Application of advanced statistical analysis for internal modeling in life insurance

Quang Dien Duong

UNIVERSITÉ PIERRE & MARIE CURIE - SORBONNE UNIVERSITÉS

Doctor of Philosophy (Ph.D.) Thesis

to obtain the title of

PhD of Science

of the University Pierre & Marie Curie
Specialty : Applied Mathematics

Defended by

Quang Dien Duong

Application of advanced statistical analysis for internal modeling in life insurance

Thesis Advisor: Agathe Guilloux and Olivier Lopez

defended on Mars 4, 2021

Jury :

Reviewers :	Mathieu Ribatet	-	École Centrale de Nantes
	Frédéric Planchet	-	Université Claude Bernard - Lyon 1
Advisors :	Agathe Guilloux	-	Université d'Évry Val d'Essonne
	Olivier Lopez	-	Université Pierre & Marie Curie
Examinators :	Michel Broniatowski	-	Sorbonne Université
	Caroline Hillairet	-	ENSAE
	Thomas Lim	-	ENSIIE Evry

Acknowledgments

First of all, I would like to thank Prof. Mathieu Ribatet and Prof. Frédéric Planchet for agreeing to review this thesis. The final version of this thesis benefited from their very careful reading and their valuable comments. I also thank all the members of the jury for agreeing to attend the presentation of this work.

Second, I owe my deepest gratitude to my PhD supervisors, Prof. Agathe Guilloux and Prof. Olivier Lopez. Agathe and Olivier, I would like to thank you for the kindness you have shown toward my work and myself over the past four years. Thank you for sharing with me your scientific knowledge and rigor; they both permeate throughout my PhD thesis today.

I would also like to thank Jean-Baptiste Monnier for being such a great PhD adviser and scientific interlocutor. Jean-Baptiste, thank you for your enlightening scientific advices, for being so welcoming and for your great sense of humor. More importantly, I would like to thank you for your constant encouragements, especially when times were tougher, and for the enjoyable times spent in your office or around a cup of coffee; I hope there will be many more.

My thanks also go to the members of the Risk and Value Measurement Services team and especially Vincent Gibrais, Emmanuel Perrin, Bastien Godrix, Didier Riche and Santiago Hector Fiallos. These meetings were very enriching and we were able to work together and share discussions on many exciting topics.

Contents

Chapter 1

Introduction

Contents

1.1 Presentation of the PwC's R&D project

Research and Development ("R&D") is a very important focus for the development of PricewaterhouseCoopers Advisory. Indeed, the consulting activity requires a constant level of excellence and needs to be constantly at the forefront of innovation. For this reason, PricewaterhouseCoopers Advisory invests in a strong R&D policy and conducts dozens of internal and clients R&D projects every year.

PricewaterhouseCoopers Advisory conducts research in various scientific and technological fields. Thus, several R&D projects are conducted in major scientific disciplines such as mathematics and financial statistics, Big Data, computer security, etc.

In addition, PricewaterhouseCoopers Advisory has set up a department called "Risk and Value Measurement Services" (or "RVMS"). It is a unique center of expertise working with the "financial industry" (major players in the insurance and banking sectors) to respond to their challenges in terms of risk and value.

The "RVMS" cluster brings together nearly 70 modeling and risk experts (actuaries, quantitative engineers and data scientists). These experts master both the quantitative techniques and the functional, operational and regulatory environment in which these techniques are used (Solvency II, Basel III, IFRS evolution, etc.). Trained in auditing methods, they also have significant experience in conducting complex consulting missions.

This team, which relies on in-house technical studies, is organized to meet the expectations of PricewaterhouseCoopers Advisory clients. It also uses PricewaterhouseCoopers' international network to be at the forefront of actuarial methods, tools and quantitative finance.

The RVMS department supervise a CIFRE thesis in partnership with the Pierre and Marie Curie University (Paris 6). This CIFRE thesis entitled "Application of advanced statistical analysis for internal modeling in life insurance" aims to develop innovative methods to effectively address the complex problem posed by the valuation of life insurance commitments and the calculation of their prudential capital cost, and is an integral part of the research project that will be described later in this document.

1.2 Context of the study

The valuation of life insurance liabilities presents real complexity, since it is based in particular on profit-sharing mechanisms which require the simultaneous modeling of these liabilities with the asset items associated with them. This complexity is naturally multiplied in the context of prudential capital calculations by internal models, given that it is then a question of obtaining the distribution of these valuations at one year, in accordance with article 121 of the Solvency 2 directive (please refer to Problem 1 in Section 1.3 for more details). To reduce the degree of complexity of this problem, various approximation, better known under the name of proxy models or loss functions, are proposed in practice. We will detail later how these proxy models are validated in practice and the efforts that remain to be made to improve their reliability. In particular, we suggest justifying the choice of the method used during the validation step of the loss functions. But first, **how are life insurance liabilities valued in practice?**

In practice, life insurance liabilities are valued using the risk-neutral Monte-Carlo method. This method consists of estimating the value of life insurance commitments as the average of the discounted values at the risk-free rate of benefits paid to policyholders for a set of financial trajectories called risk neutral scenarios. This valuation method captures the *optionality of life insurance liabilities* and gives them an economic value, also known as "market consistent". You might ask "**Where does the optionality of life insurance liabilities come from?**"

Conventional life insurance contracts with a savings component usually offer their holders a number of guarantees. Among the main guarantees are minimum guaranteed rates, profit sharing and the right of redemption. Depending on the

contracts, we can also find the cancellation option in annuity, the right to arbitrate between support in EURO and support in Unit-Linked (UL), the UL floor guarantee in the event of death, etc. These guarantees represent rights granted by life insurers to their policyholders and can be considered as options in all respects similar to those treated on the financial markets.

What is the typical calculation time for the economic value of a life insurance liability? Depending on the size of the balance sheet, the complexity of the optionality of life insurance contracts and the finesse of the cash-flow model used, the "market-consistent" valuation of a life insurer's commitments can range from several minutes to several hours. The same is therefore true for the valuation of its own funds.

The Solvency II directive offers life insurers the possibility of using a (partial) internal model in order to assess the forecast probability density of the variation in their own funds over one year (see article 121) and in particular its quantile at 99.5% called "Solvency Capital Requirement" (SCR). In practice, this quantile is estimated by the Monte-Carlo method. This method consists of simulating several tens of thousands of future economic and actuarial environments within one year, revaluing the equity (and therefore the liabilities) of the life insurance company in each of these new states of the world, and identify the annual change in equity associated with the 99.5% percentile.

Given the time required to calculate the value of a life insurance company's own funds using a cash-flow model in a given economic and actuarial environment, the abrupt calculation of the SCR by the Monte- Carlo is not possible in practice. In order to solve this problem, life insurance undertakings have developed "proxy" methods intended to reproduce the results of the cash flow model in a very short time and in any economic and actuarial environment.

So **what are the proxy methods that are used in practice?** The economic and actuarial environment of a life insurance company is in practice modeled by a vector of risk factors. The cash-flow model is therefore a function in the mathematical sense of the term which associates a specific background value with a vector of risk factors. This function will be referred to below as the "equity function". Proxy models are therefore ultimately approximations or estimators of the value of equity function. The problem of approximating the value of equity function is a complex problem. Its complexity is in practice all the greater as the dimension of the underlying risk factor vector is large and the regularity of the equity function is low. Given the inherent difficulty of this problem, no technical solution has yet emerged as being the most appropriate, and a multitude of approaches coexist, each with its own advantages and disadvantages. The best known and also the most frequently used are the Curve-Fitting, Least Square Monte Carlo (LSMC) and Replicating Portfolio methods (see Section 1.4 for more details).

How are these proxy models validated in practice? **Is there a regulatory requirement for error controlling**. As stated in Article 229-(g) of the Delegated Acts [93], deviations caused by the use of these proxy models must be measured and controlled. The validation of these proxy models therefore essentially consists

in developing robust procedures for measuring and controlling the error introduced by the use of the proxy model.

In the first part of this thesis, we will introduce a novel proxy method, which is highly practical, modular, smooth and naturally relates the approximation errors to the Monte-Carlo statistical errors. Furthermore, our approach allows insurance companies to naturally and transparently start reporting confidence levels on their prudential reporting, which is not disclosed so far by insurance companies and would be a relevant information within solvency disclosures for the industry.

In the second part of this thesis, we will deal with the quantile estimation problem when the tail distribution is heavy and covariate information is available. To this end, we rely on extreme value statistics.

In extreme value statistics, estimation of the tail-index is of importance in numerous applications since it measures the tail heaviness of a distribution. Examples include heavy rainfalls, big financial losses, high medical costs, just to name a few. When covariate information is available, we are mainly interested in describing the tail heaviness of the conditional distribution of the dependent variable given the explanatory variables and the tail-index will be thus taken as a function of this covariate information. In many practical applications, the explanatory variables can contain hundreds of dimensions. Many recent algorithms use concepts of proximity in order to estimate model parameters based on their relation to the rest of the data. However, in high dimensional space, the data is often sparse and the notion of proximity fails to retain its meaningfulness. Therefore, this implies deterioration in estimation. The main purpose of this study is thus to overcome this challenge in the context of the tail-index estimation given the explanatory variables.

1.3 Problems with calculating the distribution of basic own funds over a one-year time horizon

Recall that the Solvency Capital Requirement (SCR) is defined as the economic capital to be held to ensure that ruin occurs over a one-year horizon with a probability under 0.5%. Mathematically, the SCR is defined as follows:

$$\mathbb{P}\left(\mathrm{BOF}_{t=1} < 0 \mid \mathrm{BOF}_{t=0} \geq \mathrm{SCR}\right) \leq 0.5\%$$

where $\mathbb{P}$ is the historical probability measure and BOF stands for the Basic Own Funds defined as the difference between the Asset value within the economical balance sheet and the Best Estimate liabilities (BEL). Note that the determination of the Basic Own Funds does not include the risk margin to avoid the problem of circularity that the introduction of this notion induces. Given the implicit nature of the definition given above, Bauer et al. [8] introduced a approximately equivalent notion of the SCR. In their paper, they define the SCR as the 99.5%-quantile of the one year loss function, evaluated at $t = 0$, which is of the form $\mathrm{BOF}_{t=0} - P(0,1)\mathrm{BOF}_{t=1}$ with $P(0,1)$ the discount factor. Namely, we have

$$\mathrm{SCR} = \arg\min_{u} \left\{ \mathbb{P}\left(\mathrm{BOF}_{t=0} - P(0,1)\mathrm{BOF}_{t=1} > u\right) \leq 0.5\% \right\}$$

or equivalently

$$\mathrm{SCR} = \mathrm{BOF}_{t=0} - P(0,1) q_{0.5\%} \left(\mathrm{BOF}_{t=1} \right). \tag{1.1}$$

If we denote by $\mathrm{BOF}_1(X_1, \ldots, X_d)$ the random variable representative of the economic capital at $t = 1$ of the life insurance company exposed to risk factors $(X_1, \ldots, X_d)$, then according to the equation (1.1), what we are looking for is to seek the distribution of BOF_1. However, the vast majority of life insurance liabilities can not be valued directly via closed formulas. To circumvent this constraint, risk-neutral simulations are carried out for each initial market condition considered, and an estimator of the value of the life insurance liabilities is then obtained. This methodology, commonly called "nested simulations" method, leads to directly calculating the empirical economic capital distribution (see Figure 1.1).

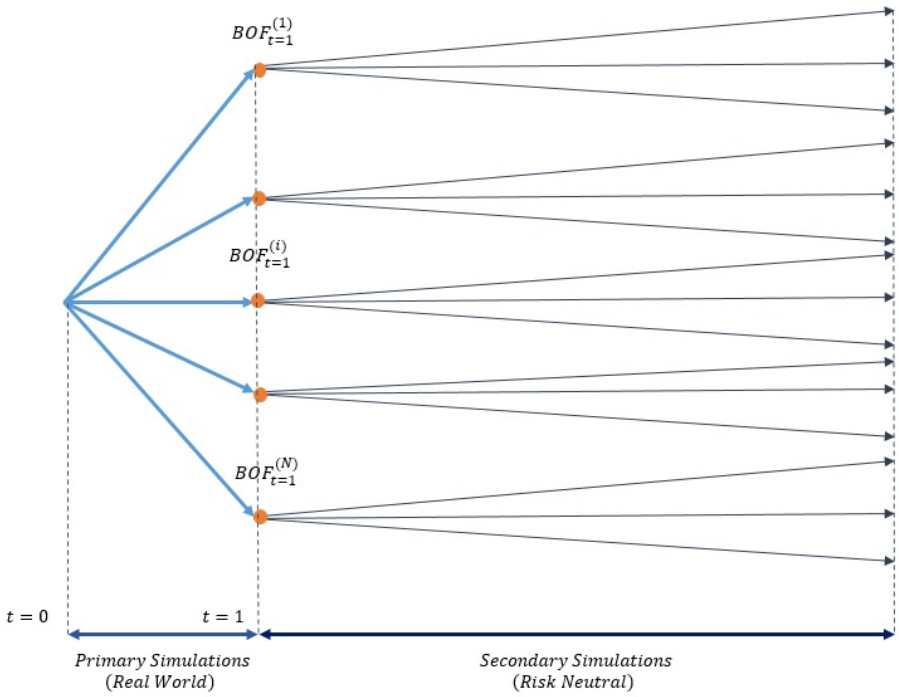


Figure 1.1: Illustration of the nested simulations method.

As a first step, a large number of real-world scenarios are generated. These scenarios are generated in a manner consistent with the distribution of risk factors to which the insurance company is exposed. These scenarios are usually called the *outer or primary scenarios*. In practice, at least 10000 outer scenarios are needed to ensure good stability of the empirical distribution, especially at its tail end.

For each of the outer scenarios considered, an estimate of the life insurance company's economic balance sheet is made through risk-neutral simulations (*inner or secondary scenarios*). At least one thousand risk-neutral scenarios are necessary to obtain a satisfactory estimate of the economic balance sheet. For particularly

extreme outer scenarios, the number of risk-neutral simulations to be carried out can be even greater.

In theory, this methodology is the one that achieves the most accurate economic own funds distribution. However, its implementation on a large scale still seems impossible today. Here we try to decompose very briefly the cycle of calculation of the economic own funds distribution.

- *Outer scenarios generation*: This step is by itself not particularly complex, as we seek in the solvency 2 regulatory framework to measure the variation in economic own funds for extreme quantiles, it is important to have sufficient outer scenarios to ensure good stability of the empirical SCR estimate.

- *Building economic balance sheets*: As previously explained, it is necessary to perform risk-neutral valuations for each outer scenarios. These valuation processes require the dissemination of risk-neutral scenarios based on the achievement of real-world risk factors. Depending on the complexity of the models used, this step may require many hours of calculation.

Today, the majority of ALM models used by insurance companies would be too slow to calculate the economic own funds distribution using this methodology. Put, end-to-end, this process does not seem possible nowadays. The constant improvement of diffusion models, projection models and the progress of information technology should enable us to improve these computing times in the future. But today we are still too far from the target to implement this methodology within life insurance companies.

In fact, we have only considered the production time so far, we must not neglect the time required to analyze these results. As it stands, life insurers encounter a production time problem which can be formulated as follows :

Problem 1: *The calculation of the Basic own Funds in life insurance is made particularly complex by many interactions between and liabilities, and is required a large number of simulations to obtain a satisfactory result as a consequence of the law of large numbers. The asset-liabilities interactions are related to the profit sharing mechanisms which are derived from both business objectives (client retention) and accounting rules. Profit sharing impacts then the liabilities through the changes in future services and the impacts it may have on policyholder behavior. The time necessary to compute a BOF corresponding to with-profit saving contracts can vary between about* 10 *min and* 1 *h depending on the computing power available and the complexity of the underlying cash-flows models. Assume that an insurer uses from* 10^4 *to* 2×10^5 *real world scenarios to derive the SCR and that the BOF is to be computed in* 10 *min, this would amount to minimum* 10^5 *min, that is* 70 *days or* 10 *weeks of computation time. As a result, the nested simulation or brute-force approach is unsuitable since it leads to significant computing times, while these processes must be implemented in a very short amount of time for reporting purposes. According to the Technical Practices Survey conducted by KPMG in 2015 [70], the majority of*

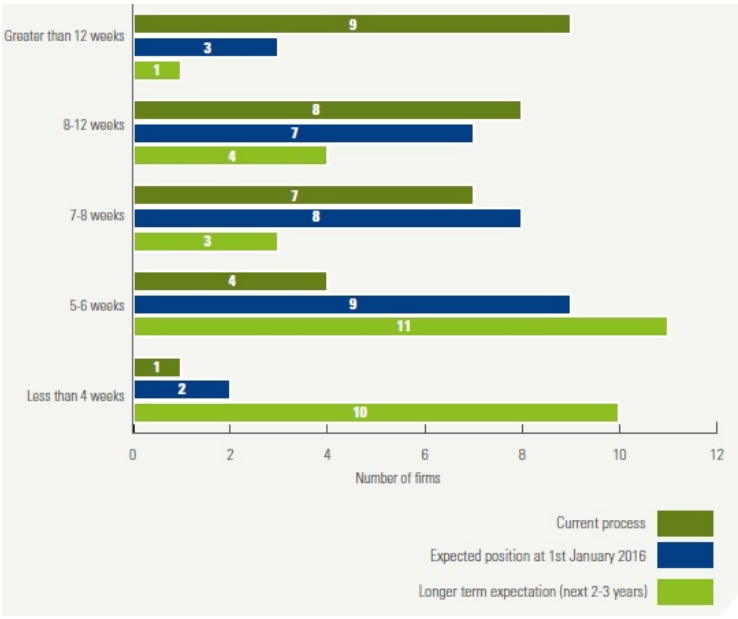


Figure 1.2: Expected production time for the Pillar 1 Balance sheet and the SCR to meet the demanding timescales.

insurers expects that the reasonable amount of time their Pillar 1 Balance Sheet and SCR take to produce for the annual processs is under 6 weeks (see Figure 1.2).

Therefore, life insurance companies have developed alternative methodologies (*proxy models*) that significantly reduce the time required to produce an economic own funds distribution. These methodologies consist in approximating the liabilities behavior in stressed conditions using closed form formulas such as loss function or financial instruments valuation formula. In general, these functions, which depend upon a certain number of coefficients, are calibrated on a limited number of simulations. Therefore, it is less expensive in computation time once calibrated.

The well-known and the most frequently used proxy approaches are the Curve-Fitting, Least Square Monte-Carlo ($LSMC$) and Replicating Portfolio methods. In the following, we will analyze each method in greater detail.

1.4 Proxy models in life insurance

In this section, we will present three major linear regression methodologies that are commonly used by insurers. These methodologies are very similar. It consists of using outputs from the projection models used by life insurance companies to derive functional forms that allow the rapid valuation of the life insurance company's economic balance sheet for any market condition. These methods differ mainly in the type of information that is used at the output of the projection models and in the functional forms used to establish the approximation.

A general description of these main approaches is summed up in Table 1.1:

	Polynomial functions		
	LSMC	Curve fitting	Replicating portfolios
Cover all risks	+++	+++	+
Accuracy	++	++	+
Objectivity	++	+	+
In line with market practice	+	++	++
Implementation time and costs	+	++	-
Less Business-as-usual effort required to perform runs	+++	+	+

Table 1.1: Benchmark on the principle existing methods to calculate the SCR

1.4.1 Curve-Fitting

This methodology consists of constructing the best parametric form (linear combination of analytic functions of real world risk factors) from a limited number of real world scenarios for which a very precise valuation was performed. This parametric curve thus passes through the calibration points (equality between the value of the parametric form and the value calculated within the projection model). To that end each interpolation points must be estimated with an extremely high precision, which demands many simulations to improve the convergence rate. Therefore, the disadvantages of this approach are that the data points must be carefully selected by expert judgements and the number of interpolation points is really limited by simulation time for each point [66].

To ensure that the estimator well replicate the economic own funds function, *out of sample scenarios* are used. The value of the parametric form is then compared with the value calculated within the projection model for these out of sample scenarios.

The criticisms relating to this methodology mainly concern its precision at the tail end. It is indeed necessary to ensure that there are enough extreme scenarios at the tail end to ensure that the effects of non-linearity at the level of the liabilities in this area is correctly anticipated.

1.4.2 Least Square Monte-Carlo

The Least Square Monte-Carlo (LSMC) method was introduced by Longstaff and Schwartz [78] in order to evaluate American Bermudan options. The difficulty of valuing these options lies in the calculation of a retrograde algorithm based on the evaluation at each iteration of a conditional expectation. To avoid the need for nested simulations to calculate the conditional expectation at each date, Longstaff and Schwartz use an approximation to calculate this conditional expectation.

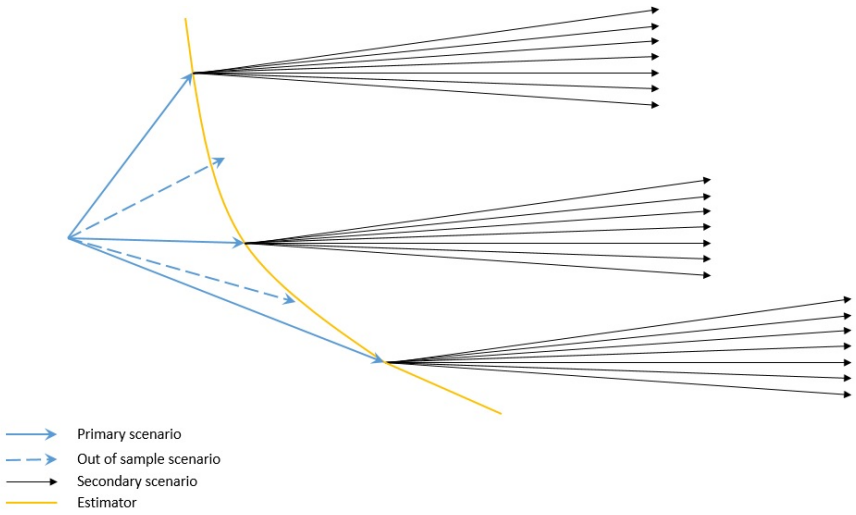


Figure 1.3: Illustration of the curve-fitting estimation method.

Bermudan options valuation

We place ourselves in a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, over a period of time $[0, T]$, where the sample space Ω is the set of possible realizations, $\mathcal{F}$ is the set of events containing the information available at each date t, $\mathcal{F} = \{\mathcal{F}_t; t \in [0, T]\}$ with $\mathcal{F}_s \subseteq \mathcal{F}_t$ for every $s \leq t$, $\mathbb{P}$ is the historical probability. It is assumed that there is no arbitrage opportunity on the market, this implies that there is a risk-neutral probability $\mathbb{Q}$ equivalent to $\mathbb{P}$.

American options can be exercised on specific dates between 0 and T, where T is the expiry date of the option, we denote these exercise dates $0 < t_1 \leq \cdots \leq t_N = T$ with $t_k = \frac{kT}{N}$, $\Delta_t = t_{k+1} - t_k$. The theoretical value of a Bermudan option under risk-neutral probability is given by

$$V_0 = \sup_{\tau \in \mathcal{T}} \mathbb{E}^{\mathbb{Q}} \left[Z_\tau e^{-r\tau} \right]$$

where τ^* is a stopping time taking the values in $\mathcal{T} = \{t_1, \ldots, t_N\}$ and Z_t is the payoff at time t.

To solve this problem, a dynamic programming method is used. The algorithm is written as follows:

$$\begin{cases} V_{t_N} = Z_T \\ V_{t_k} = \max \left(Z_{t_k}, \mathbb{E}^{\mathbb{Q}} \left[e^{-r\Delta_t} V_{t_{k+1}} \mid \mathcal{F}_{t_k} \right] \right) \end{cases}$$

We adopt the convention that there is no early exercise opportunity at time 0, hence $Z_0 = 0$. The terminal value Z_T is known and the algorithm is reiterated to determine V_0. The delicate step of this algorithm is the computation of the conditional expectation (called value function), Longstaff and Schwartz propose an approximation for this conditional expectation based on the least squares method.

We are interested here in derivative products whose payoffs are random variables belonging to the $L^2(\Omega, \mathcal{F}, \mathbb{Q})$ space, which is a Hilbert space. We know that the conditional expectation corresponds to the orthogonal projection on the Hilbert space and is then the unique solution of the following minimization problem

$$\mathbb{E}^{\mathbb{Q}}[X \mid \mathcal{F}] = \underset{Z \in L^2(\Omega,\mathcal{F},\mathbb{Q})}{\arg\min} \mathbb{E}^{\mathbb{Q}}\left[(X - Z)^2\right] \tag{1.2}$$

The calculation of the value function is based on this characterization of the conditional expectation. Let S be the underlying of the option, $\mathcal{F}_t$ is the filtration generated by S, which is a Markov process. Assume that we know how to simulate the trajectories of the underlying under $\mathbb{Q}$, so we have at each moment t_k the M simulated values $S_{t_k}^m$, $m = 1, \ldots, M$.

Denote $\mathbb{E}^{\mathbb{Q}}\left[e^{-r\Delta_t} V_{t_{k+1}} \mid \mathcal{F}_{t_k}\right] = \mathbb{E}^{\mathbb{Q}}\left[e^{-r\Delta_t} V_{t_{k+1}} \mid S_{t_k}\right] = f(S_{t_k})$. By considering an orthonormal basis of our Hilbert space, the condition expectation can then be approximated by a finite linear combination of this base that minimizes the criterion of conditional expectation (1.2). Given the basic functions $\{p_j\}, \forall j = 1, \ldots, L$, we then look for the coefficients $\alpha_{j,k}^*$, solution of the least-squares problem:

$$\alpha_{j,k}^* = \underset{\alpha_{j,k}}{\arg\min} \frac{1}{M} \sum_{m=1}^{M} \left(e^{-r\Delta_t} V_{t_{k+1}}^m - \sum_{j=1}^{L} \alpha_{j,k} p_j\left(S_{t_k}^m\right) \right)^2$$

By replacing the optimal coefficients in the linear combination of the basic functions, we obtain an approximation of the value function for each trajectory m and at each moment t_k:

$$\mathbb{E}^{\mathbb{Q}}\left[e^{-r\Delta_t} V_{t_{k+1}} \mid \mathcal{F}_{t_k}\right]^{(m)} \approx \sum_{j=1}^{L} \alpha_{j,k}^* p_j\left(S_{t_k}^m\right)$$

The following proposition provides a necessary and sufficient condition for the existence of an optimal stopping time and characterizes the smallest optimal stopping time.

Proposition 1. *There exists a stopping time $\tau^* \in \mathcal{T}$ such that $\mathbb{E}^{\mathbb{Q}}\left[Z_{\tau^*} e^{-r\tau^*}\right] = \sup_{\tau \in \mathcal{T}} \mathbb{E}^{\mathbb{Q}}\left[Z_\tau e^{-r\tau}\right]$ if and only if $\mathbb{Q}(\tau_0 < \infty) = 1$, where*

$$\tau_0 = \inf\{t \in \mathcal{T} \mid V_t = Z_t\}$$

The stopping time τ_0 is then the smallest optimal stopping time.

This corresponds to the corollary 1.3.2 in [74]. By comparing the value of the immediate exercise with the value function at each time step and for each trajectory, we are able to choose the optimal moment to exercise the option. We can now determine the option price by taking the average of the discounted cash flows on each trajectories. Noting τ_m^* the optimal stopping time corresponding to the trajectory m, we have:

$$\hat{V}_0 = \frac{1}{M} \sum_{m=1}^{M} e^{-r\tau_m^*} V_{\tau_m^*}^{(m)}.$$

1.4.2.1 Application in Life Insurance

The application in Insurance of the LSMC method is based on the fact that the own fund can be expressed as the expectation under the risk-neutral probability of the discounted value of future profits conditional on the projection of the balance sheet in the "real world" universe (see, for instance, [7, 67, 68, 105, 109]).

Indeed, basic own funds (BOF) at $t = 1$ can be expressed as follows:

$$\begin{aligned} \mathrm{BOF}_1 &= \mathrm{R}_1 + \mathrm{VIF}_1 \\ &= \mathrm{R}_1 + \mathbb{E}^{\mathbb{Q}}\left[\sum_{t=2}^{T} \mathrm{DF}(1,t) R_t \mid \mathcal{F}_1^{\text{real world}}\right] \end{aligned}$$

where $\mathrm{DF}(1,u)$ corresponds to the discount factor at the instant 1 for the time horizon u.

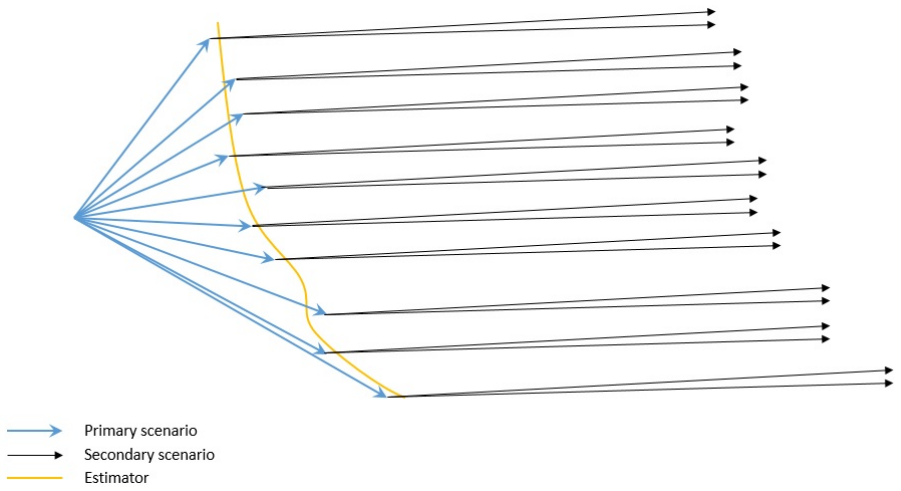


Figure 1.4: Illustration of the LSMC estimation method.

The LSMC method consists of expressing the unpredictability contained in the economic capital in a limited number of risk factors and then approximating conditional expectation by a linear combination of basic functions of these risk factors. The LSMC method is a method of reducing the number of simulations by using a large number of primary simulations and a few secondary simulations.

The goal is to apply the LSMC method to the calculation of conditional expectation: $\mathbb{E}^{\mathbb{Q}}\left[\sum_{t=1}^{T} \mathrm{DF}(1,t) R_t \mid \mathcal{F}_1^{\text{real world}}\right]$. For each primary simulation j, the empirical *net present value* (NPV) of the basic own funds is defined as the average of the sum of the discounted future results:

$$\mathrm{NPV}_1^{(k)} = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T} \mathrm{DF}^{(n)}(1,t) R_t^{(n)} \mid_{k\text{th primary scenario}}$$

Next, different stages are to be put in place:

- Assume that each primary scenario can be synthesized at time t, using d risk factors $(x_1, \ldots, x_d)$.
- Approximation of BOF_1 by a finite linear combination of basic functions $\{p_j\}_{j=1,\ldots,L}$ of these risk factors: $\widehat{\mathrm{BOF}}_1^{(k)} = \sum_{j=1}^{J} \beta_j p_j \left(x_1^{(k)}, \ldots, x_d^{(k)}\right)$.
- Calculation of the empirical $\mathrm{NPV}_1^{(k)}$ for each primary simulation k.
- Determination of optimal coefficients using the generalized least squares method:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^J}{\arg\min} \sum_{k=1}^{K} \left[\mathrm{NPV}_1^{(k)} - \sum_{j=1}^{J} \beta_j p_j \left(x_1^{(k)}, \ldots, x_d^{(k)}\right) \right]$$

- Calculation of $\widehat{\mathrm{BOF}}_1$ by replacing $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$.

The algorithm makes it possible to avoid the nested simulations since the conditional expectation is directly calibrated on the empirical NPV of the basic own funds via some secondary simulations.

1.4.3 Replicating Portfolios

A replicating portfolio of a set of liabilities is:

- a portfolio of standard financial instruments
- that has the same market consistent value as the liabilities, and
- that has similar market consistent value sensitivities to market risks drivers.

This proxy model builds a representation of the liabilities using vanilla financial instruments. This is a reasonably quick solution, relying mainly on the ability to represent exotic financial instruments (insurance) using only vanilla financial instruments. This representation is built on the projection system results. It is then combined with a line-by-line model of the assets to build a synthetic economic view of the market consistent balance sheet. The full range of initial market conditions can then be run in a very timely manner. This methodology has the advantage that it gives an understandable structure of the liabilities, which itself can

- provide insight into the business and into the financial risks
- help design hedging strategies
- help focus the calibration of the ESG to the most relevant financial instruments
- enable to challenge the results of the projection system.

If liabilities are independent from the backing assets and from the financial market, the liabilities cash flows are certain and each cash-flow can be represented by a zero-coupon bond of the same maturity and same amount. One scenario is sufficient to project all liabilities cash-flows and to determine the equivalent zero coupon bonds. However, the replicating portfolio can also be determined using the present value of cash flows under different scenarios. Therefore, for Property and Casualty (P&C) and life non-participating line of businesses, the cash flows can be perfectly replicated whatever the market conditions are. As a result, the replicating portfolio is as accurate as the projection system.

Even if liabilities do depend on the backing assets performance, for example through a profit sharing mechanism, it is possible to represent the liabilities using financial instruments. Let us consider as an example a traditional savings product with a guaranteed rate, and a guaranteed surrender value. The detail of the surrender modelling will be presented later. Here we briefly summarize the behaviour of the policyholders:

- If market rates increase above the policy guaranteed rate, the lapses will increase the policyholders will take advantage of the guaranteed surrender value (higher than the market value of the values) and re-invest at market rates; the resulting liability cash-flows can be represented by payer swaptions;
- If market rates decrease below the policy guaranteed rate, the lapses will decrease: the policyholders will take advantage of the guaranteed rate: the resulting cash-flows can be represented by receiver swaptions.

Not all policyholders will act as described-but at a portfolio level, it is possible to represent the liabilities as a combination of zero-coupon (base guarantee), receiver swaptions (lower lapses with lower rates), and payer swaptions (higher lapses at guaranteed value with higher rates). The strikes of the swaptions will depend on the lapse function.

However, it should be noted that it is often difficult to match complex liabilities well with replicating assets because the required instruments are not available in the market (see, for instance, [15,69,82,114]).Replicating portfolios only cover financial and credit (spread) risk and therefore polynomial loss functions are still needed for all other risks.

1.4.3.1 Determination of a replicating portfolio

This step is meant to find a set of financial instruments which achieves the best match of the market consistent values and sensitivities. In this section, we will however get into the details of this process since it goes beyond the scope of our objectives. In general, it will be an iterative process of:

1. Finding a candidate replicating portfolio,
2. Assessing its quality,

3. Repeating until predefined quality criteria are met.

The very first step is to define a list of candidate financial instruments. Insurance liability features will link to different types of financial instruments or to different characteristics of the financial instruments. It is therefore important to understand the features of the liability being replicated to retain relevant instruments. In most cases, the instruments will be a combination of: zero-coupons of different maturities representing the expected premiums to be received, claims to be paid, guarantees provided; receiver and payer swaptions of different maturities, tenor and strikes, representing the options given to policyholders; puts or calls on equities of diiferent maturities and strikes representing the profit-sharing given to policyholders.

The second step consists in finding the weights of all candidate instruments that will make the replicating portfolios closest to the liabilities. To this end we calculate the present values (PV) of the liabilities cash-flows resulting from a set of scenarios run through the projection system. We then define a distance on the present value of cash-flows vector space, which enables a direct resolution of the minimization to obtain the optimal weighting coefficients. Namely, denote by $\{\omega_k\}_{k=1}^K$ the weighting coefficients associate to K candidate instruments, we have

$$\omega_1^*, \dots, \omega_J^* = \underset{\omega_1,\dots,\omega_J}{\arg\min} \sum_{n=1}^{N} \left[\sum_{t=1}^{T} \left(\sum_{k=1}^{K} \omega_k \mathrm{CF}_{\mathrm{RP},k}^{(n)}(t) - \mathrm{CF}_{\mathrm{L},k}^{(n)}(t) \right) \mathrm{DF}^{(n)}(0,t) \right]^2 \quad (1.3)$$

In simple cases, representations of liabilities can be built using those financial instruments leading to high quality results for the calculation of the SCR.

1.4.4 Acceleration algorithm

Devineau and Loisel [30] develop an acceleration algorithm for the Nested Simulation method described previously. This algorithm aims to reduce the overall number of primary simulations to be carried out. The key idea of this method is to select the most adverse trajectories in terms of solvency according to the chosen risk factors and to do the simulations only along these adverse trajectories. To sum up, the acceleration algorithm is implemented in three key steps:

1. Extract the elementary risk factors that have the most impact on the items of the balance sheet for each primary simulation.

2. Define a fixed threshold confidence region: only primary simulations for which risk factors are outside the confidence region are performed.

3. Make iterations on the threshold of the region in order to integrate each step a number of additional points.

The basic idea behind this method is similar to the one of Lan et al. [72], who describe a screening procedure for expected shortfall based on nested simulations. The main advantage of this method is that it considerably reduces the calculation

times and the necessary resources. However, the speed of the algorithm rapidly decreases when the number of risk factors increases. Hence, only the risks having significant impacts on the portfolio are selected for the solvency assessment. This technique focuses on estimating economic capital and it is hard to apply to the portfolio risk management. Therefore, we observe that this method is rarely used in practice.

1.5 Error quantification for internal modeling in life insurance

Problem 2: *One may notice that none of these approaches was applied with proper control of the error implied, which is not robust in an insurance setting. Let us consider the curve fitting method for example. The error control formula is given by*[1]

$$|g(x) - \hat{g}(x)| \leq \frac{1}{2}\,(x_i - x_{i-1})^2 \,.\, \max_{x_{i-1}<x\leq x_i} |\frac{\partial^2 (g-\hat{g})}{\partial x^2}|.$$

In this formula, the risk measure depends on the second derivative of the target function which is in principle unknown. Therefore, a further estimation of this quantity is required which results in addition fitting error at this stage. Furthermore, the precision depends on the space between fitting points. This illustrates why the fitting points must be carefully selected by expert judgments. It is questionable whether applying these above approaches without proper fitting error controls will be consistent with Solvency 2 requirements for internal models. The current available information on this regulation-article 229(g) of the Commission Delegated Acts [93] indicated that using any approaches without including the estimation of the involved error would not be compliant with Solvency 2 requirements.

For many practical applications of the loss function, one usually relies upon a simpler notion of the SCR, which is approximately equivalent to (1.1). For this purpose, we define the SCR at $t = 0$ as the 99.5%-quantile of the loss function (see Chapter 3 for more details). This simplification will however generate a biased result with respect to the basic nested simulation estimator. Indeed, one of the most fundamental issues in the SCR calculation is the interplay between approximation error and estimation error. The basic nested simulation approach offers the most advantage compared to other approaches as it requires minimal assumptions on the structure of the risk model, which makes the approximation error small. However, for a life insurance company providing a complex organizational structure and portfolios where liabilities have options and guarantees, computational challenges make this approach impossible to achieve. This alternative proxy modeling technique may speed up the computation which usually leads to little estimation errors, but it

[1] This can be easily proven by setting $\delta(x) = g(x) - \hat{g}(x)$, we have $\delta(x_{i-1}) = \delta(x_i) = 0$. Using Rolle theorem and noting $M = \max_{x_{i-1}<x\leq x_i} \left|\delta^{"}(x)\right|$, we get on the segment $[x_{i-1}, x_i]$, the inequality $|\delta'(x)| \leq (x_i - x_{i-1})\, M$. The result is obtained by using integral of δ' and triangular inequality.

will generate approximation errors as we impose additional assumptions. When the number of risk drivers increases, these approximation errors may have a substantial impact on the estimated capital requirement. We will keep for further research to quantify the approximation errors in high dimensional settings. Finally, to ensure that these approximation errors in low dimensional settings are relatively small and acceptable, we prepare the box-whisker plot to see how good the approximation is.

1.6 Application of Extreme Value Theory to Solvency Capital Requirement estimation

One of the difficulties of SCR estimation is that one must evaluate quantities that depend on the tail of distribution, for which, almost by definition, one does not have observations or at least one has only very few observations. Recall that the simulation-based capital estimates are carried out as follows:

1. Generate real-world economic scenarios for all risk drivers affecting the balance sheet over one year,

2. Revalue the balance sheet under each real-world scenario (by using, for example, Monte Carlo (nested simulation), Replicating Portfolio, etc.),

3. Estimate the statistics of interest.

However, there exists many sources of uncertainty in this process. Namely, it depends on the choice of economic scenario generator models and their calibration, the liability model assumptions (e.g. dynamic lapse rules), as well as the choice of scenarios sampled (i.e. choice of real world ESG random number seed). Usually, an insurer will rely on expert judgement to define economic scenario generator models and liability model assumptions. Therefore, the first two sources of uncertainty are beyond the scope of our work and we are particularly interested in the last source of uncertainty, which is simply a statistical uncertainty. We wonder if we can estimate this statistical uncertainty. If so, how can we reduce the amount of statistical uncertainty? In our work, we will address these questions using a statistical technique known as Extreme Value Theory (EVT).

Recall that Extreme Value Theory tells us something about the shape of the distribution in the tail. The standard approaches for describing the extreme events of a stationary time series are the block maxima approach (which models the maxima of a set of blocks dividing the series) and the Peak-over-Threshold (POT) approach (which focuses on exceedances over a fixed high threshold). The POT method has the advantage of being more flexible in modeling data, because more data points are incorporated (see Chapter 4.1 for more details). Hence, the method we use in our study is the POT method.

According to this method, the distribution of liability value beyond some threshold is approximated as a Generalized Pareto distribution (GPD), which is parameterized by 2 parameters: scale σ and tail-index γ. Therefore, we can estimate the

tail of the distribution by picking a threshold and fitting the 2 parameters of the Generalized Pareto distribution to values in excess of the threshold.

In the context of financial and actuarial modeling, the observations very often depend on the other parameters, such as business line, risk profile, seniority, etc. However, all these studies assume that the tail-index is constant regardless of these variables. Many recent studies, for example [23,117], emphasized that the tail-index could be function of these explanatory variables. But none of the previously mentioned studies provide a way to estimate the tail-index parameter conditionally to these variables. As far as we can tell, in the context of financial and actuarial modeling, only three studies have been undertaken to provide methods to estimate the tail-index parameter conditionally to covariates. Beirlant and Goegebeur [9] propose a local polynomial estimator in the case of a one-dimensional covariate. When the dimension of the covariate increases, this method becomes less effective since the convergence rate of the estimator decreases rapidly. To improve the performance of the estimator, a solution would be to increase the size of data, but this would be problematic in practice since the database could not be easily enlarged. Then, Chavez-Demoulin et al. [22] propose an additive structure with spline smoothing to estimate the relationship between the GDP parameters and covariates. Recently, Heuchenne et al. [52] approach suggests a semi-parametric methodology to estimate the tail-index parameter of a GPD.

In practice, many financial and actuarial data modeling problem may depend upon several explanatory variables, which might make direct tail-index parameter estimation less accurate, or even impossible. However, it does not mean that all of these explanatory variables have more or less the same impact on the result. For example, Chernobai et al. [23] investigate the relation between frequency of operational loss events and firm-specific variables (market value of equity, firm age, cash holding ratio, etc.) as well as macroeconomic variables. They find a strong dependence between frequency and firm specific variables, but only weaker results with respect to the macroeconomic variables. This remark could also be true for the capital requirement. Therefore, it exists therefore a real need for companies to map, model and measure those risks to take proper hedging action. One technique to reduce dimension is sparse group lasso, which was introduced by Simon et al. [100]. Motivated both by the advances about the work of Chavez-Demoulin et al. [22] and the sparse group lasso method, we investigate a variable-selecting method to estimate the tail-index parameter conditionally to covariates.

1.7 Contributions and structure of the thesis

1.7.1 Contribution to the company

For PwC, the interest of sponsoring this PhD study is undeniable. Indeed, this project aims to perform a bibliographic research on recent scientific works concerning the actuarial finance, to know and understand some techniques used by insurers. and to be able to propose new commercial offers which can stand out competition. In

this sense, I think that this last objective was correctly achieved to the extent that my first research work that I have provided has made it possible to quantify and control the errors in the actuarial calculation engines.

Following the work carried out with clients in the context of either auditor's mandates or consultancy assignments, I have also established a benchmark of market practices concerning the implementation of Pillar I of the Solvency II Directive with respect to the saving contracts in euros (see Figure 1.5 in French).

Besides, I developed and put in place (i) a cross-asset Economic Scenarios Generator and (ii) an actuarial ALM simulator in life insurance. The ESG enabbles us to simulate future states of the global economy and financial markets. It uses advanced modeling and estimation technology to produce empirically validated, realistic economic scenarios which are used as inputs to the ALM simulator. These numerical tools result in numerous important contract wins for PwC. In the future, PwC would like to commercialize these numerical tools and present this work to clients. An overall introduction of these tools are given in Appendix A and B.

1.7.2 Methodological contributions

The works presented in this thesis attempt to bring a set of contributions to the performance of internal modeling in life insurance by applying advanced statistical techniques, while being easily implementable and numerically stable. In each of the simulation studies, We prove theoretical properties for the methods put in place, and we also show that these are relevant in practice and at least match the existing procedures. The results obtained allow us to consider different lines of research.

Error quantification for internal modeling in life insurance

In this work, I develop a new fitting methodology for estimating the SCR (Problem 1) and a formula for controlling the deviation of the target SCR from its estimate (Problem 2). The new method operates in the following way.

We proposed to calculate the SCR as the 99.5%-quantile of the loss function (see Section 3.3.2 for the definition of the loss function), i.e.

$$\mathrm{SCR} = \mathrm{q}_{99.5\%}(\phi) \tag{1.4}$$

The loss function $\phi(x_1, \ldots, x_d)$ is then decomposed into the stand-alone loss functions $\{\phi_j(x_j)\}_{j=1,\ldots,d}$ and the excess loss function $\phi_{1d}(x_1, \ldots, x_d)$ as follows:

$$\phi(x_1, \ldots, x_d) = \sum_{j=1}^{d} \phi_j(x_j) + \phi_{1d}(x_1, \ldots, x_d). \tag{1.5}$$

Next we apply the Bayesian penalized spline regression technique to estimate each functional component. For later use, we denote by $\hat{\phi}$ the estimate of ϕ.

The SCR can be estimated by $\widehat{SCR} = \hat{q}_{99.5\%}\left(\hat{\phi}\right)$ its empirical 99.5th-percentile derived from $\hat{\phi}$. In this stage, $\hat{\phi} \equiv \hat{\phi}(X)$ is a random variable with $X = (X_1, \ldots, X_d)$

Informations générales

Benchmark S2 Pilier I	Assureur 1	Assureur 2	Assureur 3	Assureur 4	Assureur 5	Assureur 6	Assureur 7	Assureur 8	Assureur 9
1 Activité	Epargne/ Retraite/ Prévoyance	Epargne/ Retraite/ Prévoyance/ Emprunteur	Epargne/ Retraite/ Prévoyance	Epargne	Epargne	Epargne/ Retraite/ Prévoyance	Epargne/ Retraite/ Prévoyance	Epargne/ Retraite/ Prévoyance	Epargne/ Retraite/ Prévoyance
2 Taille de bilan (en milliards d'euros)	> 50	> 50	> 50	< 15	< 15	< 15	> 15 < 50	> 50	> 15 < 50
3 Approche Solvabilité II	Formule Standard	Formule Standard	Formule Standard	Formule Standard	Formule Standard	Formule Standard	Formule Standard	Formule Standard	Formule Standard
4 Ratio Solvabilité II	>200%	>120% <200%	>120% <200%	>120% <200%	>200%	>120% <200%	>120% <200%	>120% <200%	>120% <200%

Hypothèses non économiques

Benchmark S2 Pilier I	Assureur 1	Assureur 2	Assureur 3	Assureur 4	Assureur 5	Assureur 6	Assureur 7	Assureur 8	Assureur 9
1 TMG comme maille des lois de rachats structurels	✗	✗	✓	✗	✗	✓	✗	✓	✓
2 Justification de la loi de rachats dynamiques	✗	✓	✗	✗	✗	✗	✗	✗	✓
3 Taux 10 ans pris en compte pour les rachats dynamiques	✓	✓	✓	✓	N/C	✓	✓	✓	✗
4 Hypothèses de mortalité abattues (sinon réglementaires)	✓	✓	✓	✓	✗	✓	✓	✓	✓

Figure 1.5: Benchmark Pillar 1 of Solvency II (Certain information are confidential, and thus will not be mentioned in this table).

Hypothèses économiques

Benchmark S2 Pilier I		Assureur 1	Assureur 2	Assureur 3	Assureur 4	Assureur 5	Assureur 6	Assureur 7	Assureur 8	Assureur 9
1	*ESG générant des taux négatifs*	✓	✗	✓	✗	✓	✓	✓	*N/C*	✗
2	*Modélisation des spreads*	✓	✓	✗	✗	✗	✗	✗	*N/C*	✗
3	*Prise en compte du contexte risque neutre dans le calibrage des taux de dividendes et des loyers*	✗	✗	✗	✓	*N/C*	✗	*N/C*	*N/C*	✗
4	*Distinction des OPCVM qui distribuent et de ceux qui capitalisent dans le modèle*	✗	*N/C*	✗	✓	✓	✓	*N/C*	*N/C*	✗

Modélisation

Benchmark S2 Pilier I		Assureur 1	Assureur 2	Assureur 3	Assureur 4	Assureur 5	Assureur 6	Assureur 7	Assureur 8	Assureur 9
1	*Modèle de flexing*	✗	✓	✓	✗	✗	✓	✓	✓	✗
2	*Nombre d'année de projection*	*60*	*50*	*30*	*30*	*N/C*	*40*	*40*	*40*	*40*
3	*Allocation d'actifs (cible en valeur de marché)*	✓	✓	✓	✗	✓	✓	✓	✓	✗

Benchmark S2 Pilier I		Assureur 1	Assureur 2	Assureur 3	Assureur 4	Assureur 5	Assureur 6	Assureur 7	Assureur 8	Assureur 9
4	*Taux cibles*	*Taux 10 ans avant prise en compte de la marge financière*	*Taux attendu par les assurés*	*N/C*	*Fonction des taux N-1 et de taux de marché*	*N/C*	*Fourchette définie en fonction des taux N-1 et du TME*	*Fonction des taux N-1 et de taux de marché*	*90% de la moyenne sur 2 ans des OAT 10 ans*	*Fonction des taux N-1 et de taux de marché*
5	*Non prise en compte de la PB min*	✗	✓	✓	✓	✓	✓	✗	*N/C*	✓
6	*PPE: Modélisation de la règle des 8 ans*	✗	✗	✗	✓	✗	✗	*N/C*	✗	✗

Fin de projection

Benchmark S2 Pilier I		Assureur 1	Assureur 2	Assureur 3	Assureur 4	Assureur 5	Assureur 6	Assureur 7	Assureur 8	Assureur 9
1	*Répartition des fuites de modèles de manière défavorable pour l'assureur*	✗	✓	✓	✓	✓	✓	✗	✓	✓
2	*Liquidation des PMVL de fin de projection: - 85% en cas de PVL - 0% en cas de MVL*	✗	*N/C*	✓	✓	✗	✓	✓	✓	✗
3	*Réserve de capitalisation de fin de projection affectée à l'assureur*	✓	✓	✓	✓	✓	✓	✓	✓	✓

the realistic random market state or the primary simulation state whose marginal distribution is $\mathbb{P}_X$. Let $f_{\hat{\phi}}$ denote the density function of $\hat{\phi}(X)$.

To control the probability of deviation of the target SCR from its estimate, we will need certain conditions to make the theory work. First of all, it is important to clarify that as will be seen below, the resulting confidence band will not incorporate the approximation error from the choice of the regression function.

Let us introduce some notation, definitions that will be used in the sequel. We define the (L, Ω)-Lipschitz class of functions, denoted $\Sigma(L, \Omega)$, as the set of function $g : \Omega \to \mathbb{R}$ satisfy, for any $x, x' \in \mathbb{R}^d$, the inequality:

$$|g(x') - g(x)| \leq L\|x' - x\|$$

with $\Omega \subset \mathbb{R}^d$ and $\|x\| \triangleq (x_1^2 + \cdots + x_d^2)^{1/2}$. Let $r > 0$. We define $B(a, r) = \{x \in \mathbb{R}^d \mid \|a - x\| \leq r\}$. We denote by $\bar{V}_\phi = \{x \in \mathbb{R}^d \mid \phi(x) = q_{99.5\%}(\phi)\}$ and $\bar{V}_{\hat{\phi}} = \{x \in \mathbb{R}^d \mid \hat{\phi}(x) = q_{99.5\%}(\hat{\phi})\}$ the closed set of the 99.5th-percentile scenarios for ϕ and $\hat{\phi}$ respectively.

Let Γ denote the available sampling budget used to calibrate $\hat{\phi}$. Based on the work of Aerts et al. ([2]), it is straightforward to deduce that for $\lambda_{\phi_j}(\Gamma)$ and $\lambda_{h_J}(\Gamma)$ tending to 0, the estimate $\hat{\phi}$ converges in mean square to ϕ as $\Gamma \to \infty$. Furthermore, by Markov's inequality, convergence in mean square of $\hat{\phi}$ leads to the convergence in probability of $\hat{\phi}(x)$ to $\phi(x)$ for every $x \in \mathbb{R}^d$. This implies that for every $x^* \in \bar{V}_\phi$, there exists a random sequence $x^*_{(\Gamma)} \in \bar{V}_{\hat{\phi}}$ converges in probability to x^*.

Introduce now three assumptions on ϕ, $\hat{\phi}$ and $x^*_{(\Gamma)}$ that will be used in the last step:

ASSUMPTION 1: Suppose that $\phi \in \Sigma(L, \Omega)$ where $L > 0$ and $\Omega(\supset \bar{V}_\phi)$ is an open subset of $\mathbb{R}^d$.

ASSUMPTION 2: For any $x^* \in \bar{V}_\phi$ and $r > 0$, there exists two positive constants $\xi(r, d)$, $\gamma(r, d)$ such that

$$\mathbb{P}\left(\|x^* - x^*_{(\Gamma)}\| > r\right) \leq \xi(r, d)\Gamma^{-\gamma(r,d)}$$

for large enough Γ.

ASSUMPTION 3: For any choice of $x^* \in \bar{V}_\phi$ and $\alpha \in (0, 1)$, there exists two positive constants $r(\Gamma)$ and $\Delta(\alpha, \Gamma)$, with $r(\Gamma) \xrightarrow{\Gamma \to \infty} 0$, such that

$$\mathbb{P}\left(\mid \hat{\phi}(x) - \phi(x) \mid > \Delta(\alpha, \Gamma)\right) \leq 1 - (1 - \alpha)^{\frac{d(d+3)}{2}}, \quad \forall x \in B(x^*, r(\Gamma))$$

for large enough Γ.

- SCR estimation error control: In the following, we denote by N_1 the number of the primary simulations. Note that

$$\left|\widehat{SCR} - SCR\right| \leq \left|\hat{q}_{99.5\%}\left(\hat{\phi}\right) - q_{99.5\%}\left(\hat{\phi}\right)\right| + \left|q_{99.5\%}\left(\hat{\phi}\right) - q_{99.5\%}\left(\phi\right)\right| \quad (1.6)$$

  The first term on the right-hand side corresponds to the numerical error since we appeal the empirical percentile to estimate the SCR and the second term

represents the model error. Note that the numerical error depends not only on the empirical assessment $\hat{q}_{99.5}$ but also on the fitting quality $\hat{\phi}$. To value this numerical error, we apply the Theorem in Appendix D.8. Namely, we have

$$\mathbb{P}\left(\left|\hat{q}_{99.5\%}\left(\hat{\phi}\right) - q_{99.5\%}\left(\hat{\phi}\right)\right| > z_{\alpha/2}\frac{0.07}{\sqrt{N_1}f_{\hat{\phi}}(q_{99.5\%}\left(\hat{\phi}\right))}\right) \to \alpha \tag{1.7}$$

as $N_1 \to \infty$. In the previous expression, the distribution function $f_{\hat{\phi}}$ and the evaluated point $q_{99.5\%}\left(\hat{\phi}\right)$ are however unknown and will be then replaced by their estimators. Regarding the second term, by using Assumptions (1-3), we obtain the asymptotic probability of deviation of $q_{99.5\%}\left(\hat{\phi}\right)$ from $q_{99.5\%}\left(\phi\right)$ having the form:

$$\mathbb{P}\left(\left|q_{99.5\%}\left(\widehat{\phi}\right) - q_{99.5\%}\left(\phi\right)\right| > \Delta(\alpha,\Gamma) + Lr^*\right) \leq \left[1-(1-\alpha)^{\frac{d(d+3)}{2}}\right] + \xi(r^*,d)\Gamma^{-\gamma(r^*,d)} \tag{1.8}$$

where $r^* \equiv r(\Gamma)$. The derivation of this result can be found in Appendix D.6. Combing the equations (1.7) and (1.8) leads to the control of the probability of deviation of $\widehat{SCR}$ from SCR.

The confidence interval $\Delta(\alpha,\Gamma)+Lr^*$ is however an issue as it involves the unknown parameters $\Delta(\alpha,\Gamma)$, L and r^*. In the following, we suggest a method to estimate these parameters in practice.

