TSE working paper series

10-174

Research Group: *Econometrics and Statistics*

March, 2010

Nonparametric Analysis of Hedge Funds Lifetimes

SERGE DAROLLES, JEAN-PIERRE FLORENS

AND GUILLAUME SIMON

Ecole d'économie de Toulouse (TSE) - Manufacture des Tabacs
Aile Jean-Jacques Laffont - 21, allée de Brienne - 31000 TOULOUSE
Tél : +33 (0) 5 61 12 85 89 - Fax : +33 (0) 5 61 12 86 37 - **www.TSE-fr.eu** - contact@TSE-fr.eu

Nonparametric Analysis of Hedge Funds Lifetimes*

Serge Darolles†, Jean-Pierre Florens‡, Guillaume Simon§

March 31, 2010

First Draft : Comments Welcome

Abstract

Most of hedge funds databases are now keeping history of dead funds in order to control biases in empirical analysis. It is then possible to use these data for the analysis of hedge funds lifetimes and survivorship. This paper proposes two nonparametric specifications of duration models. First, the single risk model is an alternative to parametric duration models used in the literature. Second, the competing risks model consider the two reasons why hedge funds stop reporting. We apply the two models to hedge funds data and compare our results to the literature. In particular, we show that a cohort effect must be considered. Moreover, the reason of the exit is a crucial information for the analysis of funds' survival as for a large part of disappearing funds, exit cannot be explained by low performance or low level of assets.

Keywords: Hedge funds ; Duration models ; Nonparametric specification ; Competing Risks.

*We are grateful to Christophe Boucher, Patrick Gagliardini, Olivier Scaillet, Angelo Pessaris, Vincent Pouderoux, and the participants and organizers of the conference on financial econometrics of Nanterre 2008, "Econometrics of Hedge Funds" of Paris 2009 and "Financial Econometrics Conference" of Toulouse 2009. The usual disclaimer nonetheless applies, and all errors remain ours.

†Lyxor and CREST

‡IDEI and GREMAQ, Université de Toulouse 1, 21 allée de Brienne, 31000 Toulouse, France.

§Université de Toulouse and Lyxor. Email: guillaume.simon@ensae.org.

1 Introduction

Hedge funds aim at generating specific return distributions, mainly characterized by high returns, low volatility, and low correlation with traditional financial indices. The specificity of hedge funds comes from the dynamical management of risk, allocation and exposures, but also from their low level of regulatory constraints. As the information provided is often missing, limited or subject to caution, the only way to gather information is to collect it from databases, in which funds report their performance to attract potential clients. This report relies on their willingness to provide figures, which is not compulsory and is related to the objectives of the manager. All the information was formerly provided on alive funds only. But the industry of hedge funds, contrary to mutual funds, exhibits a high level of attrition. The appearance of those dead funds bases follows a stream of academic papers pointing out the fact that unavailable data on dead funds led to numerous statistical biases. With those data, it is thus possible to analyze in a static fashion, the differences between living and dead funds by conditioning by the status (alive or dead) of the fund.

However, a dynamic approach is a more ambitious and flexible framework to model the probability to die (across time) conditional on some (potentially dynamic) fund characteristics. And in this framework, censorship is not always well taken into account in empirical studies. For example, Pojarliev and Levich (2008) presents a statistical study on funds categorized along a posterior, observable status at the date of study, depending on the fact that the funds are alive or dead. One cannot draw two distinct studies on funds depending on their current status since the information is carried by both alive and dead funds. Similarly, conditioning by the age of the fund may be not sufficient, as the whole set of variables (returns, assets, and age of the fund) are random processes that are mutually linked. From a mathematical perspective, alive funds are individuals for which the event (death) has not yet been observed (censored) but that still depend on the same framework of analysis. The dynamic approach is a growing field of research, but those publications often use parametric specification and focus on a *single risk* framework. However, failure is not the only reason why hedge funds stop reporting to databases. Starting and stopping to report depend on proper and historic characteristics of the fund. If the fund performs well, the amount of assets under management (henceforth AUM) may reach a critical size, above which the arbitrages may not be profitable. The manager can decide that the fund does not need new clients, and stops the publication of its performance. This may explain that in practice, some funds exhibit, just before their exit, a good performance and a high level of AUM. However, the exit from the database may also simply mean the end of the life of the fund (liquidation, default, etc.), or the willing of its manager to hide bad performances before liquidation.

The aim of this paper is to discuss the specification choices made in this literature on hedge funds lifetimes. First, we propose to adopt a nonparametric framework, rather than a parametric one, to avoid mis-specification biases and get robust estimation of hazard intensities of default. Second, we take into account the reasons for the exit. In this context, we take as a standard the use of covariates, including dynamic ones, as performance and AUM for instance. It is straightforward that the lifetime duration have to be cautiously defined. We must admit that we cannot take into account a self-selection bias since our data are only available thanks to voluntary publication. We do not observe funds that have never decided to appear in a database. Those funds may have been liquidated just after their creation, or may be for instance family funds that do not need to publish their performances. However, it is coherent to think that any fund appearing in a database has needed, at one moment, to collect new investors: in the opposite case, database publication would be useless. We are consequently focusing on funds needing to report performance and to collect investors.

In a Section 2, we present features of hedge funds lifetimes, including advantages and drawbacks of using hedge funds databases. We precisely define how lifetimes are calculated and present common biases studied by the literature in Section 3. Then, we present the estimation proce-

dure, particularly focusing on the nonparametric framework and the use of covariates, including competing risks model. The last section presents our results, where we underline that nonparametric estimation avoids the problems of mis-specification when studying hazard intensities. We present also a strong influence of the inception date of the fund as a covariate and accounts for the importance of identifying the cause of the exit of the fund. Section 5 concludes.

2 Description of the data and definition of lifetimes

Mutual and hedge funds differ by their reporting constraints. Indeed, it is not compulsory for hedge funds to publish their performance. Each hedge fund defines its own rule of reporting, and potentially publishes its performances in a database. This induces several classical biases, and hedge funds lifetimes and survivorship is also difficult to study precisely[1]. This section details the previous works made on the subject.

2.1 In the literature

All academic studies agree on some fundamental points. Hedge funds show a high degree of attrition within each year[2]. Fung and Hsieh (1997) study the evolution of this rate between 1990 and 1996 and find values around 19%, much higher than for mutual funds. Brown et al. (1999) find a value around 14% each year during the period 1987-1996, and Amin and Kat (2002) find values of the same order for period 1994-2001. However, these results depend on the database and on the period since it's clear that crises impact deeply death and birth processes of hedge funds. Moreover, attrition and instantaneous probability of failure are two different concepts as we will detail it further.

Numerous works have already been published concerning funds' lifetimes and survivorship. General considerations on the analysis of survival and its link with style and characteristics may be found in Liang (2000) and Barès et al. (2001). Further research estimate survival probabilities as a function of characteristics, past performance and risk statistics. Hendricks et al. (1997) assume that performance could improve the survival probability, and also propose the idea that survivorship bias may induce some patterns in hedge funds survival analysis. Brown et al. (1997) identifies a strong relation between bad performance and consecutive disappearance from the database. Gregoriou (2002) provides a technical study of hedge funds survival based on covariates taken among lagged performance, fund strategy, leverage, or characteristics (fees, redemption, etc.). This work has been extended by models using default intensity like Grecu et al. (2007), pointing out that for dying funds, performance is generally poor at the end of the fund's life. However, when controlling for covariates, this has to be done on the whole life of the fund as in Ang and Bollen (2008) who use a dynamic Cox model with dynamic covariates. All these works use a single risk approach. They focus on the exit of the fund from the database, not on the cause of the exit. In fact, there are two potential reasons to explain the exit from a database. But studying the exit from a database is useful only if we can separate between those reasons. This is a matter of identification, since if the reasons for the exit are fundamentally different, the blind mixing of the two risks leads to false interpretation, implying an over-estimated rate of attrition of the industry. Rouah (2005) uses a competing risks model to analyze hedge funds survivorship: hedge funds lifetimes are studied with time-dependent covariates, along with the cause of exit under the assumption of independent risks. He underlines that avoiding to separate liquidation from other kind of withdrawal leads to severe biases (especially over-estimation of the survivorship bias). Amin and Kat (2002) explains that it may be due, for instance, to a lack of size or performance. In some databases, the main reasons explaining the death of a fund are

[1] We give in Appendix A.1 a review of some usual database biases that have been studied in the literature, and in Appendix A.6 the overview of the academic contributions on hedge funds durations.

[2] This degree of attrition can be easily estimated by counting the proportion of funds that are alive or dead at the end of each year. This rate is particularly high for 2008.

sometimes detailed. It may have been liquidated, merged, its assets transferred in an other fund; the minimum capacity of the fund has been reached after several redemptions during a market turmoil, or the main investor may have left the fund; the company may have been closed, the manager may have left, or the team-management has decided to concentrate on other strategies.

Appropriate covariates have to be found in order to analyze hedge fund lifetimes. For this, AUM is generally assumed to be a good indicator of funds' status. Assets and performance are generally supposed to decrease the funds' default intensity. But the other potential reason to explain the withdrawal of a fund from a database is that the fund has reached a maximum capacity and then that it is closed to investment. In some successful hedge funds, the manager may think that the fund has reached a sufficient size in order to keep its ability to generate performance. This is linked with characteristics that are relative to the specificity of each fund and the decision to stop to collect money is related to the fund's objectives and possibilities. Amin and Kat (2002) explicitly don't afford much attention on this point, as Grecu et al. (2007) who explain that the first reason (dying funds) is the most frequent and that the lack of data concerning successful funds stopping to collect is not a problem in practice. Conversely, Ackermann et al. (1999), Gregoriou and Rouah (2002) or Rouah (2005) stress that this kind of exit must be considered: if not the attrition rates are too high when this separation is not made. The survivorship bias is also affected.

2.2 TASS database

A large number of hedge funds databases are available, with their own specificities and characteristics. We use in this paper the TASS database[3] for at least two reasons. First information concerns both alive and dead funds. Second, a large number of fields is available, which allows to control for biases and to improve the study. We first describe the TASS database of 2009, June the 25th. The database is made of two group of funds, depending whether they are Managed Futures/CTA funds, or not. For each category, there is a file concerning alive funds, and a "graveyard" containing dead funds. For a given file, it consists in a list of fund shares, each share being identified by an unique number. Each share corresponds to a given currency. In addition to CTA and Managed Futures, the represented strategies are among the main hedge fund strategies such as single strategies (Long/Short Equity, Event Driven, Global Macro, Fixed Income Arbitrage, Multi-Strategy, Convertible Arbitrage, Dedicated Short Bias, Emerging Markets, Equity Market Neutral, Emerging Markets, Options Strategy) and fund of funds. Information on the fund status is published concerning status, dates, liquidity and performance. For example, the domicile country or state, public opening, leverage, management and incentive fees, presence of a highwater mark, and dead reason (when the fund is dead) are displayed. Moreover, dates when the fund has been added and/or removed from the database; dates of inception, start and end of performance are indicated. Concerning liquidity and investment constraints, subscription/redemption frequency, redemption notice period, lock-up and lock-up period may also be indicated. Finally, turning to historical performances, monthly returns (published or estimated by TASS), NAV, and AUM are sometimes given (in domicile currency).

An open discussion concerns the field called *drop reason* about which academic contributors are still puzzled. As Baquero et al. (2002), or Rouah (2005), Boyson (2002) focuses first on explicitly liquidated funds only. Getmansky et al. (2004) also insist on the importance of this field. Kundro and Feffer (2003) go further as they distinguish fund's failure (external reasons forces the fund manager to stop) from fund liquidation (the collect of fees is not sufficient, the manager ceases activity before re-launching a new fund to reset the highwater mark to a new level). This suggests to use this drop reason with caution and with additional filters on the data.

[3]As Boyson (2002) (who use a database from 2002), Amin and Kat (2002), Getmansky et al. (2004), Grecu et al. (2007) (database of 2004) or Liang and Park (2008) (database of 2004) for instance.

2.3 Hedge funds lifetimes

2.3.1 Lifetimes definition

In summary, the life of a fund in a database may be decomposed in the following way. First the fund is created (inception date). After that, the fund may possibly enter the database (date added) and reports performance. Consequently, the first date of report may be anterior to the date of entry in the database, but not to the inception date. When the fund disappears from the alive database, it enters the graveyard with a dead reason, and the last performance date is equivalent to the death date of the fund. The death of the fund may be explained by the fact that the fund is closed to new investment, dormant, liquidated, no longer reporting, liquidated, merged into an other fund, that the program is closed or that it is not possible to contact the fund. If the fund stops reporting, two potential causes may explain this event. First, the fund is poorly performing and is liquidated, merged, or closed down, and ceases activity. But if the fund performs well, its manager may decide to stop from collecting new investors, and does not report performance any more. If the fund seeks new investors, it continues to report and is neither closed or dead: its performance is still observable in the database. This imply to define in a clear fashion the entry and the exit from the database.

If each information in a database was perfect, any study on hedge funds lifetimes would be easy. The first reporting date would be equal both to the entry date in the database, and to the inception date. For funds still alive, report would be updated until the current date. Each field would be cautiously reported. Between entry and exit of the database, both performance and AUM would be given at each date. Consequently, lifetime durations of hedge funds would be easy to analyze with all covariates available in the database (fund characteristics, performance, AUM). However, databases show some practical limits that constrain our study.

