FACTOR MODELS FOR CALL PRICE SURFACE WITHOUT STATIC ARBITRAGE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Fan Zhu

January 2012

FACTOR MODELS FOR CALL PRICE SURFACE WITHOUT STATIC ARBITRAGE

Fan Zhu, Ph.D.

Cornell University 2012

Although stochastic volatility models and local volatility model are very popular among the market practitioner for exotic option pricing and hedging, they have several critical defects both in theory and practice. We develop a new methodology for equity exotic option pricing and hedging within the market-based approach framework. We build stochastic factor models for the whole surface of European call option prices directly from the market data, and then use this model to price exotic options, which is not liquidly traded. The factor models are built based on Karhunen-Loeve decomposition, which can be viewed as an infinite dimensional PCA. We develop the mathematical framework of centered and uncentered versions of the Karhunen-Loeve decomposition and study how to incorporate critical shape constraints. The shape constraints are important because no static arbitrage conditions should be satisfied by our factor models. We discuss this methodology theoretically and investigate it by applying to the simulated data.

BIOGRAPHICAL SKETCH

Fan Zhu was born in Beijing, the capital of People's Republic of China and grew up there. He did his undergraduate studies in the School of Mathematical Sciences at Peking University and received his Bachelor of Science degree in Mathematics in July 2004. Then he entered the Department of Mathematics at State University of New York at Stony Brook and received the Master of Science degree in Mathematics in July 2006. That same year, he was admitted to the Ph.D. program in the field of Operations Research and Information Engineering at Cornell University. His doctoral research in Financial Mathematics, under the guidance of Professor Martin Wells, focused on exotic option pricing and hedging.

To my parents:

Yue Zhu and Hui Yang

For their love and support

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Professor Martin Wells for his invaluable guidance and encouragement. He is such a brilliant scholar and an amazing mentor. His deep knowledge and great spirit helped me overcome numerous obstacles throughout my graduate studies at Cornell. Our meetings were always filled with delightful discussions. I feel so fortunate and grateful to have had the opportunity of working with Marty.

I would like to thank Professor Robert Jarrow and Professor Michael Nussbaum and Professor Peter Jackson for serving on my special committee. I learned a lot from their well-designed courses in Finance and Statistics. I sincerely appreciate Professor Dawn Woodard for serving as a proxy in my dissertation defense.

I am very grateful to Prof. Johannes Wissel for graciously sharing his time and knowledge. The discussions with him were always delightful and inspiring to me.

I would like also to thank Prof. Alexander Schied and Prof. Stefan Weber for introducing me to the field of Mathematical Finance in their courses during their stay at Cornell University.

In addition, my thanks go to many friends I have met during my years in Ithaca, to name a few, Jie Chen, Yinan Huang, Yuohua Wu, Yuemeng Sun, Juan Li, Chao Ding, Jiawei Qian, Yi Shen, Anjie Guo.

Last but not least, I would like to thank my parents for their unconditional love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation and Background

The pricing and hedging of derivatives of (tradable) underlying assets is one of the most important and interesting topics in financial engineering for both theoretical and practical purposes. The value of a derivative depends on the underlying assets, so the modeling of the dynamics of the underlying assets is essential in the derivative pricing and hedging problems. In 1900, Louis Bachelier [5] introduced Brownian motion as a model for stock prices and Paul Samuelson [65] proposed to use geometric Brownian motion as an improvement in 1965. Black and Scholes [10] and Merton [59] derived the famous Black-Scholes formula for call (put) option prices under the geometric Brownian motion assumption for the dynamics of the underlying assets.

Although the Black-Scholes formula has been very popular among market practitioners, the geometric Brownian motion assumption is simply unrealistic. We just mention two main criticisms here. First, the geometric Brownian motion model implies that the log return must be normally distributed. However, empirical research show that the log returns of (traded) financial assets are often leptokurtic (heavy tailed) and skewed. For example, Bollerslev [11] finds leptokurtosis in monthly log S&P 500 returns, while French, Schwert and Stambaugh [35] report skewness in daily log S&P 500 returns. Engle and Gonzales-Rivera [29] find excess skewness and kurtosis in small stocks and in exchange rates as well. Secondly, for a fixed underlying asset, there is no single volatility parameter σ that causes the Black-Scholes formula to give the correct (market)

price of options written on that asset with a variety of different strikes and expiration times. This is the so-called volatility smile / smirk [22, 26, 43, 34, 44, 63]. Actually, in practice the Black-Scholes formula is very often reduced to a means of quoting option prices in term of another parameter, the (Black-Scholes) implied volatility, when applied to vanilla call and put options. The implied volatility $\sigma_t^{BS}(K,T)$ of a call (put) option with strike K and maturity date T is obtained by inverting the Black-Scholes formula given the market price of the option. Here we can get the market prices for the call and put options because call and put options on major stocks, indices or currencies are traded liquidly (at least for some strikes and maturities) on derivative markets today. Hence it is possible, and customary, to price these vanilla options according to the market, instead of using pricing models. Also these market prices of vanilla options can be used as a benchmark for the performance of pricing models.

There are lots of derivatives that have features making it more complex than vanilla options, for example, the barrier options, lookback options, and the forward start options. These derivatives are called exotic options and they are usually traded over-the-counter (OTC), which means that they are traded directly between two parties, without going through an exchange or other intermediary. Therefore we need option pricing models to price and hedge these exotic options. More concretely, on the pricing side, we need models to price exotic options consistently with the market prices of vanilla options, that is, we want models that we can calibrate to options that are liquidly traded and then use to price options that are not traded. On the hedging side, we want models that we can use to derive the hedging portfolio (static or dynamic) to make the total portfolio (book plus hedge) to be insensitive to the price changes. Black-Scholes model is not sophisticated and realistic enough for these purposes and people

have developed a considerable literature on alternative pricing models.

Here we will focus on equity derivative pricing models and the models in the continuous framework (we do not discuss the models with jumps in this thesis). Two main strands of research have been developed and some of the models are very popular in practice: stochastic volatility models and local volatility model.

Instead of assuming that the volatility of the underlying is constant (the geometric Brownian motion as in Black-Scholes model), stochastic volatility models view the it as a stochastic process and use SDEs to specify the dynamics of the volatility (or the variance). Examples are Hull and White model [45], Wiggins model [74], Stein and Stein model [72], Scott model [69], Heston model [42] and SABR model [30]. Heston model and SABR model are most popular among the market practitioner due to their tractability. The Heston model describes the dynamics of the underlying and the volatility by the following system of SDEs (in objective measure):

$$dS_t = \mu S_t dt + \sqrt{V_t} S_t dB_t^1,$$

$$dV_t = \kappa(\theta - V_t)dt + \sigma\sqrt{V_t} dB_t^2,$$

$$\langle B^1, B^2 \rangle_t = \rho_t,$$

where S_t is the underlying and B_t^1 and B_t^2 are two standard Brownian motions, $\langle B^1, B^2 \rangle_t$ is the quadratic covariation of B_t^1 and B_t^2, given by

$$\langle B^1, B^2 \rangle_t = \frac{1}{2}(\langle B^1 + B^2 \rangle_t - \langle B^1 \rangle_t - \langle B^2 \rangle_t).$$

The SABR model uses the following system of SDEs (in risk neutral measure):

$$dS_t = \sigma_t S_t^\beta dB_t^1,$$

$$d\sigma_t = \alpha \sigma_t dB_t^2,$$

$$\langle B^1, B^2 \rangle_t = \rho_t.$$

The analytic or semi-analytic formulae for European call option prices are available in both the Heston and SABR models. This makes the calibration of these models to the market possible.

Problems often arise in the calibration procedure for stochastic volatility models. First, the calibration is usually a nonlinear optimization problem that is computationally expensive. Moreover, if the number of input option prices exceeds the number of model parameters (this is often the case since in Heston model or SABR model, for example, there are 4 or 5 parameters while there are at least 10 to 20 liquidly traded call options written on, for example, S&P 500, in the market), a conflict arises between different calibration constraints. This problem, already present at the static level, becomes more acute if one examines the model dynamics with those observed in the option market. This time instability of the calibration makes the re-calibration have to be performed in a very frequent basis and leads to large variations in sensitivities and hedge parameters, which is problematic. Also many empirical evidence [6, 19, 22] has shown that one factor stochastic volatility models (which of course include the Heston model and SABR model) do not fit the (market) observed implied volatility patterns well (even after adding simple jumps). These problems make it difficult to use them in practice to price and hedge exotic options.

The concept of local volatility model was developed when Dupire [27] and Derman and Kani [24] noted that there is a unique diffusion process consistent with the risk neutral densities derived from the market prices of European call options. The idea is to find the diffusion coefficient (local volatility function) $\sigma_t = \sigma(S_t, t)$ consistent with the set of market prices of call options with all

strikes and maturities. Then the model can be used to price exotic options consistently with the observed prices of call options. The famous Dupire formula is:

$$dS_t = \sigma(S_t, t)dB_t,$$
$$\sigma(K,T) = \frac{1}{K}\sqrt{\frac{2\frac{\partial C(K,T)}{\partial T}}{\frac{\partial^2 C(K,T)}{\partial K^2}}}.$$

The term "local volatility" has subsequently been extended to cover any deterministic volatility model where forward volatilities are a function of time and the underlying price. For a detailed review of these models, see [62, 70].

While local volatility model has the advantage of market completeness (the only random source is from the underlying), it is not very convenient for pricing exotic options. As we can see, to apply Dupire formula, we need to arbitrarily pre-interpolate the input call price surface, since we can only observe discrete set of call prices from the market (with finitely many strikes and maturities) while in Dupire formula the whole surface is involved. Therefore the results would be very sensitive to the different ways of interpolation (note that there are first and second order derivatives involved). A result of the single randomness in the model is that there is no more freedom to calibrate to exotic options. Also, since the volatility is a deterministic function of the underlying in local volatility model, these kind of models are not very well used to price cliquet options or forward start options, whose values depend specifically on the random nature of volatility itself. Moreover, as the case in stochastic volatility models, local volatility surface is also very instable over time [26].

The inability of stochastic volatility models and local volatility model is due to several reasons. First, the stochastic process used for describing the underlying dynamics (together with the dynamics of the volatility) might be mis-

specified in the first place, that is, the parametric form might be wrong (we don't know the number of randomness sources and the forms). Then it is natural that there would be problems in the calibration procedures. Moreover, since the creation of organized option markets in 1973, these markets have become increasingly autonomous and option prices are driven not only by the movements in the underlying asset, but also by internal supply and demand in the option market itself. This fact is supported by some empirical evidence on the relationship between the option markets and their underlying markets [2, 7].

An alternative approach has been developed to overcome these difficulties in stochastic volatility models and local volatility model, which is called the market models for implied volatility (or the market-based approach). This approach is inspired by the fact that many European call options are liquidly traded and their prices are given by the market. The approach with the same spirit has been applied to the pricing and hedging for interest rate derivatives, that is the famous HJM methodology [41]. This idea was first applied to the equity derivatives in Lyons [58], Schönbucher [66] and Ledoit et al. [51] and later developed in Schweizer and Wissel [67, 68], Wissel [75], Jacod and Protter [46] and Carmona and Nadtochiy [16]. In the market-based model framework, the European call option prices (or Black-Scholes implied volatilities, or other parametrizations) are jointly modeled by a system of (infinitely many) SDEs, together with the underlying price dynamics.

Such a framework has several immediate advantages over the stochastic volatility models and local volatility model. First, option prices (or Black-Scholes implied volatilities) are market observables which are reasonable to be incorporated into the model. They are directly derived from market with-

out making any modeling assumptions. By contrast, the local volatility or spot volatility (which is modeled in stochastic volatility models) are not directly observable and has to be "calibrated" with option price data or assumed to have some parametric forms.

Secondly, even if a perfect calibration to market data is possible for some stochastic volatility model at a given time, the parameters may not be stable and model prices resulting from this calibration may deviate from the market prices as time evolved. By contrast, market models are not only perfectly calibrated to market prices at initial time by construction, but can also better match the future states of the markets.