In order to estimate the Lipschitz constant, we find the supremum of all slopes $|\hat{\phi}(x)-\hat{\phi}(x')|/\|x-x\|$ for distinct points x and x' within the 99.5th-percentile region. We call $\hat{x}^*$ the empirical 99.5th-percentile scenario, i.e. $\hat{\phi}(\hat{x}^*) = \hat{q}_{99.5\%}\left(\hat{\phi}\right)$. The parameter $\Delta(\alpha,\Gamma)$ will be then replaced by $\tilde{\Delta}(\alpha,\Gamma) = \sum_{j=1}^{d}\Delta_{j,\alpha}^{(\hat{x}^*)} + \sum_J \tilde{\Delta}_{J,\alpha}^{(\hat{x}^*)}$. To estimate the parameter r^*, we seek the maximum radius $\hat{r}^*$ such that for every $x^{(\nu)} \in B(\hat{x}^*,\hat{r}^*)$, the confidence intervals $\sum_{j=1}^{d}\Delta_{j,\alpha}^{(\nu)} + \sum_J \tilde{\Delta}_{J,\alpha}^{(\nu)}$ are close to $\tilde{\Delta}(\alpha,\Gamma)$. On the right-hand side of the inequality (1.8), as the true value of $\xi(r^*,d)$ and $\gamma(r^*,d)$ are unknown, it is not possible to have a direct access to the upper bound of the probability. In practice, a large number of Γ is necessary so that the term $\left[1-(1-\alpha)^{d(d+3)/2}\right]$ becomes preponderant compared to $\xi(r^*,d)\Gamma^{-\gamma(r^*,d)}$.

For many practical applications of the loss function, one usually relies upon a simpler notion of the SCR, which is approximately equivalent to Eq. 1.1. This simplification will however generate a biased result with respect to the basic nested simulation estimator. Indeed, one of the most fundamental issues in the SCR calculation is the interplay between approximation error and estimation error. The basic nested simulation approach offers the most advantage compared to other approaches as it requires minimal assumptions on the structure of the risk model, which makes the approximation error null. However, for a life insurance company providing a complex organizational structure and portfolios where liabilities have options and guarantees, computational challenges make this approach impossible to

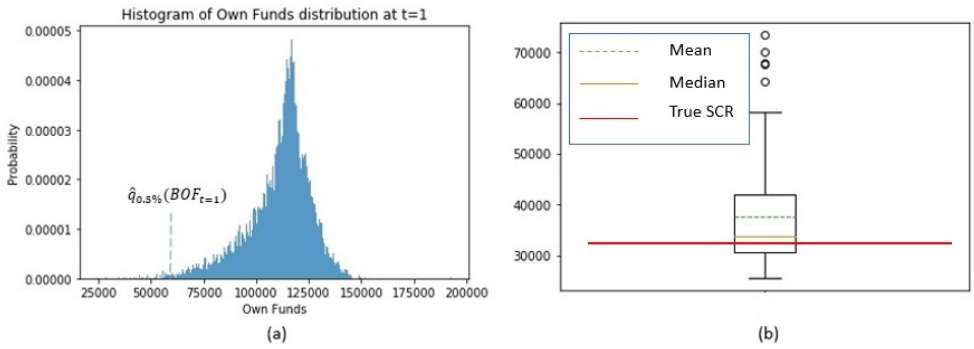


Figure 1.6: (a) Histogram of Own Funds distribution at $t = 1$. (b) Box-whisker plot of the $\widehat{SCR}$ estimated with 100 different samples of the same size.

achieve. This alternative proxy modeling technique may speed up the computation which usually leads to little estimation errors, but it will generate approximation errors as we impose additional assumptions. When the number of risk drivers increases, these approximation errors may have a substantial impact on the estimated capital requirement. We will keep for further research to quantify the approximation errors in high dimensional settings. Finally, to ensure that these approximation errors in low dimensional settings are relatively small and acceptable, we prepare the box-whisker plot and compare with the SCR estimated by the nested simulation method to see how good the approximation is.

In Figure 1.6.a, we plot the economic Own Funds distribution at $t = 1$. From this, we derive the empirical estimation of $\hat{q}_{0.5\%}(BOF_{t=1}) = 54690.16$ and the empirical SCR estimated by the Nested Simulations method is thus equal to $\widehat{SCR}_{NS} = 33743.83$. In another simulation (see Figure 1.6.b) of 100 different samples of the same size from the same two distributions of the equity risk and the interest rate level risk, we observe that the outcomes are skewed and the estimated values of SCR distribute close to the "true" SCR. All the details as well as the numerical studies can be found in Chapter 5.

Application of Extreme Value Theory to Solvency Capital Requirement estimation

Inspired by the Peaks-over-threshold method, all observations that exceed a specified high threshold $u_n(\mathbf{x})$ are used to estimate $\gamma^*(\mathbf{x})$ with $\mathbf{x} = (x^{(1)}, \ldots, x^{(p)})$. According to this approach the Generalized Parato Distribution (GPD) defined by

$$G(z; \gamma, \sigma) = 1 - \left(1 + \gamma \frac{z}{\sigma}\right)^{-\frac{1}{\gamma}}, \quad \forall z \geq 0, \quad \gamma, \sigma > 0$$

is fitted to the exceedances over a specific thresholds. Let us call $\left(\gamma^*_{u_n(\mathbf{x})}(\mathbf{x}), \sigma^*_{u_n(\mathbf{x})}(\mathbf{x})\right)$ the corresponding fitted GPD parameters. One usually encounters the curse of dimensionality problem, which leads to the rapid diminution

in convergence rate, when the covariate is high dimensional. To overcome this difficulty, we assume that $\gamma^*_{u_n(\mathbf{x})}(\mathbf{x})$ and $\sigma^*_{u_n(\mathbf{x})}(\mathbf{x})$ are approximated by a generalized additive model as follows

$$\gamma^{p,\infty}(\mathbf{x}) = \exp\left(\gamma_0 + \sum_{j=1}^{p} \gamma_j(x^{(j)})\right)$$
$$\sigma^{p,\infty}(\mathbf{x}) = \exp\left(\sigma_0 + \sum_{j=1}^{p} \sigma_j(x^{(j)})\right)$$

where each additive function $\gamma_j(\cdot), \sigma_j(\cdot)$ belongs to the Sobolev space of continuously differentiable functions. In order to ensure the identification we assume that for every $j = 1, \ldots, p$ the additive functions γ_j, σ_j are centered, i.e.

$$\sum_{i=1}^{n} \gamma_j\left(x_i^{(j)}\right) = 0, \quad \sum_{i=1}^{n} \sigma_j\left(x_i^{(j)}\right) = 0. \tag{1.9}$$

These statistical models are still nonparametric and the estimation therefore a problem of infinite dimension. We make it finite by expanding each additive functional components in natural cubic spline (NCS) bases with a reasonable amount of knots K_j for $j = 1, ..., p$. Thus, we parametrize

$$\gamma_j(\cdot) = \sum_{k=2}^{K_j} \theta_{j,k}\left(h_{j,k}(\cdot) - \frac{1}{n}\sum_{i=1}^{n} h_{j,k}\left(x_i^{(j)}\right)\right), \quad \sigma_j(\cdot) = \sum_{k=2}^{K_j} \theta'_{j,k}\left(h_{j,k}(\cdot) - \frac{1}{n}\sum_{i=1}^{n} h_{j,k}\left(x_i^{(j)}\right)\right)$$

where $h_{j,k} : \mathbb{R} \to \mathbb{R}^+$ is the natural cubic spline basis function constructed on the set of the predefined interior knots $\{\xi_1^{(j)}, \ldots, \xi_{K_j}^{(j)}\}$ satisfying $\xi_1^{(j)} \leq \cdots \leq \xi_{K_j}^{(j)}$. Clearly, this parametrization of the functional components $(\gamma_j(\cdot), \sigma_j(\cdot))$ verifies the centering conditions given in (1.9). To simplify our notation, let us define

$$\tilde{h}_{j,k}(\cdot) = \left(h_{j,k}(\cdot) - \frac{1}{n}\sum_{i=1}^{n} h_{j,k}\left(x_i^{(j)}\right)\right), \ \forall j = 1, \ldots, p, \ \forall k = 1, \ldots, K_j.$$

In the following, we denote by β_0 and θ_0 the intercept term instead of γ_0 and σ_0 to synchronize the notation with the coefficients $\theta_{j,k}, \theta'_{j,k}$ as presented previously. Finally, our statistical model is defined as

$$\gamma(\mathbf{x}) = \exp\left(\theta_0 + \sum_{j=1}^{p}\sum_{k=2}^{K_j} \theta_{j,k}\tilde{h}_{j,k}\left(x^{(j)}\right)\right)$$
$$\sigma(\mathbf{x}) = \exp\left(\theta'_0 + \sum_{j=1}^{p}\sum_{k=2}^{K_j} \theta'_{j,k}\tilde{h}_{j,k}\left(x^{(j)}\right)\right)$$

To sum up, the following diagram sets out the whole approximation scheme: Next, we denote by $\boldsymbol{\varphi} = \left(\theta_0, \boldsymbol{\theta}^T, \theta'_0, \boldsymbol{\theta}'^{,T}\right)$ the entire parameter vector where

$$\{\gamma^*(\mathbf{x})\} \xrightarrow{POT} \{\gamma^*_{u_n(\mathbf{x})}(\mathbf{x}), \sigma^*_{u_n(\mathbf{x})}(\mathbf{x})\} \xrightarrow[Approximation]{Additive} \{\gamma^{p,\infty}(\mathbf{x}), \sigma^{p,\infty}(\mathbf{x})\} \xrightarrow[Approximation]{NCS} \{\gamma(\mathbf{x}), \sigma(\mathbf{x})\}$$

$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T\right)^T$, $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}_1^{'T}, \ldots, \boldsymbol{\theta}_p^{'T}\right)^T$ with $\boldsymbol{\theta}_j = \left(\theta_{j,2}, \ldots, \theta_{j,K_j}\right)^T$ and $\boldsymbol{\theta}'_j = \left(\boldsymbol{\theta}'_{j,2}, \ldots, \boldsymbol{\theta}'_{j,K_j}\right)^T$ for every $j = 1, \ldots, p$. Clearly, the parameter vector $\boldsymbol{\varphi}$ can be structured into groups $\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_p$ and $\tilde{\mathcal{G}}_0, \tilde{\mathcal{G}}_1, \ldots, \tilde{\mathcal{G}}_p$. Each of the groups is defined in the following way:

$$\theta_0 = \boldsymbol{\varphi}_{\mathcal{G}_0}, \;\; \boldsymbol{\theta}_j = \boldsymbol{\varphi}_{\mathcal{G}_j}, \;\; \theta'_0 = \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_0}, \;\; \boldsymbol{\theta}'_j = \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}, \qquad \forall j = 1, \ldots, p.$$

Under this notation, our models can be rewritten as

$$\gamma(\mathbf{x}|\boldsymbol{\varphi}) = \exp\left(\sum_{j=0}^{p} \boldsymbol{\varphi}_{\mathcal{G}_j} \tilde{h}_{\mathcal{G}_j}\left(x^{(j)}\right)\right)$$
$$\sigma(\mathbf{x}|\boldsymbol{\varphi}) = \exp\left(\sum_{j=0}^{p} \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j} \tilde{h}_{\tilde{\mathcal{G}}_j}\left(x^{(j)}\right)\right)$$

with $\tilde{h}_{\mathcal{G}_0}(\cdot) = \tilde{h}_{\tilde{\mathcal{G}}_0}(\cdot) = 1$.

For the purpose of variable selection and eliminating perturbative effects within each group, we suggest to use the sparse group lasso technique to estimate $(\gamma(\mathbf{x}|\boldsymbol{\varphi}), \sigma(\mathbf{x}|\boldsymbol{\varphi}))$. Namely, the regression model used to estimate $(\gamma(\mathbf{x}|\boldsymbol{\varphi}), \sigma(\mathbf{x}|\boldsymbol{\varphi}))$ is defined by

$$\widehat{\boldsymbol{\varphi}}(u_n(\cdot), \boldsymbol{\lambda}, \boldsymbol{\mu}) = \arg\min_{\boldsymbol{\varphi}} \left\{P_n l(\boldsymbol{\varphi}|u_n(\cdot)) + \text{pen}\left(\boldsymbol{\varphi}|\boldsymbol{\lambda}, \boldsymbol{\mu}\right)\right\}. \tag{1.10}$$

where

$$P_n l(\boldsymbol{\varphi}|u_n(\cdot)) = -\frac{1}{n}\sum_{i=1}^{n} \log g\left(y_i - u_n(\mathbf{x}_i); \gamma(\mathbf{x}_i|\boldsymbol{\varphi}), \sigma(\mathbf{x}_i|\boldsymbol{\varphi})\right) \mathbb{I}(y_i \geq u_n(\mathbf{x}_i)).$$

with y_i a realisation of Y_i and the penalty

$$\text{pen}\left(\boldsymbol{\varphi}|\boldsymbol{\lambda}, \boldsymbol{\mu}\right) = \lambda_1 \sum_{j=1}^{p} \sqrt{G_j} \|\boldsymbol{\varphi}_{\mathcal{G}_j}\|_2 + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\varphi}_{\mathcal{G}_j}\|_1 + \mu_1 \sum_{j=1}^{p} \sqrt{G_j} \|\boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}\|_2 + \mu_2 \sum_{j=1}^{p} \|\boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}\|_1$$

with $G_j \equiv |\mathcal{G}_j| = |\tilde{\mathcal{G}}_j|$ the cardinality of the group $\mathcal{G}_j$, as well as of the group $\tilde{\mathcal{G}}_j$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T \in \mathbb{R}^2_{*,+}$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T \in \mathbb{R}^2_{*,+}$. The algorithm used to solve the equation (1.10) is summarized in Algorithm 1.

A well-known drawback of l_1-penalized estimators is the systematic shrinkage of the large coefficients towards zero. This may give rise to a high bias in the resulting

estimators and may affect the overall conclusion about the model. We then need to refit the model without any penalties on the select support

$$\mathcal{S}_{\mathcal{G}} = \{(j,k)|\widehat{\varphi}_{\mathcal{G}_{j,k}} \neq \mathbf{0}\}, \quad \mathcal{S}_{\tilde{\mathcal{G}}} = \{(j,k)|\widehat{\varphi}_{\tilde{\mathcal{G}}_{j,k}} \neq \mathbf{0}\}.$$

Finally, we perform a numerical study with different settings (i.e. $p = 2, 10$ and $n = 500, 5000$) and compare the estimating performance of our methodology with an existing method proposed by Beirlant and Goegebeur [9]. Usually in many high-dimensional studies, the dimension of the data vectors p is comparable or may be larger than the sample size n. Hence, it is obvious that our setting with $p = 10$ can not be considered as high dimensional covariate. However, we realized that it becomes computationally expensive in terms of running time required to perform estimation when the dimensionality increases. Therefore, in this thesis, we limit ourselves to the case $p = 10$. Surprisingly, we note that the proposed methodology slightly outperforms even with $p = 10$. And we hope in the near future that we can reinforce our results with higher dimensionality.

1.7.3 Structure of the thesis

As can be seen, this thesis, which is constituted of two parts, is organized in seven chapters:

- The first part deals with the error quantification problem for internal modeling in life insurance. It consists of four different chapters. Chap. 2, 3 and 4 are introductory chapters. Chap. 2 and 3 present respectively our Economic Scenario Generator (ESG) and Asset-Liability Management (ALM) cash-flows simulator, which are the main tools used to value the economic balance sheet. Chap. 4 is a general presentation: the statistical framework and the nonparametric estimation methods are introduced. All these chapters will provide us fundamental elements to achieve our findings presented in Chap. 5.

- The second part deals with the application of Extreme Value Theory to Solvency Capital Requirement estimation when the covariate information is available. Especially, when the covariate are high dimensional, we face with the curse of dimensionality problem resulting in a decrease in fastest achievable rates of convergence of regression function estimators toward their target curve. This problem refers to the phenomenon where the volume of covariate space increases so fast that the available data become sparse. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality, which is usually problematic in many practical applications. To overcome this estimating problem, we propose a new methodology for effectiveness evaluation, which is described in Chap 4.

Publication

Duong, Q.D., *Application of Bayesian penalized spline regression for internal modeling in life insurance.* European Actuarial Journal 9, 67–107(2019).

Submitted paper

Duong, Q.D., Guilloux, A. and Lopez, O., *Sparse group lasso additive modeling for Pareto-type distributions*. Submitted to Computational Statistics journal.

Chapter 2

Solvency II - Interpreting the key principles of Pillar I

Contents

2.1 History of capital requirements in the European insurance industry

2.1.1 Solvency I directive

The first European regulations on minimum capital to be held date back to the 1970s. In 1973 and 1979, two directives was published; one in the non-life insurance sector[1] and one in the life insurance sector[2]. These impose for the first time European insurers to build a layer of security in terms of own funds. In February 2002 the Solvency I directives were adopted. Recall that these directives had remained broadly close to the first European regulations.

The model developed under Solvency I to assess the solvency capital requirement is simple. According to Solvency I, the risk is either in provisions or in premiums. The calculation of the capital required is a so-called "factor-based" approach, which means that the required capital is calculated as a fraction of the elements considered

[1] First Council Directive 73/239/EEC of 24 July 1973 on the coordination of laws, regulations and administrative provisions relating to the taking-up and pursuit of the business of direct insurance other than life assurance

[2] First Council Directive 79/267/EEC of 5 March 1979 on the coordination of laws, regulations and administrative provisions relating to the taking up and pursuit of the business of direct life assurance

as risky on the balance sheet (technical provisions) or on the profit and loss account (premiums).

2.1.2 From Solvency I to Solvency II

Solvency I has the merit of being simple and can therefore be implemented at a lower cost. In addition, the regulations allow a quick comparison of the results obtained for different companies. The approach is nevertheless not adequate for several reasons, which will be discussed later. It thus justified the initiation of the new reform, called Solvency II project, hereinafter Solvency II for short.

Firstly, the level of technical provisions or the premium amounts are not in themselves good indicators of risk, for several reasons:

1. The approach does not take into account the level of prudence of the insurer in its provisioning. For example, a prudent insurer, better endowed with technical provisions, must mobilize more capital than an insurer with less provision. Such a system therefore penalizes prudential.

2. The approach highlighted in Solvency I is based only on the liabilities balance sheet of insurance companies, while other risks should be considered, such as asset risks, i.e. market and credit risks. In addition, the solvency capital requirements do not take into account, for example, the investment structure of the insurance company.

3. The risk reduction methods are also ignored: diversification between risks, risk transfer, asset-liability management, risk hedging instruments. The use of financial derivatives products, the use of reinsurance, the credit quality of re-insurers, etc., should also influence the required solvency margin.

Secondly, the assets and liabilities are valued at historical cost (or book value). However, this valuation method does not reflect the risks and the real value of the assets and liabilities. Finally, the Solvency I regime can lead to systemic risks. In fact, by way of illustration, a compulsory pricing framework for all insurance companies exposes all these companies to the same risks of errors in tariffs. To sum up, Solvency I does not adequately reflect the risk profile of each insurance companies concerned. These weaknesses justified the need for a regulatory reform.

The lessons learned from the years 2002 and 2003, during which the financial markets experienced a period of crisis, while at the same time putting the financial health of some insurance companies at a disadvantage, led the regulators to take a review of the risk valuation framework within the insurance industry. Since March 2003, the European Commission, in collaboration with the member states, had been working on developing a single reference system aimed at better integrating risk into the constraints imposed on insurers in order to ensure their ability to fulfill their commitments. This is the Solvency II Project.

2.2 Implementation of Solvency II

As one of the most crucial projects currently being carried out by the Commission and the member states in the insurance sector, Solvency II consists in developing a novel, better risk-adjusted system for assessing the overall solvency of insurance companies. Namely, this system provides the supervisory authorities with appropriate quantitative and qualitative instruments for assessing the overall solvency of insurance companies.

Solvency II has two main objectives. The first one is to create a single, competitive and open market on a European scale. The second one is to further protect insureds and counterparties. The first objective stems from the standardization of prudential constraints within each European member country. Harmonization of regulation removes the inequalities of regulatory benchmarks and allows the construction of a single and free market. The second objective is supported by the idea that an insurer must better manage, know and evaluate its risks.

It is based on a three pillar structure such as the Basel II project, Solvency emploies a risk-based approach, which encourages insurers to better measure them. This is a transition from an implicit vision of risk, that of Solvency I, to an explicit vision that integrates all risk managements gains, thus remedying the limits of the standard methods by which a flat-rate solvency margin is required and a restriction on investment in the safe, liquid, diversified and profitable assets. Each of the three pillars is synthesized in the following figure.

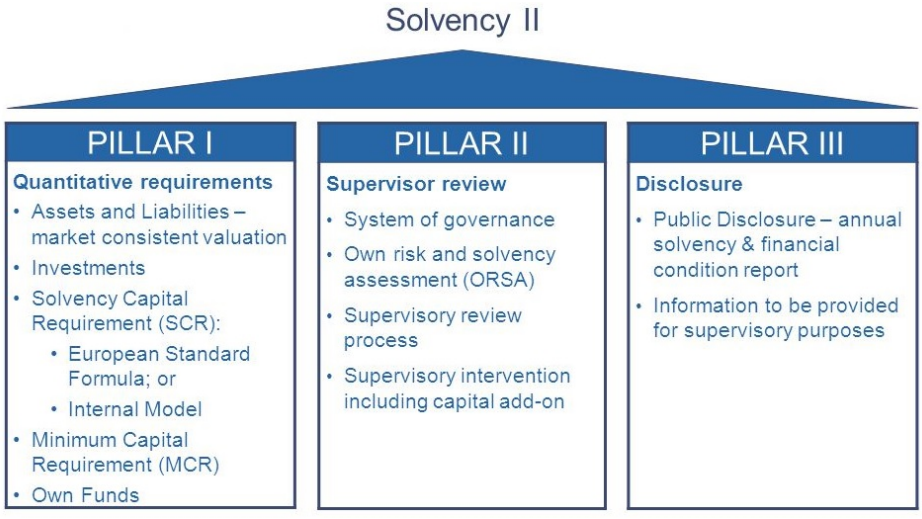


Figure 2.1: The structure of Solvency II

Solvency II aims at setting two requirements on the economic own funds or economic capital, a desirable level and a minimum level of capital. The former one must allow the company to operate with a very low probability of ruin by taking into account all the risks to which the insurance company is exposed. While the

latter one is an element of intervention of last resort, it is a minimum of capital requirement.

In fact, the technical measures of Solvency II were developed in 2004 by EIOPA in two phases. A first phase of reflection on the general principles and a second phase of detailed development of the methods of taking into account the different risks. In order to carry out this project, quantitative impact studies have been established by EIOPA to assess the applicability, consistency, comparability and implications of possible approaches for measuring the solvency of insurers. From this perspective, quantitative impact studies allow quantitative and qualitative feedback which is gathered from market participants to harmonize the management of insurance risks at European level.

However, Solvency II also gives insurance companies the possibility of adopting an partial or total internal model allowing an adequate modelling of the various risks and having an economic balance sheet over one year horizon illustrating the level of Solvency capital which is based on the notion of the distribution of own funds within this horizon. In life insurance, commitments duration are however much longer than one year. Moreover, if we take into account the assets-liabilities interaction as well as the complexity of their dependencies resulting from the profit sharing mechanisms, the interest rate guarantee, the possibility of early repayment and the buy-back behavior of the policyholders, obtaining an economic capital distribution at one year will be a delicate task.

In the following of this chapiter, we will present in more detail the quantitative requirements of the directive, namely Pillar 1. The other two pillars will not be detailed and we refer the reader to the European directive voted on 22 April 2009.

2.3 Pillar I

Pillar 1 of Solvency 2 characterizes the quantitative requirements of the Directive. These quantitative requirements are more complex than those described by Solvency 1 since they are intended to reflect an assessment of capital requirements using an economic approach.

As shown in Figure 2.2, the pillar sets out rules for the following six topics:

- Assets and Liabilities market consistent valuation.
- Investments.
- Technical Provision.
- Solvency Capital Requirement.
- Minimum Capital Requirement.
- Own Funds.

We briefly present the six topics mentioned above in order to understand the demanding regulatory context in which the work presented in this document is.

2.3.1 The quantitative requirements of Pillar 1

2.3.1.1 Assets and Liabilities market consistent valuation

Under Solvency I, liabilities are measured using conservative assumptions and assets are valued at historical cost. For example, the price of a security held in the portfolio will be recorded at its acquisition price. Thus, at each balance sheet date, the gross book value of the security does not change. Under Solvency II, assets and liabilities of insurance and reinsurance undertakings must be valued at their economic value, known as *fair value* or *market consistent value*. As explained in Kemp [64], a market consistent value of an asset or a liability, in case the liability is traded in a liquid market, is simply its market price. In this case, the market consistent valuation for a liability means that the stochastic liability cash-flows are perfectly replicated by a portfolio of liquid and deeply traded financial instruments. In absence of arbitrage, a market consistent value of a liability is thus defined as the expectation under the risk neutral measure of future liability cash-flows discounted by the value of the money market account conditional upon the economic and actuarial information available at the valuation time. Furthermore, assume that the market is complete, this market consistent value is unique thanks to the second fundamental theorem of asset pricing (see, for example, [111]).

The definition of a complete market with non arbitrage opportunities and a risk-neutral probability can be found in [111]. Here, we would like to make it clear the following notations. A *risk-less* portfolio means a portfolio with totally predictable payoff. For example, if we invest 1 euro in a risk-less bank account, then this 1 euro capitalized in the bank becomes e^{rt} euros at time t where we call r the *risk-free* rate of interest. Under the *risk-neutral* probability, the return on assets is equal to the risk-free rate r. In the risk-neutral world, an investor will ignore the risk when making decision to invest in something. This is completely different with respect to a *risk-averse* investor who prefers lower returns with known risks rather than higher returns with unknown risks. In other words, among various investments giving the same return with different level of risks, this investor always prefers the alternative with least interest. Therefore, the risk-neutral probability increases the objective probability of adverse events for the investor to take into account his risk aversion. From this point of view, the use of a risk-neutral valuation can be considered as a prudent valuation. This evaluation makes it possible to construct an economic balance sheet which will be presented in Chapter B.

2.3.1.2 Technical Provisions

Technical provisions break down into *Best Estimate* and *Risk Margin*:

The Best Estimate is defined as the probable present value of the future cash flows without any margin of caution. In other words, the Best Estimate is the discounted and probabilized sum of benefits and future costs backed by the insurer's commitments. It should be noted that Best Estimate must be based on credible current information and realistic assumptions. Note that it must be calculated in

run-off, i.e. new business is not considered in cash flows, only flows associated with current contracts are taken into account.

There are different methods to evaluate the Best Estimate. In non-life asurrance, the deterministic methods (Chain-Ladder, Borhuetter Ferguson method, ...) suffice because the assets have no influence on the insurer's liabilities and commitments. Conversely, in life insurance, assets and liabilities interact continuously. For example, the revaluation of policy liabilities (liabilities) will depend on the return rate on the asset. Therefore, it is necessary to use an Asset Liability Management (ALM) tool to capture all the interactions between assets and liabilities.

The Best Estimate varies according to the behavior of policyholders in the future (redemptions, deaths), but also according to the actions that the management will take (profit-sharing strategy, asset allocation, etc.). The modeling of the behavior of the insured as well as the management rule is therefore an important stake for a life insurer.

In this thesis, we will set up an ALM model to calculate the Best Estimate of an abstract life insurance company. The ALM model projects the company's activity over time through asset assumptions (economic scenarios, financial instrument modelling, ...) and liabilities assumptions (death, redemptions, ...). The functioning of the ALM model will be explained in the modeling part.

The *Risk Margin* is the additional amount required in relation to Best Estimate so that the liabilities can be transferred to another insurer. In other words, when an insurance company takes over the contracts of another company, it must raise the necessary capital to meet the new commitments and requirements (SCR). The risk margin is therefore interpreted as the capital cost of these assets.

2.3.1.3 Own Funds

In addition to the technical provisions which are calculated on the fair value principle, the European Commission specifies that the Own Funds must be valued at their economic value. Furthermore, the European Commission distinguishes the Own Funds into *Solvency Margin* and *Surplus*. A solvency margin is constituted so that the insurance company has a very low probability of going bankrupt within 1 year.

The capital requirement is set up at two levels:

- Minimum Capital Requirement (MCR): corresponds to the capital required to cover a probability of ruin from 10% to 20%. If the own funds is lower than this required level, the Prudential Supervisory Authority (ACP) intervenes and can implement a restructuring plan or withdraw the company's approval.

- Solvency Capital Requirement (SCR): corresponds to the capital required by an insurer to absorb unforeseen losses (extremely worst case scenario out of 200) and gives insureds certainty that benefits will be paid with a probability of 99.5% within 1 year. When the SCR is respected, the probability of ruin

of the insurance company is 0.5%. To calculate it, there are two possibilities: the standard model and an internal model.

The Directive considers the assumptions on which the SCR calculations must be based. Its calculation is based on the assumption of continuity of operation of the insurance company. In addition, the SCR must be calibrated in such a way that all quantifiable risks to which the insurance or reinsurance undertaking are taken into account. It is also specified that the SCR must cover at least the following risks:

- The subscription risk in non-life
- The subscription risk in life
- The risk of underwriting in health
- The market risk
- The credit risk: default of counterparties
- The operational risk (excluding reputation risks et strategic decision-making risks)
- The risk of intangible assets

Solvency 2 requires that the SCR has to be calculated at least once a year and notified to the supervisory authorities. However, the SCR must be continuously monitored by the insurance and reinsurance companies. Therefore, if the company's risk profile differs significantly from the last assumptions underlying the calculation, SCR must be re-evaluated without delay and its result must be notified to the supervisory authorities.

The Directive proposes two methods for calculating SCR, the choice of which is left to the company's discretion: the standard formula or the internal model. If the internal model is chosen by the company, a second calculation of the SCR by the standard formula will nevertheless be obligatory for 2 years. In addition, the internal model must be approved by the regulator.

2.3.2 Standard Formula

The standard formula is a simplified means proposed by the Solvency II Directive for the evaluation of the SCR. The global required solvency capital is calculated by aggregating specific risk-specific marginal SCRs. Thus, the standard formula is broken down into several modules and sub-modules classified as one of the first six risks mentioned above, to which are added the intangible risk, counterparty default risk and an adjustment. The adjustment proposed by CEOIPS takes into account the insurer's ability to absorb future losses via profit-sharing mechanism with the insured or via different taxes.

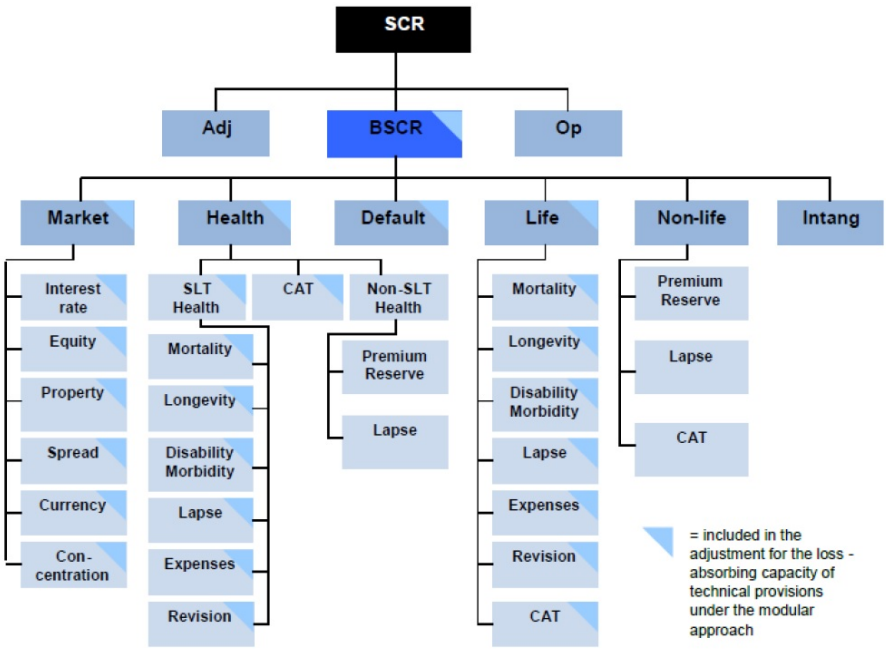


Figure 2.2: Overall structure of the SCR according to the standard formula [33].

For each module, the technical specifications of the QIS 5 [90] propose a calculation method which can be an analytical formula, a deterministic method or an estimation by simulations. Note that in the last two approaches, the standard formula need to use a cash-flow projection model. Thus, in the context of life insurance, the calculation of a SCR by the standard formula requires an ALM model.

The overall SCR can be deduced in successive steps. Each of the SCRs must first be computed for all submodules and then aggregated by correlation matrix to determine a modular SCR. All the modular SCRs are then aggregated by correlation matrix to form the *Basic Solvency Capital Requirement* or *BSCR* for short. Namely, we have

$$\text{BSCR} = \sqrt{\sum_{i,j} \text{Corr}(i,j) \times \text{SCR}_i \times \text{SCR}_j} + \text{SCR}_{intangible}$$

Finally, the adjustment denoted by Adj aiming at including in the SCR calculation the capacity to absord losses from technical provisions and deferred taxes, and operational SCR denoted by SCR_{op} are calculated separately without aggregation. Note that the value of the adjustment depends in particular on the profit-sharing mechanisms.

The global SCR is given by:

$$\text{SCR} = \text{BSCR} + \text{SCR}_{op} + \text{Adj}$$

For example, in the Standard Formula, the SCR_{equity} is determined by variation of *Net Asset Values* (NAV), floored at zero, as a result of the application of shocks.

Here, the NAV is defined as the difference between the assets market value and the Best Estimate. Note that the determination of the Net Asset Value does not include the risk margin of the technical provisions in order to avoid the problem of circularity that the introduction of this notion induces. Mathematically, we have $SCR_{Equity} = \max\left(\text{NAV}_{BE} - \text{NAV}_{shock}; 0\right)$

Schematically, the calculation of the marginal SCR by variation of NAV can be presented in Figures (2.3) and (2.4).

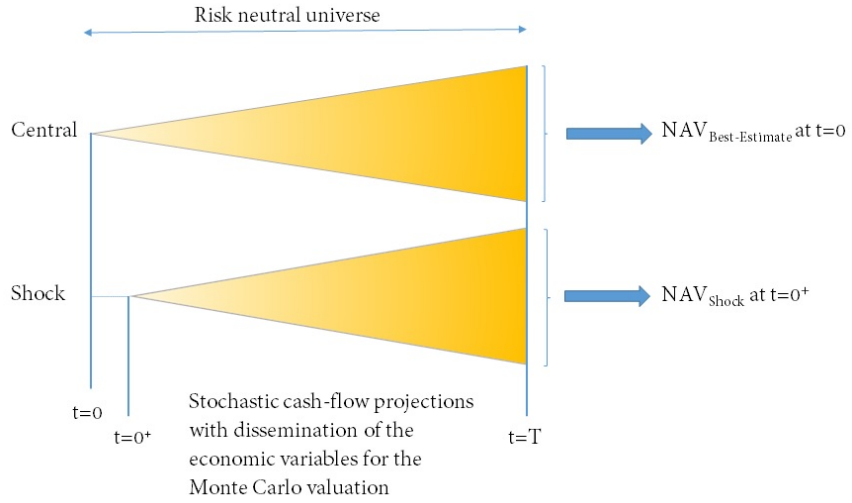


Figure 2.3: A graphical illustration of the SCR calculation by ΔNAV approach

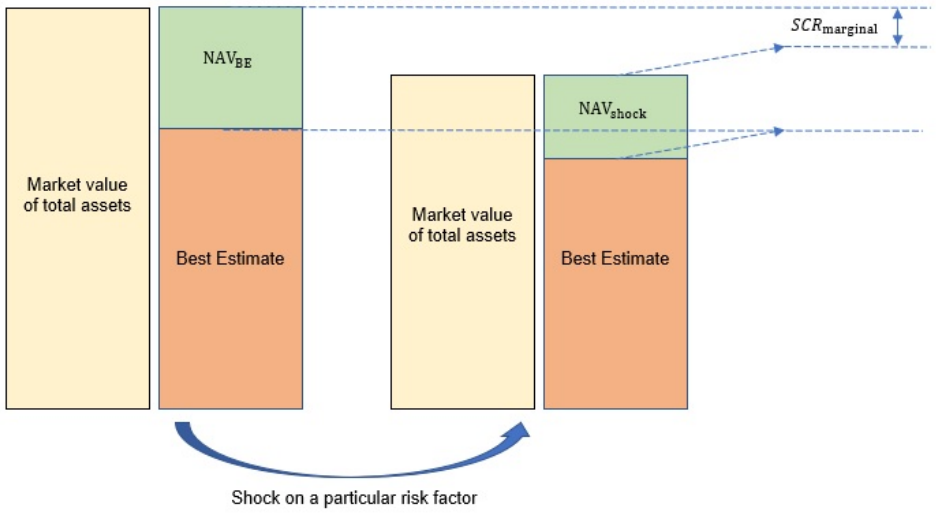


Figure 2.4: A graphical illustration of the SCR calculation by ΔNAV approach

2.3.3 Internal Model

The internal model is a model specific to the insurance or reinsurance undertaking subject to the approval of the supervisory authorities. It may be total or partial [3]. The definition of the internal model given by the CEA in Solvency II Glossary [104] is as follows:

"*Risk management system of an insurer for the analysis of the overall risk situation of the insurance undertaking, to quantify risks and/or to determine the capital requirement on the basis of the company specific risk profile.*"

The idea underlying the internal model is to carry out a customized modeling of the insurer's portfolio. As a result, the SCR and MCR are based on the underlying risks actually borne by the insurer and no longer on the standard basis of the standard formula as described above.

The advantages of an internal model, based on its own economic assumptions, are now widely shared within the insurance industry:

1. Organization interest: It helps an insurance company study and control its underlying risks;
2. Operational interest: It improves the risk management of an insurance company;
3. Competitive interest: It plays like a communication tool intended for the financial community and rating agencies.

[3] There are many risk factors which may affect an insurance firm's economic balance sheet: economic risks (interest rate risk, equity risk, credit risk, implied volatility, etc.) and non-economic risks (lapse risk, mortality risk, etc.). A partial internal model is a specific model which only target a limited number of risk factors

Chapter 3

Application of Bayesian penalized spline regression for internal modeling in life insurance

Contents

In this chapter, we develop a new fitting methodology for estimating the SCR (see Problem 1 in Chapter 1) and a formula for controlling the deviation of the target SCR from its estimate (see Problem 2 in Chapter 1). The new method operates in the following way. The loss function will be decomposed into standalone loss functions and excess loss function. Then we apply the Bayesian penalized spline regression technique to estimate each functional component. But let us first recall some basic notions of nonparametric statistics and the reason why we come up with additive models and Bayesian penalized spline regression for internal modeling.

The structure of this chapter is organized as follows: First, an overarching introduction to nonparametric regression is give in Section 3. Then in Section 3.3, we recall the loss function notion, which is largely used in life insurance internal

models. We will also impose some assumptions about the functional form of the excess loss function as well as the computation formula for SCR. In Section 3.4, we describe our statistical model as well as the mathematical framework of the method. In Section 3.5, we carry out the numerical study for illustration and elaborate the consistency of the proposed method by comparing with the LSMC method.

Part I-Overarching introduction to nonparametric regression

A classical way to estimate a regression function is to assume that the structure of the function is known, dependent on certain parameters, and is included in a finite-dimensional function space. This is the parametric approach, in which the data are used to estimate the unknown values of these parameters.

In the parametric context, estimators generally depend on few parameters, so these models are well defined even for small samples. They are easily interpretable, for example, in the linear case, the values of the coefficients indicate the influence of the explanatory variable on the response variable, and their sign describes the nature of this influence. However, a linear estimator will lead to a significant inaccuracy regardless of the size of the sample whether the true function that generated the data is not linear and can not be approached appropriately by linear functions.

The non-parametric approach does not however require a pre-determined structure of the regression function. The functional relationship between the explanatory variables and the response variable is built from the data. This flexibility makes it possible to capture the unusual or unexpected traits. However, the complexity of the estimation problem arises another issue.

In this chapter, we will briefly present different well-known nonparametric regression techniques in the scientific literature. This material is for the most part borrowed from [48,97,112]. Some of the results mentioned here will be then applied in Chapter 5.

3.1 Univariate nonparametric regression

There are several methods for obtaining a non-parametric estimator of the function f satisfying:

$$Y = f(X) + \varepsilon \tag{3.1}$$

where Y is a ramdom variable, X can be a deterministic or random variable, ε is a random variable independent of the predictor variable X such that $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$. Now, let $\{(X_i, Y_i)\}_{i=1,\dots,n}$ be the identically and independently distributed samples of (X, Y) and $\{(x_i, y_i)\}_{i=1,\dots,n}$ be its realisations. The idea of smoothing technique is to estimate $f(x_i)$ by a weighted average of $\{(y_i)\}_{i=1,\dots,n}$ in the neighborhood of x_i, that is $\hat{f}(x_i) = \sum_{k=1}^{n} W_{ik} y_k$, where $\hat{f}(x_i)$ stands for the estimation of f at x_i. The weights W_{ik} are high when $\mid x_i - x_k \mid$ is small, or x_k is

close to x_i. Otherwise, the former one approaches zero when $\mid x_i - x_k \mid$ becomes high. Among the smoothing methods, the linear smoothing technique is a particular one. A method is called linear if the weights W_{ik} depend only on $\{x_i\}_{i=1,\ldots,n}$, but not on $\{y_i\}_{i=1,\ldots,k}$. Let denote $\hat{\mathbf{f}} = (\hat{f}(x_1), \ldots, \hat{f}(x_n))^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$. From this, one has a linear relation between $\hat{\mathbf{f}}$ and $\mathbf{y}$ which is of the form $\hat{\mathbf{f}} = \mathbf{S}.\mathbf{y}$ where $\mathbf{S} = \{S_{ik}\}_{i,k=1,\ldots,n}$ is the smoothing matrix and is independent of $\mathbf{y}$. In the following of this section, we will recall the two conventional linear smoothing methods: regression by kernel functions and by spline functions.

3.1.1 Kernel smoothing method

Kernel smoothing methods are intuitive and simple from the mathematical viewpoint. These techniques use a set of local weights, defined by the kernel functions, to construct the estimator in each value. In general, the kernel function K is a continuous , bounded, non-negative and symmetric function such that:

$$\int_{\text{supp}(K)} K(x)dx = 1, \quad \int_{\text{supp}(K)} x^2 K(x)dx < +\infty$$

Here are some kernel functions which are widely used in practice:

- Gaussian: $K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$
- Epanechnikov: $K(x) = \frac{3}{4}\left(1 - x^2\right) \mathbb{I}_{|x|\leq 1}$
- Quartic: $K(x) = \frac{15}{16}\left(1 - x^2\right)^2 \mathbb{I}_{|x|\leq 1}$
- Cosine: $K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2}x\right) \mathbb{I}_{|x|\leq 1}$

The kernel estimators we can tell are the Nadaraya-Watson estimator, the Gassero-Müller estimator and local polynomial regression estimator. For the later use, we denote $K_h(x, x')$ by $\frac{1}{h}K\left(\frac{x-x'}{h}\right)$ where the parameter h is called "bandwidth". We will investigate the role of this parameter at the end of this section.

Nadaraya-Watson Estimator

The Nadaraya-Watson estimator is defined as:

$$\hat{f}_{NW}(x) = \frac{\sum_{i=1}^n y_i K_h(x, x_i)}{\sum_{i=1}^n K_h(x, x_i)} \tag{3.2}$$

It is easily noted that the idea of the Nadaraya-Watson regression consists in partitioning the set of values of X and then performing a weighted average of the values of Y in each subinterval constructed as central neighborhoods at each point x. From this, we can easily derive the smoothing matrix $\mathbf{S}$ whose elements are of the form $S_{ij} = \frac{K_h(x_i, x_j)}{\sum_{k=1}^n K_h(x_i, x_k)}$. This smoothing matrix has an eigenvalue 1 and an eigenvector $\mathbf{1}_n = (1, \ldots, 1)^T$. Therefore, the Nadaraya-Watson estimator preserves the constant functions since $\mathbf{S}\mathbf{1}_n = \mathbf{1}_n$.

Local Polynomial Regression Estimator

Local polynomial regression estimator is a generalisation of the Nadaraya-Watson estimator. Indeed, the Nadaraya-Watson estimator is the unique minimizer of the following optimization problem:

$$\hat{f}_{NW}(x) = \text{argmin}_{a\in\mathbb{R}} \sum_{i=1}^{n} (y_i - a)^2 K_h(x, x_i)$$

If $f(x)$ is p-times differentiable in a neighborhood of x, then the Taylor developpement can be applied:

$$\begin{aligned} f(x') &\approx f(x) + f'(x)(x' - x) + \cdots + \frac{f^{(p)}(x)}{p!}(x' - x)^p \\ &\approx \beta_0 + \beta_1(x' - x) + \cdots + \beta_p(x' - x)^p \end{aligned}$$

where $\beta_k = \frac{f^{(k)}(x)}{k!}$.

From the previous remarks, we can therefore consider the local polynomial regression problem in a neighborhood of x at follows. The regression function is estimated at each point by locally adjusting a polynomial of degree p by weighted least squares. The weighting at the point $x_i, i = 1, \ldots, n$ is chosen as a function of the amplitude of the kernel function centered at this point. The estimator of the regression function at each point x is the local polynomial which minimizes.

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1(x_i - x) - \cdots - \beta_p(x_i - x)^p)^2 K_h(x_i, x) \tag{3.3}$$

We denote $\mathbf{W}(x) = \text{diag}(K_h(x_1, x), \ldots, K_h(x_n, x))$, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ and

$$\mathbf{X} = \begin{bmatrix} 1 & (x_1 - x) & \ldots & (x_1 - x)^p \\ \vdots & & & \vdots \\ 1 & (x_n - x) & \ldots & (x_n - x)^p \end{bmatrix}$$

The problem(3.3) can be then reformulate as

$$\min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{W} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.4}$$

Therefore, the vector $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)^T$ minimizing the equation (3.4) is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{W}\mathbf{y} \tag{3.5}$$

The explicit expression of the estimator $\hat{f}(x)$ is then given by

$$\hat{f}(x) = \mathbf{e}_1^T \hat{\boldsymbol{\beta}} = \mathbf{s}^T(x)\mathbf{y}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ and $\mathbf{s}^T(x) = \mathbf{e}_1^T \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{W}$.

The general form of the smoothing matrix, for any p, is written as

$$\mathbf{S} = \begin{bmatrix} \mathbf{s}^T(x_1) \\ \vdots \\ \mathbf{s}^T(x_n) \end{bmatrix}$$

In particular, for $p = 0$, we refind the Nadaraya-Watson smoothing matrix as defined above.

According to the equation (3.4), it is not difficult to note that the observations close to x have more influence on the estimator at point x than those that are distant from it. In fact, this relative influence is controlled by the bandwidth parameter h. In case of small bandwidth, the local fit is strongly dependent on observations close to x. This gives rise to a very fluctuating curve which tends to interpolate the data. Otherwise, the weights given to near and distant observations tend to be equal. This gives rise to a curve obtained by the usual global least square regression. In other words, the choice of a small h corresponding to a large variance leads to an *undersmoothing*. Alternatively, with a large h we cannot control the bias, which leads to *oversmoothing*. Therefore, there exists an optimal value of h which balances the trade-off bias and variance.

3.1.2 Spline regression

The idea of the spline regression consists in constructing smoothly joining polynomials. The points of connection between the pieces of polynomials are called the knots. To represent splines, for a fixed nondecreasing set of knots, $\{\kappa_j\} j = 1, \ldots, K$, one has to determinate a basis. For example, the basis of truncated polynomial of degree p evaluated at x is defined as

$$\{b_j(x)\}_{j=1}^{K+p+1} = \{1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p\}$$

where $(\cdot)_+$ indicate the positive part function. From this, the representation of a function $f(x)$ within this basis is given by $f(x) = \sum_{j=1}^{K+p+1} \beta_j b_j(x)$. The coefficients β_j are determined by minimizing the quadratic error term, i.e.

$$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_{K+p+1}\right) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{K+p+1}}{\arg\min} \| \mathbf{B}.\boldsymbol{\beta} - \mathbf{y} \|^2$$

where

$$\mathbf{B} = \begin{bmatrix} 1 & x_1 & \ldots & x_1^p & (x_1-\kappa_1)_+^p & \ldots & (x_1-\kappa_K)_+^p \\ \vdots & & & & & & \vdots \\ 1 & x_n & \ldots & x_n^p & (x_n-\kappa_1)_+^P & \ldots & (x_n-\kappa_K)_+^p \end{bmatrix}$$

A particular case of the truncated polynomial spline is the natural cubic spline: the piecewise polynomials of degree 3 which are constrained to have continuous second order derivatives on the knots and are linear beyond the domain defined by

these knots. The natural condition of linearity on the edges implies the following expression of the natural basis of truncated polynomials for cubic splines

$$\{b_j(x)\}_{j=1}^{K} = \{1, x, d_1(x) - d_K(x), \ldots, d_{K-2}(x) - d_K(x)\}$$

where $d_k(x) = \frac{(x-\kappa_k)_+^3 - (x-\kappa_K)_+^3}{\kappa_k - \kappa_K}$ [50]. The proof of this result is quite simple. Let us consider the truncated power series representation for cubic splines with K knots

$$g(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{k=1}^{K} \theta_k (x - \kappa_k)_+^3 \tag{3.6}$$

Using the natural boundary conditions for natural cubic splines leads to

$$\beta_2 = \beta_3 = 0, \ \sum_{k=1}^{K} \theta_k = \sum_{k=1}^{K} \kappa_k \theta_k = 0$$

or equivalently $\theta_K = -\sum_{k=1}^{K-1} \theta_k$ and $\theta_{K-1} = -\sum_{k=1}^{K-2} \frac{(\kappa_k - \kappa_K)}{(\kappa_{K-1} - \kappa_K)} \theta_k$. Substituting these conditions into (3.6) implies our natural cubic splines basis.

The B-splines basis is however more suitable for calculations. This basis is obtained by linear combinations of the truncated polynomials. Namely, let $\boldsymbol{\xi} = \{\xi_0, \ldots, \xi_{N+1}\}$ be a sequence of non-decreasing real numbers such that

$$\xi_0 \leq \cdots \leq \xi_{N+1}$$

Define the augmented knot set

$$\xi_{-(m-1)} = \cdots = \xi_0 \leq \cdots \leq \xi_{N+1} = \cdots = \xi_{N+m}$$

where we have appended $m-1$ times the lower and upper boundary knots ξ_0 and ξ_{N+1}. The B-splines basis is defined by

$$B_{j,k}(x) = (\xi_{j+k} - \xi_j)[\xi_j, \ldots, \xi_{j+k}](\cdot - x)_+^{k-1}$$

for all $x \in \mathbb{R}$, $j = -(m-1), \ldots, N + m - k$ and $k = 1, \ldots, m$. In the previous definition, we used the divided differences operator $[t_0, \ldots, t_n]$ which is defined by recursion as follows

$$[t_0]f = f(t_0), \ \ [t_0, \ldots, t_n]f = \frac{[t_1, \ldots, t_n]f - [t_0, \ldots, t_{n-1}]f}{(t_n - t_0)}.$$

Since a B-spline is a linear combination of truncated power functions, so is continuous from the right. Furthermore, we can recursively define a set of real-valued functions $B_{j,k}$ as follows:

$$\begin{aligned} B_{j,1}(x) &= \mathbb{I}_{\xi_j \leq x < \xi_{j+1}} \\ B_{j,k}(x) &= \omega_{j,k}(x) B_{j,k-1}(x) + (1 - \omega_{j+1,k}(x)) B_{j+1,k-1}(x) \quad \text{for } 1 < k \leq m \end{aligned}$$

where $\omega_{j,k} = \frac{x-\xi_j}{\xi_{j+k}-\xi_j}$. For the above computation we define $0/0$ as 0. These two definitions are equivalent. (see [18, 45]). Here are some of the properties: each B-spline has support in a finite interval; the B-splines form a partition of unity, i.e. $\sum_{j=-3}^{N+m-k} B_{j,k}(x) = 1$; each B-spline $B_{j,k}(x)$ is a piecewise polynomial of order $k-1$.

These are local support functions, which implies that the corresponding matrices are strip matrices. This base is constituted by K + 2 functions, the B-splines are not natural splines, they have different restrictions on the edges.

3.2 Multivariate non-parametric regression

The multidimensional generalization of the problem (3.1) is as follows:

$$Y = f(X_1, \ldots, X_d) + \varepsilon \tag{3.7}$$

with $(\mathbf{X}, Y) = (X_1, \ldots, X_d, Y)$ a random vector, ε a random variable independent of $\mathbf{X}$ such that $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

The adjustment of Y to a d-dimensional surface can be done by generalizing the kernel smoothing [47] as

$$\mathcal{K}_{\mathbf{\Lambda}}(\mathbf{x}, \mathbf{x}') = \frac{1}{\det(\mathbf{\Lambda})} K\left(\mathbf{\Lambda}^{-1}(\mathbf{x} - \mathbf{x}')\right) \tag{3.8}$$

where $\mathbf{x} = (x_1, \ldots, x_d)^T$, $\mathbf{x}' = (x'_1, \ldots, x'_d)^T$, and $\mathbf{\Lambda}$ is a positive definite, symmetric matrix.

Many possibilities exist for defining the kernel $\mathcal{K}(\cdot)$. For example, it can be defined as the d-product of uni-dimensional kernel, i.e. $\mathcal{K}(\mathbf{t}) = \prod_{j=1}^{d} K(t_j)$, or by a single uni-dimensional kernel, i.e. $\mathcal{K}(\mathbf{t}) = K(\|t\|)$, where the choice of the norm determines the shape of the neighborhoods. Another possibility is to generalize directly the uni-dimensional kernel functions.

The generalization of the Nadaraya-Watson (3.2) is thus

$$\hat{f}_{NW}(\mathbf{x}) = \frac{\sum_{i=1}^{n} y_i \mathcal{K}_{\mathbf{\Lambda}}(\mathbf{x}, \mathbf{x}^i)}{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{\Lambda}}(\mathbf{x}, \mathbf{x}^i)} \tag{3.9}$$

where $\mathbf{x}^i = (x_{i1}, \ldots, x_{id})^T$.

The generalization of the minimization problem (3.3), in the particular case of linear local regression, is

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - (\mathbf{x} - \mathbf{x}^i)^T . \boldsymbol{\beta}_1\right)^2 \mathcal{K}_{\mathbf{\Lambda}}(\mathbf{x}, \mathbf{x}^i) \tag{3.10}$$

where $\boldsymbol{\beta}_1$ is a $d \times 1$ dimensional vector.

For cubic splines, one possibility is to generalize the penalization of the second derivative to a plate-penalty [44], i.e.

$$\int \cdots \int \left\{ \sum_{j=1}^{d} \left(\frac{\partial^2 f}{\partial x_j^2} \right)^2 + \sum_{j,k} \left(\frac{\partial^2 f}{\partial x_j \partial x_k} \right)^2 \right\} \prod_{j=1}^{d} dx_j. \tag{3.11}$$

3.2.1 Some problems in high dimensional analysis

In the multidimensional case, non-parametric regression presents several problems. First, graphical representation is not possible for more than two explanatory variables, and interpretation becomes difficult.

Second, the local methods approach fails in high dimension. This is the so-called problem of "curse of dimensionality" [12], which manifests itself in various ways. For example, assume that the observations of the explanatory variables are uniformly distributed in a d-dimensional unit cube ($d = 2, d = 10$). To recover a percentage of the data $p = 10\%$, the side length of a sub-cube should be $p^{1/d}$. The length of the side is 0.32, for $d = 2$, and 0.79, for $d = 10$. For high d, these neighborhoods are no longer local (the length of the side is very close to unity, so the sub-cube is very close to the global cube). As a result, when the dimension increases, either larger neighborhoods must be taken, implying global averages and therefore large bias, or the percentage of the data must be reduced which implies averaging over a few observations and therefore large variances of the adjustment [50].

Third, in high dimension setting, most data sets are usually "embedded" on smaller dimensional manifolds. If these manifolds are hyper-planes, we encounter the collinearity problem of the explanatory variables. If these manifolds are regular, we encounter a more general problem of concurvity [20, 49].

3.2.2 Dimension Reduction Techniques

A solution to the high dimensional problems is to assume that the regression function has a certain structure. These non-parametric techniques remain flexible tools. The price to pay is the possible erroneous specification of the model.

The techniques based on dimension reduction principles are the additive models, which assume that the regression function is a sum of mono-variate functions in each of the variables, projection pursuit models, close to multilayer perception neural networks, and regression trees.

Projection pursuit

The algorithm of projection pursuit is to build an additive regression model of the form [38, 65]:

$$Y = \sum_{k=1}^{K} f_k(\boldsymbol{\alpha}_k^T.\mathbf{X}) + \varepsilon \tag{3.12}$$

where ε is a random variable such that $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$ and independent of the explanatory variables.

The vector explanatory variables is projected on K directions $\{\boldsymbol{\alpha}_k\}_{k=1,\dots,K}$. The regression surface is constructed by estimating one-dimensional regressions f_k applied to projections. The directions $\{\boldsymbol{\alpha}_k\}_{k=1,\dots,K}$ and the number of terms K are chosen by model selection methods such as generalized cross validation.

The advantage of this technique is that it allows easy processing of low density data. The model is also little constraint. Nevertheless, for $K > 1$, this model presents difficulties of interpretation: it is difficult to evaluate the contributions of each variable. For $K = 1$, the model is known as single−index model.

Projection pursuit techniques are often compared to multilayer perception neural networks. These two methods extract linear combinations of inputs, and then model the output variable as a nonlinear function of these input variables. However, the functions f_k of the projection pursuit are different and non-parametric, whereas the neural networks use a simpler activation function, normally the softmax (or logistic) function. In the case of projection pursuit, the number of "layers" is set at two [1] and the number of functions K is also predefined, which is not the case for neural networks.