2.3.2 Initial date

It is well known that hedge funds databases suffer from numerous drawbacks, the first one being the backfilling bias. A fund can be added to a database months after its inception and backfills its history with past performance[4]. When a fund decides to backfill its track return, the manager has the choice of the first date, with the possibility to skip first returns if those are not good. In some cases, funds report performances anterior to their inception date : these are synthetic, not real performances. Funds with first report before inception are clearly backfillers and must be dropped from the sample. This assesses the choice of the inception as the initial date, as it is commonly made in the literature.

Assumption A.1 The initial date corresponds to the inception date, and only funds that display explicitly an inception date are considered.

2.3.3 Date of exit

A second drawback, named *reporting bias* is in the potential delay in the report of performances. For instance, funds classified as alive in June 2009 (and thus with data at the end of May 2009) may not have published their performances at this date[5]. The fund may still be in activity with figures that are not yet available (because of liquidity, mark-to-market reasons, or other causes). We then consider data at a former date than the last date available of our database and use this to potentially upgrade the graveyard. The date of study remains May 2009 (last data). But funds in the alive database are dropped to the graveyard if their status explicitly indicates

[4]This occurs very frequently since funds can be older than the database. This may be an important effect since young and bad-performing funds that have not been able to survive do not appear in the database.

[5]See Derman (2007) which adapts a Markov chain model to study properties of funds classified as good, sick or dead as a complement of this approach.

it[6] (liquidated, no longer reporting, or unable to contact it) or if it's not the case, that the last report is anterior to four months (no data at the end of January).

Assumption A.2 A fund is considered as dead if it is in the graveyard, including funds with too much delay or explicit dead reason. Final date for dead funds is the date of last performance report, and the fund is uncensored. Final date for alive funds is the date of study, and the fund is considered as censored.

As a consequence to assumptions A.1 and A.2, we calculate the lifetime of the fund as the time between the initial and the final date. We must add that we have to assume throughout this paper that censorship and durations are independent. If it was not the case we could not identify the law of the observed durations. This seems to be a plausible hypothesis in our framework since censor is made by the current observation date. Note that taking censorship into account is crucial in survival analysis and must always be done (contrarily to the approach of Grecu et al. (2007)).

One may also consider a *two-periods* framework. In this case (Figure 2.1) one accounts for the potential relation between two durations : the first is equal to the time elapsed between a fixed date and the fund inception date, and the second is equal to the lifetime duration of the fund.

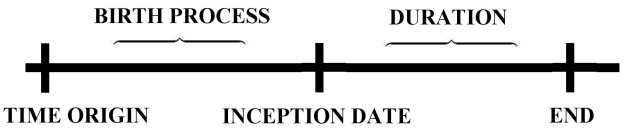

Figure 2.1: A two-period model

3 Model specification and estimation

This section discusses the specification and the estimation of the single and the competing risks models.

3.1 Model specification

Single risk models are first considered. The competing risks approach can then be seen as a multivariate extension of this framework.

3.1.1 Single risk model

A duration is defined as the length τ of a time-period spent by a fund in a given state[7] (in the database in our case). We suppose that this duration τ has a density function denoted $f(t)$, the associated cumulative distribution function F:

$$F(t) = I\!P(\tau \leq t), \quad \text{for} \quad t \geq 0,$$

and the survivor function S corresponds to $S(t) = I\!P(\tau \geq t) = 1 - F(t)$ for $t \geq 0$. The integrated (alternatively cumulative) hazard function of τ is denoted Λ and satisfies :

[6] As the TASS database is updated at given dates, it may appear that between two of them, some dead funds have not joined yet the graveyard.

[7] The assumption that $I\!P(\tau = \infty) = 0$ can be made (a fund cannot have a strictly infinite lifetime, which does not mean that this lifetime is bounded).

$$\Lambda(t) = \int_0^t \frac{f(u)}{S(u)} du = -\int_0^t \frac{dS(u)}{S(u)} = -ln(S(t)). \quad (1)$$

From Equation 1 we can easily define the hazard function λ of the duration τ through:

$$\lambda(t) = \frac{d\Lambda(t)}{dt} = \frac{f(t)}{S(t)} = -\frac{d\ln(S(t))}{dt}. \quad (2)$$

The function $\lambda(t)$ may be interpreted as an intensity, for the latent process of leaving the state, conditional on its current state at t. This function λ is not necessarily monotonous in t, and corresponds in our case to the instantaneous probability for the fund still alive to exit the database just after t :

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{I\!P[t \leq \tau < t + \Delta t | \tau \geq t]}{\Delta t}. \quad (3)$$

A duration corresponds to the time of jump of an underlying process characterized by $\lambda(t)$. The following expressions for integrated hazard, density and survival function as a function of $\lambda(t)$ are then straightforward.

Our aim is therefore to estimate the hazard function $\lambda(t)$ for the survival of hedge funds, as its interpretation is intuitive, its shape carries valuable information. Mainly, since hazard rate in t is homogenous to the intensity for each fund to exit the database at date t, the higher the hazard rate, the more likely the fund is to stop reporting performance. In general, this rate is not constant and young funds have a greater instantaneous probability to exit the database (see e.g. Brown et al. (2001), Amin and Kat (2002)). But the precise form of this shape is still subject to questions and the notion of "young funds" is not precise. The parametric framework constrains the analysis by imposing predetermined form, and misses some information at specific moments of the life of the fund.

Monotonic decreasing forms assume that fund managers benefit from their experience and that aged funds are less likely to die. Inverted U-shape assumes that funds without experience take more risk and die; old funds cease to collect assets, explaining this U-shape. But inverted U-shape patterns is the most common choice (Gregoriou (2002), Grecu et al. (2007),). The intensity increases, reach a maximum value and then decreases for aged funds.

An improvement of the previous approach is to specify this intensity as a (deterministic) function of time and also of (potentially dynamic) covariates. Cox proportional hazard models are particular cases of accelerated life models. Dependence towards covariates x is introduced via a function ψ such as the density of the duration is supposed to be $f(t\psi(x))\psi(x)$. In this framework, $\lambda(t)$ for each individual is assumed proportional to the product of a *baseline* hazard function $\lambda_0(t)$ and a function $\psi(x, \beta)$ of covariates x : $\lambda(t) = \lambda_0(t)\psi(x, \beta)$, but the model remains semi-parametric as soon as no assumption is made on $\lambda_0(t)$, and β is a parameter to be estimated. The ψ function is usually $\psi(x, \beta) = exp(\beta' x)$, which allows[8] to estimate separately the effect of the covariate, and the baseline intensity.

3.1.2 Competing risks model

When several failure types exist, the related framework is referred to as *competing risks* models. For each individual, the risks are of different natures, and several mechanisms are in *competition* to explain the causes of failures. Suppose that failure may be caused by m distinct types of risks : for each individual this implies m failure times, which are called *latent* or *potential* times. For each risk i, the potential time of failure will be T_i, and the observed time of failure is $T = min(T_1, ..., T_m)$. But the specification in those models have to be cautiously discussed.Three different problems exist (Kalbfleisch and Prentice (2002)): first, one may study

[8]Other forms are possible such as $\psi(x, \beta) = 1 + \beta' x$, or $\psi(x, \beta) = log(1 + e^{(\beta' x)})$.

the specific behavior and the probability of occurrence of each failure type; second, the objective may be to examine, under a predefined set of hypotheses, the interdependence between those failure types; finally, an interesting point may be the effect of the removal of one risk on the occurrence of the others.

Considering hedge funds lifetime durations, the motivation for using those models is straightforward. A fund may exit a database for two specific reasons. First, the fund is exposed to the risk of liquidation due to bad performances : this failure type will be labeled of type T_-. On the contrary, if the fund is a good performer, the manager may decide to stop reporting because he does not need any new investor : this exit is of type $T+$ since it implies a withdrawal from the database. If the risks respectively lead to latent times of exit from the database T_- and T_+, the observed time of failure is $T = min(T_-, T_+)$.

Basically, identification is only possible if in addition of the time of failure, the *cause* of the failure (namely the nature of the risk) is observed. Then, the set of observations has to be of the form (T, j) where T is the failure time and $j \in [1; m]$ is the cause of the failure. However, even corresponding time-independent covariates are insufficient to fully explain the interrelation between risks, and to ensure identification (Tsiatis (1975)).

Given time-dependent covariates $x(t)$, an identifiable quantity is the *type-specific hazard* which is defined as :

$$\lambda_j(t, x(t)) = lim_{\Delta t \to 0} \frac{I\!P[t \leq T \leq t + \Delta t, J = j | T \geq t]}{\Delta t}$$

where T is the observed failure time and j the cause of the failure. In our study $m = 2$, and two causes cannot occur simultaneously. Some other functions may be introduced (*overall failure rate*, *overall survivor*, *cumulative incidence function*, etc.) but even a function of the form : $S_j(t, x) = exp(-\int_0^t \lambda_j(u, x(u))du))$ cannot be interpreted as a survivor function, when no additional assumption on the interdependence of the risks is introduced.

We assume that the two risks of type 1 and 2 are independent and specify a semi-parametric form for the j-th intensity:

$$\lambda_j(t, x(t)) = \lambda_0^{(j)}(t) exp(\beta_j' x(t))$$

where $x(t)$ is a set of covariates left-continuous, with right limits, and $j \in [1; m]$. $\lambda_0^{(j)}(t)$ is left unspecified and will be nonparametrically estimated.

3.2 Estimation

We first provide in Appendix A.2 parametric and nonparametric estimation procedure in the single risk framework without covariates. We first focus here on the estimation of a dynamic Cox model using covariates, with a nonparametric baseline intensity. Then, we discuss the estimation of competing risks models.

3.2.1 Single risk models

First step : estimation of β

The parameter β can be estimated with a completely unknown baseline intensity $\lambda_0(t)$. First, let's suppose that we have a set of ordered failure times $\{\tau_1, \ldots, \tau_n\}$ without censorship. Conditional on the history $\{\tau_1, \ldots, \tau_j\}$ with corresponding index $\{i_1, \ldots, i_j\}$, the probability for individual i to fail at τ_j is equal to the probability to fail at τ_j given that there is a failure in the set $\{k | \tau_k \geq \tau_j\}$ which is:

$$\frac{\psi(x_i, \beta)}{\sum_{\{k|\tau_k \geq \tau_j\}} \psi(x_k, \beta)}$$

since the proportional form for the hazard function makes the baseline term vanish. This probability appears to be independent from $\{\tau_1, \ldots, \tau_j\}$ and the probability for the set of observations to be ordered as $\{i_1, \ldots, i_n\}$ is :

$$\prod_{j=1}^{n} \frac{\psi(x_{i_j}, \beta)}{\sum_{\{k|\tau_k \geq \tau_j\}} \psi(x_{i_k}, \beta)}.$$

With censorship, the censored individuals appear only in the survivor function and the partial likelihood becomes:

$$\prod_{j=1}^{n} \left(\frac{\psi(x_{i_j}, \beta)}{\sum_{\{k|t_k \geq \tau_j\}} \psi(x_{i_k}, \beta)} \right)^{1-\delta_j}$$

with δ_j equal to 1 if the data is censored, 0 otherwise. With the specification $\psi(x, \beta) = \exp(x'\beta)$ the estimator[9] of β is:

$$\hat{\beta} = argmax_\beta \sum_{i \in I_u} \{x_i'\beta - log[\sum_{\tau_j \geq \tau_i} exp(x_j'\beta)]\}.$$

An extension is even possible when the time-scale is discrete but the computation of the likelihood is rather tedious.

When covariates are dynamic (i.e. $x = x(t)$), the likelihood is easy to derive but those developments can only be made under strict assumptions on the nature of this time-dependence. If $\mathcal{G}_t$ is the filtration that resumes random information up to time t (failures, censoring, etc.), the chosen covariates must depend on $\mathcal{G}_t$. Moreover, the hazard function (conditioned by the risk set and $\mathcal{G}_t$) must depend at t on x only by its current value $x(t)$ and not on its entire realization. Then, β is obtained by maximizing the partial likelihood, given by:

$$\prod_{j=1}^{n} \left(\frac{\psi(x_j(\tau_j), \beta)}{\sum_{\{k|\tau_k \geq \tau_j\}} \psi(x_k(\tau_j), \beta)} \right)^{1-\delta_j}$$

where δ_j is the same indicator of censorship as before. We use for the numerator a quantity relative to effective default, and for the denominator a quantity based on the individuals still at risk. With time-dependent covariates, the variables of the individuals still at risk at date τ_j are taken at this date.

