}{\sum_{i \in s} P_i}$$

We find, as expected, the empirical variance of the interest rate x in the population of loans, estimated from the sample s. Therefore, we can conclude that, on average, the sample $\widetilde{s}$ mimics some basic features of this population.

5.1 Use of replications

Our methodology makes the results strongly dependent on the particular drawn subsample $\widetilde{s}$. We propose to avoid this drawback by averaging the results across a significant number H of independent sub-samples, in the spirit of Monte-Carlo techniques: this should significantly improve the accuracy of our estimator. However, the estimation of the key parameters $\mathcal{M}$ and λ will be kept outside the loop, due to computational constraints: we impose the values obtained from $\widetilde{s}^{(0)} = \widetilde{s}$. The different steps of the estimation methods can now be described as follows:

1. For $h = 0, \dots, H$, draw $H + 1$ independent sub-samples $\widetilde{s}^{(h)}$ from s.
2. With $\widetilde{s}^{(0)}$: perform the tests presented in sections 4.3 and 4.5 and get the (final) estimates of $\widehat{\mathcal{M}}$ and $\widehat{\lambda}$.

[14] Lower cases always refer to the "master" sample s.

3. For $h = 1, \ldots, H$, estimate the model constrained by $\widehat{\mathcal{M}}$ and $\widehat{\lambda}$ and obtain the parameters $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)}$ and the indicators $\mathcal{M}_-$ and $p, ...$ described §4.6.

4. Get the final Monte-Carlo estimates of the parameters (including p) and their asymptotic variances:

$$\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc} = \frac{1}{H}\sum_{h=1}^{H}\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)} \tag{43}$$

$$\text{var}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc}\right) = \frac{1}{H^2}\sum_{h=1}^{H}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)} - \widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc}\right)\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)} - \widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc}\right)' \tag{44}$$

We now justify briefly these results. For each h, we have:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)} - \boldsymbol{\theta}_{\mathcal{M}}\right) \Longrightarrow Y_h \tag{45}$$

with

$$Y_h \rightsquigarrow \mathcal{N}\left(0, \Sigma_{\boldsymbol{\theta}}\right) \tag{46}$$

Because the sample $s^{(h)}$ are independent, the $\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)}$ inherit this property, and the random variables Y_h are i.i.d. Therefore:

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc} - \boldsymbol{\theta}_{\mathcal{M}}\right) \underset{n\to+\infty}{\Longrightarrow} \frac{\sum_{h=1}^{H} Y_h}{H} = Y \tag{47}$$

with:

$$Y \rightsquigarrow \mathcal{N}\left(0, \frac{\Sigma_{\boldsymbol{\theta}}}{H}\right)$$

Finally:

$$\sqrt{nH}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc} - \boldsymbol{\theta}_{\mathcal{M}}\right) \underset{n\to+\infty}{\Longrightarrow} \mathcal{N}\left(0, \Sigma_{\boldsymbol{\theta}}\right) \tag{48}$$

Until this point, we just need $n \to +\infty$. On the other hand, the estimation of the asymptotic variance $\Sigma_{\boldsymbol{\theta}}$ imposes $H \to +\infty$ after $n \to +\infty$ (sequential limit as defined by Phillips & Moon (1999), see appendix 9.4 for more details). We get:

$$\widehat{\Sigma}_{\boldsymbol{\theta}} = \frac{n}{H}\sum_{h=1}^{H}\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)} - \widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc}\right)\left(\widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{(h)} - \widehat{\boldsymbol{\theta}}_{\mathcal{M}}^{mc}\right)' \tag{49}$$

It yields the effective variance given in (44).

5. Final estimation of the eviction rate:

$$p^{mc} = \frac{1}{H}\sum_{h=1}^{H} p^{(h)} \text{ and } \text{var}\left(p^{mc}\right) = \frac{1}{H^2}\sum_{h=1}^{H}\left(p^{(h)} - p^{mc}\right)^2 \tag{50}$$

6. Test of constancy of p : we use p_1^{mc} and p_2^{mc}, then:

$$\mathcal{W}_N^* = nH \times \left(p_1^{mc} - p_2^{mc}\right)^2 \times \left(\widehat{V}_{\text{asym}}\left[p_1^{mc} - p_2^{mc}\right]\right)^{-1} \tag{51}$$

$$\widehat{V}_{\text{asym}}\left[p_1^{mc} - p_2^{mc}\right] = nH \times \left\{\text{var}\left(p_1^{mc}\right) + \text{var}\left(p_2^{mc}\right)\right\} \tag{52}$$

5.2 Introduction of discrete components

The mixture model is designed for continuous distributions. However, a significant number of observations in s, not necessarily belonging to a single bank, appear to be strictly equal. This is especially the case for consuming loans issued by specialized financial institutions, for which the setting of prices appears to be standardized. The existence of clusters is then amplified by the sampling weights P_i used for the drawing of $\widetilde{s}$ from s.

Hence, we enlarge slightly model (10) in the following way:

$$f(x) = \sum_{k=1}^{\mathcal{M}} \pi_k \varphi_k(x) + \sum_{k=\mathcal{M}+1}^{\mathcal{M}+\mathcal{D}} \pi_k \delta_{x_k}(x) \tag{53}$$

$\sum_{k=1}^{\mathcal{M}+\mathcal{D}} \pi_k = 1$ and $\delta_{x_k}(x)$ is the Dirac mass at x_k.

The estimation of the parameters of the discrete part $\{\pi_k, x_k; k = \mathcal{M}+1, \ldots, \mathcal{M}+\mathcal{D}\}$ is done as follows:

1. Identification of interest rate x satisfying:

$$p(x) = \frac{\sum_{\{i \in s \text{ and } x_i = x\}} P_i}{\sum_{i \in s} P_i} > 0,05$$

   The threshold 5% is arbitrary : we seek only to identify the main clusters. From this preliminary search, we obtain D, the x_k end the empirical frequency π_k, for $k = \mathcal{M}+1, \ldots, \mathcal{M}+\mathcal{D}$:

$$\pi_k = p(x_k)$$

   Conditionally on s, these estimators have zero variance.

2. Extract from s all the data corresponding to the x_k ($\sim$ 50000 lines): we get a new sample, s_d ($\sim$ 300000 lines). Define the new sampling probabilities p_i for each loan in s_d.

3. Draw the sample $\widetilde{s}$ from s_d, and use $\widetilde{s}$ to estimate model (10):

$$f_d(x) = \sum_{k=1}^{\mathcal{M}} \pi_{d,k} \varphi_k(x) \tag{54}$$

4. Calculate the final probabilities π_k for $k \leq \mathcal{M}$ associated to the continuous part of the distribution:

$$\widehat{\pi}_k = \widehat{\pi}_{d,k} \left(1 - \sum_{j=\mathcal{M}+1}^{\mathcal{M}+\mathcal{D}} \pi_j \right) \tag{55}$$

5. Finally, the discrete components can be seen as limit cases of continuous distributions : it suffices to note that $\delta_{x_k}(x) \approx N(x_k, \varepsilon)$ with (say) $\varepsilon = 10^{-5}$. With this convention, we don't need any longer to distinguish continuous and discrete components for the calculation of indicators : we shall consider that the distribution is continuous, with $\mathcal{M}+\mathcal{D}$ regimes.

For each reference period, the clusters identified through this filter are given in the tables below. We also indicate the corresponding values of the usury rate. Some (but not all) clusters are very close to the usury rate, which gives for the corresponding categories some credit to our hypothesis H2. In 2005, thanks to the new usury legislation, we observe that one cluster is higher than the usury rate.

Category	1	1	1	2	2	7	7	7	10	10	11	11	11
x_k	**20,94**	18,0057	14,5	**16,6**	**16,81**	3,447	3,439	2,858	8,6	6,6	2,15	7,1	2,13
π_k	**0,08**	0,08	0,05	**0,12**	**0,05**	0,07	0,05	0,05	0,06	0,05	0,14	0,10	0,06
r_u	**21,25**	21,25	21,25	**16,84**	**16,84**	8,72	8,72	8,72	11,19	11,19	8,73	8,73	8,73

Table 6a : October 2003

Category	1	1	1	2	2	7	7	10	10	11	11
x_k	15	18,0057	**20,06**	**16,16**	**16,18**	3,70	3,42	**11,27**	10,1	3,15	4,58
π_k	0,11	0,07	**0,07**	**0,14**	**0,05**	0,07	0,05	**0,17**	0,07	0,07	0,06
r_u	20,13	20,13	**20,13**	**16,21**	**16,21**	8,55	8,55	**11,27**	11,27	8,2	8,2

Table 6b : October 2004

Category	1	1	1	2	7	7	7	10	10	10	10
x_k	15	18,43	18,93	16,53	3,66	3,96	**9,35**	6,9	9,1	10,1	**11,55**
π_k	0,06	0,08	0,08	0,12	0,07	0,08	**0,05**	0,05	0,06	0,07	**0,07**
r_u	19,76	19,76	19,76	17,44	8,99	8,99	**8,99**	11,55	11,55	11,55	**11,55**

Table 6c : October 2005

5.3 How informative are aggregated data ?

A direct and simple way to handle the various problems raised in the previous section (execution time, granularities in the distribution) could be to resort to aggregated data at the bank level. However, the loss of information might be severe, especially if the within banks variance is high. In this case, average interest rates can't be fully representative of the individual distribution. We will deal with this issue through an estimation of the mixture model on aggregated data, according to the same methodology than that we used for our individual database. By doing so, we hope to obtain a fair comparison between the two distributions, and notably the eviction rates.

Thanks to the reduced number of observations in each category (between 80 and 100), we set $M_{\max} = 4$ and use exclusively the combined approach described section 4.4 (subsampling techniques are prohibited with such small samples). The average interest rate are aggregated by weighting them with the related amount of the new contracts, thus making them consistent with the MIR methodology and providing an interesting connection with macroeconomic indicators used for monetary policy purposes.

6 Empirical study

The eleven categories of credits are supplemented by two additional ones. Firstly, we consider in cat. 12 leasing (movables goods) for NFC. It is often acknowledge that lease and debt are substitutes, or that leasing is used as a last resort soluton to increase the debt capacity of the firm. Indeed, leasing is not subject to the

usury regulation. Therefore, relaxing the usury law should help firms with risky projects to obtain classical loans instead of setting leasing agreement.

Secondly, we consider in cat.13, 2000 simulated observations generated according to the following distribution:

$$F(x) = \sum_{k=1}^{4} \pi_k \Phi\left(\frac{x - m_k}{\sigma_k}\right)$$

We use the numerical values:

$$\begin{cases} (\pi_1, m_1, \sigma_1^2) = (0,05 \quad 10 \quad 0,1) \\ (\pi_2, m_2, \sigma_2^2) = (0,2 \quad 5 \quad 0,5) \\ (\pi_3, m_3, \sigma_3^2) = (0,3 \quad 8 \quad 0,4) \\ (\pi_4, m_4, \sigma_4^2) = (0,45 \quad 2 \quad 1) \end{cases}$$

Through this simulated variable, we want to verify if the method we used for the estimation of the deep parameters λ and $\mathcal{M}$ provides sound results. Obviously, this is a very limited exercise which is included in this work only for illustration purpose[15].

Detailed results (estimations and standard errors) are given in appendix 9.5. In this section, we give a synthesis of the main results, including the indicators introduced in §4.6. The distributions are gathered in the appendix, section 9.6 (households) and 9.7 (non-financial corporations). For each category, the graphics are ordered as follows : 2003 (top), 2004 (middle), 2005 (bottom). We represent (solid line) the density $\widehat{f}(x)$ estimated from the mixture through (10), and extrapolated beyond the usury rate (broken line) when the usury law applies to the category. In 2004 and 2005, for loans to NFC (other than bank overdrafts), we observe interest rates higher than the usury rate, so we use a solid line for all values of x. For the sake of comparability, we also represent the non-parametric estimator $\widehat{f}_n(x)$ (3) obtained from the "master" sample $s^{(0)}$ used for the estimation of $\mathcal{M}$ and λ. Significant divergences between $\widehat{f}_n(x)$ and $\widehat{f}(x)$ may indicate that the particular sample $s^{(0)}$ was in fact not fully representative of the distribution. Finally, the vertical solid (resp. broken) lines identify the modes obtained from the continuous (resp. discrete) part of $\widehat{f}$. The dashed area identifies the values of x higher than the usury rate.

N.B : in order to allow a fair comparison between the curves, we maintain the same scales for both horizontal and vertical axis. Consequently, several peaks corresponding to small values of σ_k have been truncated.

6.1 The structural parameters

We present in tables 7a/7b the estimates of the discrete parameters[16] λ and $\mathcal{M}$ for the individual and the aggregated distributions. For the former, the number of modes arising from the continuous part of the distribution [17] is given in tables *7c/d* below, with, within brackets, the 90% confidence interval obtained from the Monte-Carlo exercise, as explained in section 5.1.

[15] A complete simulation experiment would represent a quite impossible task in out current IT. .
[16] λ = N for the model in level, λ = L for the model in log.
[17] The clusters are not taken into account in this statistic.

Category	2003		2004		2005	
	λ	$\mathcal{M}$	λ	$\mathcal{M}$	λ	$\mathcal{M}$
1	L	2	N	1	N	2
2	N	2	N	2	L	2
3	N	1	N	1	N	1
4	L	2	L	2	L	2
5	L	1	N	1	N	1
6	N	2	N	2	N	1
7	L	3	L	2	N	4
8	L	1	N	2	N	1
9	N	1	L	1	L	2
10	N	1	L	3	N	2
11	N	1	L	1	L	1
12	N	1	N	1	L	1
13	N	1	N	1	N	1

Table 7a : aggregated data

Category	2003		2004		2005	
	λ	$\mathcal{M}$	λ	$\mathcal{M}$	λ	$\mathcal{M}$
1	N	6	N	8	N	7
2	N	7	N	7	N	7
3	N	6	N	6	N	6
4	L	2	N	3	L	2
5	N	3	N	2	L	2
6	L	2	N	4	N	2
7	N	7	N	6	N	6
8	N	4	L	4	L	5
9	N	6	N	2	L	3
10	N	7	N	6	N	5
11	N	5	L	5	N	5
12	L	4	L	3	L	3
13	N	4	N	4	N	4

Table 7b : individual data

.

Category	1	2	3	4	5	6
$\mathcal{M}_-$ [2003]	5	6	4	1	2	2
	[5,6]	[4,7]	[3,5]	[1,1]	[1,2]	[1,2]
$\mathcal{M}_-$ [2004]	8	5	5	1	1	3
	[7,8]	[5,7]	[4,5]	[1,2]	[1,1]	[1,3]
$\mathcal{M}_-$ [2005]	5	6	4	1	1	2
	[5,6]	[5,6]	[4,5]	[1,2]	[1,2]	[2,2]

Table 7c : Households : number of modes (ind.)

Category	7	8	9	10	11	12	13
$\mathcal{M}_-$ [2003]	6	4	4	5	4	2	4
	[6,7]	[2,4]	[4,5]	[4,5]	[3,4]	[2,2]	[4,4]
$\mathcal{M}_-$ [2004]	4	2	1	3	4	2	4
	[4,5]	[2,3]	[1,1]	[3,4]	[3,4]	[2,2]	[4,4]
$\mathcal{M}_-$ [2005]	5	3	3	3	4	2	4
	[5,5]	[2,4]	[3,3]	[3,4]	[3,4]	[2,2]	[4,4]

Table 7d : NFC : number of modes (ind.)

Some interesting facts emerge from these results:

1. According to table *7b*, one needs at least 6 régimes for the distributions of consuming loans. Similarly, the value of $\mathcal{M}$ appears to be high, but to a lesser extent, for loans to NFC. As expected, housing loans admit more parsimonious representations (2 or 3 regimes). Most of the distributions appear to be in level, and λ is constant over time for only 7 categories. Lastly, although the number of regimes is not strictly constant, it appears to fluctuate moderately: more specifically, for loans to NFC, the relaxing of the usury regulation has not been followed by a systematic decrease - or increase - of the number of regimes, except for the of cat. 9 (loans at fixed rate with duration up to 2 years) between 2003 and 2004.

2. The results in tables *7c/7d* show that the number of modes is generally smaller than $\mathcal{M}$ and almost constant, except for loans to NFC : for these categories, we observe a decrease from 2003 to 2004/2005.

3. For the data aggregated at bank level, table *7a* indicates that the proportion of models in logarithm is larger. The variability observed in table *7b/c* (high values for the number of regimes/modes) is much less pronounced here, except for two categories of loans to NFC : credit instalment (cat.7) and to a lesser extent, bank overdrafts (cat. 10). In the other cases, the distributions are modelled with one or two regimes[18] : we will come back to this issue in § 6.2.3.

Turning now to the estimated p.d.f for individual data, we notice the following facts:

1. Results for the simulated variable (cat. 13) are rather promising because the procedure succeeded in estimating the true values of $\mathcal{M} = 4$ and $\lambda = \mathbf{N}$. Moreover, the estimated distribution is very close to its theoretical counterpart, especially in the upper side where there exists a mode with a very small value of π ($\pi_1 = 0,05$). Interestingly, the kernel estimator is too smooth, and doesn't discriminate correctly the two largest modes.

2. For leasing agreements with NFC (cat. 12), the distributions appear quite stable over time, with two well identified modes. Moreover, the largest mode seems less important in 2004 and 2005 than in 2003, a finding which could confirm a return of firms with risky projects to classical loans.

[18] In passing, we note that these results indicate that some of these aggregated distributions are normal or log-normal, although this is not true for the individual distributions.

3. In spite of a relative stability of the number of modes, the shape of the curves are often very different between the three reference months, with the notable exception of housing loans and leasing. This evolutionary feature is also present in non-parametric estimators, although to a lesser extent. This could result from the intrinsic variability of interest rate in each strata. In other words, the categories of loans are heterogeneous, and the distributions change over time according to complex structural effects. On the contrary, housing loans and leasing for NFC seem to constitute more homogeneous categories, so that macroeconomic or statistical analysis of these data is not misleading.

4. Generally, non-parametric curves follow closely the parametric curves although they appear to be smoother, as could be expected. However, we observe in a few cases some very important discrepancies which indicate probably a strong dependence of the results on the particular sample $s^{(0)}$ used for the estimation of $\mathcal{M}$. This drawback could not be avoided, since it is a limit inherent to our methodology.

5. For all categories of housing loans (cat. 4 to 6), the distributions are basically unimodal, excepted for loans at floating rate in 2003 for which a mode is estimated in the bottom of the distribution[19]. Interestingly, the split of loans by maturity as defined in MIR reports does not translate into modes in our distributions[20]. Lastly, it is clear that the usury rate doesn't distort the distribution, except for bridging loans in 2004. On the whole, we can accept hypothesis H_1.

6. Consuming loans (cat. 1 to 3) tell another story, with several concurrently features revealed by the distributions : granular behavior (cat. 1 and 2), important distortion near the usury rate (particularly for cat. 3). In addition, for cat. 1 and 2, the modes estimated from the continuous part are very close to some clusters of the discrete part.

5. With regard to loans to NFC, we notice for cat. 8 the disappearance of a mode near the usury rate in 2004 and 2005, and simultaneously the appearance of a new mode whose magnitude is limited, larger than the usury rate (in our notations, $\mathcal{M}_{\_} = \mathcal{M} - 1$). For cat. 9, we observe a similar property in 2004 compared to 2003, but the distribution obtained for 2005 indicates again a mode near the usury rate. Finally, cat. 11 shows a mode in the upper side of the distribution for all reference months, but it is distant from the usury rate : hypothesis H_1 seems to provide a good description of the data.

It is worth looking at the results we get from the alternative methods we had initially in mind (see § 4.3.1 and 4.3.2). The results are given in appendix 9.9 for October 2003 and 2004. As indicated previously, the maximum number of regimes is set to $\mathcal{M}_{\max} = 5$, but in practice, especially for models in logarithm, we have reduced this value to $\mathcal{M}_{\max} = 4$ or 3 in order to get rid of unreliable results. Implicitly, this simplification of the procedure is biased in favour of models in level, for which numerical problems are less

[19] It would be interesting here to verify if this mode arises because of one credit institution trying to win market shares.
[20] Nevertheless, one particular regime could be associated to a particular range of maturities.

frequent. In passing, we notice that the combined approach is a particular elegant way to avoid numerical problems through the annihilation of non-significant components.

