Figini, Silvia; Giudici, Paolo; Brooks, S. P.

Working Paper

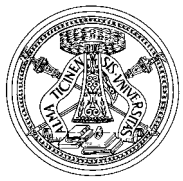
Bayesian feature selection to estimate customer survival

Quaderni di Dipartimento - EPMQ, No. 185

Provided in Cooperation with:
University of Pavia, Department of Economics and Quantitative Methods (EPMQ)

Suggested Citation: Figini, Silvia; Giudici, Paolo; Brooks, S. P. (2006) : Bayesian feature selection to estimate customer survival, Quaderni di Dipartimento - EPMQ, No. 185, Università degli Studi di Pavia, Dipartimento di Economia Politica e Metodi Quantitativi (EPMQ), Pavia

This Version is available at:
https://hdl.handle.net/10419/87130



Quaderni di Dipartimento

Building predictive models for feature selection in genomic mining

Silvia Figini
(University of Pavia)

Paolo Giudici
(University of Pavia)

S.P. Brooks
(Statistical Laboratory Centre for Mathematical Sciences)

185 (06-06)

Dipartimento di economia politica
e metodi quantitativi
Università degli studi di Pavia
Via San Felice, 5
I-27100 Pavia

Giugno 2006

Bayesian feature selection to estimate customer survival

S. Figini and P. Giudici[1] and S.P. Brooks[2]

[1] University of Pavia
Via S. Felice, 5 Pavia
silvia.figini@eco.unipv.it;giudici@unipv.it
[2] Statistical Laboratory Centre for Mathematical Sciences
Wilberforce Road, Cambridge,CB3 0WB
S.P.Brooks@statslab.cam.ac.uk

Abstract. We consider the problem of estimating the lifetime value of customers, when a large number of features are present in the data. In order to measure lifetime value we use survival analysis models to estimate customer tenure. In such a context, a number of classical modelling challenges arise. We will show how our proposed Bayesian methods perform, and compare it with classical churn models on a real case study. More specifically, based on data from a media service company, our aim will be to predict churn behaviour, in order to entertain appropriate retention actions.

1 Background and preliminaries

Our case study concerns a media service company. The main objective of such a company is to maintain its customers, in an increasingly competitive market; to evaluate the lifetime value of such customers, and to carefully design appropriate marketing actions. The company is such that most of its sales of services are arranged through a yearly contract that allows buying different packages of services at different costs. The contract of each customer with the company is thus renewed yearly. If the client does not withdraw, the contract is renewed automatically. Otherwise the client is said to churn.
In the company there are three types of churn events: people that withdraw from their contract in due time (i.e. more than 60 days before the due date); people that withdraw from their contracts overtime (i.e. less than 60 days before the due date); people that withdraw without giving notice, as in the case of bad payers. Correspondingly, the company assigns two different churn states: an 'EXIT' state to the first two classes of customers; and a 'SUSPENSION' state to the third. Concerning the causes of churn, it is possible to identify a number of components that can generate such behaviour:

- A static component, determined by the characteristics of the customers and the type/subject of contracts;
- A dynamic component, that incorporates trends and the contact of the clients with the call center of the company;

– A seasonal part, tied to the period of subscription of the contract;
– External factors, that include the course of the markets and of the competitors

1.1 Traditional churn models

Statistical models typically used to predict churn are based on logistic regression or classification trees (see e.g. Giudici 2003). Generally, all models are evaluated on the basis of a test sample (by definition not included in the training phase) and classified in terms of predictive accuracy with respect to the actual response values in it. In business terms, predictive accuracy means being able to identify correctly those individuals that will become churners during the evaluation phase (correct identification). Evaluation is thus made using a confusion, or cross validation matrix. However, there is a problem with the excessive influence of the contract deadline. For instance the fitted tree models(CART and Chaid) predict that 90% of customers whose deadline is in April are at risk. If we consider that the variable target was built from data gathered in February, the customers whose term is in April and have to regularly unsubscribe within the 60 days allowed, *must* become EXIT in February. Therefore, despite their good predictive capability, these models are useless for marketing actions, as a very simple model based on customer's deadlines will perform as well.
The use of new methods is therefore necessary to obtain a predictive tool which is able to consider the fact that churn data is ordered in calendar time. Before illustrating our new methodology to estimate churn, we introduce the association relationship and Life Time Value.

2 Churn Events and Life Time Value

We start with a simple scenario where a customer generates a margin m_t for each period t, the discount rate is i and the probability of retention rate is 1. In this case, the lifetime value of this customer is simply the present value of the future income stream, or

$$LTV = \sum_{t=0}^{\infty} \frac{m_t}{(1+i)^t}.$$

This is identical to the discounted cash flow approach of valuing perpetuities (Brealey and Myers 1996). When we account for a customer retention rate r, this formulation is modified as follows:

$$LTV = \sum_{t=0}^{\infty} m_t \frac{r_t}{(1+i)^t}.$$

Many researchers have debated the appropriate duration over which lifetime estimates should be based (Berger and Nasr 1998).

We build our model for an infinite time horizon for several reasons. First, we do not need to arbitrarily specify the number of years that a customer is going to stay with the company. Second, the retention rate can account for the fact that over time the chances of a customer staying with the company go down significantly. Third, the typical method of converting retention rates into expected lifetime and then calculating present value over that finite time period overestimates lifetime value. Fourth, both retention and discount rates ensure that earnings in a distant future contribute significantly less to lifetime value.
To estimate the lifetime value of the entire customer base of a firm, we recognize that the firm acquires new customers at each time period. Each cohort of customers goes through a defection and profit pattern.
For example, a firm acquires n_0 customers at time 0 at an acquisition cost of c_0 per customer. Over time, customers defect so that the firm is left with $n_0 \times r$. customers at the end of period 2, and so on. The profit from each customer may vary over time. For example, Reichheld (1996) suggests that profits from a customer increase over his/her lifetime. In contrast, Reinartz and Kumar (2000) find that this pattern does not hold for non-contractual settings. Therefore the lifetime value of cohort 0 at current time 0 is given by,

$$LTV_0 = n_0 \sum_{t=0}^{\infty} m_t \frac{r^t}{(1+i)^t} - n_0 c_0.$$

Cohort 1 follows a pattern similar to cohort 0 except that it is shifted in time by one period. Therefore, the lifetime value of cohort 1 at time 1 is given by,

$$LTV_1 = n_1 \sum_{t=0}^{\infty} m_{t-1} \frac{r^{t-1}}{(1+i)^t} - n_1 c_1.$$

It is easy to convert this value at the current time 0 by discounting it for one period. In other words, the lifetime value of cohort 1 at time 0 is,

$$LTV_1 = \frac{n_1}{1+i} \sum_{t=1}^{\infty} m_{t-1} \frac{r^{t-1}}{(1+i)^{t-1}} - \frac{n_1 c_1}{1+i}.$$

In general, the lifetime value for the k-th cohort at current time 0 is given by

$$LTV_k = \frac{n_k}{(1+i)^k} \sum_{t=k}^{\infty} m_{t-k} \frac{r^{t-k}}{(1+i)^{t-k}} - \frac{n_k c_k}{(1+i)^k}.$$

The value of the firm's customer base is then the sum of the lifetime value of all cohorts.

$$Value = \sum_{k=0}^{\infty} \frac{n_k}{(1+i)^k} \sum_{t=k}^{\infty} m_{t-k} \frac{r^{t-k}}{(1+i)^{t-k}} - \sum_{k=0}^{\infty} \frac{n_k c_k}{(1+i)^k}.$$

In general, an LTV model has three components: customer value over time, customer length of service and a discounting factor. Our proposal leads to a

statistical model statistical model for churn.
Given a customer, there are three factors we have to determine in order to calculate LTV:

- The customers value over time:$v(t)$ for $t > 0$, where t denotes time
- A model describing the customers churn probability over time. This is usually described by a survival function which describes the probability that the customer will still be active at time t. We can then define $f(t)$ as the customers instantaneous probability of churn at time t: $f(t) = -\frac{dS}{dt}$, where $S(t)$ is the survival function. The quantity most commonly modeled, however, is the hazard function $h(t) = f(t)/S(t)$.
- A discounting factor $D(t)$ which describes how much each euro gained in some future time t is worth for us right now. This function is usually given based on business knowledge.

Given these three components, we can write the explicit formula for a customers LTV as follows:

$$LTV = \int_0^\infty S(t)v(t)D(t)dt.$$

In other words, LTV is the total value to be gained while the customer is still active. While this formula is attractive and straight-forward, the essence of the challenge lies, of course, in estimating the $v(t)$ and $S(t)$ components in a reasonable way. We can build models of varying structural and computational complexity for these two quantities. We can use a highly simplistic model assuming constant churn rate so if we observe 0.05 churn rate in the current month, we can set $S(t) = 0.95^t$. This model ignores the different factors that can affect churning customer's individual characteristics, contracts and commitments, etc. On the other hand we can build a complex proportional hazards model, using hundreds of customer properties as predictors. Our approach is intermediate, as we shall employ hazard models based on appropriate variable selection. On the other hand the value function $v(t)$ will be left to the elicitation done by business experts.