Thirdly, market models can recognize the extra sources of randomness specific to the option market (other than randomness in the underlying price dynamics) and incorporate the statistical features of their dynamics into the model while stochastic volatility models and local volatility model don't consider randomness that may affect the derivative pricing and hedging other than the underlying dynamics.

Finally, market models can serve as a framework for hedging exotic derivatives with liquidly traded options (vanilla call options, for example).

While possessing these advantages, the market-based approach has its own shortcomings. To serve the purpose of pricing and hedging for exotic options, we need to rule out the arbitrage opportunities (both static arbitrage and dynamic arbitrage). For the static arbitrage, We will investigate in more details in Section 3.1. For the dynamic arbitrage, it was first pointed out by Lyons [58] and Schönbucher [66] that for such a model to be arbitrage-free, the coefficients

in the system of SDEs cannot be arbitrarily specified, but must be linked by certain relations, which is called drift restrictions. This is similar to the case in HJM models. If one takes these drift restrictions into account, the question whether the system of SDEs admits a solution turns out to be nontrivial.

The parametrization plays a significant role in solving this problem. We find some "good" parametrizations in which the no static arbitrage conditions and drift restrictions are simple enough to deal with and the existence of the solutions under these conditions can be proved at the same time. That means, we need to get the necessary and sufficient conditions for a no-arbitrage market model to exist in some parametrization. When we consider the whole call option price surface (with all strikes and maturities), the naive parametrization (call option price itself) and the Black-Scholes implied volatility are both not proper for this purpose. Local volatility is a relatively simple parametrization and it has some economic interpretations. This parametrization has been used in [16], in which Carmona and Nadtochiy consider the full option price surface models and give the drift restrictions (no static arbitrage condition is simply that the local volatility is nonnegative). However, if we use local volatility as the parametrization and consider the full option price surface models, the existence theorem (sufficient conditions) is too difficult to prove and there has not been this kind of results. As I know, the best result up to now is given in [75], where Wissel considers models for a family of call option prices with all strikes in $\mathcal{K} = \{K_1, \cdots K_N\}$ and maturities in $\mathcal{T} = \{T_1, \cdots T_M\}$. The parametrization used is the local implied volatility and the price level, which is very complicated mathematically and difficulty to interpret in economics. The drift restrictions and existence theorem for this kind of models are given in [75]. However, models for the full option price surface, that is, with all $K > 0$ and $T > 0$, are too

complicated and there has not been similar results in this case. Even for the finitely many strikes and maturities case in [75], the parametrization and the model itself are so complex that it has thus far prevented the implementation of such a model.

The goal of this thesis is to move a step forward in the direction of market based approach and investigate some of the properties (no static arbitrage conditions) in this class of models.

1.2 Another Framework in the Market-Based Approach

As described above, we will work in the market-based approach framework and hence we will view the European call option prices observed from the market as an input and build a models for the call price surface for exotic derivative pricing and hedging. However, the framework in this thesis is different from the market models introduced in Section 1.1 in the sense that it will not assume a system of SDEs to describe the call price surface in advance, but will try to investigate the properties of the call price surface and build a model for it directly from the market data. Then this model can be used for pricing and hedging purposes for exotic derivatives.

1.2.1 Connecting Implied Volatility and Spot Volatility

Here we present some explanation how the model for call price surface (implied volatility surface) can serve the purpose for exotic derivative pricing and hedging.

In the general framework of mathematical finance (or more specifically, continuous time option pricing), the (single underlying asset) market is modeled by filtered probability space $(\Omega, (\mathcal{F}_t), \mathcal{F}, P)$, where the filtration $(\mathcal{F}_t)$ is generated by a d-dimensional Brownian motion X_t. $S_t > 0$ is the price process for the underlying asset and the short rate process r_t is progressive (with respect to $(\mathcal{F}_t)$) and bounded. We suppose that there is at least one risk-neutral measure (that is, the market is arbitrage-free) and P is a risk-neutral measure. Then by the Martingale Representation Theorem [48], it can be shown that there exits a progressive σ_t, and 1-dimensional Brownian motion B_t, with respect to $(\mathcal{F}_t)$, such that

$$dS_t = \sigma_t S_t dB_t + r_t S_t dt,$$

where σ_t is the spot volatility of the underlying asset. If we assume the dynamics of σ_t can be describe by some SDE, then we get a stochastic volatility model.

Instead of specifying the dynamics of σ_t in advance, we would like to construct it from the market data. This is possible by the results in [28], if we have a model for the implied volatility (call option price) surface. In [28], Durrleman shows that under some relax regularity conditions, the implied volatility for at the money call option (that means, the strike is equal to the underlying asset price) will converge to the the spot volatility when the maturity goes to the current time level (that is, when $T - t \longrightarrow 0$).

1.2.2 Factor Models for the Call Option Price (Implied Volatility) Surface

Now in this framework, the plan is:

Step 1. Build a model without arbitrage for the European call price surface from the market data. We will build this model following the market-based philosophy, that means, building the model from the observed market data (call option prices) without specifying any parametric form in advance.

Step 2. Use the inverse Black-Scholes formula to transform the call price into the implied volatility, hence we get a model for the implied volatility surface.

Step 3. Apply Durrleman's result to get a description of the spot volatility from the implied volatility surface.

Once Step 3 is applied, the remaining things for pricing and hedging are the same as in stochastic volatility models (For details, see [9, 15, 38, 52], for example). Notice that in Step 1, the model is built directly from the market data, this methodology has advantages over stochastic volatility models and local volatility model in the sense that they avoid the problematic calibration procedures and capture the randomness specific to the option market it self other than the underlying asset market.

The reason we use the call price instead of the implied volatility is that the no static arbitrage conditions are much simpler for call price than for implied volatility. Since we want to apply this model for exotic option pricing and hedging, we definitely want to insist on no arbitrage in the model. We will investigate the no static arbitrage conditions in details in Section 3.1.

Since Step 2 and Step 3 are trivial (at least, theoretically), in this thesis we will focus on Step 1, that is, build a model for the call price surface without arbitrage. It can be viewed as a foundation for a market-based approach for modeling the dynamics of the options market.

The methodology we will adopt is based on a statistical technique called Karhunen-Loéve decomposition and its variants and the model for the call price surface turns out to be a factor model. Factor models have been used to describe and analyze implied volatility surface dynamics by several authors in the literature. For example, Fengler, Härdle and Mammen [31] propose a semiparametric factor model to approximate the implied volatility surface in a finite dimensional function space. In particular, their approach is tailored to the degenerated design of the market data for implied volatility. Hafner [38] uses time series method (AR(1) process) and Principle Component Analysis (PCA) to analyze and identify factors and build a four-factor model for the DAX implied volatility surface. However, these factor models do not incorporate the no static arbitrage conditions in their construction and hence the pricing and hedging with these models might be problematic. By contrast, our aim is to build factor models without static arbitrage opportunities. Also, the way to identify the factors and to construct the model in this thesis is different from the works mentioned above.

1.3 Relevant Literatures

As mentioned, our methodology to build the factor model is inspired by the Karhunen-Loéve decomposition, which can be viewed as infinite dimensional Principle Component Analysis (PCA). The features and dynamic properties of implied volatility time series have been studied using PCA and functional PCA by many authors in the literature. In particular, Avellaneda and Zhu [77] perform a PCA of the term structure of at-the-money implied volatilities and model it with a GARCH process. Fengler, Härdle and Schmidt [32] carry out a similar

study on the term structure of the VDAX and report the presence of level, shift and curvature components in the deformation of the term structure. Skiadopoulos et al. [71] perform a PCA of implied volatility smiles of S&P 500 American options traded on the CME for different maturity buckets and distinguish two significant principal components. Alexander [3] performs a similar analysis but using the deviation of implied volatilities from the at-the-money volatility. Fengler, Härdle and Villa [33] use a common principle components approach to perform a joint PCA on implied volatility smiles of different maturities. While these works focus on implied volatility with a single strike (at-the-money, for example) or a single maturity (then use common principle components to compare among several different maturities), our aim is to consider modeling the whole surface of call option price containing all available maturities and strikes in the market simultaneously.

Cont and Fonseca [20] model the joint dynamics of all implied volatilities quoted on the market and showed the randomness might be captured by a small number of random factors. Their approach is based on the standard Karhunen-Loéve decomposition. However, they also don't consider the static arbitrage opportunity in the model and hence their model can be used for risk management purpose but not for exotic option pricing and hedging. Actually all the works mentioned above deal with implied volatility but not call price (in fact the object in [20] is the logarithm of implied volatility), which makes the no static arbitrage constraints become very messy.

Our goal is to build a factor model of the call price surface without static arbitrage so that it can be used for exotic option pricing and hedging. For this purpose, we will develop some variants of the standard Karhunen-Loéve de-

composition. We will see this in more details in Sections 2.2 and 3.2.

1.4 Outline of the Thesis

Throughout this thesis, we focus on the central theme: building a model for the call price surface which is static arbitrage free. To achieve this goal, we will modify the standard Karhunen-Loéve decomposition and develop a framework and apply it to the call option price surface. The detailed organization of the thesis is as follows.

Chapter 2 gives a general, as well as rigorous mathematical framework of the standard Karhunen-Loéve decomposition and develop an "uncentered" version of it. Mathematically it is given by the spectral decomposition of some operator. The mathematical framework is built in the language of functional analysis.

Chapter 3 investigates the no static arbitrage conditions for call price surface and the violations of these conditions in the truncation of the Karhunen-Loéve decomposition and the uncentered Karhunen-Loéve decomposition when building the factor model. Then we try to modify this methodology to satisfy these conditions.

Chapter 4 describes the implementation of the Karhunen-Loéve decomposition and the uncentered Karhunen-Loéve decomposition and then applies this method to analyze the simulated data for call price options. A simulation study of the procedure is also given.

Chapter 5 includes a brief summary and some future research directions in the market-based approach framework.

CHAPTER 2

MATHEMATICAL FRAMEWORK OF CENTERED AND UNCENTERED KARHUNEN-LOÉVE DECOMPOSITION

2.1 Introduction

Karhunen-Loéve decomposition was proposed independently by Karhunen [49] and Loéve [55]. But the method itself is known under a variety of names in different fields: Empirical Component Analysis [56], Quasiharmonic Modes [13], Proper Orthogonal Decomposition [57] and others. This method is used widely in the areas of quantum physics, meteorology, geophysics, electrical engineering (statistical recognition) and computer science (image processing). Mathematically, Karhunen-Loéve decomposition / expansion is a representation of a stochastic process as an infinite linear combination of orthogonal functions with random variable coefficients, analogous to a Fourier series representation of a deterministic function on a bounded interval. It is easy to generalize this method to high-dimensional case (random field case), like the multivariate Fourier series. Remember our purpose is to build models for the call price (implied volatility) surface, which depends on two variables (the strike K and the maturity T), therefore we will consider the random surface case in the next two sections of this chapter.

When applied to a discrete and finite process, that is, random vector case, Karhunen-Loéve decomposition degenerate to the well-known statistical technique principal component analysis (PCA), for which the vast theory is quite mature (see, for example, [4, 8]). PCA can be carried out by eigenvalue decomposition of a data covariance matrix or singular value decomposition of a

data matrix, usually after mean centering the data for each attribute. While the Karhunen-Loéve decomposition can be viewed as an infinite dimensional PCA, we need to be careful because the significant concepts in the procedure, "eigenvalue", "eigenvetor" becomes much more complicated in mathematics when we work in the infinite dimensional space. Although the Karhunen-Loéve procedure is described in many literatures (textbooks, research papers, technical reports), most of these descriptions are not satisfactory mathematically. We give a rigorous mathematical framework using functional analysis language for the centered and uncentered Karhunen-Loéve procedures in this chapter. Since we want to apply this method to call price surface, we will work on the random surface case. However, the ideas are the same if we consider random fields in any dimension and there is no technical difficulty to generate this framework to other dimensional random fields cases.

2.2 Karhunen-Loéve Decomposition

Functional analysis is a powerful tool when we deal with problems in infinite dimensional space. Consider the infinite dimension nature of the Karhunen-Loéve decomposition for random surface, it is necessary to draw support from some concepts and theorems in functional analysis in order to describe the procedure rigorously.