The overall results for $\mathcal{M}$ and λ show that both alternative methods provide similar results to the combined approach (see table 7b), with a tendency to estimate lower values of $\mathcal{M}$, whereas for λ, the results are remarkably close. These methods seems promising, under the stringent condition that the sample is rich enough, and that the number of regimes is reasonable.

Lastly, the combined approach provides as a by-product a set of estimates of the eviction rate p for different values of $\mathcal{M}$ and λ. In table 7e below, we give the minimum and maximum values of p over this set, for two reference months (Oct. 2003/2005). Estimations corresponding to $\mathcal{M} = 1$ are discarded, thanks to results of section 3.2.

Category	1	2	3	4	5	6	7	8	9	10	11
p_{min} [2003]	0,5	3,5	5,0	1,1	0,9	0,3	1,4	2,4	4,3	13,8	1,9
p_{max} [2003]	2,4	5,5	8,2	1,3	1,3	1,0	5,2	4,1	5,8	16,8	3,1
p_{min} [2005]	0,1	0,0	1,9	1,5	0,5	1,1	16,6	3,6	9,9	8,0	2,8
p_{max} [2005]	3,9	6,2	9,1	2,0	0,7	2,1	18,9	5,4	15,9	14,9	8,5

Table 7e : Eviction rate

For consuming loans and loans to NFC, the range of values of p is large. This justifies ex-post the need for a careful description of the distribution of interest rates. For housing loans, the conclusion is somewhat different because of a smaller number of regimes available in our final estimates making the range of values of p much more narrow.

6.2 Detailed results

6.2.1 Loans to households

We recall briefly the meaning of the statistics given in table *8a* below:

- $r_{\cdot}^{u}$: usury rate, and p : eviction rate; within brackets : the 90% confidence interval for p resulting from the Monte-Carlo simulations (see §5.1).
- θ_{--} (%) : relative share (weighted by amount of new contracts) of loans associated to the largest mode of the distribution (including the discrete part). When several modes are very close to the largest one, they are "aggregated" together for the calculation of θ_{--}. When the distribution is close to unimodality, $\theta_{--} \approx 100\%$ and we do not report the value.
- π_{--} : probability of the largest modes used for determination of θ_{--} (relative share, without any weighting)
- Test of stability of the eviction rates (42) between 2003 and 2004 : the p-value are provided in table *8b*

t	Cat	r^u	$p\,(\%)$	$\theta_{--}\,(\%)$	π_{--}
2003/10	*1*	21,25	1,1 [0,4-1,0]	12,8	0,08
	2	16,84	4,6 [4,3-5,4]	17,9	0,27
	3	9,96	5,0 [3,4-7,6]	3,3	0,05
	4(*)	6,88	1,3 [1,1-1,7]	-	-
	5(*)	6,4	1,2 [1,3-2,1]	-	-
	6(*)	7,12	0,3 [0,3-0,9]	-	-
2004/10	*1*	20,13	0,2 [0,2-0,4]	19,2	0,14
	2	16,21	3,8 [3,5-4,8]	16,7	0,25
	3	9,12	4,7 [2,6-3,9]	4,0	0,06
	4(*)	6,56	1,4 [1,4-2,2]	-	-
	5(*)	5,85	0,5 [0,4-0,6]	-	-
	6(*)	6,68	1,4 [0,4-1,3]	-	-
2005/10	*1*	19,76	0,1 [0,1-0,4]	6,1	0,05
	2	17,44	2,6 [1,9-3,8]	1,3	0,05
	3	8,33	3,8 [3,0-5,0]	5,5	0,07
	4(*)	5,87	2,0 [1,7-2,5]	-	-
	5(*)	5,48	0,5 [0,7-1,3]	-	-
	6(*)	5,72	1,1 [0,9-1,3]	-	-

Table 8a

Category	1	2	3	4	5	6
$\mathcal{W}_n^*$	220,82	107,78	138,67	102,69	1251,57	32,97
p-value	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$

Table 8b

The eviction rates range from 3% to 5% for categories 2 and 3, and appear stable around 1% to 2% for housing loans. For category 1, the eviction rate is particularly law, smaller than 1% in 2004 and 2005. Moreover, the unweigthed share of the most expensive loans (column π_{--}) is around 10% for consuming loans (with a maximum of 15-20% attained for cat. 2 in 2003 and 2004).

Between 2003 and 2005, the eviction rates decrease steadily for consuming loans, as witnessed by the confidence intervals. Despite of the low values of interest rates (levels and volatilities), the usury rate seems to have played a role of secondary importance, probably because of the competitive interactions which took place among lenders during the period. In other words, the distribution moves to lower values faster than the usury rate. The constancy of p is always largely rejected by the formal test, a result which is not surprising since the Monte-Carlo estimates have mechanically very low variances (see 50). However, the conclusions are supported[21] by a non-parametric rank test of Wilcoxon[22] directly based on the Monte-Carlo distributions.

[21] The results are not reported here, but are available from the author upon request.

[22] The null hypothesis is the constancy of the p.d.f of p between 2004 and 2003 : $f_2(x) = f_1(x)$. Under the alternative, there is a shift between the two densities : $f_2(x) = f_1(x - \delta)$. If $\delta > 0$, the distribution is shifted to the right in 2004, from which it follows that p is likely higher in 2004 than in 2003.

Looking at market shares of consuming loans, we notice that weighted indicators (column θ_{--}) are smaller than their unweigthed counterparts (column π_{--}) for cat. 2 and 3 : this result is in line with the well-documented correlation between interest rates and amount of loans. However, this empirical finding doesn't hold for small loans : indeed, $\theta_{--} > \pi_{--}$ for all reference months.

At first sight, this result may seem puzzling because it is expected that small loans are primarily concerned by the threshold defined by the usury rate. A more refined analyse based on a specific examination of the sub-category of loans up to 500€ doesn't support this intuition : these loans are in fact not concentrated in the top of the distribution, except perhaps for instalment credits.

We conclude this analysis with the distribution for housing loans taken as a whole. This posterior p.d.f. is estimated from Bayes' rule (f_k is the distribution estimated previously for category k), and plotted fig. 9. Its expression is given by:

$$f(x) = \sum_{k=4}^{6} \mathbb{P}(\text{loan} \in \text{cat}.k)\, f_k(x) \tag{56}$$

The market share $\mathbb{P}(\text{loan} \in \text{cat}.k)$ is estimated by its empirical counterpart, using all the data available in our master sample s, *i.e*, not only the dataset used for the estimation of (f_k):

$$q_k = \frac{\sum_{i \in \text{cat}.k} P_i}{\sum_{i \in \text{cat}.\{4,5,6\}} P_i} \tag{57}$$

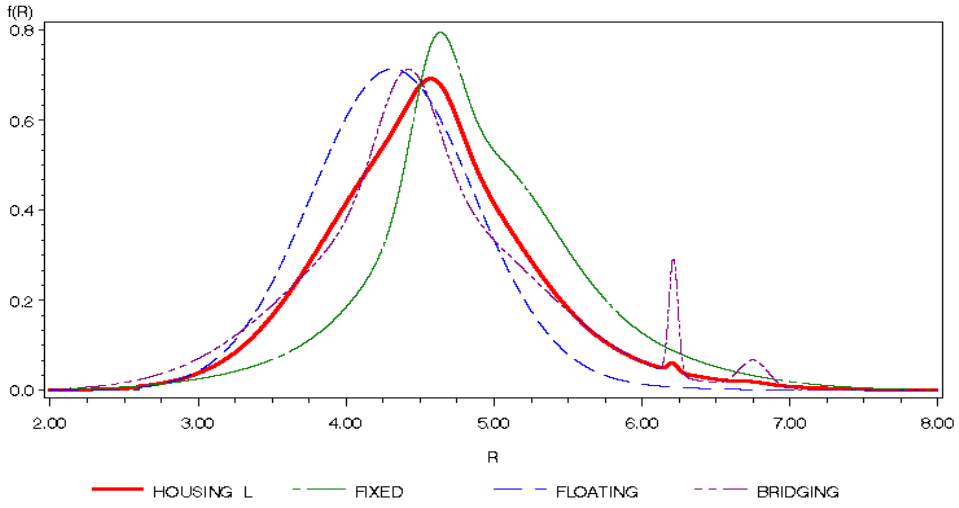


Fig. 9 : housing loans

The aggregated distribution (solid line) remains unimodal, because the distance between the 'floating rate' curve and the 'fixed rate' one is moderate. However, the distribution appears skewed, and its tails are rather thick. These departures from normality confirm that interest rates spread over a wide range of values on either side of the modal value.

6.2.2 Loans to NFC

t	Cat	r^u	$p\,(\%)$	$\theta_{--}\,(\%)$	π_{--}
2003/10	*7*	8,72	1,4 [1,4-3,1]	5,2	0,05
	8	6,49	2,4 [3,3-4,1]	15,7	22,6
	9	6,79	4,6 [4,4-5,5]	1,4	0,02
	10	11,19	13,8 [13,6-19,6]	0,7	0,06
	11	8,73	2,2 [1,9-2,5]	16,1	0,11
	12	-	-	22,8	0,27
2004/10	*7*	8,55	11,2 [16,4-18,9]	43,8	0,39
	8	6,08	4,0 [2,8-4,1]	-	-
	9	6,47	7,8 [7,4-8,8]	-	-
	10	11,27	24,1 [11,6-26,0]	4,1	0,17
	11	8,2	4,4 [3,6-4,7]	20,7	0,18
	12	-	-	21,0	0,36
2005/10	*7*	8,99	16,6 [17,0-20,0]	26,4	0,27
	8	5,52	4,5 [4,1-5,5]	-	-
	9	6,01	9,9 [9,1-11,8]	18,0	0,22
	10	11,55	11,1 [7,3-13,3]	1,1	0,18
	11	7,75	5,6 [4,4-5,8]	23,9	0,23
	12	-	-	11,4	0,25

Table 10a

Category	7	8	9	10	11
$\mathcal{W}_n^*$	15197,15	5,90	2327,73	149,45	1574,37
p-value	$<10^{-6}$	0,02	$<10^{-6}$	$<10^{-6}$	$<10^{-6}$

Table 10b

1. From table *10a*, the main finding is the increase of the (pseudo) eviction rates between 2003 and 2005 across almost all categories of loans concerned by the new regulation (cat. 8,9 and 11, as well as cat. 7 between 2003 and 2004). This pattern is confirmed by the massive rejection of the constancy test as reported table *10b* (the same conclusion holds when a Wilcoxon test is used). This result indicates a possible normalization of the distributions, with a reallocation of loans beyond the usury rates. However, this result tells us nothing about a possible increase in the volume of new contracts granted by banks due to the relaxing of the usury law.

2. For leasing, the strong decrease of θ_{--} between 2003 and 2005 confirm the visual inspection of the distribution : the largest mode appears less significant in the recent period, a fact in line with the emergence of loans with high interest rates in the other categories. Otherwise, for cat. 9 and 11, θ_{--} has increased strongly between 2003 and 2005, a fact which indicates that there is still many loans with interest rates near the usury rate.

6.2.3 Individual vs aggregated

Given the strong divergence between the number of components estimated from individual and aggregated data, it is not surprising to observe quite different distributions for consuming loans and loans to NFC. For instance, in oct. 2004, the aggregated distributions are unimodal for cat. 1,3 and 11, whereas the underlying individual distribution is highly multimodal (fig. 11a). Furthermore, these aggregated distributions aren't distorted near the usury rate, which entails that our interpretation of the effect of the usury rate on the distribution of interest rate is contingent upon the level of aggregation. However, we must keep in mind that the correspondence between these distributions is rather complicated:

- Aggregated data are weighted by amount of new business and sampling weights while individual data are only weigthed by sampling weights : therefore, the two theoretical underlying distributions have different means and variances.
- Aggregated distributions are based upon all available data, since the initial sample s is used to calculate interest rate at bank level. By contrast, we use independent samples drawn from s to estimate the parameters in the Monte-Carlo step of the analysis.

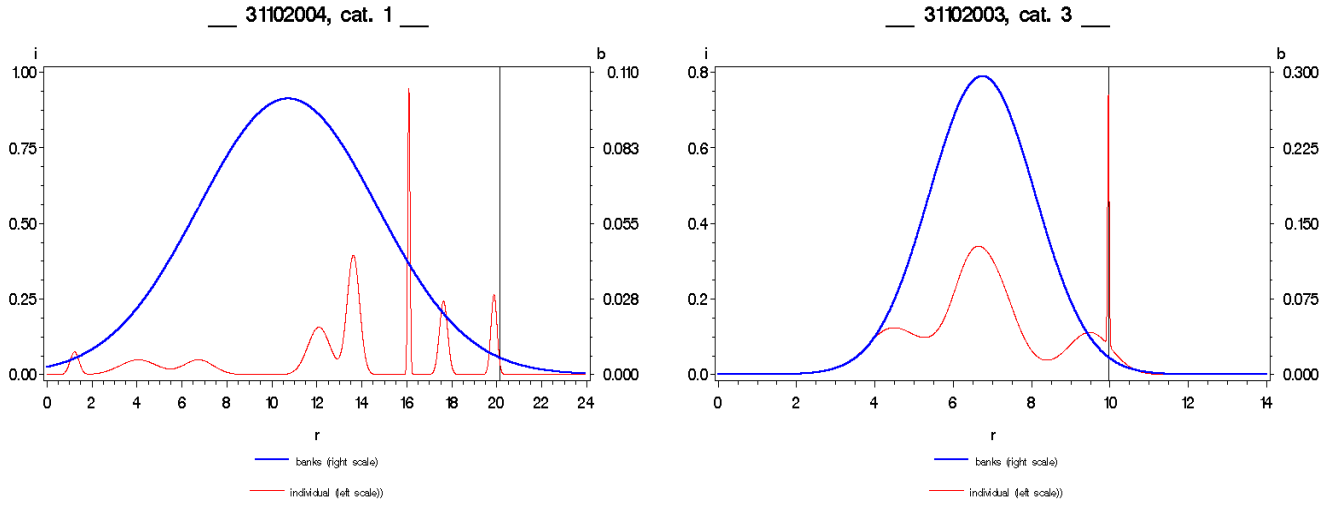


Fig. 11a

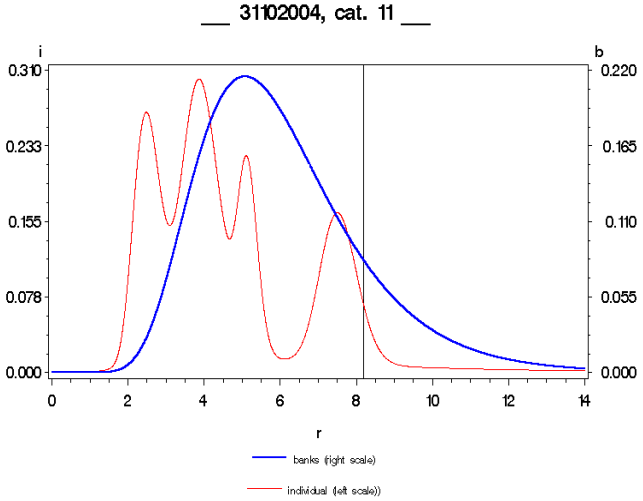


Fig. 11a (cont.)

Sometimes, the curves are very close from each other, and we can conclude that the information provided by the aggregated data are sufficient. For instance, this appears to be the case for cat. 9 (loans at fixed rate with duration over 2 years) in Oct. 2004 (fig. 11b, left). It is worth noting that this property is not structural : it doesn't extend over subsequent reference months, as we can see in Oct. 2005 (fig 11b, right) :

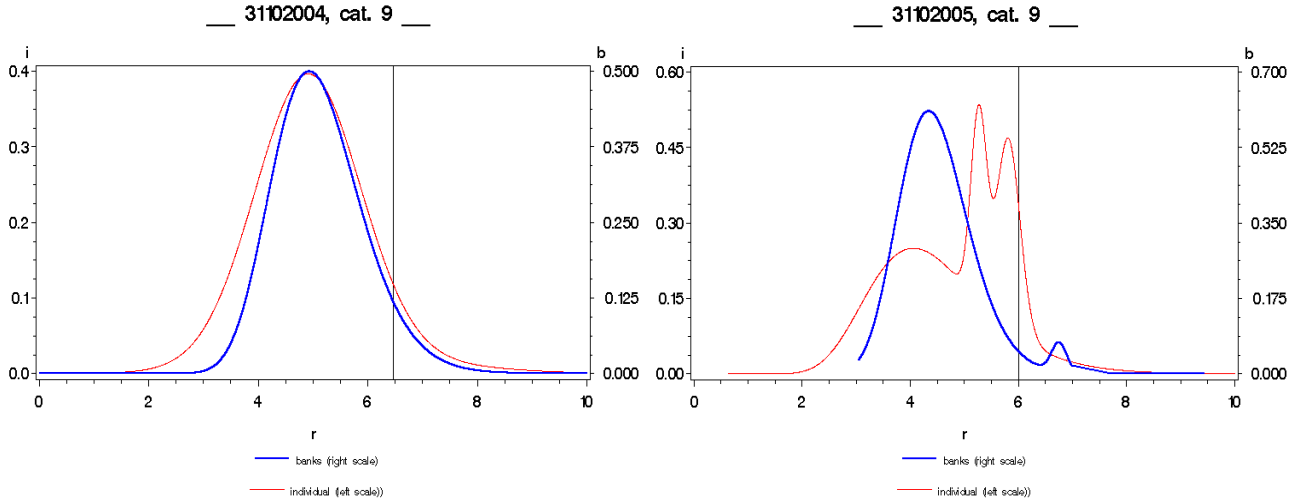


Fig. 11b : cat. 9 (Oct. 2004 & Oct. 2005)

For housing loans, we expect the results to be quite close. However, the aggregation process may induce some minor distortions, which can be observed in fig. 11c below:

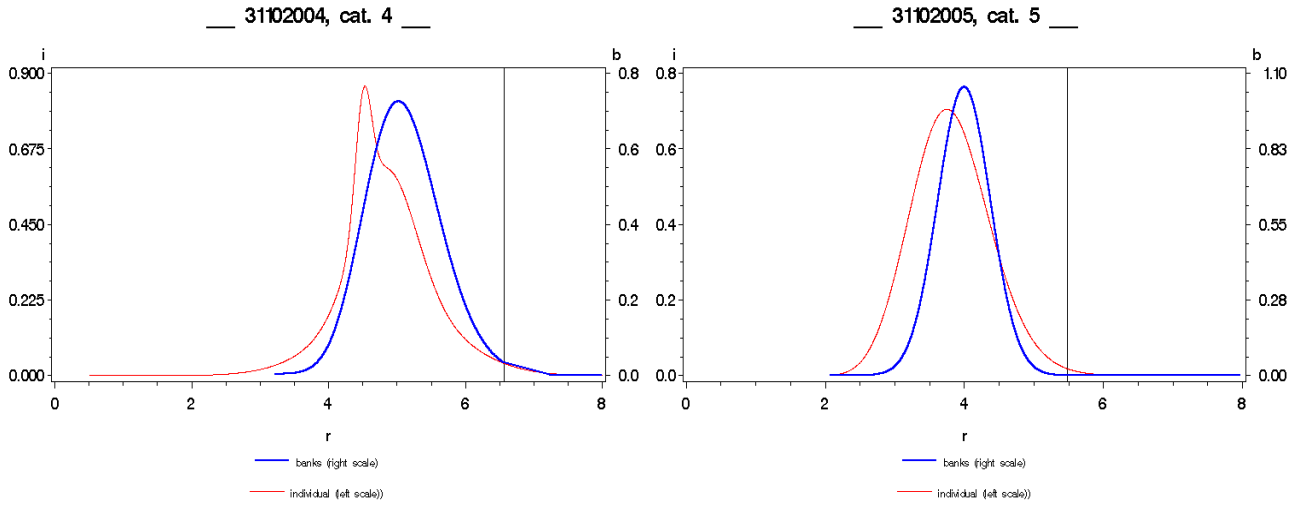


Fig. 11c : cat. 4 (Oct. 2004) & cat.5 (Oct. 2005)

This analysis confirms the empirical findings reported previously in this paper: when the within banks variance is high, individual data convey valuable information, especially for descriptive analysis purposes. More precisely, our methodology allows us to define homogeneous categories, or, in other words, categories for which data collection processes could remained at an aggregate level. Otherwise, the collection of individual data is clearly a more elegant (and efficient) way of dealing with the issues of heterogeneity than the introduction of supplementary breakdowns (per counterpart, duration, instruments) in the report.