Regression trees

Regression trees divide the space of the explanatory variables into a set of hyper-cubes. A simple model (for example, a constant) is then fitted to each hyper-cube as

$$f(\mathbf{x}) = \sum_{k=1}^{K} \bar{\alpha}_k \mathbb{I}_{\{\mathbf{x} \in R_k\}} \tag{3.13}$$

with K the number of partitions of the space of the explanatory variables, R_k disjoint regions, $\bar{\alpha}_k$ the constant that models the response in the region. The algorithm simultaneously decides the partition and the values of the parameters $\{\bar{\alpha}_k\}_{k=1,\dots,K}$.

Regression trees have the advantage of conceptual simplicity and the ability to interpret. Their limitations are instability and lack of continuity of the regression surface.

3.2.3 Additive models

Additive models assume that the regression function can be written as a sum of functions of the explanatory variables [48, 107]:

$$Y = \alpha_0 + \sum_{j=1}^{d} f_j(X_j) + \varepsilon \tag{3.14}$$

where ε is independent of $\mathbf{X} = (X_1, \dots, X_d)$, $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$; α_0 is a constant, f_j, $j = 1, \dots, d$ are the univariate functions such that $\mathbb{E}_{X_j}[f_j(X)] = 0$. This condition of identifiability implies that $\mathbb{E}_{\mathbf{X}}[Y] = \alpha_0$ [49].

Additive models can be introduced as a generalization of the linear regression models. This is the basic tool for modeling the relationship between the continuous response variable and the explanatory variables:

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_d X_d + \varepsilon \tag{3.15}$$

[1] More precisely, we have $\mathbf{X} \longrightarrow \{\boldsymbol{\alpha}_k^T.\mathbf{X}\} \longrightarrow \{f_k(\boldsymbol{\alpha}_k^T.\mathbf{X})\} \longrightarrow Y$.

where ε is independent of $\mathbf{X}$, $\mathbb{E}(\varepsilon) = 0$ and $\mathrm{Var}(\varepsilon) = \sigma^2$.

The assumption of linear dependence of $\mathbb{E}_{\mathbb{X}}[Y]$ in each of the explanatory variables is a strong assumption. When this assumption is not verified, one way to extend the linear model is the additive model. The non-parametric form of f_j gives more flexibility to the model, while the additive structure preserves the possibility of representing the effect of each variable. The model can be represented by one-dimensional functions describing the roles of explanatory variables in response modeling, which facilitates interpretation. However, the simplicity of the linear model is lost. A new problem appears: the selection of smoothing parameters, representing the complexity of each component of the model. Here we will list some properties of the additive models.

Interpretability: The joint effect of the explanatory variables on the response variable is expressed as a sum of the individual effects. These individual effects show how the expectation of the response varies when one of the components varies while the others are fixed to any values. Thus, the individual functions can be represented separately in order to visualize the effect of each explanatory variable, making the result intelligible. The possibility of representing the effects of the variables directly at the same time gives indications on the importance of each of the variables.

Scourge of dimensionality: By restricting the nature of the dependencies, the problems related to the high dimension are mitigated: the response is modeled as the sum of uni-dimensional functions of the explanatory variables, instead of being modeled by multidimensional functions. Therefore, the number of observations required increases linearly with d (and not exponentially).

Consider the estimation of the regression function (3.7). The optimal asymptotic rate for the estimate of f is $n^{-[m/(2m+d)]}$, where m is an index of the regularity of the function f is $m-1$ times continuously differentiable and its m-th directional derivatives exist) [106]. On the other hand, if f is additive, the optimal rate reaches the uni-dimensional convergence rate $n^{-[m/(2m+1)]}$ [108]. In this sense, the additive models are considered as dimension reduction techniques.

Invalid model: The model is poorly specified when the explanatory variables interact. That is, the effect of the variations of an explanatory variable on the response depends on the values adopted by the other explanatory variables.

Suppose the general multiple regression model (3.7), where the function $f(\cdot)$ is a smooth function. Assuming that the observations $\{x_{ij}\}$ are contained in a region where the curvature of the function f is small, then the additivity (and linearity) can be justified by a first-order Taylor expansion $f(\mathbf{x}) \approx f(\mathbf{x}') + Df(\mathbf{x}')(\mathbf{x} - \mathbf{x}')$, where $\mathbf{x}'$ is within the region defined by the observations and Df indicates the gradient of f. If the curvature of f is high, the Taylor expansion requires, at least, quadratic terms and cross terms in two variables. When only the former are needed, the model is always additive, although it incorporates "nonlinear" terms.

Adaptability: The interest of additive models is their ability to model the relationship between variables in an intuitive way, but also the possibility of adapting the model to simpler or more complex situations. When components do not require nonparametric modeling, they can be reduced to linear components. Also, when

interactions exist between certain variables, quadratic (or higher order) terms can be integrated into the model.

Part II-Application of Bayesian penalized spline regression for internal modeling

3.3 Notations and requirements for the fitting process

3.3.1 Risk factors

As will be seen later, we use the term "risk factors" to refer to the underlying parameters that may impact the balance sheet. As discussed in Chapter 2 and in Appendix B, we might notice that there are several risk factors, which may affect an insurance firm's economic balance sheet. These risks can generally be classified into the following categories:

1. asset-related risks (interest rates, equity and property prices, credit spreads): asset-liability impacts of variances in underlying parameters across all lines of business;

2. insurance risks: claims (mortality/morbidity/longevity), discontinuances, expenses, including effects of both actual experience over the period of assessment as well as the impact of that experience on the closing liability assessment;

3. counterparty risks: risk of default by key counterparties such as reinsurers;

4. operational risks.

There are some risks which could affect the company but have little impact on the asset position (reputation risk is an example). There are some risks which not be mitigated by holding capital against them (liquidity risk is an example). These risks should be considered in the broader risk management framework but might not feature in the calculation of Economic Capital.

The approach to quantifying risks will vary by risk type. This includes:

- factor-based methods, where a factor is applied to a driver to approximate the impact of a risk;

- stress testing, where a specific shock is defined and the impact of that shock on the balance sheet is determined;

- stochastic modelling, where a full distribution of shocks are modelled, producing a full distribution of own funds outcomes.

Like many other proxy models (e.g. Curve Fitting, LSMC), only stochastic risk factors are taken into account in this methodology.

3.3.2 Loss function

The loss function in life insurance is a function defined as the change in Basic Own Funds due to the realization of different economic states of the world. Mathematically, let us denote by $(RF_1, \ldots, RF_d)$ a d-tuple standing for the underlying risk factors and by $(x_1, \ldots, x_d)$ another d-tuple representing the shock applied at the current state. The loss function is then defined as:

$$\phi(x_1, \ldots, x_d) = \mathrm{BOF}_{t=0}\left(RF_1, \ldots, RF_d\right) - \mathrm{BOF}_{t=0^+}\left(RF_1(1+x_1), \ldots, RF_d(1+x_d)\right) \tag{3.16}$$

In the Solvency II environment, the value of each balance sheet part corresponds to the expected value of the discounted future cash-flows under a risk-neutral probability $\mathbb{Q}$. Let

- $\mathrm{DF}(0,t)$ be the stochastic discount factor in terms of a risk free instantaneous interest rate r_s, i.e. $\mathrm{DF}(0,t) = e^{\int_0^t r_s ds}$;
- R_t be the company's profit in period t.

Under this notation, the Basic Own Funds at the initial date is calculated in the following manner:

$$\mathrm{BOF}_{t=0}(RF_1, \ldots, RF_d) = \mathbb{E}^{\mathbb{Q}}\left[\sum_{u=1}^{T} \mathrm{DF}(0,u) R_u \mid (RF_1, \ldots, RF_d)\right]$$

From this we may rewrite the loss function (3.16) in terms of the conditional expected value with respect to capital loss.

Namely, let

$$\mathcal{Y} = \mathrm{BOF}_{t=0}(RF_1, \ldots, RF_d) - \sum_{u=1}^{T} \mathrm{DF}(0,u) R_u \mid_{(RF_1(1+x_1), \ldots, RF_d(1+x_d))}$$

be the potential capital loss which is a random variable whose conditional distribution depends on the $x_1, \ldots, x_d$. The loss function (3.16) can be equivalently rewritten as:

$$\phi(x_1, \ldots, x_d) = \mathbb{E}^{\mathbb{Q}}\left(\mathcal{Y} \mid x_1, \ldots, x_d\right) \tag{3.17}$$

For latter use, let us define the standalone loss functions $\phi_j(x_j)$ which is of the form:

$$\phi_j(x_j) = \mathbb{E}^{\mathbb{Q}}\left(\mathcal{Y} \mid 0, \ldots, x_j, \ldots, 0\right) \tag{3.18}$$

and the excess loss function which is expressed as:

$$\phi_{1d}(x_1, \ldots, x_d) = \mathbb{E}^{\mathbb{Q}}\left(\eta \mid x_1, \ldots, x_d\right) \tag{3.19}$$

with $\eta = \mathcal{Y} - \sum_{j=1}^{d} \phi_j(x_j)$ the residual loss of capital. From this, it is easily seen that the following relation holds:

$$\phi(x_1, \ldots, x_d) = \sum_{j=1}^{d} \phi_j(x_j) + \phi_{1d}\left(x_1, \ldots, x_d\right). \tag{3.20}$$

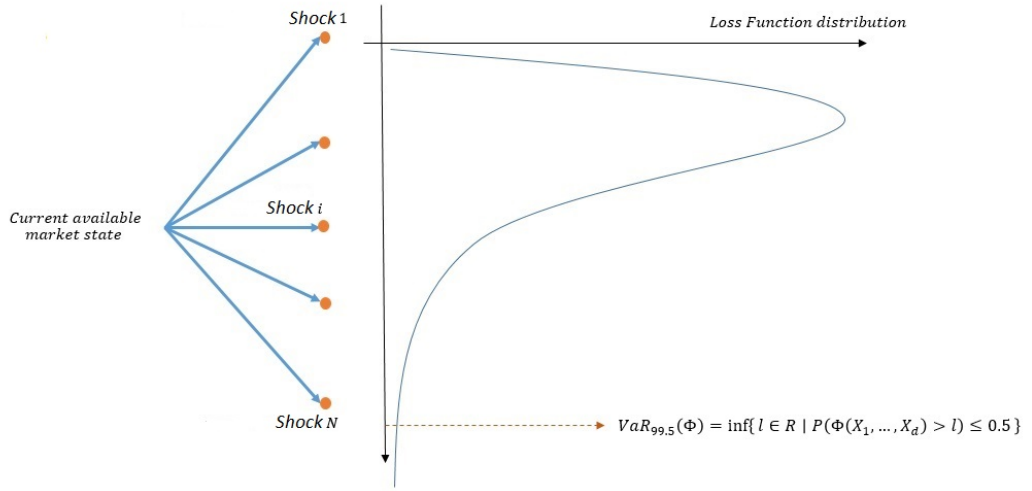


Figure 3.1: Here we illustrate the empirical SCR estimation with N_1 outer simulations.

Stand-alone loss functions are useful for risk monitoring, as they are used to analyze the contribution of each risk factor to the capital requirements. In fact, to have stand-alone loss functions allows to classify and quantify risk exposure to risk factors, which corresponds to a useful tool for steering the activity. Thus, this specific decomposition for stand-alone loss functions is fully aligned with market practice. As a result, the standalone loss functions should be independently calibrated so as not to modify the other estimators.

3.3.3 Approximation of a shock at $t = 0^+$

As seen previously, the Nested Simulation approach requires the realization of the real world scenarios between $t = 0$ and $t = 1$. In practice, one usually relies on a different approach, which consists of performing the approximation of a shock at $t = 0^+$. Regarding the market risks, this means that the market value of financial instruments are modified right after their initialization. This approximation is in line with the standard formula approach, where the defined shocks are to be applied instantaneously on the balance sheet. In the standard model, the SCR is evaluated via the "square-root" formula based on a modular approach. In our setting (internal model), the SCR is defined as the 99.5%-quantile of the loss function (see Section 3.3.2 for the definition of the loss function), i.e.

$$\text{SCR} = \text{q}_{99.5\%}(\phi) \tag{3.21}$$

As will be seen later, further analyses will be performed to highlight the reliability of our assumption (3.21) and the resulting estimate. Namely, we carry out the comparison between the SCR estimated by Nested Simulations approach and the SCR estimated by our method.

3.4 Methodology description

We assume that the functional components $\{\phi_1, \ldots, \phi_d, \phi_{1d}\}$ are regular enough and every standalone loss function can be thus approximated by polynomial spline function (see Appendix D.7). To guarantee the granularity, we will estimate independently each standalone loss functions by applying the Bayesian penalized spline regression (see Appendix D.3). Regarding the Excess Loss function, even if $\phi_{1d}(x_1, \ldots, x_d)$ is not genuinely additive, an additive approximation to ϕ_{1d} may be sufficiently accurate as well as being readily interpretable. We define ϕ^*_{1d} being of the form:

$$\phi^*_{1d}(x_1, \ldots, x_d) = h_0 + \sum_{j=1}^{d} h_j(x_j) + \sum_{1 \leq j < j' \leq d} h_{jj'}(x_j \times x_{j'}) \tag{3.22}$$

To ensure identifiability of each of the functional components, we include the intercept h_0 and the functions $h_j, h_{jj'}$ are chosen subject to the constraints $\frac{1}{n}\sum_{\nu=1}^{n} h_j(x_j^{(\nu)}) = \frac{1}{n}\sum_{\nu=1}^{n} h_{jj'}(x_j^{(\nu)} \times x_{j'}^{(\nu)}) = 0$ where $\left(x_1^{(\nu)}, \ldots, x_d^{(\nu)}\right)_{\nu=1,\ldots,n}$ are the design points as will be seen later. Then ϕ^*_{1d} is the best additive approximation to ϕ_{1d} in the sense of mean squared error. In the case that ϕ_{1d} is additive, we then have $\phi_{1d} = \phi^*_{1d}$. The Excess Loss function will be thus estimated via the additive model (3.22). With a slight abuse of notation, we use the same symbol here as in Equation (3.20), i.e.:

$$\phi(x_1, \ldots, x_d) = \sum_{j=1}^{d} \phi_j(x_j) + \phi^*_{1d}(x_1, \ldots, x_d).$$

Confidence intervals are of primary importance in many practical applications of how accurate the estimate can predict. As remarked by Nychka [84], "one limitation in applying spline methods in practice, however, is the difficulty in constructing confidence intervals or specifying other measures of the estimate's accuracy". Wahba [115] suggested a Bayesian approach to derive the point-wise asymptotic confidence interval. Surprisingly, these Bayesian asymptotic confidence intervals work well even when evaluated from the frequentist viewpoint. As it stands, we will use this approach to derive the asymptotic confidence intervals for the standalone loss functions estimation as well as the excess loss function estimation. However, we would like to point out that the confidence bands do not incorporate approximation errors from the choice of the regression function. This remark will be shown in the derivation which can be found in Appendix D.4 and Appendix D.5.

The proposed methodology is summarized in five steps:

- Fitting points selection: The first step consists of generating the design points $\{x_j^{(\nu)}\}_{j=1,\ldots,d}^{\nu=1,\ldots,n_j}$. Since the sample size is limited by calculation time, only significant and comparable point are mainly considered. In fact, we are ultimately interested in calculating a percentile on the economic own funds distribution,

it is consequently important to know with accuracy what will be the potential loss with a security level of 99.5% rather than having a perfect estimation of the loss function for every realistic scenario even though the final objective is to get the most accurate possible fitting. Therefore, a sufficient number of "tail scenarios" is required in order to estimate properly the economic own funds tail distribution. To this end, we assume that the desired tail scenarios lie in the tail of distribution of each underlying risk factor and this latter one defines the 99.5th-percentile region. The determination of the 99.5th-percentile region depends however upon the risk-factor considered. For example, we know that the falling equity markets implies the negative impact on BOF, and thus increases the loss capital. Therefore, the 99.5th-percentile for the Equity risk is placed on the left tail of its distribution which is defined as the interval between the 0.01th percentile and the 10th-percentile. Concerning the convex (or U-shaped) loss functions, such as the Interest rate loss function, the tail scenarios should be picked from both extremes of its distribution since it is not clear ex ante of whether the highest or lowest values of interest rates would be most problematic. All design points located outside of the 99.5th-percentile region can be selected randomly and uniformly.

- Standalone loss functions estimation: Using the ALM model described in Section 3.5.1.4, one values the empirical estimate of the standalone losses for each selected design points $\{x_j^{(\nu)}\}_{j=1,\dots,d}^{\nu=1,\dots,n_j}$, that is $\bar{\phi}_j(x_j^{(\nu)}) = \frac{1}{N_2}\sum_{k=1}^{N_2} \mathcal{Y}^{(k)}_{|x_j^{(\nu)}}$ with N_2 the number of inner scenarios and $\mathcal{Y}^{(k)}$ the loss of capital associated to the k-th inner scenario given the market stress condition $x_j^{(\nu)}$. Next one applies the Bayesian penalized spline regression model to smooth the data by solving the following optimization problem:

$$\min_{\beta\in\mathbb{R}^{p_j+K_j+1}} \left[\sum_{\nu=1}^{n_j}\left(\bar{\phi}_j\left(x_j^{(\nu)}\right) - B\left(x_j^{(\nu)}\right)\beta_j\right)^2 + \lambda_j\beta_j^T\mathbf{D}_j\beta_j\right], \ \forall j = 1,\dots,d \tag{3.23}$$

where $B(x) = \left(1, x, x^2, \dots, x^{p_j}, (x-\kappa_1)_+^{p_j}, \dots, (x-\kappa_{K_j})_+^{p_j}\right)^T \in \mathbb{R}^{1+p_j+K_j}$ is the truncated p_j-polynomial basis with K_j knots $\{\kappa_1, \dots, \kappa_{K_j}\}$, the symbol $(\cdot)_+$ stands for the Heaviside step function and $\mathbf{D}_j$ is the block diagonal matrix diag $\left(\mathbf{0}_{1+p_j}, \mathbf{1}_{K_j}\right)$. (For the detailed description of the Bayesian penalized spline regression, please refer to Appendix D). Let $\mathbf{B}_j$ be the $n_j \times (1+p_j+K_j)$ matrix whose i's row equals $B\left(x_j^{(i)}\right)^T$ and $\overline{\Phi}_j = \left(\bar{\phi}_j\left(x_j^{(1)}\right), \dots, \bar{\phi}_j\left(x_j^{(n_j)}\right)\right)^T$, the estimate of β_j is then given by:

$$\hat{\beta}_j = \left(\mathbf{B}_j^T\mathbf{B}_j + \lambda_j\mathbf{D}_j\right)^{-1}\mathbf{B}_j^T\overline{\Phi}_j \tag{3.24}$$

for $j = 1, \dots, d$. From this it follows that $\widehat{\Phi}_j = \left(\hat{\phi}_j\left(x_j^{(1)}\right), \dots, \hat{\phi}_j\left(x_j^{(n_j)}\right)\right) = \mathbf{B}_j\hat{\beta}_j$.

The deviation of $\hat{\phi}_j\left(x_j^{(\nu)}\right)$ from $\phi\left(x_j^{(\nu)}\right)$ is characterized by the $1-\alpha$ Bayesian asymptotic confidence interval having the form:

$$\mathbb{P}\left(|\hat{\phi}_j\left(x_j^{(\nu)}\right) - \phi\left(x_j^{(\nu)}\right)| \leq \Delta_{j,\alpha}^{(\nu)}\right) \to 1-\alpha \tag{3.25}$$

where $\Delta_{j,\alpha}^{(\nu)} = \max\left(\left|z_{\alpha/2}\sqrt{\left(\widehat{E(\mathfrak{M}_j)} - \widehat{E(\mathfrak{B}_j)}^2\right)\left[\mathbf{V}_{\hat{\Phi}_j}\right]_{\nu\nu}} \pm \widehat{E\left(\mathfrak{B}_j\right)}\sqrt{\left[\mathbf{V}_{\hat{\Phi}_j}\right]_{\nu\nu}}\right|\right)$, $z_{\alpha/2}$ is the critical point from a standard normal distribution and the explicit form of $\widehat{E(\mathfrak{M}_j)}, \widehat{E(\mathfrak{B}_j)}$ and $\mathbf{V}_{\hat{\Phi}_j}$ are given in Appendix D.4.

- Excess loss function estimation: Again we use the ALM model and the standalone loss functions estimators to value the empirical excess losses at each design points $x_{1d}^{(\nu)} = \left(x_1^{(\nu)}, \ldots, x_d^{(\nu)}\right)$ for $\nu = 1, \ldots, n$, that is

$$\bar{\phi}_{1d}\left(x_{1d}^{\nu}\right) = \frac{1}{N_2}\sum_{k'=1}^{N_2} \mathcal{Y}_{|x_{1d}^{(\nu)}}^{(k')} - \sum_{j=1}^{d}\hat{\phi}_j\left(x_j^{(\nu)}\right)$$

wherein $\mathcal{Y}_{|x_{1d}^{(\nu)}}^{(k')}$ is the capital loss associated to the k'-th inner scenario given the market stress condition $x_{1d}^{(\nu)}$. These are considered to be the responses variables. To ensure that independence is not broken, scenarios used to derive the excess loss function should be different from those used in the standalone loss functions calibration. The algorithm to derive the functional components $h_j, h_{jj'}$ estimators of excess loss function is analogue to that of the standalone loss function. The computation is tedious and can be found in Appendix D.5, so we will omit it here. For later use, we define the $1-\alpha$ Bayesian asymptotic confidence interval for $\hat{h}_j$, $\hat{h}_{jj'}$

$$\tilde{\Delta}_{J,\alpha}^{(\nu)} = \max\left(\left|z_{\alpha/2}\sqrt{\left(\widehat{E\left(\mathfrak{M}_J\right)} - \widehat{E\left(\mathfrak{B}_J\right)}^2\right)\left[\mathbf{V}_{\hat{h}_J}\right]_{\nu\nu}} \pm \widehat{E\left(\mathfrak{B}_J\right)}\sqrt{\left[\mathbf{V}_{\hat{h}_J}\right]_{\nu\nu}}\right|\right) \tag{3.26}$$

where J can be j or jj', $z_{\alpha/2}$ is the critical point from a standard normal distribution and the explicit form of $\widehat{E(\mathfrak{M}_J)}, \widehat{E(\mathfrak{B}_J)}$ and $\mathbf{V}_{\hat{h}_J}$ are given in Appendix D.5.

- Loss function estimation error control: We will now investigate the control of the deviation of $\widehat{\phi}$ from ϕ at an arbitrary design points $x^{(\nu)} = (x_1^{(\nu)}, \ldots, x_d^{(\nu)})$. Obviously, we have

$$\left\{|\widehat{\phi}(x^{(\nu)}) - \phi(x^{(\nu)})| > \sum_{j=1}^{d}\Delta_{j,\alpha}^{(\nu)} + \sum_{J}\tilde{\Delta}_{J,\alpha}^{(\nu)}\right\} \subset$$
$$\left(\bigcup_{j=1}^{d}\left\{|\hat{\phi}_j\left(x_j^{(\nu)}\right) - \phi_j\left(x_j^{(\nu)}\right)| > \Delta_{j,\alpha}^{(\nu)}\right\}\right) \cup \left(\bigcup_{J}\left\{|\hat{h}_J\left(x_J^{(\nu)}\right) - h_j\left(x_J^{(\nu)}\right)| > \tilde{\Delta}_{J,\alpha}^{(\nu)}\right\}\right) \tag{3.27}$$

where we used the notation $x_{jj'}^{(\nu)} = x_j^{(\nu)} \times x_{j'}^{(\nu)}$. From this it follows that the probability of deviation of $\widehat{\phi}(x^{(\nu)})$ from $\phi(x^{(\nu)})$ is asymptotically bounded by

$$\mathbb{P}\left(|\widehat{\phi}(x^{(\nu)}) - \phi(x^{(\nu)})| > \sum_{j=1}^{d} \Delta_{j,\alpha}^{(\nu)} + \sum_{J} \tilde{\Delta}_{J,\alpha}^{(\nu)}\right) \leq 1 - (1-\alpha)^{\frac{d(d+3)}{2}} \quad (3.28)$$

The derivation of this result can be found in Appendix D.6.

Motivated by this, we can estimate SCR by $\widehat{SCR} = \hat{q}_{99.5}\left(\hat{\phi}\right)$ its empirical 99.5th-percentile derived from $\hat{\phi}$. In this stage, $\hat{\phi} \equiv \hat{\phi}(X)$ is a random variable with $X = (X_1, \ldots, X_d)$ the realistic random market state or the primary simulation state whose marginal distribution is $\mathbb{P}_X$. Let $f_{\hat{\phi}}$ denote the density function of $\hat{\phi}(X)$.

To control the probability of deviation of the target SCR from its estimate, we will need certain conditions to make the theory work. First of all, it is important to clarify that as will be seen below, the resulting confidence band will not incorporate the approximation error from the choice of the regression function.

Let us introduce some notation, definitions that will be used in the sequel. We define the (L, Ω)-Lipschitz class of functions, denoted $\Sigma(L, \Omega)$, as the set of function $g : \Omega \to \mathbb{R}$ satisfy, for any $x, x' \in \mathbb{R}^d$, the inequality:

$$|g(x') - g(x)| \leq L\|x' - x\|$$

with $\Omega \subset \mathbb{R}^d$ and $\|x\| \triangleq (x_1^2 + \cdots + x_d^2)^{1/2}$. Let $r > 0$. We define $B(a, r) = \{x \in \mathbb{R}^d \mid \|a - x\| \leq r\}$. We denote by $\bar{V}_\phi = \{x \in \mathbb{R}^d \mid \phi(x) = q_{99.5\%}(\phi)\}$ and $\bar{V}_{\hat{\phi}} = \{x \in \mathbb{R}^d \mid \hat{\phi}(x) = q_{99.5\%}(\hat{\phi})\}$ the closed set of the 99.5th-percentile scenarios for ϕ and $\hat{\phi}$ respectively.

Let Γ denote the available sampling budget used to calibrate $\hat{\phi}$. Based on the work of Aerts et al. [2], it is straightforward to deduce that for $\lambda_{\phi_j}(\Gamma)$ and $\lambda_{h_J}(\Gamma)$ tending to 0, the estimate $\hat{\phi}$ converges in mean square to ϕ as $\Gamma \to \infty$. Furthermore, by Markov's inequality, convergence in mean square of $\hat{\phi}$ leads to the convergence in probability of $\hat{\phi}(x)$ to $\phi(x)$ for every $x \in \mathbb{R}^d$. This implies that for every $x^* \in \bar{V}_\phi$, there exists a random sequence $x_{(\Gamma)}^* \in \bar{V}_{\hat{\phi}}$ converges in probability to x^*.

Introduce now three assumptions on ϕ, $\hat{\phi}$ and $x_{(\Gamma)}^*$ that will be used in the last step:

ASSUMPTION 1: Suppose that $\phi \in \Sigma(L, \Omega)$ where $L > 0$ and $\Omega(\supset \bar{V}_\phi)$ is an open subset of $\mathbb{R}^d$.

ASSUMPTION 2: For any $x^* \in \bar{V}_\phi$ and $r > 0$, there exists two positive constants $\xi(r, d)$, $\gamma(r, d)$ such that

$$\mathbb{P}\left(\|x^* - x_{(\Gamma)}^*\| > r\right) \leq \xi(r, d)\Gamma^{-\gamma(r,d)}$$

for large enough Γ.

ASSUMPTION 3: For any choice of $x^* \in \bar{V}_\phi$ and $\alpha \in (0, 1)$, there exists two positive constants $r(\Gamma)$ and $\Delta(\alpha, \Gamma)$, with $r(\Gamma) \xrightarrow{\Gamma \to \infty} 0$, such that

$$\mathbb{P}\left(\mid \hat{\phi}(x) - \phi(x) \mid > \Delta(\alpha, \Gamma)\right) \leq 1 - (1-\alpha)^{\frac{d(d+3)}{2}}, \quad \forall x \in B(x^*, r(\Gamma))$$

for large enough Γ.

- SCR estimation error control: In the following, we denote by N_1 the number of the primary simulations. Note that

$$\left|\widehat{SCR} - SCR\right| \leq \left|\hat{q}_{99.5\%}\left(\hat{\phi}\right) - q_{99.5\%}\left(\hat{\phi}\right)\right| + \left|q_{99.5\%}\left(\hat{\phi}\right) - q_{99.5\%}\left(\phi\right)\right| \quad (3.29)$$

The first term on the right-hand side corresponds to the numerical error since we appeal the empirical percentile to estimate the SCR and the second term represents the model error. Note that the numerical error depends not only on the empirical assessment $\hat{q}_{99.5\%}$ but also on the fitting quality $\hat{\phi}$. To value this numerical error, we apply the Theorem in Appendix D.8. Namely, we have

$$\mathbb{P}\left(\left|\hat{q}_{99.5\%}\left(\hat{\phi}\right) - q_{99.5\%}\left(\hat{\phi}\right)\right| > z_{\alpha/2}\frac{0.07}{\sqrt{N_1} f_{\hat{\phi}}(q_{99.5\%}\left(\hat{\phi}\right))}\right) \to \alpha \quad (3.30)$$

as $N_1 \to \infty$. In the previous expression, the distribution function $f_{\hat{\phi}}$ and the evaluated point $q_{99.5\%}\left(\hat{\phi}\right)$ are however unknown and will be then replaced by their estimators. Regarding the second term, by using Assumptions (1-3), we obtain the asymptotic probability of deviation of $q_{99.5\%}\left(\hat{\phi}\right)$ from $q_{99.5\%}\left(\phi\right)$ having the form:

$$\mathbb{P}\left(\left|q_{99.5\%}\left(\widehat{\phi}\right) - q_{99.5\%}\left(\phi\right)\right| > \Delta(\alpha,\Gamma) + Lr^*\right) \leq \left[1 - (1-\alpha)^{\frac{d(d+3)}{2}}\right] + \xi(r^*,d)\Gamma^{-\gamma(r^*,d)} \quad (3.31)$$

where $r^* \equiv r(\Gamma)$. The derivation of this result can be found in Appendix D.6. Combing the equations (3.30) and (3.31) leads to the control of the probability of deviation of $\widehat{SCR}$ from SCR.

The confidence interval $\Delta(\alpha,\Gamma) + Lr^*$ is however an issue as it involves the unknown parameters $\Delta(\alpha,\Gamma)$, L and r^*. In the following, we suggest a method to estimate these parameters in practice.

In order to estimate the Lipschitz constant, we find the supremum of all slopes $|\hat{\phi}(x) - \hat{\phi}(x')|/\|x - x\|$ for distinct points x and x' within the 99.5th-percentile region. We call $\hat{x}^*$ the empirical 99.5th-percentile scenario, i.e. $\hat{\phi}(\hat{x}^*) = \hat{q}_{99.5\%}\left(\hat{\phi}\right)$. The parameter $\Delta(\alpha,\Gamma)$ will be then replaced by $\tilde{\Delta}(\alpha,\Gamma) = \sum_{j=1}^{d} \Delta_{j,\alpha}^{(\hat{x}^*)} + \sum_J \tilde{\Delta}_{J,\alpha}^{(\hat{x}^*)}$. To estimate the parameter r^*, we seek the maximum radius $\hat{r}^*$ such that for every $x^{(\nu)} \in B(\hat{x}^*, \hat{r}^*)$, the confidence intervals $\sum_{j=1}^{d} \Delta_{j,\alpha}^{(\nu)} + \sum_J \tilde{\Delta}_{J,\alpha}^{(\nu)}$ are close to $\tilde{\Delta}(\alpha,\Gamma)$. On the right-hand side of the inequality (3.31), as the true value of $\xi(r^*,d)$ and $\gamma(r^*,d)$ are unknown, it is not possible to have a direct access to the upper bound of the probability. In practice, a large number of Γ is necessary so that the term $\left[1 - (1-\alpha)^{d(d+3)/2}\right]$ becomes preponderant compared to $\xi(r^*,d)\Gamma^{-\gamma(r^*,d)}$.

Asset	1 130 000	Liability	1 130 000
Cash	130 000	Own Funds	100 000
		Capitalization Reserve	10 000
		Liquidity risk provision	10 000
		Profit-sharing provision	10 000
Cash	50 000	Mathematical Reserves	1 000 000
Equities/Real Estate	150 000		
Bonds	800 000		

Figure 3.2: Initial Balance Sheet used to value the prudential balance sheet.

3.5 Numerical study

Being aware of the limitations of the ALM model considered in this paper, the area of use of this cash-flow generator depending on only a few risk factors is sufficient in the context of this study. That is to say, the study serves as illustration for our proposed methodology. In practice, note that for many life insurance companies, the number of risk factors is often very large, e.g., over 100. Therefore, the efficiency as well as the performance of this approach compared with the existing approaches remain unknown for practical problems and can be the subject of future research.

3.5.1 ALM modeling

In this section, we recall in a *concise* way the main lines of the operation of an actuarial cash flow simulator that is used today by life insurers to value their prudential balance sheet. All the details of our ESG and ALM cash-flow simulator are given in Chapters A and B.

3.5.1.1 Initial balance sheet

We model the prudential balance sheet of an insurance company that sells exclusive savings contracts in euros. The initial balance sheet of the modeled insurance company is defined as follows:

Assets backing mathematical reserves consist of cash, equities/real estate and bonds, by convention, up to 5%, 15% and 80%, respectively. Assets backing liquidity risk provision (PRE), profit-sharing provision (PPE), capitalization reserve and own funds are not explicitly modeled and implicitly considered to be 100% of cash (see Figure 3.2).

The cash-flow simulator considers the risk neutral evolution of financial risk factors on the balance sheet and these trajectories run for 50 years. At each time step, the cash flows of assets and liabilities are calculated and the company's manage-

ment strategy is implemented (calculation of the profit-sharing rate, allocation of provisions, distribution of dividends to shareholders, etc.).

3.5.1.2 Liabilities

Regarding the liabilities, the contracts in the liabilities are exclusively savings contracts in euros with minimum guaranteed rates (MGR) and redemption rights. In accordance with the Solvency II Directive, the valuation of the prudential balance sheet is carried out in run-off on insurance liabilities, which means that no new future production is considered and under the assumption of continuity of activities on the assets side (maintenance of target allocation, management decisions, etc.).

Savings contracts generally allow policyholders to withdraw partially or totally their savings. We can distinguish two types of redemptions in life insurance:

1. *Conjunctural surrenders*: these are the redemptions linked to the economic situation and the performance of the insurer. They are usually estimated from the difference between the rate served by the insurer and the rate served by the competition.

2. *Structural surrenders*: these are the redemptions related to the characteristics of the contract. For example, there is usually a wave of structural surrender after the 8 year seniority. This phenomenon is explained by the taxation of life insurance, which becomes more favorable when it comes to redemptions if the contract has 8 year seniority.

The insurer makes its asset allocation according to the characteristics of its liabilities (duration, MGR ...). If actual redemptions are greater than expected redemptions, for the sake of liquidity, the insurer will be forced to sell assets that have not matured, which may be a disadvantage if these assets are unrealized losses. Similarly, the insurer has to pay a capital or an annuity in case of death of the insured to the beneficiary designated by the contract. Therefore, the risk of a buyout and mortality risk are two major risks for life insurers that results from a behavioral change of insured persons. The modeling of these behaviors is therefore a crucial issue for the asset-liability management of an insurance company.

Insured mortality is assumed to be deterministic and the death rate is given by the death table "*TH0002*" for men and "*TF0002*" for women [2]. With respect to modeling the redemptions in our setting, the total redemption rate (TR) of the model is calculated as the sum of the conjunctural surrender rate (CR), which is a function of the spread between the rate expected by the insured and the last profit sharing rate served by the insured, and the structural surrender rate (SR), which is determined on the policyholder's seniority according to a historically calibrated redemption table.

[2] The *TH-TF 0002* mortality table is built from the INSEE 2000 – 2002 table - respectively for the male population and for the female population. These are the regulatory tables for life insurance contracts (other than life annuities) The table is available at `http://www.spac-actuaires.fr/jdd/public/documents/xls/TH-TF%2000-02.xls`.

3.5.1.3 Assets

The asset portfolio of the insurance company consists of the following asset classes: cash, shares/real estate and bonds. Cash is remunerated at the risk-free rate. We choose the Hull & White one factor model to model the dynamics of short-term interest rates. Namely, under the risk neutral probability $\mathbb{Q}$, the instantaneous short-term interest rate r_t is governed by the following dynamics:

$$dr_t = (\theta_t - ar_t)dt + \sigma dW_t$$

where σ represents the instantaneous volatility of the short rate, a is the mean-reverting speed and W_t the Brownian motion under the risk neutral probability $\mathbb{Q}$. The time dependent parameter θ_t is determined by σ, a and the initial yield curve $\{R(0,T)\}$. As there is only one driving Brownian motion, all forward rates are determined by the short rate. Since the dynamics of short rates depend on the mean-reverting speed, the interest rate volatility and the initial yield curve, these latter ones will thus completely determine the shape of the forward yield curves. In this study, we fix $a = 0.35$, $\sigma = 0.5\%$ and take the risk-free interest rate term structures published by EIOPA as only input of the Hull & White model. Readers can refer to [36] for more details of this model.

Shares/real estate are modeled by a geometric Brownian movement with constant dividend rate and log-normal volatility at 17.4%. The dividend or rent rate is adjustable. Here, the dividend rate is set at 3%, the rental rate at 5%. The treasury bond portfolio consists of government bonds whose probability of default is assumed to be zero.

3.5.1.4 ALM Model

The ALM simulator projects the assets and liabilities of the insurer over time. This makes it possible to determine at each time step the balance sheet and the value of the flows distributed to the policyholders on the one hand and to the shareholders on the other hand.

At each time step, along with the risk neutral trajectory, the ALM simulator proceeds in 6 steps to forecasting the assets and liabilities by one year, calculating the cash flows of liabilities and assets, updating the balance sheet and determining the value of the outgoing flows:

1. Sale or purchase of assets to recover the target allocation at book value: 80% bonds, 15% equities and 5% monetary.

2. Disseminate the stock market values over one year; calculate the dividend and the bonds coupon received in respect of the year; construct the yield curve over one year; actualize the bonds market values; calculate the carrying amortizing amount of bonds over one year.

3. Determine death benefits and new premiums received during the past year for each point model. The death rates correspond to those of the *TF0002*

and *TH0002* mortality tables. Benefits are assumed to be paid in the middle of the year. They are revalued at the rate used for the past year for half a year. Moreover, the structural and conjunctural surrenders are evaluated during the year. Conjunctural surrender rates are valued as a function of the spread between the rate used in the previous year and the 10-year rate.

4. Periodic premiums received are invested in assets. Then, assets are eventually sold to pay for benefits. After that, we recalculate the unrealized losses associated with non-bond assets on which the PRE is acquired or taken over.

5. The minimum and maximum available resources are calculated. They can be reached by playing on the achievement of the unrealized profits and losses, and of the resumption of PPE. The expected wealth is determined according to the performance of the financial markets. Finally, one determines the rate used for each point model, considering the different MGRs.

6. The distinct items of the closing balance sheet for year N are calculated by revalorizing the Mathematical Provisions at the rate used as determined in the previous step and the various balance sheet items (basic own funds, capitalization reserve, PPE, PRE).

At the end of the trajectory, that is over 50 years, the assets are settled, the balance sheet is updated and the balances of mathematical provisions and PPE are distributed to policyholders.

3.5.2 Analysis of the loss functions

In this section, we present the results carried out to demonstrate the performance of the standalone as well as the excess loss functions fitting. As can be seen in Section (3.5.1), interest rate risk exists for all assets and liabilities for which the net asset value is sensitive to changes in the term structure of interest rates or interest rate volatility. In the standard formula, the calculations of capital requirements in the interest rate risk module are based on specified scenarios which are defined by a downward and upward stress of the term structure of interest rates. Inspired by this idea, we restrict ourselves solely to consider the risk related to the level of the initial yield curve.

To build stress scenarios, we apply a principal component analysis (explaining 98% of the variability of the annual percentage interest rate change in each of the maturities in the underlying datasets) of historical term structure data from the years $2007 - 2017$. As a result, the yield curve can be approximated as

$$R(0,T) \approx \alpha PC_1(0,T) + \beta PC_2(0,T) + \gamma PC_3(0,T)$$

where PC_1, PC_2, PC_3 are the first three forward curve loadings or principal component vectors (see Section 3.4.3 in [36] for more details). The PC_1 represents the situation that all forward rates in the yield curve move in the same direction. This

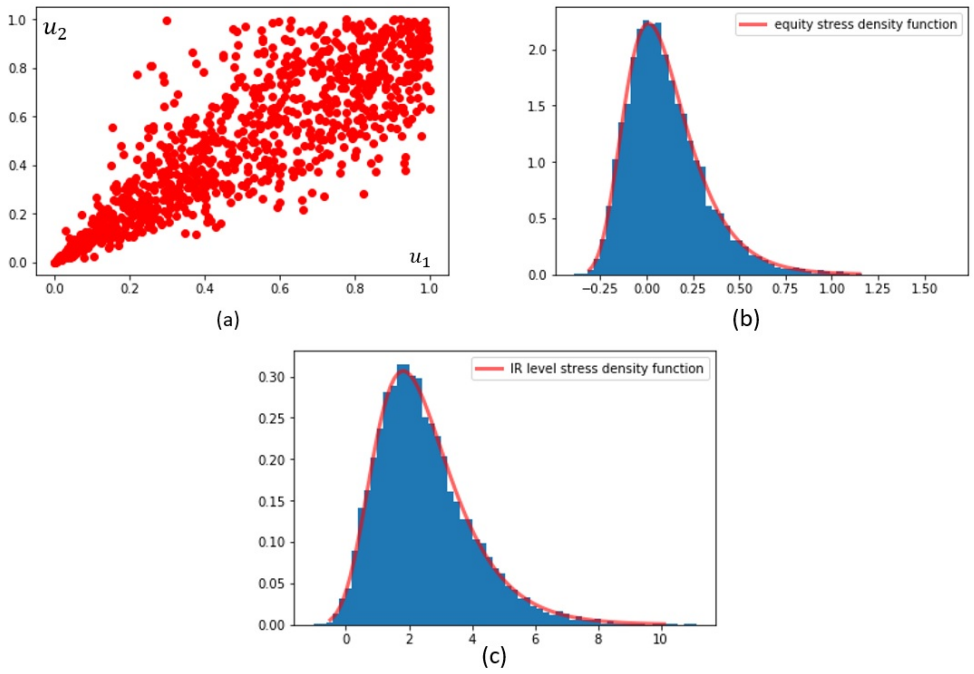


Figure 3.3: (a) Plot of the Clayton copula based on a sample size 10000 in two dimensions with parameter $\theta = 4$, (b) and (c) Plot of the Equity stress and Interest Rate stress probability density functions.

corresponds to a general rise (or fall) of all of the forward rates in the yield curve. In this study, we define the Interest Rate level risk as the shock on the coefficient α as follows $\alpha \to \alpha + x_2 (\alpha_M - \alpha_m)/7$ with α_M, α_m the maximum and minimum value of α observed during the period $2007 - 2017$.

In this study, we consider only two underlying risk factors which are equity risk and Interest rates level risk. As explained in Section 3.3.3, this method of estimating SCR requires the generation of stress realizations. The stresses are made at $t = 0^+$ and there is thus no projection of risk factors in the Economic Scenarios Generator. In order to keep the correlation between the risk factors, it is necessary to assume a model of dependence between them. In our setting, we assume that the stresses follow respectively the Gumbel distribution with the location $\mu_1 = 0.01$ and the scale paramter $\beta_1 = 0.165$ for the Equity risk, and $\mu_2 = 1.82$, $\beta_2 = 1.2$ for the Interest Rate risk. The model of dependence chosen for the realization of market stresses is that of Clayton copula with $\theta = 4$, exhibiting greater dependence in the extremely negative market situation. For the Clayton copula, we draw variates (u_1, u_2) using the conditional distribution approach. Namely, we draw two independent uniform random variables (u_1, v_2) and set $u_2 = \left[u_1^{-\theta}(v_2^{-\theta/(1+\theta)} - 1) + 1\right]^{-1/\theta}$. The following realization $F^{-1}_{\text{Gumbel}}(u_i; \mu_i, \beta_i)\,|_{i=1,2}$ then corresponds to the realization of a random variable having the corresponding Gumbel distribution.

3.5.2.1 Standalone loss function fitting

For each variable considered, we take 25 points within the 99.5th-percentile region and 25 points elsewhere. One possibility of selecting design points within the 99.5th-percentile region is to choose only comparable points. For example, one usually chooses predefined percentiles such as the 0.5th-percentile and/or the 2th-percentile. The 0.5th-percentile is of a specific interest since it gives the required capital under the assumption there is only one risk. In this paper, our choice of fitting points relies on following approach: the fitting points inside and outside of the 99.5th-percentile region are selected randomly and uniformly. Each response variables Y is empirically evaluated by $N_2 = 40$ inner scenarios. The most simple and straightforward "equally spaced" knot placement method is used inside and outside of the 99.5th-percentile zone. For the natural cubic spline space, the usual choice of the number of knots is $K = [n^{\frac{1}{5}}]$, where n is the number of observations. As a result of our hypothesis about stress realizations and the interactions between them, the 99.5th-percentile region for the Equity risk corresponds to the interval $[-0.3, -0.12]$. When running the nested simulations (see Section 3.5.3), we realize that the tail scenarios correspond to the lowest values for interest rates. Therefore, for the sake of simplicity, we only pick the points from the extreme left and the 99.5th-percent region for interest rates is associated to the interval $[-0.49, 0.82]$.

In the rest of this section, in order to the determine the optimal λ_{Eq} and λ_{IR} ridge parameters, it is recommended to perform the 10-fold cross validation as described in [40][3]. Figure 3.5 illustrates our 10-fold cross-validation for λ_{Eq} and λ_{IR} selection. Finally, the estimated standard deviation $\hat{\sigma}_{Eq}$ is found around 7.64×10^{-2} and the resulting optimal ridge parameter $\hat{\lambda}_{Eq}$ is determined to be 3.32 (see Figure 3.5). The fitting of Equity exposed loss function is presented in Figure 3.4.

Figure 3.6 presents the fitting of the Interest rate level exposed loss function. Similarly, the estimated standard deviation $\hat{\sigma}_{IR}$ is 8.02×10^{-2} and the optimal ridge parameter $\hat{\lambda}_{IR}$ turns out to be 7.39 (see Figure 3.5).

For comparison, we calculate the standalone loss functions with 10000 inner simulations at fewer fitting points and apply the Natural Cubic Spline (NCS) interpolation method to reconstruct the curve (Curve Fitting). This latter one can be then considered as the target function. On the other hand, we compare the fitting quality of our method with the Least Squares Monte-Carlo (LSMC) fitting method as described. The Hermite polynomials are chosen as regression basic functions. Regarding the LSMC regression, it is critical to have a reliable data-dependent rule

[3] In k-fold cross-validation, we partition a dataset S into k equally sized non-overlapping subsets S_i. For each fold S_i, a model is trained on $S \backslash S_i$ and is then evaluated on S_i. The cross-validation estimator of the mean squared prediction error is defined as the average of the mean squared prediction errors obtained on each fold. There is however overlap between the training sets for all $k > 2$ and the overlap is largest for *leave-one-out cross validation*. This means that the learned models are correlated implying the increasing amount of variance in the mean squared prediction error estimation. Furthermore, while two-fold cross validation does not have the problem of overlapping training sets, it also has large variance since the training sets are only half the size of the original sample. Therefore, a good compromise is usually 10-fold cross-validation (see, for instance, [13]).

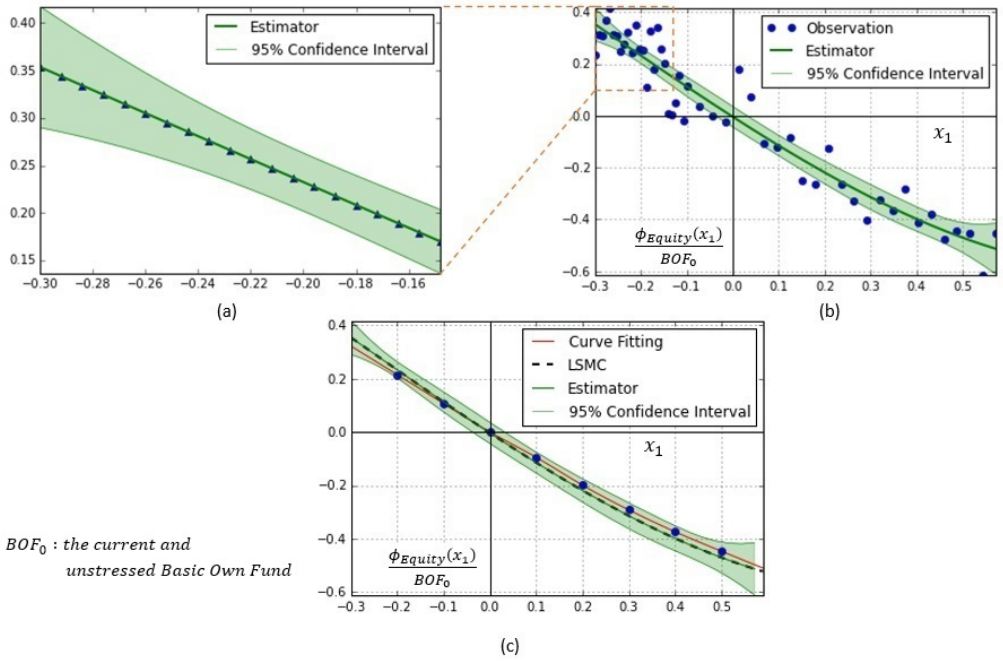


Figure 3.4: (a) Plot of the estimation of the "normalized" equity exposed loss function $\frac{\phi_{\text{equity}}}{\text{BOF}_0}$ and its corresponding 95% Bayesian asymptotic confidence interval within the 99.5th-percentile region. (b) Plot of the estimation of the "normalized" equity exposed loss function and its corresponding 95% Bayesian asymptotic confidence interval for the whole range of market stress in equity X_1. (c) A comparison between the Bayesian penalized spline regression, the LSMC fitting and the curve fitting wherein each fitting point is evaluated by 10000 inner simulations.

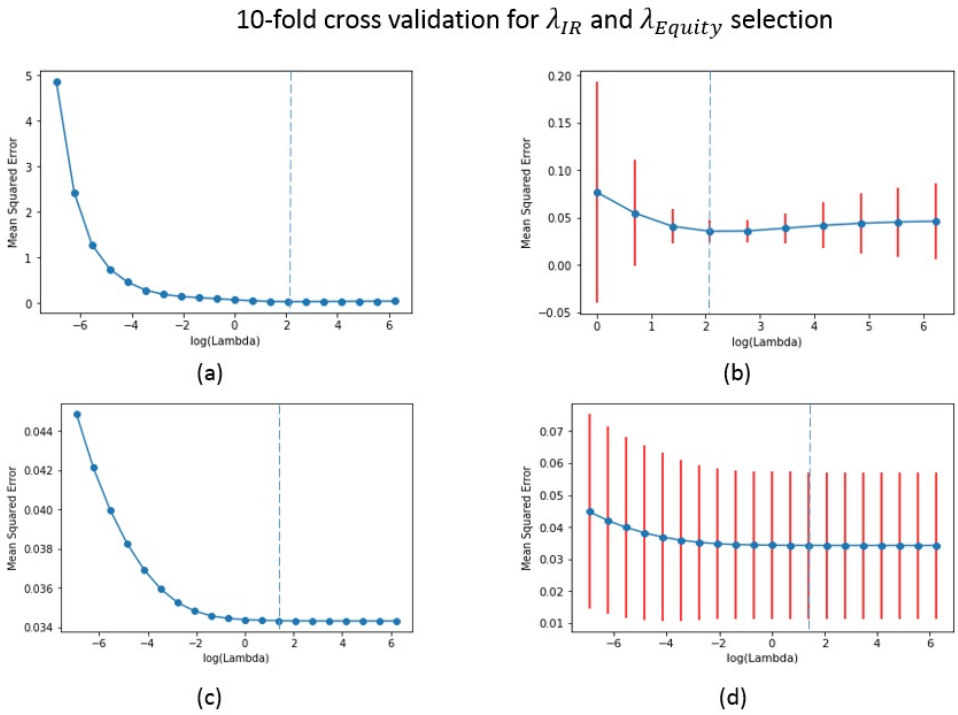


Figure 3.5: Ten-fold cross validation errors or mean squared prediction errors with error bars across different values of $\log(\lambda)$ for: (5.a, 5.b) λ_{IR} in the Interest Rate exposed loss function regression model and (5.c, 5.d) λ_{Eq} in the Equity exposed loss function regression model. The blue dashed lines show the resulting optimal ridge parameters.

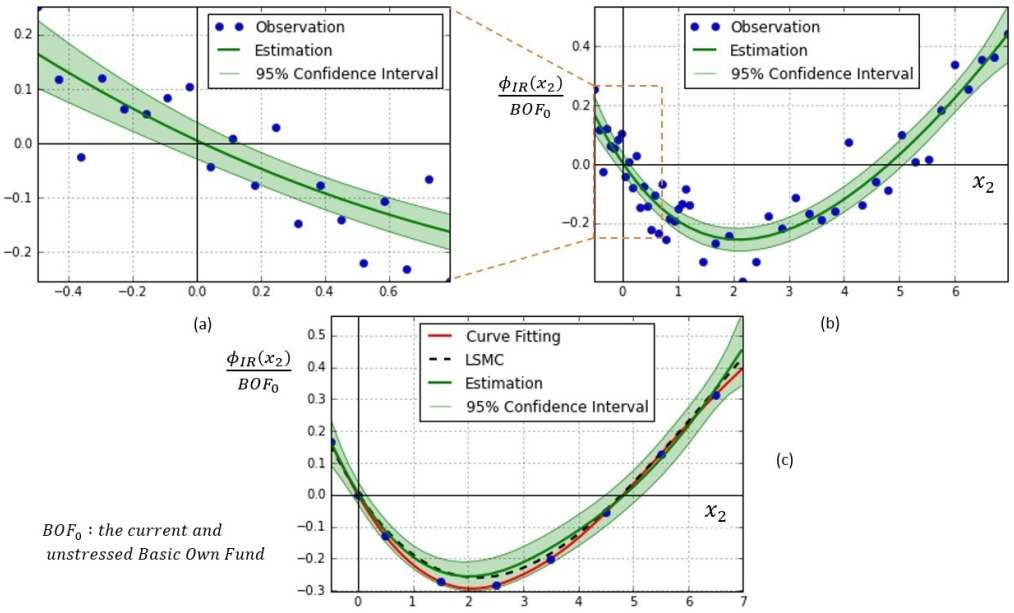


Figure 3.6: (a) Plot of the estimation of the "normalized" interest rate level exposed loss function $\frac{\phi_{\text{IR}}}{\text{BOF}_0}$ and its corresponding 95% Bayesian asymptotic confidence interval within the 99.5th-percentile region. (b) Plot of the estimation of the "normalized" interest rate level exposed loss function and its corresponding 95% Bayesian asymptotic confidence interval for the whole range of market stress in Interest Rate level x_2. (c) A comparison between the penalized spline regression, the LSMC fitting and the curve fitting wherein each fitting point is evaluated by 10000 inner simulations.

deg	CV
1	7.02×10^{-3}
2	6.84×10^{-3}
3	7.19×10^{-3}
4	7.66×10^{-3}
5	8.03×10^{-3}
6	10.41×10^{-3}

(a)

deg	CV
1	33.1×10^{-3}
2	6.99×10^{-3}
3	6.07×10^{-3}
4	6.21×10^{-3}
5	6.58×10^{-3}
6	6.78×10^{-3}
7	7.8×10^{-3}

(b)

Figure 3.7: Cross validation approach to the selection of the degree of the fitting polynomial selection: a) for the Equity exposed loss function, b) for the Interest rate level exposed loss function.

for degree of the fitting polynomial selection. Once again we rely on the 10-fold cross validation technique mentioned previously to select the optimal fitting degree $\deg_{\text{optimal}}$, which corresponds to the best bias-variance tradeoff [4]. A numerical study is carried out and it is easily seen that the best-fitting degree of polynomial equals 3 for the Interest rate level exposed loss function and equals 2 for the equity loss function (see Figure 3.7). Figures (3.4c) and (3.6c) show that our estimates are consistent with the LSMC and NCS fitting results.

To conclude this section, an interpretation of the standalone loss functions is given below. When the equities market performs well, BOF should be increased. Conversely, a fall in prices may reduce the insurer's own funds. Hence, the equity loss function should be a decreasing function of stock prices. As interest rates rise, the market value of the bond assets that make up the majority of the insurer's portfolio declines and therefore the BOF are expected to decline. However, an exponential increase in BEL dominates that of assets and reduces BOF when the interest rates fall. Hence, the interest-rate loss function should be a concave function of the interest rate.

3.5.2.2 Excess loss function fitting

When dealing with single risk loss functions, the notion of 99.5th-percentile region is straightforward. This is not the case for the excess loss function. The 99.5th-percentile region in case of the excess loss function is the smallest hypercube containing all the tail scenarios of each underlying risk factors. For example, in our setting, the 99.5th-percentile region is the rectangular $[-0.3, -0.12] \times [-0.49, 0.82]$. Below an illustration of the 99.5th-percentile region for $d = 2$ subject to a total of 40 stress points randomly and uniformly selected (Figure 3.8). To better fit the excess loss function outside of the zone, there are 60 additional stress points which

[4] In case of $\deg < \deg_{\text{optimal}}$, one misses the pattern while trying to avoid fitting the noise which leads to underfitting. On the contrary, if $\deg > \deg_{\text{optimal}}$, one tries to fit the noise in addition to the pattern which leads to overfitting.