Second step : nonparametric estimation of the baseline intensity

Estimation of the baseline intensity comes in a second step. The integrated hazard function $\Lambda(t)$ is first estimated through:

$$\hat{\Lambda}(t) = \sum_{j|\tau_j < t} \frac{n_j}{\sum_{k|\tau_k > t} \psi(x_k, \hat{\beta})}$$

where n_j is the number of defaults at date τ_j (uncensored individuals), and $\{k|\tau_k > t\}$ is the whole set of individuals still at risk at t. Having estimated $\hat{\Lambda}(t)$, a nonparametric estimation

[9]The variance of the estimator is given (see Cox and Oakes (1984)) by:

$$\mathbb{V}[\hat{\beta}] = -V^{-1} \qquad \text{with} \qquad V_{i,j} = \frac{\partial^2 log(L(\beta))}{\partial\beta_i \partial\beta_j} = -\sum_{k=1}^{n} \Big[\frac{(A_k)(B_k^{(i,j)})}{(A_k)^2} - \frac{(B_k^{(i)})(B_k^{(j)})}{(A_k)^2} \Big],$$

with, if we note $Rs(k) = \{k'|\tau_{k'} \geq \tau_k\}$:

$$A_k = \sum_{u \in Rs(k)} exp(\beta' x_u) \qquad B_k^{(i,j)} = \sum_{u \in Rs(k)} x_{u,i} x_{u,j} exp(\beta' x_u)$$
$$B_k^{(i)} = \sum_{u \in Rs(k)} x_{u,i} exp(\beta' x_u) \qquad B_k^{(j)} = \sum_{u \in Rs(k)} x_{u,j} exp(\beta' x_u).$$

of the baseline intensity is obtained as the derivative of a smoothed version of $\hat{\Lambda}(t)$. Using a gaussian kernel, $\hat{\lambda}(t)$ is estimated with :

$$\hat{\lambda}(t) = \sum_{k=1}^{n} \frac{1}{\sqrt{2\pi}h_n} exp\left(-\frac{(t-\tau_i)^2}{2h_n^2}\right) \times (\Delta\hat{\Lambda}(\tau_i)).$$

where h_n an associated bandwidth. A correction is possible using boundary kernels. When $\tau_i \leq h_n$, Li and Racine (2006) proposes the following correction for $\hat{\lambda}(t)$:

$$\hat{\lambda}(t) = \sum_{k=1}^{n} \frac{1}{\sqrt{2\pi}h_n} exp\left(-\frac{(t-\tau_i)^2}{2h_n^2}\right) \times (\Delta\hat{\Lambda}(\tau_i)) \times \left(\mathbf{1}_{\{\tau_i \geq h_n\}} + \frac{\mathbf{1}_{\{\tau_i < h_n\}}}{1+\Phi(-\frac{\tau_i}{h_n})}\right)$$

with Φ the cumulative function of the standard gaussian distribution (see also section A.2.2).

3.2.2 Competing risks model

The first step concerns the estimation of the β coefficients. The set of observations is of the form : (τ_i, j_i, δ_i) where for all individual $i \in [1; N]$, the observed time of failure is τ_i, with cause j_i, and still $\delta_i = 1$ if the observation is censored (0 otherwise). We assume that censoring and risks are independent (which is a strong and crucial, yet common assumption). Then the partial likelihood to be maximized has the following expression :

$$L(\beta_1, \ldots, \beta_m) = \prod_{j=1}^{m} \prod_{i=1}^{k_j} \left(\frac{exp(x_{j,i}(\beta_j' \tau_{j,i}))}{\sum_{\{l|\tau_l \geq \tau_{j,i}\}} exp(\beta_j' x_l(\tau_{j,i}))}\right)^{1-\delta_i}$$

where for each $j \in [1; m]$ the failures are : $\tau_{j,1} < \ldots < \tau_{j,k_j}$, x is the related process of covariate for each individual and risk, and $\{l|\tau_l \geq t\}$ is classically the risk set at t. Under those hypotheses, the estimation procedure is very similar to the single risk framework. The cumulative hazard for risk $i \in 1, 2$ is estimated through:

$$\hat{\Lambda}^{(i)}(t) = \sum_{j|\tau_j < t} \frac{n_j^{(i)}}{\sum_{k|\tau_k > t} exp(x_k, \hat{\beta})}$$

where $n_j^{(i)}$ is the number of defaults of the kind i at date τ_j (uncensored individuals of risk i), and $\{k|\tau_k > t\}$ is the whole set of individuals still at risk at t. Again $\hat{\lambda}(t)$ is estimated through :

$$\hat{\lambda}(t) = \sum_{k=1}^{n} \frac{1}{\sqrt{2\pi}h_n} exp\left(-\frac{(t-\tau_i)^2}{2h_n^2}\right) \times (\Delta\hat{\Lambda}(\tau_i)).$$

where h_n an associated bandwidth (and potential correction with the help of boundary kernels). The assumption that the latent risks are independent is very strong, and may ensure identifiability, but is untestable. Conditions for identification[10] are given in Heckman and Honoré (1989), Abbring and van den Berg (2003). Some extensions exist in the literature. Dewanjy and Sengupta (2007) develop inference in competing risks model where the failure type is not always observed and is part of a set of possible types.

4 Empirical study

We present in this section the results of our estimation procedure. In addition to the results on the whole database, we present results also by separating funds along with their declared strategies.

[10] An other way of ensuring identifiability is to study sets of data with multiple-spell data (several observations for each individual), which cannot be obtained in the case of hedge funds analysis.

4.1 Descriptive statistics

We merge the CTA file in an overall database. Then we obtain a database made of 6121 funds and a graveyard of 8102 funds. The description of the database by strategy is available in Table A.1. The histogram of raw durations is given in Figure A.1. We also present in Table A.2, separating along some main strategies, some descriptive statistics :mean duration, empirical mode and empirical quantiles. The category Long-Short Equity Hedge represents the main part of single funds (1521 funds out of 3963). We aggregate CTA and Managed Futures categories, which form one of the main categories after Multi-Strategy, Event Driven, Global Macro, Equity Market Neutral, Fixed Income Arbitrage, and Emerging Markets.

The mean duration for single funds and fund of funds is quite close, around 60 months (5 years). It's straightforward that single funds durations behavior is mainly driven by Long-Short Equity Hedge. This quantity is however coherent as for each strategy, the mean duration is comprised between 50 and 70 months. The strategies with the shortest mean durations are Global Macro, Equity Market Neutral, and Multi-Strategy, this being also verified for the upper 95% empirical quantile. Conversely, Event Driven and CTA-Managed Futures are strategies with the longest mean duration. It may be observed that in the case of CTA, there are much more dead funds, then less censored individuals, which may explain this fact. The more recent is a strategy, the younger the funds and the shortest the mean duration.

4.2 Single risk analysis

4.2.1 Nonparametric intensities

In this section, we compare parametric and nonparametric estimations of the hazard intensity for hedge funds' lifetimes. First, our aim is to check whether the specifications commonly used in the literature are justified, second, if the nonparametric estimation can improve the understanding of a hedge fund lifetime and third, to adapt the conclusions depending on the several strategies. Concerning the parametric fitting, we use a log-logistic[11] law. This distribution is commonly chosen (see e.g. Gregoriou (2002) or Grecu et al. (2007)). Results[12] are given in Table A.3.

The obtained modes for the hazard rates (that is the time at which the highest intensity of default is reached), are around 40 months, which is coherent with the values obtained in the literature (see Grecu et al. (2007)), yet a bit inferior. They are also coherent with the order of magnitude of the empirical modes[13].

The true interest of such a model is to compare these results with nonparametric estimation of the hazard intensity. For this, we must first build the confidence bounds of the parametric curve. We denote by $\theta = (\mu, \sigma)$ the parameters of the log-logistic distribution and by $\hat{\theta}$ their

[11] The density of a log-logistic distribution of parameters (μ, σ) is given by:

$$f_{\mu,\sigma}(x) = \frac{e^{\frac{ln(x)-\mu}{\sigma}}}{\sigma x(1+e^{\frac{ln(x)-\mu}{\sigma}})^2}$$

where μ is the location and σ the scale parameter. A convenient expression is obtained when $\rho = e^{-\mu}$, $\kappa = 1/\sigma$:

$$f(t) = \frac{\kappa\rho^\kappa t^{\kappa-1}}{(1+(\rho t)^\kappa)^2} \qquad S(t) = \frac{1}{1+(\rho t)^\kappa} \qquad \lambda(t) = \frac{\kappa\rho^\kappa t^{\kappa-1}}{(1+(\rho t)^\kappa)}$$

.

[12] The same study has been done for discrete durations. The results are not presented here, but they are very similar and do not modify deeply the conclusions.

[13] However, as the histogram of durations is often sparse or not well designed, the empirical mode is not very informative.

maximum likelihood estimates. If the hazard function is given by function $t \mapsto \lambda(t, \theta)$ the confidence interval for $\lambda(t)$ at level α is given by:

$$\lambda(t, \hat{\theta}) \pm \frac{q_{1-\alpha/2}}{\sqrt{N}} \sqrt{\frac{\partial \lambda}{\partial \theta} \times V \times \frac{\partial \lambda}{\partial \theta'}}$$

where N is the number of observations, $q_{1-\alpha/2}$ is the quantile of a standard gaussian, V is the estimated variance of the parameters since $\hat{\theta}$ is asymptotically gaussian. This expression is obtained with the application of a usual δ-method. The results are presented in Figures A.2 to A.12. For each strategy, it appears that the choice of the log-logistic function is coherent with the general shape of the nonparametric intensity obtained using smoothed Kaplan-Meier estimator. These estimators evokes an inverted U-shape pattern. The intensity is increasing in the first years, reach a peak and then decreases. This is at least valid for the earlier years of existence of the fund. It is not coherent to draw this intensity for very high durations (over 10 or 12 years) as there are then not enough funds in the sample for the estimation to be reliable. For each strategy, we add the evolution of the proportion of funds still in the sample across time, along with a threshold (set to 10%) to ensure that the intensities are still representative.

When we look at a finer scale to those functions, we see that several differences appear. First, the log-logistic specification is most appropriate for fund of funds than for Single Funds. For single strategies, the log-logistic specification is particularly adapted to Event Driven and Fixed Income Arbitrage funds. For Event-Driven funds, the nonparametric law has nearly exactly the same shape and the same mode than the parametric one. Long-Short Equity Hedge, Equity Market Neutral, and CTA intensities have roughly a nonparametric intensity that is of the same shape than in the parametric case, even if the intensity is not in the confidence interval during some months, before becoming close to the parametric specification after same time. For Multi-Strategy funds, the parametric specification is plausible but the nonparametric intensity decrease is stronger after 60 months than the parametric specification suggests it. The situation is more critic for Emerging Markets and Global Macro funds where the intensity is quite monotonous, increases and crosses the parametric intensity and exhibits an irregular pattern. Consequently for those two categories, the mode of the default intensity suggested by the nonparametric estimation is higher than the one obtained with the parametric estimation. Excepted Multi-Strategy and Event Driven funds, this conclusion is also valid for most strategies. Generally speaking, we cannot confirm the finding of Gregoriou (2002) that default intensity decreases after having reached a peak : except for Multi-Strategy, the behavior of the intensities after having reached a local mode is quite steady or non-monotonic.

In conclusion, the parametric specification of the log-logistic distribution is coherent with the general shape suggested by the nonparametric estimation. If the choice may be justified for several strategies, the parametric distribution systematically under-estimates the modes (up to two years) and the long-term hazard intensities. On the contrary, the real initial hazard intensity is strictly different from zero, a fact which is missed by the parametric analysis. Then, the nonparametric analysis capture specific features of the intensities and avoid specification problems (short-term under-estimation, and long-term over-estimation).

4.2.2 Including covariates

The former analysis is descriptive and useful, yet limited, the introduction of covariates is standard in the duration literature. The intensity writes : $\lambda(t) = \lambda_0(t)\phi(\underline{x}_t)$ with $\underline{x}_t$ the past and current values of appropriately chosen covariates x. When the remaining $\lambda_0(t)$ is constant equal to λ, all information is captured by the model. A "noisy" structure remains and $\lambda_0(t) = \lambda$ is the intensity of a homogenous Poisson process. it is possible to show that Poisson processes are processes with constant intensity and without memory : the future realizations of the process are independent from the former and the distribution of the number of events in a given time-interval

follows an exponential distribution whose parameter depends on the length of the time-intervals.

Performance and AUM appear as two natural covariates for the study, but other variables have been proposed. Amin and Kat (2002) examines the fund leverage, even if Barry (2002) minimizes the impact of this variable; Grecu et al. (2007) proposes to incorporate the Sharpe Ratio and the volatility of the fund return; Baquero et al. (2002) is concerned with the investment style and Gregoriou (2002) tests the influence of management fees, performance fees, minimum purchase, mean returns, or redemption periods. This is also in line with Avellaneda and Besson (2005) that develops the concept of skill-capacity i.e. the maximum amount of money a manager can expect without decreasing its arbitrages and consequently its performance. As in Couderc et al. (2008), one may also think to include business or economic variables or indexes. In our study, we mainly consider performance and AUM. But an additional variable is considered : the distance between the date of inception and a static date[14] to deal with potential cohort effects. An mistake may be found in parametric studies related to parametric or semi-parametric estimation of default intensity with covariates. When assuming an intensity $\lambda(t) = \lambda_0(t)\psi(Z)$ where Z are (potentially time-dependent) covariates, it is pointless to include time as an explanatory variable in Z, as all the time dependence will be captured by $\lambda_0(t)$.

As a fund cannot enter our analysis as soon as at a given date, some data remain undisclosed, we recall that we use the method described in Appendix A.7 to avoid to drop too much funds in our study. Moreover, we exclude backfillers, i.e. funds that present data before their inception date. The chosen form of the hazard intensity is:

$$\lambda(t) = \lambda_0(t)exp(\beta' x(t))$$

where $\lambda_0(t)$ is left unspecified. As the missing data problem shrinks the number of available funds for the study, we will mainly focus on Single and Fund of Funds. But more important is the fact that only funds in the same currencies (returns, performance, assets) can be part of the study. We decide here to consider all the funds in US Dollar. This is coherent as the funds in the database are shares, which are often shares from the same fund in different currencies. Selecting USD funds is a way of selecting some representative shares of funds (Liang and Park (2008) also focus on USD funds only). The number of funds available for the study is given in Table A.4. The coefficients of several "nested" models are given in Table A.5 for Single Funds, and Table A.6 for Funds of Funds.