We conclude this analysis with an overview of eviction rates measured with the aggregated distributions. The results are reported for 2003 and 2005 in table 11d, and should be compared to the corresponding estimates of p obtained with individual data (see tables *8a* and *10a*).

Category	1	2	3	4	5	6	7	8	9	10	11
p_{agr} [2003]	1,9	1,0	0,9	0,7	0,1	1,2	0,0	1,6	1,3	13,9	7,1
p_{agr} [2005]	0,5	4,4	1,8	2,0	0,0	0,5	3,1	1,2	4,7	11,2	13,9

Table 11d : Eviction rate (aggregated data)

We observe important divergences between the two sets of estimates, as testified by lower values of 'aggregated' estimates. These values are in addition almost always outside the confidence interval for p estimated from individual data. Both the amplitude of the truncation effect (when the usury law is in force), and the reorganization of the distribution (in 2004 and 2005 for loans to NFC) are underestimated. Moreover, p_{agr} and p don't evolve in the same way between 2003 and 2005. For instance, p_{agr} increases for cat. 2, 3 and 4, and decreases for cat. 8. Thus, the results leave no doubt that the interpretation of the distribution of interest rate depends strongly on the level of aggregation.

6.2.4 Heterogeneity of the reporting population

The numerous modes identified in the distributions, especially for consuming loans and loans to NFC, seem to indicate that the market is segmented. In this section, we go one step further with an indirect discussion of this hypothesis. To achieve this aim, the idea is to compare ex-post indicators calculated for subgroups of banks to the results concerning to the whole population of loans. By doing this way, we wish to identify modes associated to particular subsets of the credit institution population. Obviously, more direct methods are already available for this kind of problem. For instance, standard regressions allows to estimate bank effects, all things being equal, but the difficulty relies precisely on a correct specification of the explanatory variables included in the model. By contrast, our approach is model free, and nothing else than purely descriptive. For the sake of clarity, we consider three strata of banks, corresponding to the most simple partition of the MIR reporting population : commercial banks (abbreviated "G"), mutual and cooperative banks, including saving and provident institutions ("M"), and specialized financial institutions ("F"). We first tried to use indicators (36), but the results were disappointing because figures are strongly unstable over time. A better way to handle the problem is to content ourselves with a qualitative interpretation of estimated ex-post distributions for each strata. The resulting curves are plotted for october 2003 and october 2005, and for some selected categories in appendix 9.8.

For any network $R \in \{G, M, F\}$, the posterior *p.d.f.* is:

$$f_R(x) = \sum_{k=1}^{\mathcal{M}} \theta_R(k)\, \varphi_k(x) + \sum_{k=\mathcal{M}+1}^{\mathcal{M}+\mathcal{D}} \theta_R(k)\, \delta_{x_k}(x) \tag{58}$$

The relative share of each component in this mixture is estimated by:

$$\theta_R(k) = \frac{\sum_{i \in R} P_i \times \mathbf{P}(k \,|x_i)}{\sum_{i \in R} P_i} \tag{59}$$

As in (31), the sums in this expression range over all loans granted by banks belonging to network "R", while $\mathbf{P}(j \,|x_i)$ is the ex-post probability (33).

We find varying results, depending on the category of loans under investigation. As was expected, housing loans (cat. 4,5,6) don't show any strong discrimination between networks, although we reported the existence or two or three underlying regimes in the distributions. This fact provides another evidence for the homogeneity of this competitive market. For consumer loans, it is not surprising to observe very different distributions in 2003 and 2005, since this feature showed up already for the global study. However, a few patterns seem to be relatively stable, and hence should be emphasized. Firstly, specialized financial institutions appear clearly in the top of the distribution, especially for loans up to 1524€ (cat. 1) and bank overdrafts over 1524€ (cat. 2). For this category, mutual and cooperative banks seem to apply lower interest rates than commercial banks, a result which reflects a possible specialization in terms of instrument. Lastly, for personal loans over 1524€ (cat. 3), commercial banks and mutual banks can't be clearly discriminated.

Turning now to loans to NFC, we observe that specialized financial institutions seem particularly competitive for loans with duration up to 2 years. It seems also that for this category, commercial banks offer lower interest rates than mutual banks.

For loans with duration over two years at floating rate (cat. 8), the curves are very close : the three networks operate on the submarkets identified by the modes. For loans over 2 years at fixed rate (cat. 8), we observe that generalist banks appear mostly in the top of the distribution in 2005, and to a lesser extent in 2003.

These results are still preliminary, and a careful checking of their robustness should be undertaken. However, it seems that, with the exception of specialized financial institutions for some specific categories, the heterogeneity in the distributions is more related to the instruments included in the categories than to the banking population, at least when we consider a rather crude partition of this population.

7 Concluding remarks

This paper shows how individual informations permit to precise diagnosis based on aggregated data, in particular by allowing the treatment of specific issues dealing with heterogeneity. The work elaborates on finite mixtures distributions which appear to provide a fairly good description of distributions of effective interest rates, despite all the difficulties encountered when practically implementing the method. Now, this work could be pursued along three directions:

At first, the methodology provides as a by-product a typology of loans through the definition of statistical homogeneous groups (in terms of interest rate) defined from the regimes or modes identified in the distribution. In a second step, it would be interesting to analyse more precisely the loans belonging to a given group, for instance by taking into account additional variables on the supply side (characteristics of the bank such as balance sheet data, qualitative information provided by the Bank Lending Survey), and on the customer side : income (for households), size, sector, risk indicator (for NFC).

Secondly, since the comparison between individual and aggregated data sometimes display important discrepancies, we intend to build new aggregated indicators in the spirit of price index methodology. Indeed, summarizing the distribution of interest rates by a single indicator is a difficult task, and moreover, different weighting schemes may provide different interpretations for short term analysis. Estimation of an "interest rate index" means estimating in a first step structural effects deriving from usual determinants of interest rates at the micro level. Then in a second step aggregating these structural effects with constant weights, or weights from the previous quarter. This methodology could enable us to disentangle structural effects driven by the weighting process and short term effects driven by the 'price' component which are the primary concern of short term analysis. From an econometric standpoint, we could allow for flexibility for the estimation of structural effects at the individual level through a semi-linear specification.

Lastly, more work is to be done about the robustness of the estimations. Our results indicate a lack of persistence in the shape of the distributions from one quarter to another. The tentative explanation given in this paper relies on the heterogeneity of the categories and/or of the banking system. But this instability could also be linked to some weakness in the estimation procedure. For this reason, we should increase the time-series dimension of the exercise, and more importantly, seek for more robust estimators and alternative numerical optimization schemes, such as simulated annealing methods.

8 Bibliography

1. **ANDREWS D.W.K.** (1999) Estimation when a parameter is on a boundary, *Econometrica 67, 1341-1383*

2. **ANDREWS D.W.K.** (2001) Testing when a parameter is on the boundary of the maintained hypothesis, *Econometrica 69, 683-734*

3. **BABU G.J., RAO C.R.** (2004) Goodness-of fit tests when parameters are estimated, *Sankhyã 66(1), 63-74*

4. **BICKEL P.J., RITOV Y., RYDEN T.** (1998) Asymptotic normality of the maximum-likelihood estimator for general hidden markov models, *Annals of Statistics, 26(4), 1614-1635*

5. **BILLINGSLEY P.** (1968) Convergence of probability measures, *Wiley Series in Probability and Statistics, Wiley*

6. **CHENG M.Y., HALL P.** (1999) Mode testing in difficult cases, *Annals of Statistics, 27(4), 1294-1315*

7. **DACUNHA-CASTELLE D., GASSIAT E.** (1999) Testing the order of a model using locally conic parametrization: population mixtures and stationary Arma models, *Annals of Statistics, 27(4), 1178-1209*

8. **FIGUEIREDO M.A.T., LEITAO J.M.N, JAIN A.K.** (1999) On fitting mixture models, *Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer-Verlag*

9. **FIGUEIREDO M.A.T., JAIN A.K.** (2002) Unsupervised learning of finite mixture models, *IEEE Transactions on pattern analysis and machine intelligence, 24(3), 381-396*

10. **GASSIAT E.** (2002) Likelihood ratio inequalities with applications to various mixtures, *Ann. Inst.Henri Poincaré, 38, 897-906*

11. **HAMILTON.** (1991) A quasi-bayesian approach to estimating parameters for mixtures of normal distributions, *Journal of Business & Economic Statistics, 9(1), 27*-39

12. **HANSEN B.E.** (1996) Inference when a nuisance parameter is not identified under the null hypothesis,, *Econometrica, 64(2), 413-430*

13. **HATHAWAY R.J.** (1985) A constrained formulation of maximum-likelihood estimators for normal mixtures, *Annals of Statistics, 13, 795-800*

14. **KERIBIN C.** (2000) Consistent estimators of the order of mixture models, *Sankhya, Series A, vol 62, 49-66*

15. **MC LACHLAN G.J., PEEL D.** (2000) Finite mixture models, *Wiley Series in Probability and Statistics, Wiley*

16. **PHILLIPS P.C.B., MOON H.R.** (1999) Linear regresion limit theory for nonstationary panel data, *Econometrica, 67(5), 1057-1112*

17. **POLITIS D.N., ROMANO J.P.** (1994) Large sample confidence regions based on subsamples under minimal assumptions, *Annals of Statistics, 22, 2031*-2050

18. **POLITIS D.N., ROMANO J.P, WOLF M.** (1999) Subsampling, *Springer Series in Statistics, Springer-Verlag*

19. **PSARADAKIS Z., SPAGNIOLO N.** (2002) On the determination of the number of regimes in markov-switching autoregressive models, *Journal of the American Statistical Association, 83,698-708*

20. **ROMANO J.P.** (1988) A bootstrap revival of some nonparametric distance tests, *Journal of Time-Series Analysis, 24(2), 236-253*

21. **SILVERMAN, B.W.** (1981) Using kernel density estimates to investigate multimodality, *J. Roy. Statis. Soc. Ser. B, 43, 97-99*

22. **SILVERMAN, B.W.** (1986) Density estimation for statistics and data analysis, *Chapman and Hall*

9 Appendix

9.1 Normality tests

We first sketch how to obtain the limit law of the test statistics. Since F is continuous, the change of variables $u = F(x)$ yields:

$$\begin{cases} KS_n = \sqrt{n} \times \sup\limits_{0 \leq u \leq 1} \left| \widehat{F}_N \left(F^{-1}(u) \right) - u \right| \\ AD_n = n \int_0^1 \frac{\left\{ \widehat{F}_n \left(F^{-1}(u) \right) - u \right\}^2}{u(1-u)} du \end{cases}$$

$\overline{F}_n(u) = \widehat{F}_n \left(F^{-1}(u) \right)$ is the empirical distribution function of the sample obtained from the $U_k = F(X_k)$. Under the null hypothesis (truncated normal law for X_k), the U_k follow the uniform law with support on $(0,1)$, and we obtain the convergence in distribution in the functional space $D[0,1]$ (Billingsley (1968)):

$$\sqrt{n} \times \left\{ \overline{F}_n(u) - u \right\} \underset{N \to +\infty}{\Longrightarrow} B_0(u) \tag{60}$$

B_0 is the Brownian bridge on $[0,1]$. The continuous mapping theorem yields the asymptotic laws of KS_N and AD_N:

$$\begin{cases} KS_n \underset{n \to +\infty}{\Longrightarrow} \sup\limits_{t \in [0,1]} |B_0(t)| \\ AD_n \underset{n \to +\infty}{\Longrightarrow} \int_0^1 \frac{B_0^2(t)}{t\{1-t\}} dt \end{cases} \tag{61}$$

The parameters $\left(m, \sigma^2\right)$ being unknown, they are estimated by maximum likelihood. The empirical distribution function $F(x)$ is then replaced by $\widetilde{F}_n(x)$ in (8) and (7):

$$\widetilde{F}_n(x) = \Phi\left(\frac{x - \widehat{m}}{\widehat{\sigma}} \right) \Big/ \Phi\left(\frac{r^u - \widehat{m}}{\widehat{\sigma}} \right) \text{ if } x \leq r^u \tag{62}$$

We then get the feasible statistics $\widehat{KS}_n$ and $\widehat{AD}_n$. Under the null hypothesis, their asymptotic laws are not given by (61); they depend in a complex way of m and σ^2. We then use a bootstrap experiment in order to estimate the quantiles of this law. However, the classical approach (sampling with replacement in the empirical law $\mathbb{P}_n$ of $(X_k)_{1 \leq k \leq n}$) can't be used directly in our context (Babu and Rao (2004)). More precisely, let (X_k^*) be a n-sample $i.i.d$ drawn from $\mathbb{P}_n$, $\widehat{F}_n^*(x)$ its empirical distribution and $\widetilde{F}^*(x)$ the distribution function obtained with $\widehat{m}^*$ and $\widehat{\sigma}^*$ estimated from the (X_k^*). It can be shown that $\sqrt{N}\left\{ \widehat{F}_n^*(x) - \widetilde{F}^*(x) \right\}$ is a biased estimator of $\sqrt{n}\left\{ \widehat{F}_n(x) - \widetilde{F}(x) \right\}$, the bias being $O_P(1)$. Instead of trying to correct the bias, we proceed with the parametric version of the bootstrap which leads to the correct asymptotic distributions (Babu and Rao (2004), Romano (1988)). The basic principle is to draw the sample (X_k^*) from the law $\mathbf{N}\left(\widehat{m}, \widehat{\sigma}^2 \,|{-\infty}, r^u \right)$. The replication of this step through a Monte-Carlo experiment allows in a final step the estimation of the wanted distributions. The whole process can be summarized as follows:

1. Draw B (=2000) simulated independent samples[23] of size n in the law $\mathbf{N}\left(\widehat{m}, \widehat{\sigma}^2 \,|{-\infty}, r^u \right)$

[23] One obtains these samples from the simple device: $X^* = \widehat{m} + \widehat{\sigma}\Phi^{-1}\left\{ U\Phi\left(\frac{r^u - \widehat{m}}{\widehat{\sigma}} \right) \right\}$ with $U \rightsquigarrow U(0,1)$.

2. For each sample $b = 1, \ldots, B$, estimate by maximum likelihood $\widehat{m}_b^*$ and $\widehat{\sigma}_b^{*2}$, then calculate according to (62), (7) and (8), $\widehat{KS}_{n,b}^*$ and $\widehat{AD}_{n,b}^*$ with $\widehat{F}_n$ replaced by $\widehat{F}_{n,b}^*$, empirical distribution of the sample b, and F replaced by $\widetilde{F}_{n,b}^*$:

$$\widetilde{F}_{n,b}^*(x) = \Phi\left(\frac{x - \widehat{m}_b^*}{\widehat{\sigma}_b^*}\right) \Big/ \Phi\left(\frac{r^u - \widehat{m}_b^*}{\widehat{\sigma}_b^*}\right) \tag{63}$$

3. From the set of B points obtained, calculate the empirical quantiles.

It is advisable to consider one supplementary test of the Anderson-Darling type, designed for a weaker hypothesis : testing for normality by restricting ourselves to the upper side of the distribution. For the modified A-D statistic, we obtain similarly:

$$\underline{AD}_n = n \int_{F(\theta)}^1 \frac{\left\{\overline{F}_n(u) - u\right\}^2}{1-u} du \tag{64}$$

The limit law for (64) is now:

$$\underline{AD}_n \underset{n \to +\infty}{\Longrightarrow} \int_{F(\theta)}^1 \frac{B_0^2(t)}{1-t} dt \tag{65}$$

Clearly, the limit law is now depending on the category of loans. Next, we define $\widehat{\underline{AD}}_N$ in the same way as before, and use the parametric bootstrap analysis along the same lines as explained above.

$\underline{AD}_n$ and AD_n are calculated from the order statistics $U_{(k)} = F\left(X_{(k)}\right)$, with the convention at end-points given by $U_{(0)} = 0$ and $U_{(n+1)} = 1$. Then, we remark that for any function $\Psi(\cdot, \cdot)$:

$$\int_0^1 \Psi\left\{\widehat{F}_n\left(F^{-1}(u)\right), u\right\} du = \sum_{k=0}^n \int_{U_{(k)}}^{U_{(k+1)}} \Psi\left\{\widehat{F}_n\left(F^{-1}(u)\right), u\right\} du = \sum_{k=0}^n \int_{U_{(k)}}^{U_{(k+1)}} \Psi\left\{\frac{k}{n}, u\right\} du$$

- With $\Psi(x,u) = \frac{(x-u)^2}{u(1-u)}$, we obtain after integration the text-book expression:

$$AD_n = -n - \frac{1}{n}\sum_{k=1}^n (2k-1)\log\left[U_{(k)}\right] - \frac{1}{n}\sum_{k=1}^n (2k-1)\log\left[1 - U_{(n-k+1)}\right] \tag{66}$$

- With $\Psi(x,u) = \frac{(x-u)^2}{1-u}$, we obtain after some tedious manipulations ($k_0 = [nF(\theta)] + 1$) :

$$\begin{aligned} \underline{AD}_n = \quad & n \times \left\{\tfrac{1}{2} + \left(\tfrac{k_0}{n} - 1\right)^2 \log\left[1 - F(\theta)\right] + \tfrac{F^2(\theta)}{2} - \left(\tfrac{2k_0}{n} - 1\right) F(\theta)\right\} \\ & - \tfrac{1}{n}\textstyle\sum_{k=1}^{n-k_0} \left\{(2k-1)\log\left[1 - U_{(n-k+1)}\right] - 2U_{(n-k+1)}\right\} \end{aligned} \tag{67}$$

9.2 The E-M algorithm

The density of a variable X which follows a mixture model is written as:

$$f\left(x\left|\underline{\boldsymbol{\theta}}\right.\right)=\sum_{k=1}^{\mathcal{M}}\pi_k\varphi_k\left(x;\underline{\boldsymbol{\theta}}_k\right)$$

$\underline{\boldsymbol{\theta}}_k=\left(m_k,\sigma_k^2\right)$ is the parameter pertaining to the k^{th} normal law of the mixture. Let $\boldsymbol{\theta}$ be the vector of all parameters, including the $\underline{\boldsymbol{\theta}}_k$ and the π_k.