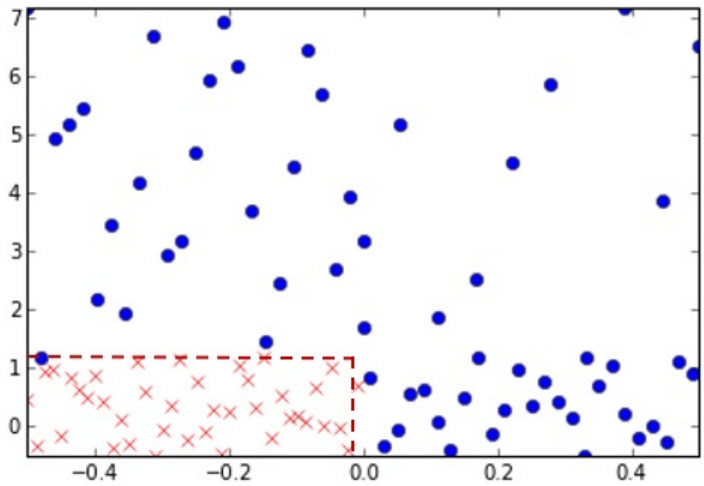


Figure 3.8: An example of the 99.5th-percentile region and the distribution of the fitting points for $d = 2$.

are also randomly and uniformly selected.

Figure 3.9 displays the fitting of three smooth component functions $h_1(X_1)$, $h_2(X_2)$ and $h_{12}(X_1 \times X_2)$ by using the third-degree P-splines. The fitting procedure is completely analogous to that of the standalone loss functions. The three optimal ridge parameters $\lambda_{h_1}, \lambda_{h_2}$ and $\lambda_{h_{12}}$ are found respectively equal to $2.33, 100.87$ and 132.82. It is clear that the practical usefulness of this method depends on its accuracy, which may be assessed via the length of a confidence interval. However, as observed in Figure 3.9 (a) and (b), these confidence intervals are quite wide at the extreme outcomes, which are most relevant. This is due to the boundary effect where the estimator does not feel the boundary, and penalizes for the lack of data beyond the boundary.

For comparison, we demonstrate the analysis for the proposed fitting method and the LSMC method over 12 distinguish stress point within the 99.5th-percentile region. To this end, we repeat the estimation process multiple times with different random-states. Then we compute the mean squared errors between the estimating excess losses and its corresponding target values evaluated by performing the Monte-Carlo simulation with 10000 inner scenarios. The same process is performed with the LSMC method. Overall results (Table 3.1) shows that the LSMC estimating method achieve slightly better convergence rate and higher efficiency than the proposed method. We suspect that the less efficient estimating performance of our approach is due to the fact that parametric models usually provide nice convergence rates of the estimators. However, the discrepancy is relative small ensuring good behavior of the estimators, except for the points $(-0.21, -0.46)$ and $(-0.3, -0.1)$ which are close to the 99.5th-percentile region and suffer thus drawback concerning the boundary effect. The last column in Table (3.1) shows the empirical probability that the estimating asymptotic confidence interval covers the target value. It is

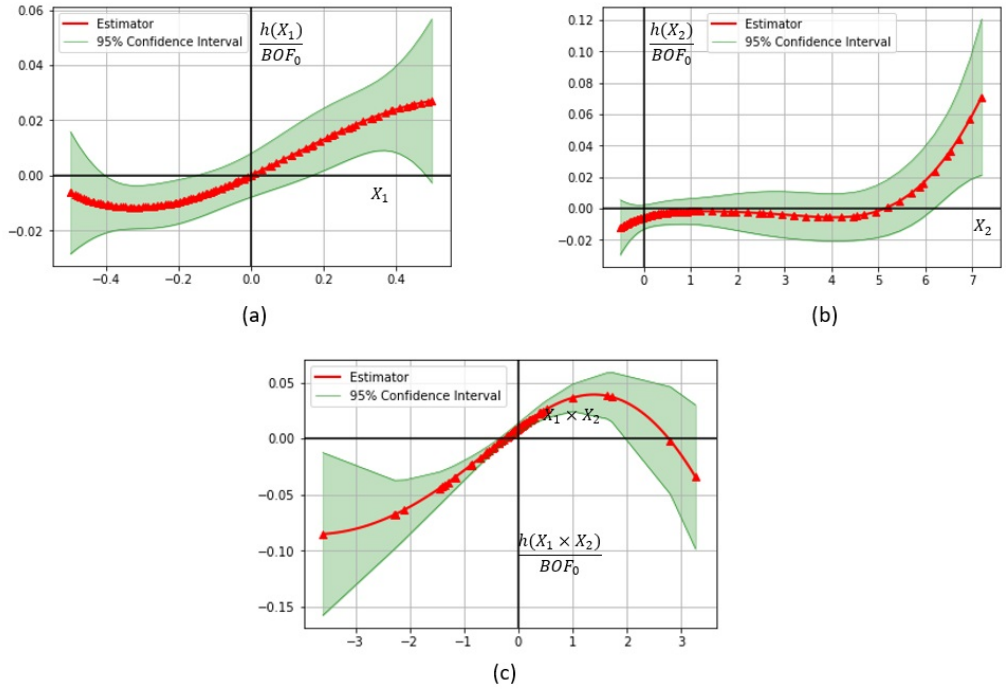


Figure 3.9: Three smooth functional components $h_1(X_1)$, $h_2(X_2)$ and $h_{12}(X_1 \times X_2)$ obtained by fitting with the third-degree P-splines and its corresponding 95% Bayesian asymptotic confidence intervals.

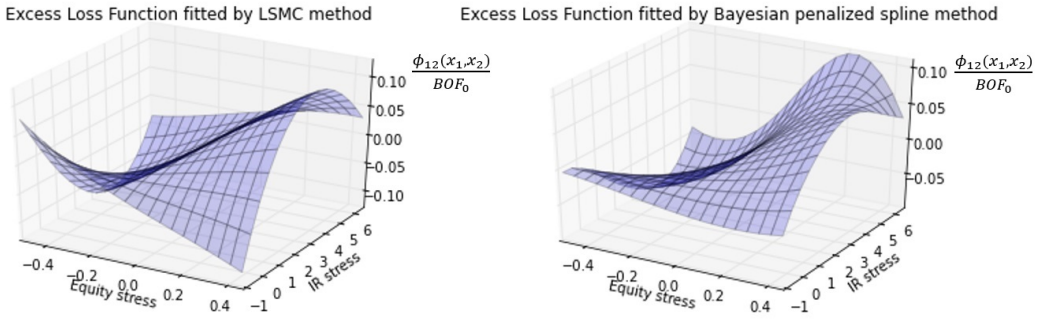


Figure 3.10: *Left*: Plot of the resulting "normalized" Excess Loss function fitted by the Least Square Monte-Carlo method and *Right*: Plot of the resulting "normalized" Excess Loss function fitted by the Bayesian penalized spline method.

Table 3.1: Empirical mean squared errors ($\times 10^{-5}$) of different estimators evaluated by the proposed fitting method and the LSMC method. The last column shows the empirical probability that the estimating asymptotic confidence interval covers the target value. Here we denote by $\delta\widehat{\phi} = \widehat{\phi}^{Spline} - \widehat{\phi}^{MC}$ the deviation from the estimator $\widehat{\phi}^{Spline}$ to the target value $\widehat{\phi}^{MC}$, m the number of repeated calibration process with different random states. In our case, we choose $m = 100$.

(X_1, X_2)	$\widehat{\mathcal{E}}^{LSMC}$	$\widehat{\mathcal{E}}^{Bayesian}$	$\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(\mid \delta\widehat{\phi}(\cdot) \mid \leq \Delta_1^{(\cdot)} + \Delta_2^{(\cdot)} + \Delta_{12}^{(\cdot)}\right)$
(-0.24,-0.11)	3.47	12.45	100%
(-0.26,-0.24)	4.44	17,11	100%
(-0.25,0.76)	3.88	13.38	100%
(-0.16,0.32)	2.83	5.23	100%
(-0.14,-0.2)	4.29	16.24	100%
(-0.2, 0.24)	2.81	9.28	100%
(-0.17, 0.8)	4.05	10.37	100%
(-0.12,0.54)	3.12	15.2	100%
(-0.3, -0.1)	3.62	23.54	100%
(-0.21,-0.46)	6.61	45.22	86.23%
(-0.22,0.28)	2.84	4.78	100%
(-0.29,0.37)	2.85	4.51	100%

easily noted that almost all the interior points are well estimated. However, there is an unexpected situation for the point $(-0.21, -0.46)$ exhibiting poor performance since it is subject to boundary effect.

3.5.3 Nested Simulations

We are ultimately interested in calculating the SCR estimated by Nested Simulations method. It is of great important to know with accuracy how well the proposed fitting method works for estimating the capital requirement.

To that end, we need:

1. A set of real world economic scenarios consistent with the market stresses used to estimate the loss function.

2. For each of these real world scenarios, a set of risk neutral scenarios are generated. The number of inner loops can vary from a scenario to another. Especially increasing the number of inner loop in the tail of the distribution increases the accuracy of the estimators calculated.

3. The total number of outer scenarios is equal to 10000. These outer scenarios are composed of: 7000 scenarios selected from the 40th-percentile to the 95th-percentile of each market stress, 3000 tail scenarios in order to estimate properly $\hat{q}_{0.5\%}(BOF_{t=1})$ the 0.5%-quantile of $\text{BOF}_{t=1}$.

4. Considering the available budget of calculation time, the number of inner simulation per outer simulation is fixed as followed to optimize the information: 20 for the scenarios 1 to 7000 and 60 for the tail scenarios.

In Figure (3.11.a), we plot the economic Own Funds distribution at $t = 1$. From this, we derive the empirical estimation of $\hat{q}_{0.5\%}(BOF_{t=1}) = 54690.16$ and the empirical SCR estimated by the Nested Simulations method is thus equal to $\widehat{SCR}_{NS} = 33743.83$. In another simulation (see Figure (3.11.b)) of 100 different samples of the same size from the same two distributions of the equity risk and the interest rate level risk, we observe that the outcomes are skewed and the estimated values of SCR distribute close to the "true" SCR.

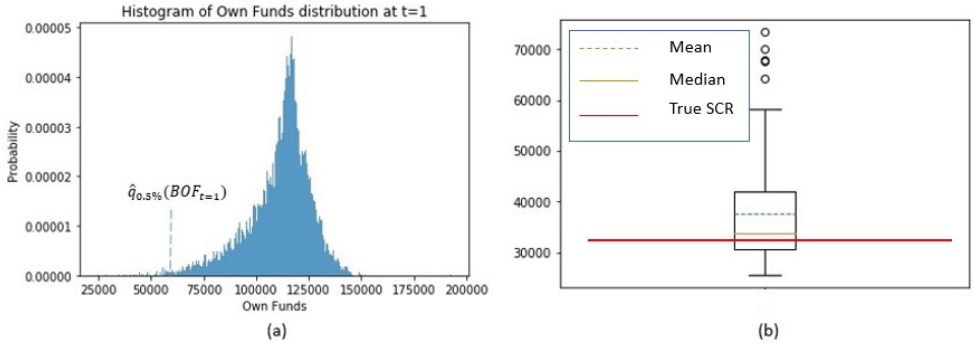


Figure 3.11: (a) Histogram of Own Funds distribution at $t = 1$. (b) Box-whisker plot of the $\widehat{SCR}$ estimated with 100 different samples of the same size.

Chapter 4

Sparse group lasso additive modeling for Pareto-type distributions

Contents

In extreme value statistics, estimation of the tail-index is of importance in numerous applications since it measures the tail heaviness of a distribution. Examples include heavy rainfalls, big financial losses, high medical costs, just to name a few. When covariate information is available, we are mainly interested in describing the tail heaviness of the conditional distribution of the dependent variable given the explanatory variables and the tail-index will be thus taken as a function of this covariate information. In many practical applications, the explanatory variables can contain hundreds of dimensions. Many recent methods use concepts of proximity in order to estimate model parameters based on their relation to the rest of the data. However, in high dimensional space, the data is often sparse and the notion of proximity fails to retain its meaningfulness. Therefore, this implies deterioration in estimation. Having this problematic, we aim to overcome this challenge in the context of the tail-index estimation given the explanatory variables.

This chapter consists of two parts: The first part presents an overarching introduction of extreme values theory, which is served as base for our proposed methodology. Then the mechanism of this methodology will be detailed in the second part.

4.1 Part I - Overview of Extreme Values Theory

Nowadays, modeling extreme events (hurricane, earthquake, floods, financial crises, oil shocks, etc.) is a particularly active research field. In recent years, there has been a growing interest in the application of Extreme Values Theory (EVT) for modeling such events.

Predict certain events or behaviors from the study of extreme values of a sequence, is therefore one of the main goals for those trying to apply EVT. This theory emerged between 1920 and 1940, thanks to Frechet, Fisher and Tippett, Gumbel and Gnedenko. When modeling the maximum of a set of random variables, then, under certain conditions that we will specify later, its distribution can only belong to one of the three following laws: Weibull (with bounded support), Gumbel (with unbounded support and with fine tails) and Fréchet (with unbounded support and thick tails). These three laws define a family of statistical distributions called "generalized extreme value distribution", whose applications are innumerable and very diverse. We will limit ourselves in this report to related insurance risks.

The extreme values theory makes it possible to evaluate the rare events and the losses associated with their appearance. In other words, when a significant loss occurs, this theory makes it possible to evaluate its magnitude. Moreover, this theory plays a particularly important role since it is directly interested in the tail of the distribution. In fact, only the extreme data are used to estimate the parameters of the EVT models which ensures a better fit of the model to the tail of the distribution and therefore a better estimate of the Value-at-Risk (VaR). VaR is a concept commonly used to measure the market risk of a portfolio of financial instruments. It corresponds to the amount of losses that should only be exceeded with a given probability on a given time horizon.

In the following, the material is for the most part borrowed from [27]. Therefore, proofs are most often not given and readers are rather referred to the above references for detailed proofs.

4.1.1 Generalized extreme value distribution

The extreme values theory aims to study the law of the maximum of a sequence of real random variables even if, and especially if, the law of the phenomenon is unknown. Formally, let us consider $(X_1, \ldots, X_n)$ a sequence of n independent and identically distributed random variables of distribution function F_X.

To study the behavior of extreme events, let us consider the random variable $M_n = \max(X_1, \ldots, X_n)$. Since the random variables are independent and identically

distributed, then the distribution function of M_n is given by:

$$F_{M_n}(x) = \mathbb{P}(M_n \leq x) = \mathbb{P}(X_1 \leq x; \ldots; X_n \leq x) = \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = F_X^n(x). \quad (4.1)$$

Equation (4.1) is of very limited interest. Moreover, the law of a random variable X is rarely known precisely and, even if the law of this random variable is known exactly, the law of the maximum term is not always easily calculable. For these reasons, it is interesting to consider the asymptotic behaviors of the appropriately standardized maximum.

Definition: *We say that two real random variables X and Y are of the same type if there are two real constants a and b such that Y and $aX + b$ follow the same law of distribution.*

In a similar way to the central limit theorem, can we find normalization constants $a_n > 0$ and b_n and a non-degenerate law H such that:

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) = F_X^n(a_n x + b) \to H(x) \quad (4.2)$$

as $n \to \infty$?

Fisher and Tippett [37] find a solution to this problem by means of a theorem which bears their name and which is one of the foundations of the theory of extreme values.

Theorem 1 (Fisher - Tippett Theorem)**.** *If there are two sequences of normalization constants with $(a_n) > 0$ and $(b_n) \in \mathbb{R}$ and a non degenerate law of distribution H such that*

$$\lim_{n\to\infty} F_{\frac{M_n - b_n}{a_n}}(x) = H(x),$$

then $H(x)$ is one of these three limits:

- *Gumbel distribution:* $G_0(x) = \exp(-\exp(-x))$, $x \in \mathbb{R}$
- *Fréchet distribution:* $\Phi_\alpha(x) = \exp(-x^{-\alpha})\,\mathbb{I}_{(0,\infty)}(x)$
- *Gumbel distribution:* $\Psi_\alpha(x) = \mathbb{I}_{[0,\infty)}(x) + \exp(-(-x)^{\alpha})\,\mathbb{I}_{(-\infty,0)}(x)$.

Although the behavior of these laws of distribution is completely different, they can be combined in a single parametrization containing a single parameter that controls the thickness of the tail of distribution, which is called the tail-index of extreme values:

$$H_\gamma(x) = \begin{cases} \exp\left(-(1+\gamma x)^{-\frac{1}{\gamma}}\right), & \text{if } \gamma \neq 0;\ \ 1+\gamma x > 0 \\ \exp(-\exp(-x)), & \text{if } \gamma = 0 \end{cases}$$

where H is a non-degenerate function. This law of distribution is called the generalized extreme values distribution (GEV). By introducing the location parameters

μ and the dispersion σ into the parameterization, we obtain the most general form of the generalized extreme value distribution (GEV):

$$H_{\gamma,\mu,\sigma}(x) = \exp\left(-\left(1+\gamma\frac{x-\mu}{\sigma}\right)^{-\frac{1}{\gamma}}\right), \quad \gamma \neq 0, \quad 1+\gamma\frac{x-\mu}{\sigma} > 0 \tag{4.3}$$

where γ is the shape parameter.

The Fisher-Tippett Theorem provides the counterpart of the Central Limit Theorem (CLT) in the case of extreme events. However, unlike the CLT, where the normal distribution is the only possible limiting distribution. In the case of extremes, three types of limiting distribution are possible:

- Gumbel distribution: $\gamma = 0$,
- Fréchet distribution: $\gamma > 0$ corresponds to the Fréchet parameter $\alpha = \frac{1}{\gamma}$,
- Weibull distribution: $\gamma < 0$ corresponds to the Weibull parameter $\alpha = -\frac{1}{\gamma}$.

The GEV-based approach has been criticized as the use of a single maxima leads to a loss of information contained in the other large values of the sample. To overcome this problem, the Peak-over-Threshold method (POT) was introduced in Pickands [89].

4.1.2 Peak-over-threshold method

The Peak-Over-Threshold (POT) method is based on the behavior of observed values beyond a given threshold. In other words, it consists in observing not the maximum or the greatest values but all the values of the realizations which exceed a certain high threshold. The basic idea of this approach is to choose a sufficiently high threshold and study the excesses beyond this threshold.

We define a threshold $u \in \mathbb{R}$, $N_u = \text{card}\{i : i = 1, \ldots, n, X_i > u\}$, and $Y_j = X_i - u > 0$ for $0 \leq j \leq N_u$ where N_u is the number of exceedances over the threshold u by the $\{X_i\}_{i=1,\ldots,n}$ and $\{Y_j\}_{j=1,\ldots,N_u}$ are the corresponding excesses.

We seek from the distribution F_X to define a conditional distribution F_u with respect to the threshold u for the random variables exceeding this threshold. We then define the conditional law of excess F_u by:

$$F_u(y) = \mathbb{P}(X - u \leq y \mid W > u) = \frac{F_X(y+u) - F_X(u)}{1 - F_X(u)}$$

The Pickands-Balkema-de Haan theorem [6,89] below gives the form of the limiting distribution for extreme values: under certain convergence conditions, the limiting distribution is a generalized Pareto distribution that we note GPD.

Theorem 2 (Pickands-Balkema-de Haan theorem)**.** *A distribution function F belongs to the maximum domain of attraction of H_γ if and only if, there exists a positive function $\sigma(u)$ such that:*

$$\lim_{u \to x_F} \sup_{0 \leq y \leq x_F - u} \left|F_u(y) - G_{\gamma,\sigma(u)}(y)\right| = 0 \tag{4.4}$$

where F_u is the conditional distribution function of the excesses for the threshold u, x_F is the end point of F_X, $x_F = \sup\{x \in \mathbb{R} : F_X(x) < 1\}$ and $G_{\gamma,\sigma(u)}(y)$ is the GPD given by:

$$G_{\gamma,\sigma(u)}(y) = \begin{cases} 1 - \left(1 + \gamma \frac{y}{\sigma(u)}\right)^{-\frac{1}{\gamma}}, & \text{if } \gamma \neq 0 \\ 1 - \exp\left(-\frac{y}{\sigma(u)}\right), & \text{if } \gamma = 0 \end{cases} \tag{4.5}$$

with $y \geq 0$ for $\gamma \geq 0$ and $0 \leq y \leq -\frac{\sigma(u)}{\gamma}$ for $\gamma < 0$.

This theorem shows the existence of a close relationship between the GPD and the GEV (Generalized Extreme Value); Pickands [89] has shown that for any distribution F_X, the GPD approximation defined above is verified only if there are normalization constants and a non-degenerate law such as Eq. (4.2) is satisfied. In this case, if H is written in the form of a GEV, then the tail index γ is the same as that of the GPD.

Similarly, for the GPD, the case where $\gamma > 0$ corresponds to the distributions with thick tails, for which $1 - G$ behaves like a power $x^{-\frac{1}{\gamma}}$ for x large enough. If $\gamma = 0$, we have $1 - \exp\left(-\frac{y}{\sigma(u)}\right)$: it is an exponential law of parameter σ and finally $\gamma < 0$, it is the type-II Pareto distribution with bounded support.

The GPD has the following properties:

$$\mathbb{E}(Y) = \frac{\sigma}{1-\gamma}, \ \ \gamma < 1 \tag{4.6}$$

$$V(Y) = \frac{\sigma^2}{(1-\gamma)^2(1-2\gamma)}, \ \ \gamma < \frac{1}{2}. \tag{4.7}$$

In practice, the choice of the threshold constitutes a difficulty. In fact, u must be large enough for the GPD approximation to be valid, but not too high to keep enough overruns to estimate model parameters. The threshold must be chosen so as to make a traditional arbitration between the bias and the variance.

Generally, u is determined graphically by exploiting the linearity of the mean excess function $e(u)$ for the GPD [34]. The function of average excess is given by the relation:

$$e(u) = \mathbb{E}\left(X - u \mid X > u\right) = \frac{\sigma + \gamma u}{1-\gamma}, \ \ \gamma < 0.$$

This technique provides valuable help. However, one should not expect from it the good value of u. In practice, several values of u must be tested. This problem of choice has aroused many works in the literature. Beirlant et al. [10] suggest choosing the threshold u that minimizes the asymptotic mean squared error of the Hill index estimator, while assuming that F_X belongs to Fréchet's maximum attraction domain.

4.1.3 Example of limiting distributions

In this subsection, we propose three examples illustrating how the limit distributions of the GEV and the GPD manifest themselves in practice, taking into account different assumptions about the distribution F_X.

Exponential distribution

For the exponential law of parameter $\lambda = 1$, the distribution function is $F_X(x) = 1 - e^{-x}$ for $x \geq 0$. By posing $b_n = \ln(n)$ and $a_n = 1$ then,

$$F_X^n(a_n x + b) = \left(1 - \frac{e^{-x}}{n}\right)^n \to \exp\left(-e^{-x}\right) = G_0(x) \tag{4.8}$$

as $n \to \infty$.

This shows that the normalized maximum $(M_n - b_n)/a_n$ of the exponential distribution converges to Gumbel distribution. With regard to the POT method, taking $\sigma_u = 1$, then, for all $y > 0$,

$$\begin{aligned} F_u(y) &= \frac{F_X(y+u) - F_X(u)}{1 - F_X(u)} \\ &= 1 - e^{-y}. \end{aligned}$$

Also, the limiting distribution is the GPD of parameter $\gamma = 0$ with $\sigma_u = 1$. Note that in this case, the GPD is not simply the limiting distribution, but it is the exact distribution for every u.

Pareto distribution

For the distribution function $F_X(x) = 1 - cx^{-\alpha}$, where $c > 0$ and $\alpha > 0$. By posing $b_n = 0$ and $a_n = (nc)^{1/\alpha}$ then we have for $x > 0$:

$$F_X^n(a_n x + b) = \left(1 - \frac{x^{-\alpha}}{n}\right)^n \to \exp\left(-x^{-\alpha}\right) = \Phi_\alpha(x) \tag{4.9}$$

which is the Fréchet distribution. Pareto distribution belongs to Fréchet's domains of attraction.

Based on the POT method with the threshold u and considering $\sigma_u = ub$ for $b > 0$, then we have

$$\begin{aligned} F_u(y) &= \frac{F_X(u + uby) - F_X(u)}{1 - F_X(u)} \\ &= 1 - (1 + by)^{-\alpha} \end{aligned}$$

which is the GPD with $\gamma = 1/\alpha$ and $b = \gamma$.

Normal distribution

Let $F_X(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$ be the normal cumulative distribution function. A mathematical result says that $1 - F(x) \sim \frac{1}{x\sqrt{2\pi}e^{-x^2/2}}$ in the neighborhood of $+\infty$, therefore:

$$\lim_{u\to\infty} \frac{1 - F_X\left(u + z/u\right)}{1 - F_X(u)} = \lim_{u\to\infty} \left[\left(1 + \frac{z}{u^2}\right)^{-1} \exp\left(-\frac{1}{2}\left(u + \frac{z}{u}\right)^2 + \frac{1}{2}u^2\right)\right] = e^{-z} \tag{4.10}$$

If we assume at first that $\beta_u = \frac{1}{u}$, then

$$1 - \frac{1 - F_X(u + z/u)}{1 - F_X(u)} = \frac{F_X(u + \beta_u z) - F_X(u)}{1 - F_X(u)} \to 1 - e^{-z}, \quad u \to +\infty$$

and subsequently the limiting distribution of the excesses beyond a threshold u is the exponential distribution.

In a second time if we consider b_n, the solution of the equation $F_X(b_n) = 1 - \frac{1}{n}$, and $a_n = \frac{1}{b_n}$, we obtain

$$n\left[1 - F_X(a_n x + b_n)\right] = \frac{1 - F_X(a_n x + b_n)}{1 - F_X(b_n)} \to e^{-x}.$$

And then

$$\lim_{n\to\infty} F_X^n(a_n x + b_n) = \lim_{n\to\infty} \left(1 + \frac{e^{-x}}{n}\right)^n \to \exp\left(-e^{-x}\right) = G_0(x)$$

4.1.4 Statistical Estimation

Referring to the literature, various methods that have been proposed to estimate the parameters of the GEV and GPD laws are noted, the maximum likelihood method [103], the method of moments [24] the probability weighted moments method [4], or Bayesian methods [79]. There are also nonparametric approaches for estimating the tail index. The most used in practice are the Pickands estimator [89], the Hill estimator [53] (for the case of Frechet type laws only) and the Dekkers-Einmahl De Hann estimator [28]. The most popular method that under certain conditions is the most effective is the maximum likelihood method.

In what follows, we will first present this last parametric estimation method for the GPD. Subsequently, we will present the value-at-risk estimate using an approach based on the Peak Over Thershold (POT) method. And finally, we will present another non-parametric method, the McNeil and Frey [81] model that applies to financial data. For a more complete description, see Embrechts et al. [34].

4.1.4.1 Estimation of GPD parameters by maximum likelihood

Consider again the GPD whose the density function is given by:

$$g(y) = \frac{1}{\sigma}\left(1 + \gamma\frac{y}{\sigma}\right)^{-\frac{1}{\gamma}-1} \tag{4.11}$$

for $y \geq 0$ if $\gamma > 0$, and $0 \leq y \leq -\frac{\sigma}{\xi}$ if $\gamma < 0$.

The estimation of the GPD, by the method of maximum likelihood, relates to the tail index γ as well as the scale parameter σ. The expression of the log-likelihood is therefore

$$l(Y;\gamma,\sigma) = -N_u \ln(\sigma) - \left(1 + \frac{1}{\gamma}\right)\sum_{i=1}^{N_u} \ln\left(1 + \gamma\frac{y_i}{\sigma}\right) \tag{4.12}$$

for a sample of excesses $Y = (y_1, \ldots, y_{N_u})$. From this, by taking the derivatives with respect to each parameter, we obtain the maximum likelihood estimator (MLE), from $\hat{\theta} = (\hat{\gamma}, \hat{\sigma})$.

For $\gamma > -\frac{1}{2}$, Smith [101, 102], Hosking and Wallis [55] prove that the regularity conditions of the likelihood function are fulfilled and the maximum likelihood estimator results in an unbiased, asymptotically normal estimator.

4.1.4.2 Estimate of the Value-at-Risk or the extreme quantile

Recall that the distribution of excesses beyond sufficiently high threshold u is

$$F_u(y) = \mathbb{P}\left(X - u \leq y \mid X > u\right) = \frac{F(u+y) - F(u)}{1 - F(u)} = \frac{\bar{F}(u) - \bar{F}(u+y)}{\bar{F}(u)}, \quad y \geq 0 \tag{4.13}$$

where $\bar{F} = 1 - F$. This can be rewritten as

$$F_u(y)\bar{F}(u) = \bar{F}(u) - \bar{F}(u+y). \tag{4.14}$$

This is equivalent to

$$\bar{F}(u+y) = \bar{F}(u) - F_u(y)\bar{F}(u) = \bar{F}(u)\bar{F}_u(y). \tag{4.15}$$

Thanks to the Pickands-Balkema-de Haan theorem, we have

$$\bar{F}_u(y) \approx \left(1 + \gamma \frac{y}{\sigma(u)}\right)^{-\frac{1}{\gamma}}, \quad y \geq 0 \tag{4.16}$$

as $u \to \infty$. This approximation makes it possible to propose an estimator for $\bar{F}_u(y)$, which is of the form

$$\hat{\bar{F}}_u(y) = \left(1 + \hat{\gamma} \frac{y}{\hat{\sigma}(u)}\right)^{-\frac{1}{\hat{\gamma}}}. \tag{4.17}$$

A natural estimate of $\bar{F}(u)$ is the empirical estimator

$$\hat{\bar{F}}_n(u) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{X_i > u\}} = \frac{N_u}{n} \tag{4.18}$$

where N_u is the number of exceedances.

The estimator results from the tail $\bar{F}(u+y) = \bar{F}(x)$ (for) and therefore has the form:

$$\hat{\bar{F}}(u+y) = \hat{\bar{F}}_n(u)\hat{\bar{F}}_u(y) = \frac{N_u}{n}\left(1 + \hat{\gamma} \frac{y}{\hat{\sigma}(u)}\right)^{-\frac{1}{\hat{\gamma}}}. \tag{4.19}$$

By inverting this equation, we obtain the quantile estimator

$$\hat{x}_p = u + \frac{\hat{\sigma}}{\hat{\gamma}}\left[\left(\frac{n}{N_u}(1-p)\right)^{-\hat{\gamma}} - 1\right] \tag{4.20}$$

for $p > F(u)$. Finally, the Value-at-Risk (VaR) is nothing other than the extreme quantile calculated from the asymptotic extreme distribution (Generalized Pareto Distribution), obtained by modeling extreme losses (or profits) by the POT method.

4.1.5 Characterisation of Maximum Domains of Attraction

Before jumping into the second part of this chapter, let us recall the characterisation of maximum domains of attraction. However, we will mainly focus on the maximum domain of attraction of Fréchet, or MDA(Fréchet), since we are only interested in heavy-detailed distributions. For more information, readers can refer to [95].

The characterisation of maximum domains of attraction actually relies on the theory of regular-varying functions [16]. A positive function U is regularly-varying with index $\delta \in \mathbb{R}$ at infinity if

$$\lim_{x\to\infty} \frac{U(\lambda x)}{U(x)} = \lambda^{\delta} \tag{4.21}$$

for all $\lambda > 0$. This property is denoted by $U \in \mathcal{RV}_{\delta}$.

The following theorem (see Theorem 4 in [42]) tells us how the distribution function F looks like if F belongs to MDA(Fréchet).

Theorem 3. *F belongs to MDA(Fréchet) if and only if $\bar{F} = 1 - F$ is regularly varying with index $-1/\gamma$. The associated extreme-value index is γ. Moreover, a possible choice for the normalizing sequences is $a_n = F^{\leftarrow}(1 - 1/n)$ and $b_n = 0$.*

Let us highlight that necessarily the endpoint of F is infinite. The distribution F is called a Pareto-type distribution if F has the following form:

$$\bar{F}(y) = 1 - F(y) = y^{-\frac{1}{\gamma}} L(y), \quad y > 0 \tag{4.22}$$

for some slowly varying function $L : (0, \infty) \to (0, \infty)$ measurable so that

$$\frac{L(\lambda y)}{L(y)} \to 1 \text{ as } y \to \infty, \forall \lambda > 0.$$

Interestingly, the theorem above shows that the all Pareto-type distributions belong to MDA(Fréchet).

4.2 Part II - Sparse group lasso additive modeling for conditional Pareto-type distributions

In the context of financial and actuarial modeling, the observations very often depend on the other parameters, such as business line, risk profile, seniority, etc. However, all these studies assume that the tail-index is constant regardless of these variables. Many recent studies, for example [23,117], emphasized that the tail-index could be function of these explanatory variables. But none of the previously mentioned studies provide a way to estimate the tail-index parameter conditionally to these variables. As far as we can tell, in the context of financial and actuarial modeling, only three studies have been undertaken to provide methods to estimate the tail-index parameter conditionally to covariates. Beirlant and Goegebeur [9] propose a local polynomial estimator in the case of a one-dimensional covariate. When the

dimension of the covariate increases, this method becomes less effective since the convergence rate of the estimator decreases rapidly. To improve the performance of the estimator, a solution would be to increase the size of data, but this would be problematic in practice since the database could not be easily enlarged. Then, Chavez-Demoulin et al. [22] propose an additive structure with spline smoothing to estimate the relationship between the GDP parameters and covariates. Recently, Heuchenne et al. [52] approach suggests a semi-parametric methodology to estimate the tail-index parameter of a GPD.

In practice, many financial and actuarial data modeling problem may depend upon several explanatory variables, which might make direct tail-index parameter estimation less accurate, or even impossible. One technique to reduce dimension is sparse group lasso, which was introduced by Simon et al. [100]. Motivated both by the advances about the work of Chavez-Demoulin et al. [22] and the sparse group lasso method, we investigate a variable-selecting method to estimate the tail-index parameter conditionally to covariates.

Here is the section layout. We recall first some general results regarding the Peaks-over-Threshold (POT) methodology given covariates, and present the generalized additive model (GAM) in Section (4.2.1). In Section 4.2.1.4, we introduce the sparse group lasso regression and propose a computational algorithm, which is built upon a theoretical property of our statistical model. Finally, we conduct a simulation study to assess the finite sample performance of the proposed method in Section 4.2.2. At the end of this section, we carry out a comparative study with the local polynomial estimation proposed by Beirlant and Goegebeur [9]. Some concluding remarks are made in Section **??**.

4.2.1 Methodology

4.2.1.1 Asymptotic conditional distribution in the POT technique

In this section, we will recall the POT method (see, for instance, [26, 75]) when covariate information is available. Define a set of covariate $\mathcal{X} \subset \mathbb{R}^p$. In this paper, we assume that the design points $\mathbf{x}_i = (x_i^{(1)}, \ldots, x_i^{(p)}) \in \mathcal{X}$ for $i = 1, \ldots, n$ are fixed. Let us consider $(Y_i, \mathbf{x}_i^T)_{1 \leq i \leq n}$ where Y_i is a random variable whose distribution function is of the form $F(y|\mathbf{x}_i) = \mathbb{P}(Y_i \leq y|\mathbf{x}_i)$ of the type (4.22) with some $L(y|\mathbf{x}_i)$. Namely,

$$1 - F(y|\mathbf{x}_i) = y^{-1/\gamma^*(\mathbf{x}_i)} L(y|\mathbf{x}_i). \tag{4.23}$$

Moreover, for some threshold function $u_n(\mathbf{x}_i) > 0$, we define the conditional distribution of $Y_i - u_n(\mathbf{x}_i)$ given $Y_i > u_n(\mathbf{x}_i)$ as follows :

$$F_{u_n(\mathbf{x}_i)}(z|\mathbf{x}_i) = \mathbb{P}(Y_i - u_n(\mathbf{x}_i) \leq z | Y_i \geq u_n(\mathbf{x}_i)) = \frac{F(u_n(\mathbf{x}_i) + z|\mathbf{x}_i) - F(u_n(\mathbf{x}_i)|\mathbf{x}_i)}{1 - F(u_n(\mathbf{x}_i)|\mathbf{x}_i)}.$$

Gnedenko (see Theorem 4 in [42]) showed the equivalent between (4.23) and $F(\cdot|\mathbf{x}_i) \in \mathcal{D}(H_{\gamma^*(\mathbf{x}_i)})$. Then, according to the Pickands theorem [89], we have,

for $\forall i \in \{1, \ldots, n\}$,

$$\lim_{u_n(\mathbf{x}_i)\to\infty} \sup_{0\leq z\leq\infty} |F_{u_n(\mathbf{x}_i)}(z|\mathbf{x}_i) - G\left(z; \gamma^*(\mathbf{x}_i), \sigma^*(\mathbf{x}_i)\right)| = 0 \tag{4.24}$$

where $G(z;\gamma,\sigma)$ is the GPD. This means that, by taking $u_n(\mathbf{x}_i)$ large enough, the distribution of the excesses over $u_n(\mathbf{x}_i)$ is sufficiently close to a GPD with the parameters $\gamma^*(\mathbf{x}_i)$ and $\sigma^*(\mathbf{x}_i)$. Hence, we approximate the condition distribution of $Y_i - u_n(\mathbf{x}_i)$ given $Y_i > u_n(\mathbf{x}_i)$ by a GPD with the parameters $\gamma^*(\mathbf{x}_i)$ and $\sigma^*(\mathbf{x}_i)$ and all observations that exceed a specified high threshold are used to estimate $\gamma^*(\mathbf{x}_i)$. However, since the conditional distribution of $Y_i - u_n(\mathbf{x}_i)$ given $\mathbf{x}_i$ is not exactly a GPD, this consideration will imply a misspecification error in the estimation, which is more difficult to assess.

To be more precise, let us denote by $g(z;\gamma,\sigma)$ the density function of $G(z;\gamma,\sigma)$ being of the form

$$g(z;\gamma,\sigma) = \frac{1}{\sigma}\left(1+\gamma\frac{z}{\sigma}\right)^{-\frac{1}{\gamma}-1},$$

Let us define $M_{\mathbf{x}_i}(\gamma,\sigma) = \mathbb{E}_{(\gamma^*(\mathbf{x}_i),\sigma^*(\mathbf{x}_i))}\left[\log g\left(Z_i;\gamma,\sigma\right)|\mathbf{x}_i\right]$ the minus information cross entropy [98] where Z_i given $\mathbf{x}_i$ exactly follows a GPD with the shape parameters $\gamma^*(\mathbf{x}_i)$ and $\sigma^*(\mathbf{x}_i)$ and $\mathbb{E}_{(\gamma^*(\mathbf{x}_i),\sigma^*(\mathbf{x}_i))}$ denotes the expectation with respect to the true parameters $(\gamma^*(\mathbf{x}_i),\sigma^*(\mathbf{x}_i))$. Clearly, we have

$$(\gamma^*(\mathbf{x}_i),\sigma^*(\mathbf{x}_i)) = \underset{(\gamma,\sigma)\in\mathbb{R}_+^*\times\mathbb{R}_+^*}{\arg\max}\ M_{\mathbf{x}_i}(\gamma,\sigma) \text{ for every } i=1,\ldots,n \tag{4.25}$$

as a result of the Kullback–Leibler divergence [71] between $g(z;\gamma,\sigma)$ and $g(z;\gamma^*(\mathbf{x}_i),\sigma^*(\mathbf{x}_i))$. Following the idea mentioned previously, we define $M_{u_n(\mathbf{x}_i)}(\gamma,\sigma) = \mathbb{E}\left[\log g(Y_i - u_n(\mathbf{x}_i);\gamma,\sigma)|\mathbf{x}_i, Y_i \geq u_n(\mathbf{x}_i)\right]$ the expectation of the approximative log-likelihood $\log g(Y_i - u_n(\mathbf{x}_i);\gamma,\sigma)$ given $Y_i > u_n(\mathbf{x}_i)$. Thanks to the equations (4.24) and (4.25), one can see that $\left(\gamma^*_{u_n(\mathbf{x}_i)}\left(\mathbf{x}_i\right), \sigma^*_{u_n(\mathbf{x}_i)}\left(\mathbf{x}_i\right)\right)$, which are defined by

$$\left(\gamma^*_{u_n(\mathbf{x}_i)}\left(\mathbf{x}_i\right), \sigma^*_{u_n(\mathbf{x}_i)}\left(\mathbf{x}_i\right)\right) = \underset{(\gamma,\sigma)\in\mathbb{R}_+^*\times\mathbb{R}_+^*}{\arg\max}\ M_{u_n(\mathbf{x}_i)}(\gamma,\sigma), \text{ for every } i=1,\ldots,n$$

are the approximations of $(\gamma^*(\mathbf{x}_i),\sigma^*(\mathbf{x}_i))$ for every $i=1,\ldots,n$. In order to obtain the consistency and asymptotic normality of $\left(\gamma^*_{u_n(\mathbf{x}_i)}\left(\mathbf{x}_i\right), \sigma^*_{u_n(\mathbf{x}_i)}\left(\mathbf{x}_i\right)\right)$, we have to impose a further condition on the behavior of the function $L(y|\mathbf{x}_i)$ as follows.

Condition $(\mathcal{S})$: For every $i=1,\ldots,n$, $\frac{L(tz|\mathbf{x}_i)}{L(z|\mathbf{x}_i)} = 1 + \phi(z|\mathbf{x}_i)c(\mathbf{x}_i)\int_1^t s^{\rho(\mathbf{x}_i)-1}ds + o\left(\phi(z|\mathbf{x}_i)\right)$ as $z\to\infty$ for each $t>0$, with $\phi(z|\mathbf{x}_i)>0$ and $\phi(z|\mathbf{x}_i)\to 0$ as $z\to\infty$ and $\rho(z|\mathbf{x}_i)\leq 0$.

The above condition corresponds to the condition C.6 in Beirlant and Goegebeur [9]. Furthermore, this condition is equivalent to the second order condition (see Definition 2.3.1 and Theorem 2.3.9 in [27]). Under the second order condition, de Haan and Ferreira [27] showed that the asymptotic normality of the maximum likelihood estimates holds for $\gamma^*(\mathbf{x}_i) > -\frac{1}{2}$ (please refer to Section 3.4 in [27] for more details).

With a slight abuse of notation, we will use $(\gamma^*(\mathbf{x}), \sigma^*(\mathbf{x}))$ to represent the misspecified shape parameter $(\gamma^*_{u_n(\mathbf{x})}(\mathbf{x}), \sigma^*_{u_n(\mathbf{x})}(\mathbf{x}))$ for the rest of this paper.

4.2.1.2 Generalized Additive Model (GAM)

Recall that our main purpose is to describe the tail-heaviness of the conditional distribution of the dependent variable Y given the predictor $\mathbf{x} \in \mathbb{R}^p$. As consequence, the tail-index is taken as functions of the covariate $\mathbf{x}$. As previously mentioned in the introduction, Beirlant and Goegebeur [9] considered the POT approach and proposed the technique of local polynomial estimation to fit the shape parameter $(\gamma^*(\mathbf{x}), \sigma^*(\mathbf{x}))$ and their corresponding derivatives up to the degree of the chosen polynomial. However, it is difficult to reproduce the forms of $\gamma^*(\mathbf{x})$ and $\sigma^*(\mathbf{x})$ with high-dimensional covariates. This is the so-called "curse of dimensionality" problem, which is due to the fact that data points are isolated in their immensity and the notion of nearest points vanishes with such data. This thus implies the rapid deterioration in convergence rate. On the other hand, the more regular $\gamma^*(\mathbf{x})$ and $\sigma^*(\mathbf{x})$ are, the easier the regression functions are to estimate. The absence of hypothesis on the form of the regression functions leads to a speed of convergence depending on the number of explanatory variables. To overcome this difficulty, we can make stronger assumptions about the form of $\gamma^*(\mathbf{x})$ and $\sigma^*(\mathbf{x})$, which brings us back to the case of parametric models and methods. However, these models lack flexibility for our problem.

Generalized additive models, introduced by Hastie and Tibshirani [49], can compromise the flexibility of non-parametric models and the non-dependence of the speed of convergence of estimators with respect to the number of components of parametric models. Since $\gamma^*(\mathbf{x})$ and $\sigma^*(\mathbf{x})$ are positive functions, we then introduce our generalized additive model as follows:

$$\gamma^{p,\infty}(\mathbf{x}) = \exp\left(\gamma_0 + \sum_{j=1}^{p} \gamma_j(x^{(j)})\right) \tag{4.26}$$

$$\sigma^{p,\infty}(\mathbf{x}) = \exp\left(\sigma_0 + \sum_{j=1}^{p} \sigma_j(x^{(j)})\right) \tag{4.27}$$

where each additive function $\{\gamma_j(\cdot), \sigma_j(\cdot)\}_{j=1}^p$ belongs to the Sobolev space of continuously differentiable functions. In order to ensure the identification we assume that for every $j = 1, \ldots, p$ the additive functions $\{\gamma_j, \sigma_j\}$ are centered, i.e.

$$\sum_{i=1}^{n} \gamma_j\left(x_i^{(j)}\right) = 0, \quad \sum_{i=1}^{n} \sigma_j\left(x_i^{(j)}\right) = 0 \tag{4.28}$$

Supposing that $\log \gamma^*(x)$ and $\log \sigma^*(x)$ are additive will introduce a bias in the estimation, but this assumption is less restrictive than assuming a parametric form on the regression functions, so the modelling error is lower (see, for example [107], for more details about the additive approximation error).

4.2.1.3 Natural cubic splines expansion

The model presented in (4.26) and (4.27) is still nonparametric and the estimation is therefore a problem of infinite dimension. We make it finite by expanding each additive functional components in natural cubic spline (NCS) bases with a reasonable amount of knots K_j for $j = 1, \ldots, p$. Indeed, as pointed in Section 4.3.2, for any regular functions f, we can always find a best spline approximation $\ddot{f}$ of f to minimize $\|\ddot{f} - f\|_\infty$. The error in approximating f by $\ddot{f}$ is usually small, thus in practice we estimate $\tilde{f}$ instead of f. An usual choice would be to use $K_j - 4 \asymp \sqrt{n}$ interior knots. For the sake of simplicity, we consider that every coupled-additive function $\{(\gamma_j(\cdot), \sigma_j(\cdot))\}_{j=1}^p$ will be expanded in the same base. Thus, we parametrize

$$\gamma_j(\cdot) = \sum_{k=2}^{K_j} \theta_{j,k} \left(h_{j,k}(\cdot) - \frac{1}{n} \sum_{i=1}^{n} h_{j,k}\left(x_i^{(j)}\right) \right), \quad \sigma_j(\cdot) = \sum_{k=2}^{K_j} \theta'_{j,k} \left(h_{j,k}(\cdot) - \frac{1}{n} \sum_{i=1}^{n} h_{j,k}\left(x_i^{(j)}\right) \right)$$

where $h_{j,k} : \mathbb{R} \to \mathbb{R}^+$ is the natural cubic spline basis function constructed on the set of the predefined interior knots $\{\xi_1^{(j)}, \ldots, \xi_{K_j}^{(j)}\}$ satisfying $\xi_1^{(j)} \leq \cdots \leq \xi_{K_j}^{(j)}$. Namely, these natural cubic spline basis functions are of the form

$$h_{j,1}(x) = 1, \;\; h_{j,2}(x) = x, \;\; h_{j,k+2}(x) = d_k(x) - d_{K_j}(x) \qquad \forall k = 1, \ldots, K_j - 2$$

with $d_k(x) = \frac{(x-\xi_k)_+^3 - (x-\xi_{K_j})_+^3}{\xi_{K_j} - \xi_k}$. Clearly, this parametrization of the functional components $(\gamma_j(\cdot), \sigma_j(\cdot))$ verifies the centering conditions given in (4.28). To simplify our notation, let us define

$$\tilde{h}_{j,k}(\cdot) = \left(h_{j,k}(\cdot) - \frac{1}{n} \sum_{i=1}^{n} h_{j,k}\left(x_i^{(j)}\right) \right), \;\; \forall j = 1, \ldots, p, \; \forall k = 1, \ldots, K_j.$$

In the following, we denote by β_0 and θ_0 the intercept term instead of γ_0 and σ_0 to synchronize the notation with the coefficients $\theta_{j,k}, \theta'_{j,k}$ as presented previously. Finally, our statistical model is defined as

$$\gamma(\mathbf{x}) = \exp\left(\theta_0 + \sum_{j=1}^{p} \sum_{k=2}^{K_j} \theta_{j,k} \tilde{h}_{j,k}\left(x^{(j)}\right) \right) \tag{4.29}$$

$$\sigma(\mathbf{x}) = \exp\left(\theta'_0 + \sum_{j=1}^{p} \sum_{k=2}^{K_j} \theta'_{j,k} \tilde{h}_{j,k}\left(x^{(j)}\right) \right) \tag{4.30}$$

To sum up, the following diagram sets out the whole approximation scheme:

$$\{\gamma^*(\mathbf{x})\} \xrightarrow{POT} \{\gamma^*_{u_n(\mathbf{x})}(\mathbf{x}), \sigma^*_{u_n(\mathbf{x})}(\mathbf{x})\} \xrightarrow[Approximation]{Additive} \{\gamma^{p,\infty}(\mathbf{x}), \sigma^{p,\infty}(\mathbf{x})\} \xrightarrow[Approximation]{NCS} \{\gamma(\mathbf{x}), \sigma(\mathbf{x})\}$$

4.2.1.4 Sparse group lasso estimation

For notational simplicity, we denote by $\boldsymbol{\varphi} = \left(\theta_0, \boldsymbol{\theta}^T, \theta_0', \boldsymbol{\theta}'^{,T}\right)$ the entire parameter vector where $\boldsymbol{\theta} = \left(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_p^T\right)^T$, $\boldsymbol{\theta}' = \left(\boldsymbol{\theta}_1'^{,T}, \ldots, \boldsymbol{\theta}_p'^{,T}\right)^T$ with $\boldsymbol{\theta}_j = \left(\theta_{j,2}, \ldots, \theta_{j,K_j}\right)^T$ and $\boldsymbol{\theta}_j' = \left(\boldsymbol{\theta}_{j,2}', \ldots, \boldsymbol{\theta}_{j,K_j}'\right)^T$ for every $j = 1, \ldots, p$. This high-dimensional parameter vector carries a group structure where the parameter is partitioned into disjoint pieces. This usually occurs when dealing with expansions in high-dimensional additive models as discussed in Section 4.2.1.2. The goal is high-dimensional estimation in generalized additive models being sparse with respect to whole group. Clearly, the parameter vector $\boldsymbol{\varphi}$ can be structured into groups $\mathcal{G}_0, \mathcal{G}_1, \ldots, \mathcal{G}_p$ and $\tilde{\mathcal{G}}_0, \tilde{\mathcal{G}}_1, \ldots, \tilde{\mathcal{G}}_p$ which build a partition of the index set $\{1, \ldots, 2 + 2\sum_{j=1}^{p}(K_j - 1)\}$. That is,

$$\bigcup_{j=0}^{p}(\mathcal{G}_j \cup \tilde{\mathcal{G}}_j) = \{1, \ldots, 2 + 2\sum_{j=1}^{p}(K_j - 1)\}$$

and the intersection of any distinct groups is an empty set. Each of the groups is defined in the following way:

$$\theta_0 = \boldsymbol{\varphi}_{\mathcal{G}_0}, \;\; \boldsymbol{\theta}_j = \boldsymbol{\varphi}_{\mathcal{G}_j}, \;\; \theta_0' = \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_0}, \;\; \boldsymbol{\theta}_j' = \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}, \qquad \forall j = 1, \ldots, p.$$

Under this notation, the equations (4.29) and (4.30) can be rewritten as

$$\gamma(\mathbf{x}|\boldsymbol{\varphi}) = \exp\left(\sum_{j=0}^{p} \boldsymbol{\varphi}_{\mathcal{G}_j} \tilde{h}_{\mathcal{G}_j}\left(x^{(j)}\right)\right) \tag{4.31}$$

$$\sigma(\mathbf{x}|\boldsymbol{\varphi}) = \exp\left(\sum_{j=0}^{p} \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j} \tilde{h}_{\tilde{\mathcal{G}}_j}\left(x^{(j)}\right)\right) \tag{4.32}$$

with $\tilde{h}_{\mathcal{G}_0}(\cdot) = \tilde{h}_{\tilde{\mathcal{G}}_0}(\cdot) = 1$.

Let $y_i \in \mathbb{R}$ be a realisation of Y_i. We define in the sequel the empirical loss function as follows

$$P_n l(\boldsymbol{\varphi}|u_n(\cdot)) = -\frac{1}{n}\sum_{i=1}^{n} \log g\left(y_i - u_n(\mathbf{x}_i); \gamma(\mathbf{x}_i|\boldsymbol{\varphi}), \sigma(\mathbf{x}_i|\boldsymbol{\varphi})\right) \mathbb{I}(y_i \geq u_n(\mathbf{x}_i)). \tag{4.33}$$

By minimizing this empirical loss function, we could obtain an estimate of the model parameters $\boldsymbol{\varphi}$. However, there are two main reasons why a practitioner is often not satisfied with this estimating approach. The first reason is prediction accuracy: the estimators often have low bias but large variance, especially when $p \gg n$. Shrinking some coefficients to 0 could improve the estimators quality. Indeed, we sacrifice a little bias to reduce the variance of the estimators. This allows to balance the bias-variance trade-off which may improve the overall prediction accuracy. The second reason is related to the interpretation. We prefer to bring out a smaller subset among a large number of predictors that exhibits the strongest effects. In

other words, we would like to identify the principal explanatory variables having the strongest impact on the determination of the tail index parameters. To this end, Yuan and Lin [120] suggested the group lasso penalty for this problem. Moreover, we would like not only sparsity of groups but also within each group. Indeed, there are so many coefficients to calibrate in the model. By doing so it allows to omit negligible coefficients and eliminate perturbative effects. Therefore, we combine both the group lasso criterion and the l_1 penalty proposed by Tibshirani [110].

Namely, for some constants $\lambda_1, \lambda_2, \mu_1, \mu_2 > 0$, defining

$$\text{pen}\,(\boldsymbol{\varphi}|\boldsymbol{\lambda}, \boldsymbol{\mu}) = \lambda_1 \sum_{j=1}^{p} \sqrt{G_j} \|\boldsymbol{\varphi}_{\mathcal{G}_j}\|_2 + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\varphi}_{\mathcal{G}_j}\|_1 + \mu_1 \sum_{j=1}^{p} \sqrt{G_j} \|\boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}\|_2 + \mu_2 \sum_{j=1}^{p} \|\boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}\|_1 \tag{4.34}$$

where $G_j \equiv |\mathcal{G}_j| = |\tilde{\mathcal{G}}_j|$ denotes the cardinality of the group $\mathcal{G}_j$, as well as of the group $\tilde{\mathcal{G}}_j$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$.

The regression model that we consider to estimate $(\gamma(\mathbf{x}|\boldsymbol{\varphi}), \sigma(\mathbf{x}|\boldsymbol{\varphi}))$ is defined by

$$\widehat{\boldsymbol{\varphi}}(u_n(\cdot), \boldsymbol{\lambda}, \boldsymbol{\mu}) = \arg\min_{\boldsymbol{\varphi}} \left\{ P_n l(\boldsymbol{\varphi}|u_n(\cdot)) + \text{pen}\,(\boldsymbol{\varphi}|\boldsymbol{\lambda}, \boldsymbol{\mu}) \right\}. \tag{4.35}$$

Note that this latter one is not exactly the penalized log-likelihood estimation since the true conditional distribution of $Y_i - u_n(\mathbf{x}_i)$ given $Y_i > u_n(\mathbf{x}_i)$ is not a GPD as mentioned in Section 4.2.1.1.

4.2.1.5 Algorithm for the sparse group lasso

In this section, we will use $\widehat{\boldsymbol{\varphi}}$ to designate the estimator $\widehat{\boldsymbol{\varphi}}(u_n(\cdot), \boldsymbol{\lambda}, \boldsymbol{\mu})$ for notational simplicity. Furthermore, for the later use, we need to define the following parameters: $\mathbf{s}_j = \widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j} / \|\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\|_2$ if $\hat{\boldsymbol{\varphi}}_{\mathcal{G}_j} \neq 0$ (i.e. not equal to the 0-vector) and $\mathbf{s}_j$ is a vector satisfying $\|\mathbf{s}_j\|_2 \leq 1$ if $\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j} \equiv 0$, and $t_{j,k} \in \text{sign}((\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j})_k)$ if $(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j})_k \neq 0$ and $t_{j,k} \in [-1, 1]$ otherwise. By interchanging $\mathcal{G}_j$ with $\tilde{\mathcal{G}}_j$, we obtain the similar definition for $\mathbf{u}_j$ and $v_{j,k}$. Besides, we denote by $\widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j}$ the $\widehat{\boldsymbol{\varphi}}$-vector whose components in $\mathcal{G}_j$ are set to zero, by $\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j,-k}$ the $\widehat{\boldsymbol{\varphi}}$-vector where only the kth component in the group $\mathcal{G}_j$ is set to zero.

For later use, let us denote by Φ a nonempty subset of $\mathbb{R}^{2(1+\sum_{j=1}^{p}(K_j-1))}$ containing the optimal vector of model parameters $\widehat{\boldsymbol{\varphi}}(u_n(\cdot), \boldsymbol{\lambda}, \boldsymbol{\mu})$. As a consequence of the Karush-Kuhn-Tucker (KKT) conditions (see, for example, [14]), we have the following result, which is an important characterization of the optimal solution $\widehat{\boldsymbol{\varphi}}$ in (4.35).