3 Survival analysis models to estimate churn

Survival analysis is concerned with studying the time between entry to a study and a subsequent event (churn). Let T be a continuous nonnegative random variable representing the survival times of individuals in some population. Let $f(t)$ denote the probability density function (pdf) of T and let the distribution function be

$$F(t) = P(T \leq t) = \int_0^t f(u)du.$$

The probability of an individual surviving until time t is given by the survivor function

$$S(t) = 1 - F(t) = P(T \geq t).$$

We note that $S(t)$ is a monotone decreasing function with $S(0) = 1$ and $S(\infty) = 0$. All functions, unless stated otherwise, are defined over the interval $[0, \infty]$. The hazard function, $h(t)$ is an instantaneous rate of failure at time t and is defined by

$$h(t) = \lim_{\epsilon_t \to 0} \frac{Pr(t \leq T \leq t + \epsilon_t | T \geq t)}{\epsilon_t} = \frac{f(t)}{S(t)}.$$

The functions $f(t)$, $F(t)$, $S(t)$, and $h(t)$ give mathematically equivalent specifications of the distributions of T. It is easy to derive expressions for $S(t)$ and $f(t)$ in terms of $h(t)$ (see e.g. Kalbfleisch and Prentice, 1980).
In our case, to explain *h(t)* we have chosen to implement Cox's model, see e.g. Cox 1972. In survival models (see e.g. Hougaard, 1995), the hazard function for a given individual describes the instantaneous risk of experiencing an event of interest within an infinitesimal interval of time, given that the individual has not yet experienced that event. The Cox hazard function for fixed-time covariates, x, is

$$\lambda(t; x) = \lambda_0(t) exp(x'\beta). \tag{1}$$

Due to the construction of the previous equation, the baseline hazard $\lambda_0(t)$ is defined as the hazard function for that individual with zero on all covariates. Because the baseline hazard is not assumed to be of a parametric form, Cox model (1) is referred to as a semi-parametric model for the hazard function. The survival function corresponding to (8) is then (see e.g. Klein and Moeschberger, 1997)

$$S(t; x) = exp\left[-exp(x'\beta) \int_0^t \lambda_0(u) du\right]. \tag{2}$$

The integral in (2) is called the baseline cumulative hazard function. Several methods are available for estimating the baseline cumulative hazard function. Cox model has become the most used procedure for modeling the relationship of covariates to a survival or other censored outcome (see e.g. Singer and Willet 2003). Its form is flexible enough to allow time-dependent covariates as well as frailty terms and stratification. However, it has some restrictions. One of the restrictions in using the Cox model with time-fixed covariates is its proportional hazards (PH) assumption; that is, that the hazard ratio between two covariate values is constant over time. This is due to the common baseline hazard function canceling out in the ratio of the two hazards. Thus, for fixed-time covariates, the exponent of a coefficient describes the relative change in the baseline hazard due to that covariate.
The baseline hazard is typically considered a nuisance parameter, and estimation of β is done by maximizing a profile likelihood, with $\lambda_0(t)$ being substituted for an expression involving β and x , as well as the times at which failures occur. This expression is called the profile maximum likelihood estimate of $\lambda_0(t)$. The

likelihood with $\lambda_0(t)$ profiled out is called the *partial likelihood* by Cox. For fixed-time covariates and independent observations, the partial likelihood is given by

$$L(\beta) = \prod_{i=1}^{D} \frac{exp(x_i'\beta)}{\left[\sum_{j \in R(t_i)} exp(x_j'\beta)\right]^{d_i}}, \tag{3}$$

where D is the total number of events, d_i is the number of events at time t_i and $R(t_i)$ is the risk set at time at time t_i (the number of customers in the data set who have censored times later than or equal to time t_i). The value of β that maximizes (3) is called the maximum partial likelihood estimate (MPLE).
In Cox model building the objective is to identify the variables that are most associated with the churn event. This implies that a model selection exercise, aimed at choosing the statistical model that best fits the data, is to be carried out. The statistical literature presents many references for model selection; most models are based on the comparison of model scores. The main score functions to evaluate models are related to the Kullback-Leibler principle. This occurs for criteria that penalize for model complexity, such as AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion).
The tenure prediction models we have developed generate, for a given customer *i*, a hazard function, that indicates the probability $h_i(t)$ of cancellation at a given time t in the future. A hazard curve can be converted to a survival curve or to a survival function which plots the probability $S_i(t)$ of 'survival' (non-cancellation) at any time t, given that customer was 'alive' (active) at time *t-1*, i.e.,

$$S_i(t) = S_i(t-1) \times [1 - h_i(t)],$$

with $S_i(1) = 1$.

4 Criticism of the classical Cox Model

A very crucial aspect of causal models in survival analysis is the preliminary stage, in which a set of explanatory variables must be properly chosen and designed, usually among, as in our real case, a very large number of alternatives. This part of the analysis is typically accomplished with the help of descriptive tools, such as plots of the observed hazard rates at the covariate values. However, it is often the case that such tools are not sufficiently informative. As a consequence, a large number of variables are included as predictors and a model selection procedure needs to be run in order to find a parsimonious linear combination.
Our claim is that classical Cox proportional hazard models may not be the best strategy for Customer Lifetime Value modelling . Some criticisms are:

- If repeated events occur, as in our case, a different model structure (e.g. based on counting processes) should be adopted.

- The Cox model assumes that every subject experiences an event at most once, and that the event times are independent. In our context, a subject can experience multiple events (e.g. a churn event in different times and locations), possibly with dependencies among the event times of the same individual. Modelling multiple event time data requires a different approach. An example of modelling multiple event time data was given by Gail, Santer and Brown (1980) with an application to mammary tumors.
- When many explanatory variables, possibly correlated, are specified, the *efficiency* of Cox's model selection and estimation becomes heavily dependent on the number of available observations. Variable selection is thus needed in a model selection step. However classical model selection chooses a model and then inferences on quantities of interest, such as $\lambda(t|z)$ are made *conditionally* upon the selected model. Consequently, model uncertainty is not taken into account and, thus, inference may be seriously biased.
- It may be difficult, particularly in observational studies, to have *complete* information on all relevant covariates. Furthermore, random effects, expressing accident proneness or *frailties* may affect inferences on fixed effects.

In this paper we shall show how to improve the classical Cox model for Customer Lifetime Value in two ways:

- Considering Bayesian Variable Selection and Bayesian Model Averaging to correctly take model uncertainty into account
- Introducing a multilevel multivariate survival model via stratification.

5 A two step Bayesian lifetime value model

5.1 Our Bayesian Variable Selection approach

To illustrate clearly our variable selection methodology, we shall assume first an exponential survival time, such that, for $i = 1, \ldots, n$: $\lambda_i(t) = \lambda_i$. It can then be shown that, given the observed data $\underline{y} = (y_1, \ldots, y_n)$, the likelihood of $\underline{\lambda} = (\lambda_1, \ldots, \lambda_n)$ is:

$$L(\underline{\lambda}) = \prod_{i \in \mathcal{U}} \lambda_i \exp\{-\sum_{i=1}^{n} \lambda_i t_i\},,$$

where $\mathcal{U} = \{i : \delta_i = 1\}$ are the uncensored subjects. Now, let g indicate a *partition* of the index set $\mathcal{I} = \{1, \ldots, n\}$, with d_g subsets $S_k(g)$, for $k = 1, \ldots, d_g$. Clearly, given the correspondence between $\mathcal{I}, \underline{y}$ and $\underline{\lambda}$, g also defines a partition of the data and of the hazard functions. Notice that the likelihood in (4) assumes all λ_i to be distinct and, thus, is in fact conditional on the *independence* partition $g_{ind} = \{\{1\}, \{2\}, \ldots, \{n\}\}$, containing $d_g = n$ separate subsets S_i, each with $n(S_i) = 1$ observations. For this reason, it can be indicated by $L(\underline{\lambda}|g_{ind})$.
A different likelihood arises when all hazards can be set equala fixed equal to a common rate, say μ. This situation occurs when *no* covariate or frailty affects the

survival times and corresponds to considering all data to be *exchangeable*. The resulting likelihood can be seen as conditional on the partition $g_{exc} = \{1, \ldots, n\}$, containing a single subset S_1 (with $n(S_1) = n$):

$$L(\mu|g_{exc}) = \mu^d exp\{-\mu V\},,$$

with $d = \sum_{i=1}^{n} \delta_i$ the total number of failures and $V = \sum_{i=1}^{n} t_i$ the overall time at risk.

Apart from the above situations, which can be regarded as somewhat extreme, survival analysis is typically concerned with a plurality of effects that may induce dependencies among survival times. Such effects may be either observable (possibly with some missing values) or unobservable. In any case, when relevant, they *induce* a partition of the observations, by associating different hazards to individuals having the same level of the factor. In our approach, we will entertain several partition structures, each induced by the levels of a potential prognostic factor. This amounts to considering a collection of alternative *partial exchangeability structures* for the survival times. Our model consists of two parts: a *likelihood* specification and a *hierarchical prior* distribution on the partition structure as well as on the corresponding set of hazards. Conditionally on a *general* partition g, let $\lambda_i = \mu_k$, $\forall i \in S_k(g)$. Consequently, the likelihood of the hazards $\underline{\mu} = (\mu_1, \ldots, \mu_{d_g})$ is the following:

$$L(\underline{\mu}|g) = \prod_{k=1}^{d_g} \mu_k^{d_k} \exp\{-\mu_k V_k\}, \tag{4}$$

where, for $k = 1, \ldots, d_g$: μ_k, $d_k = \sum_{i \in S_k(g)} \delta_i$ and $V_k = \sum_{i \in S_k(g)} t_i$ are the hazard, death and risk set of the k-th partition subset. On the other hand, the prior specification requires the definition of a class of possible partitions $\mathcal{G} = \{1, \ldots, G\}$.