Classically, we introduce the state vector $(Y_1,\ldots,Y_N)$ such that:

$$Y_i=k\in\{1,\ldots,\mathcal{M}\}\text{ if individual }i\text{ arises from component }k$$

Clearly, $\mathbb{P}\left(Y=k\right)=\pi_k$ for all $k=1,\ldots,\mathcal{M}$. Obviously, the Y_i are unobservable, but we may write the joined law of X and Y :

$$f\left(x,y\left|\underline{\boldsymbol{\theta}}\right.\right)=f\left(x\left|y,\underline{\boldsymbol{\theta}}\right.\right)f\left(y\right)$$

The likelihood of the variables $\left(X_i,Y_i\right)_{1\leq i\leq n}$ is:

$$f\left(X,Y\left|\underline{\boldsymbol{\theta}}\right.\right)=\prod_{i=1}^{n}f\left(X_i\left|Y_i,\underline{\boldsymbol{\theta}}\right.\right)\pi_{Y_i}\tag{68}$$

Moreover:

$$\begin{aligned}f\left(Y\left|X,\underline{\boldsymbol{\theta}}\right.\right)&=\prod_{i=1}^{n}f\left(Y_i\left|X_i,\underline{\boldsymbol{\theta}}\right.\right)\\&=\prod_{i=1}^{n}\frac{\pi_{Y_i}f\left(X_i\left|Y_i,\underline{\boldsymbol{\theta}}\right.\right)}{f\left(X_i\left|\underline{\boldsymbol{\theta}}\right.\right)}\end{aligned}\tag{69}$$

and:

$$f\left(X_i\left|\underline{\boldsymbol{\theta}}\right.\right)=\sum_{k=1}^{\mathcal{M}}\pi_k f\left(X_i\left|Y_i=k,\underline{\boldsymbol{\theta}}\right.\right)$$

We suppose that a preliminary estimate of $\underline{\boldsymbol{\theta}}$, $\underline{\boldsymbol{\theta}}^{(1)}$ is available. The basic idea is to improve the estimate by considering the criterion:

$$Q\left(\underline{\boldsymbol{\theta}}\right)=\mathbb{E}_{\underline{\boldsymbol{\theta}}^{(1)}}\left\{\log f\left(X,Y\left|\underline{\boldsymbol{\theta}}\right.\right)\left|X\right.\right\}\tag{70}$$

The calculus of this criterion is the "E" step of the algorithm. It enables us to replace the unknown quantities Y by their theoretical mean; (70) can be written as:

$$Q\left(\underline{\boldsymbol{\theta}}\right)=\int\log f\left(X,y\left|\underline{\boldsymbol{\theta}}\right.\right)f\left(y\left|X,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)dy$$

We use (68) to calculate $\log f\left(X,y\left|\underline{\boldsymbol{\theta}}\right.\right)$ and (69) for the second term. We then get, with $\delta_{l,y_i}=1$ and $y_i=l$, zero otherwise:

$$Q\left(\underline{\boldsymbol{\theta}}\right)=\left(\sum_{y_1=1}^{\mathcal{M}}\cdots\sum_{y_N=1}^{\mathcal{M}}\right)\left(\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\right)\delta_{l,y_i}\log\left\{\pi_l\times f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}\right.\right)\right\}\prod_{j=1}^{n}f\left(y_j\left|X_j,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

$$Q\left(\underline{\boldsymbol{\theta}}\right)=\left(\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\right)\log\left\{\pi_l\times f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}\right.\right)\right\}\left(\sum_{y_1=1}^{\mathcal{M}}\cdots\sum_{y_N=1}^{\mathcal{M}}\right)\delta_{l,y_i}\prod_{j=1}^{n}f\left(y_j\left|X_j,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

We observe now that:

$$\left(\sum_{y_1=1}^{\mathcal{M}}\cdots\sum_{y_N=1}^{\mathcal{M}}\right)\delta_{l,y_i}\prod_{j=1}^{n}f\left(y_j\left|X_j,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)=\left\{\left(\sum_{\substack{y_1=1\\k\neq i}}^{\mathcal{M}}\cdots\sum_{y_N=1}^{\mathcal{M}}\right)\prod_{j\neq i}f\left(y_j\left|X_j,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)\right\}\times f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

The expression between brackets reads as follows:

$$\left\{\prod_{j\neq i}\left(\sum_{y_j=1}^{\mathcal{M}}f\left(y_j\left|X_j,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)\right)\right\}\times f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)=f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

Finally,

$$Q\left(\underline{\boldsymbol{\theta}}\right)=\left(\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\right)\log\left\{\pi_l\times f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}\right.\right)\right\}f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

and the we get the more tractable expression:

$$\begin{aligned}Q\left(\underline{\boldsymbol{\theta}}\right) &= \left(\textstyle\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\right)\log\left(\pi_l\right)f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)\\ &+\left(\textstyle\sum_{i=1}^{n}\sum_{l=1}^{\mathcal{M}}\right)\log\left\{f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}\right.\right)\right\}f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)\end{aligned}\tag{71}$$

In this expression, $f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}\right.\right)=\varphi_l\left(X_i;\underline{\boldsymbol{\theta}}_l\right)$ and for all couple (l,i):

$$f\left(y_l\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)=\frac{\pi_l^{(1)}f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}{f\left(X_i\left|\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}=\frac{\pi_l^{(1)}f\left(X_i\left|y_i=l,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}{\sum_{k=1}^{\mathcal{M}}\pi_k^{(1)}f\left(X_i\left|y_i=k,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}$$

Since $Q\left(\underline{\boldsymbol{\theta}}\right)$ is perfectly calculable from the data, we define the updated parameter according to:

$$\underline{\boldsymbol{\theta}}^{(2)}=\arg\max Q\left(\underline{\boldsymbol{\theta}}\right)\tag{72}$$

(72) is the "M" step of the algorithm. It is particularly simple when the laws composing the mixture are Gaussian. Then, the updating relations become:

$$\pi_k^{(2)}=\frac{1}{N}\sum_{i=1}^{N}f\left(y_k\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

$$m_k^{(2)}=\frac{\sum_{i=1}^{N}X_i f\left(y_k\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}{\sum_{i=1}^{N}f\left(y_k\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}$$

$$\sigma_k^{2(2)}=\frac{\sum_{i=1}^{N}\left(X_i-m_k^{(1)}\right)^2\times f\left(y_k\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}{\sum_{i=1}^{N}f\left(y_k\left|X_i,\underline{\boldsymbol{\theta}}^{(1)}\right.\right)}$$

Now, we briefly consider the typical case when sampling weights P_i are attached to each observation i. In such case, each observation must be duplicated how many times as indicated by P_i.Then, it is easily seen that the updating equations are only slighted modified according to:

$$\pi_k^{(2)} = \frac{1}{N}\sum_{i=1}^{N} P_i f\left(y_k \left| X_i, \underline{\boldsymbol{\theta}}^{(1)}\right.\right)$$

$$m_k^{(2)} = \frac{\sum_{i=1}^{N} P_i X_i f\left(y_k \left| X_i, \underline{\boldsymbol{\theta}}^{(1)}\right.\right)}{\sum_{i=1}^{N} P_i f\left(y_k \left| X_i, \underline{\boldsymbol{\theta}}^{(1)}\right.\right)}$$

$$\sigma_k^{2(2)} = \frac{\sum_{i=1}^{N} P_i \left(X_i - m_k^{(1)}\right)^2 \times f\left(y_k \left| X_i, \underline{\boldsymbol{\theta}}^{(1)}\right.\right)}{\sum_{i=1}^{N} P_i f\left(y_k \left| X_i, \underline{\boldsymbol{\theta}}^{(1)}\right.\right)}$$

9.3 Estimation through subsampling analysis

Subsampling techniques are valid in a very general context, especially when the observations at hand, $(X_1, \ldots, X_n)$ can be supposed to be independent and equidistributed. Indeed, the only requirement is that the statistical problem under investigation leads to a convergence in distribution such as:

$$n^H (\theta_n - \theta) \underset{n\to+\infty}{\Longrightarrow} X \tag{73}$$

θ is some unknown parameter, $\theta_n = \theta_n (X_1, \ldots, X_n)$ is a statistic function of the observations, and H describes the speed of convergence towards the limit law X. Generally, H is known, but in some cases H depends on some structural unknown parameters (e.g. long memory models). The method can be adapted so that it allows the estimation of both H and θ.

The standard maximum likelihood estimator fulfills obviously this very weak hypothesis, with $H = 1/2$. But this is also the case for the test statistics used in the paper. For instance, the test (18) may be written as:

$$\begin{aligned} \mathcal{LR}_n (\mathcal{M} \,|\mathcal{M}+1) &= -2\left\{\mathcal{L}_n (\mathcal{M}) - \mathcal{L}_n (\mathcal{M}\mathcal{M}+1)\right\} \\ &= -2n\left\{\tfrac{\mathcal{L}_n(\mathcal{M})}{n} - \tfrac{\mathcal{L}_n(\mathcal{M}+1)}{n}\right\} \\ &\equiv n\theta_n \end{aligned} \tag{74}$$

Here $\theta_n = 2\left\{\frac{\mathcal{L}_n(\mathcal{M}+1)}{n} - \frac{\mathcal{L}_n(\mathcal{M})}{n}\right\}$ $H = 1$ and $\theta = 0$. The probability limit of θ_n is given by the Kullback contrast:

$$\theta_n \to -2 \times \mathbb{E}_{\mathcal{M}+1}\left(\log \frac{f(x;\boldsymbol{\theta}_{\mathcal{M}})}{f(x;\boldsymbol{\theta}_{\mathcal{M}+1})}\right) \tag{75}$$

This limit is always ≥ 0 and zero when $f(x;\boldsymbol{\theta}_{\mathcal{M}}) = f(x;\boldsymbol{\theta}_{\mathcal{M}+1})$ a.e. for the measure $f(x;\boldsymbol{\theta}_{\mathcal{M}+1})$, that is when the null hypothesis $\mathbf{H}_0$ is satisfied ($\theta = 0$). Conversely, under $\mathbf{H}_a, \theta > 0$, and θ is omitted from expression (74) because the statistic must diverge under the alternative.

The statistics used in section 3.2 can be handled in the same way. For instance:

$$\begin{aligned} AD_n &= n\int_{-\infty}^{r^u} \frac{\left\{\widehat{F}_n(x) - F(x)\right\}^2}{F(x)\{1-F(x)\}} dF(x) \\ &\equiv n\theta_n \end{aligned} \tag{76}$$

Under the null hypothesis, $\theta_n \longrightarrow 0$, whereas under the alternative (we note F_a the c.d.f. of the X_k):

$$\theta_n \longrightarrow \theta \equiv \int_{-\infty}^{r^u} \frac{\{F_a(x) - F(x)\}^2}{F(x)\{1 - F(x)\}} dF(x) > 0$$

The basic idea of subsampling methodology is to draw with replacement B independent samples of size m from $(X_1, \ldots, X_n)$. Then, the test statistics of interest are calculated for all the subsamples : we obtain $\theta_{n,m,b}$ for $b = 1, \ldots, B$. Suppose now that θ_n is the Maximum Likelihood estimator of parameter θ. We approximate the c.d.f F_X of X by its empirical counterpart obtained from the B subsamples:

$$\widehat{F}_{n,X}(x) = \frac{1}{B} \sum_{b=1}^{B} 1\left\{m^H(\theta_{n,m,b} - \theta_n) \leq x\right\}$$

It can be shown (theorem 2.2.1 from Politis *and alii* (1999)) that, if $\frac{1}{B} + \frac{1}{m} + \frac{m}{n} \longrightarrow 0$ when $n \to +\infty$ and if F_X is continuous, then:

$$\sup_x \left|\widehat{F}_{n,X}(x) - F_X(x)\right| \underset{n \to +\infty}{\xrightarrow{P}} 0$$

We turn now to the case when θ_n is a test statistic, as in (74) or (76). Since we work under the null hypothesis, $\theta = 0$ and we approximate the c.d.f F_X of X with:

$$\widehat{F}_{n,X}(x) = \frac{1}{B} \sum_{b=1}^{B} 1\left\{m^H \theta_{n,m,b} \leq x\right\}$$

Under $\mathbf{H}_0$ and the same set of hypothesis as before, $\sup_x \left|\widehat{F}_{n,X}(x) - F_X(x)\right| \underset{n \to +\infty}{\xrightarrow{P}} 0$ (theorem 2.6.1 from Politis *and alii* (1999)).

Finally, let $X_{n,1-\alpha}$ be the quantile of order $1 - \alpha$ from the (empirical) law $\widehat{F}_{n,X}$. Politis *and alii* (1999) have shown that:

$$\left[\begin{array}{l} \text{Under } \mathbf{H}_0 : \mathbb{P}(n\theta_n > X_{n,1-\alpha}) \underset{n \to +\infty}{\longrightarrow} \alpha \\ \text{Under } \mathbf{H}_a : \mathbb{P}(n\theta_n > X_{n,1-\alpha}) \underset{n \to +\infty}{\longrightarrow} 1 \end{array}\right.$$

These properties say : 1) that the asymptotic test associated to θ_n and based on the "subsampling" quantiles $X_{n,1-\alpha}$ has the correct nominal size α, and 2) that the test is consistent.

9.4 Estimation of the asymptotic variance

For the ease of exposition, we consider only the case of a scalar parameter. The estimator is:

$$\widehat{\sigma}^2_{\boldsymbol{\theta}} = \frac{n}{H} \sum_{h=1}^{H} \left(\widehat{\boldsymbol{\theta}}^{(h)} - \widehat{\boldsymbol{\theta}}^{mc}\right)^2$$

with:

$$\widehat{\boldsymbol{\theta}}^{mc} = \frac{1}{H} \sum_{h=1}^{H} \widehat{\boldsymbol{\theta}}^{(h)}$$

and

$$\sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(h)} - \boldsymbol{\theta}\right) \Longrightarrow Y_h$$

with

$$Y_h \rightsquigarrow \sigma_{\boldsymbol{\theta}} \mathcal{N}(0,1)$$

For H fixed and $n \to +\infty$, and thanks to the joined convergence of the $\widehat{\boldsymbol{\theta}}^{(h)}$ which is a direct consequence of the independency of these variables, we get the following convergences:

$$\begin{aligned}
\widehat{\sigma}^2_{\boldsymbol{\theta}} &= \tfrac{1}{H}\sum_{h=1}^{H} n\left(\widehat{\boldsymbol{\theta}}^{(h)} - \boldsymbol{\theta}\right)^2 + n\left(\widehat{\boldsymbol{\theta}}^{mc} - \boldsymbol{\theta}\right)^2 - \tfrac{1}{H} \times \sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{mc} - \boldsymbol{\theta}\right) \times \sum_{h=1}^{H} \sqrt{n}\left(\widehat{\boldsymbol{\theta}}^{(h)} - \boldsymbol{\theta}\right) \\
\widehat{\sigma}^2_{\boldsymbol{\theta}} &\Longrightarrow \tfrac{1}{H}\sum_{h=1}^{H} \sigma^2_{\boldsymbol{\theta}} Y_h^2 + \tfrac{\sigma^2_{\boldsymbol{\theta}}}{H}\left(\sum_{h=1}^{H} Y_h\right)^2 - \tfrac{\sigma^2_{\boldsymbol{\theta}}}{H}\left\{\tfrac{1}{H}\sum_{h=1}^{H} Y_h\right\} \times \left\{\sum_{h=1}^{H} Y_h\right\} \\
&\Longrightarrow \sigma^2_{\boldsymbol{\theta}} \times \left\{\tfrac{1}{H}\sum_{h=1}^{H} Y_h^2 - \left(\tfrac{1}{H}\sum_{h=1}^{H} Y_h\right)^2\right\}
\end{aligned}$$

Now, since the Y_h are $iid\ \mathcal{N}(0,1)$, when $H \to +\infty$, we have:

$$\frac{1}{H}\sum_{h=1}^{H} Y_h^2 - \left(\frac{1}{H}\sum_{h=1}^{H} Y_h\right)^2 \Rightarrow 1$$

Thus,

$$\widehat{\sigma}^2_{\boldsymbol{\theta}} \Longrightarrow \sigma^2_{\boldsymbol{\theta}} \text{ in sequential limit when } (n, H \to +\infty)_{\text{seq}}$$

9.5 Estimated parameters

Note :

The estimators are all pertaining to the continuous part of the distribution.

TETA_E, TETA_S : Monte-Carlo estimate of m_k and its standard error.

PI_E, PI_S : Monte-Carlo estimate of π_k and its standard error.

SIG2_E, SIG2_S : Monte-Carlo estimate of σ_k^2 and its standars error.

October 2003

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
Consumer <1524€	R1	2.7995	0.0223	0.0678	0.0007	2.6342	0.0264
	R2	6.6853	0.0132	0.1198	0.0009	2.3112	0.0233
	R3	12.6625	0.0096	0.4064	0.0018	2.8968	0.0167
	R4	14.9247	0.0096	0.2271	0.0019	0.1255	0.0659
	R5	16.9756	0.0104	0.1223	0.0006	0.1802	0.0165
	R6	19.8523	0.0173	0.0567	0.0002	0.5696	0.0686
Bk over. > 1524€	R1	4.5863	0.0146	0.2342	0.0024	0.4108	0.0124
	R2	6.5662	0.0177	0.1577	0.0016	0.3214	0.0095
	R3	8.7631	0.0164	0.2694	0.0022	1.3811	0.0191
	R4	11.5716	0.0146	0.1013	0.0008	0.4639	0.0241
	R5	13.8538	0.0090	0.0985	0.0006	0.9345	0.0159
	R6	15.5839	0.0104	0.0622	0.0004	0.0014	0.0037
	R7	17.8038	0.0135	0.0768	0.0006	0.4951	0.0043
Personal > 1524€	R1	2.8712	0.0215	0.1083	0.0012	0.3249	0.0047
	R2	5.7399	0.0060	0.2819	0.0031	2.2923	0.0775
	R3	6.6663	0.0098	0.2000	0.0023	0.1988	0.1187
	R4	7.9816	0.0122	0.2264	0.0015	0.2538	0.0126
	R5	9.5598	0.0070	0.1491	0.0010	0.3030	0.0104
	R6	11.4120	0.0291	0.0343	0.0003	0.0004	0.3197

October 2003

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
Housing, fix. rate	R1	0.7564	0.0148	0.0542	0.0005	0.1536	0.0686
	R2	1.5729	0.0001	0.9458	0.0005	0.0230	0.0000
Housing, var. rate	R1	3.4596	0.0034	0.1770	0.0030	0.0157	0.0013
	R2	4.5631	0.0045	0.7323	0.0028	0.3601	0.0018
	R3	6.2359	0.0204	0.0907	0.0008	1.5196	0.0133
Housing, brid.	R1	1.1923	0.0018	0.0913	0.0008	0.0043	0.0004
	R2	1.5681	0.0001	0.9087	0.0008	0.0214	0.0000

October 2003

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
NFC, instal.	R1	3.6052	0.0012	0.1823	0.0004	0.1181	0.0005
	R2	3.6738	0.0002	0.1632	0.0004	0.0002	0.0004
	R3	4.0586	0.0021	0.0851	0.0004	0.0022	0.0002
	R4	4.8672	0.0033	0.1706	0.0003	0.0130	0.0004
	R5	6.3984	0.0031	0.1243	0.0003	0.2626	0.0016
	R6	7.9654	0.0017	0.2154	0.0004	0.3038	0.0011
	R7	8.6835	0.0001	0.0589	0.0002	0.0002	0.0000
NFC,Bk over.	R1	0.7726	0.0117	0.0554	0.0006	0.0000	0.0130
	R2	2.8053	0.0195	0.0608	0.0006	0.1183	0.0084
	R3	5.2203	0.0237	0.2112	0.0020	0.7747	0.0123
	R4	8.7395	0.0140	0.4065	0.0024	1.5139	0.0169
	R5	10.7224	0.0074	0.1253	0.0015	0.0001	0.0087
	R6	12.2753	0.0088	0.0697	0.0003	0.0411	0.0037
	R7	14.2878	0.0082	0.0712	0.0005	0.3702	0.0080
NFC, leasing	R1	1.7321	0.0004	0.0656	0.0002	0.0008	0.0001
	R2	1.8460	0.0004	0.5607	0.0011	0.0811	0.0002
	R3	2.2776	0.0005	0.2359	0.0010	0.0060	0.0003
	R4	2.5188	0.0013	0.1378	0.0006	0.0491	0.0003

October 2003

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
NFC,> 2 years (var.)	R1	1.7732	0.0185	0.2925	0.0040	0.0775	0.0115
	R2	4.4045	0.0115	0.4377	0.0035	0.3728	0.0188
	R3	5.5981	0.0074	0.1634	0.0018	0.0541	0.0020
	R4	7.0841	0.0159	0.1064	0.0013	0.1102	0.0090
NFC,> 2 years (fixed)	R1	0.7291	0.0019	0.0082	0.0000	0.1496	0.0011
	R2	3.8618	0.0032	0.0295	0.0003	0.0000	0.0004
	R3	3.9419	0.0029	0.0596	0.0003	0.0327	0.0005
	R4	4.6567	0.0003	0.0919	0.0007	0.0187	0.0006
	R5	5.0973	0.0014	0.7902	0.0007	1.1922	0.0010
	R6	6.7684	0.0022	0.0206	0.0000	0.0011	0.0003
NFC,< 2 years	R1	2.1261	0.0000	0.1451	0.0001	0.0002	0.0000
	R2	3.6076	0.0006	0.4472	0.0015	0.3299	0.0013
	R3	4.7899	0.0047	0.1826	0.0018	0.0740	0.0043
	R4	6.5596	0.0078	0.1117	0.0005	5.2723	0.0412
	R5	8.4148	0.0086	0.1133	0.0004	0.2755	0.0477

October 2004

cat	Regime	TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
Consumer <1524€	R1	1.2245	0.0173	0.0370	0.0007	0.0384	0.0139
	R2	4.6300	0.0125	0.1018	0.0004	0.6824	0.0061
	R3	7.2230	0.0170	0.0812	0.0005	0.4323	0.0070
	R4	10.4358	0.0268	0.1929	0.0019	0.2387	0.0258
	R5	13.5211	0.0153	0.2715	0.0022	0.0755	0.0257
	R6	15.4375	0.0100	0.1171	0.0015	0.0024	0.0167
	R7	17.6267	0.0051	0.1097	0.0005	0.0325	0.0040
	R8	19.9222	0.0028	0.0889	0.0002	0.0180	0.0018
Bk over. > 1524€	R1	3.7798	0.0054	0.1827	0.0003	0.0031	0.0000
	R2	5.0056	0.0034	0.1209	0.0010	0.0411	0.0053
	R3	6.8200	0.0086	0.2650	0.0013	2.4624	0.0136
	R4	10.6498	0.0155	0.2285	0.0010	5.8258	0.0305
	R5	14.1704	0.0128	0.0861	0.0006	0.0186	0.0181
	R6	15.8804	0.0036	0.0917	0.0005	0.0272	0.0045
	R7	17.9280	0.0011	0.0250	0.0001	0.0000	0.0004
Personal > 1524€	R1	4.2532	0.0041	0.0999	0.0005	0.0087	0.0015
	R2	4.4574	0.0008	0.1198	0.0007	0.0003	0.0021
	R3	6.1080	0.0036	0.1603	0.0024	8.8632	0.0346
	R4	6.6141	0.0029	0.5009	0.0022	1.2578	0.0402
	R5	8.8185	0.0007	0.0636	0.0002	0.0691	0.0004
	R6	9.1187	0.0027	0.0555	0.0001	0.0008	0.0030