Lemma 1. *Assume that $P_n l(\boldsymbol{\varphi})$ is locally convex[1] on Φ. Then, the necessary and sufficient conditions for $\widehat{\boldsymbol{\varphi}}$ to be a solution of (4.35) are*

$$\frac{\partial P_n l(\widehat{\boldsymbol{\varphi}})}{\partial \boldsymbol{\varphi}_{\mathcal{G}_0}} = 0 \tag{4.36}$$

[1]Regarding the definition of a locally convex function and its related details, readers can refer to, for example, [76].

$$\frac{\partial P_n l(\widehat{\boldsymbol{\varphi}})}{\partial \boldsymbol{\varphi}_{\tilde{\mathcal{G}}_0}} = 0 \tag{4.37}$$

$$\left[\nabla P_n l(\widehat{\boldsymbol{\varphi}})_{\mathcal{G}_j}\right]_k + \lambda_1 \sqrt{G_j}(\boldsymbol{s}_j)_k + \lambda_2 t_{j,k} = 0 \tag{4.38}$$

$$\left[\nabla P_n l(\widehat{\boldsymbol{\varphi}})_{\tilde{\mathcal{G}}_j}\right]_k + \mu_1 \sqrt{G_j}(\boldsymbol{u}_j)_k + \mu_2 v_{j,k} = 0 \tag{4.39}$$

for every $j = 1, \ldots, p$ *and* $k = 2, \ldots, K_j$ *where* $\nabla P_n l(\widehat{\boldsymbol{\varphi}})_{\mathcal{G}_j}$ *(respectively for* $\nabla P_n l(\widehat{\boldsymbol{\varphi}})_{\tilde{\mathcal{G}}_j}$*) denotes the gradient vector of* $P_n l(\boldsymbol{\varphi})$ *with respect to* $\boldsymbol{\varphi}_{\mathcal{G}_j}$ *(respectively for* $\boldsymbol{\varphi}_{\tilde{\mathcal{G}}_j}$*) at* $\widehat{\boldsymbol{\varphi}}$*, and* $\boldsymbol{s}_j, \boldsymbol{u}_j, t_{j,k}, v_{j,k}$ *are defined above.*

The proof of this lemma will be given in Appendix 4.3.1. These first derivative tests (4.36 - 4.39) can give insight into the sparsity of groups and within each group. Indeed, the necessary and sufficient condition for $\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j} \equiv 0$ is that the equation $\left[\nabla P_n l(\widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j})_{\mathcal{G}_j}\right]_k + \lambda_1 \sqrt{G_j}(\mathbf{s}_j)_k + \lambda_2 t_{j,k} = 0$ has a solution with $\|\mathbf{s}_j\|_2 \leq 1$ and $t_{j,k} \in [-1, 1]$ for every $k \in \mathcal{G}_j$. To this end, we define $J(\mathbf{t}_j; \widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j}) = \frac{1}{\lambda_1^2 G_j} \sum_{k \in \mathcal{G}_j} \left(\left[\nabla P_n l(\widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j})_{\mathcal{G}_j}\right]_k + \lambda_2 t_{j,k}\right)^2 = \|\mathbf{s}_j\|_2^2$. Let us denote by $\hat{\mathbf{t}}_j$ the minimizer of $J(\mathbf{t}_j, \widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j})$. If $J(\hat{\mathbf{t}}_j, \widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j}) \leq 1$, then $\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j} \equiv 0$. Otherwise, $\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}$ is not identically equal to the 0-vector. Moreover, it is easily seen that the minimizer is of the form:

$$\hat{t}_{j,k} = \begin{cases} -\frac{\left[\nabla P_n l(\hat{\varphi}_{-\mathcal{G}_j})_{\mathcal{G}_j}\right]_k}{\lambda_2}, & \text{if } \left|\frac{\left[\nabla P_n l(\hat{\varphi}_{-\mathcal{G}_j})_{\mathcal{G}_j}\right]_k}{\lambda_2}\right| \leq 1 \\ -\text{sign}\left(\left[\nabla P_n l(\hat{\varphi}_{-\mathcal{G}_j})_{\mathcal{G}_j}\right]_k\right), & \text{otherwise} \end{cases} \tag{4.40}$$

With a little bit of algebra, we can show that $J(\hat{\mathbf{t}}_j, \widehat{\boldsymbol{\varphi}}_{-\mathcal{G}_j}) \leq 1$ is equivalent to

$$\|S\left(\nabla P_n l(\hat{\boldsymbol{\varphi}}_{-\mathcal{G}_j})_{\mathcal{G}_j}, \lambda_2\right)\|_2 \leq \lambda_1 \sqrt{G_j}$$

with $S(\cdot)$ the coordinate-wise soft thresholding operator:

$$\left(S\left(z, \lambda\right)\right)_i = \text{sign}(z_i)\left(|z_i| - \lambda\right)_+, \quad z \in \mathbb{R}^{G_j}, \lambda \in \mathbb{R}_+.$$

If $\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j} \neq \mathbf{0}$, we apply the coordinate descent algorithm to find its element $\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k$. The logic of the coordinate descent procedure is as follows: if $\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k \neq 0$, then the equation $\left[\nabla P_n l(\widehat{\boldsymbol{\varphi}})_{\mathcal{G}_j}\right]_k + \lambda_1 \sqrt{G_j}\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k / \|\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\| + \lambda_2 \text{sign}(\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k) = 0$ must have a solution. This latter one leads to the inequality $|\left(\nabla P_n l(\widehat{\boldsymbol{\varphi}})_{\mathcal{G}_j}\right)_k| > \lambda_2$. This follows easily by examining the case where $\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k$ is strictly positive and negative. Therefore, check if $|\left(\nabla P_n l(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j, -k})_{\mathcal{G}_j}\right)_k| \leq \lambda_2$ and if so set $\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k = 0$. Otherwise, we minimize the equation (4.35) over $(\boldsymbol{\varphi}_{\mathcal{G}_j})_k$ by a one-dimensional optimization to get $\left(\widehat{\boldsymbol{\varphi}}_{\mathcal{G}_j}\right)_k$.

It is natural to think of a generalized gradient descent method to get the optimal solution. This consideration thus leads to the computation presented in 4.3.3. According to this algorithm, the optimal solution can be found by cycling through the groups $\mathcal{G}_0 \to \mathcal{G}_1 \to \cdots \to \tilde{\mathcal{G}}_p \to \mathcal{G}_0$. Within each iterative steps, we optimize the objective function (4.35) by solving the equations (4.36 - 4.39) with respect to the

current group $\mathcal{G}_j$ (or $\tilde{\mathcal{G}}_j$) while keeping all except for the group fixed. This is called the block coordinate descend algorithm, as proposed by Friedman et al. [39].

As presented earlier, we considered the coordinate descent procedure to fit the model within group. As pointed out by Simon et al. [100], this algorithm provides a poor performance in terms of timing and accuracy[2]. To overcome this drawback, they propose a block-wise descent algorithm which makes stride in performance. Inspired by this idea, we extend the fitting algorithm to our statistical model. Since our penalty (4.34) is separable between groups, we will only focus on an arbitrary one, saying $\mathcal{G}_j$ and the algorithm will be applied on the same principal for the others. Therefore, we consider the other group coefficients as fixed and ignore the penalties corresponding to these groups. With a slight abuse of notation, we denote in the following our loss function $P_n l(\boldsymbol{\varphi}_{\mathcal{G}_j})$ taking $\boldsymbol{\varphi}_{\mathcal{G}_j}$ as parameter to minimize over.

We start with the majorization minimization scheme. This means that we majorize the empirical loss function and then minimize the upper bound, together with the penalty. Namely, the empirical loss function is majorized by

$$P_n l\left(\boldsymbol{\varphi}_{\mathcal{G}_j}\right) \leq P_n l\left(\boldsymbol{\varphi}^0_{\mathcal{G}_j}\right) + \left(\boldsymbol{\varphi}_{\mathcal{G}_j} - \boldsymbol{\varphi}^0_{\mathcal{G}_j}\right)^T . \nabla P_n l\left(\boldsymbol{\varphi}^0_{\mathcal{G}_j}\right) + \frac{1}{2t}\|\boldsymbol{\varphi}_{\mathcal{G}_j} - \boldsymbol{\varphi}^0_{\mathcal{G}_j}\|_2^2$$

where $\boldsymbol{\varphi}^0_{\mathcal{G}_j}$ is a vector parameter to be determined at a later point and t is sufficiently small so that the quadratic term dominates the Hessian of the loss function for every $\boldsymbol{\varphi}_{\mathcal{G}_j} \in \Phi_j$ with Φ_j the set of vectors parameter containing the target vector of coefficients.

We will omit the demonstration since it is given in [100]. Finally, we get that if $\|S\left(\boldsymbol{\varphi}^0_{\mathcal{G}_j} - t\nabla P_n l(\boldsymbol{\varphi}^0_{\mathcal{G}_j}), t\lambda_2\right)\|_2 \leq t\lambda_1\sqrt{G_j}$, then $\hat{\boldsymbol{\varphi}}_{\mathcal{G}_j} \equiv 0$. Otherwise,

$$\hat{\boldsymbol{\varphi}}_{\mathcal{G}_j} = \mathcal{F}(\boldsymbol{\varphi}^0_{\mathcal{G}_j}, t) = \left(1 - \frac{t\lambda_1\sqrt{G_j}}{\|S\left(\boldsymbol{\varphi}^0_{\mathcal{G}_j} - t\nabla P_n l(\boldsymbol{\varphi}^0_{\mathcal{G}_j}), t\lambda_2\right)\|_2}\right)_+ S\left(\boldsymbol{\varphi}^0_{\mathcal{G}_j} - t\nabla P_n l(\boldsymbol{\varphi}^0_{\mathcal{G}_j}), t\lambda_2\right)$$

To get the optimal solution, we cyclically iterate the procedure through the blocks. At each iterative step, we update $\left(\boldsymbol{\varphi}^0_{\mathcal{G}_j}\right)^{(m)} = \hat{\boldsymbol{\varphi}}^{(m-1)}_{\mathcal{G}_j}$. By introducing a momentum term in the gradient updates, Nesterov [83] showed that this modification can have a huge improvement in terms of convergence rate. As also suggested by Simon et al. [100], we present here Algorithm 1 for the blockwise descent fitting method.

4.2.1.6 Refitting step

A well-known drawback of l_1-penalized estimators is the systematic shrinkage of the large coefficients towards zero. This may give rise to a high bias in the resulting estimators and may affect the overall conclusion about the model (see, for example,

[2]For the reason mentioned above, we will no longer discuss the performance of the Block Coordinate descent algorithm (or the accelerated generalized gradient descent algorithm) in the rest of this paper. However, interested readers can can refer to Appendix 4.3.3 where the pseudo-code version of this algorithm is provided, in order to facilitate its implementation

Algorithm 1 Block-wise Descent Algorithm

1: Set up with the initial parameter vector $\hat{\boldsymbol{\varphi}}^{(0)}$ and the loop index $m = 0$.

2: Increase m by one: $m \leftarrow m + 1$ and cycle the optimization procedure through the groups:

(2.1) Set $\hat{\boldsymbol{\varphi}}^{(m)} = \hat{\boldsymbol{\varphi}}^{(m-1)}$.

(2.2) Regarding $j = 0$, if $\nabla P_n l(\widehat{\varphi}^{(m)}_{-\mathcal{G}_0})_{\mathcal{G}_0} = 0$: set $\widehat{\varphi}^{(m)}_{\mathcal{G}_0} = \mathbf{0}$, and for $j = 1, \ldots, p$, if $\|S\left(\nabla P_n l(\widehat{\boldsymbol{\varphi}}^{(m)}_{-\mathcal{G}_j})_{\mathcal{G}_j}, \lambda_2\right)\|_2 \leq \lambda_1\sqrt{G_j}$: set $\widehat{\boldsymbol{\varphi}}^{(m)}_{\mathcal{G}_j} = \mathbf{0}$. Otherwise, set counter $l = 1$, step size $t = 1$ and $\widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j} = \boldsymbol{\mu}^{(m,l)}_{\mathcal{G}_j} = \widehat{\boldsymbol{\varphi}}^{(m)}_{\mathcal{G}_j}$ and repeat the following until convergence:

(2.2.1) Update gradient $g = \nabla P_n l\left(\widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j}\right)$.

(2.2.2) Estimate optimal step size by iterating $t \leftarrow 0.8 * t$ until

$$P_n l\left(\mathcal{F}(\widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j}, t)\right) \leq P_n l\left(\widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j}\right) + \left(\Delta^{(m,l)}_t\right)^T . g + \frac{1}{2t}\|\Delta^{(m,l)}_t\|_2^2$$

with $\Delta^{(m,l)}_t = \mathcal{F}(\widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j}, t) - \widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j}$.

(2.2.3) Update $\boldsymbol{\mu}^{(m,l)}_{\mathcal{G}_j}$ by $\boldsymbol{\mu}^{(m,l+1)}_{\mathcal{G}_j} \leftarrow \mathcal{F}(\widehat{\boldsymbol{\varphi}}^{(m,l)}_{\mathcal{G}_j}, t)$.

(2.2.4) Update $\widehat{\boldsymbol{\varphi}}^{(m,l)}$ by

$$\widehat{\boldsymbol{\varphi}}^{(m,l+1)} \leftarrow \boldsymbol{\mu}^{(m,l)}_{\mathcal{G}_j} + \frac{l}{l+3}\left(\boldsymbol{\mu}^{(m,l+1)}_{\mathcal{G}_j} - \boldsymbol{\mu}^{(m,l)}_{\mathcal{G}_j}\right).$$

(2.2.5) Increase l by one: $l \leftarrow l + 1$.

(2.3) Repeat the procedure for the groups $\tilde{\mathcal{G}}_j$ for $j = 0, \ldots, p$.

3: Repeat the entire step (2) until convergence.

[11]). A simple remedy is to treat sparse group lasso as a variable selection tool and to perform a refitting step on the select support. Namely, let us define the active set by

$$\mathcal{S}_{\mathcal{G}} = \{(j,k)|\widehat{\varphi}_{\mathcal{G}_{j,k}} \neq \mathbf{0}\}, \quad \mathcal{S}_{\tilde{\mathcal{G}}} = \{(j,k)|\widehat{\varphi}_{\tilde{\mathcal{G}}_{j,k}} \neq \mathbf{0}\}$$

Next, we define

$$\gamma'(\mathbf{x}|\boldsymbol{\varphi}) = \exp\left(\sum_{(j,k)\in\mathcal{S}_{\mathcal{G}}} \varphi_{\mathcal{G}_{j,k}} \tilde{h}_{\mathcal{G}_{j,k}}\left(x^{(j)}\right)\right) \tag{4.41}$$

$$\sigma'(\mathbf{x}|\boldsymbol{\varphi}) = \exp\left(\sum_{(j,k)\in\mathcal{S}_{\tilde{\mathcal{G}}}} \varphi_{\tilde{\mathcal{G}}_{j,k}} \tilde{h}_{\tilde{\mathcal{G}}_{j,k}}\left(x^{(j)}\right)\right) \tag{4.42}$$

Our refitted estimator is thus the only minimizer of the following equation

$$\hat{\hat{\boldsymbol{\varphi}}} = \arg\min_{\boldsymbol{\varphi}} P_n l'(\boldsymbol{\varphi}|u_n(\cdot)) \tag{4.43}$$

where

$$P_n l'(\boldsymbol{\varphi}|u_n(\cdot)) = -\frac{1}{n}\sum_{i=1}^{n} \log g\left(y_i - u_n(\mathbf{x}_i); \gamma'(\mathbf{x}_i|\boldsymbol{\varphi}), \sigma'(\mathbf{x}_i|\boldsymbol{\varphi})\right) \mathbb{I}\{y_i \geq u_n(\mathbf{x}_i)\}.$$

4.2.2 Simulation Study

How well does the sparse group lasso procedure described above estimate the tail-index function $\gamma^*(\mathbf{x})$? To answer this question, we conduct a small simulation study of the block-wise descent estimator where $\{y_i\}_{i=1}^n$ are generated from the Burr$(\eta, \tau(\mathbf{x}), \xi)$ distribution [21] for which the distribution function is given by

$$F_{\text{Burr}}(y) = 1 - \left(\frac{\eta}{\eta + y^{\tau(\mathbf{x})}}\right)^{\xi}. \tag{4.44}$$

Let us recall the Hall class of Pareto-type distributions [46] which is of the form

$$1 - F(y) = a y^{-\frac{1}{\gamma^*}(\mathbf{x})}\left[1 + b y^{-\theta(\mathbf{x})} + o\left(y^{-\theta(\mathbf{x})}\right)\right].$$

Note that this class of distribution satisfies the condition 1 with $c(\mathbf{x}) = -\theta(\mathbf{x})b$, $\rho(\mathbf{x}) = -\theta(\mathbf{x})$ and $\phi(z|\mathbf{x}) = z^{-\theta(\mathbf{x})}$. It is easily seen that the Burr$(\eta, \tau(\mathbf{x}), \xi)$ distribution belongs to the Hall class of Pareto-type distribution with $\gamma^*(\mathbf{x}) = 1/(\xi\tau(\mathbf{x}))$, $a = \eta^{\xi}$, $b = -\eta\xi$ and $\theta(\mathbf{x}) = \tau(\mathbf{x})$ since its survival function can be written as

$$1 - F_{\text{Burr}}(y) = y^{-\xi\tau(\mathbf{x})}\eta^{\xi}\left(1 - \xi\eta y^{-\tau(\mathbf{x})} + o(y^{-\tau(\mathbf{x})})\right)$$

as $y \to \infty$. Therefore, the condition $(\mathcal{S})$ is satisfied with $c(\mathbf{x}) = \eta\xi\tau(\mathbf{x})$, $\rho(\mathbf{x}) = -\tau(\mathbf{x})$ and $\phi(z|\mathbf{x}) = z^{-\tau(\mathbf{x})}$.

In this simulation study, we consider two sample sizes $n = 500, 5000$ and two p values $p = 2$ and 10. Usually in many high-dimensional studies, the dimension of

the data vectors p is comparable or may be larger than the sample size n. Hence, it is obvious that our setting with $p = 10$ can not be considered as high dimensional covariate. However, we realized that it becomes computationally expensive in terms of running time required to perform estimation when the dimensionality increases. Therefore, in this paper, we limit ourselves to the case $p = 10$. Surprisingly, we note that the proposed methodology slightly outperforms the local polynomial maximum likelihood regression proposed by Beirland and Goegebeur [9] even with $p = 10$.

The data are generated from the Burr$(\eta, \tau(\mathbf{x}), \xi)$ distribution with $\xi = \eta = 1$ and

$$\tau(\mathbf{x}) = \frac{20}{3\left[(x^{(1)})^2 - (x^{(2)})^2 + 4\right]}$$

where $\mathbf{x} = \left(x^{(1)}, x^{(2)}\right)^T$ for $p = 2$ and $\mathbf{x} = \left(x^{(1)}, x^{(2)}, \ldots, x^{(10)}\right)^T$ for $p = 10$. Clearly, there are only two active variables for both cases. From this, it follows that the tail-index function $\gamma^*(\mathbf{x})$ is then given by

$$\gamma^*(\mathbf{x}) = \frac{1}{\xi\tau(\mathbf{x})} = 0.15\left[(x^{(1)})^2 - (x^{(2)})^2 + 4\right]. \tag{4.45}$$

Each explanatory variable $x^{(j)}, j = 1, \ldots, 10$ takes value from the 10-equally spaced samples in the closed interval $[0, 1]$. For each simulated dataset, we apply the proposed methods to estimate $\gamma^*(\mathbf{x})$.

A hurdle in the Peaks-over-threshold approach for analyzing extreme values is the selection of the threshold. The misdetermination of the threshold value will have a non negligible impact on the performance of the estimator. Indeed, threshold selection constitutes a trade-off situation between bias and variance. If we set the threshold value too low, the GPD approximation is not suitable which implies a large bias. On the other hand, if we set the threshold value too high, a small number of observations is used which leads to an increasing variance in the estimated GPD parameters. In the previous section, one considers that threshold $u_n(\mathbf{x})$ depends on both the covariates $\mathbf{x}$ and the sample size n, with $u_n(\mathbf{x}) \to \infty$ as $n \to \infty$. However, this ideal threshold selection framework will be addressed in this section since it goes beyond the scope of our paper. Hence, we assume that the threshold is constant in terms of the explanatory variables $\mathbf{x}$, but still depends the sample size.

Instead of the regularization parameters $(\lambda_1, \lambda_2, \mu_1, \mu_2)$ as in (4.34), we consider a modification which allows for more efficient computation. Namely, we take $\lambda_1 = (1 - \alpha_1)\lambda$, $\lambda_2 = \alpha_1\lambda$, $\mu_1 = (1 - \alpha_2)\mu$ and $\mu_2 = \alpha_2\mu$ where $\alpha_1, \alpha_2 \in [0, 1]$ are the mixing parameters − a convex combination of the lasso and group lasso penalties. In practice, cross-validation or generalized cross-validation has been widely used to search for the optimal tuning parameters $\lambda, \mu, \alpha_1, \alpha_2$ and the threshold u in order to maximize its performance. However, since there are too many tuning parameters, this calibration process could be a computational burden and hardly be useful for many practical applications. Therefore, we consider the reduction of the "degrees of freedom" by taking $\alpha_1 = \alpha_2 = \alpha$. Furthermore, since we expect strong group-wise sparsity, we would thus use $\alpha = 0.05$. This condition clearly does not give

practitioners correct guidance to find the optimal regularization parameters since different problems will possibly be better fitted by different values of α and there is no reason for α_1, α_2 to be the same.

We also perform a series of simulations to compare our method with that of Beirlant and Goegebeur [9]. Their approach is based on the technique of local polynomial maximum likelihood estimation. Namely, in order to give more importance to the log-likelihood function contributions of observations close to $\mathbf{x}$, a weighting function governed by a kernel function K is introduced. Given K and a bandwidth parameter h, we denote $K_h(\mathbf{x}) = (1/h).K(\|\mathbf{x}\|/h)$. Much of our attention will be devoted to the local linear estimation of the functions $\ln \gamma^*(\mathbf{x})$ and $\ln \sigma^*(\mathbf{x})$, which is a different approach with respect to the one proposed by Beirlant and Goegebeur, since we know that the tail index $\gamma^*(\mathbf{x})$ and the scale parameter $\sigma^*(\mathbf{x})$ must be positive. Secondly, this parameterization allows avoiding the constrained optimization in the presence of constraints on those variables. For $\mathbf{x}$ sufficiently close to $\mathbf{x}_i$, we may write

$$\ln \gamma^*(\mathbf{x}_i) \approx \ln \gamma^*(\mathbf{x}) + \sum_{j=1}^{p} \frac{\partial}{\partial x^{(j)}} \ln \gamma^*(\mathbf{x}).(x_i^{(j)} - x^{(j)}) = \boldsymbol{\beta}_1.\Delta\mathbf{x}_i$$

and

$$\ln \sigma^*(\mathbf{x}_i) \approx \ln \sigma^*(\mathbf{x}) + \sum_{j=1}^{p} \frac{\partial}{\partial x^{(j)}} \ln \sigma^*(\mathbf{x}).(x_i^{(j)} - x^{(j)}) = \boldsymbol{\beta}_2.\Delta\mathbf{x}_i$$

where $\Delta\mathbf{x}_i = (1, x_i^{(1)} - x^{(1)}, \ldots, x_i^{(p)} - x^{(p)})^T$, $\boldsymbol{\beta}_1 = (\ln \gamma^*(\mathbf{x}), \frac{\partial \ln \gamma^*(\mathbf{x})}{\partial x^{(1)}}, \ldots, \frac{\partial \ln \gamma^*(\mathbf{x})}{\partial x^{(p)}})^T$ and $\boldsymbol{\beta}_2 = (\ln \sigma^*(\mathbf{x}), \frac{\partial \ln \sigma^*(\mathbf{x})}{\partial x^{(1)}}, \ldots, \frac{\partial \ln \sigma^*(\mathbf{x})}{\partial x^{(p)}})^T$.

We can therefore define the local linear log-likelihood estimator $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ as the maximizer of the weighted log-likelihood being of the form:

$$L_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \log g \left(y_i - u, \exp\left(\beta_{10} + \sum_{j=1}^{p} \beta_{1j}.(x_i^{(j)} - x^{(j)}) \right), \right.$$
$$\left. \exp\left(\beta_{20} + \sum_{j=1}^{p} \beta_{2j}.(x_i^{(j)} - x^{(j)}) \right) \right) K_h(\mathbf{x}_i - \mathbf{x}) \mathbb{I}(y_i \geq u) \tag{4.46}$$

where $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}, \ldots, \beta_{1p})^T$, $\boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}, \ldots, \beta_{2p})^T$ and $g(y; \gamma, \sigma)$ is the GPD density function. Denote $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2) = \arg\max_{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2} L_n(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \mathbf{x})$. From this, one obtains the local linear log-likelihood estimator of $\gamma^*(\mathbf{x})$ (resp. $\sigma^*(\mathbf{x})$) as $\hat{\gamma}(\mathbf{x}) = \exp(\hat{\beta}_{10})$ (resp. $\hat{\sigma}(\mathbf{x}) = \exp(\hat{\beta}_{20})$).

In this paper, the data-driven cross-validated negative log-likelihood scheme is applied as a performance metric to select the optimal hyper-parameters. The selection process is done via grid search, which is simply an exhaustive searching through a manually specified subset of the hyper-parameter space of a learning algorithm. Regarding the sparse group lasso approach, we define a finite set of

"reasonable" values for u, λ and μ as follows $u \in \{0.0, 0.1, \ldots, 3.9, 4.0\}$, $\lambda, \mu \in \{10^{-2.0}, 10^{-2.1}, \ldots, 10^{-3.3}, 10^{-3.4}\}$. Similarly, we set $u \in \{0.0, 0.1, \ldots, 3.9, 4.0\}$ and $h \in \{10^1, 10^{1.05}, \ldots, 10^2\}$ for the local polynomial approach.

On each knot, we compute the 5-fold cross validation error which is defined as follows:

For the sparse group lasso approach:

$$\mathrm{CV}_{\mathrm{SGL}}(\lambda, \mu; u) = \frac{1}{5} \sum_{k=1}^{5} CV^{[k]}(\lambda, \mu; u) \tag{4.47}$$

For the local polynomial approach:

$$\mathrm{CV}_{\mathrm{LP}}(h; u) = \frac{1}{5} \sum_{k=1}^{5} CV^{[k]}(h; u) \tag{4.48}$$

where $CV^{[k]}(\lambda, \mu; u)$ as well as $CV^{[k]}(h; u)$ is the cross-validation error in predicting the kth part (testing set), which is given by $-\sum_{i \in \text{testing set}} \log g\left(y_i - u; \hat{\gamma}^{[-k]}(\mathbf{x}_i), \hat{\sigma}^{[-k]}(\mathbf{x}_i)\right) \mathbb{I}(y_i \geq u)$ with $g(y; \gamma, \sigma)$ the GPD density function, $\hat{\gamma}^{[-k]}, \hat{\sigma}^{[-k]}$ calibrated on the training set.

The use of (4.47) is illustrated in Figure 4.1 for $n = 5000$ and $p = 10$ where we denote by $\mathrm{CV}_{\mathrm{optimal}}(u) = \min_{\lambda,\mu} CV(\lambda, \mu; u)$ the optimal cross validation metrics given the threshold u.

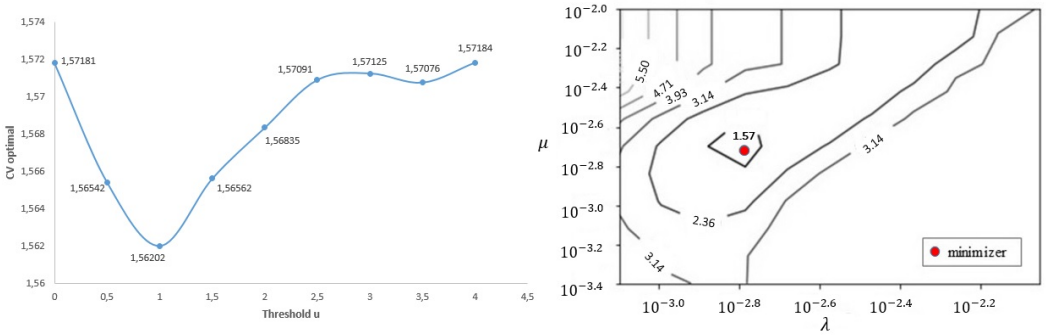


Figure 4.1: (left) $\mathrm{CV}_{\mathrm{optimal}}(u) = \min_{\lambda,\mu} CV(\lambda, \mu; u)$ versus the threshold u; (right) Contour plot of 5-fold CV error as a function of λ and μ at $u = 1$ and $\alpha = 0.05$ for $n = 5000$ and $p = 10$.

Regarding the sparse group lasso estimation, we are also interested in how well we can estimate the non-zero patterns of the κ_j's with others observations. To this end, we repeat such a procedure a total of 100 times with 100 independent samples of size $n = 5000$ for $p = 10$. In theory, we have to re-evaluate the optimal hyper-parameters u, λ and μ for each scenario since all these parameters depend upon the

number of exceedances over threshold. However, in practice, this latter step requires multiple iterative calculations which could be a computational burden. Therefore, for sake of simplicity, we assume that the optimal hyper-parameters remain constant with respect to different scenario. However, we would like to note that in some trials we were unable to make the sparse-group lasso regression select the right zero coefficients. This is due to the misspecified optimal hyper-parameters and the ideal threshold value with respect to the given sample, and due to the grouping effects. Overall we found that the sparse-group lasso reaches about 62% prediction accuracy, which is the proportion of correct nonzero functional components identifications over the initial total number of samples, i.e. 100.

Once the optimal hyper-parameters for the sparse group lasso estimation and the local polynomial estimation have been calibrated, our final step consists in checking how well these estimations perform. For this purpose, we compute the Mean Integrated Squared Error (MISE) which is defined as

$$MISE = \int_{[0,1]^p} MSE(\mathbf{x}) d\mathbf{x} \tag{4.49}$$

where we denote by $MSE(\mathbf{x})$ = empirical mean of $\left\{ \left| \hat{\gamma} \left(\mathbf{x}; \{\mathbf{x}_i\}_{i=1}^n, \{y_i^{(k)}\}_{i=1}^n \right) - \gamma^*(\mathbf{x}) \right|^2 : 1 \leq k \leq 100 \right\}$ the (empirical) mean squared error at point $\mathbf{x} \in [0,1]^p$. As a reminder, the mean squared error allows us, in a single measurement, to capture the ideas of bias and variance in our models, as well as showing that there is some uncertainty in our models that we cannot get rid. Therefore, the mean squared error is arguably the most important criterion used to evaluate the performance of an estimator at a particular point and the Mean Integrated Squared Error is thus considered as global metric for the performance of an estimation method. To value the equation (4.49) is a difficult operation, we will thus replace it by its *roughly* approximated version being of the form

$$\widetilde{MISE} = \Delta^p \mathbf{x} \sum_{m_1=1}^{5} \sum_{m_2=1}^{5} \sum_{m_3 \in \{1,2\}} \cdots \sum_{m_p \in \{1,2\}} MSE(x_{m_1}^{(1)}, \ldots, x_{m_p}^{(p)}) \tag{4.50}$$

where $x_{m_j}^{(j)}$ takes values in $\{0, 0.25, 0.5, 0.75, 1\}$ for $j = 1, 2$ and $\{0.25, 0.75\}$ for $j = 3, \ldots, p$, $\Delta^p \mathbf{x} = (0.25)^2.(0.5)^{p-2}$. As can be seen, the discretization is more granular with respect to the first two explanatory variables in (4.50) since these are two active variables in our model (4.45). Our key findings are reported in Table 1.

Surprisingly, we find that the thresholds u for different settings, obtained by using the cross-validation optimization method (4.47, 4.48) are more or less the same levels whether the local polynomial approach or the sparse group lasso approach. As observed from our numerical studies, the time required to calculate the cross-validation metrics in the sparse group lasso approach is much more longer than that in the local polynomial approach. This result tells us that the optimal threshold obtained from the criterion (4.48) could be directly applied in the sparse group lasso

		Local Polynomial			Sparse Group Lasso			
p	n	u	h	$\widetilde{MISE}$ (Std) ($\times 10^{-2}$)	u	λ	μ	$\widetilde{MISE}$ (Std) ($\times 10^{-2}$)
2	500	0.8	10^{3}	4.82 (2.08)	0.5	0.5×10^{-3}	0.63×10^{-3}	6.22 (3.37)
	5000	1.5	$10^{1.3}$	0.76 (0.47)	1.5	0.5×10^{-3}	0.63×10^{-3}	1.44 (0.65)
10	500	0.5	10^{4}	8.54 (0.42)	0.5	5.01×10^{-3}	5.01×10^{-3}	6.81 (4.31)
	5000	1.0	$10^{0.8}$	2.75 (0.59)	1.0	1.58×10^{-3}	2.0×10^{-3}	1.56 (0.75)

Table 4.1: Summary of optimal hyper-parameters, as well as the approximated Mean Integrated Squared Error $\widetilde{MISE}$, for the local polynomial estimation and the sparse group lasso estimation, with respect to different settings ($p = 2, 10$ and $n = 500, 5000$). To measure how spread out the calculation of $\widetilde{MISE}$ is, we compute its corresponding standard deviation (Std), which is the square root of the numerical value obtaine while calculating $var\left(\widetilde{MISE}\right) = \Delta^{p}\mathbf{x} \sum_{m_1=1}^{5} \sum_{m_2=1}^{5} \sum_{m_3 \in \{1,2\}} \cdots \sum_{m_p \in \{1,2\}} var\left(MSE(x_{m_1}^{(1)}, \ldots, x_{m_p}^{(p)})\right)$.

estimation, or at least as an indicator to check or to reinforce the reliability of the optimal threshold obtained from the criterion (4.47).

Table 1 shows that the local polynomial estimation has a better fit for low dimensional data ($p = 2$). The explanation of these results is twofold. First, the local polynomial estimation for low dimensional data is less subject to the curse of dimensionality (reference). Second, except for the GDP approximation (4.24), there is no any other approximation error that could occur. This is however not the case for the sparse group lasso estimation where we consider the natural cubic spline approximation (4.29, 4.30) even though these approximation errors could be small. This thus results in a better performance of the local polynomial estimation for low dimensional data. By contrast, the sparse group lasso estimation globally gives a better result for high dimensional data ($p = 10$). Clearly, in this case, the local polynomial estimation is affected by the curse of dimensionality, caused by the sparsity of data in a high dimensional space, resulting in a decrease in fastest achievable rate of convergence. As a result, this leads to a bad performance for the local polynomial estimators. The natural cubic spline assumption and the sparse group lasso algorithm could prevent those estimators suffering from the curse of dimensionality by partially or totally eliminating the non-active explanatory variables. As mentioned in Section (4.2.1.6), imposing a penalty term based on the l_1-norm will generate a high bias in our estimators. This latter one implies an increase in estimation errors. Therefore, we need to re-calibrate the coefficients on the active set. Besides, if we focus on the last column in Table 1, we will observe that the $\widetilde{MISE}$ for $p = 10$ is slightly greater than the $\widetilde{MISE}$ for $p = 2$ whether $n = 500$ or 5000. As explained above, these prediction inaccuracies come from the fact that the sparse group lasso method is somehow not able to identify all the non-active coefficients. Consequently, this drawback will create a small fluctuation in our estimators.

So far we compare the (rougly approximated) mean integrated squared error calculated by the local polynomial approach to that calculated by the sparse group lasso approach with different settings. In the following we carry out the out-of-sample test to elaborate the efficiency and forecasting capability of our estimators at different points with the same settings. As it involves only two active explanatory variables $(x^{(1)}, x^{(2)})$ constituting a two dimensional plane $[0, 1] \times [0, 1]$, This plane is divided into four quadrants. The first quadrant is the upper left-hand corner of the plane $[0, 0.5] \times [0.5, 1]$. The second quadrant is the upper right-hand corner $[0.5, 1] \times [0.5, 1]$. The third quadrant is the lower left-hand corner $[0, 0.5] \times [0, 0.5]$. Finally, the fourth quadrant is the lower right-hand corner $[0.5, 1] \times [0, 0.5]$. At each quadrant, we take unintentionally a testing point as shown in Table 2. For $p = 10$, we will concatenate the active part $(x^{(1)}, x^{(2)})$ and the inactive part $\mathbf{x}^{-(1,2)} \in \mathbb{R}^8$. Let us define $\mathbf{u}^+ = (0.15, 0.25, 0.35, \ldots, 0.85)^T$ and $\mathbf{u}^- = (0.85, 0.75, 0.65, \ldots, 0.15)$. The inactive part $\mathbf{x}^{-(1,2)}$ will take value in either $\mathbf{u}^+$ or $\mathbf{u}^-$. Finally we come up with 4 testing points as shown in Table 2 for $p = 10$. In this study, we will focus on the coverage probability and the average of confidence intervals at these testing points. Recall that the coverage probability is the proportion of the time that the

confidence interval contains the true value. To this end, we proceed the simulation as follows:

1. Reuse the samples of size $n = 100$, which are generated in the $\widetilde{MISE}$ calculation, i.e. $\left\{\{y_i^{(k)}\}_{i=1}^n : 1 \leq k \leq 100\right\}$.

2. Compute the 95% confidence interval (CI) for each sample by applying the boostrap sampling with replacement method from $n_{\text{boot}} = 500$ boostrapped data sets (see, for example, [29]).

3. Compute the proportion of samples for which the true tail index $\gamma^*(\mathbf{x})$ is contained in the confidence interval. That proportion is an estimate for the empirical coverage probability for the CI.

4. Compute the average of confidence intervals $\overline{\text{CI}}$.

Why is this necessary? Isn't the coverage probability always 95%? The answer is negative since the estimators $\left\{\hat{\gamma}(\mathbf{x}; \{\mathbf{x}_i\}_{i=1}^n, \{y_i^{(k)}\}_{i=1}^n) : 1 \leq k \leq 100\right\}$ are not normally distributed and the sample sizes are not large enough that we can invoke the Central Limit Theorem. Finally, our key findings are reported in Table 2.

N = 500, p = 2

	Local Polynominal		Sparse Group Lasso	
	Coverage Probability	$\overline{\text{CI}}$	Coverage Probability	$\overline{\text{CI}}$
(0.12, 0.86)	35%	0.2306	44%	0.361
(0.76, 0.21)	53%	0.4082	90%	0.5612
(0.22, 0.01)	46%	0.5097	64%	0.6843
(0.92, 0.96)	10%	0.3673	58%	0.597

N = 500, p = 10

	Local Polynominal		Sparse Group Lasso	
	Coverage Probability	$\overline{\text{CI}}$	Coverage Probability	$\overline{\text{CI}}$
(0.12, 0.86, $\mathbf{u}^+$)	0%	0.049	44%	0.3924
(0.76, 0.21, $\mathbf{u}^+$)	0%	0.049	4%	0.361
(0.22, 0.01, $\mathbf{u}^-$)	1%	0.1009	0%	0.025
(0.92, 0.96, $\mathbf{u}^-$)	0%	0.025	19%	0.3619

N = 5000, p = 2

	Local Polynominal		Sparse Group Lasso	
	Coverage Probability	$\overline{\text{CI}}$	Coverage Probability	$\overline{\text{CI}}$
(0.12, 0.86)	67%	0.1941	85%	0.2591
(0.76, 0.21)	75%	0.1939	80%	0.3086
(0.22, 0.01)	77%	0.2575	81%	0.3681
(0.92, 0.96)	74%	0.2593	71%	0.3547

N = 5000, p = 10

	Local Polynominal		Sparse Group Lasso	
	Coverage Probability	$\overline{\text{CI}}$	Coverage Probability	$\overline{\text{CI}}$
(0.12, 0.86, $\mathbf{u}^+$)	53%	0.2157	81%	0.2361
(0.76, 0.21, $\mathbf{u}^+$)	68%	0.2926	78%	0.3152
(0.22, 0.01, $\mathbf{u}^-$)	70%	0.3016	80%	0.3710
(0.92, 0.96, $\mathbf{u}^-$)	67%	0.2925	71%	0.3589

Table 4.2: Comparison of the coverage probability and the average of confidence intervals $\overline{\text{CI}}$ for the local polynomial estimators and the sparse group lasso estimators with different settings. For $p = 10$, we denote by $\mathbf{u}^+ = (0.15, 0.25, 0.35, \ldots, 0.85)^T$ and $\mathbf{u}^- = (0.85, 0.75, 0.65, \ldots, 0.15)$ the inactive covariates in our setting.

First, for a high-dimensional setting $p = 10$ and for a small sample size $N = 500$, we find that the local polynomial estimators have a poor performance; This is due to the curse of dimensionality as mentioned previously. Regarding the sparse group lasso estimators, their not-so-accurate performance can be explained as follows: a small sample budget does not allow to accurately identify the non-zero patterns and generates an important estimation error while doing the refitting.

Second, according to the simulation results, we notice that the expected confidence intervals estimated by the sparse group lasso approach is always wider than that estimated by the local polynomial approach. It is clear that the width of a confidence interval is related to its coverage probability. This is to say that wider confidence intervals have higher coverage probabilities, and narrower confidence intervals have lower coverage probabilities. This explains why the local polynomial estimator has a lower coverage probability than the sparse group lasso estimator. In other words, the sparse group lasso estimator is more conservative.

Third, in case of $p = 10$, when we increase the sample budget to $N = 5000$, it can be seen that the sparse group lasso estimators outperform the local polynomial estimators, which coincides with the previous results that the MISE of the sparse group lasso estimator may have a faster rate of convergence. Indeed, the expected length of confidence interval $\overline{\text{CI}}$ of the sparse group lasso estimator is close to that of the local polynomial estimator, but the sparse group lasso estimator has a higher

coverage probability.

4.3 Appendix

4.3.1 Proof of Lemma 1

The conditions (4.36) and (4.37) are trivial. To prove (4.38) and (4.39), we first recall the definition of the *subgradient* and *subdifferential* of a locally convex function on Ω, $f : \Omega \to \mathbb{R}$, at $x \in \Omega$ where Ω is a nonempty subset of $\mathbb{R}^m$. A vector $d \in \mathbb{R}^m$ is called a subgradient of f at point x if $f(y) \geq f(x)+(y-x)^T.d$ for all $y \in \Omega$. The collection of all subgradients of f at x is called the subdifferential of f at x, denoted by $\partial f(x)$. Then, a neccesary and sufficient conditions for x to be a minimum of f is that $\mathbf{0} \in \partial f(x)$. For any further information about the subgradient and subdifferential as well as the optimality theorem, the interested readers can refers to [14]. Let us return to our proof. We can easily verify (the same arguments for the groups $\tilde{\mathcal{G}}_j$) that

$$\partial\|\boldsymbol{\varphi}_{\mathcal{G}_j}\|_2 = \{\mathbf{e} \in \mathbb{R}^{G_j};\ \ \mathbf{e} = \frac{\boldsymbol{\varphi}_{\mathcal{G}_j}}{\|\boldsymbol{\varphi}_{\mathcal{G}_j}\|_2} \ \text{ if } \ \boldsymbol{\varphi}_{\mathcal{G}_j} \neq 0 \ \text{ and } \ \|\mathbf{e}\|_2 \leq 1 \ \text{ if } \ \boldsymbol{\varphi}_{\mathcal{G}_j} \equiv 0\}.$$

And the subdifferential set for $|(\boldsymbol{\varphi}_{\mathcal{G}_j})_k|$ obviously equals

$$\partial|(\boldsymbol{\varphi}_{\mathcal{G}_j})_k| = \{t \in \mathbb{R};\ \ t = \text{sign}((\boldsymbol{\varphi}_{\mathcal{G}_j})_k) \ \text{ and } \ |t| \leq 1 \ \text{ if } \ (\boldsymbol{\varphi}_{\mathcal{G}_j})_k = 0\}$$

Due to local convexity and differentiability of $P_n l(\boldsymbol{\varphi})$, we obtain finally the conditions (4.38) and (4.39).

4.3.2 Best approximation by splines

Let us first introduce two functional spaces.

Definition 1 (Polynomial Spline Space $\Phi_s^{a,b}$)**.** *Letting ξ_l for $l \in \{1, \ldots, K\}$ be K-interior knots satisfying the condition $a = \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{K+1} = b$. We define $\Phi_s^{a,b}$ the space of functions whose element is a polynomial of at most degree p on each of the intervals $[\xi_l, \xi_{l+1})$ for $l = 0, 1, \ldots, K$ and is $p-1$ continuously differentiable on $[a, b]$ if $p \geq 1$.*

Definition 2 (Empirically Centered Polynomial Spline Space $\bar{\Phi}_s^{a,b}$)**.** *Given the design points $(x_1, \ldots, x_n) \in [a, b]^n$, a polynomial spline space is centered if for every $g \in \Phi_s^{a,b}$ the following identity holds:*

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) = 0.$$

We denote by $\bar{\Phi}_s^{a,b}$ the empirically centered polynomial spline space.

According to de Boor (p. 149 in [18]), for every $f(x) \in C^{p+1}([a,b])$, there exists a constant $c > 0$ and a spline function $\dot{f} \in \bar{\Phi}_s^{a,b}$, such that $\|f - \dot{f}\|_\infty \leq c\|f^{(p+1)}\|_\infty \delta^{p+1}$ with $\delta = \max_{1 \leq l \leq K}(\xi_{l+1} - \xi_l)$.

Given the design points $(x_1, \ldots, x_n) \in [a,b]^n$, we assume furthermore that

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) = 0.$$

By defining $\ddot{f}(x) = \dot{f}(x) - \frac{1}{n}\sum_{i=1}^{n} \dot{f}(x_i) \in \bar{\Phi}_s^{a,b}$, it is straightforward to show that there exists a positive constant c' such that $\|f - \ddot{f}\|_\infty \leq c'\|f^{(p+1)}\|_\infty \delta^{p+1}$.

4.3.3 Block Coordinate Descent Algorithm

Algorithm 2

1: Set up with the initial parameter vector $\widehat{\boldsymbol{\varphi}}^{(0)}$ and the loop index $m = 0$.

2: Increase m by one: $m \leftarrow m + 1$ and cycle the optimization procedure through the groups:

(2.1) Set $\widehat{\boldsymbol{\varphi}}^{(m)} = \widehat{\boldsymbol{\varphi}}^{(m-1)}$.

(2.2) Regarding $j = 0$, if $\nabla P_n l(\widehat{\boldsymbol{\varphi}}^{(m)}_{-\mathcal{G}_0})_{\mathcal{G}_0} = 0$: set $\widehat{\boldsymbol{\varphi}}^{(m)}_{\mathcal{G}_0}$ and for $j = 1, \ldots, p$, if $\|S\left(\nabla P_n l(\widehat{\boldsymbol{\varphi}}^{(m)}_{-\mathcal{G}_j})_{\mathcal{G}_j}, \lambda_2\right)\|_2 \leq \lambda_1\sqrt{G_j}$: update $\widehat{\boldsymbol{\varphi}}^{(m)}_{\mathcal{G}_j} = \mathbf{0}$. Otherwise, cycle the optimization procedure with respect to each coordinate within the group fixed. That is, if $|\left(\nabla P_n l(\widehat{\boldsymbol{\varphi}}^{(m)}_{(\mathcal{G}_j,-k)})_{\mathcal{G}_j}\right)_k| \leq \lambda_2$: update $(\widehat{\boldsymbol{\varphi}}^{(m)}_{\mathcal{G}_j})_k = 0$. Otherwise, minimize the objective function over $(\boldsymbol{\varphi}_{\mathcal{G}_j})_k$ by a one-dimensional optimization. Cyclically iterate this coordinate-wise optimization process until convergence.

(2.3) Repeat the procedure for the groups $\tilde{\mathcal{G}}_j$ for $j = 0, \ldots, p$.

3: Repeat the entire step (2) until convergence.

Chapter 5

Conclusion

The objective of this thesis is twofold. Firstly, we aim to define the SCR estimation error related to the use a proxy in the context of the Solvency II regime, to establish the various causes of this error and to propose a methodology allowing it to be quantified in order to assess and control it.

Namely, we suggest to decompose the loss function into marginal and residual loss functions and apply the Bayesian penalized spline smoothing analysis on each functional components and showed how to control its errors. We also carried out several numerical tests on a simplified life insurance ALM simulator aiming to put into practice this methodology of quantification of the model error and the result is considered satisfactory. But, how well does this method perform with respect to the others in low and high dimensions (number of underlying risk-factors)? The optimal rate of convergence is typically of the form Γ^{-2r} where $r = p/(2p + d)$, Γ being the available sampling budget, p being a measure of the assumed smoothness of the loss function, d being the dimension of underlying risk-factors. The rate of convergence becomes slower when d increases. This is caused by the sparsity of data in high-dimensional spaces, resulting in a decrease in fastest achievable rates of convergence of regression function. This phenomenon is called the "curse of dimensionality". In the context of portfolio risk measurement, Hong et al. [54] in particular show the same issue of non-parametric approaches in high dimensional settings.

Pelsser and Schweizer [88] discussed the pros and cons of the LSMC and Replicating Portfolios methods in insurance liability modeling. As pointed out by the authors, both methods also suffer the curse of dimensionality problem as a result of using a multivariate basis constructed as the tensor product of the univariate bases. Alternative basis constructions must be considered to overcome this drawback. To the best of our knowledge, it is still a major challenge for the LSMC method to choose a functional form, which correctly approximate the conditional expected value. In practice, the conditional expected value of the cash flows can be calculated analytically in portfolio replication and closed form solutions can be obtained by combining the standard financial instruments that provides the same structure of cash-flows. However, finding a portfolio that replicates the strong path-dependent payoff functions is a more difficult problem.

Stone [107] revealed multiple advantages of the additive models. One of the interesting points is that we can achieve asymptotically the univariate-like optimal rate of convergence. This latter one leads us to consider the two-factor additive model (3.22) for the estimation of the excess loss function. However, this approach contains an approximation error, which cannot be eliminated and is probably non-

negligible. As we have seen in the derivation of the confidence interval, the error control is limited to the estimation error, but not to the approximation error. Within our application, we only consider two risk drivers: equity risk and interest rate risk, and the approximation error is thus relatively small compared to the estimation error. If one keeps all risk drivers of an insurance group, number of dimensions becomes incredibly large. With more risk drivers, it is not sure that the approximation (3.22) is still relevant. We will keep this for further research. There are several possibilities for further improving the procedure's efficiency in practice. Hong et al. [54] propose a decomposition technique for portfolio risk measurement, through which the loss of a portfolio is a linear combination of losses depending on only a small number of common risk factors. Another possibility is to use the variance reduction techniques into the simulation to improve the rate of convergence and to get a better performance.

Secondly, we would like to propose a methodology to estimate the tail-index of a heavy-tailed distribution when covariate information is available. In general, the goodness of a regression model is determined based on three fundamental aspects: high flexibility, less curse of dimensionality and strong interpretability. A model is flexible if it could provide accurate fits in a wide range of applications. Curse of dimensionality refers to various phenomena that the variance in estimation increases rapidly with increasing dimensionality. A model is interpretable if it could reveal the underlying structure of the problem that we want to solve. These are the criteria we can look at to give us a sense of what will be a reasonable approach to start with the tail-index estimation problem. Based on our simulation study, we can see that the proposed methodology has all of these properties.

We have seen that both the Local Polynomial maximum likelihood modelling and the Sparse Group Lasso modelling provide a correct estimate of the tail-index of a heavy-tailed distribution when covariate information is available. Another remark that we would like to point out is that both methods are relatively simple to use and to programme. According to the results of our numerical study, we notice that the Sparse Group Lasso approach allow for a more stable estimation of the tail-index parameter. This phenomenon can be interpreted as a result of the additive assumptions (4.26, 4.27). Indeed, Stone [107] showed that, under some mild auxiliary conditions, the additive regression can achieve the same optimal rate of convergence as that in a unidimensional setting. However, it happens that the Sparse Groupe Lasso estimation suffers a practical issue compared to the Local Polynomial maximum likelihood estimation. Indeed, this additive regression technique may speed up the computation which usually leads to little estimation errors, but it will generate a non-negligible approximation error as we impose an additional assumption. The quantification of this approximation error is however out of the scope of the current paper.

Nevertheless, there is a major gap between the computation and theoretical analysis due to the non-convex behavior of the negative log-likelihood objective function. This drawback will in some situations lead to the inconsistent results. We do not provide an answer to this issue in this paper and will keep this for further

research.

Concerning the interpretability of two methods, it is clear that the Sparse Group Lasso modelling becomes predominant in selecting the most relevant predictors contributing to the tail-heaviness of a distribution. However, there is still room for improvement regarding the proposed methodology. First, we will work toward the theoretical validation of this method by showing that the resulting estimate has oracle properties. Second, we will enrich our simulation part with other higher dimensional datasets where the dimensionality is comparable or even larger than the sample size.

Appendix A

Economic Scenarios Modeling

In this chapter, we discuss the modeling of the financial assets evolution, via our Economic Scenario Generator (ESG), that intervene in our Asset-Liability Management (ALM) model which will be presented in Chapter B.

Recall that an economic scenario generator (ESG) is a computer-based model of an economic environment that is used to produce simulations of joint behavior of financial market values and economic variables. Two common applications are driving the increased utilization of ESGs:

1. Market-consistent (risk-neutral) valuation work for pricing complex financial derivatives and insurance contracts with embedded options. These applications are mostly concerned with mathematical relationship within and among financial instruments are less concerned with forward-looking expectations of economic variables.

2. Risk management work for calculating business risk, regulatory capital and rating agency requirements. These applications apply real-world models that are concerned with forward-looking potential paths of economic variables and their potential influence on capital and solvency.

In our setting, our ESG is a support that allows us to simulate evolution of:

- Interest-Rate curves
- Discount factors
- Equity index
- Credit risk.

These economic variables and their interrelationships are modeled through a correlated Brownian motions generated by a correlated random vectors generator to maintain model integrity.

As mentioned previously, the simulations generated by our ESG have to verify the two following properties:

1. They must be market-consistent, that is to reflect the economic conditions of the valuation moment

2. They must be risk-neutral. The expected return is equal to the risk-free rate for every asset class.

Here, a question arises: "Which model will be used to model the curve of short rates ?". Two approaches are possible: equilibrium models and no-arbitrage models. The major difference between these two approaches is that the no-arbitrage models consider the observed yield curve as inputs while this is not the case for the equilibrium models. The observed yield curve used as input for the model is the yield curve provided by EIOPA. We decide to use the Hull and White one-factor (HW) model to model the short-rate curve. Regarding the equity index and credit risk, we use respectively the Black-Scholes (BS) model and the Jarrow, Lando and Turnbull (JLT) model. Each section will be organized as follows: Theoretical framework, model calibration and test on market consistency.

But first, let us introduce the correlated random vectors generator.

A.1 Correlated random vectors generator

Recall that a vector $\mathbf{X} = (X_1, \ldots, X_d)$ is Gaussian if any linear combination of its components $\sum_{i=1}^{d} a_i X_i$ has the Gaussian law. A Gaussian vector $\mathbf{X}$ is characterized by its mean $\mathbf{m}$ and its covariance matrix $\mathbf{V}$. We denote $\mathbf{X} \sim \mathcal{N}(\mathbf{m}, \mathbf{V})$.

In general, a Gaussian vector is simulated by the affine transformation of independent reduced Gaussian random variables, i.e. $\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.

Proposition 2. *Let d and d_0 be two non-zero integers, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$, $\boldsymbol{m} \in \mathbb{R}^d$ and $\boldsymbol{L}$ be a matrix of dimension $d \times d_0$. Then we have*

$$\boldsymbol{m} + \boldsymbol{L}\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{L}\boldsymbol{L}^T)$$

i.e. $\boldsymbol{m} + \boldsymbol{L}\boldsymbol{X}$ is a Gaussian vector of mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{V} = \boldsymbol{L}\boldsymbol{L}^T$

Conversely, a symmetric positive covariance matrix $\mathbf{V}$ of size d can always decompose in a non-unique way being of the form $\mathbf{V} = \mathbf{L}\mathbf{L}^T$, thanks to the spectral theorem [51]. Therefore this latter one allows us to simulate any Gaussian vector by being reduced to the previous case.

Theorem 4 (Spectral Theorem)**.** *Suppose $\boldsymbol{A}$ a Hermitian matrix. Then*

- *The eigenvalues of $\boldsymbol{A}$ are real.*
- *There is an orthogonal basis of eigenvectors for $\boldsymbol{A}$; in particular, $\boldsymbol{A}$ is diagonalizable over $\mathbb{C}$ (and even over $\mathbb{R}$ if $\boldsymbol{B}$ has real entries).*

To compute $\mathbf{L}$, we can use the Cholesky decomposition method, providing a lower triangular matrix $\mathbf{L}$, which when applied to a vector of uncorrelated samples, $\mathbf{u}$, produces the covariance vector of the system.

Cholesky decomposition assumes that the matrix being decomposed is *Hermitian* and *positive-definitive*. Since we are only interested in real-valued matrices, we can replace the property of Hermitian with that of *symmetric* (i.e. the matrix equals its own transpose).