Once $\mathcal{G}$ is specified, it is necessary to assign a probability distribution on both $\underline{\lambda}|g \in \mathcal{R}^{d_g}$ and $g \in \mathcal{G}$. Specifically, conditionally on a partition g we shall take, for $k = 1, \ldots, d_g$ and $\forall i \in S_k(g)$:

$$\mu_k \overset{\text{ind}}{\sim} Gamma(r_k m_k, r_k), \tag{5}$$

with m_k and r_k known positive constants. Finally, a simple probability function on $\mathcal{G}$ would take $p(g)$ to be uniformly spread among partitions, i.e. $p(g) = G^{-1}$. Our first aim is to evaluate the importance of each prognostic factor. This can be achieved by calculating, given the observed evidence $\underline{y}$, the posterior probability of each partition, $p(g|\underline{y})$. Following (4) and (5) it can be shown that:

$$p(\underline{y}|g) = \prod_{k=1}^{d_g} \frac{(r_k)^{r_k m_k}}{\Gamma(r_k m_k)} \frac{\Gamma(r_k m_k + d_k)}{(V_k + r_k)^{r_k m_k + d_k}},$$

Furthermore, Bayes theorem gives $p(g|\underline{y}) \propto p(\underline{y}|g)p(g)$, from which $p(g|\underline{y})$ is obtained by normalisation.

Our second aim is to estimate the hazard function, in order to make predictions on survival times. This task can be performed in two steps: first we work conditionally on a partition, and determine a Bayesian estimate of each individual hazard, by calculating the posterior mean $E(\lambda_i|\underline{y}, g)$. Computationally, following (4) and (5), it turns out that, for $i \in S_k(g)$:

$$E(\lambda_i|\underline{y}, g) = \frac{r_k\, m_k + d_k}{V_k + r_k},$$

The above expression shows that r_k and $r_k\, m_k$ can be interpreted space respectively, as pre-experimental total time at risk and observed events (e.g. coming from a meta-analysis). When no prior information is available, they may be taken in an appropriate *uninformative* manner. The second step of the estimation procedure involves using $p(g|\underline{y})$ to calculate the marginal posterior expectation of each individual hazard $E(\lambda_i|\underline{y})$, via the law of total probabilities:

$$E(\lambda_i|\underline{y}) = \sum_{g=1}^{G} E(\lambda_i|g, \underline{y}) p(g|\underline{y}).$$

As shown, for instance, in Raftery *et al* (1995), using the marginal posterior expectation via the above model averaging procedure leads to predictions better than those based on conditioning on a single partition, such as that associated with the best model. The procedure that we have shown for a constant hazard can be easily generalized for a counting process framework, as are shall show in the next section.

5.2 Survival analysis in the Point Processes framework

Several authors have discussed Bayesian inference for censored survival data where the integrated baseline hazard function is to be estimated nonparametrically, see e.g. Kalbfleisch and Prentice (1980). In particular, Clayton (1994) formulates the Cox Model using the counting process notation introduced by Andersen and Gill (1982) and discusses estimation of the baseline hazard and regression parameters using a Bayesian approach based on Markov Chain Monte Carlo. Although his approach may appear somewhat contrived, it forms the basis for extensions to random effects frailty models, time-dependent covariates, smoothed hazards, multiple events and so on.
Here we follow Clayton's guidelines and propose a methodology based on counting processes. In particular the counting process associated with a point process is characterized by a dynamic process (intensity), and a special pattern of incompleteness of observations (right-censoring or left-truncation in our case). This characterization is an application of the well known Doob-Meyer decomposition theorem. Having defined the intensity process, one is interested in estimation of its parameters.
Inferential procedures in this framework were first presented in Aalen (1975), and turned out to be very fruitful. For further developments, see Andersen et al.

(1993). A counting process is a stochastic process $\{N(t) : t \geq 0\}$ adapted to a self-exciting filtration $Im_t : t \geq 0$ with $N(0) = 0$ and $N(t) < \infty$ a.s., and whose paths are, with probability one, right-continuous, piecewise constant, and have only jump discontinuities, with jumps of size one.
For the derivation of the likelihood function we follow a well developed theory leading to a Poisson type of likelihood (see e.g. Andersen et. al. (1993), or Fleming and Harrington (1991)). The justification is based on Jacod's Formula for the likelihood ratio.
Suppose that the y^{th} individual in the i^{th} cluster survival time T^{ij} is an absolutely continuous random variable conditionally independent of a right censoring time Z_{ij} given the covariates x_{ij} and frailty w_i. Let $Vij = min(T^{ij}, Z_{ij})$ and $\delta_{ij} = I(T_{ij} \leq Z_{ij})$ denote the time to the end-point event and the indicator for the event of interest to take place, respectively. Suppose that $(Vij, \delta_{ij}, x_{ij}, w_i)$ are i.i.d, for $i = 1, ..., n$ and $j = 1, ...m_i$, and the conditional hazard function of T_{ij} given x_{ij} and w_i satisfies the additive exponential linear hazard model. For subject j in cluster i, let $N_{ij}(t) = 1$ if $\delta_{ij} = 1$ in interval $[0, t]$ and $N_{ij}(t) = 0$ otherwise, and let $Y_{ij}(t) = 1$ if the subject is still exposed to risk at time t and $Y_{ij}(t) = 0$ otherwise.
Hence, we have a set of $N = \sum_n^{i=1} m_i$ subjects such that the counting process $\{N_{ij}(t); t \geq 0\}$ for the j^{th} subject in the i^{th} cluster set, records the number of observed events up to time t. Letting $dN_{ij}(t)$ denote the increment on $N_{ij}(t)$ over the small interval $[t, t + dt]$, the likelihood (6) of the data conditional on w_i is proportional to:

$$\prod_{i=1}^{n}\prod_{j=1}^{m_i}\left(\prod_{t\geq 0} Y_{ij}(t)w_i\left[h_0(t|\lambda_0) + h_1(t|\theta)\right]^{dN_{ij}(t)}\right)$$
$$\times exp\left(-\int_{t\geq 0} Y_{ij}(t)w_i\left[h_0(t|\lambda_0) + h_1(t|\theta)\right]\right). \quad (6)$$

Since we allow each $N_{ij}(t)$ to take at most one jump for each subject, note that $dN_{ij}(t)$ contribute to the likelihood in the same manner as independent Poisson random variables even though $dN_{ij}(t) \leq 1$ for all i, j and t.
Suppose that the time axis is partitioned into $g + 1$ disjoint intervals $I_1, ..., I_{g+1}$ where $I_k = [a_{k-1}, a_k)$ for $K = 1, 2, ..., g + 1$, with $a_0 = 0$ and $a_{g+1} = \infty$. In the K^{th} interval, given w_i, the j^{th} subject in the i-th cluster has hazard form $w_i\{h_0(t_{ij}|\lambda_{0k}) + h_1(t_{ij}|\theta_{ij})\}$, $K = 1, ..., g_{ij}$ where g_{ij} denotes the number of partitions of the time interval for the j^{th} subject in the i^{th} group.
Given the complete data (T, w), where $T = \{t_{ij} : i = 1, ..., n_i; j = 1, ..., m_i\}, w = (w_1, ..., w_n)$, the likelihood can be re-expressed as

$$\prod_{i=1}^{n}\prod_{j=1}^{m_i}\prod_{k=1}^{g_{ij}}\prod_{t\in(a_{k-1},a_k)}\left[Y_{ij}(t)w_i\left[h_0(t|\lambda_0) + h_1(t|\theta)\right]^{dN_{ijk}}\right]$$
$$\times exp\left(-\int_{t\geq 0} Y_{ij}(t)w_i\left[h_0(t|\lambda_0) + h_1(t|\theta)\right]\right), \quad (7)$$

where dN_{ijk} is the change in the count function for j^{th} subject in the i^{th} group in the interval k. Note that $h_0(t|\lambda_0)$ is the baseline hazard function and $h(t|\theta)$ is the parametric part, with $\theta = X\beta$. Under the assumption that the risk occuring in the interval I_k is small, i.e.,

$$\int_{a_k}^{a_{k-1}} Y_{ij}(t)\left[h_0(t|\lambda_0) + h_1(t|\theta)\right] dt \approx 0,$$

for all i,j,k, the likelihood contribution across this interval for individuals at risk is approximately

$$\left\{ w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right] \right\}^{dN_{ijk}} \times exp\left(-w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right] \right),$$

where $dH_{0k} = \int_{ak-1}^{a_k} h_0(t)dt$ is the usual cumulative baseline intensity function for the k^{th} interval.
We remark that the previous likelihood is essentially Poisson in form, reflecting the fact that the likelihood may be thought of as being generated by independent contributions of many data atoms each concerned with the observation of an individual over a very short interval during which the intensity may be regarded as constant and approximately zero (for a review of this point, see e.g. Clayton, 1994). Therefore, we replace the previous equation with:

$$\begin{aligned} &\prod_{n}^{i=1}\prod_{m_i}^{j=1}\prod_{Y_{ijk}=1} \left\{ w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right] \right\}^{dN_{ijk}} \\ &\times exp\left(-w_i \left[dH_{0k} + \frac{1}{\theta_{ij}}(a_k - a_{k-1}) \right] \right), \end{aligned} \tag{8}$$

where $Y_{ijk} = 1$ if the j^{th} subject in the i^{th} group is exposed to risk at time $t \in (a_{k-1}, a_k]$, and $Y_{ijk} = 0$. We now present a Bayesian version of the counting process model introduced before. To formulate a Bayesian specification of the model, prior distributions are needed for the vector parameters λ_0 and θ and the hyperparameters β, σ_θ^2.
We assume for the elements of λ_0 , $\lambda_0 = (\lambda_{00}, ..., \lambda_{0g})'$, independent Gamma priors, i.e.,

$$(\lambda_{0k}|a_{0k}, b_{0k}) \overset{\text{ind}}{\sim} Ga(a_{0k}b_{0k}), K = 1, 2, ..., g_{ij},$$

where g_{ij} denotes the number of partitions of the time interval for the j^{th} subject in the i^{th} group, a_{0k}/b_{0k} is the prior expectation for λ_{0k} and a_{0k}/b_{0k}^2 is the prior variance. For β we choose the usual Normal-Inverse Gamma conjugate priors, i.e.

$$\beta|\sigma_\theta^2 \overset{\text{ind}}{\sim} N_p(m_\theta, \sigma_\theta^2 V_\theta),$$

with $\sigma_\theta^2 \overset{\text{ind}}{\sim} Ga(a_\theta, b_\theta)$. In order to estimate the posterior distributions we have implemented a Gibbs sampling procedure.