Among the chosen variables we consider the monthly returns (*Returns*), the level of equity[15] (*Equity*), the logarithm of the level of AUM[16] (*Ln(AUM)*), the monthly log-return of AUM (Δ *Ln(AUM)*) and the difference (in days) between the inception date of the fund and[17] a static date, taken here as the 1st of January 1990 (*Date Ref*). Results are provided in Table A.5 for Single funds, and in Table A.6 for Funds of Funds.

For single funds and fund of funds, we see that for models with one or two variables, without *Date Ref*, all the coefficients are significant. They are all negative, suggesting that a higher level of the covariate decreases the intensity of default. This is intuitive as high levels of performance, return and AUM may be indicators of good health of the fund. However, when we include the *Date Ref* variable, this variable captures most of the dynamic of the intensity. Both for single and fund of funds, return and log-return of AUM are no more significant. For single funds, equity and AUM are plausible explicative variables, whereas for fund of funds, only the AUM

[14] We take here arbitrarily this date as the 1st January 1990 but any static reference date could be used. However, a reference date too far from the sample of the inception dates could lead to potential scale effects that shrink the mean level of the baseline intensity.

[15] If R_t denotes the monthly return of a fund, we define the level E_T of equity at time T as $E_T = \Pi_{t=1}^{T}(1 + R_t)$.

[16] In millions of USD.

[17] We specifically here focus on a limited set of variables. Nonlinear functions of the return or financial indicators as in Couderc et al. (2008) have also been included but appeared to be not really informative.

variable appears to affect the intensity. In conclusion, the only variables of concern are AUM, Equity and the "reference date". This last variable accounts for a cohort effect, suggesting that younger funds are more at risk than old ones, other variables being equal.

A parametric specification in this context of the baseline intensities (non-monotonic and not always with an inverted U-shape) would be difficult since it would be too constraining and lead us to miss the specific feature of the baseline (especially its flatness). The obtained nonparametric intensities are given in Figure A.15 and A.15.

4.3 Competing risks analysis

In the single risk framework, no distinction is made on the nature of the exit. When a fund is still reporting, one may be interested in the potential issues he is going to face in a near future. The probability of exit from the database is interesting in itself, but what is more pertinent, is to forecast or to assess the reason for it. Bad performers are liquidated, whereas good performers may be still in activity after their exit from the database.

4.3.1 Competing risks model

A simple simulation exercise can be made to illustrate the importance of using competing risks model. Draw a sequence of N couples of parameters $(\mu_i, \sigma_i)_{i\in[0;N]} \in [-r; +r] \times [s_m; s_M]$ with $r, s_m, s_M > 0$. Then, make for each i a random simulation path of independent monthly gaussian returns R_t with mean μ_i and variance σ_i. For each i, simulate two independent durations T_i^+ and T_i^-, with the respective intensity processes: $\lambda_+(t) = \lambda_0(t)exp(\beta \times R_t)$ and $\lambda_-(t) = \lambda_0(t)exp(-\beta \times R_t)$. The easiest way to run the simulation is to draw for each individual i two independent realization E_i^+ and E_i^- of $Exp(1)$ variables. Then T^+ and T^- are obtained as :

$$\int_0^{T_i^+} \lambda_0(t)exp(\beta R_t^i)dt = E_i^+ \quad \text{and} \quad \int_0^{T_i^-} \lambda_0(t)exp(-\beta R_t^i)dt = E_i^-.$$

The simulations are made with the same positive value for β but with a positive sign for exit T^+ and negative sign for T^-. λ_0 is chosen identical for both risk (log-logistic for instance).
If we "drop" in our analysis the true generating process of the durations, and that we only consider the times $T_i = min(T_i^+, T_i^-)$ (regardless of the cause, T_+ or T_-), with the return as a dynamic covariate and the exact model $\lambda(t) = \lambda_0(t)exp(\gamma R_t)$, we find a value of γ which is close to zero, very different from β. As two populations with opposite risks are aggregated in similar proportions, both effects compensate the other in the estimation. Then, letting aside the cause of the failure leads to estimated coefficients that are severely biased.

4.3.2 Estimation

To estimate a competing risks model, we have to identify the cause of the failure of the fund. This one is not always observed and we have to set a procedure to decide whether the fund exits the database because of bad or good performance. The decision is made along two criterions : the death reason of the fund in the database and/or its level of AUM at the time of exit.

First, all the funds that classified as "closed to new investment" in May 2009 are considered as exiting the database because of good performance (exit of type "T_+"). For the other uncensored funds, we consider that the funds have an exit of type "T_-", when their AUM at the time of exit is below 200 millions of US dollars or that this value is less than 75% of the maximum value reached by the total AUM. In the opposite case (high AUM, above 200 millions, and more than 75% of the maximum value) the exit is considered as of type T_+ and that the fund is exiting

from the database because it has reached a sufficient capacity[18]. Those two values are chosen to represent a selective, upper quantile of the distribution of AUM. It may be observed empirically that the distribution of the variable made by the ratio of the last AUM on the maximum AUM of the fund is quite uniform except in one where it reaches a peak. This suggests that a significant part of funds exit the database without experiencing a large decollect of assets. The number of resulting funds are given in Table A.7. The number of funds for the T_+ label are not very sensitive at this level to changes in the quantitative criterions on the AUM selected here.

The result of the estimation are given in Table A.8. The effect of the reference date is quite homogenous for single and fund of funds. However its effect is modulated for single funds as the effect is more important for the T_+ exit. For single funds and T_+ exit, we find that both equity and AUM have a positive influence in the increase of the positive default intensity. This is coherent since better performance and more AUM increases the possibility of the fund to be closed to new investment, and then to stop reporting. Concerning the T_- exit, we recover the former interpretation concerning AUM, but equity seems to be not significant in this case. For fund of funds, we recover for the negative exit the former interpretation for AUM, but which is not significant for the positive exit. The resulting baseline intensities are given in Figure A.15 and A.16.

5 Conclusion

We define in this paper a precise framework to study hedge funds lifetimes. The more convenient way to study durations is to define them as the difference between the last date of performance report (after consolidation of the database) and the inception date (when available) of each fund. Inclusion of censorship is crucial. We obtain that the single risk framework is greatly improved by nonparametric estimation since even if the choice of the log-logistic function is justified for some hedge funds strategies, the parametric hazard intensity are systematically under-estimating the probability of exit in the early times, and behavior on the long-term depends on the strategy, with modes of distributions that are too low, up to two years. The use of covariates is a crucial step, and the inclusion of the absolute date of inception of the fund allows to account for a cohort effect capturing most of the information. Finally, competing risks model are a necessary improvement since a large part of the exiting funds with disclosed data have an exit which may not be explained by poor performance. Not taking the nature of this exit into account would lead to a severe bias in the estimation. Further research is needed on the subject, on the identifiability conditions of the competing risks (by including for instance a rule taking into account the highwatermark) and on the potential heterogeneity arising through the dynamic nature of the link between performance, AUM, liquidity and survival.

[18]This method is close to the one of Liang and Park (2008) who underlines that AUM and performance are better indicators of failure than status. They identify real failure when three criterions are met: fund has stopped reporting, performance is deteriorating on the last six months, and AUM is decreasing during the last year. They also use highwater mark to improve the identification of true liquidation.

A Appendix

A.1 Hedge Funds databases biases in the literature

We summarize here some major contributions of the literature studying Hedge Funds databases biases. One of the main difficulty comes from the nature of the data and the multiplicity of their sources. There are several databases with common drawbacks, and the conclusions may differ from one study to one other.

One of these drawbacks is called *backfilling*: some hedge funds communicate a posteriori past performances to databases. For instance, an existing fund in 2004 survives until 2008 and publishes now a performance relative to the 2004 period. The bias is that we dispose now of a past track of its performances, conditionally on the fact that we know that the fund still exists. We miss the funds that have not entered yet the database (see Amin and Kat (2002)) and that the fund may have selected attractive past returns or produce proforma (that is to say, virtual or historically re-estimated) figures.

One other known bias is the *survivorship bias*. Several definitions are possible (Brown et al. (1999), Fung and Hsieh (2000), Amin and Kat (2002)). One is the difference between average return of all existing funds at the end of the sample period and the average return of a portfolio containing all the funds of the sample. A second and more restrictive definition is the difference between the average return of funds that have specifically survived to all the sample period and the mean return of a portfolio containing all the funds of the sample. Amin and Kat (2002) shows that estimating quantities only on survivors induces an upward bias of 2% for average returns. The potential hazard for investors interested in hedge funds is then to over-allocate in this asset class. This bias is even more pronounced when considering small funds (bias that may be close to 5%). For large funds, this bias is close to zero. This confirms that funds with huge assets under management that stop to report, are probably not liquidated. Some studies estimate this bias depending on the investment style, or try to quantify similar effects on higher order moments. For more developments, see Ackermann et al. (1999) and Liang (2000).

Finally, the look-ahead bias arises when studies are made conditional on survival on further consecutive periods (see Baquero et al. (2002)).

A.2 Single risk estimation

Censorship can be of two kinds. First, there is a right-censorship[19]. For dead funds, the duration can be observed. For the other ones, their potential exit cannot be observed and then their lifetime duration is truncated. The second kind of censorship is due to the fact that we observe only discrete durations (in months).

A.2.1 Parametric estimation

Suppose that the durations have a continuous support, and are potentially censored. Working with a parametric family for the density f_θ where θ is the parameter to be estimated, we get that a non-censored observation t has a contribution to the likelihood equal to $f_\theta(t)$. However, for a censored observation t, the contribution to the likelihood is equal to:

$$\int_t^\infty f_\theta(u)du.$$

The total likelihood for a set of observations $(\tau_i)_{i\in[1;n]}$ with a censor variable $(\delta_i)_{i\in[1;n]}$ (with δ_i equal to 1 if the observation is censored, 0 otherwise) is equal to:

$$L(\tau_1,\dots,\tau_n,\theta)=\prod_{i=1}^{n}(f_\theta(\tau_i))^{\delta_i}\Big(\int_{\tau_i}^\infty f_\theta(u)du\Big)^{1-\delta_i}. \tag{4}$$

Then, a maximum likelihood estimation can be set up to estimate the parameters that fit best the distribution.

The general form of the likelihood is easy to derive in the most general case. We note generically $f(t,x,\theta)$ the density, $\lambda(t,x,\theta)$ the hazard function and $S(t,x,\theta)$ the survivor function, depending on time t, observation of the covariate x and parameters θ (we suppose that x includes the possibility to give information on censorship). The log-likelihood is:

$$\begin{aligned} log(L(\tau_1,\dots,\tau_n,x_1,...,x_n,\theta)) &= \sum_{i\in I_u} log(f(\tau_i,x_i,\theta)) - \sum_{i\in I_c} log(S(\tau_i,x_i,\theta)) \\ &= \sum_{i\in I_u} log(\lambda(\tau_i,x_i,\theta)) - \sum_{i} log(S(\tau_i,x_i,\theta)) \end{aligned} \tag{5}$$

where I_c is the set of censored observations, I_u the uncensored ones. The second expression can be obtained thanks to the link between the density, survivor and hazard function, as $f(t)=\lambda(t)S(t)$[20].

Now, we must define an extension to the case where observed durations are discrete (second kind of censorship). We face this situation since our Hedge Fund lifetimes are estimated as a number of months. A duration τ means that the Hedge Fund has failed during the time interval $[\tau;\tau+1[$. We propose to replace in the likelihood of expression (4) the contribution of uncensored data $f_\theta(t_i)$, by the quantity $\int_{\tau_i}^{\tau_i+1} f_\theta(u)du$. Then the likelihood to be maximized becomes:

$$L(\tau_1,\dots,\tau_n,\theta)=\prod_{i=1}^{n}\Big(\int_{\tau_i}^{\tau_i+1} f_\theta(u)du\Big)^{\delta_i}\Big(\int_{\tau_i}^\infty f_\theta(u)du\Big)^{1-\delta_i}.$$

In terms of survival function, if $S_\theta(t)=\int_t^{+\infty} f_\theta(t')dt'$ the likelihood becomes:

$$L_{dis}((t_i)_{i\in[1;n]},\theta)=\prod_{i=1}^{n}(S_\theta(t_i)-S_\theta(t_i+1))^{\delta_i}(S_\theta(t_i))^{1-\delta_i}$$

[19] We could also include the problem of left-censored data when we are doubtful on the exact beginning of the lifetime. On this point see the work of Patilea and Rolin (2006).

[20] In an advanced modelling framework, this formula holds only if the conditioning set is well defined and regularity conditions on the survivor function are satisfied. See Hautsch (2004)

This form of likelihood is straightforward and is also presented in Hautsch (2004). This approach could also be compared to the discrete choice model developed in Lunde et al. (1999), even if it's not similar.