In order to solve for the lower triangular matrix, we will make use of the Cholesky-Banachiewicz algorithm. First, we calculate the values for $\mathbf{L}$ on the main diagonal. Subsequently, we calculate the off-diagonals for the elements below the diagonal:

$$l_{kk} = \sqrt{v_{kk} - \sum_{j=1}^{k-1} l_{kj}^2}$$

$$l_{ik} = \frac{1}{l_{kk}} \left(v_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj} \right), \quad i > k$$

In a general sense, Monte Carlo methods involve the use of sampling from distribution(s) during the calculation of numerical approximations, most commonly to evaluate high-dimensional integrations. Although most schemes use random sampling, Monte Carlo does not necessarily imply the use of random numbers. In some situations there are better ways, most notably via the use of low-discrepancy numbers. There are many formal definitions of what "random" means. In a simplistic sense, we can say that in a sequence of truly random numbers each variate has no correlation with any other, on any scale.

In reality, a deterministic method must be used to generate variates, and by its vary nature this can never actually be random. Hence, we should really be using the term pseudo-random to describe computer generated numbers that resemble random numbers.

Discrepancy refers to the clustering of values that occurs when a sequence of samples are drawn from the uniform interval for the 1-dimensional case, the unit square for the 2-dimensional case, and the unit hypercube in higher dimensions. It is a measure of how inhomogeneously (non-uniformly) the values fit into the hypercube. Low-discrepancy numbers are deterministic sequences drawn to minimise their discrepancy - or equivalently, maximise their uniformity.

Formally, for a set S of N points in the d-dimensional hypercube $[0,1]^d$ we can define the discrepancy of S as :

$$D_N(S) = \sup_{\Omega \in [0,1]^d} \left| \frac{\text{Card}(\Omega; S)}{N} - v(\Omega) \right|$$

where $v(\Omega)$ is the volume of a sub-region Ω of the unit hypercube, $\text{Card}(\Omega; S)$ is the number of points in S that fall into Ω. In a general sense, more uniformly distributed sets of points have lower discrepancy than less uniformly distributed sets of points.

Sobol numbers are low-discrepancy, quasi-random numbers. They are highly uniform, much more so than standard uniform distribution generators as illustrated by the diagrams below [1]: Expectations obtained from Monte Carlo schemes using

[1] NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. For any further information about this package, please refer to https://www.numpy.org/

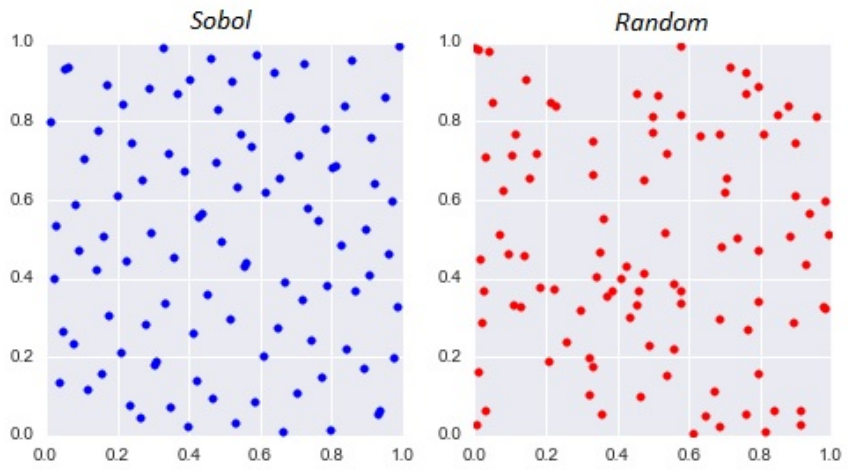


Figure A.1: Sampling the fitting space: Sobol pseudo-random numbers vs Numpy pseudo-random numbers.

ideal random numbers are expected to converge asymptotically as a function of the number of trials as $\frac{1}{\sqrt{N}}$. In contrast, low-discrepancy numbers should converge as $\frac{\ln(N)^d}{N}$ where d is the dimensionality of the problem. In low dimensions, we expect low-discrepancy numbers to converge substantially faster than pseudo-random numbers. This increased convergence speed allows us to achieve a greater accuracy with the same number of simulations, or equivalently the same accuracy with fewer simulations and reduced computational expense.

For a futher evaluation of low-discrepancy numbers and their use within stochastic scenario generation, please refer to [92].

A.2 Hull White Model

John Hull and Alan White introduced the one-factor Hull-White interest rate model in 1990 (see, e.g. [56–60]). The model is no-arbitrage yield curve model, meaning that it can reproduce exactly the initial yield curve implied by bond prices. The model assumes that the short rate r_t is governed by the following dynamics:

$$dr_t = (\theta_t - ar_t)dt + \sigma_n dW_t^n \tag{A.1}$$

where σ represents the instantaneous volatility of the short rate, and a is the mean-reverting speed. The time dependent parameter θ_t is determined by σ, a and the initial yield curve.

Namely, let us denote by $f(0,t)$ the instantaneous forward rate given by

$$f(0,t) = -\frac{\partial \ln P(0,t)}{\partial t}$$

with $P(0,t)$ the price of the zero coupon bond paying 1 at time T, the factor θ_t is given by

$$\theta_t = \frac{df(0,t)}{dt} + af(0,t) + \frac{\sigma_n^2}{2a}\left(1 - e^{-2at}\right).$$

The parameter θ_t is in fact derived from the no arbitrage condition of the model on the price of the discount factors.

Namely, one can work out the expression of the parameter θ_t so that we have the following relationship:

$$P(0,T) = \mathbb{E}\left(e^{-\int_0^T r_s ds}\right)$$

Since the process $\int_t^T r_s ds$ is Gaussian, we have

$$\begin{aligned} P(0,T) &= \exp\left(-\mathbb{E}\left[\int_t^T r_s ds \mid \mathcal{F}_t\right] + \frac{1}{2}\mathrm{Var}\left[\int_t^T r_s ds \mid \mathcal{F}_t\right]\right) \\ &= A(t,T)e^{-B(t,T)} \end{aligned}$$

where

$$B(t,T) = \frac{1}{a}\left(1 - e^{-2a(T-t)}\right)$$

and

$$A(t,T) = \frac{P(0,T)}{P(0,t)} \exp\left(B(t,T)f(0,t) - \frac{\sigma^2}{4a}\left(1 - e^{-2at}\right)B^2(t,T)\right)$$

For a full derivation of the parameter θ_t, please refer to Appendix C.

The Hull-White model is used widely in derivative pricing as well as in risk management. Hull and White published a series of papers that discuss the procedure to construct a Hull-White interest rate tree as well as applications of the model.

One main advantage of the Hull-White model is its tractability. More specifically, given the initial yield curve and the model parameters, analytic formula is available for the distribution of short term and long term interest rates at a future time. In addition, vanilla bonds and European options can also be valued analytically.

The tractability of the Hull-White model makes it convenient to calibrate the parameters using bond or options prices. This is the primary reason we use the model in this study. Meanwhile, it is also important to understand that the one-factor Hull-White model has certain limitations. For instance, the Hull-White model assumes that the short rate is normally distributed. As a consequence, both the spot and forward interest rates can be negative, which is usually considered unrealistic. Another drawback of the one-factor Hull-White model is that, as there is only one driving Brownian motion, all forward rates are determined by the short rate. As a consequence, the shape of the yield curve is completely determined by the short rate. Therefore, the model is not flexible enough to account for twists in the yield curve. However, various other models can resolve one or both issues. Here, we provide a few examples of these models and explain the reason why these models are not used in our study. A comprehensive review of these models is beyond the scope of this study and can be found in Andersen and Piterbarg [3] or Brigo and Mercurio [19].

The Black-Derman-Toy model and Black-Karasinski model assume that the short rate follows log-normal distribution and is always non-negative. However, log-normal short rate models typically don't have analytic bond pricing formula and numerical technique is required to calibrate the model to the initial yield curve. The lack of flexibility associated with one-factor models is a major concern when pricing exotic options. As pointed out in Andersen and Piterbarg, "... as a general rule, all derivatives that have payouts exhibiting significant convexity to non-parallel moves of the forward curve must not be priced in a one-factor model." Jagannathan, Kaplin and Sun [62] pointed out that pricing error can be relatively large even for multi-factor models such as the three-factor CIR model. On the other hand, one-factor models remain popular for other purposes, such as risk management. Since our study does not focus on pricing exotic derivatives, the one-factor Hull-White model is considered sufficiently flexible, especially when the parameters are allowed to be time variant.

Model calibration

As we have already pointed out, the model's diffusion depends on two different parameters: the volatility and the mean reversion. Calibrating the model finding values for these two parameters, consistent with some market prices. These market prices should obviously be actively traded options, i.e. financial instruments used by the trader to effectively hedge his portfolio. Caps and swaptions are the two main markets in the interest rate derivatives world. However, in our setting, we limit ourselves to the model calibration based on the cap pricing.

A.2.1 Cap pricing

In the Hull and White framework, the price of the European call priced at t, of maturity T, with strike K and written of a zero-coupon bond of maturity S is:

$$\mathbf{ZBC}(t,T,S,K) = P(t,S)\Phi(h) - KP(t,T)\Phi(h-\tilde{\sigma})$$

where $\Phi(\cdot)$ is the standard normal cumulative function, and

$$\tilde{\sigma} = \sigma\sqrt{\frac{1-e^{-2a(T-t)}}{2a}}B(T,S)$$
$$h = \frac{1}{\tilde{\sigma}}\ln\left(\frac{P(t,S)}{P(t,T)}\right) + \frac{\tilde{\sigma}}{2}.$$

The price of the put contract having a similar formula, which is

$$\mathbf{ZBP}(t,T,S,K) = KP(t,T)\Phi(\tilde{\sigma}-h) - P(t,S)\Phi(-h).$$

We get now the price of the cap with settlement dates $t_0 \equiv T, t_1, \ldots, t_n \equiv S$, with strike K and of nominal N:

$$\mathbf{Cap}(t,T,S,K) = N\sum_{i=1}^{n}(1+K\tau_i)\mathbf{ZBP}\left(t,t_{i-1},t_i,\frac{1}{1+K\tau_i}\right)$$

where we denote by τ_i the fraction of year between two settlement date t_{i-1} and t_i.

We carry out the calibration of the Hull and White model from ATM[2] Euribor 6 month Caps for the maturaties $\mathcal{T} = \{3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 20\}$. Regarding the market data, we retrieve on Bloomberg:

1. the prices of ATM caps at the valuation date 29/12/2017 [3],
2. the ATM strike K_{atm} at the same date [4].

Let us denote by $\mathbf{Cap}^{Mkt}$ the market price observed on Bloomberg and by $\mathbf{Cap}^{Mdl}$ its corresponding theoretical price. To calibrate the parameters a and σ_n, we have to solve the following optimization problem:

$$\hat{a}, \hat{\sigma}_n = \underset{a,\sigma_n}{\arg\min} \sum_{T_i \in \mathcal{T}} \left(\mathbf{Cap}^{Mkt}(T_i, K_{atm}) - \mathbf{Cap}^{Mdl}(0, 0, T_i, K_{atm}; a, \sigma_n) \right)^2 \quad \text{(A.2)}$$

We realize this minimization problem in two steps:

1. First, one fix a and solves the optimization problem (A.2) only on σ_n by using the gradient descent method developed on Python.
2. We again carry out this optimization problem for each value of a within a predefined interval in a way to find the couple $(\hat{a}, \hat{\sigma}_n)$ minimizing the mean-square error (A.2).

Here is the result of our calibration.

Maturity	**Market Price**	**Model Price**	**Abs. Error (%)**
3Y	0.0046	0.008882	0.428234
4Y	0.0093	0.014044	0.474377
5Y	0.0151	0.020086	0.498567
6Y	0.0219	0.026825	0.492522
7Y	0.0296	0.034158	0.455831
8Y	0.0380	0.041948	0.394783
9Y	0.0469	0.050244	0.334361
10Y	0.0562	0.058975	0.277498
12Y	0.0753	0.077757	0.245738
15Y	0.1040	0.104103	0.010276
20Y	0.1480	0.140244	0.775627

Table A.1: Overall difference between the market price and the model price.

[2] ATM stands for At-The-Money
[3] Bloomberg ticker: EUCPAM** where ** is replaced by the maturity
[4] Bloomberg ticker: EUCPST** where ** is replaced by the maturity

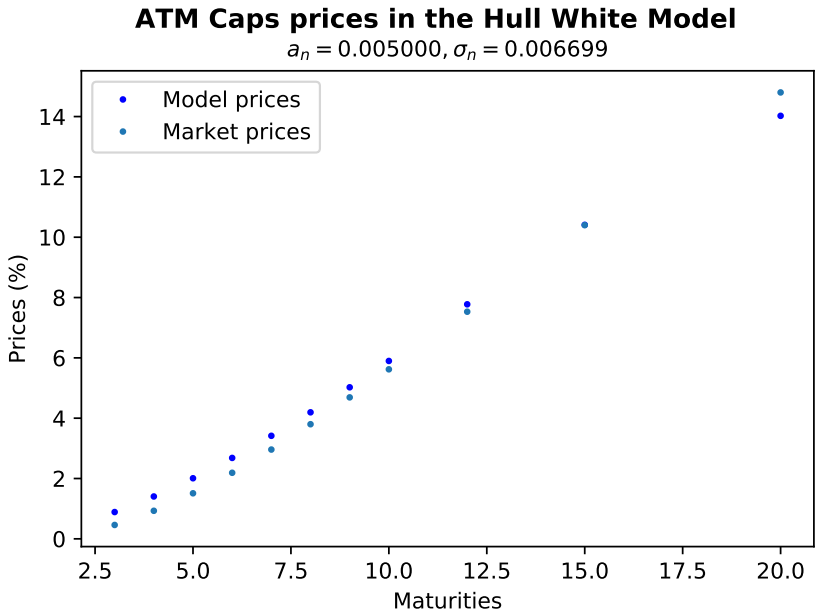


Figure A.2: Calibration results given the optimal model parameters $\hat{a}$ and $\hat{\sigma}_n$.

A.3 Black Scholes Model

The standard Black-Scholes formula [17] has been obtained under the assumptions that the stock price S_t follows a lognormal diffusion with constant volatility σ_S:

$$\frac{dS_t}{S_t} = rdt + \sigma_S dW_t^S, \quad S_0 = F, \tag{A.3}$$

where r is the risk-free interest rate. The results established in the Black Scholes model still hold under more general hypothese, namely when the volatility is a time-dependent determistic function $\sigma_S(t)$:

$$\frac{dS_t}{S_t} = (r_t - q_t)dt + \sigma_S(t)dW_t^S, \quad S_0 = F, \tag{A.4}$$

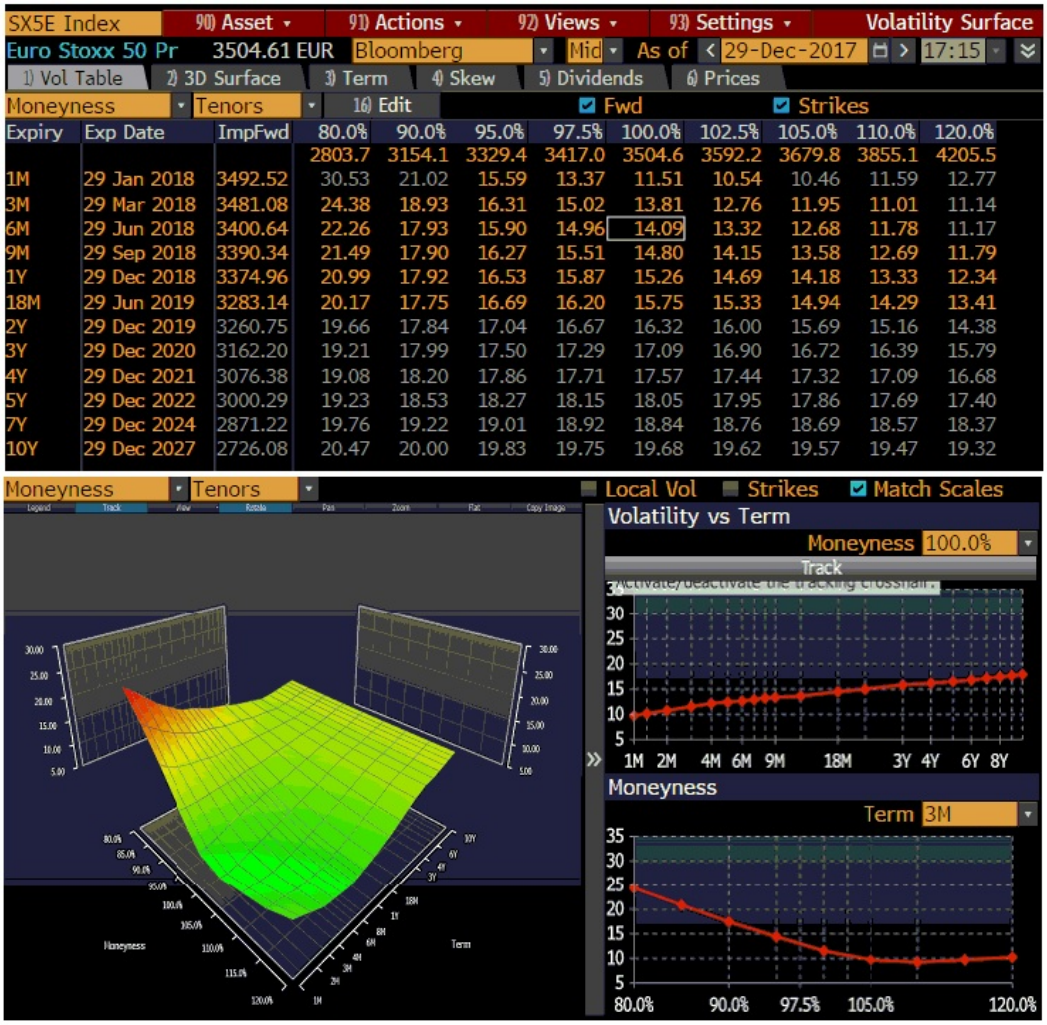
where r_t and q_t are also time-dependent functions, being the risk-free interest rate and the dividend rate of S_t, respectively.

To get a simple and general picture of the Black Scholes model, we list here its advantages and limitations.

Advantages: Closed form formula can be obtained to price Calls, Puts and many other European contracts. As a result, computations are instantaneous.

Limitations: The Black-Scholes hypothesis of a constant volatility is unrealistic under real market conditions. Recall that using market option prices, we can invert Black Scholes formula to compute the implied volatility. For different option strikes K and maturities T, we get different volatilities, therefore the Black-Scholes hypothesis of constant volatility does not hold. Moreover, empirical evidence of

the markets shows that the implied volatility is shaped like a smile or a skew. An example of the implied volatility surface is given in Figure (A.3).



Expiry	Exp Date	ImpFwd	80.0%	90.0%	95.0%	97.5%	100.0%	102.5%	105.0%	110.0%	120.0%
			2803.7	3154.1	3329.4	3417.0	3504.6	3592.2	3679.8	3855.1	4205.5
1M	29 Jan 2018	3492.52	30.53	21.02	15.59	13.37	11.51	10.54	10.46	11.59	12.77
3M	29 Mar 2018	3481.08	24.38	18.93	16.31	15.02	13.81	12.76	11.95	11.01	11.14
6M	29 Jun 2018	3400.64	22.26	17.93	15.90	14.96	14.09	13.32	12.68	11.78	11.17
9M	29 Sep 2018	3390.34	21.49	17.90	16.27	15.51	14.80	14.15	13.58	12.69	11.79
1Y	29 Dec 2018	3374.96	20.99	17.92	16.53	15.87	15.26	14.69	14.18	13.33	12.34
18M	29 Jun 2019	3283.14	20.17	17.75	16.69	16.20	15.75	15.33	14.94	14.29	13.41
2Y	29 Dec 2019	3260.75	19.66	17.84	17.04	16.67	16.32	16.00	15.69	15.16	14.38
3Y	29 Dec 2020	3162.20	19.21	17.99	17.50	17.29	17.09	16.90	16.72	16.39	15.79
4Y	29 Dec 2021	3076.38	19.08	18.20	17.86	17.71	17.57	17.44	17.32	17.09	16.68
5Y	29 Dec 2022	3000.29	19.23	18.53	18.27	18.15	18.05	17.95	17.86	17.69	17.40
7Y	29 Dec 2024	2871.22	19.76	19.22	19.01	18.92	18.84	18.76	18.69	18.57	18.37
10Y	29 Dec 2027	2726.08	20.47	20.00	19.83	19.75	19.68	19.62	19.57	19.47	19.32

Figure A.3: Implied volatility surface for the index Eurostoxx 50 extracted from Bloomberg as of date of calibration 29/12/2017

To respect the market-consistent and risk-free properties, and to simplify our simulation, the return rate r_t follows the Hull-White model (A.1) and the volatility σ_S and the dividend rate q are supposed to be constant, i.e.

$$\frac{dS_t}{S_t} = (r_t - q)dt + \sigma_S dW_t^S, \quad S_0 \equiv F \tag{A.5}$$

Model calibration

The only parameter to calibrate in is the volatility σ_S. To this end, we appeal the following result to obtain the closed formula for the vanilla European Call option.

Proposition 3. *Assume that the stock price S_t follows the Black-Scholes model (A.5) with the return rate r_t following the Hull-White model (A.1) and that the correlation between two Brownian motions dW_t^n and dW_t^S is ρ. Then we have*

$$\boldsymbol{Call}(T,K) = \mathbb{E}^{\mathbb{Q}}\left(e^{-\int_0^T r_s ds}(S_T - K)_+\right) = F\Phi(d_1) - KP(0,T)\Phi(d_2)$$

where $d_1 = \frac{1}{\ddot{\sigma}\sqrt{T}}\left[\ln F/(P(0,T)K) + \frac{1}{2}\ddot{\sigma}^2 T\right]$ *and* $d_2 = d_1 - \ddot{\sigma}^2\sqrt{T}$ *with*

$$\ddot{\sigma}^2 = \sigma_S^2 + \frac{2\rho\sigma_n\sigma_S}{aT}\left[T - \frac{1}{a}\left(1 - e^{-aT}\right)\right] + \frac{\sigma_n^2}{a^2 T}\left[T - \frac{1}{2a}e^{-2aT} + \frac{2}{a}e^{-aT} - \frac{3}{2a}\right]$$

Proof. First we rewrite the equations (A.1) and (A.5) differently as follows:

$$\begin{aligned} dr_t &= (\theta_t - ar_t)dt + \sigma_n dW_t^1 \\ dS_t &= S_t\left(r_t dt + \sigma_S(\rho dW_t^1 + \sqrt{1-\rho}dW_t^2)\right) \end{aligned}$$

where $\{W_t^1, t \geq 0\}$ and $\{W_t^2, t \geq 0\}$ are two independent standard Brownian motions.

Following the results obtained in the Hull White model, the zero-coupon bond price is given by

$$\begin{aligned} P(t,T) &= \mathbb{E}^{\mathbb{Q}}\left(e^{-\int_t^T r_s ds} \mid \mathcal{F}_t\right) \\ &= \exp\left(-B(0,T)r_t - \int_t^T \theta(s)B(s,T)ds + \frac{1}{2}\int_t^T \sigma_0^2 B(s,T)^2 ds\right) \end{aligned}$$

Then we have

$$d\ln P(t,T) = \left(r_t - \frac{1}{2}\sigma_n^2 B(t,T)^2\right)dt - \sigma_n B(t,T)dW_t^1$$

or

$$dP(t,T) = P(t,T)(r_t dt - \sigma_0 B(t,T)dW_t).$$

Let us denote by $\mathbb{Q}^T$ the T-forward measure with the corresponding numeraire $P(t,T)$ defined by

$$\begin{aligned} \frac{d\mathbb{Q}^T}{d\mathbb{Q}}\mid_t &= \frac{P(t,T)}{P(0,T)}e^{-\int_0^t r_s ds} \\ &= \exp\left(-\frac{1}{2}\int_0^t \sigma_n^2 B(s,T)^2 ds - \int_0^t \sigma_n B(s,T)dW_s^1\right). \end{aligned}$$

By the Girsanov theorem [41], under $\mathbb{Q}^T$, the process $\{(\widehat{W}_t^1, \widehat{W}_t^2), t \geq 0\}$ where

$$\begin{aligned} \widehat{W}_t^1 &= W_t^1 + \int_0^t \sigma_n B(t,T)ds \\ \widehat{W}_t^2 &= W_t^2 \end{aligned}$$

are two standard Brownian motions. Clearly, under $\mathbb{Q}^T$,

$$\begin{aligned} dP(t,T) &= P(t,T)\left[(r_t + \sigma_n^2 B(t,T)^2)dt - \sigma_n B(t,T)d\widehat{W}_t^1\right] \\ dS_t &= S_t\left[(r_t - \rho\sigma_n\sigma_S B(t,T))\,dt + \sigma_S\left(\rho d\widehat{W}_t^1 + \sqrt{1-\rho^2}d\widehat{W}_t^2\right)\right] \end{aligned}$$

and $S_u/P(u,T)$ is a martingale. Therefore, the forward price has the form

$$F(t,T) = \mathbb{E}^{\mathbb{Q}^T}\left[S_T \mid \mathcal{F}_t\right] = \frac{S_t}{P(t,T)}$$

and

$$\begin{aligned} dF(t,T) &= \frac{dS_t}{P(t,T)} - \frac{S_t}{P^2(t,T)}dP(t,T) - \frac{d\langle S_t, P(t,T)\rangle}{P^2(t,T)} + \frac{S_t}{P^3(t,T)}d\langle P(t,T), P(t,T)\rangle \\ &= F(t,T)\left[(\rho\sigma_S + \sigma_n B(t,T))d\widehat{W}_t^1 + \sigma_S\sqrt{1-\rho^2}d\widehat{W}_t^2\right] \end{aligned}$$

Define $\ddot{\sigma}$ an effective volatility being of the form:

$$\begin{aligned} T\ddot{\sigma}^2 &= \int_0^T \left[(\rho\sigma_S + \sigma_n B(t,T))^2 + \sigma_S^2(1-\rho^2)\right] ds \\ &= \sigma_S^2 T + \frac{2\rho\sigma_S\sigma_n}{a}\left[T - \frac{1}{a}(1-e^{-aT})\right] + \frac{\sigma_n^2}{a^2}\left[T - \frac{1}{2a}e^{-2aT} + \frac{2}{a}e^{-aT} - \frac{3}{2a}\right] \end{aligned}$$

Then

$$F(T,T) = F(0,T)\exp\left(-\frac{1}{2}\ddot{\sigma}^2 T + \ddot{\sigma}\sqrt{T}\xi\right)$$

where ξ is a standard normal random variable. Consequently, we have

$$\begin{aligned} \mathbb{E}^{\mathbb{Q}}\left(e^{-\int_0^T r_s ds}(S_T - K)_+\right) &= \mathbb{E}^{\mathbb{Q}}\left(e^{-\int_0^T r_s ds}(F(T,T) - K)_+\right) \\ &= \mathbb{E}^{\mathbb{Q}^T}\left(e^{-\int_0^T r_s ds}(F(T,T) - K)_+ \frac{d\mathbb{Q}}{d\mathbb{Q}^T}\mid_T\right) \\ &= P(0,T)\mathbb{E}^{\mathbb{Q}^T}\left((F(T,T) - K)_+\right) \\ &= F\Phi(d_1) - KP(0,T)\Phi(d_2) \end{aligned}$$

□

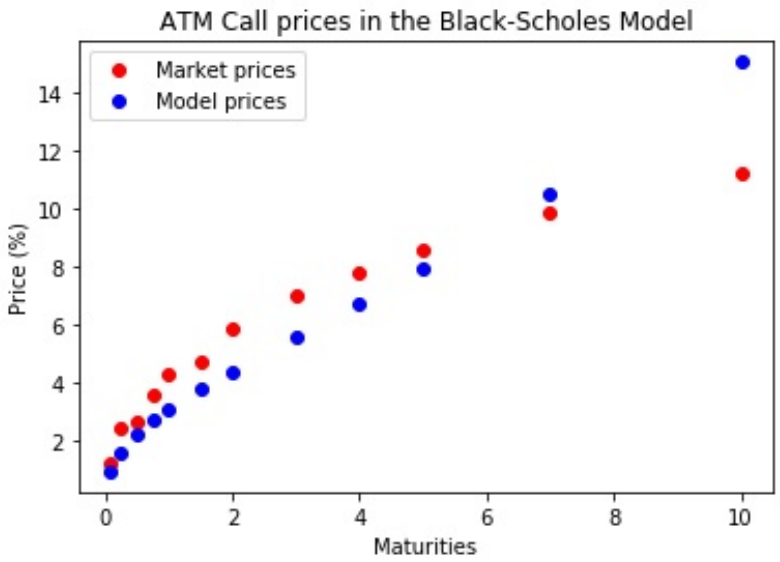


Figure A.4: Calibration results given the optimal model parameter $\hat{\sigma}_S = 8.03\%$

Maturity	Market Price	Model Price	Abs. Error (%)
1M	0.0117	0.0091	0.2615
3M	0.0240	0.0156	0.8421
6M	0.0261	0.0219	0.4173
9M	0.0357	0.0266	0.9011
1Y	0.0430	0.0304	1.2399
18M	0.0471	0.0375	0.9587
2Y	0.0583	0.0436	1.4720
3Y	0.0697	0.0554	1.4331
4Y	0.0782	0.0671	1.1035
5Y	0.0858	0.0792	0.6563
7Y	0.0984	0.1054	0.7026
10Y	0.1122	0.1509	3.8706

Table A.2: Overall difference between the market price and the model price.

A.4 Jarrow, Lando and Turnbull Model

The stochastic credit development allows the ALM model to reach more accuracy in bond simulated cash-flows. Bonds are rated by credit agency. This rating evolves over time and leads many changes for the bond such as the market value and the credit spread.

The ratings of the bond are given by credit agency with a transition matrix which contains the probabilities to jump from one state to another. These probabilities are computed as the historical average of such migrations. However the market anticipates different migration probabilities. They are the risk neutral migration probabilities and are those used in the model to simulate the bond rating dynamics.

The bonds projection in stochastic model introduces the following components:

1. Bond's rating at time t: η_t
2. Recovery rate δ: The same recovery rate is used for all the bonds. At maturity, the cash flow is 1 if the bond has not defaulted and δ otherwise.

The risky zero-coupon bond price $P^{\eta_t}(t,T)$ can be expressed thanks to the risk-free zero-coupon price $P(t,T)$, the recovery rate and the risk neutral default probability $\mathbb{Q}\left(\tau_D^{\eta_t} > T\right)$ and assumption that risk free rate $\{r_s\}_{t\leq s\leq T}$ and rating process are independent as

$$\begin{aligned} P^{\eta_t}(t,T) &= \mathbb{E}^{\mathbb{Q}}\left[\exp\left(-\int_t^T r_s ds\right)\left(\mathbb{I}(\tau_D^{\eta_t} > T) + \delta\mathbb{I}(\tau_D^{\eta_t} \leq T)\right)\right] \\ &= P(t,T)\left(\delta + (1-\delta)\mathbb{Q}\left(\tau_D^{\eta_t} > T\right)\right) \end{aligned}$$

where $\mathbb{Q}$ is the risk neutral probability measure and $\tau_D^{\eta_t}$ is the stopping time whether an event of default occurs.

The historical transition matrices are average of past migrations. Risk neutral transition matrices are expectations of future migrations. In $t = 0$ a time-dependant factor-the risk premium-is computed to transform the historical transition process in risk-neutral one. This factor is then used to simulate the dynamics of the risk-neutral transition matrix over time in each scenario.

The central variable of the model is the bond rating. His evolution is described as a Markov chain. The spreads are computed from this dynamic. A process that follows a Markov chain over a set of state is a process that jumps from one state to another over time. For example, we can have a set of two states A and B. The process η_t following the Markov chain will value A or B and changes over time. If the changes occur at discrete time, the Markov chain is named discrete Markov chain. We can then define the probabilities to jump from one state to another. With our example, at the time t of jump, the process have a probability $p_t(A \to B)$ to jump from the state A to the state B between the time t and $t+1$, $p_t(A \to A)$ to jump from the state A to the state A (to stay in the same state), $p_t(B \to A)$ to jump from B to A and $p_t(B \to B)$ to jump from B to B. At time t the process η_t is only in one state so there are only two possibilities, but at time $t+1$ it is not known

in which state the process will be at time t and then it is necessary to define the four probabilities. It is convenient to present these probabilities under matrix form.

$$P_{t,t+1} = \begin{bmatrix} p_t(A \to A) & p_t(A \to B) \\ p_t(B \to A) & p_t(B \to B) \end{bmatrix}$$

The rows represent the start state which is the state wherein the process is at time t. The columns represent the final state which is the state where the process jumps and then the state of the process at time $t + 1$. Then on the call which is on the row A and the column B we consequently have the probability to jump from the state A to the state B.

As we can see the probabilities of transition have an index for the time, the transition matrix is then time dependant. In this case the process is named inhomogeneous Markov chain. This is an usual feature. For example the one year transition matrix given by credit rating agency changes every year which means that the historical rating process is an inhomogeneous Markov chain.

We can also define the transition matrix over two time periods $P_{t,t+2}$. The important property of Markov chain is then

$$P_{t,t+2} = P_{t,t+1}P_{t+1,t+2}$$

If the transition matrix is not time dependant, the process follows an homogeneous Markov chain. The transition matrix can then be indexed only with number of time period over which the transition matrix runs. This is possible thanks to the previous property.

Because of the homogeneity, we have

$$P_{t,t+1} = P_1$$

And then

$$P_{t,t+2} = P_{t,t+1}P_{t+1,t+2} = P_1^2 = P_2$$

Until now the Markov chain is discrete. The jumps occur only at some points of time. We now introduce continuous Markov chain where the jumps can occur at any time. To define the transition matrix over an infinitesimal length of time dt, we introduce the generator Λ_t of the Markov chain. The generator is a matrix such as

$$P_{t,t+dt} = I + \Lambda_t dt$$

Mathematical properties allow us to write

$$P_{t,t+T} = \exp\left(\int_t^{t+T} \Lambda_s ds\right)$$

As in the discrete time case, we can have homogeneous continuous Markov chain. The generator is then no more time dependant and the transition matrix over a time period of length T is given by

$$P_T = P_{t,t+T} = P_1^T, \ \ \forall t > 0.$$

We see clearly the link between discrete and continuous model. Thanks to all these definitions, we can now describe the rating process modelling.

A.4.1 Transition process

Credit rating agencies compute transition matrices. These matrices are the average over historical data of the transition that happened. Then we named these matrices historical transition matrices.

There is another probability of default that allows computing the market price of risky bonds. This probability is named risk free probability. The pricing formula will be developed in the next section. We are now interested in how compute this risk free default probability. Under the no-arbitrage condition, there exists a unique risk free probability which is equivalent to the historical probability. Moreover, we also assume that the risk free rate and the rating process are independent. These assumptions are analysed in Jarrow et al. [63]. Under this probability, the risk free bonds and risky bonds discounted prices are martingales.

We make assumption that under the historical probability the rating process follows a homogeneous Markov chain. This is not true since the one year historical transition matrices provide by credit rating agencies change each year. But this simplification is required for the sake of computability and the changes in the transition matrices are little.

Always for the sake of computability we will work with the continuous framework. The one year historical transition matrix can be written:

$$P_1^{historical} = \exp\left(\Lambda^{historical}\right)$$

We could compute the risk free transition matrix with a time dependant matrix of risk premium Π_t. The relation between historical process and risk free process would be

$$\Lambda_t^{riskfree} = \Pi_t \Lambda^{historical}$$

With no particular assumption, the risk free Markov chain is then inhomogeneous. The risk premium as a matrix is not bearable from a computability point of view. This is due to the lack of data and the impossibility to ensure that the risk free generator remains a generator matrix over time. The trade-off between computability and flexibility is then to use a time dependant scalar risk premium. This is discussed in Lando (2004) [73] (see Chapter 6). Then we have

$$\Lambda_t^{riskfree} = \pi_t \Lambda^{historical}$$

Since the risk premium has to be positive, it is convenient to use a CIR (Cox, Ingersoll and Ross) process to model this factor.

$$d\pi_t = \alpha(\mu - \pi_t)dt + \sigma_\pi \sqrt{\pi_t} dW_t^\pi.$$

The risk neutral transition matrix over the time interval $[t, t + \Delta t]$ is

$$P_{t,t+\Delta t} = e^{\int_t^{t+\Delta t} \Lambda_s^{riskfree} ds} = e^{\Lambda^{historical} \int_t^{t+\Delta t} \pi_s ds}.$$

This matrix is stochastic since the risk premium is. To compute this matrix, we have to know the risk premium path over the time interval $[t, t + \Delta t]$. Thus this

matrix is named conditional transition matrix. Note that it is still a Markov chain transition matrix. The problem is that at time t we do not know the future path of the risk premium. So we define the unconditional transition matrix as:

$$P_{t,T}^{\pi_t} = \mathbb{E}^{\mathbb{Q}}\left[P_{t,T} \mid \pi_t\right]$$

The unconditional transition matrix is only the expectation of the conditional transition matrix. This expression allows us to understand that the risk free transition matrix at time t depends only on the risk premium at time t, the time length over which the matrix runs and the parameters of the risk premium. And among these parameters solely the risk premium at t stochastic.

In practice the risk neutral transition matrix can be obtained by a closed formula. The transition matrix can be diagonalized. For the proof we can see Israel et al. [61]. This diagonalization exists and is unique. We also make the assumption that the diagonalization basis is not time dependant. This assumption is empirically approved in Arvanatis et al. [5]. We then write

$$\Lambda^{historical} = \Sigma\,\mathrm{Diag}(d_1, \ldots, d_K)\Sigma^{-1}$$

With $\{d_i\}_{i=1}^K$ the eigenvalues and Σ the matrix of eigenvectors. Then

$$\begin{aligned} P_{t,t+\Delta t} &= e^{\Sigma\,\mathrm{Diag}(d_1,\ldots,d_K)\Sigma^{-1}\int_t^{t+\Delta t}\pi_s ds} \\ &= \Sigma\,\mathrm{Diag}\left(e^{d_1\int_t^{t+\Delta t}\pi_s ds}, \ldots, e^{d_K\int_t^{t+\Delta t}\pi_s ds}\right)\Sigma^{-1} \end{aligned}$$

The unconditional transition matrix is then given by

$$P_{t,t+\Delta t}^{\pi_t} = \Sigma\,\mathrm{Diag}\left(\mathbb{E}^{\mathbb{Q}}\left[e^{d_1\int_t^{t+\Delta t}\pi_s ds} \mid \pi_t\right], \ldots, \mathbb{E}^{\mathbb{Q}}\left[e^{d_K\int_t^{t+\Delta t}\pi_s ds} \mid \pi_t\right]\right)\Sigma^{-1}$$

Each terms of the diagonal matrix can be computed as

$$\mathbb{E}^{\mathbb{Q}}\left[e^{d_i\int_t^{t+\Delta t}\pi_s ds} \mid \pi_t\right] = e^{A_i(\Delta t) - \pi_t B_i(\Delta t)}$$

with

$$\begin{aligned} A_i(\Delta t) &= \frac{2\alpha\mu}{\sigma^2}\ln\left(\frac{2v_i e^{1/2(\alpha+v_i)\Delta t}}{(\alpha_i+v_i)(e^{v_i\Delta t}-1)+2v_i}\right) \\ B_i(\Delta t) &= -\frac{2d_i(e^{v_i\Delta t}-1)}{(\alpha+v_i)(e^{v_i\Delta t}-1)+2v_i} \\ v_i &= \sqrt{\alpha^2 - 2d_i\sigma^2} \end{aligned}$$

The previous equations give a close formula to compute unconditional transition matrix as long as the risk premium is known. The default probabilities are then easy to deduce. Namely, we have

$$\mathbb{Q}\left(\tau_D^{\eta_t} > T\right) = 1 - \sum_j^{K-1}\sigma_{\eta_t,j}\mathbb{E}^{\mathbb{Q}}\left[e^{d_j\int_t^T\pi_s ds} \mid \pi_t\right](\sigma^{-1})_{j,K} \tag{A.6}$$

where:

- K the index of the default rating in the transition matrix,
- $\sigma_{\eta_t,j}$ the value of Σ on the row η_t and column j,
- $(\sigma^{-1})_{j,K}$ the value of Σ^{-1} on the row j and the column K.

A.4.2 Spread

In this section, we express the risky bond price using the rating process and then define the spread for a rating and a maturity at any time.

We consider a risky zero-coupon bond of maturity T and rated η_t at time t. We define the recovery rate δ as the proportion of the nominal that the owner of the risky zero-coupon bond earns at maturity if default occurs. Then at maturity, the cash flow is 1 if the bond has not defaulted and δ otherwise.

Since the price of a bond is the present value of expected future cash flow and using $\mathrm{DF}(t,T) = \exp\left(-\int_t^T r_s ds\right)$ the deflator from maturity to time t, we can express the price $P^{\eta_t}(t,T)$ of the bond previously defined as

$$P^{\eta_t}(t,T) = \mathbb{E}^{\mathbb{Q}}\left[\mathrm{DF}(t,T)\left(1-(1-\delta)\mathbb{I}(\tau_D^{\eta_t} \leq T)\right)\right]$$

In case of default, the bond becomes a risk free bond with the same maturity but with a nominal reduced by the loss rate which is $(1-\delta)$.

We assume independence between the risk free rate and the rating process. Then we can write

$$P^{\eta_t}(t,T) = P(t,T)\left(1-(1-\delta)\mathbb{Q}(\tau_D^{\eta_t} \leq T)\right)$$

where $\mathbb{Q}(\tau_D^{\eta_t} \leq T)$ is derived from Equation A.6.

We define the spread $s_{t,T}^{\eta_t}$ between t and T for the rating η_t at t as follows:

$$\exp\left(-s_{t,T}^{\eta_t}(T-t)\right) = 1-(1-\delta)\mathbb{Q}(\tau_D^{\eta_t} \leq T)$$

or equivalently

$$s_{t,T}^{\eta_t} = -\frac{1}{T-t}\ln\left(1-(1-\delta)\mathbb{Q}(\tau_D^{\eta_t} \leq T)\right) \tag{A.7}$$

This expression allows us to understand that the spread is stochastic thanks to the risk premium at time t.

A.4.3 Model Calibration

A.4.3.1 Moody's historical transition matrix

The first step consists in recovering the historical transition matrix $P_{t,T}^{historical}$: for example, we retrieve the one from Moody's over the period 1983-2013. The latter one is deduced directly from the market data

However, in the process of estimating credit spreads and default probabilities, it is not this transition matrix that is directly used, but rather its generator. The

	AAA	AA	A	BBB	BB	B	CCC	D
AAA	96,79%	3,30%	0,04%	0,10%	0,00%	0,00%	0,00%	0,00%
AA	3,86%	92,76%	1,69%	0,89%	0,13%	0,00%	0,00%	0,00%
A	0,00%	4,52%	90,51%	3,49%	1,40%	0,08%	0,00%	0,00%
BBB	0,00%	0,00%	6,11%	89,10%	4,13%	0,61%	0,05%	0,00%
BB	0,00%	0,00%	0,00%	8,59%	84,93%	5,31%	0,34%	0,67%
B	0,00%	0,00%	0,00%	0,00%	5,64%	87,89%	2,74%	3,24%
CCC	0,00%	0,00%	0,00%	0,00%	0,00%	13,33%	48,33%	38,33%
D	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	100,00%

Table A.3: Moody's historical transition matrix from 1983 to 2013

second step is therefore the generator estimates associated with the historical transition matrix provided by Moody's. Recall that the equation connecting the historical transition matrix to its generator is given by

$$P_{t,T}^{historical} = e^{(T-t)\Lambda} = \sum_{n=0}^{\infty} \frac{((T-t)\Lambda)^n}{n!}$$

A problem arises: what are the conditions of existence and/or uniqueness to find such a matrix Λ?

We will join here the works of Israel, Rosenthal and Wei [61] which allowed in particular to identify the conditions under which a real generator exists, and how to choose the right generator, that will be compatible with the behavior of the credit ratings. They state in particular a theorem allowing, under the sufficient condition that the diagonal terms of the matrix $P_{t,T}^{historical}$ are strictly greater than 0.5, to express Λ according to the matrix $P_{t,T}^{historical} - I$:

$$\Lambda = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{(P-I)^k}{k}$$

where P stands for the historical transition matrix $P_{t,T}^{historical}$ for the sake of simplicity.

However, this condition does not guarantee the non-negativity of terms located off the diagonal of Λ, preventing it from being a true generator of P. As these terms are generally very small, it is customary to correct the problem:

- by replacing them with 0,
- then adding their initial value to the corresponding diagonal element to preserve the property that the sum of a line must be zero.

This new matrix will therefore have many positive non-diagonal elements and summing lines at 0, guaranteeing the good properties of the generator. This is what we

will choose to do, thus obtaining the next generator Λ as follows:

$$\Lambda = 10^{-2} \begin{bmatrix} -3,336 & 3,199 & 0,011 & 0,095 & 0 & 0 & 0 & 0 \\ 4,079 & -7,836 & 1,814 & 0,932 & 0,115 & 0 & 0 & 0 \\ 0 & 4,937 & -10,254 & 3,8 & 1,502 & 0,034 & 0 & 0 \\ 0,005 & 0 & 6,826 & -12,106 & 4,683 & 0,544 & 0,053 & 0 \\ 0 & 0,011 & 0 & 9,912 & -17,108 & 6,106 & 0,375 & 0,551 \\ 0 & 0 & 0,015 & 0 & 6,564 & -13,777 & 4,139 & 2,532 \\ 0 & 0 & 0 & 0,05 & 0 & 20,244 & -73,957 & 53,72 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

A.4.3.2 Recovery rate

Moody's annually publishes a default risk study, called "*Moody's annual default study*". It includes the history of recovery rates, classified according to the seniority of the debt (from the most secure to the least secure):

1. Senior secured
2. Senior unsecured
3. Senior Subordinated
4. Subordinated
5. Junior Subordinated

For our model, we particularly note a recovery rate of $\delta = 35\%$.

A.4.3.3 Merril Lynch spreads

We recover on Bloomberg the historical data of spreads on the Merrill Lynch bond indices, in the valuation date 31/12/2017.

	AAA	AA	A	BBB
1	0,0016	0,0036	0,0044	0,0068
3	0,0017	0,0046	0,0054	0,0088
5	0,0023	0,0044	0,0067	0,0113
7	0,0026	0,0067	0,0083	0,013
10	0,0069	0,0097	0,0109	0,0177

Table A.4: Spreads on Merrill Lynch bond indices as of 31/12/2017

A.4.3.4 Calibration of the risk premium on spreads

Let us consider the historical spreads presented in section (A.4.3.3). These spreads therefore represent our market spreads as of date of calibration $t = 0$, which will be

noted $s^{mkt}_{k,T}$ where k tracks all the K ratings of maturities $T \in \{T_1, \dots, T_m\}$. The theoretical forward spread is given by the equation (A.7).

In our setting, we choose to fix values of the parameters α, μ, σ_π according to values provided by Moody's analytics in a technical note published in 2003, where they have already performed a historical calibration of the CIR process $\pi(t)$. The study then suggests considering the following parameters:

Parameter	Historical Calibration
α	0,1
σ_π	0,75
μ	3

Table A.5: Historical calibration of the CIR process $\pi(t)$ provided by Moody's

The last parameter to be calibrated is therefore the initial risk premium, which we will choose as the solution of the optimization problem

$$\pi^*(0) = \arg\min_{\pi(0)} \sum_{k\in\{1,\dots,K\}} \sum_{T\in\{T_1,\dots,T_m\}} \left[s^{mkt}_{k,T} - s^{model}_{k,T}(\pi(0)) \right]$$

Thanks to the Python function *scipy.optimize.minimize*[5], we obtain $\pi^*(0) = 3.3615$ and the following spread curves:

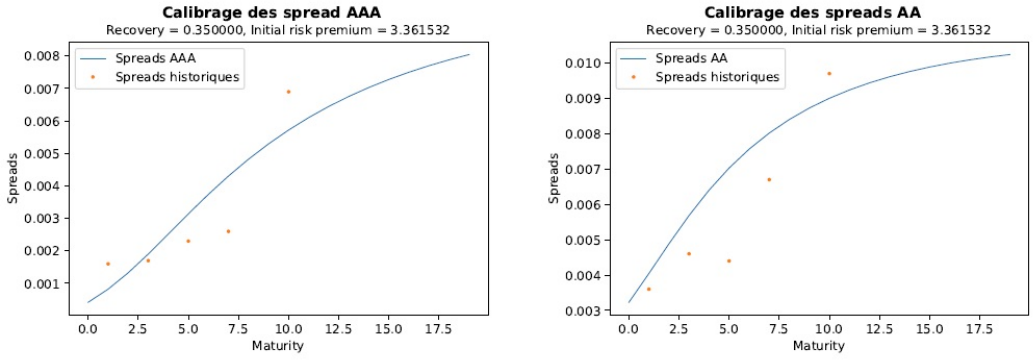


Figure A.5: Credit spread curves for the rating AAA and AA

[5]For the documentation of this Python function, please refer to https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.minimize.html

Appendix B

Asset-Liability Management

B.1 Introduction

The insurer's job is to provide financial protection to people who want to transfer some of their risks for a premium. Its role is to better manage the risks received to be able to honor its commitments at any time. For this, it has two levers: technical lever that is on the liabilities side and financial lever that is on the assets side, but cannot be separated from liabilities. To this end, the insurer need a Asset-Liability Management (ALM) simulator. This tool simulates the economic and accounting behavior of a "Euro Funds" type life insurance fund. The model takes into account for each fund (liability scenario) various assumptions on the evolution of markets (asset scenario) and a number of management rules:

1. Liability scenario: can be stochastic or deterministic. If it is deterministic, it still has correlations with the asset, such as profit-sharing, dynamical redemption. In addition, the model make it possible to treat different tranches of liabilities within the same fund. These slices are called *model points*. Segmentation is done following factors that may influence the fund's behavior (Age, gender, TMG[1] type contract, behavior of the insured, etc.) and thus increases the accuracy of modeling. Each model point is managed individually by the model although the allocation assets and financial products are pooled.

2. Asset scenario: can be also deterministic or stochastic. These scenarios return market performance for a number of asset classes (bonds, equities, etc). All these scenarios are provided by the ESG presented in Chapter A.

3. Management rules: These are all the rules for managing the fund. These rules are various type (accounting, economic, contractual, etc).

In our setting, we limit ourselves to the deterministic liability scenario, but the stochastic asset scenarios. The simulations are carried out over a period of 50[2] years by annual time step. In practice, the projection horizon should cover the entire life of all incoming cash flows that are required to fulfill contractual commitments. In our case, the determination of the projection horizon should match the farthest expected lapse of the contract of the portfolio (i.e. run-off mode).

[1] Guaranteed minimum rate

[2] The projection horizon is modifiable in our modeling tool.

In this chapter we will present the implementation of our ALM model. It's a Python-coded tool that can estimate the Best Estimate Liabilities of a life insurance company marketing euro savings products using a stochastic approach. Our ALM model was built based on our benchmark of market practices concerning the implementation of Pillar I of the Solvency II Directive as introduced in Section 1.7.1. As it concerns the privileged and confidential documents, we will not cite any relative references in this report. Regarding other materials such as mathematical provision, capitalization reserve, etc., readers can refer to the "Code des Assurances" [3]

We will detail in a first time the operation of the tool, the different models implemented and the simplifications made. Then we will present and analyze the results obtained.

B.2 Saving contract

Saving is simply a matter of placing money that becomes unavailable for immediate payments and current consumption. The investment can be made on products offered by financial institutions or insurers. Their return varies depending on the type of investment, the lock-up period and the rate of pay set by the contract.

Savings product are the answer to different needs:

1. save without a specific goal or just as a precaution
2. finance the short and the medium term
3. value or grow a capital
4. provide additional income for retirement

In order for everyone to find the product in line with their needs, there is a very diversified range of products. We will focus only on saving contracts in the life insurance business. In particular, we will study its characteristics and the accounting mechanisms involved.

The life insurance savings contract looks like a financial investment and is close to a capitalization contract. But it is still a contract of insurance, which is not a fixed-term product and is intended to cover you until you pass away. Life insurance is based on the lifetime capitalization technique. This means that during the term of the contract, the subscriber does not receive any income, apart from the possible payment of interest and profit sharing. The premiums paid by insureds are thus immediately reinvested and incorporated into savings, thus becoming interest-bearing. However this is not a purely financial investment since it involves both a lifetime parameter (the mortality rate) and a financial parameter (the profit-sharing rate). Indeed, the benefits are conditioned by the occurrence of certain events such as the death of the insured during the term of the contract or by the redemption of the contract. It is therefore necessary for the modeling of a contract to have available

[3] Available at `https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000006073984`

mortality tables indicating the number of living at each age of human life, and also the laws of redemptions depending on the type of contract. In the following, we will briefly present the different types of life insurance saving contract.

Secure Funds in Euros

These are contracts with minimal risk as they are mostly invested in bonds. They therefore have a yield directly linked to bond rates and are therefore not very sensitive to the vagaries of the stock market. They also have a double guarantee: 1) a guaranteed minimum return and 2) a "ratchet effect" that allows the subscriber to keep definitively the annual interest credited on the contract.

Unit-linked life insurance contract

Unit-linked contracts are contracts that do not refer to a currency but to units of account, i.e. shares , securities or real estate. These contracts provide diversified investment in the financial and real estate markets. They are chosen by long-term investors who are willing to accept the risks inherent in financial market fluctuations to obtain a higher expectation of earnings than a conventional bond-type contract like the secure funds in euros.

Multi-vehicle life insurance contract

In this type of contract, investments are made in several supports or funds (in euros and / or in unit-linked contracts). Depending on the contract, the distribution of the investment is free, imposed or pre-established. These contracts benefit from more than one possibility of arbitrage between the support in unit-linked and the supports in euros (the arbitrage is an operation which consists of modifying the distribution of the capital between the various supports of the contract).

It is thus possible to divide its investments between more or less risky support. Several risky profiles are often proposed: prudential, dynamic, balanced. The subscriber then entrusts the financial experts to manage its payments according to the chosen profile.

B.2.1 Characteristics of a saving contract

The subscriber pays premiums which are capitalized at the guaranteed minimum rate to constitute the guaranteed capital. It also revalued taking into account profit-sharing. It should be noted that in case of death before term, the capital is paid to the beneficiary designated in the contract.

Premiums

Premiums can be made in different forms:

1. Scheduled periodic contracts: a payment schedule (monthly, quarterly, annual) is set up with most of the time the possibility of making additional payments; this is a payment option and not a firm commitment since the insured can stop payments at any time.

2. Flexible payment contracts: there is no payment schedule but the insured is often subject to a minimum amount of contributions.

3. Single payment contracts: the payment takes place at the time of subscription.

The insurer cannot demand the payment of premiums. Non-payment from the insured entails either the reduction of the contract (continuation of the contract, but reduction of the amount of the guaranteed benefits), or the outright cancellation of the contract.

Expenses

They can be very different from one company to another. It exists in particular (this is a non-exhaustive list):

1. Acquisition fees: this may be a percentage taken from payments or a lump-sum per policy.

2. Administration fees.

3. Management fees related to the investment of the fund: they are deducted from the savings (i.e. on the policy liabilities), during the annual capitalization, on the interest generated by the fund.

4. Commission fees related to distribution networks.

5. Arbitrage fees (in case of multi-support contract): they are calculated on the sums transferred in the event of a change of support.

Redemption option or policy loan (advance) during the contract

In case of need of money before the end of the contract, it is possible to request a partial or total surrender (i.e. repurchase agreement) insofar as the contract has a cash value. However, there may be considerable penalties (expressed as percentages of the mathematical provision) depending on the residual life of the contract. This indemnity intended for the insurer cannot, however, exceed 5% of the mathematical provision and becomes nil after a period of ten years from the effective date of the contract.

Partial Surrender: corresponds to the payment by the insurer of a part of the mathematical provision.

Total Surrender: terminates the contract and allows the insured to recover the value of his fund before the end of the contract.

Advance: allows the insured to obtain a sum of money without reducing the savings. The insurer agrees to advance funds in the form of a loan that will have to be repaid by the insured. The amount that can be borrowed is capped at 1% of the mathematical provision. The advance is granted at an interest rate and for a variable amount depending on the contract. It should be noted that an advance cannot be granted on a periodic premium contract.

B.2.2 Accounting in insurance companies-Basic concepts

In the following, we will discuss some accounting elements that relate to life insurance and that will be useful for understanding management decisions and asset-liability management.

Mathematical provision (PM)

Insurance companies have to set up sufficient technical provisions for the full settlement of their commitments in relation to policyholders or beneficiaries of the contracts. The mathematical provision consists of the funds that life insurance companies set aside to meet the commitments they made to their policyholders. It is defined as the difference between the commitment of the insurer and that of the insured. In other words, it represents the net insurer's liability for policyholders' liabilities.

Profit-sharing reserve (PPE)

The profit-sharing reserve is defined as the amount of profit sharing attributed to policyholders which is not repaid immediately. In accordance with the Insurance Code, all amounts allocated to the profit-sharing reserve must be returned to the insured within 8 years. The temporal distribution of this provision is thus left to the discretion of the insurers, which allows them to smooth the profit-sharing rate served to policyholders or to attract new customers by proposing a revaluation rate higher than the average during the first years of the contract thanks to an attractive profit-sharing rate.

Capitalization reserve (RC)

The capitalization reserve is a reserve fueled by the capital gains realized on bond sales and taken over symmetrically only in the event of realized capital losses on this type of asset. This makes it possible to smooth the results corresponding to the gains or losses realized on bonds sold before their term, in case of movements of interest rates. Thus, insurance companies are not encouraged, in case of falling interest rates, to sell their bonds with high coupons and to generate one-off profits while buying other, less performing bonds at a later date. This special reserve, considered as a provision in relation to the hedging requirements of the commitments, forms part of the solvency margin.