6 A one step Bayesian lifetime value model

Methods for analyzing survival data, in contrast with the previous Section, a Bayesian model averaging approach allows to derive a one-step procedure for estimation of a lifetime value model, focus on modeling the hazard rate through proportional hazards model.
Since the integrals required for BMA do not have a closed-form solution for Cox models, Raftery, Madigan and Volinsky (1996) and Volinsky et. al (1997) adopted a number of approximations for Bayesian Model averaging. In particular they showed that it is possible to use the MLE approximations:

$$p(\Delta|M_k, D) \approx p(\Delta|M_k, \hat{\beta}_k, D),$$

and the Laplace approximation,

$$\log p(D|M_k) \approx \log p(D|\hat{\beta}_k, M_k) - d_k \log(n),$$

where d_k is the dimension of β_k and n is usually taken to be the total number of cases. This is the Bayesian Information Criterion (BIC) approximation.
To implement BMA for Cox Models, we have followed Raftery et. al (1999) and used an approach similar to the Occam's window method, implemented in a set of R routines provided for Bayesian Model Averaging. To efficiently identify good models, we adapted the leaps and bounds algorithm of Furnival and Wilson (1974) which was originally created for linear regression model selection. The leaps and bounds algorithm provides the top q models of each model size, where q is designated by the user. The MLE $\hat{\beta}_k$ and $var(\hat{\beta}_k)$ for each model M_k are also returned. Lawless and Singhal (1978) and Kuk (1984) provide a modified algorithm for nonnormal regression models that gives an approximate likelihood ratio test statistic and hence an approximate BIC value. With BMA it is possible to have for each model a BIC, the posterior probability and for each parameter the relative mean, the variance and also the posterior probability that a Cox regression coefficient for a variable is nonzero (posterior effect probability) as the sum of posterior probabilities of the models which contain that variable.
The posterior mean (9), following Raftery et al. (1999), of a regression coefficient can be shown to be:

$$\begin{aligned}
\hat{\theta}_{BMA} &= E_M(\hat{\theta}) = \sum_{i=1}^{K} \hat{\theta}_i P(M_i|D) \\
&= \frac{\sum_{i=1}^{K} \hat{\theta}_i P(M_i|D)}{\sum_{i:\theta_i \in M_i} P(M_i|D)} \times \sum_{i:\theta_i \in M_i} P(M_i|D) \\
&= E(\hat{\theta}|\theta_i \in M_i) \times P(\theta \neq 0), \qquad (9)
\end{aligned}$$

which is the conditional posterior mean of θ multiplied by its posterior probability.
Similarly, to calculate the variance of the regression coefficient, let $p_i = P(M_i|D)$

and $V_i = Var(\hat{\theta}|M_i, D)$. Then:

$$
\begin{aligned}
V(\hat{\theta}) &= E(\hat{\theta}^2) - (\sum_{i=1}^{K} p_i \hat{\theta}_i)^2 \\
&= \sum_{i=1}^{K} p_i (V_i + \theta_i^2) - (\sum_{i=1}^{K} p_i \hat{\theta}_i)^2 \\
&= \sum_{i=1}^{K} p_i V_i + [\sum_{i=1}^{K} p_i \hat{\theta}_i^2 - (\sum_{i=1}^{K} p_i \hat{\theta}_i)^2] \\
&= \sum_{i=1}^{K} p_i V_i + \sum_{i=1}^{K} p_i (\hat{\theta}_i - \sum_{i=1}^{K} p_i \hat{\theta}_i)^2, \qquad (10)
\end{aligned}
$$

Note that the first term is the weighted variance over models, but the overall variance is affected by the second term, which depends on how stable the estimates are across models. The more these estimates differ across models, the higher the posterior variance. In this way the standard errors reported for variables directly take into account model uncertainty.
Prior probabilities on both model space and parameter space are defined by this procedure. All models are considered equally likely *a priori* by the leaps and bounds algorithm. Using the BIC approximation to the integrated likelihood defines an implicit prior on all the regression parameters, as described before.
When there is little prior information about the relative plausibility of the models considered, the assumption that all models are equally a priori is a reasonable neutral choice. However, Spiegelhalter, Dawid, Lauritzen and Cowell (1993) provide a detailed analysis of the benefits of incorporating informative prior distributions in Bayesian Knowledge - based systems and demonstrate improved predictive performance with informative priors. When prior information about the importance of a variable is available for model structures the prior probability for model M_i can be specified as

$$
p(M_i) = \prod_{j=1}^{p} \pi_j^{\delta_{ij}} (1 - \pi_j)^{1-\delta_{ij}},
$$

where $\pi_j \in [0, 1]$ is the prior probability that $\beta_j \neq 0$ in a regression model and δ_{ij} is an indicator of whether or not variable j is included in model M_i. Assigning $\pi_j = 0.5$ for all j corresponds to a uniform prior across model space, while $\pi_j < 0.5$ for all j imposes a penalty for large models. Using $\pi_j = 1$ ensures that variable j is included in all models.
This approach was used to specify model priors for variable selection in linear regression in Dobson (1990) and suggested for model priors for BMA in Cox models by Raftery et al. (1999).

7 Stratified Cox Models

The Cox model has become the most used procedure for modeling the relationship of covariates to a survival or other censored outcomes (Therneau and Grambsch, 2000). Its form is flexible enough to allow time-dependent covariates as well as frailty terms and stratification, but it has some restrictions.
One of the restrictions of using a Cox model with fixed in time is its proportional hazards (PH) assumption, that is, that the hazard ratio between two covariate values has to be constant over time (this is due to the common baseline hazard function canceling out in the ratio of the two hazards).
We now review tools available to assess whether hazards can be considered proportional (PH assumption) for all covariates. For binary covariates, as in our case, a comparison of nonparametric survival curve estimates may be sufficient to decide on PH because if the hazards were proportional, the survival curves for the two conditions would separate exponentially, and the two curves would not cross each other. Non-PH would imply that the relative risk changes over time for subjects who churn versus subjects who do not churn during the temporal period of study.
For continuous covariates it is not sufficient to rely only on stratified survival estimates to assess PH because the choice of stratification points is subjective. In this case an alternative is via the use of time-varying coefficients (Grambsch and Therneau, 1994). That is, one or more coefficients multiplying their respective covariates varies with time. If the coefficient multiplying a covariate is not constant over time, the impact of that covariate on the hazard varies over time, leading to non-PH. Instead if PH holds, a plot of the coefficient versus time will be a horizontal line. Therefore, we can perform formal tests for specific forms of departure from PH. To illustrate formal tests of time-varying coefficients, we first describe the Schoenfeld (1982) residual, using the notation of Therneau and Grambsch (2000).
Let $t_1, ..., t_d$ be d unique ordered event times, and let $X_i(s)$ be the $p \times 1$ covariate vector for the $i-th$ individual at time s. For time-fixed covariates, this is just X_i. Also, define the weighted mean of the $X_i(s)$ over those still at risk at time s as:

$$\overline{x}(\hat{\beta}, s) = \frac{\sum Y_i(s) exp(X_i(s)\hat{\beta}) X_i(s)}{\sum Y_i(s) exp(X_i(s)\hat{\beta})},$$

where $Y_i(s)$ is the predictable variation process indicating whether observation i is at risk at time s, so that $Y_i(s) = 1$ if observation i is still at risk at time s and is zero otherwise. The estimate $\hat{\beta}$ comes from fitting a Cox PH model. In particular, a Schoenfeld residual is a $p \times 1$ vector that is defined at the $k-th$ event time as:

$$s_k = \int_{t_{k-1}}^{t_k} \sum_i \left[X_i(s) - \overline{x}(\hat{\beta}, s) \right] dN_i(s),$$

where $N_i(s)$ is a counting process that counts the number of events for observation i at time s. Thus, s_k sums the quantities $X_i(t_k) - \overline{x}(\hat{\beta}, t_k)$ over observa-

tions that have experienced the event by time t_k. With no tied event times, the $k-th$ Schoenfeld residual is the sum of contributions to the derivative of the log partial likelihood by subjects who have experienced events by t_k (Hosmer and Lemeshow, 1998).
There are several options for attempting to correct non-PH or to use alternatives to a PH model. An option is to use an accelerated failure time (AFT) model. Therneau and Grambsch (2000) show these models can be detected by the time-varying coefficient tests mentioned in this section. AFT models are most appropriate in settings in which the time scale of the hazard function is either slower or faster (multiplicatively) than the time scale on which the measurements are made, as the covariates act by expanding or contracting time by a factor $exp(X\beta)$.
Another alternative is to stratify the model across levels of one or more covariates, leading to a Stratified Cox model.