October 2004

cat	Regime	TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
Housing, fix. rate	R1	4.5070	0.0064	0.1060	0.0013	0.0136	0.0128
	R2	4.8269	0.0005	0.4439	0.0017	0.1852	0.0072
	R3	5.0670	0.0038	0.4501	0.0022	0.7737	0.0117
Housing, var. rate	R1	4.2714	0.0028	0.9312	0.0024	0.3255	0.0007
	R2	7.7935	0.0170	0.0688	0.0024	0.3958	0.0140
Housing, brid.	R1	2.3788	0.0181	0.2008	0.0031	0.1014	0.0087
	R2	4.2552	0.0046	0.4787	0.0038	0.4599	0.0107
	R3	5.4348	0.0084	0.2419	0.0035	0.0937	0.0078
	R4	6.2975	0.0076	0.0786	0.0014	0.1448	0.0065

October 2004

cat	Regime	TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
NFC, instal.	R1	3.6607	0.0001	0.1982	0.0003	0.0013	0.0001
	R2	3.9659	0.0032	0.2552	0.0004	0.1422	0.0058
	R3	5.1167	0.0079	0.0700	0.0004	0.0034	0.0083
	R4	7.6438	0.0041	0.1302	0.0008	0.2231	0.0466
	R5	7.8870	0.0058	0.1257	0.0008	0.4352	0.0483
	R6	9.0807	0.0171	0.2207	0.0007	0.6513	0.0024
NFC,Bk over.	R1	2.4458	0.0245	0.0882	0.0010	0.0623	0.0092
	R2	6.0016	0.0133	0.1821	0.0009	0.9469	0.0051
	R3	7.9898	0.0077	0.1825	0.0010	0.3165	0.0566
	R4	9.5251	0.0045	0.4393	0.0016	1.2279	0.0214
	R5	12.8582	0.0138	0.0544	0.0010	1.3763	0.0121
	R6	15.0007	0.0001	0.0534	0.0001	0.0000	0.0002
NFC, leasing	R1	1.7857	0.0002	0.6146	0.0006	0.0479	0.0000
	R2	2.2472	0.0003	0.2057	0.0007	0.0113	0.0005
	R3	2.3461	0.0008	0.1796	0.0005	0.0831	0.0004

October 2004

cat	Regime	TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
NFC,> 2 years (var.)	R1	0.7104	0.0004	0.0229	0.0001	0.0000	0.0000
	R2	1.3439	0.0022	0.6762	0.0038	0.0339	0.0002
	R3	1.6483	0.0014	0.2794	0.0038	0.0022	0.0002
	R4	2.0744	0.0010	0.0214	0.0009	0.0000	0.0000
NFC,> 2 years (fixed)	R1	4.8998	0.0003	0.9066	0.0007	0.9149	0.0012
	R2	6.6489	0.0138	0.0934	0.0007	3.2688	0.0186
NFC,< 2 years	R1	0.9282	0.0032	0.2372	0.0007	0.0217	0.0058
	R2	1.3559	0.0008	0.3480	0.0007	0.0157	0.0001
	R3	1.6285	0.0004	0.1175	0.0010	0.0024	0.0002
	R4	1.7931	0.0009	0.1094	0.0005	0.3454	0.0028
	R5	2.0296	0.0005	0.1879	0.0004	0.0042	0.0005

October 2005

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
Consumer <1524€	R1	3.4015	0.0294	0.0727	0.0009	2.2015	0.0282
	R2	7.7186	0.0505	0.0775	0.0010	1.8880	0.0348
	R3	11.3285	0.0279	0.2116	0.0023	1.2402	0.0594
	R4	13.5561	0.0237	0.1339	0.0033	0.0183	0.0479
	R5	14.8441	0.0104	0.3460	0.0035	3.0738	0.0316
	R6	17.9143	0.0001	0.0902	0.0001	0.0017	0.0000
	R7	19.7087	0.0001	0.0682	0.0001	0.0003	0.0000
Bk over. > 1524€	R1	3.6378	0.0001	0.1221	0.0002	0.0026	0.0000
	R2	4.5921	0.0043	0.1668	0.0012	0.4609	0.0054
	R3	7.0096	0.0150	0.3027	0.0025	3.7904	0.0382
	R4	11.4090	0.0372	0.1759	0.0023	4.6560	0.0815
	R5	14.8817	0.0086	0.0961	0.0016	0.0015	0.0330
	R6	16.4858	0.0156	0.0843	0.0011	0.1241	0.0239
	R7	17.6344	0.0117	0.0521	0.0008	0.0000	0.0049
Personal > 1524€	R1	3.8314	0.0237	0.3207	0.0026	0.6546	0.0045
	R2	5.5922	0.0138	0.0620	0.0015	0.4020	0.1115
	R3	6.1150	0.0070	0.1570	0.0044	0.4487	0.0953
	R4	6.6009	0.0098	0.3071	0.0043	0.8599	0.1115
	R5	8.0961	0.0014	0.0886	0.0004	0.0371	0.0004
	R6	8.4433	0.0140	0.0646	0.0003	0.0004	0.0027

October 2005

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
Housing, fix. rate	R1	1.0242	0.0110	0.0734	0.0016	0.1325	0.0055
	R2	1.4328	0.0002	0.9266	0.0016	0.0264	0.0001
Housing, var. rate	R1	1.1946	0.0040	0.4167	0.0089	0.0237	0.0002
	R2	1.3951	0.0013	0.5833	0.0089	0.0173	0.0002
Housing, brid.	R1	2.7282	0.0107	0.0574	0.0004	0.0000	0.0002
	R2	4.0142	0.0008	0.9426	0.0004	0.5691	0.0009

October 2005

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
NFC, instal.	R1	3.6052	0.0002	0.3699	0.0004	0.0251	0.0000
	R2	4.7334	0.0067	0.1204	0.0012	0.1078	0.0114
	R3	7.2489	0.0044	0.1689	0.0011	0.5146	0.0120
	R4	8.2244	0.0048	0.0895	0.0014	0.0000	0.0782
	R5	8.6854	0.0140	0.1125	0.0010	7.0308	0.0933
	R6	9.4400	0.0242	0.1387	0.0008	0.2346	0.0465
NFC,Bk over.	R1	4.3599	0.0065	0.1808	0.0012	0.6346	0.0088
	R2	7.5902	0.0011	0.1351	0.0022	0.0099	0.0037
	R3	8.9571	0.0165	0.5288	0.0039	4.4872	0.0522
	R4	11.1629	0.0372	0.1338	0.0021	0.4239	0.0308
	R5	15.5320	0.0170	0.0216	0.0002	0.4981	0.0104
NFC, leasing	R1	1.5368	0.0065	0.5255	0.0021	0.0525	0.0034
	R2	2.0883	0.0013	0.1932	0.0014	0.0072	0.0002
	R3	2.4261	0.0010	0.2813	0.0012	0.0503	0.0002

October 2005

		TETA_E	TETA_S	PI_E	PI_S	SIG2_E	SIG2_S
cat	Regime						
NFC,> 2 years (var.)	R1	1.0879	0.0028	0.0905	0.0041	0.0007	0.0002
	R2	1.2544	0.0016	0.0943	0.0019	0.0004	0.0002
	R3	1.3795	0.0027	0.7087	0.0056	0.0451	0.0033
	R4	1.5950	0.0068	0.0963	0.0027	0.0015	0.0034
	R5	2.4431	0.0075	0.0102	0.0001	0.0617	0.0122
NFC,> 2 years (fixed)	R1	1.4542	0.0011	0.6569	0.0029	0.0628	0.0033
	R2	1.6545	0.0009	0.1420	0.0023	0.0008	0.0024
	R3	1.7582	0.0003	0.2011	0.0012	0.0013	0.0000
NFC,< 2 years	R1	3.2975	0.0015	0.5382	0.0010	0.3338	0.0010
	R2	4.1662	0.0005	0.0897	0.0007	0.0089	0.0001
	R3	4.9986	0.0043	0.1078	0.0011	0.3142	0.0123
	R4	6.7632	0.0092	0.1324	0.0008	6.1719	0.0641
	R5	7.3812	0.0076	0.1319	0.0007	0.0849	0.0799