For simplicity, we will not give the rigorous definition for most of the basic concepts and the detailed proof for the theorem in functional analysis needed in this thesis. For the detailed treatments of these concepts and theorems, we refer to the standard references [64, 76].

Consider a random surface $U(x)$ in some probability space $(\Omega, \mathcal{F}, P)$, where $x \in [a,b] \times [c,d] = A \subset R^2$. We can think of $U(x)$ as the call option price surface at some time point and $x = (K,T)$ where K is the strike and T is the maturity. We wish to rigorously define the Karhunen-Loéve expansion for $U(x)$. For this purpose, we need a further assumption: $U(x)$ is square-integrable, which means that $\int_{\Omega \times A} U^2(\omega, x) d\omega dx < \infty$. Using functional analysis language, this assumption says that $U(x)$ is an element of the Hilbert space $L^2(\Omega \times A) = L^2(\Omega) \otimes L^2(A)$. For any fixed $\omega_0 \in \Omega$, $U(\omega_0, x)$ gives a square-integrable surface ($U(\omega_0, x) \in L^2(A)$) and for any fixed $x_0 \in A$, $U(\omega, x_0)$ gives a square-integrable random ($U(\omega, x_0) \in L^2(\Omega, \mathcal{F}, P)$).

Notice that PCA in finite dimensional space is done after mean centering, we do the same thing to $U(x)$, that is, subtracting $E(U(x))$ from $U(x)$. (We will discuss a similar decomposition without mean centering in the next section.) Since $E(U(x) - E(U(x))) = 0$, we can consider a new $U(x)$ with $E(U(x)) = 0$ without loss of generality.

Consider the covariance function $K(x,y) = Cov(U(x), U(y))$. It is a function defined on $A \times A \subset R^4$. Since $E(U(x)) = 0$, we have

$$K(x,y) = Cov(U(x), U(y)) = E(U(x)U(y)) = \int_{\Omega} U(\omega, x) U(\omega, y) d\omega.$$

Note that $K(x,y)$ is symmetric by definition. Furthermore, $K(x,y)$ defines a linear operator $\mathcal{K}$ on the Hilbert space $L^2(A)$ in the way that

$$\mathcal{K}(f)(x) = \int_A K(x,y) f(y) dy$$

for any element $f(x)$ in $L^2(A)$ (that is, any L^2 surface defined on A). Linear operator can be viewed as an analogue of matrix. Here the linearity of $\mathcal{K}$ is guaranteed by the linearity of integration.

The operator $\mathcal{K}$ has some nice properties:

1. $\mathcal{K}$ is symmetric (or more precisely, $\mathcal{K}$ is a self-adjoint operator). This means that

$$\langle f, \mathcal{K}(g)\rangle = \langle \mathcal{K}(f), g\rangle, \ \forall f, g \in L^2(A).$$

This can be seen immediately from the definition of $\mathcal{K}$ and the symmetry of $K(x, y)$

$$\begin{aligned}
\langle f, \mathcal{K}(g)\rangle &= \int_A f(x)\mathcal{K}(g)(x)dx \\
&= \int_A f(x) \int_A K(x, y)g(y)dydx \\
&= \int_{A\times A} K(x, y)f(x)g(y)dxdy \\
&= \int_A \int_A K(y, x)f(x)dxg(y)dy \\
&= \langle \mathcal{K}(f), g\rangle.
\end{aligned}$$

2. $\mathcal{K}$ is positive semidefinite (or more precisely, $\mathcal{K}$ is a positive operator). This means that

$$\langle f, \mathcal{K}(f)\rangle \geq 0 \ \forall f \in L^2(A).$$

This is easy to proof by some calculation

$$\begin{aligned}
\langle f, \mathcal{K}(f)\rangle &= \int_A f(x)\mathcal{K}(f)(x)dx \\
&= \int_{A\times A} K(x, y)f(x)f(y)dxdy \\
&= \int_{A\times A} f(x)f(y) \int_\Omega U(\omega, x)U(\omega, y)d\omega dxdy \\
&= \int_\Omega \int_A U(\omega, x)f(x)dx \cdot \int_A U(\omega, y)f(y)dyd\omega \\
&= \int_\Omega (\int_A U(\omega, x)f(x)dx)^2 d\omega \\
&\geq 0.
\end{aligned}$$

3. $\mathcal{K}$ is a compact operator, which means that $\overline{\mathcal{K}(B_1)}$ is a compact set in $L^2(A)$, where B_1 is the unit ball in $L^2(A)$. This is a standard result but not that easy to prove and we omit the proof here. However, this property is very important for us since there is an elegant theory for the spectrum of compact operator in functional analysis (Rieze-Schauder theory) and the spectrum can be viewed as a correspondence concept for eigenvalues. We shortly investigate this theory and use it to describe the Karhunen-Loéve expansion.

In the finite dimensional case, we know that every $n \times n$ matrix A can be viewed as a linear operator on the Euclidean space R^n. For any $\lambda \in R$, one and only one of the following statements holds:

(1) $(\lambda I - A)$ is invertible. This means $(\lambda I - A)^{-1}$ exists as a matrix. Here I is the identity matrix.

(2) λ is an eigenvalue of A. This means $\exists v_0 \in R^n$, s.t. $Av_0 = \lambda v_0$.

However, when we consider the infinite dimensional case in functional analysis, things becomes much more complicated. Let $\mathcal{A}$ be a linear operator on a Hilbert space $\mathcal{X}$. (To be rigorous, $\mathcal{A}$ should be not only linear, but also bounded. It is easy to show that the operator $\mathcal{K}$ defined above is bounded, so there is no problem.) For any $\lambda \in R$, one and only one of the following statements holds

(1) $(\lambda I - \mathcal{A})$ is invertible. This means $(\lambda I - \mathcal{A})^{-1}$ is a well-defined operator on $\mathcal{X}$;

(2) $(\lambda I - \mathcal{A})$ is not one-to-one;

(3) $(\lambda I - \mathcal{A})$ is a one-to-one mapping of $\mathcal{X}$ onto a dense proper subset of $\mathcal{X}$, that is, $\overline{(\lambda I - \mathcal{A})} = \mathcal{X}$;

(4) λ does not satisfy any of (1) to (3), that is, $(\lambda I - \mathcal{A})$ is one-to-one, but $\overline{(\lambda I - \mathcal{A})} \neq \mathcal{X}$.

The set of λ for which (2) or (3) or (4) hold are called the spectrum of $\mathcal{A}$, usually denoted by $\sigma(\mathcal{A})$. $\sigma(\mathcal{A})$ can be divided into three disjoint sets: the point spectrum, the continuous spectrum and the residual spectrum, corresponding to λ satisfy (2), (3) and (4).

The Rieze-Schauder theory tells us that, if $\mathcal{A}$ is a compact operator, then it has only countably many point spectrum (no continuous and residual spectrum) except for 0. Hence for the spectrum of a compact operator $\mathcal{A}$, there are only three possibilities:

(1) $\sigma(\mathcal{A}) = \{0\}$;

(2) $\sigma(\mathcal{A}) = \{0, \lambda_1, \lambda_2, \cdots, \lambda_n\}$; or

(3) $\sigma(\mathcal{A}) = \{\lambda_1, \lambda_2, \cdots, \lambda_n, \cdots, 0\}$, and $\lambda_n \longrightarrow 0$.

Moreover, if $\lambda \in \sigma(\mathcal{A})$ and $\lambda \neq 0$, then λ is an eigenvalue of $\mathcal{A}$ and the corresponding eigenspace is $ker(\lambda I - \mathcal{A})$. Each $v \in ker(\lambda I - \mathcal{A})$ (except $v = 0$) is an eigenvalue of $\mathcal{A}$, satisfying $\mathcal{A}v = \lambda v$.

We have mentioned that the operator $\mathcal{K}$ defined by the covariance function is compact, so now we can consider the eigenvalues of $\mathcal{K}$. From the discussions above, we know that there are at most countably many eigenvalues of $\mathcal{K}$. Rieze-Schauder theory also tells us that the invariant space for each eigenvalue λ of a compact operator is finitely many dimensional. (similar to the case in linear algebra, the dimension of this invariant space are called the multiple of λ.) We denote them by $\{\lambda_1, \lambda_2, \cdots, \lambda_n, \cdots\}$ and assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots$. Here

we count eigenvalues with multiples greater than one many times.

Moreover, since $\mathcal{K}$ is a positive operator, each eigenvalue λ_i will be nonnegative. This is easy to see. For any λ_i and a corresponding eigenvector $f \neq 0$, we have

$$0 \leq \langle \mathcal{K}(f), f \rangle = \lambda_i \langle f, f \rangle.$$

We know that $\langle f, f \rangle > 0$ since $f \neq 0$, so that $\lambda_i \geq 0$.

Note that $\mathcal{K}$ is also symmetric, by the Hilbert-Schmit Theorem in functional analysis, we know that for $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq \cdots \geq 0$, there is a corresponding orthonormal basis $\{f_i\}$ for $L^2(A)$, such that

$$\mathcal{K}(f_i) = \lambda_i f_i,$$

that is, f_i is the corresponding eigenvetor for λ_i. Note that it is possible that linearly independent f_i's are corresponding to the same λ_i (multiple greater than one case). Moreover, for each $v \in L^2(A)$, we have

$$v = \sum_{i=1}^{\infty} \langle v, f_i \rangle f_i,$$

$$\mathcal{K}(v) = \sum_{i=1}^{\infty} \lambda_i \langle v, f_i \rangle f_i.$$

This representation is the mathematical foundation of Karhunen-Loéve decomposition.

Remember that for any fixed $\omega_0 \in \Omega$, $U(\omega_0, x)$ is an element in the Hilbert space $L^2(A)$. We can expand it with respect to the orthonormal basis $\{f_i\}$

$$U(\omega_0, x) = \sum_{i=1}^{\infty} \langle U(\omega_0, x), f_i(x) \rangle f_i.$$

Therefore we have generally an expansion for the random surface $U(\omega, x)$:

$$U(\omega, x) = \sum_{i=1}^{\infty} \langle U(\omega, x), f_i(x) \rangle f_i.$$

Now we define

$$U_i = \langle U, f_i \rangle = \int_A U(x) f_i(x) dx, \quad i = 1, 2, 3, \cdots.$$

Easy to see that $U_i \in L^2(\Omega)$ and we can write $U(x) = \sum_{i=1}^{\infty} U_i f_i(x)$. This is the Karhunen-Loéve decomposition of the random surface $U(x)$.

This expansion has several nice properties:

1. In this expansion, the randomness is separated from dependence of the parameter x. Note that each term of the expansion is a product of a random variable and a deterministic function (surface). This gives a factor model representation (with infinitely many factors) of the random surface $U(x)$ and we can truncate the series to get an N-factor model.

2. The eigenmodes f_i's are orthonormal. This means:

$$\langle f_i, f_j \rangle = \int_A f_i(x) f_j(x) dx = 0, \quad \forall i \neq j$$

$$\| f_i \| = \int_A f_i^2(x) dx = 1, \quad \forall i.$$

3. The random coefficients U_i's have mean 0 and they are uncorrelated with each other. The mean 0 property is from the assumption $E(U(x)) = 0$:

$$E(U_i) = E(\int_A U(x) f_i(x) dx) = \int_A E(U(x)) f_i(x) dx = 0, \quad \forall i.$$

The U_i's are uncorrelated by the orthogonality of $\{f_i\}$:

$$\begin{aligned} Cov(U_i, U_j) &= E(U_i U_j) \\ &= \int_\Omega (\int_A U(\omega, x) f_i(x) dx \int_A U(\omega, y) f_j(y) dy) d\omega \\ &= \int_{A \times A} \int_\Omega U(\omega, x) U(\omega, x) d\omega f_i(x) f_j(y) dx dy \end{aligned}$$

$$
\begin{aligned}
&= \int_{A\times A} K(x,y) f_i(x) f_j(y) dxdy \\
&= \int_A \int_A K(y,x) f_i(x) dx f_j(y) dy \\
&= \langle \mathcal{K}(f_i), f_j \rangle \\
&= \lambda_i \langle f_i, f_j \rangle \\
&= 0, \quad \forall i \neq j.
\end{aligned}
$$

This property together with property 2, show that $(U_i \cdot f_i)$ are orthogonal both as random variables (elements in $L^2(\Omega)$) and as surfaces (elements in $L^2(A)$), which makes the Karhunen-Loéve decomposition very convenient for computational purposes.