7.1 Bayesian Stratified fixed effects Cox models

The Classical Cox model (1) can be extended to account for stratification. When a factor does not affect the hazard multiplicatively, stratification may be useful in model building. The strata divide the subjects into disjoint groups, each of which has a distinct (arbitrary) baseline hazard function but common values for the coefficients β (Therneau and Grambsch, 2000). The hazard function for an individual i who belongs to stratum k is then:

$$\lambda(t; x_i) = \lambda_k(t) exp(x_i'\beta),$$

Typically, strata are naturally defined within the context of the problem. For example, in medical research, multi-center clinical trials typically stratify on the clinic in which they are conducted (Therneau and Grambsch, 2000).
The stratified Cox model also allows a deviation from proportional hazards, and as such provides an alternative to the assumption of proportional hazards. The hazard functions for two different strata do not have to be proportional to one another. However, within a stratum, proportional hazards are assumed to hold. We take advantage of this use of stratification for our data.
The partial likelihood for stratified Cox models with K strata becomes a product of K terms, each of the form of (3), but where i ranges over only the subjects in stratum k, $k = 1, , K$. Stratification entails fitting separate baseline hazard functions across strata. A baseline hazard function represents the hazard rate over time for an individual with all modeled covariates set to zero. With a stratified Cox model, a proportional hazards structure does not necessarily hold for the combined data, but is assumed to hold within each stratum. However, the coefficients on the included covariates are common across strata so that the relative effect of each predictor is the same across strata, unless there is a significant strata-by-covariate interaction, which means that the effect of the covariate differs within strata.
The estimated coefficients of a stratified Cox model can be computed using the

entire data set. A formal test of overall goodness-of-fit for the stratified Cox model was proposed by Parzen and Lipsitz (1999) and independently by May and Hosmer (1998). The test compares observed and (model-based) expected numbers of events within covariate risk groups and computes a chi-square test. Here we propose a Bayesian version for the Stratified Cox Model:

$$h_i(t) = h_{0i}(t)exp(\beta' x),$$

Stratum-specific baseline hazards, $h_{0i}(t)$ are assumed to be drawn from the Weibull family:

$$h_{0i}(t) = \rho_i t^{\rho_i - 1} exp(\rho_i \beta_{0i}),$$

where β_{0i} is a unit specific time-scale accelerator with prior $N(\mu_0, \sigma_0^2)$, μ_0 is a flat prior and σ_0^2 is an Inverse Gamma with specific value of parameter (e.g. 3,0.5). We remark that ρ_i is a unit-specific shape parameter. In particular, if $\rho_i > 1$ there is an increasing hazard, $\rho_i \sim Ga(\alpha, \alpha^{-1})$, with $\alpha \sim Ga(c, d)$. In our case we have chosen c=3, d=10. The posterior distribution function of the coefficients can be derived via Gibbs Sampling.

8 Application of Bayesian survival models

We now turn our attention towards the application of the presented methodologies for modelling survival risk. In our case study the risk concerns the value that derives from the loss of a customer. The objective is to determine which combination of covariates affect the risk function, studying specifically the characteristics and the relation with the probability of survival for each customer.

8.1 The available data

The data available for our analysis contains information that can affect the event time, such as demographic variables, variables about the contract, the payment, the contacts and geomarketing variables. The response variable, used as a dependent variable to build predictive models, includes two different types of customers: those who during the survey are active and those, instead, who regularly cancelled their subscription.
We remark that, due to the different nature of the withdrawal, SUSPENSION status customers, who have not paid the subscription, although not cancellers, cannot be simply included in a classical analysis but, rather, require a specific treatment, as in the survival analysis context.
The target variable has been observed 3 months after the extraction of the data set used for the model implementation phase, in order to verify correctly the effectiveness and predictive power of the models themselves. We have available 606 variables and 3.4 Million observations (customers), extracted from the company database.

Concerning explanatory variables, the variables used were taken from different databases used within the company, which contained, respectively: socio-demographic information about the customers; information about their contractual situation and about its change over time; information about contacting the customers (through the call centre, promotion campaigns, etc) and, finally, geo-marketing information (divided into census, municipalities and larger geographical sections information).
The variables regarding customers contain demographic information (age, gender, marital status, location, number of children, job and degree) and other information about customer descriptive characteristics: hobbies, PC possession at home, address changes.
The variables regarding the contract contain information about its chronology (signing date and starting date, time left before expiration date), its value (fees and options) at the beginning and at the end of the survey period, about equipment needed to use services (if they are rented, leased or purchased by the customer) and binary variables which indicate if the customer has already had an active, cancelled or suspended contract. There is also information about invoicing (invoice amount compared to different period of time 2, 4, 8, 12 months).
The variables regarding payment conditions include information about the type of payment of the monthly subscription (postal bulletin, account charge, credit card), as well as other info about the changes of the type of payment. The data set used for the analysis also includes variables which give information about the type of the services bought, about the purchased options, and about specific ad-hoc purchases, such as number and total amount of specific purchases during the previous month and the last 2 months.
The variables regarding contacts with the customer contain information about any type of contact between the customer and the company (mostly through calls to the call centre). They include many types of calling categories (and relatives subcategories). They also include information about the number of questions made by every customer and temporal information, such as the number of calls made during the last month, the last two months and so on. Finally, geomarketing variables are available although a great amount of work is involved in their pre-processing and definition.
Regardless of their provenence, all variables have gone through a pre-processing feature selection step aimed at reducing their very large number (equal to 606). This step has been perfomed using a combination of different techniques, going from dimensionality reduction to association measure ranking and stepwise selection. For more details, see Figini, 2006.

8.2 Classical Survival Analysis

In order to build a survival analysis model, we have constructed two variables: one variable of status (that distinguishes between active and non active customers) and one of duration (indicator of customer seniority) . The first step in the analysis of survival data consists of a plot of the survival function and of the hazard.

The application of the Kaplan Meier estimator to our data leads to the estimated survival function in Figure 1.

Figure 1 about here

From Figure 1 we note that the survival function has varying slopes, corresponding to different periods. When the curve decreases rapidly we have time periods with high churn rates; when the curve decreases slowly we have periods of loyalty. We remark that the final jump is due to a distortion caused by a few data in the tail of the lifecycle distribution.
In Figure 2 we show the hazard function, which shows how the instantaneous risk rate varies in time.

Figure 2 about here

From Figure 2 we note two peaks, corresponding to months 4 and 12, the most risky ones. Note that the risk rate is otherwise almost constant along the lifecycle. Of course there is a peak in the end, corresponding to what we observed in Figure 1.
Of great value, in business terms, is the calculation of the life expectancy of the customers. This can be obtained as a sum over all observed event times:

$$\sum_{j=1}^{T} \hat{S}(t_j) \times (t_j - t_{j-1}),$$

where $\hat{S}(t_j)$ is the estimate of the survival function at the j-th event time, obtained using the Kaplan Meier method, and t is a duration indicator. We remark that life expectancy tends to be underestimated if most observed event types are censored (i.e., no more observable).
We now move to the building of a full predictive model. We have chosen to implement first the classical Cox model. The number of variables available are 606. The result, following a stepwise model selection procedure, is a set of about twentyfive explanatory variables. Such variables can be grouped into three main categories, according to the sign of their association with the churn rate, represented by the hazard ratio:

- variables that show a positive association (e.g. wealth of the geographic regions, the quality of the call center service, the sales channel)
- variables that show a negative association (e.g. number of technical problems, cost of service bought, payment method)
- variables that have no association (e.g. equipment rental cost, age of customer, number of family components).

To better interpret the previous associations we consider the values of the hazard ratio under different covariate values. For example, for the variable indicating the number of technical problems we have compared the hazard function for those that have called at least once with those that have not made any. As the resulting ratio turns out to be equal to 0.849, the risk of becoming a churner is

lower for callers than for non callers.
A very important remark is that the Cox model generates survival functions that are adjusted for covariate values. More precisely, the survival function is computed according to the following

$$S(t, X) = S_0(t)exp(\sum_{i=1}^{p} \beta_i X_i),$$

Figure 3 shows a comparison between the survival curve obtained without covariates (the baseline, as in Figure 1) and the same curve adjusted for the presence of covariates.