9.6 Estimated distributions : households

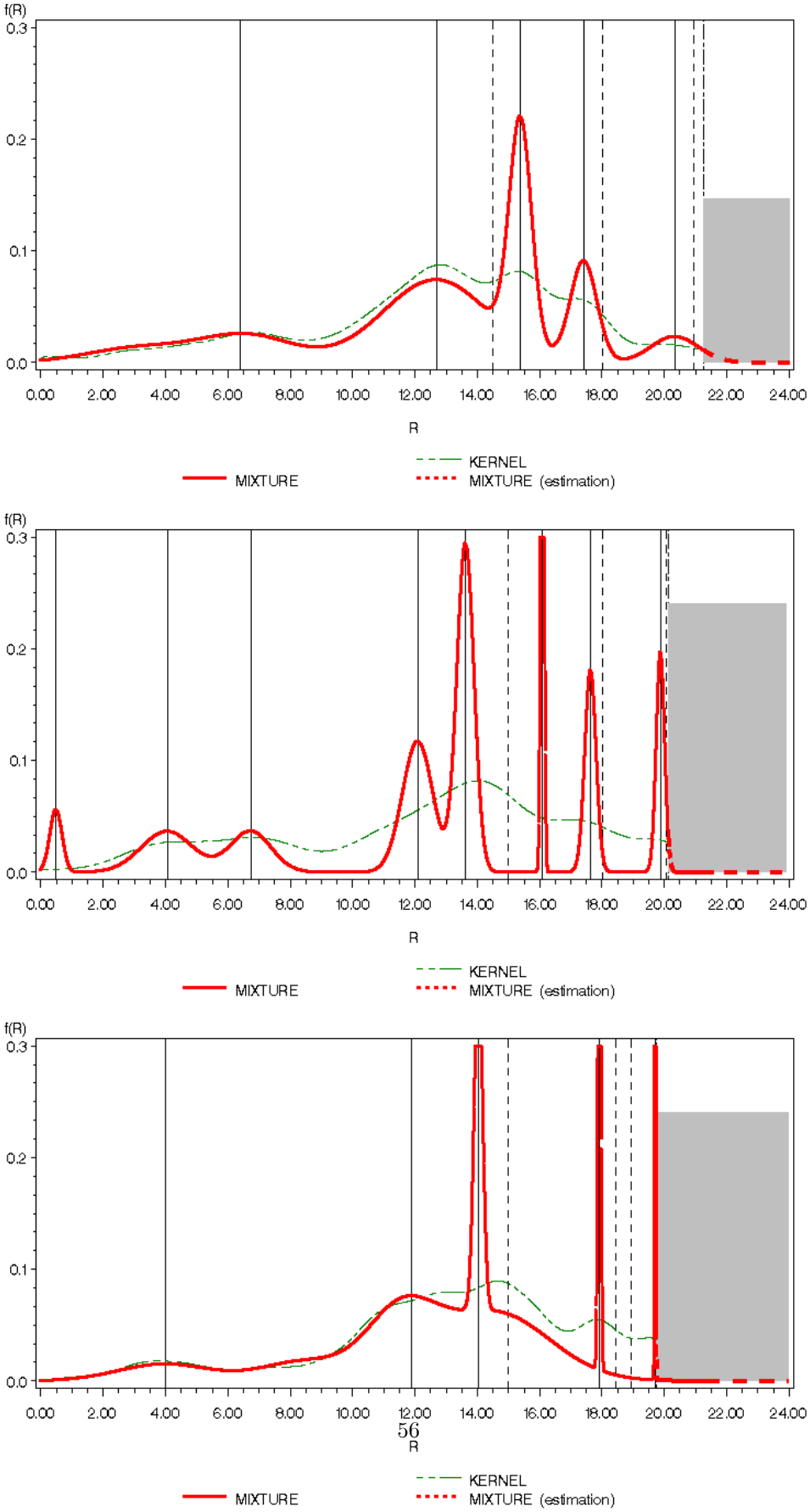


Fig. 9. Category 1

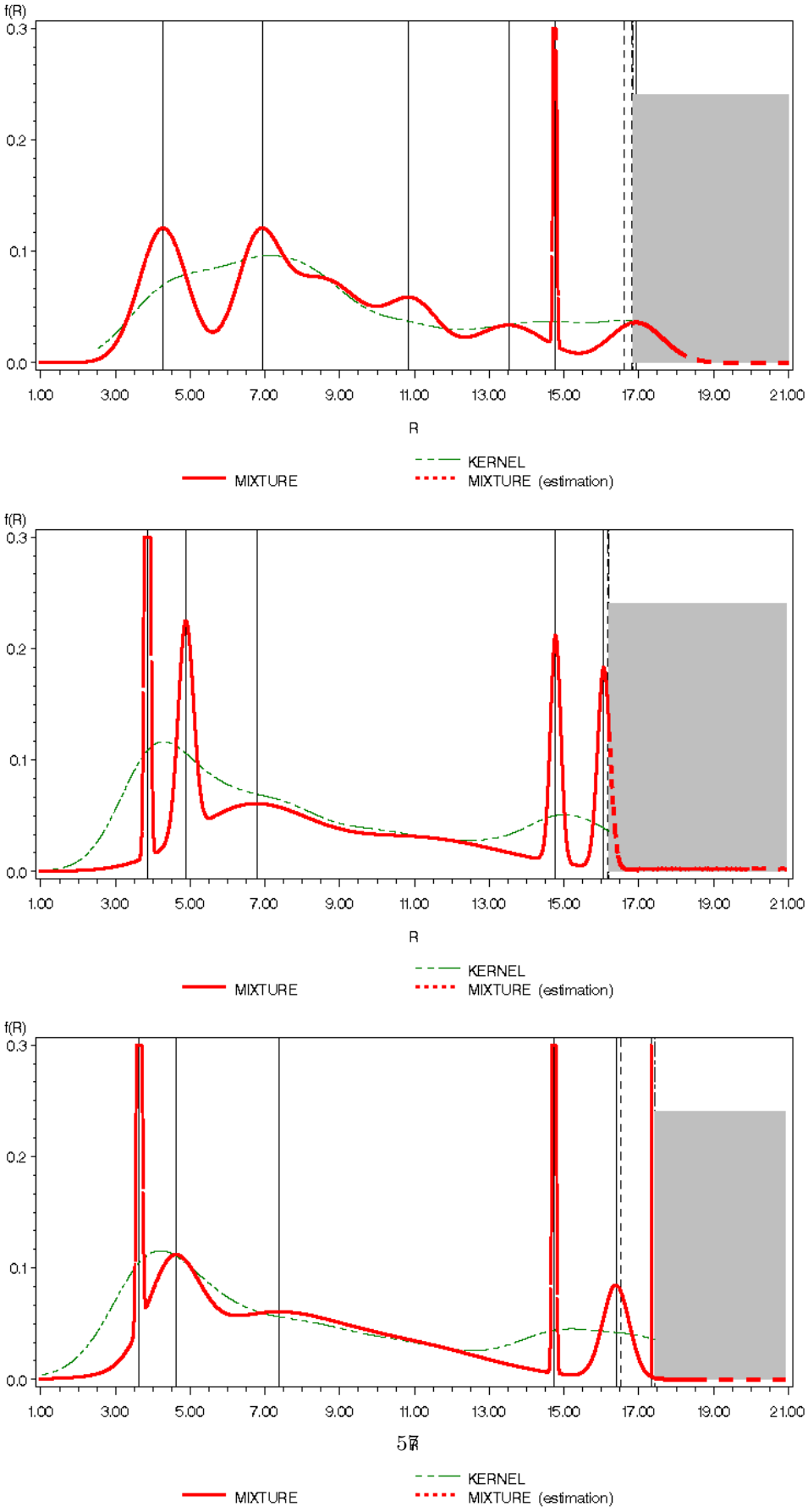


Fig. 10. Category 2

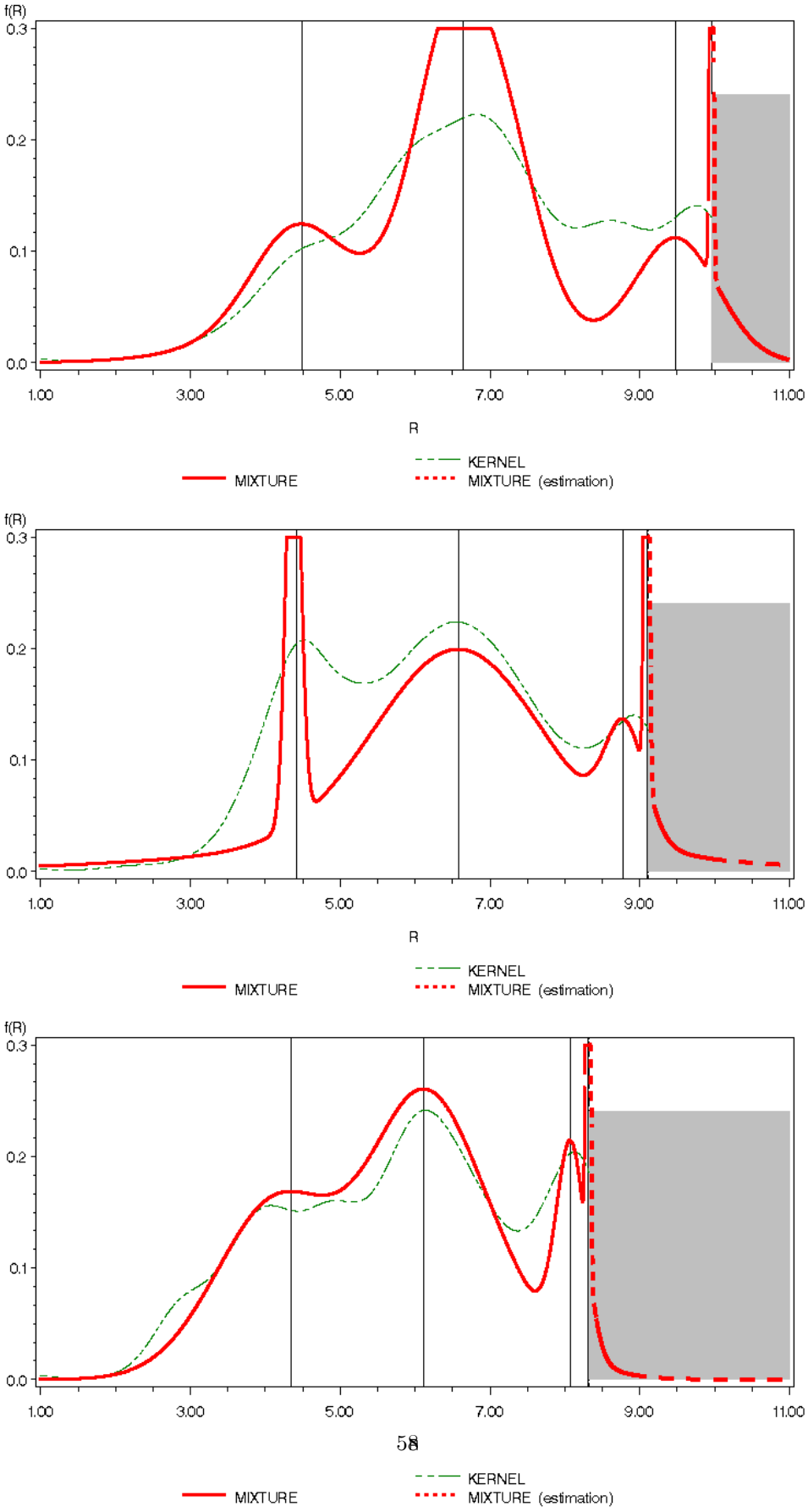


Fig. 11. Category 3

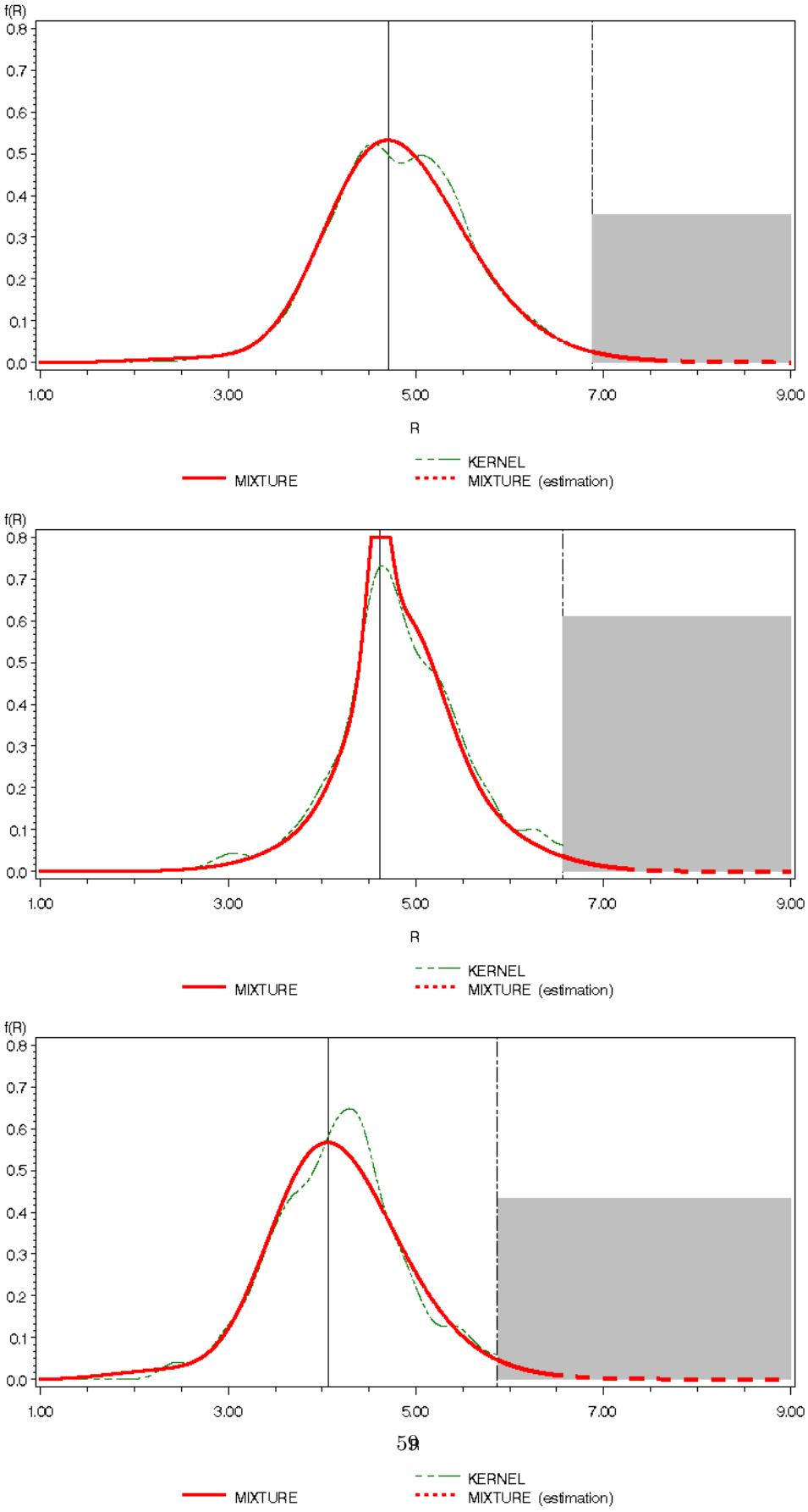


Fig. 12. Category 4

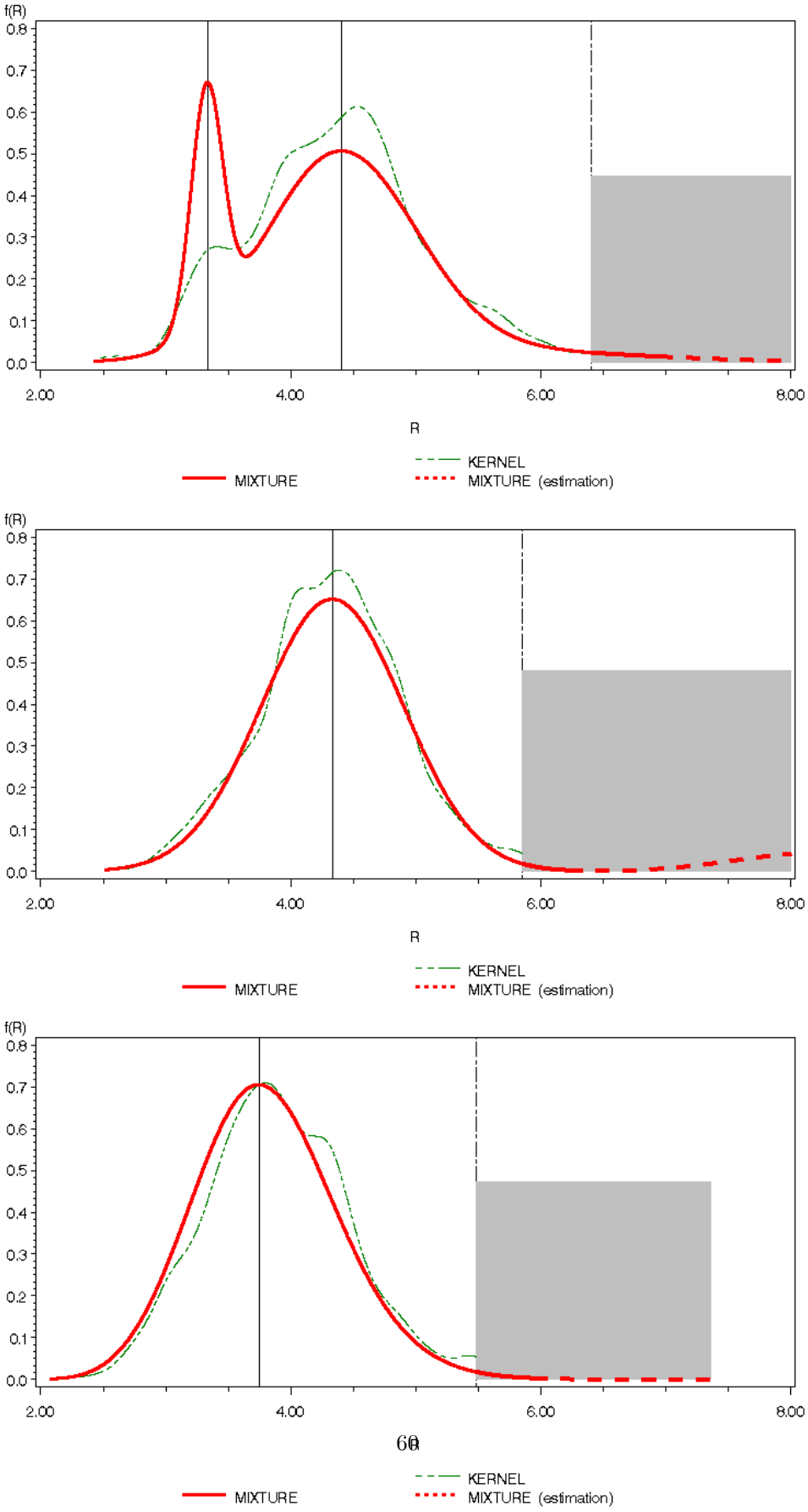


Fig. 13. Category 5

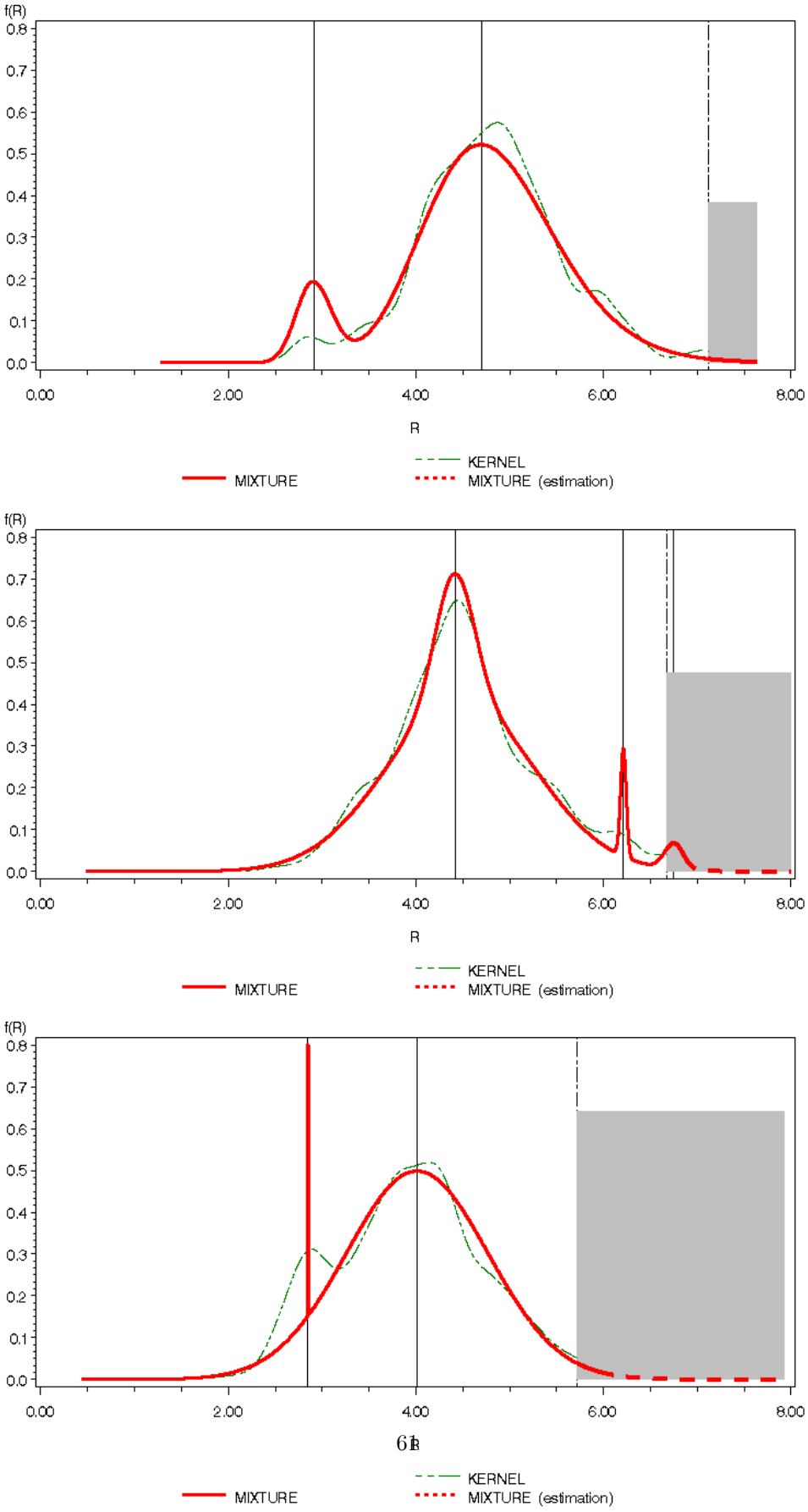


Fig. 14. Category 6

9.7 Estimated distributions : firms

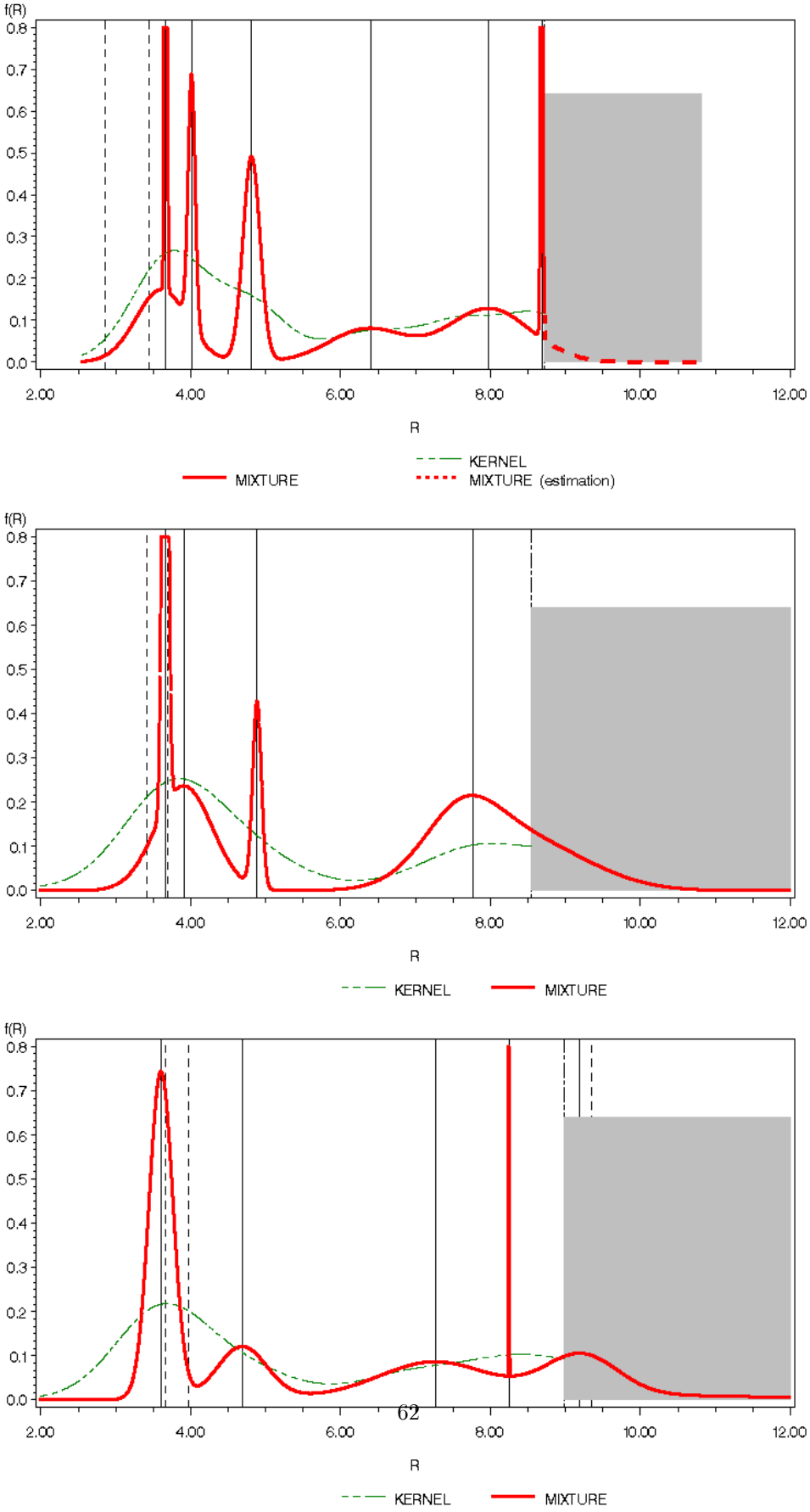


Fig. 15. Category 7