4. As an infinite dimensional analogue of PCA, Karhunen-Loéve decomposition has the distinction of being the optimal orthogonal transformation for keeping the subspace that has largest "variance". So if we truncate the series from this decomposition and get an N-factor model, it captures greater variance structure of the random surface $U(x)$ than any other N-factor model with property 2. and 3. This property is again a consequence of a property of compact operator. From functional analysis, we know that for the compact positive operator $\mathcal{K}$, $\exists g_0 \in L^2(A)$, such that

$$\langle \mathcal{K}(g_0), g_0 \rangle = \sup_{\|f\|=1} \langle \mathcal{K}(f), f \rangle$$

and

$$\mathcal{K}(g_0) = \lambda_1 g_0, \quad \|g_0\| = 1.$$

Now we can choose $f_1 = g_0$ and we get

$$\lambda_1 = \sup_{\|f\|=1} \langle \mathcal{K}(f), f \rangle.$$

Similarly, we can repeat this procedure to find

$$\lambda_i = \sup_{\|f\|=1} \{\langle \mathcal{K}(f), f\rangle \mid f \perp \text{span}\{f_1, \cdots, f_{i-1}\}\}.$$

Also note that $\forall f \in L^2(A)$, we have

$$\begin{aligned}
Var(\langle U, f\rangle) &= E(\langle U, f\rangle^2) \\
&= \int_\Omega \left(\int_A U(\omega, x) f(x) dx \right)^2 d\omega \\
&= \int_\Omega (\int_A U(\omega, x) f(x) dx \int_A U(\omega, y) f(y) dy) d\omega \\
&= \int_{A\times A} \int_\Omega U(\omega, x) U(\omega, x) d\omega f(x) f(y) dx dy \\
&= \int_{A\times A} K(x, y) f(x) f(y) dx dy \\
&= \int_A \int_A K(y, x) f(x) dx f(y) dy \\
&= \langle \mathcal{K}(f), f\rangle.
\end{aligned}$$

Now we see that for every deterministic surface f, the projection of the random surface U on f will have variance $\langle \mathcal{K}(f), f\rangle$. This together with the property of λ_i

$$\lambda_i = \sup_{\|f\|=1} \{\langle \mathcal{K}(f), f\rangle \mid f \perp \text{span}\{f_1, \cdots, f_{i-1}\}\}$$

shows that Karhunen-Loéve decomposition captures the maximum variance structure of $U(x)$.

Now we briefly summarize the key idea of Karhunen-Loéve decomposition.

We can represent the random surface $U(\omega, x) \in L^2(\Omega \times A)$ as the infinite sum

$$U(x) = u(x) + \sum_{i=1}^{\infty} U_i f_i(x),$$

where $u(x) = E(U(x))$ and f_i's are the normalized eigenvectors for eigenvalues of $\mathcal{K}$, such that

$$\mathcal{K}(f_i) = \lambda_i f_i.$$

The operator $\mathcal{K}$ is defined by

$$\mathcal{K}(f)(x) = \int_A K(x,y)f(y)dy, \quad \forall f \in L^2(A)$$

where $K(x,y)$ is the covariance function

$$K(x,y) = Cov(U(x)U(y)) = E(U(x)U(y)) - E(U(x))E(U(y)).$$

The U_i's are orthogonal as random variables and f_i's are orthogonal as deterministic surfaces. This expansion captures the maximum variance structure of $U(x)$.

We can compute the variance of $U(x)$ at each $x \in L^2(A)$ as

$$\begin{aligned}
Var(U(x)) &= Var(u(x) + \sum_{i=1}^{\infty} U_i f_i(x)) \\
&= Var(\sum_{i=1}^{\infty} U_i f_i(x)) \\
&= \sum_{i=1}^{\infty} f_i^2 Var(U_i) + \sum_{i \neq j} f_i f_j Cov(U_i, U_j) \\
&= \sum_{i=1}^{\infty} f_i^2 Var(U_i) \\
&= \sum_{i=1}^{\infty} f_i^2 Var(\langle U, f_i \rangle) \\
&= \sum_{i=1}^{\infty} f_i^2 \langle \mathcal{K}(f_i), f_i \rangle \\
&= \sum_{i=1}^{\infty} f_i^2 \lambda_i \langle f_i, f_i \rangle \\
&= \sum_{i=1}^{\infty} \lambda_i f_i^2
\end{aligned}$$

and the total variance of $U(x)$ is

$$\int_A Var(U(x))dx = \int_A \sum_{i=1}^{\infty} \lambda_i f_i^2 dx$$

$$
\begin{aligned}
&= \sum_{i=1}^{\infty} \int_A \lambda_i f_i^2 dx \\
&= \sum_{i=1}^{\infty} \lambda_i \| f_i \| \\
&= \sum_{i=1}^{\infty} \lambda_i < +\infty.
\end{aligned}
$$

Here $\sum_{i=1}^{\infty} \lambda_i < +\infty$ means that $\mathcal{K}$ belongs the so-called trace class operators.

We can truncate this infinite series and get an N-factor model for $U(x)$

$$U(x) \approx u(x) + \sum_{i=1}^{N} U_i f_i(x).$$

The total variance of this N-factor model is

$$
\begin{aligned}
\int_A Var(u(x) + \sum_{i=1}^{N} U_i f_i(x)) dx &= \int_A Var(\sum_{i=1}^{N} U_i f_i(x)) \\
&= \sum_{i=1}^{N} \lambda_i \| f_i \| \\
&= \sum_{i=1}^{N} \lambda_i.
\end{aligned}
$$

The ratio $\sum_{i=1}^{N} \lambda_i / \sum_{i=1}^{\infty} \lambda_i$ can be interpreted as how much total variance is explained by this N-factor model. If this ratio is close to 100%, this model is a good approximation.

2.3 Uncentered Karhunen-Loéve Decomposition

We investigated the Karhunen-Loéve decomposition for random surfaces in last section. In this section we develop a variant of it—-the uncentered version of this expansion.

Remember that in the Karhunen-Loéve decomposition, we first subtract the mean from the random surface $U(x)$ and then expanded $U(x) - E(U(x))$ with respect to the basis $\{f_i\}$. The main reason for the centering is that we want the function $K(x, y)$ used for defining the operator $\mathcal{K}$ to be equal to the variance:

$$K(x, y) = E(U(x)U(y)) = \int_{\Omega} U(\omega, x)U(\omega, y)d\omega.$$

Then the variance of the projection of $U(x)$ on any element in $L^2(A)$ can be represented as

$$Var(\langle U, f\rangle) = \langle \mathcal{K}(f), f\rangle$$

and we can detect how much total variance can be explained by the first N factors where we truncate the infinite expansion.

We should notice that one price for the centering is that some important features, for example, the shape of the original random surface $U(x)$ will not be captured by the basis functions (eigenmodes) in the truncated model. This is not an issue for most problems where Karhunen-Loéve decomposition is used. In most problems the random object (random vector, stochastic process, random surface, or random field) to be analyzed has no shape constraint, hence in these problems the shape of truncated model does not matter and there is no need for investigating the shape of the eigenmodes.

However, for the purpose of option pricing and hedging, the shapes of the random surface $U(x)$ and the its model are not arbitrary. There are no static arbitrage conditions for the call price surface $U(x)$ and these conditions are nothing but shape constraints for $U(x)$. These shape constraints should be satisfied in the factor model for the call price surface. Otherwise the existence of statics arbitrage in the model may lead to unreasonable prices and imperfect hedging for exotic options.

In this section we investigate what happens if we give up the centering (hence the explanation of total variance) in the Karhunen-Loéve decomposition. We still use notations in last section.

Consider the random surface $U(\omega, x) \in L^(\Omega \times A)$ and the function

$$M(x,y) = E(U(x)U(y)) = \int_\Omega U(\omega,x)U(\omega,y)d\omega.$$

Now $M(x,y)$ may not be the covariance function anymore since the mean of $U(x)$ may not be zero. But it still defines a linear operator $\mathcal{M}$ on $L^2(A)$ by

$$\mathcal{M}(f)(x) = \int_A M(x,y)f(y)dy, \quad \forall f \in L^2(A).$$

$M(x,y)$ is symmetric and $\mathcal{M}$ is a symmetric positive compact operator. The argument is the same as the case for $\mathcal{K}$

$$\begin{aligned}
\langle f, \mathcal{M}(g)\rangle &= \int_A f(x)\mathcal{M}(g)(x)dx \\
&= \int_A f(x)\int_A M(x,y)g(y)dydx \\
&= \int_{A\times A} M(x,y)f(x)g(y)dxdy \\
&= \int_A\int_A M(y,x)f(x)dxg(y)dy \\
&= \langle \mathcal{M}(f), g\rangle
\end{aligned}$$

and

$$\begin{aligned}
\langle f, \mathcal{M}(f)\rangle &= \int_A f(x)\mathcal{M}(f)(x)dx \\
&= \int_{A\times A} M(x,y)f(x)f(y)dxdy \\
&= \int_{A\times A} f(x)f(y)\int_\Omega U(\omega,x)U(\omega,y)d\omega dxdy \\
&= \int_\Omega\int_A U(\omega,x)f(x)dx\cdot\int_A U(\omega,y)f(y)dyd\omega \\
&= \int_\Omega(\int_A U(\omega,x)f(x)dx)^2 d\omega \\
&\geq 0.
\end{aligned}$$

The compactness again depends on some functional analysis argument and we do not give the details of the proof here. We refer to [64, 76] for the rigorous proof.

Therefore $\mathcal{M}$ has at most countably many positive eigenvalues $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_n \geq \cdots \geq 0$. Again, by Hilbert-Schmidt Theorem, we choose a orthonormal basis $\{g_i\}$ for $L^2(A)$, s.t. $\forall v \in L^2(A)$, we have

$$\mathcal{M}(g_i) = \eta_i g_i, \quad \forall i$$

$$v = \sum_{i=1}^{\infty} \langle v, g_i \rangle g_i$$

$$\mathcal{M}(v) = \sum_{i=1}^{\infty} \eta_i \langle v, g_i \rangle g_i$$

and moreover,

$$\eta_i = \sup_{\|f\|=1} \{\langle \mathcal{M}(f), f \rangle \mid f \perp \text{span}\{g_1, \cdots, g_{i-1}\}\}.$$

We can expand $U(x)$ with respect to $\{g_i\}$ as

$$U(x) = \sum_{i=1}^{\infty} V_i g_i(x), \tag{2.1}$$

where V_i is defined by

$$V_i = \langle U, g_i \rangle = \int_A U(x) g_i(x) dx, \quad i = 1, 2, 3, \cdots.$$

(2.1) can be called uncentered Karhunen-Loéve decomposition. In this expansion, the random coefficients (factors) may not be uncorrelated anymore, since now we have

$$E(V_i V_j) = \int_\Omega \left(\int_A U(\omega, x) g_i(x) dx \int_A U(\omega, y) g_j(y) dy \right) d\omega = \eta_i \langle g_i, g_j \rangle = 0, \quad \forall i \neq j,$$

although $E(V_i V_j)$ may not be the covariance anymore. However, the eigenmodes g_i's are still orthogonal to each other.