Figure 3 about here

Figure 3 shows that covariates affect the survival time substantially up for to two years of lifetime; indeed the Cox survival curve (described by the symbols '+') is greater with respect to the baseline (described by the continuous curve). After this period the survival probability declines abruptly and turns out to be much lower for the remaining lifespan. Once a Cox model has been fitted it is advisable to produce diagnostic statistics based on the analysis of residuals to verify if the hypotheses underlying the model are correct. In our case they were found to be correct, so we could proceed with predictive modelling.
In the following prediction step the goodness of the model will be evaluated in terms of predictive accuracy in a cross-validation exercise. We first split the dataset in the two usual subsets: training and test. Both have been proportionally sampled with respect to the status variable. All sampled data contain information on all finally chosen explanatory variables (about twenty). In order to evaluate the predictive performance of the model, and compare it with classification trees (routinely used by the company), we have focused our attention to a 3 month ahead prediction. We have devised and implemented a procedure based on the estimated survival probabilities, aimed at building the confusion matrix (see e.g. Giudici 2003) and, correspondingly, the percentage of captured true churners of the model. We remark that this is not a fair comparison, as survival models predict more than a point; however company experts tipically ask for this type of model benchmarking.
In this paper, for brevity, we do not report full results on this model, but only the summary predictive performances; for more details see Figini (2006). Corresponding to each estimated probability decile, the percentage of true churners captured is : 0.104 in the first decile, 0.0745 in the second decile. In general, while in the first decile (that is, among the customers with the highest estimated churn probability) 0.10 of the clients are effective churners, the same percentage reduces in susequent deciles, thus giving an overall picture of good performance of the model. Indeed the lift of the model, as measured by the ratio between the captured true responses between the model and a random allocation, does not turn out to be substantially better with respect to what was obtained with the tree models. However, we remark that, in constrast to what occurred with the latter models, the customers with the highest estimated churn rate are now

not necessarily those whose contract is close to the deadline. This is the most beneficial advantage of the survival analysis approach, which, in turn, leads to substantial gains in campaign costs. A further advantage of the survival analysis approach lies in its immediate translation in terms of lifetime value analysis, as we shall see in the next subsection.

8.3 Bayesian Variable selection: results

We now show the results concerning our proposal for Bayesian Feature selection, in a two step approach. We run this procedure for each variable in the data set. In Table 1 we show the most important covariate to predict Customer Lifetime Value.

Table 1 about here

We observe that the most important variable for predicting churn risk are about information on disconnection; this means that when a customer contacts the call-center to stop the contract. The second concerns the decoder's usage, and then covariates for payment method, promotion and special offers.

8.4 Two step Bayesian survival analysis

We now proceed with Bayesian modelling, in the two step context. We apply the method explained earlier to select variables. We have built and evaluated, for each covariate, a measure of importance. We report the results in Table 2 for the most important variables. It is possible to see that the most important variables to explain churn, concern information on disconnection, decoder rental , payment method, promotions, sale channel and contact with the call center.
After feature selection we used WinBUGS to implement a Bayesian counting process model using the Gibbs sampler. Table 2 shows the results.

Table 2 about here

In particular for each covariate selected by our Bayesian feature selection approch we have calculated, for each parameter, the mean , the standard deviation, the Monte Carlo error, the median and the Bayesian credible interval. We have estimated our models with different MCMC chains. The most stable result is with 10000 iterations and 500 iterations as a burn-in.
We have then used the idea of parallel multiple chains to check the convergence of the Gibbs sampler, following Gelman and Rubin (1992). In particular, to generate the Gibbs posterior samples, we have used three parallel chains. Monitoring convergence of the chains, has been done via the Brooks and Gelman (1998) convergence-diagnostic-graph.
For each of the 3 chains WinBUGS provides estimated parameters as a function of the iteration number (see e.g. Figure 4 and Figure 5).

Figure 4 and Figure 5 about here

Inspection of the diagnostic graphs for the two most important covariates (Figure 5 and 6), show that BGR converges to one. This shows that convergence is achieved for the two most important variables. We remark that this result is achieved for all covariates in the model. As is well known the WinBUGS software offers also a graph of the autocorrelation function (ACF); the autocorrelation plot illustrates dependence between subsequent simulated observations. In our case, the ACF indicate fairly rapid mixing and thus good convergence of the parameter space with a reasonably small number of iterations. They are suppressed in this paper for lack of space. We also remark that for the model in Table 2, the estimated correlations between parameters are quite low. In order to compare classical and Bayesian Cox models we have run both model comparison approaches. The results are shown in Table 3.

Table 3 about here

In Table 3 the first column is the variable, the second and the third are the estimated mean and standard deviation in the Bayesian model and the last two column are relative to the classical estimation for each variable, reporting the parameter MLE, the standard deviation and the p-value. As it is possible to see from the p-value, the variables β_3, β_4 and β_7 are equal to zero. The parameters β_3, β_4 and β_6 have different estimation in the two approaches. The BIC for the Bayesian Cox model is equal to 6583.411 and for the classical semi-parametric Cox model is equal to 6165.974. In particular, in Table 3, the variance of the estimates is quite high in the Bayesian model. To demonstrate the consistency of our proposed method for feature selection we next compare the previous results with the results from the one step approach.

8.5 One step Bayesian survival analysis

This section focuses on the application of Bayesian Model Averaging to our dataset. For computational reasons we have preselected 25 covariates which correspond to those that would be selected in a classical feature selection approach. We have then compared the feature selection obtained from BMA with our proposed approach. We recall that there are 2^{25} possible models; we fit the models and averaged over them to get parameter estimates and posterior probabilities of the parameters. Table 3 shows the top 3 models. Note that this models include % of the overall posterior probability, so that not much information is lost by reducing the model space. Table 4 shows the Bayesian model averaging computation for the covariates selected by our approach.

Table 4 and Table 5 about here

From Table 5 we can see the posterior probabilities for each model and the number of selected variables for each model.

Table 6 about here

From Table 6 we can see the posterior probabilities for each model and the number of selected variables for each model after the feature selection process.

Table 7 about here

As we can see, after Bayesian feature selection the model results are better in terms of posterior probabilitity.

8.6 Bayesian stratified survival models: results

Before we consider the application of stratified models, we show the results concerning the residuals, in Figure 6 and Figure 7.

Figure 6 about here

Figure 7 about here

In particular as we can see from Figure 7 there is evidence of violations of the Proportional Hazard assumption. In fact, there are some customers with different characteristics for the decoder and for the special offers. This information can be used to build a new model with stratification. First, in Table 7 we present results from a Classical Stratified Cox Model.

Table 7 about here

As we can see in Table 7, the first column is the variable selected by the classical stratified procedure, the second is the relative estimate for each variable, then the hazard ratio and finally the sign of the association between each variable and the target variable (a mixture of ralathionship duration and customer status). As we can see, for some variables there are computational problems for the estimation. In particular, for smart card offers and old customer, the estimate and the hazard ratio are undefined. This is because the algorithm does not always converge.

Table 7 about here

We now implement a new procedure, where we propose a natural extension of the classical stratified Cox Model in a Bayesian paradigm, as shown in Section 7.
The results are provided in Table 8 where p is the probability of inclusion for each variable across the models, EV is the expected value for each variable, derived from the Bayesian Model, and finally for each model we provide the parameter estimates. At the last rows of the table, we have estimates for each stratus variables.

Table 8 about here

We can compare the results in Table 2 with the results shown in Table 8, based on one step Bayesian Model Averaging. In particular the variance of estimates across models is lower than in the previous one step Bayesian Model Averaging. If we look at the results, the estimates across the models in Table 8, are very similar. This suggest some new theoretical field of research in order to improve this approach.
Table 9 shows the best 5 models found by our Stratified fixed effects Cox Bayesian Model averaging.

Table 9 about here

Table 9 presents for each model its posterior probability and its dimension based on the number of covariates. As we can see the stratified model is more parsimonious and therefore better in a business context.

8.7 Estimation of customer lifetime value

We now employ the results from Bayesian stratified survival analysis modelling to create models that allow us to estimate the lifetime value of each customer, or, in a perhaps more useful aggregated analysis, for each group of customers.
In other words, survival analysis is useful to quantify, in precise monetary terms, how much is gained or how much is lost by moving through different strata corresponding to different survival curves. For instance,how much is gained/lost if 8% of the clients, say, switch from buying service A to buying service B. Or, similarly, the relative gains when a certain percentage of clients change method of payment (e.g. moving between direct debit, credit card and postal order).
In order to quantify gains and losses, a simple measure is to calculate the area between the two corresponding survival curves, as shown in Figure 8 below. Suppose the two survival curves correspond to two different services bought, say black and grey, corresponding to the colors of the two curves.

Figure 8 about here

In order to determine exactly the area in Figure 8 we need to specify a temporal period ahead, e.g. 13 months. In Figure 8, the difference between survival probabilities after 13 months of life of the customers (e.g. 13 months since the first contact), is equal to 0.078. This value should be multiplied by the difference in business margin between the two methods of payment, as given, for example, by the difference in costs. Such costs can be described by a gain table as in Table 10.

Table 10 about here

From Table 10, a value of A is the relative gain if the client switches from PO to CC and, similarly, B and C corresponds to relative gains switching from PO to RID, and CC to RID where PO = postal order, CC= payment through credit card, RID= payment through banking account.
In terms of Figure 8, if we assume that we start with an acquired client base of 1000 customers in both categories (product black buyers and product grey buyers), the results say that, after 13 months we will remain with 934 black and 856 grey. If the finance department tell us that product black is worth 10 euros and product grey 20 euros we have that, after 13 months, we lose 660 euros for black churners and 2880 for grey churners. In other words, the priority of the marketing department should be to build targeted campaigns for grey product clients. From a different perspective, if black and grey correspond to two different selling channels of the same product, or to two different geographical areas, it is clear that the black channel (or area) is much better in terms of

customer retention. Often promotional campaigns are conducted looking only at increasing the customer base. Our results show that the number of captured clients should be traded with their survival or, better, lifetime value profile.