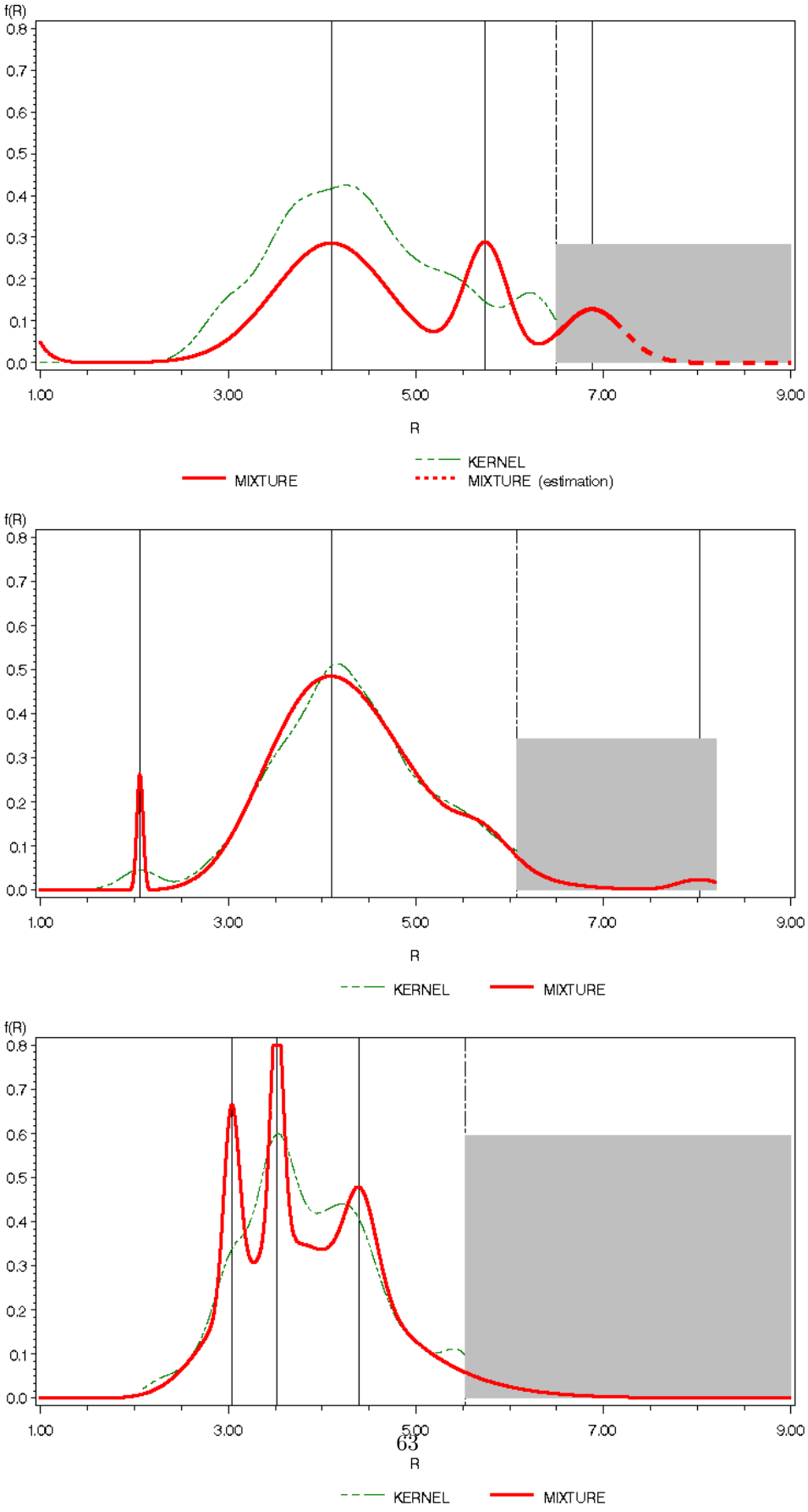


Fig. 16. Category 8

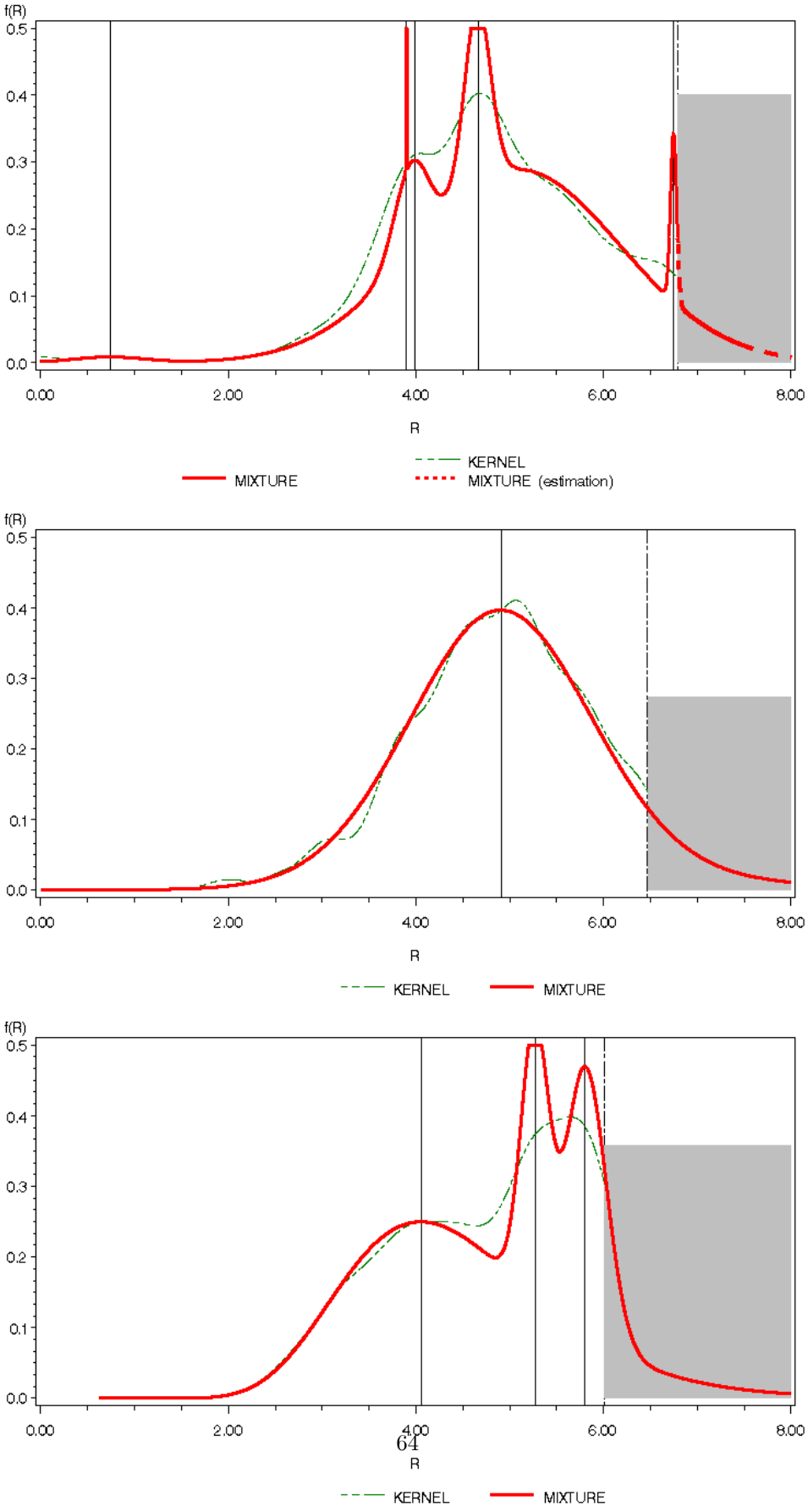
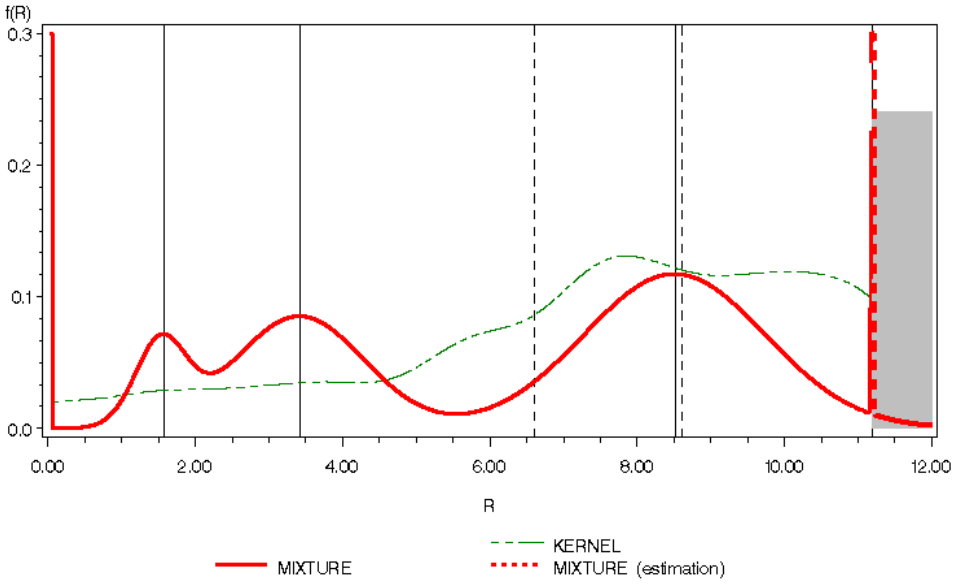
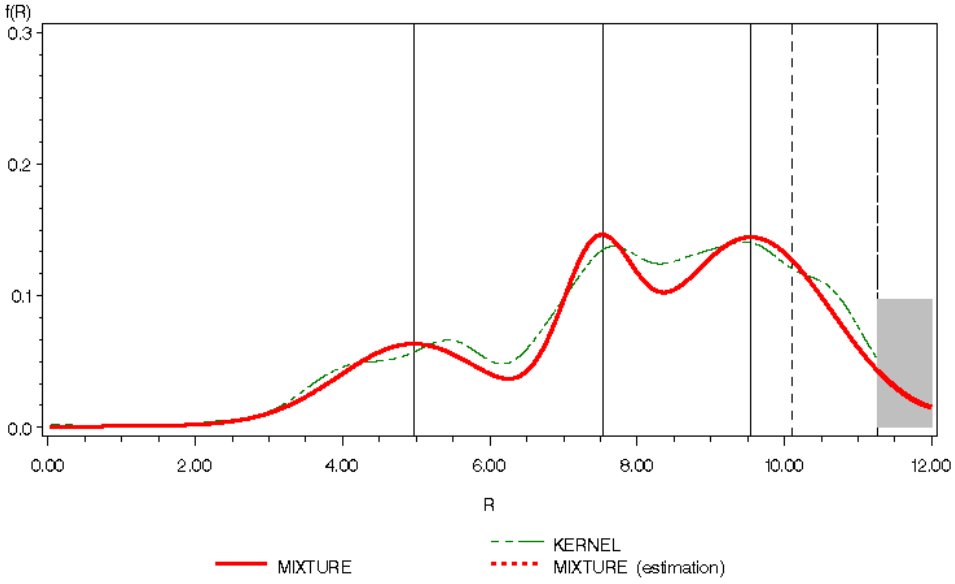


Fig. 17. Category 9





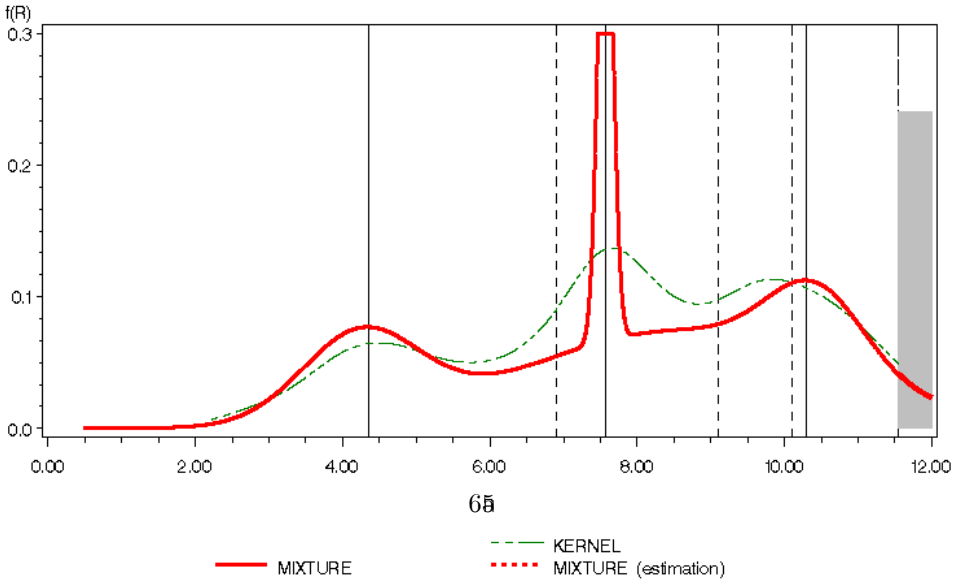


Fig. 18. Category 10

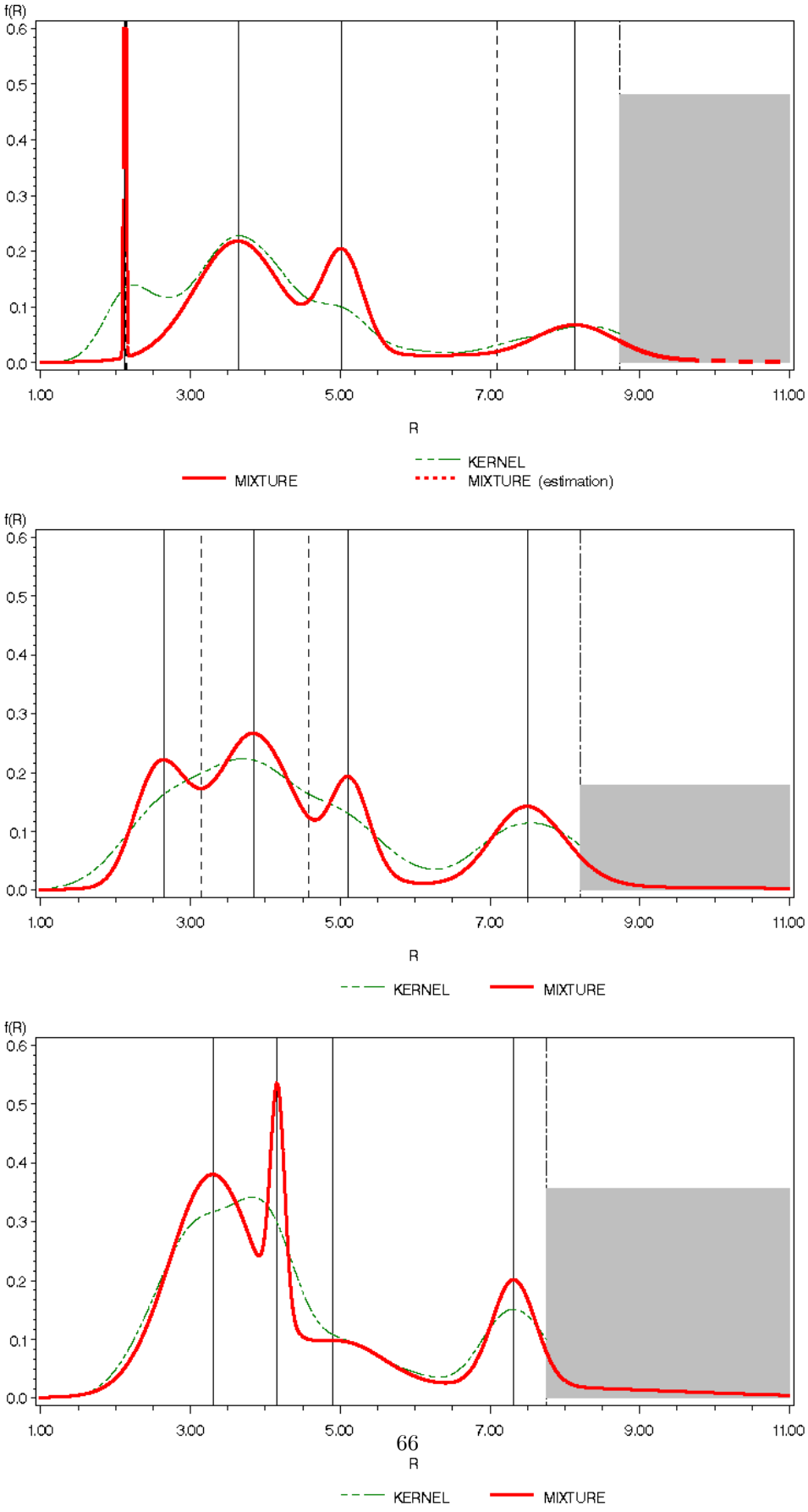
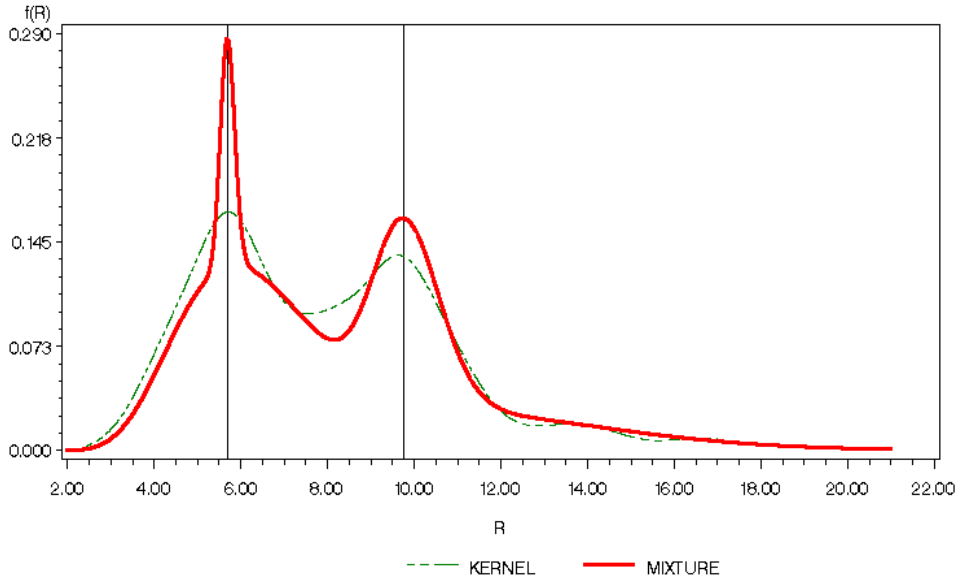
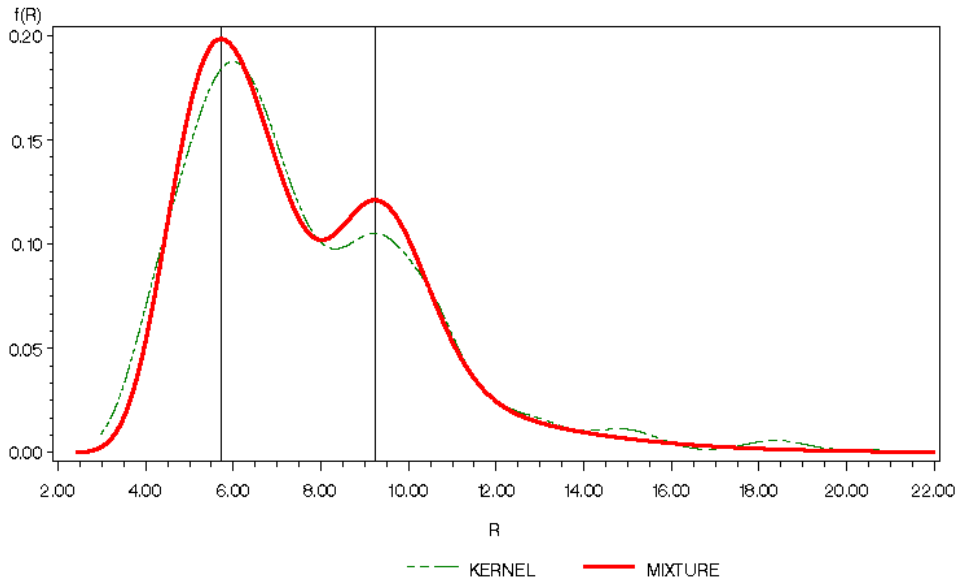


Fig. 19. Category 11





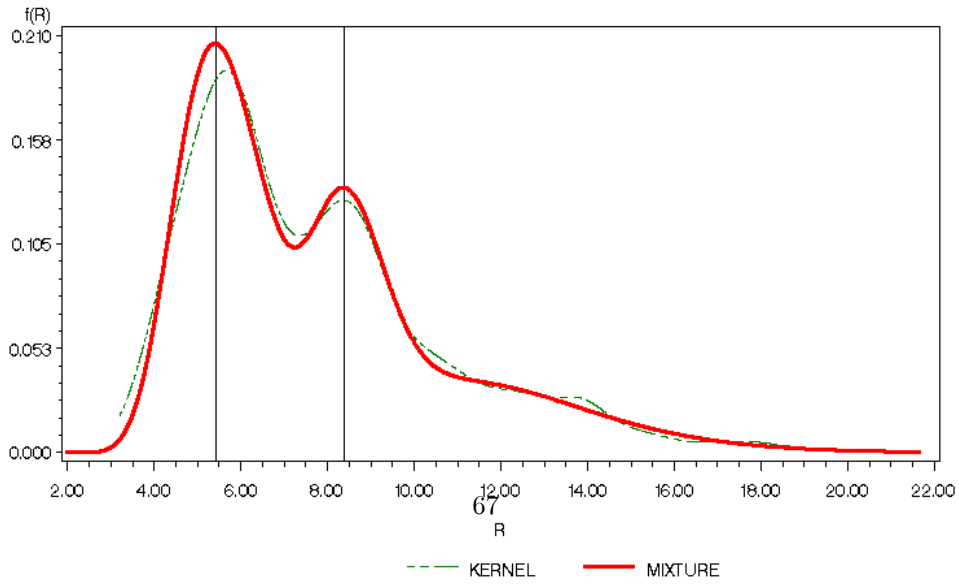
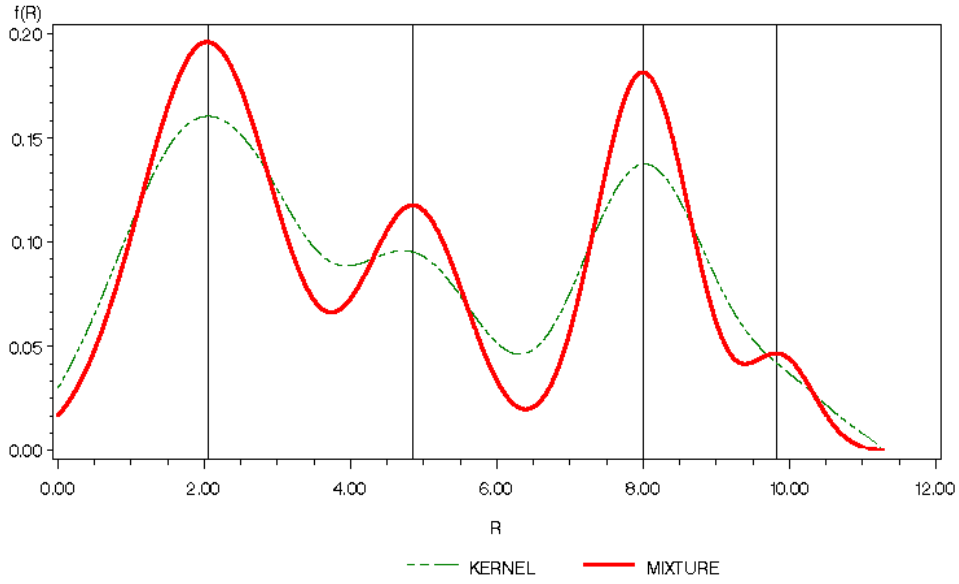
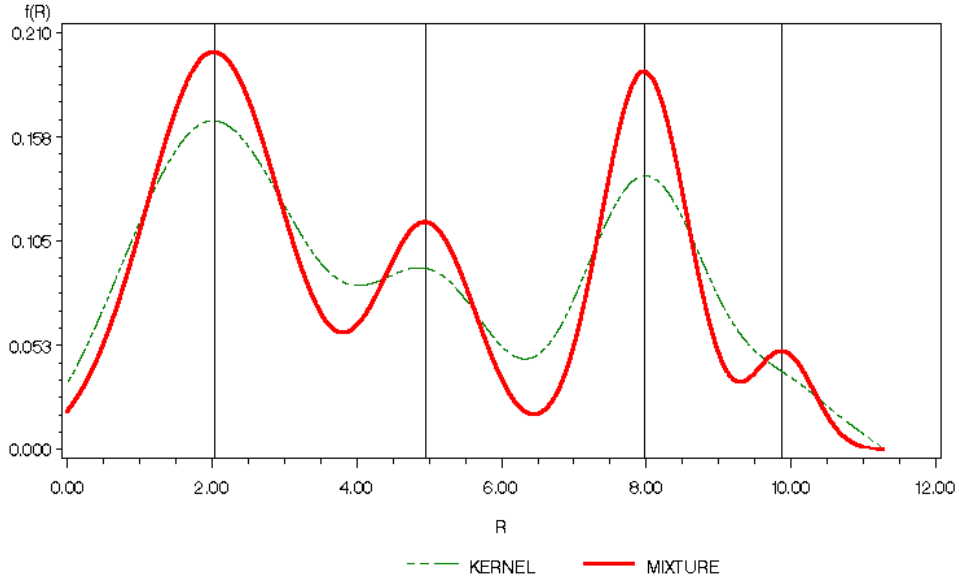