The variance of $\langle U, f\rangle$ may not be equal to $\langle \mathcal{M}(f), f\rangle$, we only have

$$E(\langle U, f\rangle^2) = \langle \mathcal{M}(f), f\rangle.$$

Similarly, we now have

$$\begin{aligned}
E(U^2(x)) &= E[(\sum_{i=1}^{\infty} V_i g_i(x))^2] \\
&= \sum_{i=1}^{\infty} g_i^2 E(V_i^2) + \sum_{i\neq j} g_i g_j E(V_i V_j) \\
&= \sum_{i=1}^{\infty} g_i^2 E(V_i^2) \\
&= \sum_{i=1}^{\infty} g_i^2 E(\langle U, g_i\rangle^2) \\
&= \sum_{i=1}^{\infty} g_i^2 \langle \mathcal{M}(g_i), g_i\rangle \\
&= \sum_{i=1}^{\infty} g_i^2 \eta_i \langle g_i, g_i\rangle \\
&= \sum_{i=1}^{\infty} \eta_i g_i^2
\end{aligned}$$

and

$$\begin{aligned}
\int_A E(U^2(x))dx &= \int_A \sum_{i=1}^{\infty} \eta_i g_i^2 dx \\
&= \sum_{i=1}^{\infty} \int_A \eta_i g_i^2 dx \\
&= \sum_{i=1}^{\infty} \eta_i \|g_i\| \\
&= \sum_{i=1}^{\infty} \eta_i.
\end{aligned}$$

Again we can truncate this infinite series and get an N-factor model for $U(x)$:

$$U(x) \approx \sum_{i=1}^{N} V_i g_i(x).$$

Therefore it follows that

$$\int_A E[(\sum_{i=1}^{N} V_i g_i(x))^2]dx = \sum_{i=1}^{N} \eta_i.$$

It is easy to see that in the uncentered Karhunen-Loéve expansion, the ratio $\sum_{i=1}^{N} \eta_i / \sum_{i=1}^{\infty} \eta_i$ cannot be interpreted as how much total variance is explained by the first N factors anymore.

However, while unable to interpret variance by its factors, the uncentered Karhunen-Loéve decomposition has another elegant interpretation. The truncated uncentered Karhunen-Loéve expansion was the best approximation of the original random surface $U(x)$ in the sense that it reduces the total mean-square error resulting of its truncation. Because of this property, it is often said that the Karhunen-Loéve expansion optimally compacts the energy.

More specifically, given any orthonormal basis $\{h_i\}$, we may expand $U(x)$ as:

$$U(x) = \sum_{i=1}^{\infty} X_i h_i(x),$$

where

$$X_i = \langle U, h_i \rangle = \int_A U(x) h_i(x) dx, \quad i = 1, 2, 3, \cdots$$

and we may approximate $U(x)$ by the truncated finite sum

$$U(x) \approx \sum_{i=1}^{N} X_i h_i(x)$$

for some positive integer N. We claim that of all such approximations, the approximation given by the truncated uncentered Karhunen-Loéve decomposition is the one that minimizes the total mean square error (provided we have

arranged the eigenvalues in decreasing order). Here the total mean square error is the distance in the Hilbert space $L^2(\Omega \times A)$, which is defined by

$$\|Z_1 - Z_2\| = \int_A E[(Z_1(x) - Z_2(x))^2]dx = \int_{\Omega \times A} (Z_1(\omega, x) - Z_2(\omega, x))^2 d\omega dx$$

for any elements Z_1 and Z_2 in $L^2(\Omega \times A)$.

Now we show why this claim is true. First, for any such $\{h_i\}$ and $\{X_i\}$, we have

$$\langle h_i, h_j \rangle = \int_A h_i(x)h_j(x)dx = 0, \quad \forall i \neq j$$

and hence

$$E(X_i X_j) = \int_\Omega (\int_A U(\omega, x)h_i(x)dx \int_A U(\omega, y)h_j(y)dy)d\omega = \eta_i \langle h_i, h_j \rangle = 0, \quad \forall i \neq j.$$

Also we have

$$E(X_i^2) = E(\langle U, h_i \rangle^2) = \langle \mathcal{M}(h_i), h_i \rangle.$$

Moreover,

$$\begin{aligned} E(U^2(x)) &= E[(\sum_{i=1}^{\infty} X_i h_i(x))^2] \\ &= \sum_{i=1}^{\infty} h_i^2 E(X_i^2) + \sum_{i \neq j} h_i h_j E(X_i X_j) \\ &= \sum_{i=1}^{\infty} h_i^2 E(X_i^2) \\ &= \sum_{i=1}^{\infty} h_i^2 E(\langle U, h_i \rangle^2) \\ &= \sum_{i=1}^{\infty} h_i^2 \langle \mathcal{M}(h_i), h_i \rangle \end{aligned}$$

and hence

$$\int_A E(U^2(x))dx = \int_A \sum_{i=1}^{\infty} \langle \mathcal{M}(h_i), h_i \rangle h_i^2 dx = \sum_{i=1}^{\infty} \langle \mathcal{M}(h_i), h_i \rangle.$$

Similarly, for the difference $U(x) - \sum_{i=1}^{N} X_i h_i(x) = \sum_{i=N+1}^{\infty} X_i h_i(x)$, we have

$$\begin{aligned} E[(\sum_{i=N+1}^{\infty} X_i h_i(x))^2] &= \sum_{i=N+1}^{\infty} h_i^2 E(X_i^2) + \sum_{i,j>N, i\neq j} h_i h_j E(X_i X_j) \\ &= \sum_{i=N+1}^{\infty} h_i^2 E(X_i^2) \\ &= \sum_{i=N+1}^{\infty} h_i^2 E(\langle U, h_i\rangle^2) \\ &= \sum_{i=N+1}^{\infty} h_i^2 \langle \mathcal{M}(h_i), h_i\rangle \end{aligned}$$

and hence

$$\int_A E[(U(x) - \sum_{i=1}^{N} X_i h_i(x))^2]dx = \int_A \sum_{i=N+1}^{\infty} \langle \mathcal{M}(h_i), h_i\rangle h_i^2 dx = \sum_{i=N+1}^{\infty} \langle \mathcal{M}(h_i), h_i\rangle.$$

Remember that $\{g_i\}$ and $\{V_i\}$ have the nice property

$$\eta_i = \sup_{\|f\|=1} \{\langle \mathcal{M}(f), f\rangle \mid f \perp \text{span}\{g_1, \cdots, g_{i-1}\}\},$$

we have

$$\sum_{i=1}^{N} \langle \mathcal{M}(h_i), h_i\rangle \leq \sum_{i=1}^{N} \eta_i = \sum_{i=1}^{N} \langle \mathcal{M}(g_i), g_i\rangle.$$

Note that

$$\int_A E[(U(x) - \sum_{i=1}^{N} X_i h_i(x))^2]dx = \int_A E(U^2(x))dx - \sum_{i=1}^{N} \langle \mathcal{M}(h_i), h_i\rangle$$

and

$$\int_A E[(U(x) - \sum_{i=1}^{N} V_i g_i(x))^2]dx = \int_A E(U^2(x))dx - \sum_{i=1}^{N} \langle \mathcal{M}(g_i), g_i\rangle.$$

Therefore it follows that

$$\int_A E[(U(x) - \sum_{i=1}^{N} V_i g_i(x))^2]dx \leq \int_A E[(U(x) - \sum_{i=1}^{N} X_i h_i(x))^2]dx.$$

We have shown that the uncentered Karhunen-Loéve decomposition can be viewed from another perspective. That is, the truncated (first N-term) uncentered Karhunen-Loéve expansion is the solution of a optimization problem over

an infinite dimensional space $L^2(\Omega \times A)$:

$$\min \int_A E[(Y(x) - U(x))^2]dx, \text{ such that } Y(x) \text{ has "rank" } N.$$

Here $U(x)$ is the original random surface which we want to model and "rank" N means that $Y(\omega, x)$ can be represented as $Y(\omega, x) = \sum_{i=1}^{N} X_i(\omega)h_i(x)$. Even though h_i here may not be a part of an orthonormal basis, we can always orthogonalize and normalize it and then expand it to an orthonormal basis of $L^2(A)$.

We have presented the mathematical framework of centered and uncentered Karhunen-Loéve decomposition. As mentioned, to apply this methodology to our problem, we need to incorporate some shape constraints (the no static arbitrage conditions). We will develop these ideas in the next chapter.

CHAPTER 3

STATIC ARBITRAGE AND SHAPE CONSTRAINTS

3.1 Static Arbitrage Conditions for Call Price Surface

When building models for the European call price surface (or the Black-Scholes implied volatility) for exotic option pricing, one main challenge is to guarantee that these models are arbitrage-free. An arbitrage opportunity is a costless trading strategy which at some future time provides a positive profit with positive probability, but has no possibility of a loss. When we investigate the whole surface of the European call price, the prices of call options with different strikes and maturities have to satisfy some relations, otherwise there will be arbitrage opportunity no matter what model is assumed. These conditions are called static arbitrage conditions (or no static arbitrage conditions). We use the term "static" here since the this kind of arbitrage opportunity (if exists) are model independent. However, Carr, Gémen, Madan and Yor [17] use of the term "static arbitrage" has a more delicate meaning and we do not consider it in this thesis. In this section we investigate the static arbitrage conditions in details. A technical note here is that we assume, for simplicity, the interest rate is zero in this thesis.

Recall that a European call option is a contract that gives the owner the right to buy one unit of a risky asset (underlying) at an expiration time (maturity) T at a strike price K agreed upon today (time 0). We use $C(K,T)$ to denote the price of the call option with strike K and maturity T. When we consider the whole surface $C(K,T)$, $(K,T) \in [0,+\infty] \times [0,+\infty]$, certain conditions have to be satisfied if we want to rule out several arbitrage opportunities.

Define calendar spread (also called a time spread or horizontal spread) as,

$$C(K, T_1) \leq C(K, T_2), \ \forall\, T_1 \leq T_2,\ K.$$

This follows from Jensen's inequality if we view $C(K, T)$ as the conditional expectation of the future payoff. However, here we give a model independent argument since the conditional expectation argument may not work in some situations (specifically, in some local martingale models, as we will see below).

Suppose there exists $T_1 < T_2, K$, such that $C(K, T_1) > C(K, T_2)$. We can implement the following strategy and get an arbitrage. (We use S to denote the underlying asset). At time 0, short 1 share $C(K, T_1)$ and long 1 share $C(K, T_2)$. This gives us positive amount of money since $C(K, T_1) > C(K, T_2)$. At T_1, there are 2 scenarios.

Scenario (1): $S_{T_1} \leq K$. In this case, the call option with maturity T_1 is out of money and our short position costs us 0. The long position on the call option with maturity T_2 will always have nonnegative value. So we get positive profit.

Scenario (2): $S_{T_1} > K$. In this case, we can borrow 1 share underlying from the market and change it with the buyer of the option for K dollars. Then at time T_2, we can use this K dollar to buy 1 share underlying (we can do this since we hold the call option with strike K and maturity T_2) and return it to the market. So we also get positive profit.

Calendar spread shows that the European call price is increasing with respect to the maturity T.

We should mention here that in some research papers (for example, [21, 61]), it is shown that in some models in which the underlying is a strict local martingale instead of a martingale under the risk-neutral measures, $C(K, T_1) > C(K, T_2)$

is possible. According to the first fundamental theorem of asset pricing [23, 39], if the underlying asset price is a local martingale, then the model should be arbitrage-free. This seems to contradict the calendar spread at the first look. However, after careful investigation, we find that the arbitrage-free here means that admissible arbitrage opportunity does not exist. Any admissible arbitrage strategy have to be self-financing and bounded below (since we want to avoid infinite loss). We can see that if the underlying asset dynamics follows a strict local martingale, then the strategy we carry in the argument above may not be bounded below and hence is not admissible. There is no contradiction.

Next, define vertical spread as

$$-1 \leq \frac{C(K_2, T) - C(K_1, T)}{K_2 - K_1} \leq 0, \ \forall\, T,\ K_1 \neq K_2.$$

Again we show this by construct arbitrage strategy in the case that this condition is violated.

(1) Suppose there exists T, $K_1 \leq K_2$, such that $\frac{C(K_2,T)-C(K_1,T)}{K_2-K-1} > 0$. Assume $K_2 > K_1$, this gives $C(K_2, T) - C(K_1, T) > 0$. At time 0, we long 1 share $C(K_1, T)$ and short 1 share $C(K_2, T)$. This gives us positive amount of money. At the maturity T, the value of our portfolio is $(S_T - K_1)^+ - (S_T - K_2)^+ \geq 0$. So we get positive profit. The case that $K_2 < K_1$ is the same.