A.2.2 Nonparametric estimation

One first estimates $\Lambda(t)$ (or at least its increments) and then one tries to find a smooth expression for its derivative in order to get $\hat{\lambda}(t)$ (see Ramlau-Hansen (1983)). Suppose that we have an estimator $\hat{\Lambda}(t)$ at dates $0 = \tau_0 \leq \tau_1 \leq \ldots \leq \tau_n$. Using a gaussian kernel, $\hat{\lambda}(t)$ is estimated through :

$$\hat{\lambda}(t) = \sum_{k=1}^{n} \frac{1}{\sqrt{2\pi}h_n} exp\left(-\frac{(t-\tau_i)^2}{2h_n^2}\right) \times (\Delta\hat{\Lambda}(\tau_i))$$

with h_n an associated bandwidth. We can improve this formula by using boundary kernels : for the terms corresponding to little values of τ_i ($\tau_i \leq h_n$) we adopt the correction proposed by Li and Racine (2006) :

$$\hat{\lambda}(t) = \sum_{k=1}^{n} \frac{1}{\sqrt{2\pi}h_n} exp\left(-\frac{(t-\tau_i)^2}{2h_n^2}\right) \times (\Delta\hat{\Lambda}(\tau_i)) \times \left(\mathbf{1}_{\{\tau_i \geq h_n\}} + \frac{\mathbf{1}_{\{\tau_i < h_n\}}}{1 + \Phi(-\frac{\tau_i}{h_n})}\right)$$

with Φ the cumulative function of the standard gaussian distribution. We have not yet defined how to estimate the integrated hazard $\Lambda(t)$. Several equivalent approaches are available. First, the Kaplan-Meier estimator is a product-limit estimator. It is obtained by specifying that being alive at time t is equivalent to being alive just before t and not dying at this date. Its expression is:

$$\hat{\Lambda}_{KM}(t) = -ln(\hat{S}_{KM}(t)) \qquad \text{with} \qquad \hat{S}_{KM}(t) = \prod_{\{j|\tau_j<t\}} \left(1 - \frac{f_j}{a_j}\right)$$

where f_j is the number of individuals failing at date τ_j (thus only uncensored individuals), and a_j is the total number of individuals which have attained at least τ_j, then including censored individuals.

An other possibility is to use the Nelson-Aalen estimator (see Aalen (1978)) which writes:

$$\hat{\Lambda}_{NA}(t) = \sum_{i=1}^{n} \frac{(1-\delta_i)\mathbf{1}_{\{\tau_i \leq t\}}}{\sum_{j=1}^{n} \mathbf{1}_{\{\tau_i \leq \tau_i\}}}$$

where $\delta_i = 1$ if the observation is censored, 0 otherwise.

Finally, an other estimation is used in Couderc et al. (2008) which uses Gamma-Ramlau Hansen estimator. Its expression is (see Couderc (2005)) :

$$\hat{\lambda}_{GRH}(t) = \sum_{i=1}^{n} \frac{\tau_i^{t/h_n} exp(-\tau_i/h_n)}{h_n^{t/(h_n+1)}\Gamma(\frac{t}{h_n}+1)} \times (\Delta\hat{\Lambda}(\tau_i))$$

where h_n is again the smoothing parameter and δ_i the censorship indicator defined as above.

A.3 Descriptive Statistics

Strategy	**Alive**	**Dead**	**Total**
All Funds	6121	8102	14223
Fund of Funds	2158	1653	3811
Single Funds	3963	6449	10412
Long/Short Equity Hedge	1529	2024	3553
Equity Market Neutral	212	415	627
Event Driven	281	489	770
Dedicated Short Bias	16	33	49
Fixed Income Arbitrage	159	310	469
Convertible Arbitrage	64	198	262
Emerging Markets	307	323	630
Multi-Strategies	580	517	1097
Options Strategies	9	2	11
CTA-Managed Futures	497	1784	2281
Global Macro	252	344	596
Other-Undefined	57	10	67

Table A.1: Number of funds for single risk estimation

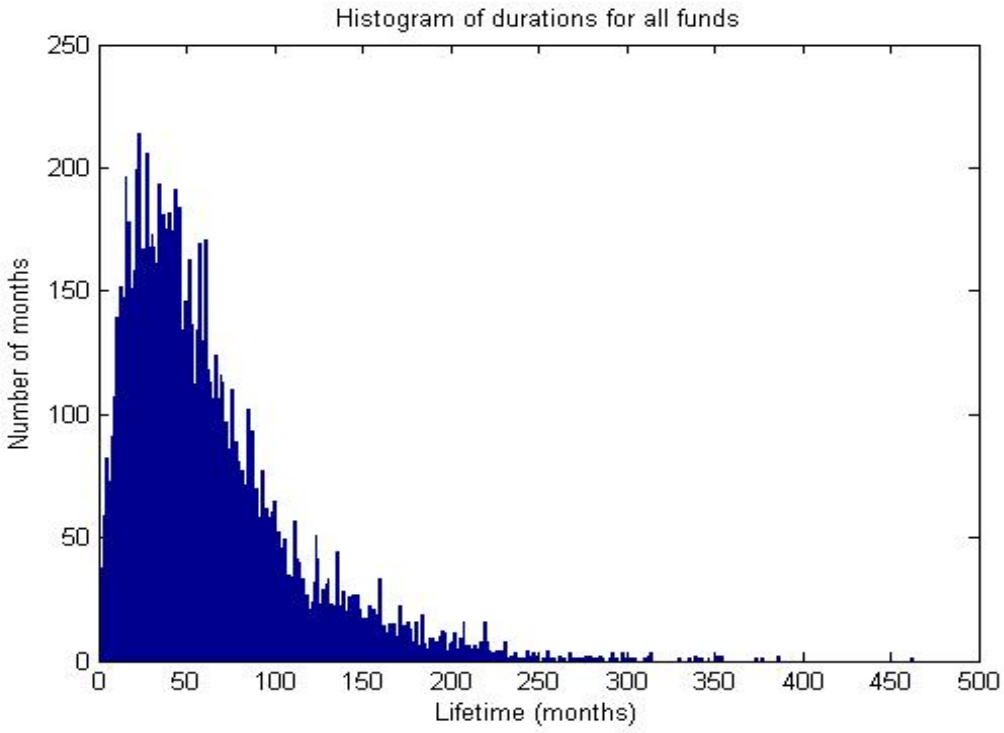

Figure A.1: Empirical distribution of durations (in months).

Strategy	**Mean**	**Qu.**(5%)	**Qu.**(95%)	**Mode**	**Alive**	**Dead**
All Funds	62, 0	11	158	23	6121	8102
Fund of Funds	61, 0	12	149	23	2158	1653
Single Funds	62, 4	11	160	35	3963	6449
Long/Short Equity Hedge	62, 1	12	154	35	1529	2024
Equity Market Neutral	54, 8	12	134	24	212	415
Event Driven	70, 4	12	184	61	281	489
Fixed Income Arbitrage	59, 1	12	144	42	159	310
Emerging Markets	61, 2	12	160	17	307	323
Multi-Strategies	53, 3	8	145	22	580	517
CTA-Managed Futures	69, 3	10	189	29	497	1784
Global Macro	51, 8	7	135	44	252	344

Table A.2: Durations (in months) statistics and corresponding number of funds

A.4 Single risk estimation : empirical results

A.4.1 Parametric estimation

Strategy	μ	σ	**Mode**	ρ	κ
All Funds	$4,2762$	$0,5638$	$35,0$	$0,0139$	$1,7737$
	$[4,2581; 4,2944]$	$[0,5538; 0,5739]$			
Fund of Funds	$4,4859$	$0,5637$	$43,2$	$0,0113$	$1,7740$
	$[4,4469; 4,5250]$	$[0,5419; 0,5864]$			
Single Funds	$4,2073$	$0,5615$	$32,9$	$0,0149$	$1,7808$
	$[4,1866; 4,2279]$	$[0,5504; 0,5729]$			
Long/Short Equity Hedge	$4,2733$	$0,5669$	$34,6$	$0,0139$	$1,7639$
	$[4,2369; 4,3097]$	$[0,5470; 0,5876]$			
Equity Market Neutral	$4,0348$	$0,5136$	$31,6$	$0,0177$	$1,9470$
	$[3,9598; 4,1099]$	$[0,4746; 0,5558]$			
Event Driven	$4,3104$	$0,5382$	$39,0$	$0,0134$	$1,8579$
	$[4,2385; 4,3824]$	$[0,5004; 0,5790]$			
Fixed Income Arbitrage	$4,1287$	$0,4818$	$37,4$	$0,0161$	$2,0754$
	$[4,0465; 4,2109]$	$[0,4399; 0,5278]$			
Emerging Markets	$4,3617$	$0,5499$	$39,7$	$0,0128$	$1,8184$
	$[4,2727; 4,4508]$	$[0,5042; 0,5998]$			
Multi-Strategies	$4,2603$	$0,5979$	$31,0$	$0,0141$	$1,6725$
	$[4,1852; 4,3353]$	$[0,5574; 0,6413]$			
CTA-Managed Futures	$4,0821$	$0,5834$	$27,2$	$0,0169$	$1,7140$
	$[4,0390; 4,1251]$	$[0,5615; 0,6062]$			
Global Macro	$4,0984$	$0,5575$	$29,9$	$0,0166$	$1,7938$
	$[4,0103; 4,1865]$	$[0,5115; 0,6077]$			

Table A.3: Single risk parametric estimation for a log-logistic specification.

A.4.2 Comparison between parametric and nonparametric intensities

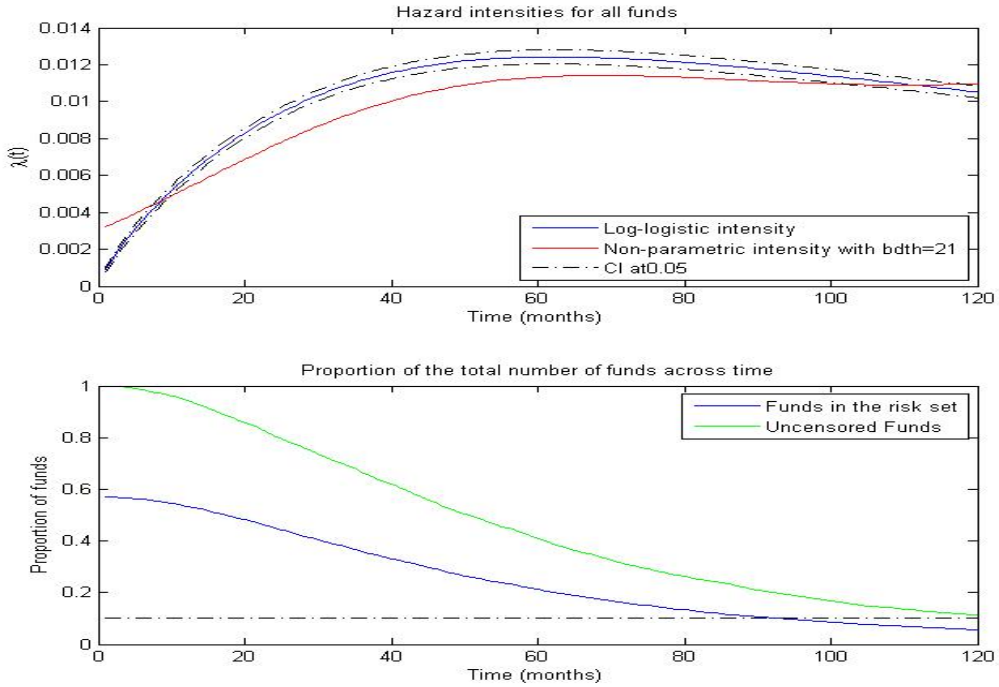

Figure A.2: Parametric and nonparametric intensities for all funds.

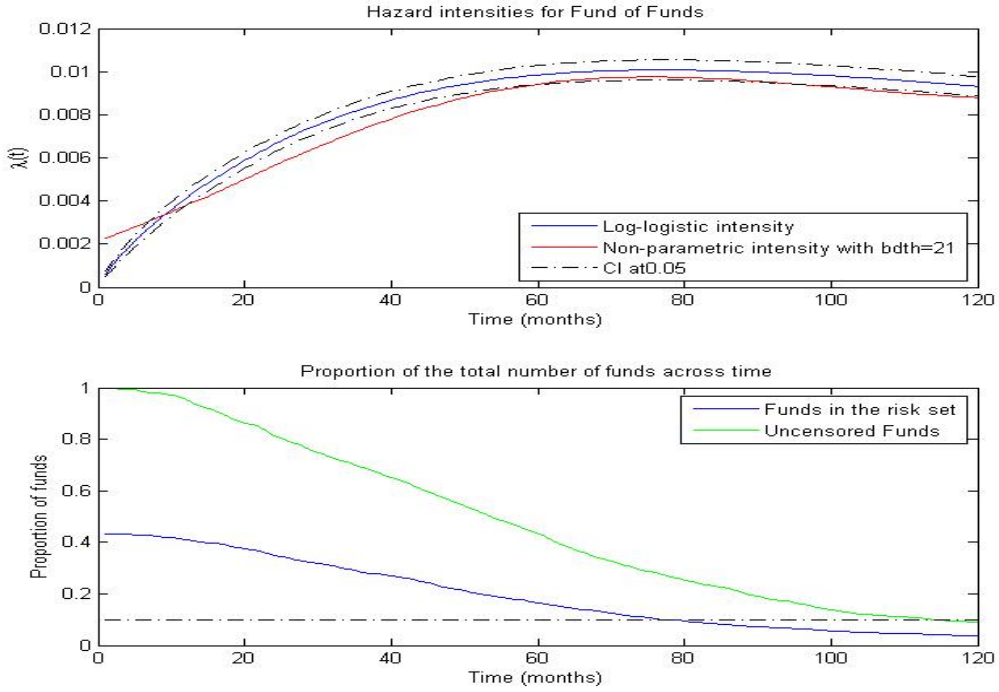

Figure A.3: Parametric and nonparametric intensities for Fund of Funds.