Liquidity risk provision (PRE)

This is a regulated technical provision in insurance, which arises when the non-amortizable investments are in a total net unrealized loss position. As will be presented below in Section B.3.1, the only non-amortizable investment in our assets is the equity portfolio. This provision therefore corresponds to that for risk on the equity portfolio and will be calculated as follows:

$$\mathrm{PRE}_t = \min\left(\mathrm{PRE}_{t-1} + \frac{1}{3}\mathrm{MVL}_t^{\text{equity}}; \mathrm{MVL}_t^{\text{equity}}\right). \tag{B.1}$$

From this it is easily seen that a payment or a drawn-down is made on the liquidity risk provision according to its value last year and the unrealized capital losses on the equity portfolio. Namely, we have

$$\Delta\mathrm{PRE} = \begin{cases} < 0, & \text{if } \mathrm{PRE}_{t-1} > \mathrm{MVL}_t^{\text{equity}} \\ \geq 0, & \text{if } \mathrm{PRE}_{t-1} \leq \mathrm{MVL}_t^{\text{equity}} \end{cases}$$

As will be seen later, we assume that the recovery or the provisioning of the liquidity risk provision is made before the decision on profit-sharing rate. This will thus impact the financial incomes as well as the profit-sharing which will be distributed to the insureds.

B.3 General presentation of the ALM simulator

Our ALM modeling tool makes it possible to summarize the annual cash flows such as premiums, claims, changes in provisions, financial income, management expenses, etc. and calculate the margins and then record them in the income statement and balance sheet. In addition, the balance sheet shows the stocks of provisions as well as the market and book values of the assets that are backed by the liabilities at each time step.

The simulation technique chosen in our tool is the Monte-Carlo method based on the law of large numbers. This involves performing a large number of scenarios independently, in order to obtain an approximation close to the true Best Estimate. The tool takes as input N neutral risk economic scenarios generated by our ESG. For each scenario, the tool projects the assets and liabilities of the insurance company over 50 years, while performing the Asset/Liability interactions according to a pre-defined algorithm. The Best Estimate Liabilities (BE) is thus calculated according to the following formula:

$$\mathrm{BE} = \mathbb{E}^{\mathbb{Q}}\left[\sum_{t=1}^{50} \mathrm{DF}(0,t)\mathrm{CF}(t)\right] \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{50} \mathrm{DF}^{(i)}(0,t)\mathrm{CF}^{(i)}(t) \tag{B.2}$$

where:

1. $\mathrm{DF}^{(i)}(0,t)$ is the discount factor in scenario i,

2. $\mathrm{CF}^{(i)}(t)$ is the liability cash flow at time t in scenario i.

3. N is the number of simulations.

In our model, we do not simulate the expenses and the taxes. Therefore, the scope of the liability cash flows only recovers: the claims (redemptions and death benefits) Cl_t and the premiums P_t. Namely we have

$$\mathrm{CF}(t) = \mathrm{Cl}_t - \mathrm{P}_t.$$

In order to facilitate the change of the model assumptions, the different elements of the balance sheet are developed in different classes. Their interactions are shown in Figure B.3.

B.3.1 Description of the Asset

In order not to burden the modeling, we simplified the modeling of the asset with the following assumptions:

1. The insurance undertaking's asset portfolio consists solely of the following asset classes:

- Cash remunerated at risk free rate,
- Equities
- Fixed-rate government and corporate bonds

2. The financial market where our assets are located is supposed to be perfectly liquid. In addition, the assets are infinitely divisible and can be purchased without transaction costs. In other words, we can sell or buy assets at any time in the quantities desired.

3. The assets allocation is defined at the beginning of the projection and the insurance company keeps the same allocation of the portfolio throughout the projection. Thus, at the end of each year, the portfolio allocation in terms of market value is identical to the initial one. Finally, the asset portfolio is allocated as follows:

- 10% cash,
- 20% equities,
- 70% bonds.

B.3.1.1 Bonds valuation

Recall the equation (A.7), defining the forward spreads directly according to the default probability. In particular, we will generate the actuarial spreads for each rating, defined as

$$S^{\eta_t}(t,T) = P(t,T)^{-\frac{1}{T-t}} \left(e^{s^{\eta_t}_{t,T}} - 1 \right) \tag{B.3}$$

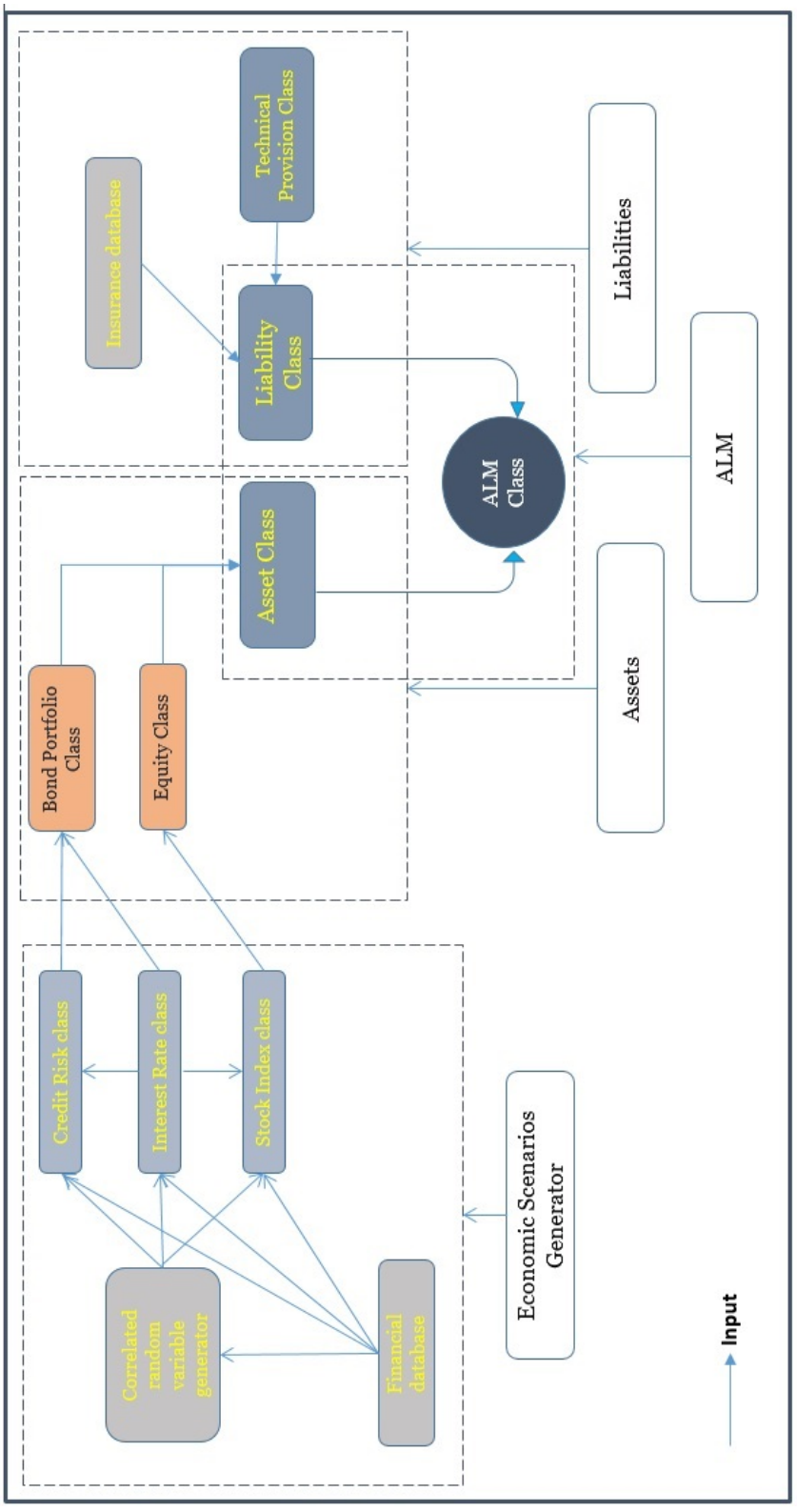


Figure B.1: ALM modeling structure

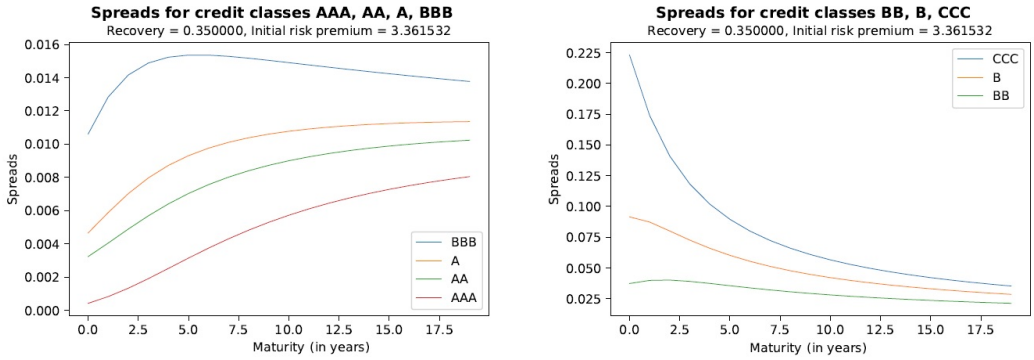


Figure B.2: Credit spread curves

In particular, these credit spreads make it possible to value risk bonds by adjusting the value of the deflators. More precisely, the deflator $\mathrm{DF}^{\eta_t}(t,T)$ for the rating η_t adjusted for the corresponding credit spread $S^{\eta_t}(t,T)$, is therefore expressed as:

$$\mathrm{DF}^{\eta_t}(t,T) = \frac{1}{\left(1 + R(t,T) + S^{\eta_t}(t,T)\right)^{T-t}}. \tag{B.4}$$

We can also draw the curves of the deflators, from zero coupon rates of different maturities $R(0,T)$ and spread values $S^{\eta_0}(0,T)$. These deflators will be particularly useful for carrying out the martingale test in the following part.

The market value in $t = 0$ of an obligation for the rating class η, maturity T, coupon $\{c_t\}_{t=1,\dots,T}$ and nominal 1 is none other than

$$V^{\eta}(0,T) = \mathrm{DF}^{\eta}(0,T) + \sum_{t=1}^{T} \mathrm{DF}^{\eta}(0,t) c_t \tag{B.5}$$

We choose, for the sake of simplicity, to simulate ratings migrations and defaults proportionally to the simulated transition matrices at each time step. This allows us in particular to rewrite $V^{\eta}(0,T)$ according to the recovery rate δ and the elements of the transition matrix such as

$$V^{\eta}(0,T) = \mathrm{DF}^{\eta}(0,T) f_T + \sum_{t=1}^{T} \mathrm{DF}^{\eta}(0,t) \left[\delta f^d_{t-1,t} + c_t f_t\right] \tag{B.6}$$

where f_t represents the proportion of the obligation that is not in default in t, and $f^d_{t-1,t}$ is the proportion of the obligation that was not in default in $t-1$ but that is in t.

Finally, let us also remind that, theoretically, the price of this obligation is expressed as a function of the recovery rate δ and the probability of default as

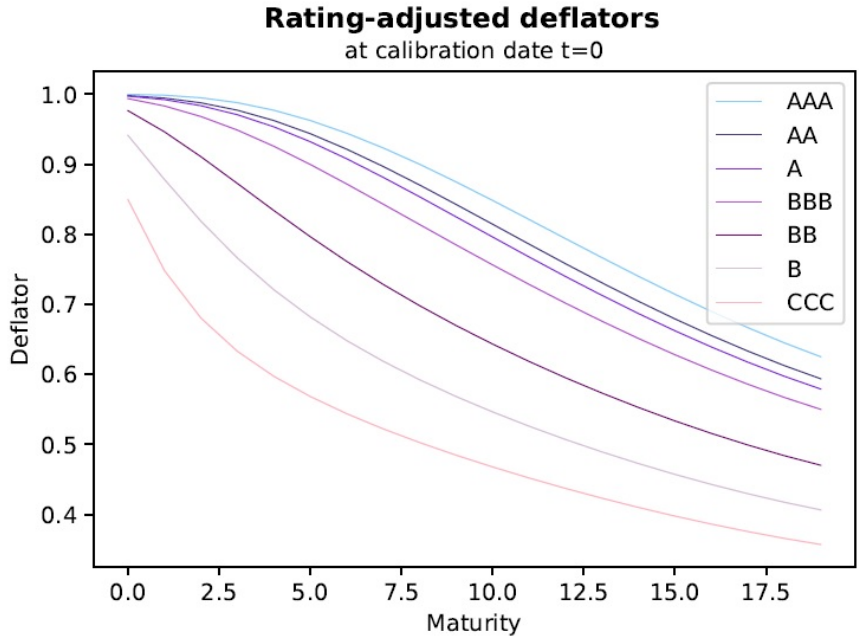


Figure B.3: Adjusted deflator curves

follows:

$$v^{\eta}(0,T) = P(0,T)\,(1 - q_{\eta,K}(0,T)) \\ + \sum_{t=1}^{T} P(0,t)\,(c_t[1 - q_{\eta,K}(0,t)] + \delta[q_{\eta,K}(0,t) - q_{\eta,K}(0,t-1)])$$

where $q_{\eta,K}(0,u)$ is the probability of default given the initial rating η which is of the form (see Section (A.4.1) for more detail):

$$q_{\eta,K}(0,u) = \sum_{j}^{K-1} \sigma_{\eta,j}\,(A_j(u) - \pi_0 B_j(u))\,(\sigma^{-1})_{jK}.$$

Martingale test The last step in implementing a neutral risk model is to check the quality of the simulation and the parameters. For this, we carry out a martingale test, aiming to test if the price of the assets is equal to their discounted future flows simulated under the risk-neutral probability. This approach makes it possible to check the concept of market consistency, that is to say the property ensuring that the discounted prices are indeed martingales.

Note that for $i \in [1, N]$ with N the number of simulated trajectories, the present value of the bonds price is

$$V_i^{\eta}(0,T) = D_i^{\eta}(0,T) + \sum_{t=1}^{T} D_i^{\eta}(0,t) c_{i,t}$$

where $D_i^\eta(0,T)$ and $c_{i,t}$ respectively correspond to the deflator between $t=0$ and T for scenario i, and the simulated cash flow at time t for scenario i. The law of large numbers allows us in particular to affirm that

$$\frac{1}{N}\sum_{i=1}^{N} V_i^\eta(0,T) \xrightarrow{N\to\infty} \mathbb{E}^{\mathbb{Q}}\left[V_1^\eta(0,T)\right]$$

or equivalently

$$\widehat{V}^\eta(0,T) = \frac{1}{N}\sum_{i=1}^{N} V_i^\eta(0,T) \xrightarrow{N\to\infty} v^\eta(0,T).$$

Thus, we try to verify that we have, for each T maturity, the convergence of the empirical average of the discounted cash flows towards the price of the corresponding maturity bonds. On the other hand, the central limit theorem states that

$$\sqrt{N}\left(\widehat{V}^\eta(0,T) - v^\eta(0,T)\right) \xrightarrow{N\to\infty} \mathcal{N}\left(0, \sigma^2_{\eta,T}\right)$$

where $\sigma^2_{\eta,T}$ is the variance of $V_1^\eta(0,T)$. So we can easily build the associated confidence intervals.

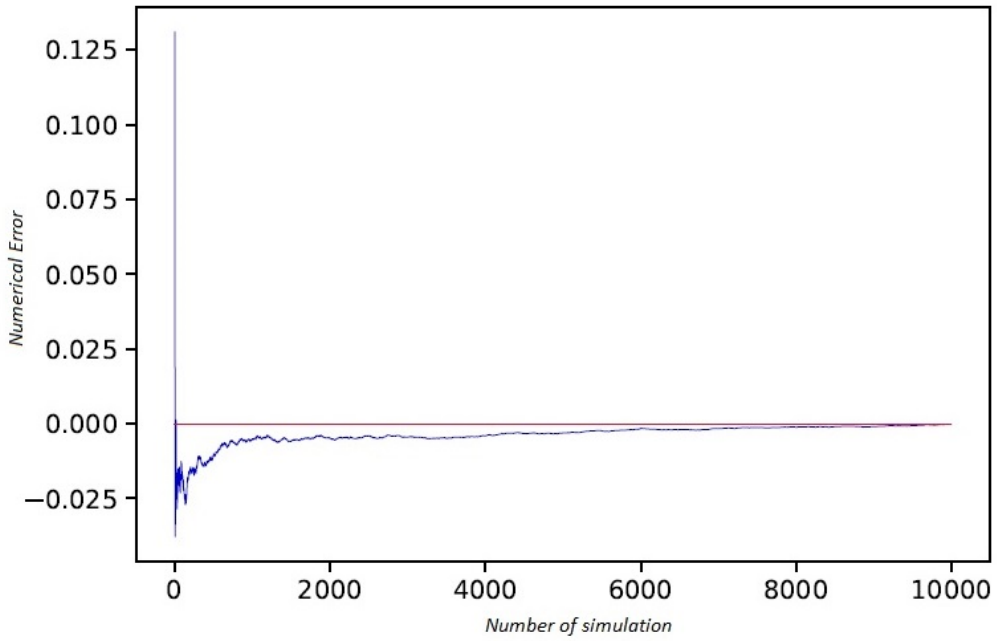


Figure B.4: Martingale test on a 10 year maturity AAA-rated bond

B.3.2 Description of the Liability

This subsection concerns the liabilities modeling. We begin by exposing the simplifying assumptions on the liabilities modeling. In order to determine the Best

Estimate, no new policy is underwritten during the projection. It is said that the insurance company operates in mode run-off. Therefore, only future premiums for current policies should be taken into account. Finally, we consider that the insurance company does not use reinsurance to give up part of its risk.

To model the liabilities evolution, the following assumptions are applied:

- The mortality rates are assumed to be deterministic depending on the age of the insured.
- Deaths and redemptions occur mid-year.
- The insurance policies are grouped into homogeneous groups (Model Points) according to discriminating criteria as will be seen later.
- At the end of 50 years, the activity of the company comes to the end, that is to say that all remaining policyholders will buy their contract and the mathematical provision of the insurer will become null.

Model Point

Our liability portfolio is entirely fictitious. It was built from ten saving products close to those that could be found in the liabilities of a life insurer today.

id	Age	Sexe	Subscription _Date	Valuation_ Date	Premium	PM_ouverture	TMG	Rate sensibility	Margin Rate	Number_of _Contract	Lapse
1	30	TH	01/06/2013	20/01/2016	300	100000	TMGA	1%	1%	5229,61581	1%
2	35	TH	01/06/2012	20/01/2016	300	100000	TMGA	1%	1%	4374,21798	1%
3	40	TH	01/06/2011	20/01/2016	300	100000	TMGA	1%	1%	3783,69509	1%
4	45	TH	01/06/2010	20/01/2016	300	100000	TMGA	1%	1%	3831,37181	1%
5	50	TH	01/06/2009	20/01/2016	300	100000	TMGA	1%	1%	3340,31351	1%
6	55	TH	01/06/2008	20/01/2016	300	100000	TMGX	1%	1%	2353,80966	1%
7	65	TH	01/06/2007	20/01/2016	300	100000	TMGX	1%	1%	2104,18238	1%
8	75	TH	01/06/2006	20/01/2016	300	100000	TMGX	1%	1%	2918,1644	1%
9	85	TH	01/06/2005	20/01/2016	300	100000	TMGX	1%	1%	3012,2169	1%
10	95	TH	01/06/2004	20/01/2016	300	100000	TMGX	1%	1%	2672,2479	1%

Figure B.5: Summary of our liability portfolio

To further simplify our model points, the following characteristics, apart from the age, subscription date and initial number of contract, necessary for the projection of the liabilities are identical. Therefore, for each model point, we have:

- Opening mathematical provision: 100.000 euros,
- There are two types of annual guaranteed minimum rate: one for TMGA: 2% and another for TMGX 0.1%,
- Periodic premium: 300 euros,
- Structural Lapse rate: 1%,
- Rate sensibility, margin rate will be detailed in the following.

Lapse or surrender

In our setting we only consider the partial surrender. The amount of redemptions is modeled as a ratio, called the lapse rate, of the mathematical provision per model point. There are two main types of surrender to be distinguished according to the elements that trigger them:

Structural surrenders: They are often linked to tax benefits because the capital gains made in life insurance are subject to income tax with a declining scale. Indeed, we often see a redemptions peak at 8 years of seniority. Structural surrenders can also come from an imminent need for liquidity. Policyholders buy back their capital to finance their personal projects or to cover an unforeseen risk, even if the buyout is not favorable to them in terms of profitability. An example of the experimented structural lapse rate for the saving contracts is presented in Figure B.6. For the sake of simplicity, the structural lapse rate (RS) is supposed to be constant and deterministic for every model points in our model.

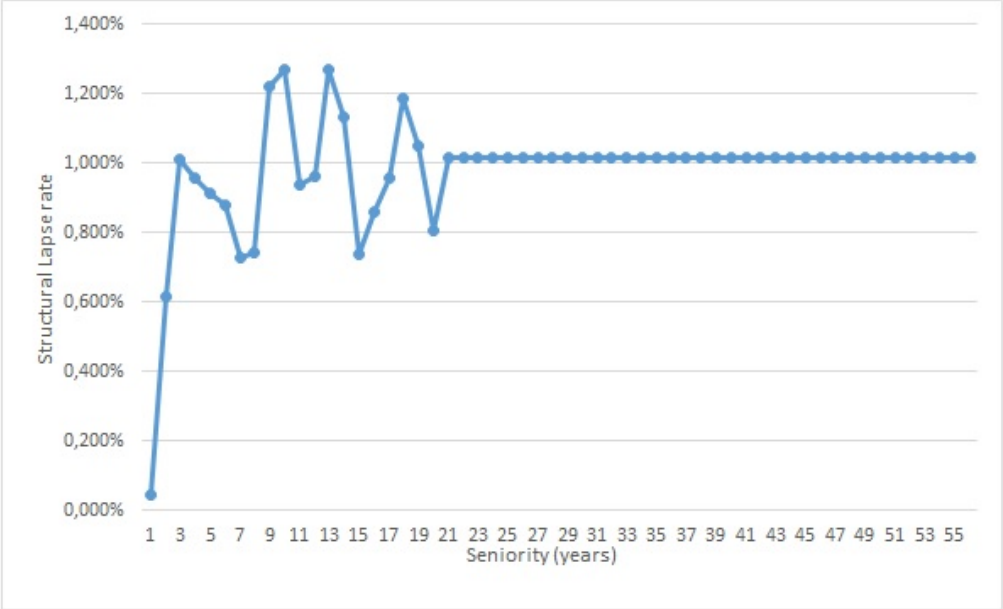


Figure B.6: Here we plot an example of the experimented structural lapse rate

Cyclical surrenders: They are closely linked to market conditions and also depend on the macroeconomic context, the legislation or the reputation of the insurance company. Cyclical surrender modeling is a problem for insurance companies. They cannot establish a historical calibration as they do for structural surrender. To overcome this obstacle, we build a cyclical buyback law based on the spread between the profit-sharing rate and the rate expected by the insured. Namely, the curve is the "average" between the upper and lower cyclical buyback laws proposed by ACPR (French Prudential Supervision and Resolution Authority) [1].

It is considered that the cyclical behavior of the insured is triggered by the finding of a difference between the profit-sharing rate and the rate expected by the insured or target rate for short.

The cyclical behavior and its intensity are triggered by thresholds and are based in particular on three reference rates: the target rate, the floor rate (plafond min) and the ceiling rate (plafond max), as shown in Figure B.7.

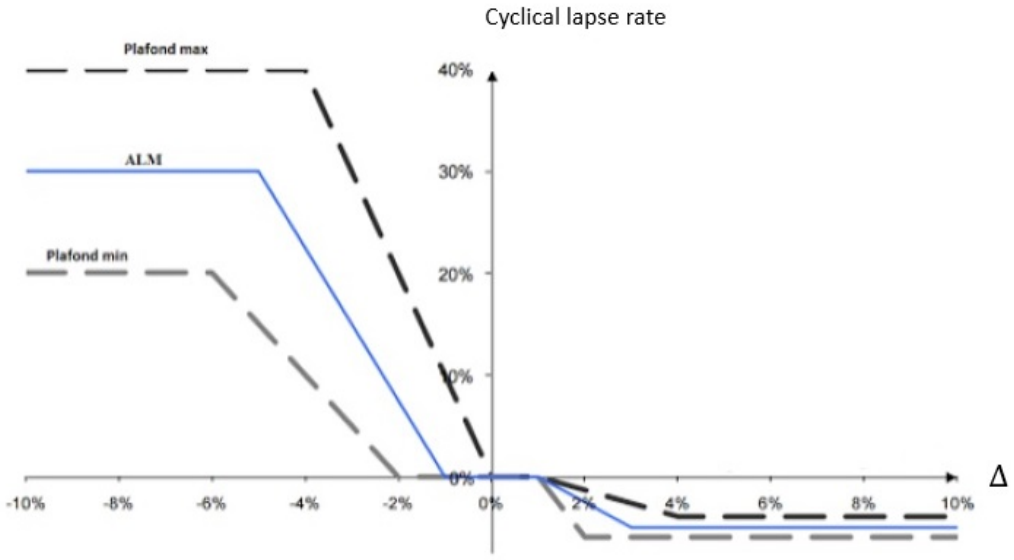


Figure B.7: Cyclical lapse rate curve

To be more precise, the cyclical lapse rate (RC) for the model point j at a given point in time t is defined as

$$\mathrm{RC}_{j,t} = \begin{cases} RC_{\max} & \text{if } \Delta_{j,t} < \alpha \\ RC_{\max}\frac{\Delta_{j,t}-\beta}{\alpha-\beta} & \text{if } \alpha \leq \Delta_{j,t} < \beta \\ 0 & \text{if } \beta \leq \Delta_{j,t} < \gamma \\ RC_{\min}\frac{\Delta_{j,t}-\gamma}{\delta-\gamma} & \text{if } \gamma \leq \Delta_{j,t} < \delta \\ RC_{\min} & \text{if } \Delta_{j,t} \geq \delta \end{cases}$$

where $\Delta_{j,t}$ is the spread between the profit sharing rate (PSR) and the target rate (TR),i.e. $\Delta_{j,t} = \mathrm{PSR}_{j,t-1} - \mathrm{TR}_{j,t}$ and the parameters $\alpha, \beta, \gamma, \delta$ are given by

Parameter	Plafond min	Plafond max	ALM
α	-6%	-4%	-5%
β	-2%	0%	-1%
γ	1%	1%	1%
δ	2%	4%	3%
RC_{min}	-6%	-4%	-5%
RC_{max}	20%	40%	30%

This redemption law is divided into three zones reflecting three behaviors of policyholders according to the difference between the profit-sharing rate and the expected rate:

- *Unfavorable situation*: the profit-sharing rate is below the target rate, i.e. $\Delta < 0$, which is below the insured's expectations. Positive cyclical surrenders are triggered.
- *Favorable situation*: the profit-sharing rate is included in the $[\text{TR}+\beta, \text{TR}+\gamma]$ interval, which is close to the rate expected by the insured. Cyclical surrenders are void.
- *Very favorable situation*: The profit-sharing rate is beyond the target rate, i.e. $\Delta > 0$, which is beyond the expectations of the insured. Cyclical surrenders are negative and offset all or part of the structural buybacks.

Finally, the lapse rate (LR) is calculated as the sum of the structural lapse rate and the cyclical lapse rate as follows:

$$LR_{j,t} = \min\left(1, \max\left(0, RS + RC_{j,t}\right)\right) \tag{B.7}$$

B.3.3 Chronology of the Asset-Liability interactions

In the framework of the calculation of the BEL, the ALM model makes it possible to project over a given time horizon the assets and the liabilities of the balance sheet. It allows to determine at each time step the balance sheet and the value of the cash flows distributed to the policyholders on the one hand and to the shareholders on the other hand.

During each projection period, the following operations are performed:

- *Liability cash flows*: The insurer collects the premiums and pays the benefits relating to the contracts taken out by the insureds. In practice, the liabilities cash flows are collected and paid throughout the year. In order to facilitate the calculation of cash flows, we assume that the liabilities cash flows occur mid-period within the stochastic projection model. The modeling of liability flows within the stochastic projection model is presented in Section (B.3.2).
- *Financial production*: The insurer receives the returns from the assets held in the portfolio.

The interactions between assets and liabilities occur at the end of each period over the duration of the stochastic projection. These interactions are as follows:

- *Purchase and sale of assets*: Invest the premiums collected during the period (purchase of assets); finance the claims paid during the period (sale of assets).
- *Profit-sharing strategy*: profit-sharing rates are calculated for contracts eligible for profit-sharing which takes into account regulatory and contractual constraints. The profit-sharing strategy is described in Section B.3.4.
- *Calculation of liabilities at the end of the period*: the contracts are revalued via the corresponding profit-sharing rates calculated previously. The explicit calculation of the liabilities at the end of the period is presented in Section [?].

  Note that at the end of the projection (after the profit-sharing strategy), the residual general reserves deemed to belong to the insurer are included in the profit, those deemed to belong to the insured are included in the BEL.
- *Assets reallocation*: In accordance with the target asset allocation, the insurer will buy or sell assets to meet the predefined target asset allocation at the end of the period.

The modeled general reserves (capitalization reserve (RC), provision for risk on the equity portfolios (PRE) and profit-sharing provision (PPE)) are recalculated following the purchases and sales of assets.

The following illustration (Fig. B.8) shows the chronology of operations performed during a projection period.

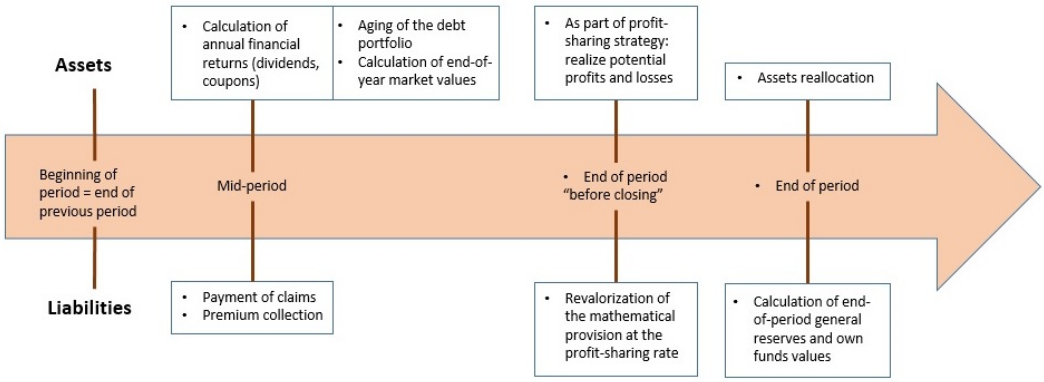


Figure B.8: Chronology of the Asset-Liability interactions

B.3.4 Profit-sharing strategy

The profit-sharing strategy is a central element of the ALM model. On the one hand, it plays a role of leverage in the business development of the insurance company,

because it has an impact on the satisfaction level of policyholders, and on the other hand, it is subject to regulatory constraints aimed at protecting the interest of the insured.

In the following, we will give the details of this profit-sharing strategy.

Calculation of Profit-sharing rate

The insurer must at least serve the guaranteed minimum rate (TMG) as defined in the contract with its insureds. This amount can only be financed by the financial results of the company. In case of insufficient incomes, the insurer will draw on its own funds and will realize a loss.

In any case, the insurer cannot use the PPE from previous years to serve the TMG. But once the wealth acquired in this year makes it possible to serve the TMG, the insurer has the right to take over the PPE to provide policyholders a better revaluation rate. More specifically, the insurer will serve the insureds a so-called *net desired rate* (NDR), which is defined by

$$\mathrm{NDR}_{j,t} = \max\left(\mathrm{TR}_{j,t} - \mathrm{RS}_j; \mathrm{TMG}_j\right) \tag{B.8}$$

where RS_j is the rate sensibility for the model point j mentioned in Section B.3.2. But what is exactly the meaning of the rate sensibility? In fact, the target rate corresponds to the desired rate on products subject to strong commercial pressure (flagship products). On less exposed (non-flagship) contracts, the policyholder will be less sensitive and the expected rate may be reduced by a spread. This latter is set according to the policyholder's sensitivity to its profit-sharing rate. This is why the spread is called the rate sensitivity. For the flagship products, this spread is set to 0.

For the latter use, we define the *gross desired rate* (GDR) and the *gross TMG* (GTMG) for the model point j as

$$\mathrm{GDR}_{j,t} = \mathrm{NDR}_{j,t} + \mathrm{MR}_j \tag{B.9}$$

and

$$\mathrm{GTMG}_j = \mathrm{TMG}_j + \mathrm{MR}_j \tag{B.10}$$

where MR_j is the margin rate mentioned in Section (B.3.2).

Now the question is what profit-sharing rate the insurer will be able to serve. To this end we define at first different notions of wealth.

First, we define the targets that the insurer seeks to achieve: the "TMG wealth" and the "desired wealth". They are obtained with a capitalization of the mathematical provisions with respectively the TMG and the net desired rate. Namely, we have

$$\begin{aligned}\mathrm{Wealth}_{j,t}^{\mathrm{TMG}} = {} & \mathrm{PM}_{j,t-1}(1+\mathrm{GTMG}_j) + (\mathrm{P}_{j,t} - \mathrm{Cl}_{j,t})\sqrt{1+\mathrm{GTMG}_j} \\ & - \left(\mathrm{PM}_{j,t-1} + \mathrm{P}_{j,t} - \mathrm{Cl}_{j,t}\right)\end{aligned}$$

and

$$\begin{aligned}\text{Wealth}_{j,t}^{\text{desired}} = \text{PM}_{j,t-1}(1+\text{GDR}_j) + (\text{P}_{j,t} - \text{Cl}_{j,t})\sqrt{1+\text{GDR}_j} \\ - \ (\text{PM}_{j,t-1} + \text{P}_{j,t} - \text{Cl}_{j,t})\end{aligned}$$

By definition (B.8), it is clear that $\text{Wealth}_{j,t}^{\text{desired}} \geq \text{Wealth}_{j,t}^{\text{TMG}}$. Then we define the total TMG wealth and the total desired wealth as:

$$\text{Wealth}_t^{\text{TMG}} = \sum_{\substack{j\in \\ \text{Model Points}}} \text{Wealth}_{j,t}^{\text{TMG}}$$

and

$$\text{Wealth}_t^{\text{desired}} = \sum_{\substack{j\in \\ \text{Model Points}}} \text{Wealth}_{j,t}^{\text{desired}}$$

Once the targets were identified, we seek to achieve them by using the available wealth. This latter one is complemented by the achievement of unrealized gains (PVL) or unrealized losses (MVL). We can therefore define three levels of wealth that the insurer can achieve through the realization of unrealized gains and losses (PMVL): minumum wealth, maximum wealth and available wealth. However, in the case of the realization of MVL, it is necessary to take into account the partial (or total) compensation of the MVL realized by the release of the capitalization reserve, as illustrated in Figure B.9.

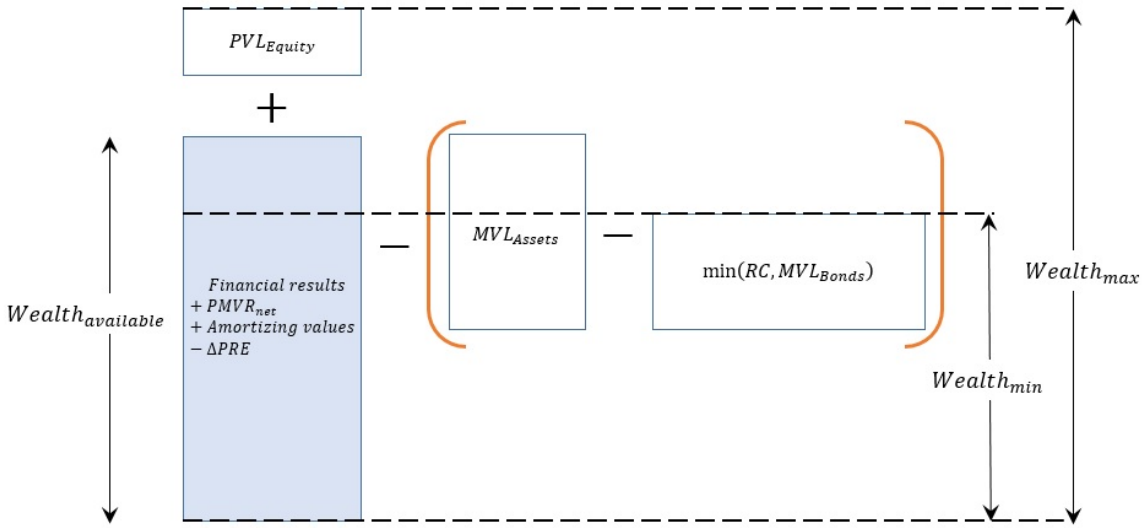


Figure B.9: Graphical illustration of the available wealth, minimum wealth and maximum wealth.

In the following, we list the representative cases of how we calculate the profit-sharing rate.

Case 1: $\text{Wealth}_t^{\text{TMG}} \geq \text{Wealth}_t^{\max}$:

In this case the insurer has to first draw on its own funds a certain amount in order to serve the TMG and then use the PPE to reach a) the desired rate or b) the corresponding maximum profit-sharing rate that we can serve (see Figure B.10).

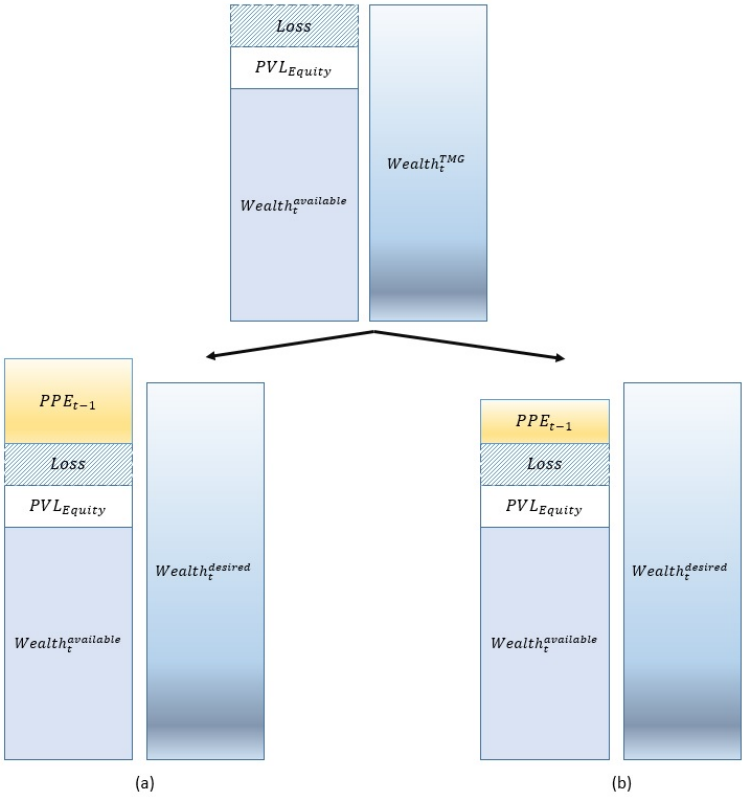


Figure B.10: Case 1: $\text{Wealth}_t^{\text{TMG}} \geq \text{Wealth}_t^{\max}$

Case 2: $\text{Wealth}_t^{\text{available}} \leq \text{Wealth}_t^{\text{TMG}} \leq \text{Wealth}_t^{\max}$ and $\text{Wealth}_t^{\text{desired}} \leq \text{Wealth}_t^{\max} + \text{PPE}_{t-1}$:

The insurer must first realize a portion of unrealized gains to serve the TMG before the resumption of the PPE to serve the desired rate. (see Figure B.11). More precisely, the unrealized gains needed to be realized is given by

$$\begin{aligned}\text{PVR}_{\text{Equity}} = \min\{&\max(\text{Wealth}_t^{\text{TMG}} - \text{Wealth}_t^{\text{available}};\\ &\text{Wealth}_t^{\text{desired}} - \text{Wealth}_t^{\text{available}} - \text{PPE}_{t-1});\\ &\text{PVL}_{\text{Equity}}\}.\end{aligned}$$

Case 3: $\text{Wealth}_t^{\min} \leq \text{Wealth}_t^{\text{TMG}} \leq \text{Wealth}_t^{\text{available}}$ and $\text{Wealth}_t^{\text{desired}} \leq \text{Wealth}_t^{\text{available}} + \text{PPE}_{t-1}$:

The insurer is able to serve the TMG and the desired rate with the available wealth. Moreover, we can realize in this case a portion of unrealized loss (see Figure

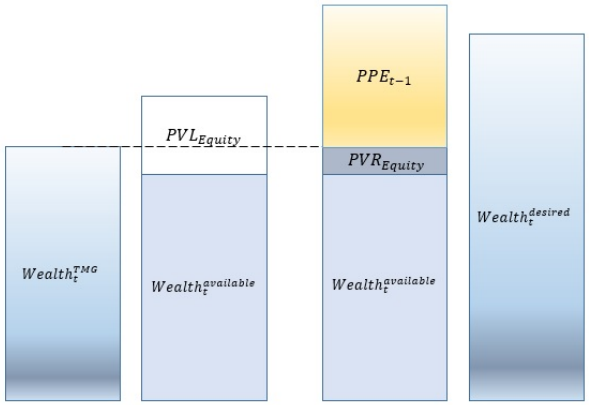


Figure B.11: Case 2: $\text{Wealth}_t^{\text{available}} \leq \text{Wealth}_t^{\text{TMG}} \leq \text{Wealth}_t^{\text{max}}$ and $\text{Wealth}_t^{\text{desired}} \leq \text{Wealth}_t^{\text{max}} + \text{PPE}_{t-1}$

B.12). In the same way as previously, the realized loss is determined by comparing different levels of wealth

$$\begin{aligned}\text{MVR} = \min\{&\text{Wealth}_t^{\text{available}} - \text{Wealth}_t^{\text{TMG}};\\ &\text{Wealth}_t^{\text{available}} + \text{PPE}_{t-1} - \text{Wealth}_t^{\text{desired}};\\ &\text{MVL}_{\text{realizable}}\}.\end{aligned}$$

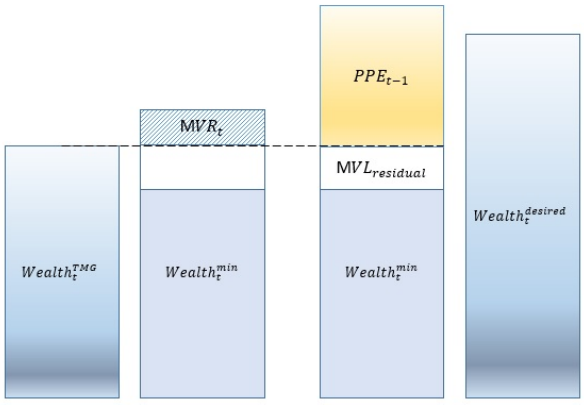


Figure B.12: Case 3: $\text{Wealth}_t^{\text{min}} \leq \text{Wealth}_t^{\text{TMG}} \leq \text{Wealth}_t^{\text{available}}$ and $\text{Wealth}_t^{\text{desired}} \leq \text{Wealth}_t^{\text{available}} + \text{PPE}_{t-1}$

.

Case 4: $\text{Wealth}_t^{\text{TMG}} \leq \text{Wealth}_t^{\text{max}}$ and $\text{Wealth}_t^{\text{max}} + \text{PPE}_{t-1} < \text{Wealth}_t^{\text{desired}}$:

The insurer can not serve the desired rate despite the realization of all its unrealized gains. We will therefore serve a profit-sharing rate corresponding to its

maximum level of wealth (see Figure B.13).

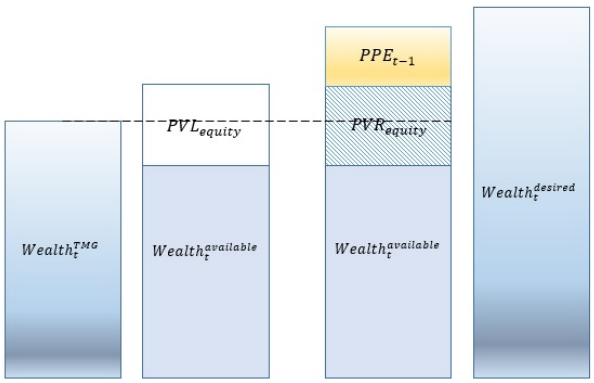


Figure B.13: Case 4: $\text{Wealth}_t^{\text{TMG}} \leq \text{Wealth}_t^{\text{max}}$ and $\text{Wealth}_t^{\text{max}} + \text{PPE}_{t-1} < \text{Wealth}_t^{\text{desired}}$

.

Case 5: $\text{Wealth}_t^{\text{desired}} \leq \text{Wealth}_t^{\text{min}}$:

In the best case, the insurer can not only serve the desired rate but only make a profit (see Figure B.14).

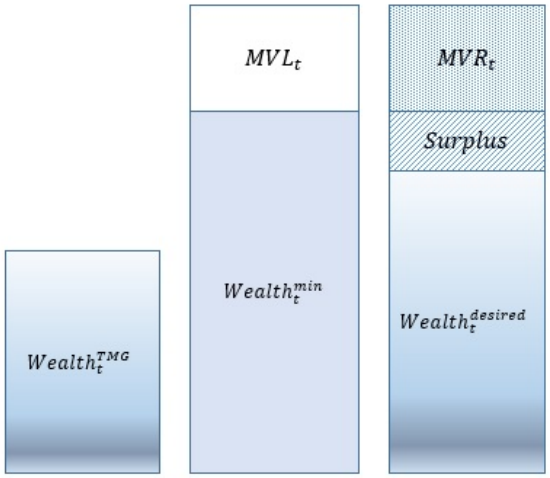


Figure B.14: Case 5: $\text{Wealth}_t^{\text{desired}} \leq \text{Wealth}_t^{\text{min}}$

.

In regulation, insurance companies must redistribute at least 85% of the financial products (if any) to the insureds. It is therefore necessary to verify at the end of

the wealth distribution process that this regulatory condition is well verified. In practice, we will cap the margin to be distributed to shareholders at 15% of all financial income by putting the surplus in the PPE.

B.3.5 End-of-period liabilities modeling

Own funds

In practice, during the projection of the activity of the insurance companies, the profits are never redistributed to the shareholders. Therefore, we assume that the final own funds BOF_t, in our model, becomes the previous own funds BOF_{t-1} to which all the results R_t are added. This assumption then makes it possible to calculate the own funds at the end of each year as follows

$$\mathrm{BOF}_t = \mathrm{BOF}_{t-1} + \mathrm{R}_t.$$

B.4 ALM modeling consistency - Leakage test

In order to verify that the ALM model is consistent and correctly implemented, we verify that there is no value leakage during projections. This test is called the leakage test. It consists of comparing the market value of liabilities with the market value of assets at $t = 0$. Namely, we have to ensure that the following equation holds $\mathrm{BEL}_{t=0} + \mathrm{VIF}_{t=0} = \mathrm{Assets}_{t=0}$, where $\mathrm{VIF}_{t=0}$ stands for the value-of-in-force business at $t = 0$. Recall that the value-of-in-force business (VIF) is a concept used within insurance that essentially refers to the future profits expected to emerge from a particular life insurance portfolio. Mathematically, it is defined as

$$\mathrm{VIF}_{t=0} = \mathbb{E}^{\mathbb{Q}}\left[\sum_{t=1}^{T} DF(0,t)\left(\Phi_t + \mathrm{P}_t - \mathrm{Cl}_t - \Delta\mathrm{PM}_t - \Delta\mathrm{PPE}_t - \Delta\mathrm{RC}_t - \Delta\mathrm{PRE}_t\right)\right]$$

where

- Cl_t the policyholder claims occured over the period of t to $t+1$,
- P_t the periodic primes paid by the policyholders over the period of t to $t+1$,
- Φ_t the financial result between the time t and $t+1$.

To better understand the leakage test, a derivation of the previous equation in case of a *deterministic projection* is given in the following. By definition, we have $\mathrm{BEL}_{t=0} = \sum_{t=1}^{T} P(0,t)\left(\mathrm{Cl}_t - \mathrm{P}_t\right)$ and $\mathrm{VIF}_{t=0} = \sum_{t=1}^{T} P(0,t)\left(\Phi_t + \mathrm{P}_t - \mathrm{Cl}_t - \Delta\mathrm{PM}_t - \Delta\mathrm{PPE}_t - \Delta\mathrm{RC}_t - \Delta\mathrm{PRE}_t\right)$. Therefore,

$$\mathrm{BEL}_{t=0} + \mathrm{VIF}_{t=0} = \sum_{t=1}^{T} P(0,t)\left(\Phi_t - \Delta\mathrm{PM}_t - \Delta\mathrm{PPE}_t - \Delta\mathrm{RC}_t - \Delta\mathrm{PRE}_t\right).$$

For later use, let us denote by

- Cash_t the cash return on capital deposited on saving account,
- Div_t the dividend paid to shareholders,
- Cpn_t the coupon collected on the investment bonds,
- Amortization_t the carrying value of a bond portfolio,
- $\text{PMVR}_t^{bonds}, \text{PMVR}_t^{eq}$ the realized capital gains and losses arising from financial operations on bonds and equities respectively.

The financial result Φ_t includes the dividend Div_t, the cash return Cash_t, the coupon Cpn_t, the carrying value Amortization_t and the realized capital gains and losses PMVR_t, i.e.

$$\Phi_t = \text{Div}_t + \text{Cash}_t + \text{Cpn}_t + \text{Amortization}_t + \text{PMVR}_t^{bonds} + \text{PMVR}_t^{eq}.$$

We denote by $\text{MV}_{t^+}, \text{MV}_{t^-}$ the market values of a financial product at the beginning and at the end of the year t, and by α, β the allocations (in percentage) of equities and bonds. It is not difficult to note that

$$\begin{aligned}
\text{Div}_t &= q\text{MV}^{eq}_{(t-1)^-} \\
\text{Cash}_t &= R(t-1,t)\left[(1-\alpha-\beta)\text{PM}_{t-1} + \text{PPE}_{t-1} + \text{RC}_{t-1} + \text{PRE}_{t-1}\right] \\
\text{PMVR}_t^{eq} &= \alpha\left(\text{PM}_t - \text{PM}_{t-1}\right) + \left(\text{MV}^{eq}_{t^+} - \text{MV}^{eq}_{t^-}\right) \\
&= \alpha\left(\text{PM}_t - \text{PM}_{t-1}\right) + (1 + R(t-1,t) - q)\text{MV}^{eq}_{(t-1)^-} - \text{MV}^{eq}_{t^-} \\
\text{Amortization}_t + \text{PMVR}_t^{bonds} &= \beta\left(\text{PM}_t - \text{PM}_{t-1}\right) + \left(\text{MV}^{bonds}_{t^+} - \text{MV}^{bonds}_{t^-}\right)
\end{aligned}$$

and

$$\begin{aligned}
\text{MV}^{bonds}_{t^-} &= P(t-1,t)\left(\text{MV}^{bonds}_{(t+1)^+} + \text{Cpn}_{t+1}\right) \\
&= \frac{P(0,t+1)}{P(0,t)}\left(\text{MV}^{bonds}_{(t+1)^+} + \text{Cpn}_{t+1}\right)
\end{aligned}$$

Combining all these equations together, we get finally

$$\begin{aligned}
P(0,t)(\Phi_t - \Delta\text{PM}_t - \Delta\text{PPE}_t - \Delta\text{RC}_t - \Delta\text{PRE}_t) \\
&= P(0,t-1)[\text{MV}^{eq}_{(t-1)^-} + \text{MV}^{bonds}_{(t-1)^-} + (1-\alpha-\beta)\text{PM}_{t-1} \\
&+ \text{PRE}_{t-1} + \text{RC}_{t-1} + \text{PPE}_{t-1}] - P(0,t)[\text{MV}^{eq}_{t^-} + \text{MV}^{bonds}_{t^-} \\
&+ (1-\alpha-\beta)\text{PM}_t + \text{PRE}_t + \text{RC}_t + \text{PPE}_t] \qquad \text{(B.11)}
\end{aligned}$$

Since we liquidate all the liabilities at the end of projection, we have to add $P(0,T)[\text{MV}^{eq}_{T^-} + \text{MV}^{bonds}_{T^-} + (1-\alpha-\beta)\text{PM}_T + \text{PRE}_T + \text{RC}_T + \text{PPE}_T]$ on the right-hand side of the equation (B.11) at $t = T$. With a little bit of algebra, it is easy to show that

$$\begin{aligned}
\text{BEL}_{t=0} + \text{VIF}_{t=0} &= \text{MV}_0^{eq} + \text{MV}_0^{bonds} + (1-\alpha-\beta)\text{PM}_0 + \text{PRE}_0 + \text{RC}_0 + \text{PPE}_0 \\
&= \text{Asset}_{t=0}.
\end{aligned}$$

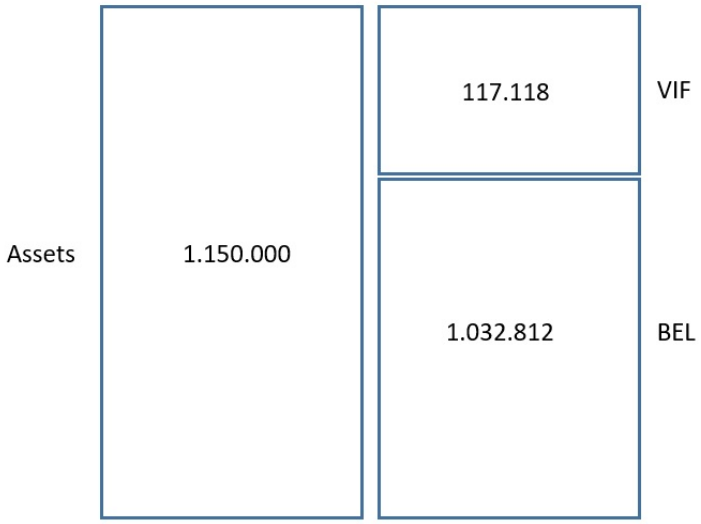


Figure B.15: Comparison of the market value of assets and the market value of liabilities at $t = 0$ given the deterministic scenario.

By freezing all the volatility terms to be zero (deterministic scenario), we find that there is no leakage in our ALM model (see Fig B.15). We also perform the leakage test for the probabilistic simulations. In this case, we will observe the leaks in our simulations mainly explained by the finite numbers of simulations. To this end, we define the leakage ratio as

$$L_\varepsilon = \frac{|\mathrm{BEL}_{t=0} + \mathrm{VIF}_{t=0} - \mathrm{Asset}_{t=0}|}{\mathrm{Asset}_{t=0}}$$

We iterate the simulation with 100 different random seeds. For each simulation, we compute the BEL and the VIF using Monte-Carlo method with N_{MC} number of simulations. Then we compute the mean squared error defined by

$$MSE = \frac{1}{100}\sum_{j=1}^{100}\left(L_\varepsilon^{(j)}\right)^2$$

where the index j stands for the jth random state. Finally, we obtain the following results

N_{MC}	$MSE(\times 10^{-4})$
10^3	83.2
10^4	11.5
10^5	0.98

In the above table we can see that the leaks are significantly reduced as the number of simulations increases.

Appendix C

Demonstration of the θ_t equation

The Hull White model reproduces exactly the zero-coupon rate curve, if

$$\theta_t = \frac{d}{dt}f(0,t) + a.f(0,t) + \frac{\sigma^2}{2a}\left(1 - e^{-2at}\right) \tag{C.1}$$

Proof. We assume that the zero-coupon price at maturity T is written as a function of t and the short rate r_t. So we have $P(t,T) = h(t,r_t)$. Let's apply the formula of Itô to $h(t,r_t)$, in order to find a partial differential equation verified by h:

$$dP(t,T) = \left(\frac{dh}{dt} + \frac{dh}{dr_t}(\theta_t - ar_t) + \frac{1}{2}\frac{d^2h}{dr_t^2}\sigma^2\right)dt + \frac{dh}{dr_t}\sigma dW_t$$

However, under the neutral risk probability, $dP(t,r_t) = r_t P(t,r_t)dt$ since the zero-coupon bond is considered as the risk-free asset. By uniqueness, we obtain the following partial differential equation

$$\frac{dh}{dt} + \frac{dh}{dr_t}(\theta_t - ar_t) + \frac{1}{2}\frac{d^2h}{dr_t^2}\sigma^2 = r_t P(t,r_t) \tag{C.2}$$

In addition, $f(T,r_T) = P(T,T) = 1$ by definition of a zero-coupon bond.

We then seek to find a solution of the equation (C.2) of form: $f(t,r_t) = P(t,T) = A(t,T)e^{-B(t,T)r_t}$. By differentiating P in terms of t and r_t, and by simplifying by $e^{-B(t,T)r_t}$:

$$\begin{cases} \frac{dA(t,T)}{dt} - \theta_t A(t,T)B(t,T) + \frac{1}{2}\sigma^2 A(t,T)B^2(t,T) = 0 \\ 1 + \frac{dB(t,T)}{dt} - aB(t,T) = 0 \end{cases}$$

with $A(T,T) = 1$ and $B(T,T) = 0$.