(2) Suppose there exists T, $K_1 \leq K_2$, such that $\frac{C(K_2,T)-C(K_1,T)}{K_2-K_1} < -1$. Assume $K_2 > K_1$, this gives $C(K_1, T) - C(K_2, T) > K_2 - K_1 > 0$. At time 0, we long 1 share $C(K_2, T)$ and short 1 share $C(K_1, T)$. This gives us positive amount of money greater than $K_2 - K_1$. At the maturity T, the value of our portfolio is $(S_T - K_2)^+ - (S_T - K_1)^+ \geq K_1 - K_2$. We have $[(S_T - K_2)^+ - (S_T - K_1)^+] + (K_2 - K_1) \geq 0$. This means we get positive profit. The case that $K_2 < K_1$ is the same.

Vertical spread shows that the European call price is decreasing with respect to the strike K and the decreasing rate cannot be too high.

Lastly, define butterfly spread as

$$C(\frac{K_1 + K_2}{2}, T) \leq \frac{1}{2}[C(K_1, T) + C(K_2, T)], \ \forall T, K_1, K_2.$$

This follow from the convexity of the payoff function of European call option.

Suppose there exists T, $K_1 \leq K_2$, such that $C(\frac{K_1+K_2}{2}, T) > \frac{1}{2}[C(K_1, T)+C(K_2, T)]$. At time 0, we long $\frac{1}{2}$ share $C(K_1, T)$ and $\frac{1}{2}$ share $C(K_2, T)$ and short 1 share $C(\frac{K_1+K_2}{2}, T)$. This gives us positive amount of money. At the maturity T, the value of our portfolio is $(S_T - \frac{K_1+K_2}{2})^+ - \frac{1}{2}[(S_T - K_1)^+ + (S_T - K_2)^+] > 0$. So we get positive profit.

Butterfly spread shows that the European call price is convex with respect to the strike K.

Besides the three spread conditions above, we still need some other conditions to exclude static arbitrage. One obvious condition is

$$C(K, T) \geq 0, \ \forall K, T$$

and boundary conditions are

$$C(0, T) = S, \ \forall T,$$

$$C(K, 0) = (S - K)^+, \ \forall K.$$

Other than static arbitrage conditions, a further economically reasonable requirement is that

$$\lim_{K \to +\infty} C(K, T) = 0, \ \forall T,$$

which gives another boundary condition.

We end this section by a result for the other direction of the above conditions. Fora given function $C(K,T) \geq 0$, if it satisfies

1. $C(K,T)$ is increasing with respect to T for any K,

2. $C(K,T)$ is decreasing and convex with respect to K for any T, and

3. $\lim_{K\to+\infty} C(K,T) = 0$ for any T and $C(0,T) = S_0$.

Under these conditions there exists a martingale S_t, such that $C(K,T) = E[(S_T - K)^+]$, for all K,T. This means that $C(K,T)$ is the price of the European call option with strike K and maturity T, written on the underlying asset S_t. The proof of this result depends on very deep theorems and very complicated argument in analysis. Instead of giving the proof here, we refer to [25, 50, 73].

3.2 Shape Constraints from Static Arbitrage Conditions

We do want to exclude static arbitrage in our model for the call price surface. Therefore some shape constraints are put on the call price surface by the static arbitrage conditions. If we further assume some smoothness of the call price surface $C(K,T)$, these constraints can be written as

1. $\dfrac{\partial C(K,T)}{\partial T} \geq 0,$
2. $\dfrac{\partial^2 C(K,T)}{\partial K^2} \geq 0,$
3. $-1 \leq \dfrac{\partial C(K,T)}{\partial K} \leq 0,$ **and**
4. $C(0,T) = S,\ C(K,0) = (S-K)^+,\ \lim\limits_{K\to+\infty} C(K,T) = 0,\ \forall\, K,T.$

Constraints 1, 2 and 3 are the shape constraints and constraint 4 is the boundary conditions.

These constraints are necessary conditions for a model of call price surface to have no static arbitrage opportunity. What we care about more is the converse, that is, the sufficient condition for a model without static arbitrage. According to Gyöngy-Dupire theory [27, 37] for local volatility, if we have a surface $C(K,T)$ satisfying constraints 1, 2 and 4 above and some relax regularity conditions, then we can define the local volatility:

$$\sigma(K,T) = \frac{1}{K}\sqrt{\frac{2\frac{\partial C(K,T)}{\partial T}}{\frac{\partial^2 C(K,T)}{\partial K^2}}},$$

(note that here constraints 1 and 2 can guarantee $\sigma(K,T)$ is well-defined) and set

$$d\hat{S}_t = \sigma(\hat{S}_t, t)\hat{S}_t dB_t$$

and then $C(K,T)$ can be viewed as the price of a European call option written on underlying asset S_t with strike K and maturity S, where S_t has the same marginal distribution as $\hat{S}_t$. This coincides with the result in [12] that the 1-dimensional marginal density (under risk-neutral measure) of the underlying asset S_t is fully determined by the call price surface $C(K,T)$

$$f_{S_T}(s) = \left.\frac{\partial^2 C(K,T)}{\partial K^2}\right|_{K=s}.$$

Actually after studying the 4 constraints above, we find that constraint 3 can be implied from constraints 1, 2 and 4.

Proposition 1 *Let $C(K,T)$ be a twice differentiable function defined on $[0,+\infty)\times[0,+\infty)$. Suppose $C(K,T)$ satisfies $\frac{\partial C(K,T)}{\partial T} \geq 0$, $\frac{\partial^2 C(K,T)}{\partial K^2} \geq 0$ and $C(0,T) = S$, $C(K,0) = (S-K)^+$, $\lim_{K\to+\infty} C(K,T) = 0$, $\forall\, K,T$ for some constant $S \geq 0$. Then $C(K,T)$ must have satisfy $-1 \leq \frac{\partial C(K,T)}{\partial K} \leq 0$.*

Proof. First, we claim that

$$\frac{\partial C(K,T)}{\partial K} \geq -1.$$

From constraint 2, we know that $\frac{\partial C(K,T)}{\partial K}$ is increasing. So we just need to show that

$$\left.\frac{\partial C(K,T)}{\partial K}\right|_{K=0} \geq -1.$$

Note that $C(K,T)$ is increasing with respect to T, then

$$\frac{C(K,T)-C(0,T)}{K-0} = \frac{C(K,T)-S}{K} \geq \frac{(S-K)^+ - S}{K} \geq -1, \ \forall K > 0.$$

Hence

$$\left.\frac{\partial C(K,T)}{\partial K}\right|_{K=0} = \lim_{K\to 0+} \frac{C(K,T)-C(0,T)}{K-0} \geq -1.$$

Then, we claim that

$$\frac{\partial C(K,T)}{\partial K} \leq 0.$$

Suppose there $\exists K_0$, s.t.

$$\left.\frac{\partial C(K,T)}{\partial K}\right|_{K=K_0} > 0.$$

By constraint 2, we have

$$\frac{\partial C(K,T)}{\partial K} > 0, \ \forall K \geq K_0.$$

This contradicts the condition in constraint 4 that

$$\lim_{K\to+\infty} C(K,T) = 0.$$

□

The goal becomes to build factor models for the call price surface $C(K,T)$ satisfying shape constraints

1. $\frac{\partial C(K,T)}{\partial T} \geq 0$;

2. $\dfrac{\partial^2 C(K,T)}{\partial K^2} \geq 0$;

3. $C(0,T) = S,\ C(K,0) = (S-K)^+,\ \lim\limits_{K\to+\infty} C(K,T) = 0,\ \forall\, K,T.$

Now the reason why we choose call price surface instead of Black-Scholes implied volatility to model is obvious. By some simple computations, we see that if we choose implied volatility, the shape constraints 1 and 2 become

$$\frac{S_t\sqrt{T-t}e^{-\frac{d_1^2}{2}}}{\sqrt{2\pi}}\frac{\partial\sigma_t^{BS}}{\partial T} + \frac{\sigma_t^{BS}S_t}{2\sqrt{2\pi(T-t)}} \geq 0$$

and

$$\frac{S_t\sqrt{T-t}e^{-\frac{d_1^2}{2}}}{\sqrt{2\pi}}\frac{\partial^2\sigma_t^{BS}}{\partial K^2} + \frac{S_t d_1 e^{-\frac{d_1^2}{2}}}{\sqrt{2\pi}K\sigma_t^{BS}}\frac{\partial\sigma_t^{BS}}{\partial K} + \frac{S_t e^{-\frac{d_1^2}{2}}}{\sqrt{2\pi(T-t)}K^2\sigma_t^{BS}} \geq 0,$$

where $\sigma_t^{BS}$is the Black-Scholes implied volatility computed by inverting the Black-Scholes formula for call price.

These shape constraints are much more complicated and difficult to deal with than in the call price surface case.

We look at in the next section how to incorporate these shape constraints with the centered and uncentered Karhunen-Loéve decomposition.

3.3 Incorporating Shape Constraints with Karhunen-Loéve Decomposition

Following the notations in chapter 2, we use $U(x)$ to denote the call price surface. Here $x = (x_1, x_2) = (K,T) \in [a,b]\times[c,d] = A \subset R^2$. Here we assume $(K,T) \in [a,b]\times[c,d] = A$ instead of $[0,+\infty]\times[0,+\infty]$, because we can only observe call

price with $(K,T) \in \{(K_i,T_j),\ i=1,2,\cdots M,\ j=1,2,\cdots N\} \subset [a,b]\times[c,d]$. Also to apply the results in Chapter 2, $K(x,y)$ in section 2.1 and $M(x,y)$ in Section 2.2 must be defined over bounded range to make the operators $\mathcal{K}$ and $\mathcal{M}$ compact.

Now we have the call price surface $U(x) = U(\omega, x_1, x_2) \in L^2(\Omega \times A)$. However we still need some smoothness for $U(x)$ so that we can represent the no static arbitrage conditions into the shape constraints in Section 3.2. This is to say that $U(x)$ has to be at least twice differentiable. It is tempting to require that $U(x) \in L^2(\Omega) \otimes (L^2(A) \cap C^2(A))$. However this is not correct since $L^2(A) \cap C^2(A)$ is not a well-defined Hilbert space. Strictly speaking, we need to require that $U(x)$ is an element in the Sobolev space $L^2(\Omega) \otimes H^{2,2}(A)$ and the derivative is taken in the Schwartz-Sobolev sense instead of the usual sense. The details requires much further knowledge in functional analysis and very complicated. To make things easy and clear, we simply assume that $U(x) = U(\omega, x_1, x_2) \in L^2(\Omega \times A)$ and it has second order derivative with respect to x_1 and first order derivative with respect to x_2. This will not affect the procedure and the results. We refer the readers interested in these functional analysis details to [64, 76].

In the centered Karhunen-Loéve expansion, $U(x)$ is represented as

$$U(x) = u(x) + \sum_{i=1}^{\infty} U_i f_i(x)$$

and we truncate this infinite series to get an N-factor model

$$U(x) \approx u(x) + \sum_{i=1}^{N} U_i f_i(x),$$

where $u(x) = E(U(x))$, U_i's are random variables and f_i's are orthogonal normalized functions.

From the discussions in Section 3.2, we know that $U(x_1, x_2)$ satisfies shape

constraints

$$\frac{\partial^2 U}{\partial x_1^2} \geq 0, \quad \frac{\partial U}{\partial x_2} \geq 0.$$

However, these constraints might be violated by the N-factor model $u(x) + \sum_{i=1}^{N} U_i f_i(x)$ since it is just an approximation to the original $U(x)$ with difference $\sum_{i=N+1}^{\infty} U_i f_i(x)$.