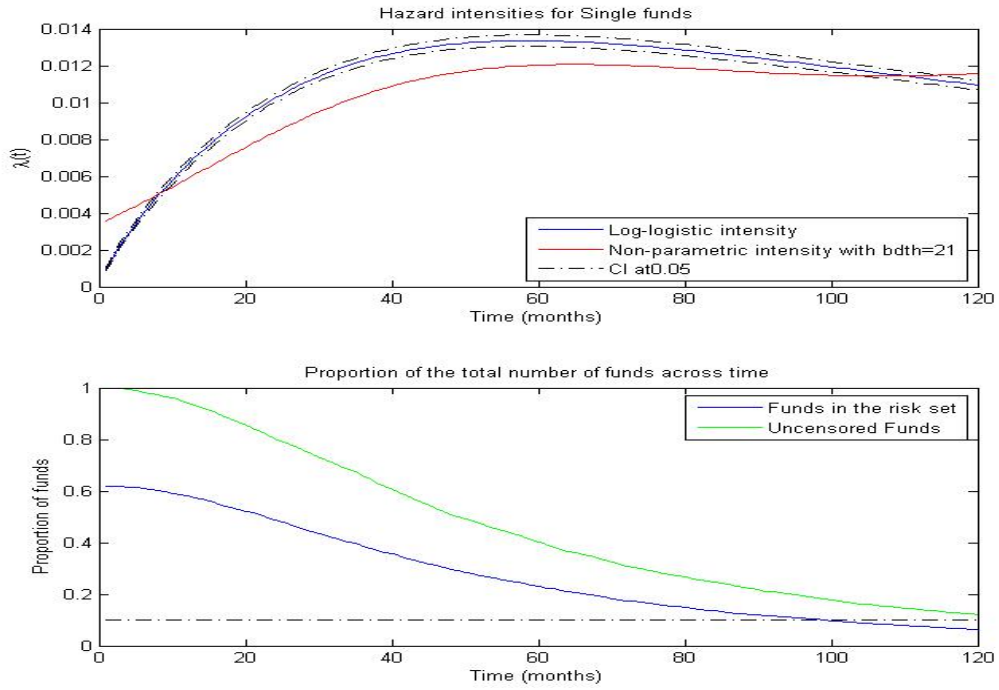

Figure A.4: Parametric and nonparametric intensities for Single Funds.

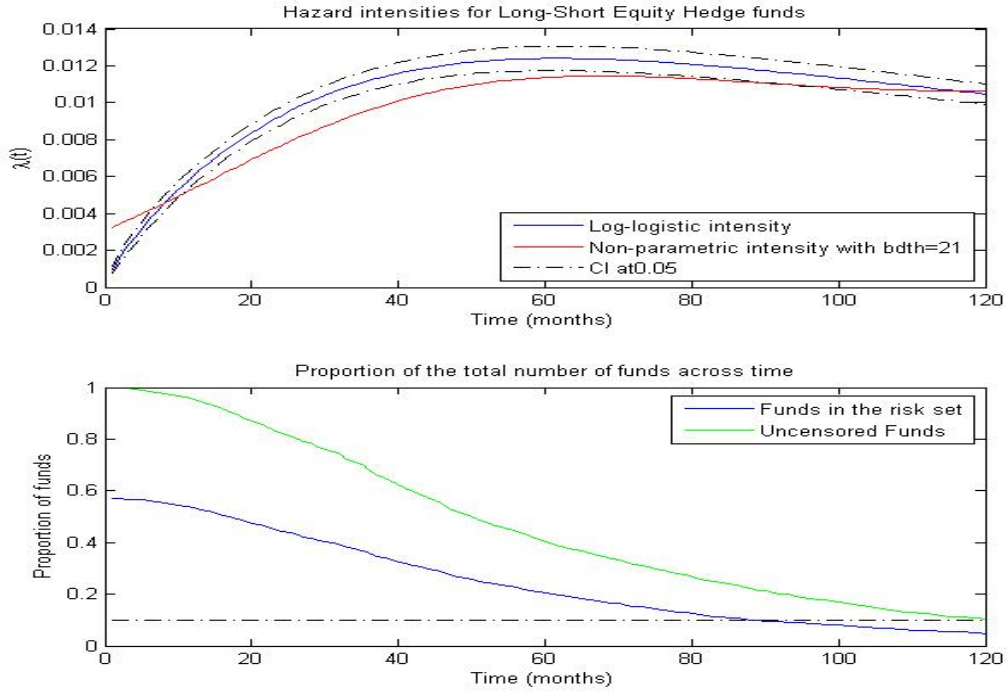

Figure A.5: Parametric and nonparametric intensities for Long-Short Equity Hedge funds.

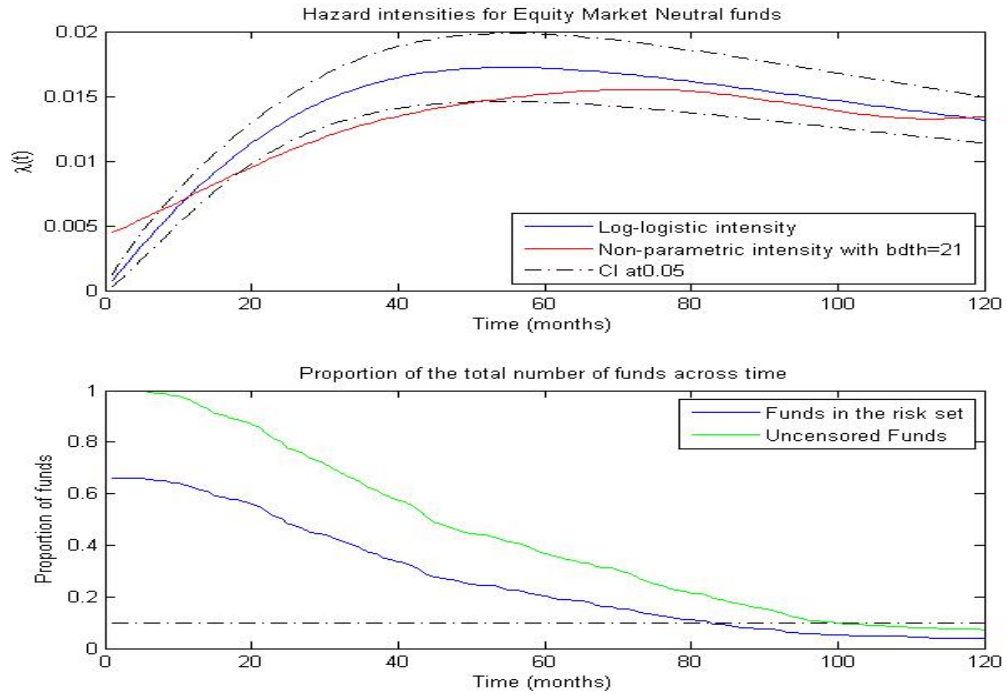

Figure A.6: Parametric and nonparametric intensities for Equity Market Neutral funds.

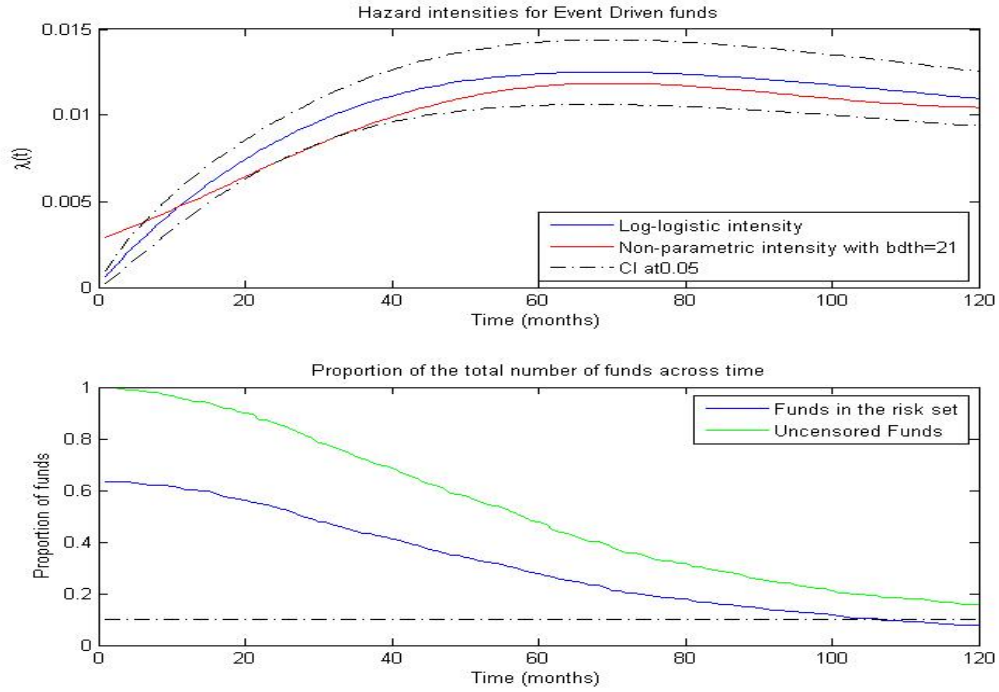

Figure A.7: Parametric and nonparametric intensities for Event Driven funds.

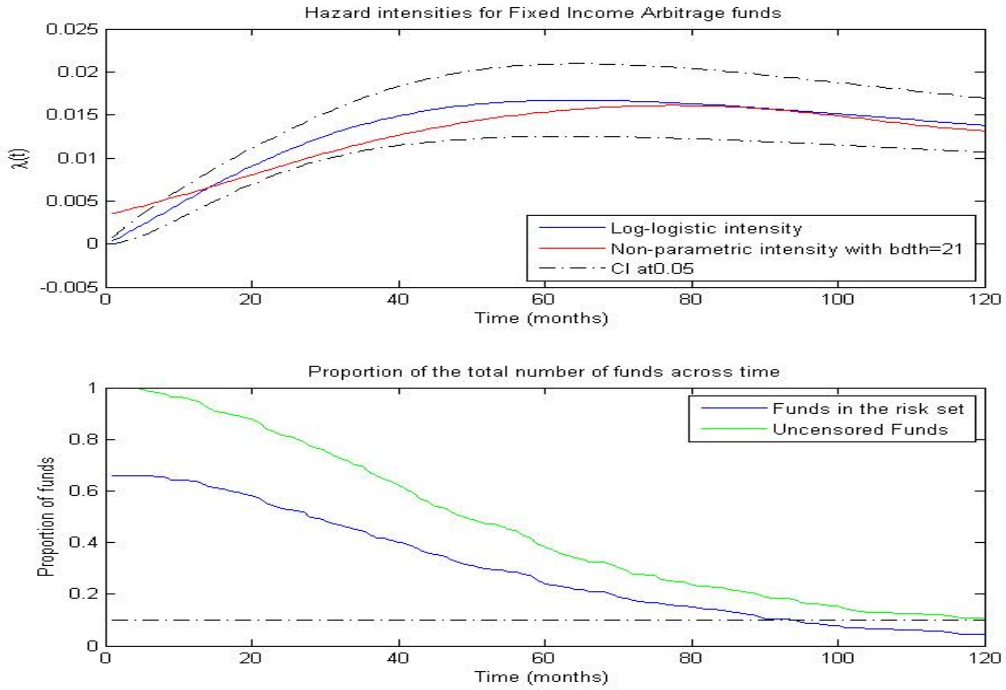

Figure A.8: Parametric and nonparametric intensities for Fixed Income Arbitrage funds.

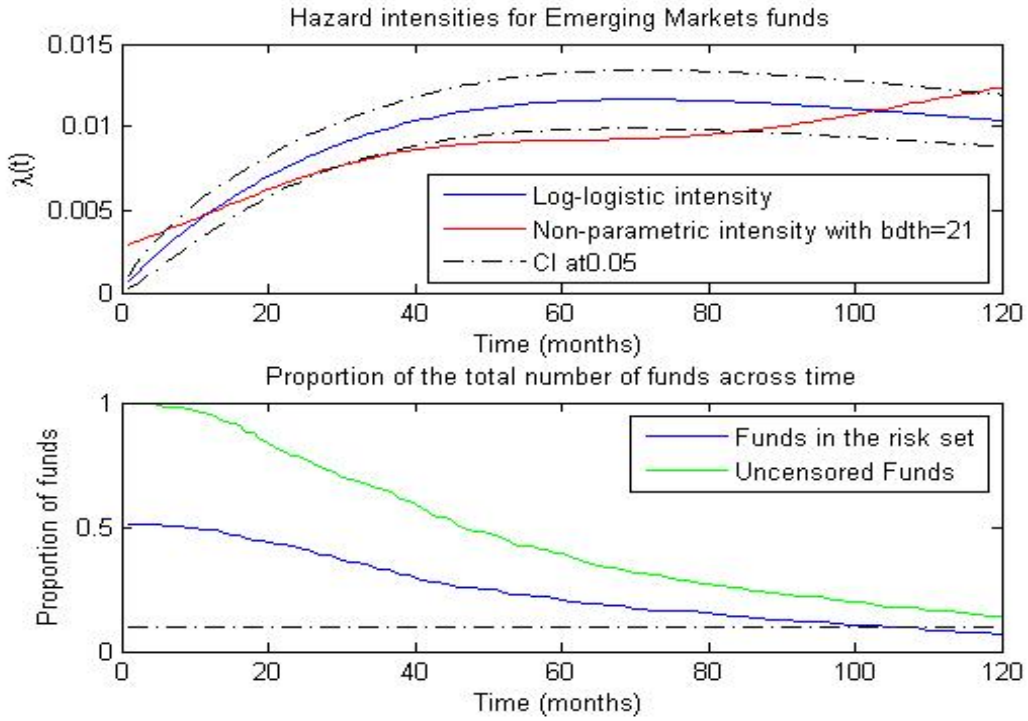

Figure A.9: Parametric and nonparametric intensities for Emerging Markets funds.