9 Future Research

In this paper we have presented a comparison between classical and Bayesian methodology to predict rates of churn of customers. Our conclusions show that the Bayesian approaches we have proposed, based on survival analysis modelling, lead to more robust conclusions and can be extendedmore easily to more complex frameworks. Our results also show that our Bayesina survival analysis models are a much powerful tool for lifetime value analysis and, consequently, for the actual planning of a range of marketing actions that impact on both perspective and actual customers.
However we believe that although Bayesian survival analysis is a very promising tool in the area, further research is indeed needed, both in applied and methdological terms. From an applied viewpoint, directions to be further investigated cocern the application of the methodology to a wider range of companies (we have studies in progress in the banking sector). From a methodological viewpoint further research is needed on the robustification of Bayesian Cox model (two step model and one step model). We are investigating the usage of a random effect stratified approach within this context.
In our future research we plan to experiment Bayesian feature selection via penalized likelihood approach.

10 Acknowledgment

This work has been supported by MUSING 2006 contract number 027097, 2006-2010 and PRIN Project 2005-2006 on Data Mining for E-business applications.

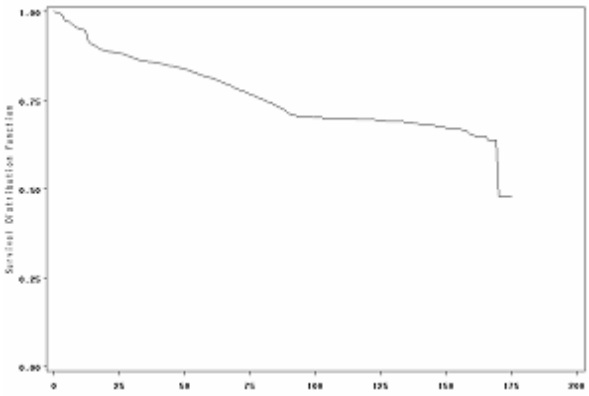


Fig. 1. Descriptive Survival function

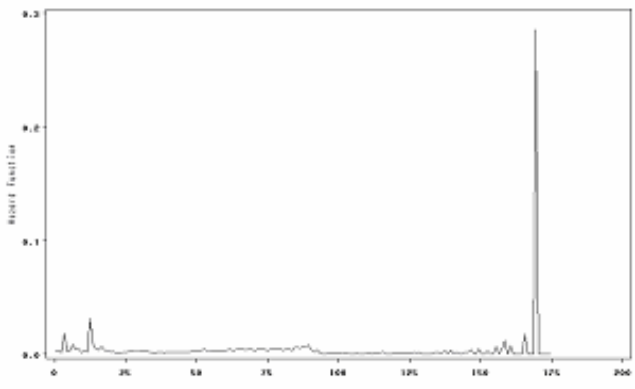


Fig. 2. Hazard function

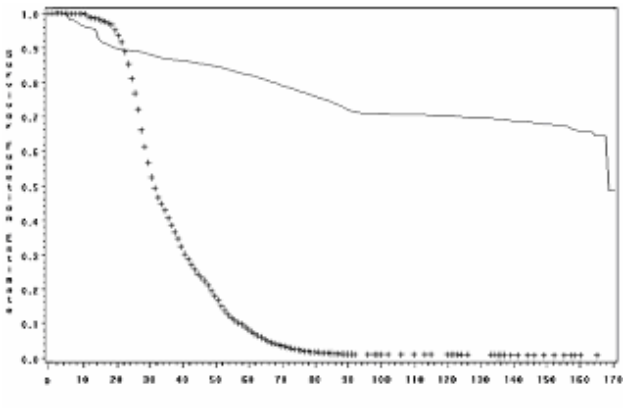


Fig. 3. Comparison between survival functions

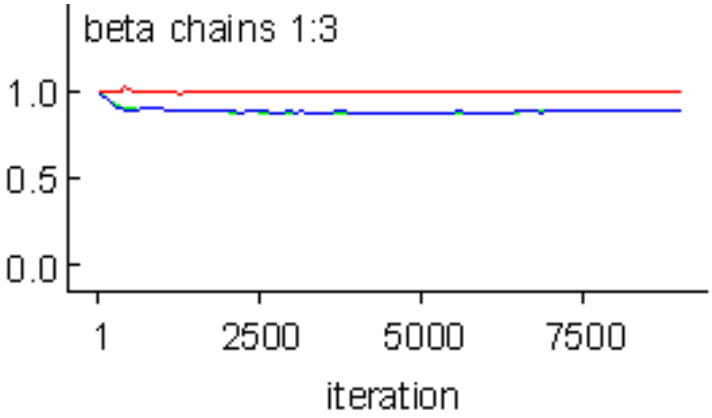


Fig. 4. Diagnostic for information on disconnection

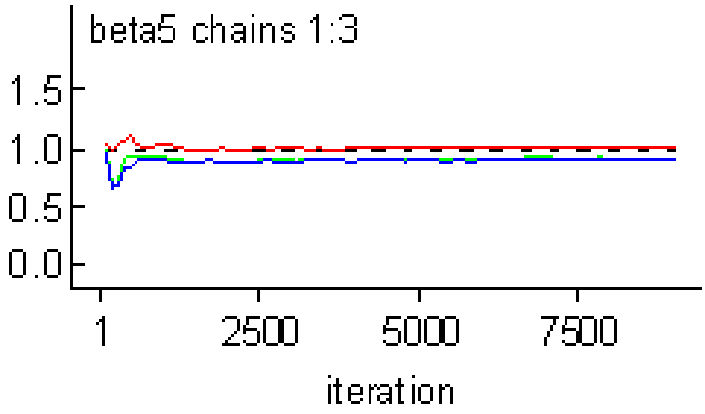


Fig. 5. Diagnostic for method of payment

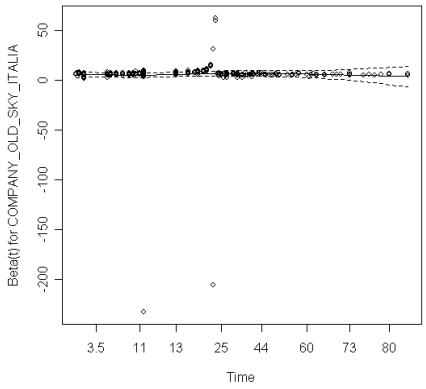


Fig. 6. Shoenfeld residuals for company history

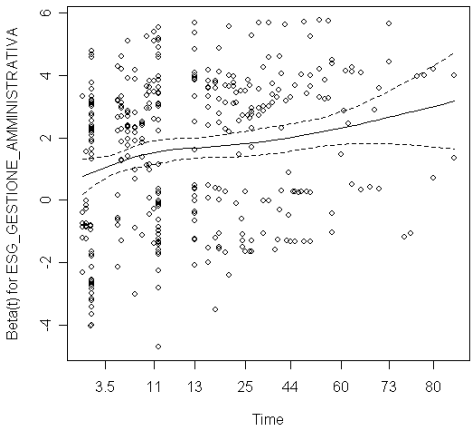


Fig. 7. Shoenfeld residuals for amministrative esigence

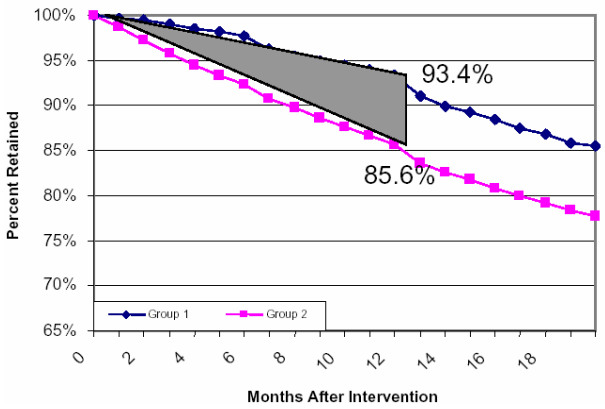


Fig. 8. Evaluation of gain/losses by comparing survival curves

$Variable$	$p(y\|g)$	$p(g\|y)$
β info disconnection	0.2451	0.0472
β_2 decoder sold	0.2452	0.0472
β_3 decoder rental	0.2466	0.0475
β_4 payment credit card	0.2497	0.0481
β_5 promotion	0.2514	0.0484
β_6 chanell of sell	0.2588	0.0491
β_7 ex decoder rental	0.2521	0.0488
β_8 special offers	0.2835	0.0546

Table 1. Bayesian Feature Selection results

$Variable$	$Mean$	Sd	$MCerror$	0.25	$Median$	0.975
β	0.7769	0.2123	0.00139	0.3547	0.7831	1.162
β_2	-1.632	2.223	0.08101	-5.938	-1.688	3.186
β_3	-1.731	0.6359	0.0308	-2.991	-1.718	-0.4818
β_4	-2.203	0.8412	0.04174	-3.715	-2.25	-0.3793
β_5	-1.368	0.6166	0.02468	-2.514	-1.396	-0.1127
β_6	-0.7287	1.626	0.09111	-3.206	-0.9579	3.382
β_7	-1.494	0.6678	0.02963	-2.845	-1.48	-0.215
β_8	0.67	2.141	0.1202	-3.957	0.6207	4.817