It is easy to see that the center $u(x)$ still satisfy the shape constraints

$$\frac{\partial^2 u}{\partial x_1^2} = \frac{\partial^2}{\partial x_1^2} E(U(x_1, x_2)) = E\left[\frac{\partial^2 U(x_1, x_2)}{\partial x_1^2}\right] \geq 0,$$

$$\frac{\partial u}{\partial x_2} = \frac{\partial}{\partial x_2} E(U(x_1, x_2)) = E\left[\frac{\partial U(x_1, x_2)}{\partial x_2}\right] \geq 0.$$

However, the eigenmodes $f_i(x_1, x_2)$ generally does not satisfy these constraints. We need to modify this N-factor model from Karhunen-Loéve expansion to exclude the static arbitrage opportunity. This is done by modifying the eigenmodes. We first investigate each f_i, $i = 1, 2, \cdots, N$ to see if they satisfy the shape constraints. For those violating the constraints, we look at how they violate them and use modified eigenmodes to replace them. These modified eigenfunctions are computed by constrained spline functions. We discuss this in details on the simulated data set in the next chapter.

In the uncentered Karhunen-Loéve expansion, things are similar. $U(x)$ is represented as

$$U(x) = \sum_{i=1}^{\infty} V_i g_i(x)$$

and we truncate this infinite series to get an N-factor model

$$U(x) \approx \sum_{i=1}^{N} V_i g_i(x).$$

We do not have the $u(x)$ term as in the centered case. However, we can show that the first eigennmode g_1 satisfies the shape constraints.

First, from the construction of g_1 we know that $g_1(x) \geq 0$, $\forall x \in A$ or $g_1(x) \leq 0$, $\forall x \in A$. Otherwise g_1 will not correspond to the largest eigenvalue η_1 of $\mathcal{M}$. We omit the details for this step here since it involves the construction in the Hilbert-Schmidt procedure. Again we refer to [64, 76]. Hence we can choose $g_1 \geq 0$, otherwise we can simply take $-g_1$.

Note that $\mathcal{M}(g_1) = \eta_1 g_1$, so we have

$$\begin{aligned}
\frac{\partial g_1}{\partial x_1^2} &= \frac{1}{\eta_1}\frac{\partial \mathcal{M}(g_1)}{\partial x_1^2} \\
&= \frac{1}{\eta_1}\frac{\partial}{\partial x_1^2}\int_A M(x_1, x_2, y_1, y_2) g_1(y_1, y_2) dy_1 dy_2 \\
&= \frac{1}{\eta_1}\int_A \frac{\partial}{\partial x_1^2} M(x_1, x_2, y_1, y_2) g_1(y_1, y_2) dy_1 dy_2.
\end{aligned}$$

We know that

$$U(x_1, x_2) \geq 0, \ \frac{\partial^2 U}{\partial x_1^2} \geq 0.$$

Then we have

$$\begin{aligned}
\frac{\partial}{\partial x_1^2} M(x_1, x_2, y_1, y_2) &= \frac{\partial}{\partial x_1^2} E\left[U(x_1, x_2) U(y_1, y_2)\right] \\
&= E\left[\frac{\partial}{\partial x_1^2} U(x_1, x_2) U(y_1, y_2)\right] \\
&\geq 0.
\end{aligned}$$

This together with $g_1(x_1, x_2) \geq 0$ and $\eta_1 \geq 0$ give

$$\frac{\partial g_1}{\partial x_1^2} = \frac{1}{\eta_1}\int_A \frac{\partial}{\partial x_1^2} M(x_1, x_2, y_1, y_2) g_1(y_1, y_2) dy_1 dy_2 \geq 0.$$

The constraint $\frac{\partial g_1}{\partial x_2} \geq 0$ can be proved similarly.

Also notice that

$$V_1 = \langle U(x), g_1(x) \rangle = \int_A U(x) g_1(x) dx \geq 0.$$

We can see that if we use 1-factor model $U(x) \approx V_1 g_1(x)$ for the call price surface $U(x)$, then the shape constraints for no static arbitrage are satisfied. However, 1-factor model is hardly expected to have more capability than stochastic models there are 2 random factor drive the call price in most stochastic models, one from the Brownian motion drives the underlying and the other from the Brownian motion drives the spot volatility. Hence we should use N-factor model with $N \geq 3$. Then the N-factor model $\sum_{i=1}^{N} V_i g_i(x)$ needs to be modified since the eigenmodes $g_i, i \geq 2$ would not satisfy the shape constraints generally.

We can incorporate the shape constraints into the optimization formulation in section of uncentered Karhunen-Loéve decomposition and investigate this problem in a more geometric way. The N-factor model we want can be viewed as the solution of a constrained optimization problem over an (infinite dimensional) Hilbert space $L^2(\Omega \times A)$ (actually, more accurately, the Sobolev space $L^2(\Omega) \otimes H^{2,2}(A)$):

$$\begin{aligned} \min_{Y \in L^2(\Omega \times A)} \quad & \int_A E[(Y(x) - U(x))^2] dx \\ \text{s.t.} \quad & Y(x) \text{ has "rank" } N \\ & \tfrac{\partial^2 Y(\omega, x_1, x_2)}{\partial x_1^2} \geq 0, \quad \tfrac{\partial Y(\omega, x_1, x_2)}{\partial x_2} \geq 0. \end{aligned} \tag{3.1}$$

Consider the subset corresponding to the first constraint in (3.1): $\mathcal{R}_N = \{Y \in L^2(\Omega \times A) \,|\, Y \text{ has rank } N\} \subset L^2(\Omega \times A)$. If the first constraint is the only constraint, then the solution would be the projection of $U(x)$ on the subset $\mathcal{R}_N$. Although it is to see that $\mathcal{R}_N$ is not convex, we can still solve this problem by the Karhunen-Loéve procedure—-the solution is just the N-truncated expansion.

Consider the subset corresponding to the second constraint in (3.1): $C = \{Y \in L^2(\Omega \times A) \,|\, \frac{\partial^2 Y}{\partial x_1^2} \geq 0, \ \frac{\partial Y}{\partial x_2} \geq 0\} \subset L^2(\Omega \times A)$. If the first constraint is the only constraint,

then the solution would be the projection of $U(x)$ on the subset C. C is a closed and convex. While the convexity is obvious, the closeness need some technical argument in functional analysis, which we omit here.

When combining these two constraints together, the solution is the projection of $U(x)$ on the subset $\mathcal{R}_N \cap C$. It seems that the subset $\mathcal{R}_N \cap C$ does not have any good property. A heuristic approach is then the alternating projection. The idea is to start from $U(x)$ and iteratively project it first onto one subset (C or $\mathcal{R}_N$) and then onto the other. Since we have the Karhunen-Loéve procedure and the projection onto a closed convex subset is relatively easy, each step in this alternating projection procedure would not be difficult. However, the problem is that there is no guarantee for the convergence with this approach. Existing results show that under some regularity conditions, alternating projection method converges locally with a linear rate [53, 54]. But these results are for finite dimensional spaces / manifolds and need some smoothness / differential structures, which is not the case in our problem setup. Therefore we do not apply the alternating projection method but we use the same strategy as in the centered Karhunen-Loéve case, that is, modifying the eigenmodes g_i, $\forall i = 2, 3, \cdots, N$. More details about this is given in the next chapter.

CHAPTER 4

SOME NUMERICAL ANALYSIS

4.1 Introduction

In the last two chapters, we built the mathematical framework for the centered and uncentered Karhunen-Loéve decomposition and discussed the shape constraints that should be put on the eigenmodes in the model for all price surface. Now we investigate how this method works on a simulated data set, which might enlighten the application in practice. The whole procedure is:

1. Apply centered and uncentered Karhunen-Loéve decompositions to analyze a simulated data set for samples of European call option price with different strikes and maturities, finding the factors and the eigenmodes.

2. See how the eigenmodes violate the shape constraints, justifying our discussions in Chapter 3.

3. Use constrained spline method to modify these eigenmodes.

Details of the simulation procedure and the data set is given in Section 4.2. Then we discuss some issues about the implementation of Karhunen-Loéve decomposition in Section 4.3. Section 4.4 contains the details about how to modify the eigenmodes which violate the shape constraints. Finally, some numerical results are summarized in Section 4.5.

4.2 Description of Simulated Data

We simulate the samples of the underlying asset price and the European call price from the well-known Heston model [42]. The main advantage with this model is that there is a semi-analytical formula for the European call price if the underlying price dynamics follows Heston model, which is the main reason why this model is popular in financial industry. Here semi-analytical formula means that an explicit expression for the Fourier transform (or characteristic function) of the European call price with strike K and maturity T is available. Carr and Madan [18] propose a method to value call options with fast Fourier transform (FFT) when the characteristic function of the call price is known analytically. This method is much quicker than using a numerical integration especially when we deal with many options with different strikes and maturities simultaneously. We give a description (without proofs) of this method and apply it to simulate the call price data.

Recall that the underlying dynamics in Heston model is described by

$$\begin{aligned} dS_t &= \mu S_t dt + \sqrt{V_t} S_t dB_t^1, \\ dV_t &= \kappa(\theta - V_t)dt + \sigma\sqrt{V_t} dB_t^2, \\ \langle B_t^1, B_t^2 \rangle &= \rho_t. \end{aligned} \tag{4.1}$$

Using FFT, the price of European call option is given by

$$C(K,T) = \frac{\exp(-\alpha \log K)}{\pi} \int_0^{+\infty} \exp(-iv \log K) \frac{\psi(v - (\alpha+1)i, \kappa, \theta, \sigma, \rho, S, V)}{\alpha^2 + \alpha - v^2 + i(2\alpha+1)v} dv.$$

In this formula, i is the imaginary unit, α is a dampening parameter for FFT, $\psi(u, \kappa, \theta, \sigma, \rho, S, V)$ is a function of u with parameters $\kappa, \theta, \sigma, \rho, S, V$. The parameters $\kappa, \theta, \sigma, \rho$ are given in Heston model in (4.1) and S, V are the level of underlying and the variance at the time point that the call price is evaluate. The function

ψ is defined as

$$\psi(u, \kappa, \theta, \sigma, \rho, S, V) = \exp(iu \log S + C + DV),$$

where

$$C = \frac{\theta\kappa}{\sigma^2}\left[(\kappa - \rho\sigma iu - d)T - 2\log(\frac{1 - g\exp(-dT)}{1 - g})\right],$$
$$D = \frac{(\kappa - \rho\sigma iu - d)(1 - \exp(-dT))}{\sigma^2(1 - g\exp(-dT))},$$
$$d = \sqrt{(\kappa - \rho\sigma iu)^2 + \sigma^2(iu + u^2)},$$
$$g = \frac{\kappa - \rho\sigma iu - d}{\kappa - \rho\sigma iu + d}.$$

The data set is divided into 2 groups. In each group, 10000 samples of underlying price and variance are simulated with parameters $\kappa = 2, \theta = 0.04, \sigma = 0.1, \rho = -0.25, S_0 = 1, V_0 = 0.04$ at time $t = 0.25$ (year) (group 1), and $\kappa = 2, \theta = 0.04, \sigma = 0.2, \rho = -0.4, S_0 = 1, V_0 = 0.04$ at time $t = 0.25$ (group 2). Note that here the parameter ρ should be negative to match the empirically observed stylized facts in equity derivative market [42]. Then call option prices are computed using the formula above based on these 2 groups of simulated underlying price and variance. Strikes are chosen to be from 0.8 to 1.2 with grid 0.1 and maturities are chosen to be from 0.5 to 2.75 with grid 0.25. Actually, if standing at time $t = 0$, the maturities is from 0.75 to 3.

Now in each group, we have 10000 samples of call price at $t = 0.25$ with 50 different pairs of strike and maturity (5 strikes and 10 maturities). These can be viewed as samples of 50 points on a random surface $C(K, T)$. We simulate samples of 50 points instead of the whole surface because we can only observe call price with finitely many strikes and maturities in the market. After simulating 10000 samples, we find that there are several (less than 10) small negative values for the call price. These unreasonable values are due to the numerical errors in

the computation procedure, especially in the numerical integration procedure. We use 5000 samples (each sample containing 50 call prices with different strikes and maturities), which are all positive, out of the total 10000 for the proceeding numerical analysis and compare the results between 2 groups with different paramters.

The data set above is simulated with Matlab. All the codes are presented in the Appendix A.