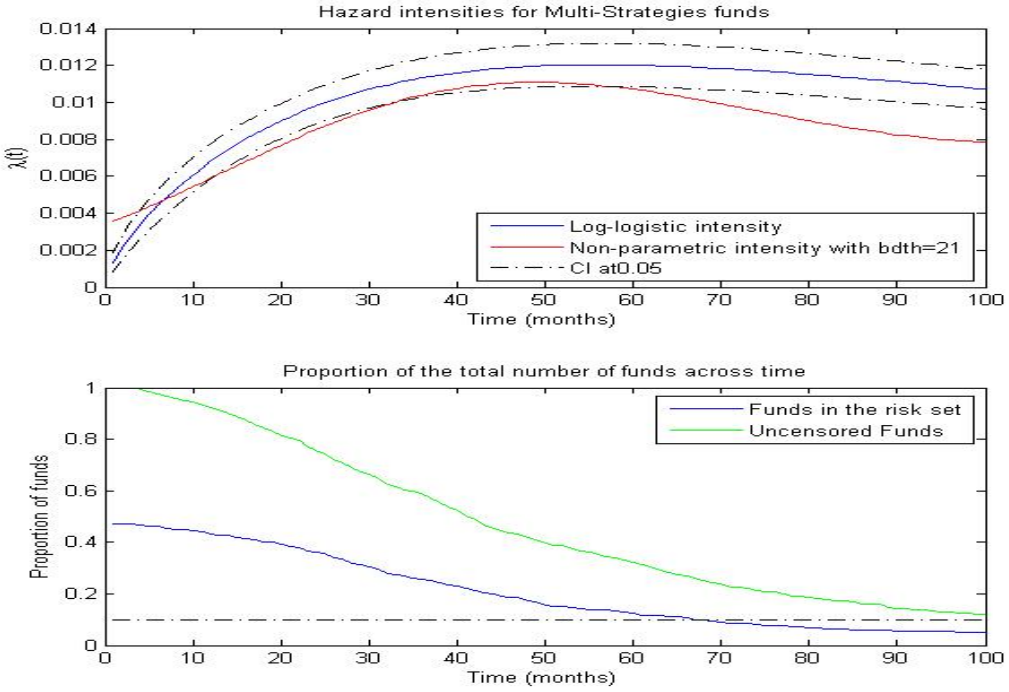

Figure A.10: Parametric and nonparametric intensities for Multi-Stategies funds.

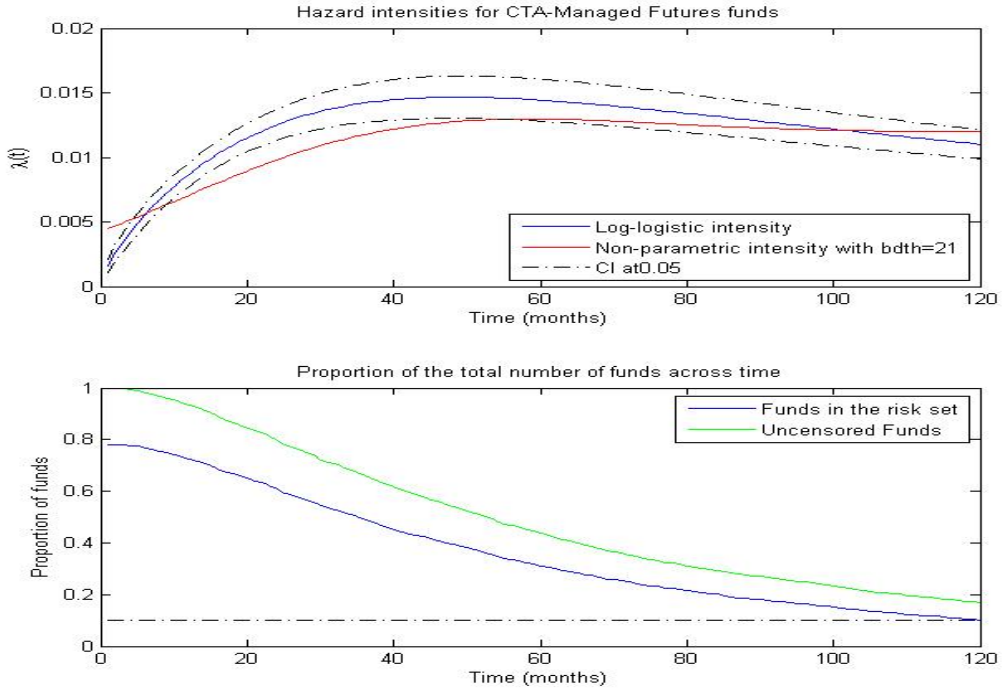

Figure A.11: Parametric and nonparametric intensities for CTA-Managed Futures funds.

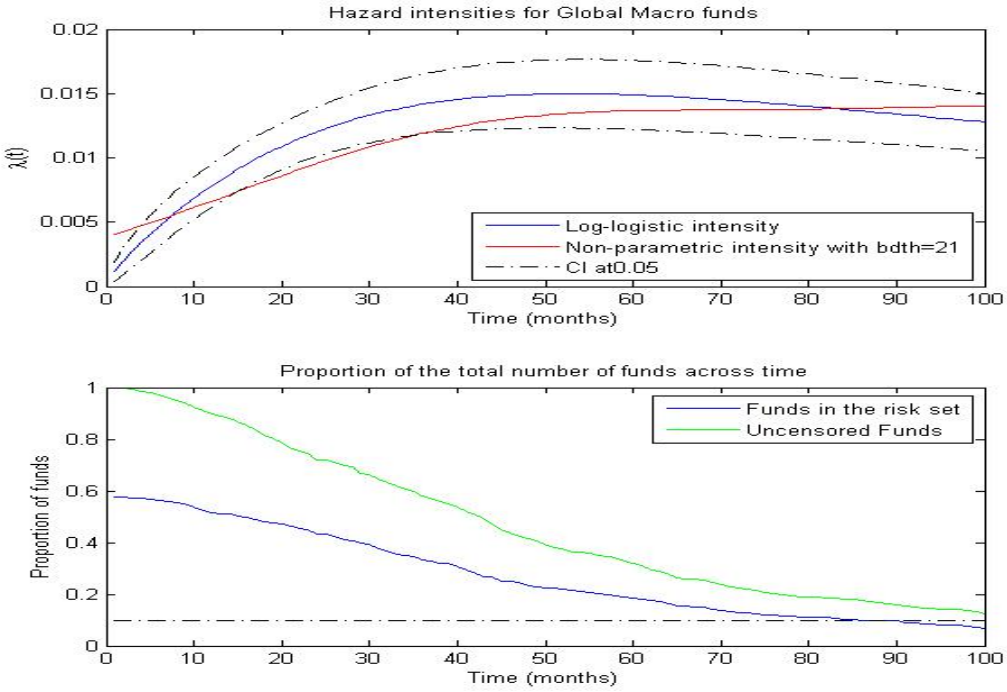

Figure A.12: Parametric and nonparametric intensities for Global Macro funds.

A.4.3 Single risk estimation with covariates

Strategy	**Alive**	**Dead**	**Total**
Single Funds	1075	2487	3562
Fund of Funds	281	415	696

Table A.4: Number of funds for a single risk analysis with dynamic covariates.

Var. 1	**Var. 2**	**Var. 3**	β_1	β_2	β_3
Equity	-	-	$-0,216$	-	-
			$[-0,259; -0,174]$		
ln(AUM)	-	-	$-0,239$	-	-
			$[-0,297; -0,182]$		
Equity	*ln(AUM)*	-	$-0,144$	$-0,218$	
			$[-0,176; -0,112]$	$[-0,279; -0,157]$	
Equity	*ln(AUM)*	*Date Ref*	$-0,0946$	$-0,2863$	$4,20.10^{-4}$
			$[-0,123; -0,066]$	$[-0,351; -0,221]$	$[3,14.10^{-4}; 5,26.10^{-4}]$
Returns	Δ *ln(AUM)*	-	$-3,61$	$-0,716$	-
			$[-5,12; -2,10]$	$[-1,06; -0,367]$	-
Returns	Δ *ln(AUM)*	*Date Ref*	$1,41.10^{-4}$	$-2,03.10^{-4}$	$3,65.10^{-4}$
			$[-2,23; 13,6]$	$[-0,841; 0,841]$	$[2,63.10^{-4}; 4,65.10^{-4}]$

Table A.5: Single risk estimation for Single Funds with dynamic covariates

Var. 1	**Var. 2**	**Var. 3**	β_1	β_2	β_3
Equity	-	-	-0.035	-	-
			$[-0,057; -0,014]$		
ln(AUM)	-	-	$-0,202$	-	-
			$[-0,263; -0,141]$		
Equity	*ln(AUM)*	-	$-0,015$	$-0,199$	-
			$[-0,026; -0,004]$	$[-0,263; -0,135]$	
Equity	*ln(AUM)*	*Date Ref*	$0,0031$	$-0,2948$	$4,69.10^{-4}$
			$[-0,003; 0,009]$	$[-0,363; -0,227]$	$[3,53.10^{-4}; 5,84.10^{-4}]$
Returns	Δ *ln(AUM)*	-	$-6,38$	$-0,925$	-
			$[-9,26; -3,49]$	$[-1,23; -0,618]$	-
Returns	Δ *ln(AUM)*	*Date Ref*	$-2,18.10^{-4}$	$-2,06.10^{-4}$	$3,93.10^{-4}$
			$[-4,13; 4.13]$	$[-0,770; 0,770]$	$[2,77.10^{-4}; 5,10.10^{-4}]$

Table A.6: Single risk estimation for Fund of Funds with dynamic covariates

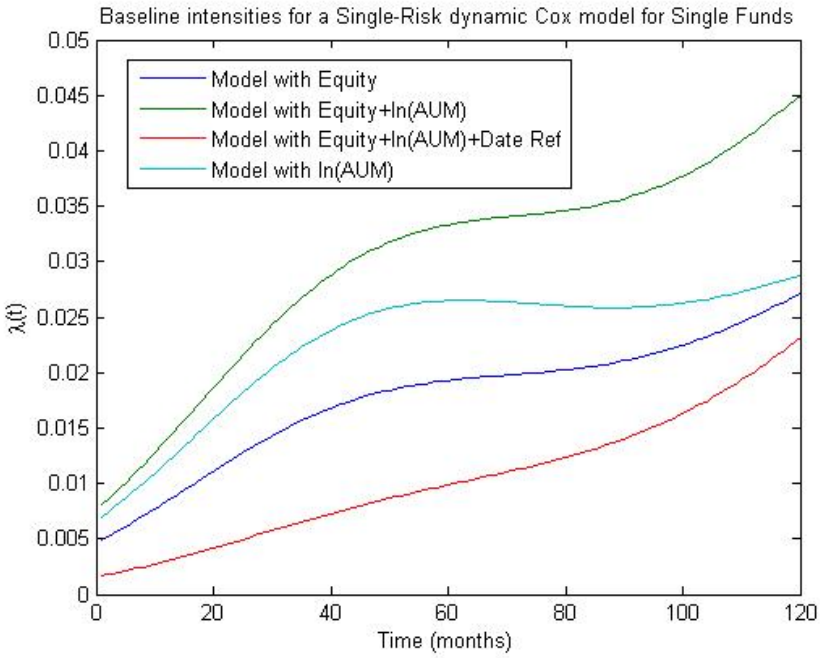

Figure A.13: Nonparametric baseline intensities of nested dynamic Cox models for single risk estimation for Single Funds.

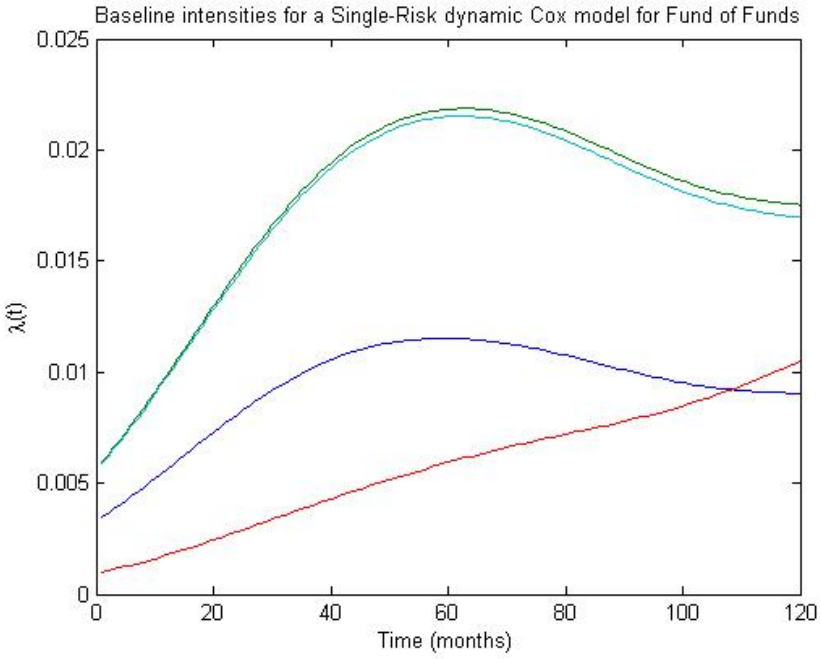

Figure A.14: Nonparametric baseline intensities of nested dynamic Cox models for single risk estimation for Fund of Funds.