Fig. 20. Category 12





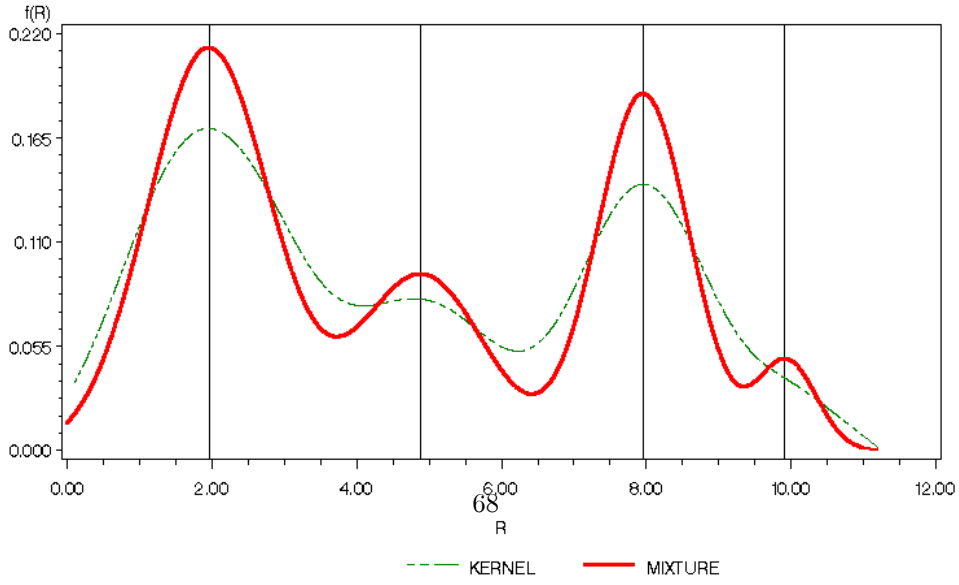
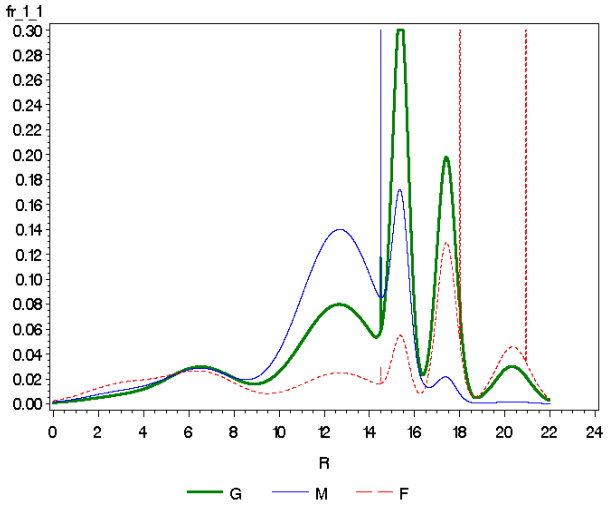


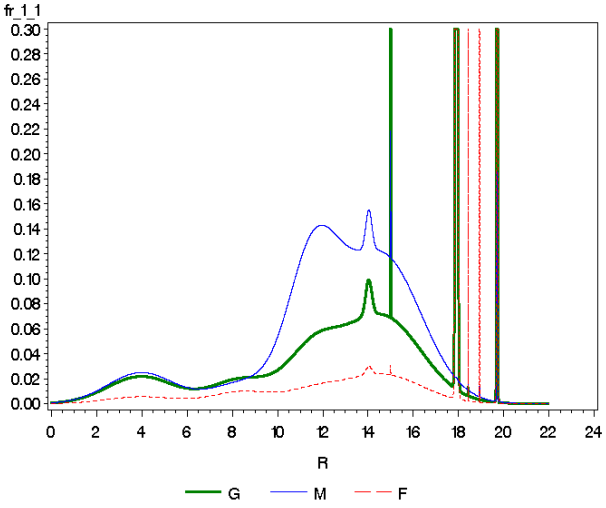
Fig. 21. Category 13

9.8 Distributions per network

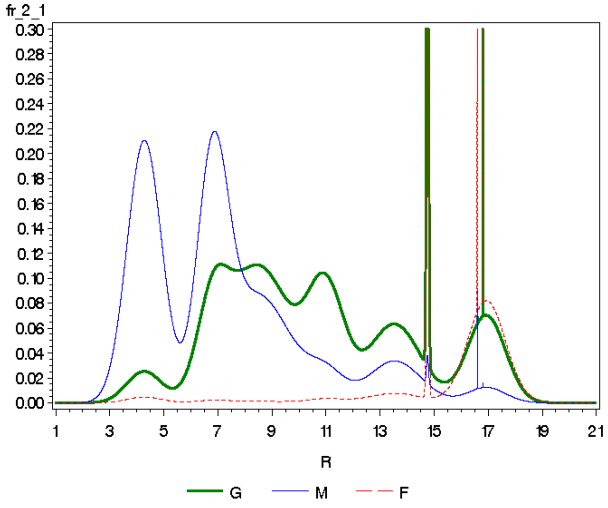
9.8.1 Consuming loans



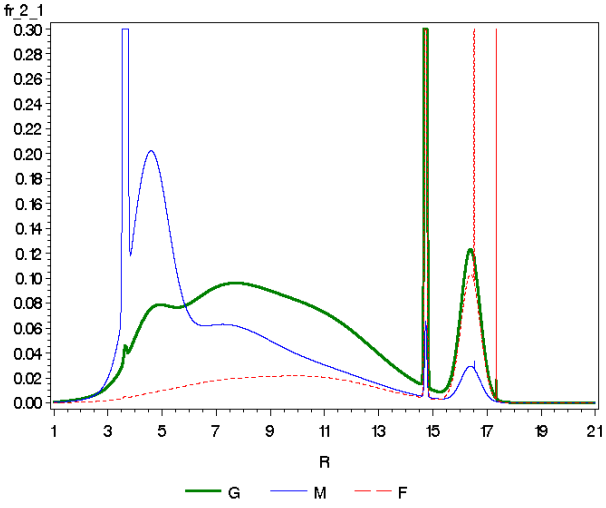
Cat. 1 : 2003



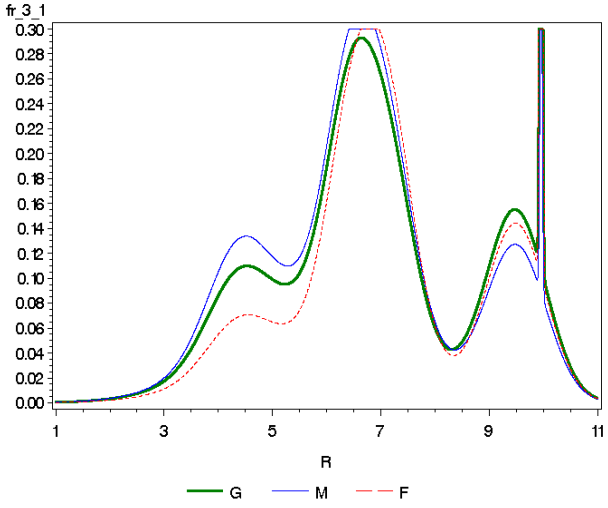
Cat. 1 : 2005



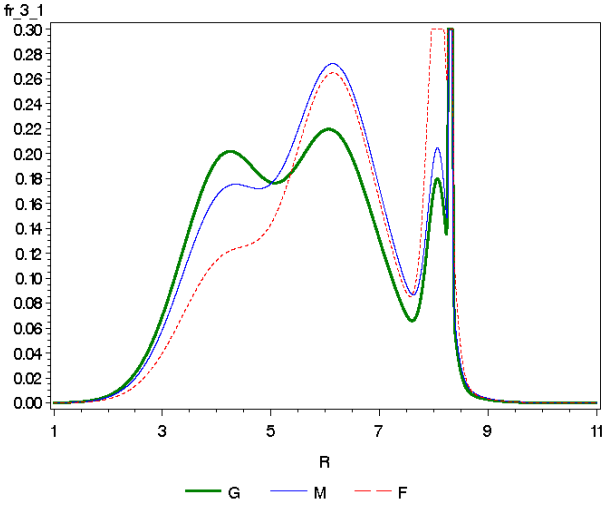
Cat. 2 : 2003



Cat. 2 : 2005

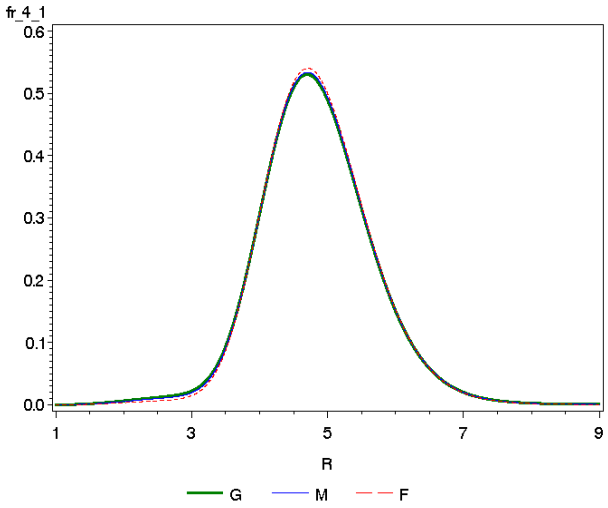


Cat. 3 : 2003

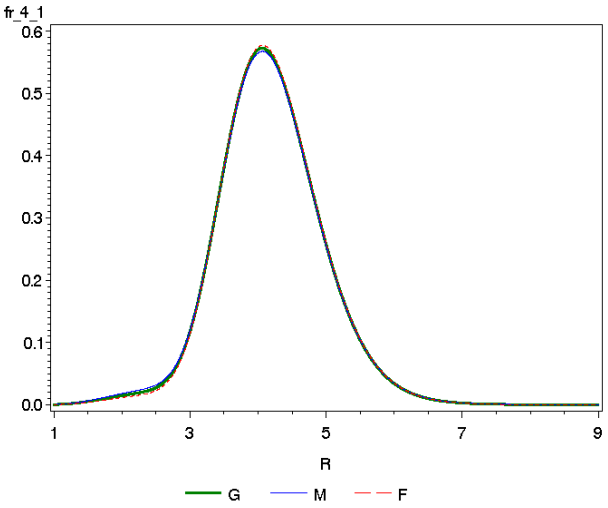


Cat. 3 : 2005

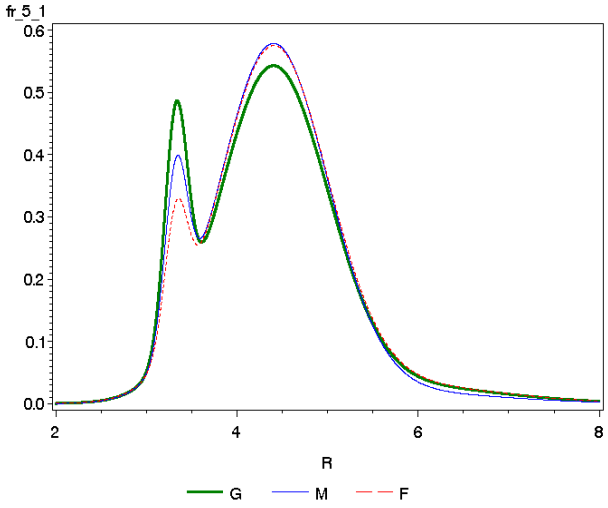
9.8.2 Housing loans



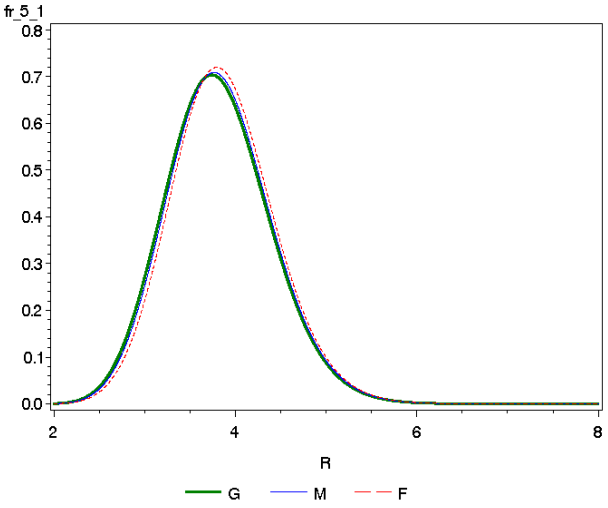
Cat. 4 : 2003



Cat. 4 : 2005

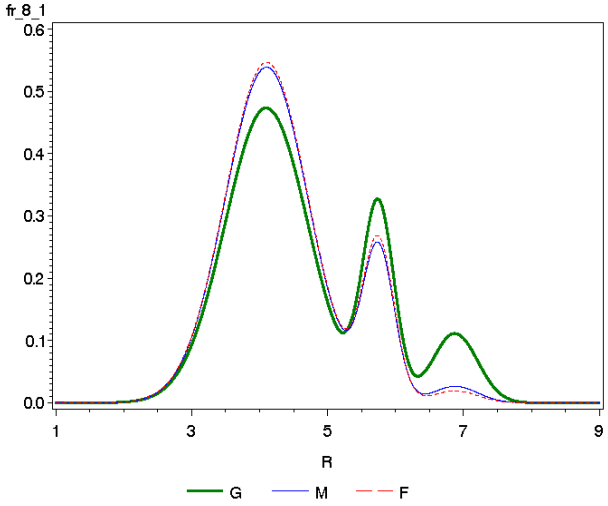


Cat. 5 : 2003

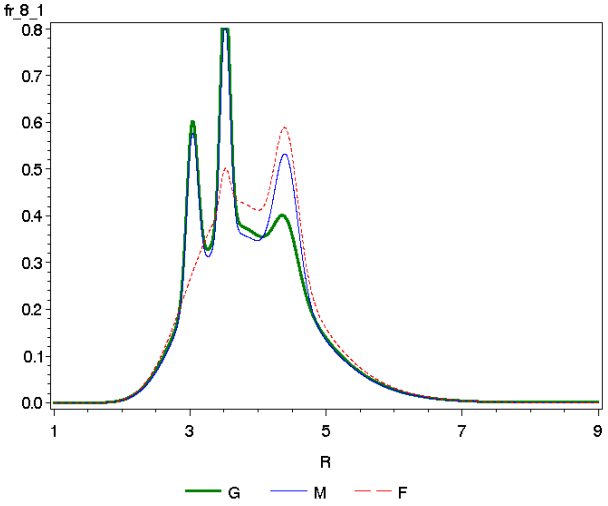


Cat. 5 : 2005

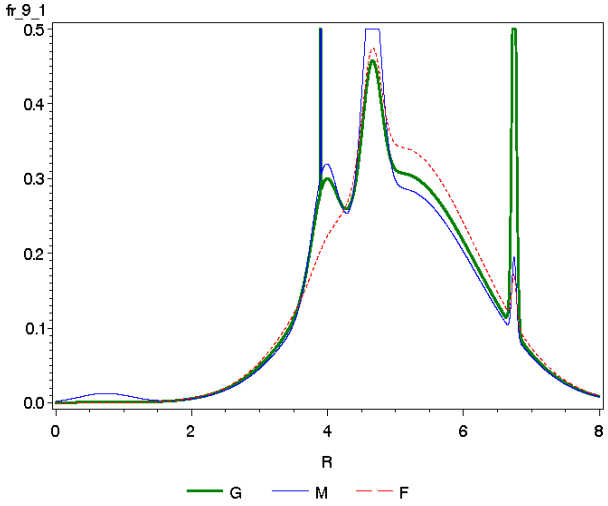
9.8.3 Loans to NFC



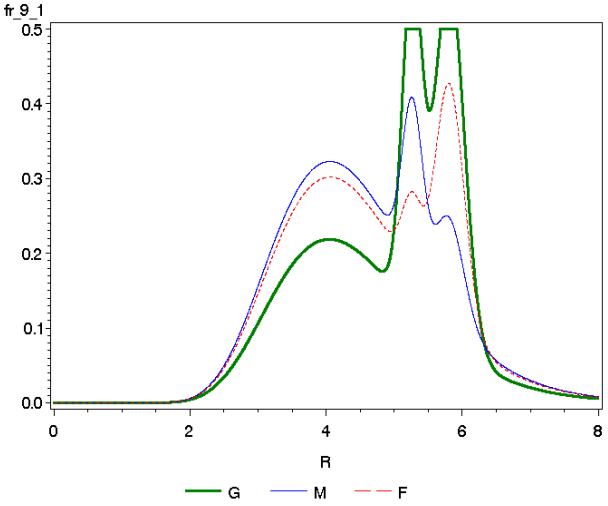
Cat. 8 : 2003



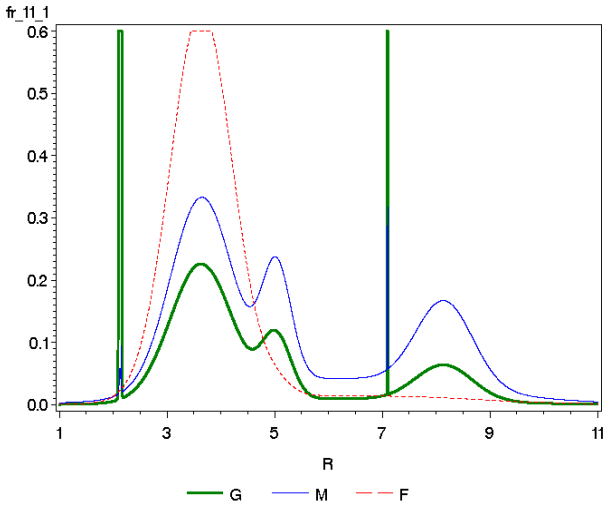
Cat. 8 : 2005



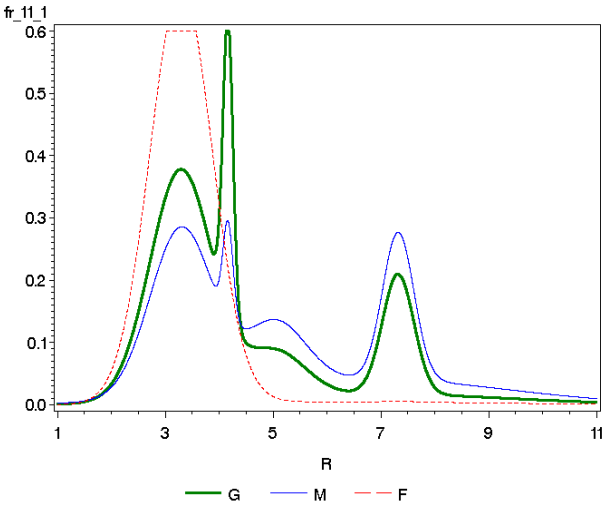
Cat. 9 : 2003



Cat. 9 : 2005



Cat. 11 : 2003



Cat. 11 : 2005

9.9 Structural parameters : alternative methods

For the method labelled "LRT+subsampling", the number of regimes is determined on the basis of the distribution of the likelihood ratio test, obtained from subsampling analysis. For the method labelled "BIC", the model is chosen among the 5 competitive models according to the usual BIC criterion. For both methods, the final choice, between model in level *vs* model in log is achieved through the BIC criterion.

Category	2003		2004	
	λ	$\mathcal{M}$	λ	$\mathcal{M}$
1	N	3	N	5
2	N	5	N	5
3	N	3	N	2
4	L	2	N	3
5	N	3	L	2
6	L	2	N	3
7	N	5	N	5
8	N	4	N	4
9	N	3	N	4
10	N	2	N	4
11	N	5	N	5
12	L	3	L	3
13	N	5	N	5

Method : LRT+subsampling

Category	2003		2004	
	λ	$\mathcal{M}$	λ	$\mathcal{M}$
1	N	3	N	5
2	N	4	N	5
3	N	2	N	4
4	L	2	N	2
5	L	1	L	2
6	L	2	N	2
7	N	5	N	5
8	N	3	N	2
9	N	3	N	2
10	N	4	N	4
11	N	5	N	5
12	N	3	L	3
13	N	4	N	4

Method : BIC