We then obtain

$$B(t,T) = \frac{1}{a}\left(1 - e^{-a(T-t)}\right)$$

and

$$\ln A(t,T) = -\int_t^T \theta_u B(u,T)du - \frac{\sigma^2}{2a}\left(B(t,T) - (T-t)\right) - \frac{\sigma^2}{4a}B^2(t,T)$$

With the expression of A and B, we have the zero coupon price expression in the Hull & White model, we can infer an expression of today's instantaneous forward rates for all maturities T

$$f(0,T) = \int_0^T \theta_u \frac{dB(u,T)}{dT}du + \frac{dB(0,T)}{dT}r_0 - \frac{\sigma^2}{2a}B(0,T)\left(1 - \frac{dB(0,T)}{dT}\right)$$

Let us differentiate the above equation by T, we have

$$\frac{df(0,T)}{dT} = \theta_T - af(0,T) - \frac{\sigma^2}{2a}\left(1 - e^{-2aT}\right)$$

□

Appendix D

Bayesian P-spline regression and Bayesian asymptotic confidence interval

In this section, a brief description of smoothing and penalized splines (or P-splines) will be presented. Some interesting additional information about smoothing splines and P-splines can be found for example in the work of Reinsch [94], Duchon [31], Green and Silverman [43], Hastie and Tibshirani [49], Eubank [35], Eilers and Marx [32] and Ruppert and Carroll [96].

Consider the regression model $Y_i = m(X_i) + \varepsilon_i$, $i = 1, \ldots, n$ where ε_i are the independent random variables with mean 0 and variance σ^2. We assume that the design points $X_i \in [a, b]$ with $a, b < \infty$. As pointed out in Appendix (D.7), for any regular functions $m(x)$, we can always find a best spline approximation $\tilde{m}(x)$ of $m(x)$ to minimize $\|m - \tilde{m}\|_\infty$. The error in approximating $m(x)$ by $\tilde{m}(x)$ is usually negligible compared to the estimator error, thus in practice we estimate $\tilde{m}(x)$ instead of $m(x)$. In the following, we denote by $B(x) = \{B_1(x), \ldots, B_N(x)\}^T \in \mathbb{R}^N$, $N \leq n$ a spline basis.

D.1 Smoothing Splines

The key idea of this regression method is to approximate the target function $m(x)$ by a natural cubic spline $\tilde{m}(x)$ with knots at the distinct X_i values. The basis functions for natural cubic splines are

$$\{1, x, d_1(x) - d_{n-1}(x), \ldots, d_{n-2}(x) - d_{n-1}(x)\}$$

where $d_k(x) = \frac{(x-X_k)_+^3 - (x-X_n)_+^3}{X_n - X_k}$ for $k = 1, \ldots, n-1$.

A smoothing spline estimator arises as the minimizer of the penalized sum of squares,

$$\sum_{i=1}^{n} (Y_i - \tilde{m}(X_i))^2 + \lambda \int_a^b \left(\tilde{m}^{"}(x)\right)^2 dx \tag{D.1}$$

for $\lambda > 0$.

The integral term in the previous expression is a roughness penalty with the smoothing parameter λ. We write the vector $\tilde{m} = (\tilde{m}(X_1), \ldots, \tilde{m}(X_n))^T \in \mathbb{R}^n$. The theorem 2.1 in Green and Silverman [43] shows that there exists a $n \times n$

dimensional matrix $\mathbf{K}$ of rank $n-2$ such that the penalty term $\int_a^b \left(\tilde{m}^{"}(x)\right)^2 dx$ can be written as $\tilde{m}^T\mathbf{K}\tilde{m}$. The smoothing spline estimate at the design points $\hat{m} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^T \in \mathbb{R}^n$ is thus explicitly given by $\hat{m} = (\mathbf{I} + \lambda\mathbf{K})^{-1} Y$ where $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$. This regression method is however less practical when the number of design points n becomes large since it uses n knots.

D.2 Regression Penalized Splines or P-Splines

The idea of penalized spline smoothing with basic functions can track back to O'Sullivan [87], see also Eilers and Marx [32] and Ruppert and Carroll [96] for additional information. Two common basis used in practical, up to our knowledges, are the B-spline basis (Eilers and Marx [32], De Boor [25]) and the truncated power basis (Ruppert and Carroll [96]). The B-spline basis is preferable for computation because of its better numerical properties. For formulation and theoretical study the truncated power basis is preponderant because of its simplicity.

The penalized spline model specifies that the estimate $\tilde{m}(x)$ is expressed as $B(x)^T\beta$ for some N-dimensional space. Hence, the estimation of $\tilde{m}(x)$ is equivalent to that of β. Let $\mathbf{D}$ be a fixed, symmetric, positive semidefinite $N \times N$ matrix, which is equivalent to the matrix $\mathbf{K}$ in the smoothing spline model. The penalized spline estimator $\hat{\beta}$ arises as the minimizer of

$$\sum_{i=1}^{n} (Y_i - B(X_i)\beta)^2 + \lambda\beta^T\mathbf{D}\beta \tag{D.2}$$

Define $\mathbf{B}$ the $n \times N$ matrix whose i-th row equals $B(X_i)^T$. The penalized spline estimator is thus written as $\hat{\beta} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}^T Y$. From this, we obtain:

$$E\left(\hat{\beta}\right) = \left(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}\right)^{-1}\mathbf{B}^T\mathbf{B}\beta \tag{D.3}$$

and the covariance of the estimate

$$\mathbf{V}_{\hat{\beta}} = \sigma^2\left(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}\right)^{-1}\mathbf{B}^T\mathbf{B}\left(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}\right)^{-1} \tag{D.4}$$

It follows that $\hat{\beta} \sim \mathcal{N}\left(E\left(\hat{\beta}\right), \mathbf{V}_{\hat{\beta}}\right)$. It is clear that $E\left(\hat{\beta}\right) \neq \beta$ except for $\beta = 0$ which leads to the difficulty in using this result for calculating confidence intervals.

In this paper, we choose the truncated p-polynomial basis as mentioned previously, i.e. $B(x) = \left(1, x, x^2, \ldots, x^p, (x-\kappa_1)_+^p, \ \ldots, (x-\kappa_K)_+^p\right)^T$ for $p \geq 1$ and $\mathbf{D}$ is the diagonal matrix $diag(\mathbf{0}_{p+1}, \mathbf{1}_K)$, indicating that only the spline coefficients are penalized. One can easily verify that $\int_a^b \left[\tilde{m}^{(p+1)}(x)\right]^2 dx = (p!)^2 \sum_{j=1}^{K} \beta_{j+p}^2$ by using the fact that the derivative of an indicator function is a Dirac delta function. Therefore, the equation (D.2) can be rewritten as

$$\sum_{i=1}^{n} (Y_i - \tilde{m}(x))^2 + \lambda' \int_a^b \left[\tilde{m}^{(p+1)}(x)\right]^2 dx \tag{D.5}$$

which is a generalization of equation (D.1).

D.3 Bayesian Analysis for Penalized Splines Regression

The Bayesian analysis for penalized spline regression is an alternative approach to calculate the confidence intervals. The initial works for this approach is mainly due to Wahba [115] and Silverman [99]. The main idea behind this method is to partition the coefficient vector β into the coefficients of the monomial basis functions of the truncated power functions by letting $\beta = (\beta_1^T, \beta_2^2)^T$ where $\beta_1 \in \mathbb{R}^{1+p}$ has an improper uniform prior density and $\beta_2 \in \mathbb{R}^K$ has a proper prior equal to $(\lambda/\sigma^2)^{K/2} \exp\left(-(\lambda/2\sigma^2)\beta_2^T \beta_2\right)$.

Following the idea of Wood (section 4.8.1 in [119]), we obtain the Bayesian posterior covariance matrix for the parameter β:

$$\mathbf{V}_\beta = \left(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D}\right)^{-1} \sigma^2 \tag{D.6}$$

and its corresponding posterior distribution:

$$\beta_{|Y} \sim \mathcal{N}\left(\hat{\beta}, \mathbf{V}_\beta\right) \tag{D.7}$$

where $\hat{\beta} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}^T Y$. The penalized least squares estimator is the mean of the posterior distribution of β. This posterior on β induces a posterior on $\tilde{m}(\cdot)$ and then the posterior distribution of $\tilde{m} = (\tilde{m}(X_1), \ldots, \tilde{m}(X_n))^T$, i.e. $\tilde{m}_{|Y} \sim \mathcal{N}(\mathbf{A}Y, \sigma^2\mathbf{A})$ with $\mathbf{A} = \mathbf{B}(\mathbf{B}^T\mathbf{B} + \lambda\mathbf{D})^{-1}\mathbf{B}^T$.

D.4 Bayesian Asymptotic Confidence Interval

Nychka [84] showed that if $m(x)$ is estimated using a cubic smoothing spline for which the smoothing parameter is sufficiently reliably estimated that the bias in the estimates is a modest fraction of the mean squared error for $m(x)$, then the average coverage probability (ACP)

$$\text{ACP} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(m(X_i) \in BI_\alpha(X_i)\right)$$

is very close to the nominal level $1-\alpha$, where $BI_\alpha(x)$ indicates the $(1-\alpha)100\%$ Bayesian interval for $m(x)$ and α the significance level. This is due to the fact that the average posterior variance for the spline is similar to a consistent estimate of the average squared error and that the average squared bias is relatively small with respect to the total average squared error.

Marra and Wood [80] modified Nychka's [84] approach to obtain the confidence interval for Generalized Additive Model (GAM), which is also applicable in the Bayesian penalized spline models. Here we will sketch the main steps to obtain confidence interval of variable width.

Given some constants C_i, which will be defined later, the primary purpose is to find a constant A, such that

$$\text{ACP} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{P}\left(|\hat{m}(X_i) - \tilde{m}(X_i)| \leq z_{\alpha/2} A/\sqrt{C_i}\right) = 1-\alpha \tag{D.8}$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical point from a standard normal distribution.

Letting $\hat{m} = (\hat{m}(X_1), \ldots, \hat{m}(X_n))^T = \mathbf{B}\hat{\beta}$ and $\tilde{m} = (\tilde{m}(X_1), \ldots, \tilde{m}(X_n))^T = \mathbf{B}\beta$. We write the covariance of $\hat{m}$ as $\mathbf{V}_{\hat{m}} = \mathbf{B}\mathbf{V}_{\hat{\beta}}\mathbf{B}^T$ and the same as that of $\tilde{m}$, i.e. $\mathbf{V}_{\tilde{m}} = \mathbf{B}\mathbf{V}_{\beta}\mathbf{B}^T$. Define $b \equiv (b(X_1), \ldots, b(X_n))^T = E(\hat{m}) - \tilde{m}$ and $v \equiv (v(X_1), \ldots, v(X_n))^T = \hat{m} - E(\hat{m})$. We have $v \sim \mathcal{N}(0, \mathbf{V}_{\hat{m}})$ following from multivariate normality of $\hat{m}$. Equivalently, the equation (D.8) can be written as:

$$\begin{aligned} \text{ACP} &= \mathbb{P}\left(|b(X_I) + v(X_I)| \leq z_{\alpha/2} A/\sqrt{C_I}\right) \\ &= \mathbb{P}\left(|\mathfrak{B} + \mathfrak{V}| \leq z_{\alpha/2} A\right) = 1 - \alpha \end{aligned} \tag{D.9}$$

where I is a random variable uniformly distributed on $\{1, 2, \ldots, n\}$, $\mathfrak{B}$ and $\mathfrak{V}$ are respectively the random scaled bias and random scaled variance defined as follows:

$$\mathfrak{B} = \sqrt{C_I} b(X_I) \text{ and } \mathfrak{V} = \sqrt{C_I} v(X_I).$$

This means that we need to know the distribution of $\mathfrak{B} + \mathfrak{V}$ in order to find the constant A.

Clearly, by definition, we have $\mathbb{E}(\mathfrak{B}) = c^T\mathbf{B}\left(\mathbf{F}\beta - \beta\right)$, $\mathbb{E}(\mathfrak{V}) = 0$ and $\text{var}(\mathfrak{V}) = \text{tr}(\mathbf{C}\mathbf{V}_{\hat{m}})/n$ where $c = \left(\sqrt{C_1}, \ldots, \sqrt{C_n}\right)^T$, $\mathbf{F} = (\mathbf{B}^T\mathbf{B} + \lambda\mathbf{B})^{-1}\mathbf{B}^T\mathbf{B}$ and $\mathbf{C}$ is the diagonal matrix $\text{diag}\,(C_1, \ldots, C_n)$. Since $v \sim \mathcal{N}(0, \mathbf{V}_{\hat{m}})$, $\mathfrak{V}$ is a mixture of normals. However, if we choose $C_i^{-1} = [\mathbf{V}_{\hat{m}}]_{ii}$, the random scaled variance $\mathfrak{V}$ then has a normal distribution. Since its distribution no longer depends on i, it implies the independence of $\mathfrak{B}$ and $\mathfrak{V}$.

We call $\mathfrak{M}$ the scaled average mean squared error which is given by:

$$\mathfrak{M} = \frac{1}{n}\sum_{i=1}^{n} C_i \left(\hat{m}(X_i) - \tilde{m}(X_i)\right)^2. \tag{D.10}$$

The mean squared error $E\left(\mathfrak{M}\right)$ can be then determined as follows:

$$E(\mathfrak{M}) = \frac{1}{n}\text{Tr}\left(\mathbf{B}^T\mathbf{C}^2\mathbf{B}\mathbf{V}_{\hat{\beta}}\right) + \frac{1}{n}\|\mathbf{C}\mathbf{B}\left(\mathbf{F} - \mathbf{I}\right)\beta\|^2. \tag{D.11}$$

By construction, we have $\mathbb{E}(\mathfrak{B} + \mathfrak{V}) = \mathbb{E}(\mathfrak{B})$ and $\text{var}(\mathfrak{B} + \mathfrak{V}) = \mathbb{E}(\mathfrak{M}) - \mathbb{E}(\mathfrak{B})^2$. Nychka's [84] simulation results showed that $\mathfrak{B} + \mathfrak{V}$ will be approximately normally distributed, provided that $\mathfrak{B}$ is small relative to $\mathfrak{V}$, i.e.

$$\mathfrak{B} + \mathfrak{V} \sim \mathcal{N}\left(\mathbb{E}(\mathfrak{B}), \mathbb{E}(\mathfrak{M}) - \mathbb{E}(\mathfrak{B})^2\right)$$

The expectations $\mathbb{E}(\mathfrak{B})$ and $E(\mathfrak{M})$ can be estimated by substituting β by $\hat{\beta}$ which yields

$$\widehat{E(\mathfrak{B})} = c^T\mathbf{B}(\mathbf{F}\hat{\beta} - \hat{\beta})/n$$

and

$$\begin{aligned} \widehat{E\left(\mathfrak{M}\right)} &= \frac{1}{n}\text{Tr}\left(\mathbf{B}^T\mathbf{C}^2\mathbf{B}\mathbf{V}_{\hat{\beta}}\right) + \frac{1}{n}\|\mathbf{C}\mathbf{B}\left(\mathbf{F} - \mathbf{I}\right)\hat{\beta}\|^2 \\ &= 1 + \frac{1}{n}\|\mathbf{C}\mathbf{B}\left(\mathbf{F} - \mathbf{I}\right)\hat{\beta}\|^2 \end{aligned} \tag{D.12}$$

Therefore, we have the approximate result

$$\mathfrak{B}+\mathfrak{V}\sim\mathcal{N}\left(\widehat{E(\mathfrak{B})},\widehat{E(\mathfrak{M})}-\widehat{E(\mathfrak{B})}^2\right)$$

under the assumption that $\mathfrak{B}$ is small relative to $\mathfrak{V}$, i.e. $\overline{b^2}\ll\overline{v^2}$.

As the definition (D.9), it follows that $A=\sqrt{\widehat{E(\mathfrak{M})}-\widehat{E(\mathfrak{B})}^2}$ and then we obtain

$$\hat{m}(X_i)-\widehat{E(\mathfrak{B})}\sqrt{[\mathbf{V}_{\hat{m}}]_{ii}}\pm z_{\alpha/2}\sqrt{\left(\widehat{E(\mathfrak{M})}-\widehat{E(\mathfrak{B})}^2\right)[\mathbf{V}_{\hat{m}}]_{ii}} \tag{D.13}$$

as the definition of $1-\alpha$ Bayesian asymptotic confidence intervals at the point X_i.

Finally, we use the estimator $\hat{\sigma}^2=\frac{\|Y-\mathbf{B}\hat{\beta}\|^2}{n-\mathrm{Tr}(\mathbf{F})}$ to estimate σ^2.

D.5 Additive model and Asymptotic confidence interval for each functional components

In this section, the additive model will be briefly presented. In this model the relation between the response Y_i, $i\in\{1,\ldots,n\}$ and the d-explanatory variables $X_{1i},X_{2i},\ldots,X_{di}$ is expressed through arbitrary univariate functions f_j as follows:

$$Y_i=\beta_0+\sum_{j=1}^{d}f_j\left(X_{ji}\right)+\varepsilon_i \tag{D.14}$$

where the errors ε_i are independent and identically distributed with mean zero and the variance σ^2. Several estimation strategies have been developed to fit such a model, e.g. backfitting algorithm (Hastie and Tibshirani [49]) as well as its asymptotic statistical properties (Opsomer and Ruppert [85], Opsomer [86] and Wand [116]), and a marginal integration approach (Linton and Nielsen [77]). Due to the computational expediency, the penalized splines regression to ordinary additive models (Hastie and Tibshirani [49]) have been widely used in practice. As consequent, we will apply this approach to estimate the excess loss function.

Like the univariate case, each of the functional components f_j is modeled as $\mathbf{B}^T\beta_j$ a degree p_j penalized spline estimator with smoothing parameters λ_j and $\mathbf{B}_j$ a $n\times(p_j+K_j)$ matrix whose i-th row is $(X_{ji},X_{ji}^2,\ldots,X_{ji}^{p_j},(X_{ji}-\kappa_{j1})_+^{p_j},\ldots,(X_{ji}-\kappa_{jK_j})_+^{p_j})$. However, for the identifiability reasons, we replace $f_j(\cdot)$ by $f_j(\cdot)-1/n\sum_{i=1}^n f_j(X_{ji})$ which leads to the condition on $f_j(\cdot)$ of the form $\sum_{i=1}^n f_j(X_{ji})=0$. Therefore, the estimate of β_0 is $\hat{\beta}_0=\bar{Y}\equiv\frac{1}{n}\sum_{i=1}^n Y_i$ which is independent of X_{ji}'s and the matrix $\mathbf{B}_j$ is adjusted to $\mathbf{B}_j^*=\frac{1}{n}(\mathbf{I}-\mathbf{1}.\mathbf{1}^T)\mathbf{B}_j$ where $\mathbf{I}$ is the identity matrix and $\mathbf{1}$ is a $n\times 1$ column of ones. This adjustment is called the "centering effect". Letting $Y^*=(Y_1-\hat{\beta}_0,\ldots,Y_n-\hat{\beta}_0)^T$, $\ddot{\mathbf{B}}=[\mathbf{B}_1^*,\ldots,\mathbf{B}_d^*]$ and $\mathbf{D}_\lambda=\mathrm{blockdiag}_{1\le j\le d}(\lambda_j\mathbf{D}_j)$ with $\lambda_j>0$. We then have the estimate of $\beta^T=(\beta_1^T,\ldots,\beta_d^T)$ is given by

$$\hat{\beta}=\left(\ddot{\mathbf{B}}^T\ddot{\mathbf{B}}+\mathbf{D}_\lambda\right)^{-1}\ddot{\mathbf{B}}Y^*\in\mathbb{R}^{\sum_{j=1}^d(p_j+K_j+1)} \tag{D.15}$$

From this we derive the estimate of β_j, $\hat{\beta}_j = \mathbf{P}_j\hat{\beta}$ where $\mathbf{P}_j = [0,\ldots,\mathbf{1}_{(p_j+K_j+1)},\ldots,0] \in \mathbb{R}^{(p_j+K_j+1)\times\sum_k(p_k+K_k+1)}$.

Regarding the Bayesian asymptotic confidence interval for each of the functional components, the derivation is analogue as for the univariate case. Namely, assume that β_0 has an improper uniform prior distribution and the prior for β_j is given by $(\lambda_j/\sigma^2)^{K_j/2}\exp\left(-(\lambda_j/2\sigma^2)\beta_j^T\beta_j\right)$ for $j=1,\ldots,d$. By using Bayes rule, it has been shown ([119]) that

$$\beta_{|Y} \sim \mathcal{N}\left(\hat{\beta}, \mathbf{V}_\beta\right)$$

where $\mathbf{V}_\beta = (\ddot{\mathbf{B}}^T\ddot{\mathbf{B}} + \mathbf{D}_\lambda)^{-1}\sigma^2$. Thanks to the equation (D.15), it is then routine to show that $\mathbf{V}_{\hat{\beta}} = (\ddot{\mathbf{B}}^T\ddot{\mathbf{B}} + \mathbf{D}_\lambda)^{-1}\ddot{\mathbf{B}}^T\ddot{\mathbf{B}}(\ddot{\mathbf{B}}^T\ddot{\mathbf{B}} + \mathbf{D}_\lambda)^{-1}\sigma^2$. It follows immediately that the variance of $\hat{f}_j$ is $\mathbf{V}_{\hat{f}_j} = (\mathbf{B}_j^*\mathbf{P}_j)\mathbf{V}_{\hat{\beta}}(\mathbf{B}_j^*\mathbf{P}_j)^T$. The $1-\alpha$ Bayesian asymptotic confidence interval computation for $\hat{f}_j(X_{ji})$ is similar to that can be found in (D.4). Routine manipulation then results in

$$\hat{f}_j\left(X_{ji}\right) - \widehat{E(\mathfrak{B}_j)}\sqrt{\left[\mathbf{V}_{\hat{f}_j}\right]_{ii}} \pm z_{\alpha/2}\sqrt{\left(\widehat{E(\mathfrak{M}_j)} - \widehat{E(\mathfrak{B}_j)}^2\right)\left[\mathbf{V}_{\hat{f}_j}\right]_{ii}} \tag{D.16}$$

as the definition of $1-\alpha$ Bayesian asymptotic confidence interval for $\hat{f}_j(X_{ji})$. Here, we denoted $\widehat{E(\mathfrak{B}_j)} = \frac{1}{n}c^T\mathbf{B}_j^*\mathbf{P}_j\left(\ddot{\mathbf{F}}\hat{\beta} - \hat{\beta}\right)$ with $c = \left([\mathbf{V}_{\hat{f}_j}]_{11}^{-1/2},\ldots,[\mathbf{V}_{\hat{f}_j}]_{nn}^{-1/2}\right)$ and $\ddot{\mathbf{F}} = (\ddot{\mathbf{B}}^T\ddot{\mathbf{B}} + \mathbf{D}_\lambda)^{-1}\ddot{\mathbf{B}}^T\ddot{\mathbf{B}}$, $\widehat{E(\mathfrak{M}_j)} = 1 + \frac{1}{n}\|\mathbf{C}_j\mathbf{B}_j^*\mathbf{P}_j(\ddot{\mathbf{F}}\hat{\beta} - \hat{\beta})\|$ with $\mathbf{C}_j = \mathrm{diag}([\mathbf{V}_{\hat{f}_j}]_{11}^{-1},\ldots,[\mathbf{V}_{\hat{f}_j}]_{nn}^{-1})$. Readers can refer to [80] for more details.

D.6 Upper bound of the probabilities of deviation

1. Let us denote $S_j = \{|\hat{\phi}_j\left(x_j^{(\nu)}\right) - \phi_j\left(x_j^{(\nu)}\right)| > \Delta_{j,\alpha}^{(\nu)}\}$ and $\tilde{S}_j = \{|\hat{h}_J\left(x_J^{(\nu)}\right) - h_J\left(x_J^{(\nu)}\right)| > \tilde{\Delta}_{J,\alpha}^{(\nu)}\}$. Since $\{S_j\}$ and $\{\tilde{S}_J\}$ are mutually independent, it implies that (by De Morgan's law)

$$\begin{aligned}\mathbb{P}\left(\left(\bigcup_j S_j\right)\bigcup\left(\bigcup_J \tilde{S}_J\right)\right) &= 1 - \mathbb{P}\left(\left(\bigcap_j S_j^c\right)\bigcap\left(\bigcap_J \tilde{S}_J^c\right)\right)\\ &= 1 - \left(\prod_j \mathbb{P}(S_j^c)\right)\times\left(\prod_J \mathbb{P}(\tilde{S}_j^c)\right)\end{aligned}$$

which asymptotically tends to $1-(1-\alpha)^{d(d+3)/2}$ as $\Gamma\to\infty$. With this the equation (3.28) is then a straightforward consequence of the relation

$$\mathbb{P}\left(|\widehat{\phi}(x^{(\nu)}) - \phi(x^{(\nu)})| > \sum_{j=1}^d \Delta_{j,\alpha}^{(\nu)} + \sum_J \tilde{\Delta}_{J,\alpha}^{(\nu)}\right) \leq \mathbb{P}\left(\left(\bigcup_j S_j\right)\bigcup\left(\bigcup_J \tilde{S}_J\right)\right)$$

thanks to the equation (3.27).

2. For a given value of Γ, we call $r^* \equiv r(\Gamma)$ the positive constant mentioned in Assumption (3). Furthermore, we assume that $B(x^*, r^*) \subset \Omega$ for large enough Γ.

For notional simplicity, let us denote $\delta\phi(x, y) = |\hat{\phi}(y) - \phi(x)|$. We have

$$\begin{aligned}
&\mathbb{P}\left(\delta\phi(x^*, x^*_{(\Gamma)}) > \Delta(\alpha, \Gamma) + Lr^*\right) \\
&= \mathbb{P}\left(\delta\phi(x^*, x^*_{(\Gamma)}) > \Delta(\alpha, \Gamma) + Lr^* \mid x^*_{(\Gamma)} \in B(x^*, r^*)\right) \mathbb{P}\left(x^*_{(\Gamma)} \in B(x^*, r^*)\right) \\
&+ \mathbb{P}\left(|\hat{\phi}(x^*_{(\Gamma)}) - \phi(x^*)| > \Delta(\alpha, \Gamma) + Lr^* \mid x^*_{(\Gamma)} \notin B(x^*, r^*)\right) \mathbb{P}\left(x^*_{(\Gamma)} \notin B(x^*, r^*)\right) \\
&\leq \mathbb{P}\left(\delta\phi(x^*_{(\Gamma)}, x^*_{(\Gamma)}) > \Delta(\alpha, L) \mid x^*_{(\Gamma)} \in B(x^*, r^*)\right) + \xi(r^*, d)\Gamma^{-\gamma(r^*, d)}
\end{aligned}$$

since $|\hat{\phi}(x^*_{(\Gamma)}) - \phi(x^*)| \leq |\hat{\phi}(x^*_{(\Gamma)}) - \phi(x^*_{(\Gamma)})| + |\phi(x^*_{(\Gamma)}) - \phi(x^*)|$ and

$$\mathbb{P}\left(|\phi(x^*_{(\Gamma)}) - \phi(x^*)| > Lr^* \mid x^*_{(\Gamma)} \in B(x^*, r^*)\right) = 0$$

thanks to Assumption (1).

Finally, by applying Assumption (3), we obtain

$$\mathbb{P}\left(\delta\phi(x^*, x^*_{(\Gamma)}) > \Delta(\alpha, \Gamma) + Lr^*\right) \leq \left[1 - (1 - \alpha)^{d(d+3)/2}\right] + \xi(r^*, d)\Gamma^{-\gamma(r^*, d)}.$$

D.7 Best approximation by splines

Let us first introduce two functional spaces.

Definition 3 (Polynomial Spline Space $\Phi_s^{a,b}$)**.** *Letting κ_l for $l \in \{1, \ldots, K\}$ be K-interior knots satisfying the condition $a = \kappa 0 \leq \kappa_1 \leq \cdots \leq \kappa_K \leq \kappa_{K+1} = b$. We define $\Phi_s^{a,b}$ the space of functions whose element is a polynomial of at most degree p on each of the intervals $[\kappa_l, \kappa_{l+1})$ for $l = 0, 1, \ldots, K$ and is $p - 1$ continuously differentiable on $[a, b]$ if $p \geq 1$.*

Definition 4 (Empirically Centered Polynomial Spline Space $\overline{\Phi}_s^{a,b}$)**.** *Given the design points $(x_1, \ldots, x_n) \in [a, b]^n$, a polynomial spline space is centered if for every $g \in \Phi_s^{a,b}$ the following identity holds:*

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) = 0.$$

We denote by $\overline{\Phi}_s^{a,b}$ the Empirically centered polynomial spline space.

According to de Boor (p.149 in [25]), for every $\phi(x) \in C^{p+1}([a, b])$, there exists a constant $c > 0$ and a spline function $\phi^*(x) \in \Phi_s^{a,b}$, such that $\|\phi - \phi^*\|_\infty \leq c\|\phi^{(p+1)}\|_\infty \delta^{p+1}$ with $\delta = \max_{1 \leq l \leq K}(\kappa_{l+1} - \kappa_l)$.

Given the design points $(x_1, \ldots, x_n) \in [a, b]^n$, we assume furthermore that

$$\frac{1}{n}\sum_{i=1}^{n} \phi(x_i) = 0$$

.

By defining $\phi^{**}(x) = \phi^*(x) - \frac{1}{n}\sum_{i=1}^{n} \phi^*(x_i) \in \overline{\Phi}_s^{a,b}$, it is straightforward to show that there exists a positive constant c' such that $\|\phi - \phi^{**}\|_\infty \leq c'\|\phi^{(p+1)}\|_\infty \delta^{p+1}$.

D.8 Asymptotic distribution of empirical quantiles

Theorem 5. *Let $X_1, \ldots, X_n$ be n-real valued observations with unknown distribution function F, p be a real number defined in the interval $[0, 1]$, $x_p = x_p(F)$ be the pth-percentile of F and $F_n(x)$ be the empirical distribution function.*

If we suppose that F is continuous and differentiable at x_p of derivation $f(x_p)$, then

$$\sqrt{n}\,(x_p(n) - x_p) \to \mathcal{N}\left(0, \frac{p(1-p)}{f^2(x_p)}\right) \tag{D.17}$$

as $n \to \infty$.

This is a well-known result for the asymptotic convergence of empirical quantiles. We will thus omit the proof here. For any further and detailed information, the interested reader can refer to [113] (see Section 3.9.21)

Bibliography

[1] ACPR, *Préparation à Solvabilité II*. Enseignements des annexes techniques vie remises en 2013. (Cited on page 137.)

[2] Aerts, M., Claeskens, G. and Wand, M. P., *Some theory for penalized spline generalized additive models* J. Stat. Plan. Inference, 103, 455−470, 2002. (Cited on pages 22 and 55.)

[3] Anderson, L. B. G. and Piterbarg, V. V., *Interest Rate Modeling*. Atlantic Financial Press, Business & Economics, 2010. (Cited on page 109.)

[4] Arthur Greenwood, J., Maciunas Landwehr, J., Matalas, N. C. and Wallis, J. R., *Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form*. Water Resources Research, 15, 1979. (Cited on page 77.)

[5] Arvanitis A., Gregory J., and Laurent J. P., *Building Models for Credit Spreads*. The Journal of Derivatives, 6 (3) 27−43, 1999. (Cited on page 120.)

[6] Balkema, A. A. and de Haan, L., *Residual Life Time at Great Age*. Annals of probability, 2(5): 792−804, 1974. (Cited on page 74.)

[7] Bauer, D., Bergmann, D. and Reuss, A., *Solvency II and Nested Simulations-A Least Square Monte Carlo Approach*. Proceedings of the 2010 ICA congress, 2009. (Cited on page 11.)

[8] Bauer, D., Reuss, A. and Singer, D., *On the Calculation of the Solvency Capital Requirement based on Nested Simulations*. ASTIN Bulletin, 42, 453−499, 2012. (Cited on page 4.)

[9] Beirlant, J. and Goegebeur, Y., *Local polynomial maximum likelihood estimation for Pareto-type distributions*. Journal of Multivariate Analysis, 89, 97−119, 2004. (Cited on pages 17, 27, 79, 80, 81, 82, 90 and 91.)

[10] Beirlant, J., Goegebeur, Y., Teugels, J. and Segers, J., *Statistics of Extremes: Theory and Applications*. John Wiley & Sons, Inc. Published 2004. (Cited on page 75.)

[11] Belloni, A. and Chernozhukov, V., *Least squares after model selection in high dimensional sparse model*. Bernoulli, 19(2):521−547, 2013. (Cited on page 89.)

[12] Bellman, R. E., *Adaptive Control Processes*. Princeton University Press, 1961. (Cited on page 46.)

[13] Bengio, Y. and Grandvalet, Y., *No unbiased estimator of the variance of k-fold cross-validation*. J Mach Learn Res, 5, 1089−1105, 2004. (Cited on page 62.)

[14] Bertsekas, D., *Nonlinear programming*. Athena Scientific, 1999. (Cited on pages 85 and 98.)

[15] Beutner, E., Pelsser, A. and Schweizer, J., (1993) *Theory and validation of replicating portfolios in insurance risk management*. Available at `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2557368` Accessed 20 June 2018. (Cited on page 13.)

[16] Bingham, N.H, Goldie, C.M. and Teugles, J.L., *Regular Variation*. Cambridge University Press, 1987. (Cited on page 79.)

[17] Black, F. and Scholes, M.. *The Pricing of Options and Corporate Liabilities*. Journal of Political Economy. 81(3): 637˘654, 1973. (Cited on page 112.)

[18] de Boor, C., *A Practical Guide to Splines, Revised Edition*. volume 27, Applied Mathematical Sciences, Springer, New York, 2001. (Cited on pages 45 and 99.)

[19] Brigo, D. and Mercurio, F., *Interest Rate Model: Theory and Practice*. Springer Science & Business Media, 2013. (Cited on page 109.)

[20] Buja, A., Hastie, T. J. and Tibshirani, R. J., *Linear smoothers and additive models*. Annals of Statistics, 17: 453−510, 1989. (Cited on page 46.)

[21] Burr, I. W., *Cumulative frequency functions*. Ann of Math Statist, 13, 215−232, 1954. (Cited on page 89.)

[22] Chavez-Demoulin, V., Embrechts, P. and Sardy, S., *An extreme value approach for modeling operational risk losses depending on covariates*. J Risk Insur, 2014. (Cited on pages 17 and 80.)

[23] Chernobai, A., Jorion, P. and Yu, F., *The determinants of operational risk in u.s. financial institutions*. Journal of Financial and Quantitative Analysis, 46(8):1683−1725, 2011. (Cited on pages 17 and 79.)

[24] Christopeit, N., *Estimating parameters of an extreme value distribution by the method of moments*. Journal of Statistical Planning and Inference, 41(2):173−186, 1994. (Cited on page 77.)

[25] de Boor, C., *A Practical Guide to Splines*, revised ed. Springer, New York, 2001. (Cited on pages 154 and 159.)

[26] de Haan, L. and Rotzen, H., *On the estimation of high quantiles*. J Stat Plan Inference, 35(1), 1−13, 1993. (Cited on page 80.)

[27] de Haan, L. and Ferreira, A. F., *Extreme Value Theory - An Introduction*. Springer Series in Operations Research and Financial Engineering, 2006. (Cited on pages 72 and 81.)

[28] Dekkers, A. L. M., Einmahl, J. H. J. and de Hann, J., *A moment estimator for the index of an extreme-value distribution*. Annals of Statistics, 17:1833−1855, 1989. (Cited on page 77.)

[29] Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P. and Meester, L.E., *A Modern Introduction to Probability and Statistics: Understanding Why and How*. Springer Texts in Statistics, 2005. (Cited on page 96.)

[30] Devineau, L. and Loisel, S., *Construction of an acceleration algorithm of the Nested Simulations method for the calculation of the Solvency II economic capital*. Bulletin Français d'Actuariat, 2009. (Cited on page 14.)

[31] Duchon, J., *Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Construction Theory of Functions of Several Variables*. Springer, Berlin, 1977. (Cited on page 153.)

[32] Eilers, P. H. C. and Marx, B. D., *Flexing smoothing with B-splines and Penalties*. Stat Sci, 11, 89-102, 1996. (Cited on pages 153 and 154.)

[33] EIOPA-14/209, Technical Specification for the Preparatory Phase (Part I), 30 April 2014, p.120. (Cited on page 36.)

[34] Embrechts, P., Klüpplberg, C. and Mikosch, T., *Modelling Extremal Events*. Berlin: Springer, 1997. (Cited on pages 75 and 77.)

[35] Eubank, R. L., *Nonparametric Regression and Spline Smoothing*. (2^{nd} ed.) Marcel Dekker, New York, 1999. (Cited on page 153.)

[36] Filipovic, D., *Term-Structure Models: A Graduate Course*. Springer Finance Textbooks, 2009. (Cited on pages 59 and 60.)

[37] Fisher, R. A. and Tippett, L. H. C., *Limiting forms of the frequency distribution of the largest or smallest member of a sample*. Mathematical Proceedings of the Cambridge Philosophical Society, 24(2): 180−190, 1928. (Cited on page 73.)

[38] Friedman, J. H. and Stuetzle, W., *Projection pursuit regression*. Journal of the American Statistical Association, 76:817−823, 1981. (Cited on page 46.)

[39] Friedman, J., Hastie, T. and Tishirani, R., (2010) *A note on the Group Lasso and a Sparse Group Lasso*. arXiv:1001.0736.22. (Cited on page 87.)

[40] Giraud, C., *Introduction to High-Dimensional Statistics*. Chapman and Hall, CRC Monographs on Statistics & Applied Probability, 2014. (Cited on page 62.)

[41] Girsanov, I. V., *On transforming a certain class of stochastic processes by absolutely continuous substitution of measures*. Theory of Probability and its Applications, 5 (3): 285−310, 1960. (Cited on page 114.)

[42] Gnedenko, B., *Sur la distribution limite du terme maximum d'une serie aleatoire*. Ann Stat, 44(3), 423−453, 1943. (Cited on pages 79 and 80.)

[43] Green, P. J. and Silverman, B. W., *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall, London, 1994. (Cited on page 153.)

[44] Gu, C., *Multivariate spline regression*. In Schimek, M. G., editor: *Smoothing and Regression: Approaches, Computation and Applications*. Wiley Series in Probability and Mathematical Statistics, 229−356. John Wiley & sons, 2000. (Cited on page 45.)

[45] Gu, C., *Smoothing spline ANOVA Models*. Springer Series in Statistics, Springer, New York, 2002. (Cited on page 45.)

[46] Hall, P., *On some simple estimates of an exponent of regular variation*. J Roy Statist Soc Ser B, 44, 37−42, 1982. (Cited on page 89.)

[47] Härdle, W. and Muller, M., *Multivariate and semiparametric kernel regression*, In Schimek, M. G., editor: *Smoothing and Regression: Approaches, Computation and Application*. Wiley Series in Probability and Mathematical Statistics, 357−392, John Wiley & sons, 2000. (Cited on page 45.)

[48] Hastie, T. and Tibshirani, R., *Generalized additive models*. Statistical Science, 1:297−318, 1986. (Cited on pages 40 and 47.)

[49] Hastie, T. and Tibshirani, R., *Generalized Additive Models*. volume 43, Monographs on Statistics and Applied Probability, Chapman & Hall, 1990. (Cited on pages 46, 47, 82, 153 and 157.)

[50] Hastie, T. J., Tibshirani, R.J., and Friedman, J., *The elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics, Springer, New York, 2001. (Cited on pages 44 and 46.)

[51] Hawkins, T. *Cauchy and the spectral theory of matrices*. Historia Mathematica. 2:1˘29, 1975. (Cited on page 106.)

[52] Heuchenne, C., Lopez, O. and Hambuckers, J., *A semiparametric model for generalized pareto regression based on a dimension reduction assumption*. Unpublised work, 2014. (Cited on pages 17 and 80.)

[53] Hill, B. M., *A simple general approach to inference about the tail of a distribution*. Annals of Statistics, 3:1163−1174, 1975 (Cited on page 77.)

[54] Hong, L.J., Juneja, S. and Liu, G., (2017) *Kernel smoothing for nested estimation with application to portfolio risk measurement*. Operations Research, 65, 657−673, 2017. (Cited on pages 101 and 102.)

[55] Hosking, J. R. M. and Wallis, J. R., *Parameter and quantile estimation for the generalized Pareto distribution*. Technometrics, 29:339−349, 1987. (Cited on page 78.)

[56] Hull, J. and White, A., *Pricing interest-rate derivative securities*. The Review of Financial Studies, 3(4): 573−592, 1990. (Cited on page 108.)

[57] Hull, J. and White, A., *One factor interest rate models and the valuation of interest rate derivative securities*. Journal of Financial and Quantitative Analysis, 28(2): 235−254, 1993. (Cited on page 108.)

[58] Hull, J. and White, A., *Numerical procedures for implementing term structure models I*. Journal of Derivatives, 7–16, Fall 1994. (Cited on page 108.)

[59] Hull, J. and White, A., *Numerical procedures for implementing term structure models I*. Journal of Derivatives, 37–48, Winter 1994. (Cited on page 108.)

[60] Hull, J. and White, A., *Using Hull–White interest rate trees* Journal of Derivatives, 3(3), 26˘36, 1996. (Cited on page 108.)

[61] Israel, R. B., Rosenthal, J. S. and Wei, J. Z., *Finding generator for Markov chains via empirical transition matrices with application to credit ratings*. Mathematical Finance, 11(2): 245−265, 2001. (Cited on pages 120 and 122.)

[62] Jagannathan, R., Kaplin, A. and Sun, S., *An evaluation of multi-factor CIR models using LIBOR, swap rates, and cap and swaption prices*. Journal of Econometrics, 116(1−2):113−146, 2003. (Cited on page 110.)

[63] Jarrow, R. A., Lando, D. and Turnbull, S. M., *A Markov Model for the Term Structure of Credit Risk Spreads*. The Review of Financial Studies, 10 (2): 481−523, 1997. (Cited on page 119.)

[64] Kemp, M., *Market Consistency*. Wiley Finance, 2009. (Cited on page 33.)

[65] Klinke, S. and Grassmann, J., *Projection pursuit regression*, In Schimek, M. G., editor: *Smoothing and Regression: Approaches, Computation and Application*. Wiley Series in Probability and Mathematical Statistics, 471−496, John Wiley & sons, 2000. (Cited on page 46.)

[66] Koursaris, A., *Improving capital approximation using the curve-fitting approach*. Barrie & Hibbert (working paper), 2011. (Cited on page 8.)

[67] Koursaris, A., *The advantages of least squares Monte Carlo*. Barrie & Hibbert. Available at `http://www.barrhibb.com/documents/downloads/The_Advantages_of_Least_Squares_Monte_Carlo.pdf`. Accessed 22 June 2018, 2011. (Cited on page 11.)

[68] Koursaris, A., *A least squares Monte Carlo approach to liability proxy modelling and capital calculation*. Barrie & Hibbert. Available at `http://www.barrhibb.com/documents/downloads/Least_Square_Monte_Carlo_Approach_to_Liability_Proxy_Modelling_and_Capital_Calculation.pdf`. Accessed 22 June 2018, 2011. (Cited on page 11.)

[69] Koursaris, A., (2011) *A primer in replicating portfolios*. Barrie & Hibbert. Available at `http://www.barrhibb.com/documents/downloads/Primer_in_Replicating_Portfolios.pdf`. Accessed 22 June 2018, 2011. (Cited on page 13.)

[70] KPMG, *Technical Practices Survey 2015 Solvency II*, Available at `https://assets.kpmg.com/content/dam/kpmg/pdf/2016/04/TPS_2015.pdf`, 2015. (Cited on page 6.)

[71] Kullback, S. and Leibler, R. A., *On information and sufficiency*. Ann. Math. Stat., 22(1):79−86, 1951. (Cited on page 81.)

[72] Lan, H. and Nelson, B. L. and Staum, J., *A confidence interval procedure for expected shortfall risk measurement via two-level simulation*. Operations Research, 58, 1481−1490, 2007. (Cited on page 14.)

[73] Lando, D. *Credit Risk Modelling: Theory and Practice*. Princeton University Press, 2004. (Cited on page 119.)

[74] Lamberton, D., *Optimal stopping and American options*. Lecture note, available at `https://www.fmf.uni-lj.si/finmath09/ShortCourseAmericanOptions.pdf` (Cited on page 10.)

[75] Leadbetter, M. R., *On a basis for Peaks over Threshold modeling*. Statistics and Probability Letters. 12(4): 357˘362, 1991. (Cited on page 80.)

[76] Li, Y. C. and Yeh, C. C., *Some characterizations of convex functions*. Computers & Mathematics with Applications, 59(1), 327−337, 2010. (Cited on page 85.)

[77] Linton, O. and Nielsen, J. P., *A kernel method of estimating structured nonparametric regression based on marginal integration*. Biometrika, 82, 93−100, 1995. (Cited on page 157.)

[78] Longstaff, F. and Schwartz, E. *Valuing American Options by Simulation: A Simple Least-Squares Approach*. Finance, Anderson Graduate School of Management, UC Los Angeles, 2001. (Cited on page 8.)

[79] Lye, L. M., Hapuarachchi, K. P. and Ryan, S., *Bayes Estimation of the Extreme-Value Reliability Function*. IEEE Transactions on Reliability, 42, 1993. (Cited on page 77.)

[80] Marra, G. and Wood, S. N., *Coverage Properties of Confidence Intervals for Generalized Additive Model Components*. Scand Stat Theory Appl, 39, 53−74, 2012. (Cited on pages 155 and 158.)

[81] McNeil, J. A. and Frey, R., *Estimation of tail-related risk measures for heteroscedastic financial time series: an extreme value approach*. Journal of Empirical Finance, 7, 271−300, 2000. (Cited on page 77.)

[82] Natolski, J. and Werner, R., *Mathematical analysis of different approaches for replicating portfolios*. Eur Actuar J, 4(2), 411−435, 2014. (Cited on page 13.)

[83] Nesterov, Y., *Gradient methods for minimizing composite objective function*. CORE, 2007. (Cited on page 87.)

[84] Nychka, D., *Bayesian confidence intervals for smoothing splines*. J Am Stat Assoc, 83, 1134−1143, 1988. (Cited on pages 52, 155 and 156.)

[85] Opsomer, J. D. and Ruppert, D., *Fitting a bivariate additive model by local polynomial regression*. Ann Stat, 25, 186−211, 1997. (Cited on page 157.)

[86] Opsomer, J. D., *Asymptotic properties of backfitting estimators*. Journal of the American Statistical Association, 93:605−619, 2000. (Cited on page 157.)

[87] O'Sullivan, F., *A statistical perspective on ill-posed inverse problems*. Stat Sci, 1, 505−527, 1986. (Cited on page 154.)

[88] Pelsser, A. and Schweizer, J., *The difference between LSMC and replicating portfolio in the insurance liability modeling*. Eur Actuar J, 6:441−494, 2016. (Cited on page 101.)

[89] Pickands, J., *Statistical Inference Using Extreme Order Statistics*. Annals of Statistics, 3:119−131, 1975. (Cited on pages 74, 75, 77 and 80.)

[90] European Commission, *QIS5 Technical Specifications, Annex to Call for Advice from CEIOPS on QIS5*. `https://eiopa.europa.eu/Publications/QIS/QIS5-technical_specifications_20100706.pdf` (Cited on page 36.)

[91] Ramsay, T. O., Burnett, R. T. and Krewski, D., *The effect of concurvity in generalized additive models linking mortality to ambient particulate matter*. Epidemiology, 14(1):18−23, 2003. (Not cited.)

[92] Redfem, D., *Low discrepancy numbers and their use within the ESG, Moody's Analytics*, 2010. (Cited on page 108.)

[93] Règlement délégué (UE) 2015/35 de la commission du 10 octobre 2014 complétant la directive 2009/138/CE du Parlement européen et du Conseil sur l'accès aux activités de l'assurance et de la réassurance et leur exercise (solvabilité II) [French version] `https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32015R0035&from=EN` (Cited on pages 3 and 15.)

[94] Reinsch, C., *Smoothing by spline functions II*. Numer Math, 16, 451−454, 1971. (Cited on page 153.)

[95] Resnick, S., *Extreme values, regular variation, and point process*. Springer, 2008. (Cited on page 79.)

[96] Ruppert, D. and Carroll, R. J., *Spatially Adaptive Penalties for Spline Fitting*. Aust N Z J Stat, 42, 205−233, 2000. (Cited on pages 153 and 154.)

[97] Schimek, M. G. and Turlach, B. A., *Additive and generalized additive models*, In Schimek, M. G., editor: *Smoothing and Regression: Approaches, Computation and Application*. Wiley Series in Probability and Mathematical Statistics, 229−276, John Wiley & sons, 2000. (Cited on page 40.)

[98] Shannon, C. E., *A Mathematical Theory of Communication*. Belle System Technical Journal, 27(3) : 379−423, 1948. (Cited on page 81.)

[99] Silverman, B. W., *Some aspects of the spline smoothing approach to non-parametric regression curve fitting*. J R Stat Soc Series B Stat Methodol, 47, 1−52, 1985. (Cited on page 155.)

[100] Simon, N., Friedman, J., Hastie, T. and Tibshirani, R., *A Sparse-Group Lasso*. Journal of Computational and Graphical Statistics, 22, 2013. (Cited on pages 17, 80 and 87.)

[101] Smith, R. L., *Maximum Likelihood Estimation in a Class of Non-Regular Cases*. Biometrika, 72(1):67−90, 1985. (Cited on page 78.)

[102] Smith, R. L., *Estimating tails of probability distributions*. Annals of Statistics, 15:1174−1207, 1987. (Cited on page 78.)

[103] Smith, R., *Extreme Value Analysis of Environmental Time Series: an Application to Trend Detection in Ground-Level Zone*. Statistical Science, 4: 367−393, 1989. (Cited on page 77.)

[104] *Solvency II Glossary*, `http://ec.europa.eu/internal_market/insurance/docs/solvency/impactassess/annex-c08d_en.pdf` (Cited on page 38.)

[105] Stentoft, L., *Convergence of the Least Square Monte-Carlo Approach to American Option Valuation*. Manage Sci, 50(9), 2004. (Cited on page 11.)

[106] Stone, C., *Optimal global rates of convergence for nonparametric regression*. Annals of Statistics, 10:1040−1053, 1982. (Cited on page 48.)

[107] Stone, C., *Additive regression and other nonparametric models*. Annals of Statistics, 13(2):689−705, 1985. (Cited on pages 47, 82, 101 and 102.)

[108] Stone, C., *The dimensionality reduction principle for generalized additive models*. Annals of Statistics, 14:590−606, 1986. (Cited on page 48.)

[109] Teuquia, O. N., Ren, J. and Planchet, F., *Internal model in life insurance: application of least squares monte-carlo in risk assessment*, 2014. (Cited on page 11.)

[110] Tibshirani, R., *Regression shrinkage and selection via the lasso*. J R Stat Soc Series B, 58:267−288, 1996. (Cited on page 85.)

[111] Tomas, B., *Arbitrage Theory in Continuous Time*. (Fourth Edition), Oxford Finance Series. (Cited on page 33.)

[112] Tsybakov, A. B., *Introduction to Nonparametric Estimation*. Springer Series in Statistics, 2003. (Cited on page 40.)

[113] van der vaart, A. W. and Wellner, Jon, *Weak Convergence and Empirical Process*. Springer Series in Statistics, 1996. (Cited on page 160.)

[114] Vidal, E. G. and Daul, S., *Replication of insurance liabilities*. RiskMetrics Journal, 9, 79−96, 2009. (Cited on page 13.)

[115] Wahba, G., *Bayesian confidence intervals for the cross validated smoothing spline*. J R Stat Soc Series B Stat Methodol 45, 133-150, 1983. (Cited on pages 52 and 155.)

[116] Wand, M. P., *Central Limit Theorem for Local Polynomial Backfitting Estimators*. J Multivar Anal, 70, 57−65, 1999. (Cited on page 157.)

[117] Wang, T. and Hsu, C., *Board composition and operational risk events of financial institutions*. Journal of Banking & Finance, 37(6):2042−2051, 2013. (Cited on pages 17 and 79.)

[118] Wetherill, G. B., *Regression Analysis with Applications*. Monographs on Statistics and Applied Probability, volume 27, Chapman & Hall, 1986. (Not cited.)

[119] Wood, S. N., *On confidence intervals for generalized additive models based on penalized regression splines*. Aust N Z J Stat, 48, 445−464, 2006. (Cited on pages 155 and 158.)

[120] Yuan, M. and Lin, Y., *Model selection and estimation in regression with grouped variables*. J R Stat Soc Series B, 68(1):49−67, 2007. (Cited on page 85.)

Résumé: Les accords de Bale et les directives européennes associées ont conduit à conditionner les capitaux prudentiels des banques à leur profil de risque plutôt qu'à leur taille ou chiffre d'affaires. La directive Solvabilité 2 (ci-après la "directive") répète ce processus pour les assureurs et réassureurs européens. Elle constitue un changement total de paradigme pour la majorité des assureurs européens. Elle définit les grands principes réglementaires visant à encadrer leur activité et en particulier à déterminer le montant des capitaux prudentiels associés aux risques inhérents à leur activité.

Conformément à la directive, le capital prudentiel correspond en principe pour un assureur au quantile à 99.5% de la variation de ses fonds propres sur l'année à venir. Une telle mesure de risque prospective requiert pour un assureur la capacité d'adresser deux problèmes : un problème de valorisation et un problème de simulation. En pratique, le quantile à 99.5% de la variation de ses fonds propres est estimé par méthode de Monte-Carlo. Il est particulièrement sensible à la loi jointe à un an retenue pour le vecteur de facteurs de risque $\mathbf{x}$. Son évaluation par méthode de Monte-Carlo nécessiterait idéalement de simuler m réalisations du facteur de risque $\mathbf{x}$ à un an et d'évaluer les valeurs des fonds propres associées. Compte tenu du temps de calcul important nécessaire à l'évaluation numérique, cette approche s'avère en pratique inadaptée. De manière à contourner ce problème, les opérationnels ont mis au point de nombreuses méthodes d'approximation ou « proxys » qui permettent d'en approximer la valeur de manière instantanée. Aujourd'hui, ces méthodes sont rarement accompagnées de contrôles d'erreur qui permettraient d'en mesurer la qualité. Plus précisément, les méthodes actuellement utilisées par les opérationnels ne permettent pas de contrôler naturellement l'erreur d'approximation engendrée par l'utilisation du modèle proxy en lieu. Les contrôles d'erreur proposés sont donc toujours empiriques et trop approximatifs.

Afin de résoudre cette problématique, nous proposons, dans une première partie de cette thèse, une nouvelle méthode de construction du proxy à la fois économe en ressources informatiques et offrant un contrôle d'erreur rigoureux. La deuxième partie de cette thèse a pour l'objectif d'appliquer la théorie de la valeur extrême à l'estimation du capital prudentiel lorsque l'information sur la covariable est disponible. En particulier, lorsque la covariable est de grande dimension, nous sommes confrontés au problème de la "curse of dimensionality", qui se traduit par une diminution des taux de convergence les plus rapides possibles des estimateurs de la fonction de régression vers leur courbe cible. Ce problème fait référence au phénomène où le volume de la covariable augmente si rapidement que les données disponibles deviennent rares. Pour obtenir un résultat statistiquement fiable, la quantité de données nécessaire à l'appui du résultat augmente souvent de manière exponentielle avec la dimensionnalité, ce qui est généralement problématique dans de nombreuses applications pratiques. Pour surmonter ce problème d'estimation, nous proposons une nouvelle méthodologie d'évaluation efficace en combinant le modèle additif généralisé et la méthode de sparse group lasso.

Mots-clé: Solvabilité 2, Assurance Vie, regression bayésienne par splines pénalisées, sparse group lasso, modèle additif généralisé.

Discipline: Mathématiques

Abstract: The Basel agreements and the associated European directives have made banking prudential capital contingent on their risk profile rather than on their size or turnover. The Solvency 2 Directive (hereinafter the "Directive") repeats this process for European insurers and reinsurers. It constitutes a total paradigm shift for the majority of European insurers. It defines the main regulatory principles aimed at regulating their activity and in particular determining the amount of prudential capital associated with the risks inherent to their activity.

In accordance with the directive, the prudential capital corresponds in principle to an insurer with a 99.5% percentile of the change in its basic own funds over the coming year. Such a prospective risk measure requires for an insurer the ability to address two problems: a valuation problem and a simulation problem. In practice, the 99.5% percentile of the change in basic own funds is estimated using the Monte Carlo method. It is particularly sensitive to the one-year law retained for the risk factors vector. Its Monte Carlo valuation would ideally require the simulation of one-year risk factor vector $\mathbf{x}$ and the valuation of the associated equity values. Given the significant calculation time required for numerical evaluation, this approach is in practice unsuitable. In order to circumvent this problem, the insurers have developed many approximate methods or "proxies" which make it possible to approximate the basic own funds value instantaneously. Today, these methods are rarely accompanied by error controls that would measure the simulation quality. More precisely, the methods currently used by the insurers do not make it possible to control naturally the approximation error generated by the use of the proxy model instead. The proposed error checks are therefore always empirical and too approximate.

In order to solve this problematic, we propose, in a first part of this thesis, a new method of constructing the proxy that is both resource-efficient and offers rigorous error control. The second part of this thesis aims at applying the extreme value theory to the prudential capital estimate when information on the covariate is available. In particular, when the covariate is high dimensional, we are confronted with the problem of the curse of dimensionality, which translates into a decrease in the fastest possible convergence rates of estimators of the regression function to their target curve. This problem refers to the phenomenon where the volume of the covariate increases so rapidly that available data become sparse. To obtain a statistically reliable result, the amount of data needed to support the result often increases exponentially with dimensionality, which is generally problematic in many practical applications. To overcome this estimation problem, we propose a new efficient evaluation methodology by combining the generalized additive model and the sparse group lasso method.

Key Words: Solvency 2, Life Insurance, Bayesian penalized spline regression, sparse group lasso, generalized additive model.