A.5 Competing risks

Strategy	Censored	T_+	T_-
Single Funds	1075	1018	1469
Fund of Funds	281	215	200

Table A.7: Resulting number of funds for each category of risk.

Single Funds					
T_+					
Var.1	**Var.2**	**Var.3**	β_1	β_2	β_3
Equity	*Ln(AUM)*	*Date Ref*	$0,0592$	$0,186$	$5,42.10^{-4}$
T_-					
Var.1	**Var.2**	**Var.3**	β_1	β_2	β_3
Equity	*Ln(AUM)*	*Date Ref*	$-3,61.10^{-3}$	$-0,364$	$3,36.10^{-4}$
Fund of Funds					
T_+					
Var.1	**Var.2**	**Var.3**	β_1	β_2	β_3
Equity	*Ln(AUM)*	*Date Ref*	$-0,115$	$3,00.10^{-2}$	$4,85.10^{-4}$
T_-					
Var.1	**Var.2**	**Var.3**	β_1	β_2	β_3
Equity	*Ln(AUM)*	*Date Ref*	$0,0175$	$-0,568$	$4,50.10^{-4}$

Table A.8: Parameter estimates for a competing risks model with covariates.

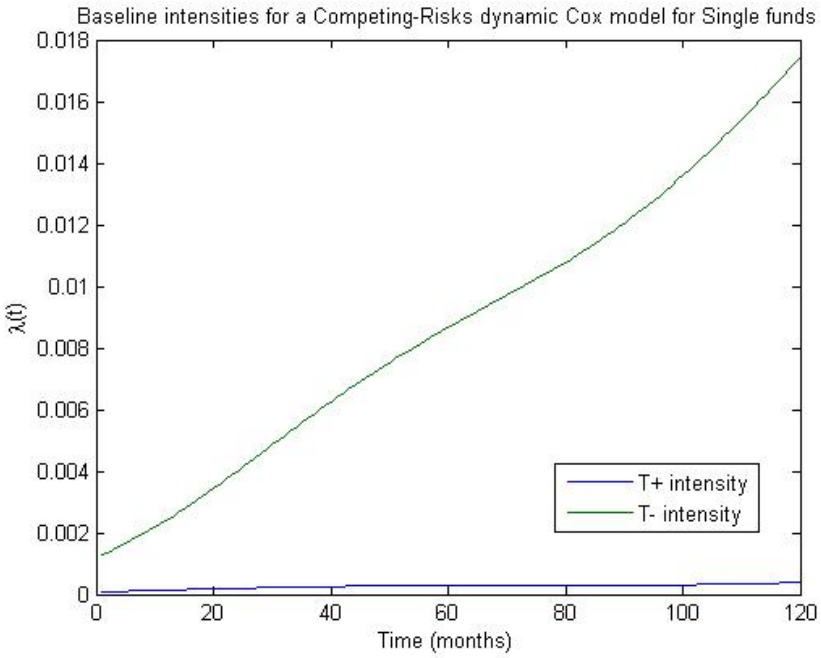

Figure A.15: Nonparametric baseline intensities of competing risks for Single Funds for a dynamic Cox model (covariates : Equity, Ln(AUM), Date Ref).

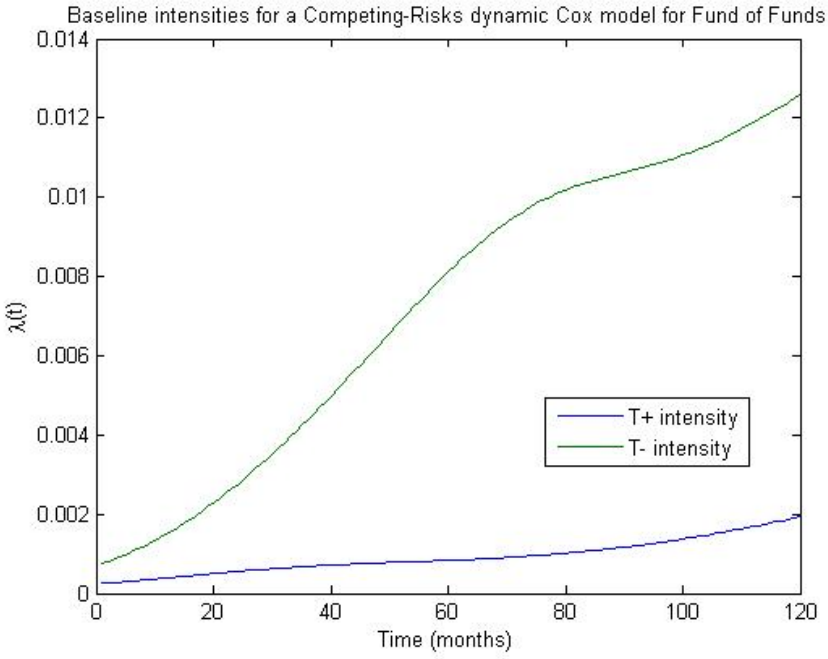

Figure A.16: Nonparametric baseline intensities of competing risks for Fund of Funds for a dynamic Cox model (covariates : Equity, Ln(AUM), Date Ref).

A.6 Academic Contributions

We sum-up here the approaches of the main academic contributions on hedge fund lifetimes, depending on the chosen model (parametric or nonparametric approach, use of covariates, dynamic or static covariates, etc.). The legend is the following :

- NDM : No Duration Model
- PWC : Parametric without Covariates
- NPWC : Nonparametric Without Covariates
- PSC : Parametric with Static Covariates
- PDC : Parametric with Dynamic Covariates
- SPSC : Semi-Parametric with Static Covariates
- SPDC : Semi-Parametric with Dynamic Covariates
- CR : Competing Risks
- DCM : Discrete Choice Model

	NDM	PWC	NPWC	PSC	PDC	SPSC	SPDC	CR	DCM
Amin and Kat (2002)	√								
Lunde et al. (1999)							√		√
Baquero et al. (2002)	√								√
Boyson (2002)							√		
Fung and Hsieh (1997)	√								
Ang and Bollen (2008)					√		√		
Barès et al. (2001)			√						
Brown et al. (2001)							√		
Gregoriou (2002)			√			√			
Grecu et al. (2007)		√				√			
Rouah (2005)							√	√	
Liang and Park (2008)							√		

A.7 Empirical method for missing data

Working with dynamic covariates, as soon as a fund does not provide data for a given month, it cannot enter the estimation, as the partial likelihood procedure needs at any date the value of the covariate for all fund in the risk set. Undisclosures may be caused by the fact that funds may forget to report at some dates, due to exogenous reasons. More precisely, and contrary to performance which is quite well reported, AUM may be often missing. Our correction method will then only be focused on treating missing AUM.

We correct data in two cases. First, when AUM is explicitly provided at the beginning and at the end of the lifetime of the fund. Then it happens that AUM is missing for isolated dates. If missing data represent less than 33% of the total length of the track, they are replaced by the first next available AUM. If AUM is missing at the end of the track, the data are forward-filled and replaced by the last available AUM.

An important lack of data may also occur at the beginning of the life of the fund. As in this period of the life of the fund AUM is particularly volatile, we only keep funds with missing data inferior to six months.

References

Aalen, O.: 1978, Nonparametric Inference for a family of counting processes, *The Annals of Statistics* **6**(4), 701–726.

Abbring, J. and van den Berg, G.: 2003, The Nonparametric Identification of Treatment Effects in Durations Models.

Ackermann, C., McEnally and Ravenscraft: 1999, The Performance of Hedge Funds : Risk, Return and Incentives, *Journal of Finance* **54**(3), 833–874.

Amin, G. and Kat, H.: 2002, Hedge Fund Attrition and Survivorship bias over the period 1994-2001, *Cass Business School Research Centre Working Paper* .

Ang, A. and Bollen, N.: 2008, Locked Up by a Lockup : Valuing Liquidity as a Real Option, *Working Paper* .

Avellaneda, M. and Besson, P.: 2005, Hedge Funds : How big is big, *Working Paper* .

Baquero, G., ter Horst, J. and Verbeek, M.: 2002, Survival, Look-Ahead bias and the performance of Hedge Funds, *Working Paper* .

Barry, R.: 2002, Hedge Funds : a walk through the graveyard, *Applied Finance Center, Macquarie University, Sydney, Australia, Working Paper* .

Barès, P., Gibson, R. and Gyger, S.: 2001, Style Consistency and Survival Probability in the Hedge Fund's Industry, *Working Paper* .

Boyson, N.: 2002, How are Hedge Fund Managers Characteristics Related To Performance, Volatility and Survival, *Ohio State University, Working Paper* .

Brown, S., Goetzmann, W. and Ibbostson, R.: 1999, Offshore Hedge Funds : Survival and Performance, *Journal of Business* **72**, 91–117.

Brown, S., Goetzmann, W. and Park, J.: 1997, Conditions for Survival : changing risk and the performance of hedge fund managers and CTAs, *Working Paper* .

Brown, S., Goetzmann, W. and Park, J.: 2001, Careers and Survival : Competition and Risk in the Hedge Fund and CTA industry, *Journal of Finance* **56**(5), 1869–1886.

Couderc, F.: 2005, Understanding Default Risk Through Nonparametric Intensity Estimation, *Working paper* .

Couderc, F., Renault, O. and Scaillet, O.: 2008, Business and Financial Indicators : What are the Determinants of Default Probability Changes?, *in Credit Risk : Models, Derivatives, and Management, Chapman and Hall, Financial Mathematics Series* pp. 235–268.

Cox, D. and Oakes, D.: 1984, Analysis of Survival Data, *Chapman & Hall* .

Derman, E.: 2007, A Simple Model for the Expected Premium for Hedge Fund Lockups, *Journal Of Investment Management* **5**, 5–15.

Dewanjy, A. and Sengupta, D.: 2007, Estimation of Competing Risks with General Missing Pattern in Failure Types, *Biometrics* **59**, 1063–1070.

Fung, W. and Hsieh, D.: 1997, Survivorship bias and investment style in the return of CTAs, *Journal of Portfolio Management* **24**(1), 30–41.

Fung, W. and Hsieh, D.: 2000, Performance Characteristics of Hedge Funds and Commodity Funds: Natural vs. Spurious Biases, *Journal of Financial and Quantitative Analysis* **35**, 291–307.

Getmansky, M., Lo, A. and Mei, S.: 2004, Sifting Through the Wreckage: Lessons from Recent Hedge-Fund Liquidations, *Journal of Investment Management, Fourth Quarter* **2**(4), 6–38.

Grecu, A., Malkiel, B. and Saha, A.: 2007, Why do Hedge Funds stop reporting their performance?, *Journal of Portfolio Management* **34**(1).

Gregoriou, G.: 2002, Hedge Fund survival lifetime, *Journal of Asset Management* **3**(3), 237–252.

Gregoriou, G. and Rouah, F.: 2002, Is Size a Factor in Hedge Fund Performances?, *Derivatives Use, Trading and Regulation* **7**(4), 301–306.

Hautsch, N.: 2004, Modelling Irregularly Spaced Financial Data, *Springer* .

Heckman, J. and Honoré, B.: 1989, The Identifiability if the competing risks model, *Biometrika* **76**, 325–330.

Hendricks, D., Patel, J. and Zeckhauser, R.: 1997, The J-Shape Of Performance Persistence Given Survivorship Bias, *The Review of Economics and Statistics, MIT Press* **79**(2), 161–166.

Kalbfleisch, J. and Prentice, R.: 2002, *The Statistical Analysis of Failure Time Data, 2nd Edition.*

Kundro, C. and Feffer, S.: 2003, Understanding and Mitigating Operational Risk in Hedge Fund, *Capco White Paper* .

Li, Q. and Racine, J.-S.: 2006, *Nonparametric Econometrics : Theory and Practice.*

Liang, B.: 2000, Hedge Funds : the Living and the Dead, *Journal of Financial and Quantitative Analysis* **35**(3), 309–335.

Liang, B. and Park, H.: 2008, Share Restrictions, Liquidity Premium, and Offshore Hedge Funds , *Working Paper* .

Lunde, A., Timmermann, A. and Blake, D.: 1999, The Hazards of Mutual Fund Underperformance : A Cox Regression Analysis, *Journal of Empirical Finance* **6**, 121–152.

Patilea, V. and Rolin, J.-M.: 2006, Product-Limit Estimators of the Survival Function with Twice-Censored Data, *The Annals of Statistics* **34**(2), 925–938.

Pojarliev, M. and Levich, R.: 2008, Trades of the Living Dead : Style Differences, Style Persistece and Performance of Currency Fund Managers, *Working Paper* .

Ramlau-Hansen, H.: 1983, Smoothing counting processes intensities by means of kernel functions, *The Annals of Statistics* **11**(2), 453–466.

Rouah: 2005, Competing Risks in Hedge Funds survival, *Working Paper* .

Tsiatis: 1975, A nonidentifiability aspect of competing risks, *Proc. Nat. Acad. Sci USA* **72**(1), 20–22.