Table 2. Two step model: parameter estimation from the Bayesian Cox Model

$Variable$	$Mean$	Sd	$Estimate$	Sd	$p-value$
β	0.7769	0.2123	0.9396	0.2052	0.0001
β_2	-1.632	2.223	-1.4215	0.2647	0.0001
β_3	-1.731	0.6359	0.1164	0.1159	**0.3155**
β_4	-2.203	0.8412	0.2396	0.1356	**0.0772**
β_5	-1.368	0.6166	-0.8086	0.1510	0.0001
β_6	-0.7287	1.626	1.9636	0.1748	0.0001
β_7	-1.494	0.6678	-0.3392	0.1542	**0.0278**
β_8	0.67	2.141	1.0876	0.1206	0.0001

Table 3. Comparison of estimates from classical and Bayesian Cox models

$Variable$	p	EV	$Model_1$	$Model_2$	$Model_3$
info activation	100	1.0783	1.1152	1.0793	1.0826
info amministrative	100	1.5323	1.5274	1.5343	1.5134
β **info disconnection**	100	0.8512	0.8640	0.8596	0.8703
technical problem	100	-0.5071	-0.5159	-0.5098	-0.5078
β_5 **promotion**	100	-0.8985	-0.8963	-0.8920	-0.8749
β_6 **channel of sell**	100	1.6203	1.6415	1.6243	1.6220
β_4 **payment with credit card**	100	-0.6285	-0.6356	-0.6223	-0.6435
geographical area	100	0.3976	0.3991	0.3899	.
β_8 **special offers**	100	3.1730	3.1294	3.1790	3.1676
β_7 **ex decoder rental**	100	-2.1571	-2.7982	-1.6625	-2.8616
β_3 **decoder rental**	50.1	-0.6230	-1.3120	.	-1.3717
β_2 **decoder sold**	59.7	-0.3894	-0.9544	.	-1.0077

Table 4. One step model: Bayesian Model averaging results

$Model$	$PosteriorProbability$	$nVar$
Model_1	0.299	12
Model_2	0.166	10
$Model_3$	0.150	11
$Model_4$	0.148	10
$Model_5$	0.139	9

Table 5. One step model: the best 5 models

$Model$	$PosteriorProbability$	$nVar$
$Model_1$	0.456	5
$Model_2$	0.395	6
$Model_3$	0.071	6
$Model_4$	0.054	7
$Model_5$	0.023	6

Table 6. One step model: the best 5 models after feature selection

Variable	*Estimate*	*Hazardratio*	*Association*
info activation	1.018	2.77	+
β info disconnection	1.1577	3.18	+
technical problem	-0.4131	0.662	-
contractual variation	-0.1605	0.852	NA
Moovie package	0.1889	1.210	NA
β_8 special offers	3.2696	26.3	+
special discount offers	3.4625	31.9	+
info amministrative	1.6123	5.01	+
payment methods	-0.1493	0.861	NA
β_5 promotion	-0.9596	0.383	-
Sport package	-0.086	0.917	NA
payment with bancomat	0.8063	2.24	+
geographical area	0.4246	1.53	+
smart card offers	19.184	∞	∞
old customer	6.1058	∞	∞

Table 7. Classical Stratified Cox Model: results

$Variable$	p	EV	$Model_1$	$Model_2$	$Model_3$
info activation	100	1.0977	1.1	1.1	1.1
β **info disconnection**	100	0.8904	0.89	0.9	0.88
technical problem	100	-0.5073	-0.51	-0.51	-0.51
contractual variation	5.7	-0.0071	.	.	.
Moovie package	8.9	0.0191	.	0.21	.
β_8 **special offers**	100	3.2182	3.2	3.2	3.2
special discount offers	100	2.9434	2.9	2.9	3.0
info amministrative	100	1.5115	1.5	1.5	1.5
payment methods	4.2	0.0045	.	.	.
β_5 **promotion**	100	-0.9436	-0.94	-0.94	-0.96
Sport package	6.0	-0.0077	.	.	-0.13
payment with bancomat	100	0.7970	0.8	0.79	0.79
geographical area	100	0.4168	0.42	0.41	0.41
special card offers	100	3.4814	3.5	3.5	3.5
old customer	100	5.4121	5.4	5.4	5.4
$Rental_1$	.	-1.4896	-1.5	-1.5	-1.5
$Rental_2$	.	0.0215	0.022	0.029	0.014
$Rental_3$	.	0.3732	0.37	0.37	0.37
$Rental_4$	.	1.2731	1.3	1.3	1.3
$Channel_1$	.	-1.0705	-1.1	-1.1	-1.1
$Channel_2$	.	-1.0963	-1.1	-1.1	-1.1
$Channel_3$	.	-0.7992	0.8	0.8	0.78
$Channel_4$	.	0.2743	-0.27	-0.28	-0.27
$Channel_5$	.	-1.2890	-1.3	-1.3	-1.3

Table 8. Fixed effects Stratified Cox Bayesian Model averaging

$Model$	$PosteriorProbability$	$nVar$
$Model_1$	0.752	11
$Model_2$	0.089	12
$Model_3$	0.06	12
$Model_4$	0.057	11
$Model_5$	0.042	11

Table 9. Stratified fixed effects Cox Bayesian Model averaging: the best 5 models

	PO	CC	RID
PO		A	B
CC			C
RID			

Table 10. relative gains between different methods of payment

References

1. Aalen O. Nonparametric Inference for a Family of counting processes ,Annals of Statistics, **4**, 701–726 (1978)
2. Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. Statistical Models based on counting processes, Springer Verlag, NY (1993)
3. Andersen, P. K. and Gill, R. D. Cox's regression model for counting process: a large sample study,Annals of Statistics, **10**, 1100-1120 (1982)
4. Berger, P.D. and N.I. Nasr Customer Lifetime Value: Marketing Models and Applications, Journal of Interactive Marketing, **12**, 17–30 (1998)
5. Brealey, R.A. and Myers S.C. Principles of Corporate Finance, McGraw Hill, NY (1996)
6. Brooks, S.P and Gelman A. General methods for monitoring convergence of iterative simulations, Journal of Computational and Graphical Statistics, **7**, 434– 456 (1998)
7. Clayton D.G. Bayesian analysis of frailty models, Technical report, Medical Research Council Biostatistics Unit, Cambridge (1994)
8. Cox, D.R. Regression Models and Life Tables, Journal of the Royal Statistical Society, Series B, **34** 187–220 (1972)
9. Dobson A. An introduction to generalized linear models, Chapman Hall (1990)
10. Figini S. Customer relationship: a survival analysis approach, (to appear in Proceedings of Compstat, Roma)(2006)
11. Fleming, T.R. and Harrington, D.P. Counting processes and survival analysis, Wiley, NY (1991)
12. Furnival G.M., Wilson R.W. Regressions by Leaps and Bounds, Technometrics, **16** (1974)
13. Gelman, A. and Rubin, D.R. A single series from the Gibbs sampler provides a false sense of security in Bayesian Statistics 4, Oxford University Press (1992)
14. Giudici,P. Applied data mining. Wiley (2003)
15. Grambsch, P.M. and Therneau T.M. Proportional Hazards Tests and Diagnostics Based on Weighted Residuals, Biometrika, **81**, 515–526 (1994)
16. Hoeting J.A., Madigan D., Raftery A.E., Volinsky C.T. Bayesian model averaging: a tutorial, Statistical Science,**14**, 382–417 (1999)
17. Hosmer, D.W. and Lemeshow, S. Applied Logistic Regression, Wiley NY (1989)
18. Hougaard P. Analysis of Multivariate Survival Data. Springer Verlag (1995)
19. Lawless J.F., Singhal K. Efficient Screening of Nonnormal Regression Models, Biometrics **34**, 318–327 (1978)
20. May S.L., Hosmer D.W.L. A Simplified Method of Calculating an Overall Goodness-of-Fit Test for the Cox Proportional Hazards Model, Lifetime Data Analysis, **4** (1998)

21. Kalbfleisch, J.D. and Prentice R.L. The Statistical Analysis of Failure Time Data, Wiley NY (1980)
22. Klein J.P., and Moeschberger M.L. Survival analysis: Techniques for censored and truncated data, Springer Verlag (1997)
23. Kuk A.Y.C. All subsets regression in a proportional hazards model, Biometrika **71**, 587–592 (1984)
24. Parzen M., Lipsitz S.R. A Global Goodness-of-Fit Statistic for Cox Regression Models, Biometrics **55**, 580–588 (1999)
25. Raftery, A.E, Madigan, D. and Volinsky, C.T. Accounting for model uncertainty in survival analysis improves predictive performance, Bayesian Statistics 5, Oxford University Press (1996)
26. Reichheld,F. The Loyalty Effect: The Hidden Force Behind Growth, Profits and Lasting Value, HBS Press, Boston (1996)
27. Reinartz, W.J. and Kumar V. On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing, Journal of Marketing, **64**, 17–35 (2000)
28. Schoenfeld, D. Partial Residuals for the Proportional Hazards Regression Model, Biometrika,**69**, 239–241 (1982)
29. Singer J.D., Willet J.B. Applied Longitudinal Data Analysis : Modeling Change and Event Occurrence, Oxford University Press (2003)
30. Spiegelhalter D.J., Dawid A.P., Lauritzen S.L. and Cowell G.: Bayesian Analysis in Expert Systems, Statistical Science, **8** 219–247 (1993)
31. Therneau T.M. and Grambsch P. Modeling Survival Data: Extending the Cox Model, Springer-Verlag (2000)
32. Volinsky C.T, Madigan D., Raftery A.E., Kronmal R.A Bayesian Model Averaging in Proportional Hazard Models: Assessing the Risk of a Stroke, Applied Statistics **46** 433–448 (1997)