4.3 Numerical Implementation Details for Karhunen-Loéve Procedure

4.3.1 Standard Implementation

As discussed in Section 2.1 and Section 2.2, performing centered and uncentered Karhunen-Loéve decompositions are formally eigenvalue problems for operators $\mathcal{K}$ and $\mathcal{M}$ in $L^2(\Omega \times A)$. The standard method for solving this kind of problem is to reduce them to finite dimensional problems with the stochastic Galerkin procedure [36].

Taking the operator $\mathcal{K}$ and the basis $\{f_i\}$ for example, the idea of Galerkin method is to expand each eigenfunction f_i on some basis $\{h_i\}$ and to take the truncated sum of first J terms

$$f_i(x) \approx \sum_{j=1}^{J} a_{ij} h_i(x).$$

Substituting the truncated sum into the equation

$$\mathcal{K}(f_i)(x) = \int_A K(x,y) f_i(x,y)dy = \lambda_i f_i(x)$$

yields an error term

$$\varepsilon_{iJ} = \int_A K(x,y) \sum_{j=1}^{J} a_{ij}h_i(y)dy - \lambda_i \sum_{j=1}^{J} a_{ij}h_i(x) = \sum_{j=1}^{J} a_{ij}\left(\int_A K(x,y)h_i(y)dy - \lambda_i h_i(x)\right).$$

The Galerkin method consists in requiring that the error term ε_{iJ} be orthogonal to $h_1, h_2, \cdots, h_J$. The orthogonality condition can be written as a system of linear equations, the solution of which gives the coefficients a_{ij} and an approximation $\hat{\lambda}_i$ of the true eigenvalue λ_i.

When applied to our problem, there are several defects with this standard implementation procedure.

1. The results of stochastic Galerkin procedure heavily depends on the choice of the number and forms of the basis functions in the expansion. However, there is no optimal way to select $\{h_i\}$ and J.

2. It is possible that the operator $\mathcal{K}$ has infinitely many (countable) eigenvalues. However, the Galerkin method can only give at most J approximations of them. This causes problem when we analyze the variances.

3. It is difficult to control the error in the results given by solving the orthogonality condition.

4. Remember that we do not observe the samples of the whole random surface $U(x_1, x_2) = C(K,T), (K,T) \in [a,b] \times [c,d]$, but only the sample of the random matrix $U_{ij} = C(K_i, T_j)$ where the entries are points on the surface. Especially, in the simulated data set, there are 5000 samples for a 5×10 random matrix (5

strikes and 10 maturities). In order to obtain $\mathcal{K}$ and $\mathcal{M}$, we need to first interpolate or smooth the discrete data. This makes the results depend heavily on the choice of the interpolation / smoothing.

Due to these defects, we apply another implementation procedure, which is described in the next section.

4.3.2 2-Dimensional Array PCA Implementation

Notice that in the simulated data set, as well as in practice, what we have is discrete data, instead of whole surface of call price. This makes it reasonable to reduce the centered and uncentered Karhunen-Loéve procedures for random surfaces to the PCA and spectral decomposition procedures for random matrices. However, what is different from the usual PCA is that we deal with samples for random matrix instead of random vector. Formally, we expand the 5×10 random matrix $C(K_i, T_j)$ as

$$C(K_i, T_j) = c(K_i, T_j) + \sum_{l=1}^{50} U_l f_l(K_i, T_j)$$

and

$$C(K_i, T_j) = \sum_{l=1}^{50} V_l g_l(K_i, T_j)$$

where $c(K_i, T_j) = E(C(K_i, T_j))$, f_l and g_l are the "eigenmatrix" of the four dimensional arrays $Cov(C(K_i, T_j)C(K_m, T_n))$ and $E(C(K_i, T_j)C(K_m, T_n))$.

We should mention here that one important reason why this reduction can be performed is that all the samples are with the same grid of strikes and maturities. This might not be true in the real market observed data, since the strikes

and maturities, especially the strikes, of the liquidly traded call options in the market might change everyday.

Since $R^{5\times10}$ is equivalent to R^{50} as inner product spaces (finite dimensional Hilbert spaces) and hence $L^2(\Omega) \otimes R^{5\times10}$ is equivalent to $L^2(\Omega) \otimes R^{50}$ as Hilbert spaces, the procedures can be further reduced to the centered and uncentered PCA for the 50×1 random vector $\mathbf{U}$. This is done by viewing those 5×10 matrices as 50×1 vectors. Results are presented in Section 4.4.

After computing the eigenvectors, we change them back to the matrix form and interpolate or smooth them to get the eigenmode surfaces. In this procedure, there is no information lost again because all the samples are with the same grid of strikes and maturities.

All data analysis in Section 4.5 is performed with the statistical software R. All the codes are presented in the Appendix B and C.

4.4 Shape Constrained Eigenmodes

What remains now is to interpolate or smooth the eigenmode matrices to surfaces. This is not the usual interpolation / smoothing, since the resulting surface are supposed to satisfy the shape constraints in Section 3.2. First and second order derivatives are involved in these constraints.

We do not develop new methods or techniques for constrained interpolation or smoothing, since this is not the focus of our research. Some methods has been proposed for interpolate or smooth discrete data set of call prices or implied volatilities to the whole surface satisfying some shape constraints [1, 14, 47]. We

apply the constrained quantile B-splines [40, 60] and implement this method with the "cobs" package in R.

A note here is that this "cobs" method, as well as other method, are mostly for smoothing 1-dimensional data to curves, but not for high-dimensional cases. However, we can still apply this method since our final goal is to getting the description of the spot volatility σ_t from implied volatility $\sigma_t^{BS}(K,T)$, as mentioned in section 1.2.1. The connection comes from the relation

$$\lim_{T-t\to 0} \sigma_t^{BS}(S_t, T) = \sigma_t$$

where S_t is the price of the underlying (that is, the implied volatility is got from at the money call option). Hence we only need to analyze and modify the eigenmodes $f_i(K,T)$ and $g_i(K,T)$ at $K = S_0$ and get modified eigenmode functions $\hat{f}_i(S,T)$ and $\hat{g}_i(S,T)$.

Again we need to mention that this method is possible for simulated data since the the samples for at the money call prices are with the same strike S_0. However, when considering real market data, one problem is that the strike for at the money call is always changing since the underlying price is not constant.

4.5 Numerical Analysis Results

4.5.1 Centered Karhunen-Loéve Decomposition

We apply the usual (centered) Karhunen-Loéve expansion to the simulated data set described in Section 4.2.

For the first group of data ($\kappa = 2, \theta = 0.04, \sigma = 0.1, \rho = -0.25, S_0 = 1, V_0 = 0.04$), the centering matrix is shown in Table 4.1 and Table 4.2.

	0.5	0.75	1.0	1.25	1.5
0.8	0.206236715	0.211036421	0.2159182	0.220755026	0.225496158
0.9	0.126418897	0.135023409	0.14284645	0.150060693	0.156785685
1.0	0.067869918	0.078427562	0.087722099	0.096119989	0.103837349
1.1	0.031841285	0.041461899	0.050213214	0.058285776	0.065811279
1.2	0.013178614	0.020123966	0.026995352	0.033681049	0.040147742

Table 4.1: Centering (mean) matrix in PCA (part): Strikes: 0.8, 0.9, 1.0, 1.1, 1.2; Maturities: 0.5, 0.75, 1.0, 1.25, 1.5. Group 1

	1.75	2.0	2.25	2.5	2.75
0.8	0.23012068	0.234622983	0.239004135	0.243267778	0.24741926
0.9	0.163107236	0.169088504	0.174777693	0.18021255	0.185423413
1.0	0.111015334	0.117751677	0.124117418	0.13016629	0.135940059
1.1	0.072883853	0.079573771	0.085934739	0.092009094	0.097830758
1.2	0.04639123	0.052419399	0.058244382	0.063879683	0.069338641

Table 4.2: Centering (mean) matrix in PCA (part): Strikes: 0.8, 0.9, 1.0, 1.1, 1.2; Maturities: 1.75, 2.0, 2.25, 2.5, 2.75. Group 1

As discussed in Section 3.3, the centering matrix satisfies the shape constraints. This can be justified from the above tables. We then interpolate the mean (a 5×10 matrix) into a surface with the usual (unconstrained) 2-dimensional spline functions, which is illustrated in Figure 4.1.

It is easy to see that this surface satisfies the shape constraints, that is, increasing with respect to the maturity and convex with respect to the strike.

The principal components might violate the shape constraints. For example, it is easy to see from Figure 4.3 that the at the money ($K = S_0 = 1$) part of the first component in PCA is decreasing with respect to T. Constrained spline functions are used for smoothing to avoid the violation of the shape constraints. The result is shown in Figure 4.4. Notice that in Figure 4.4, the curve becomes a

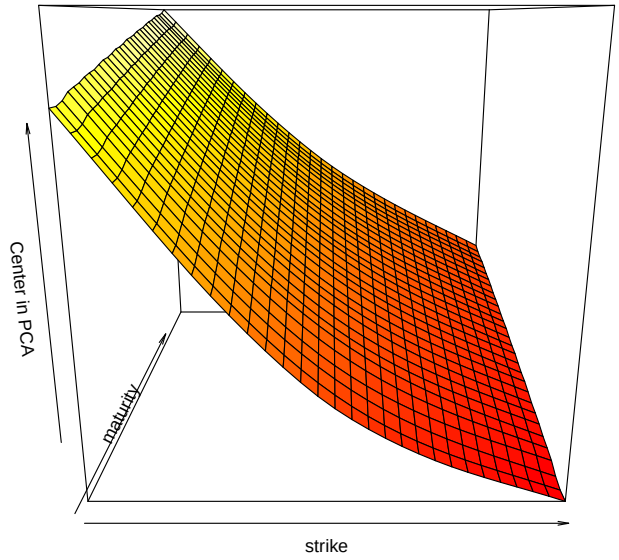

Figure 4.1: Surface interpolated from the center matrix in PCA, with unconstrained spline functions. Group 1

horizontal line. The discrete data is strictly decreasing, for which an increasing curve is fitted in. Hence the best result is a constant line.

For the second component, Figure 4.7 shows that at the money part is increasing with respect to T, satisfying the constraint. Since the shape constraints are not violated, usual (unconstrained) spline functions can be applied for smoothing and the result is shown in Figure 4.8.

	Component 1	Component 2	Component 3
Standard deviation	0.4224074	0.034211851	0.0057996510
Proportion of Variance	0.9932649	0.006515611	0.0001872431
Cumulative Proportion	0.9932649	0.999780482	0.9999677254

Table 4.3: Summary of the first 3 components in PCA. Group 1

The case for the third component is similar to the first one above. From

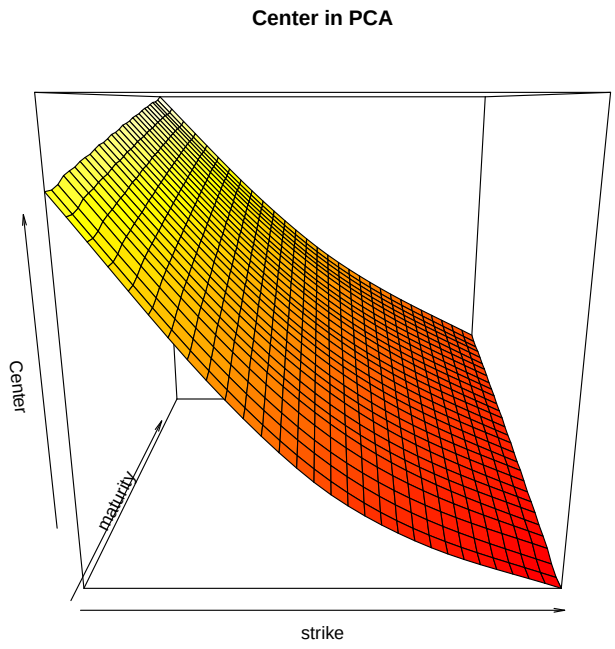

Figure 4.2: Surface interpolated from the center matrix in PCA, with unconstrained spline functions. Group 2

Table 4.3, we find that the first 3 components can explain more than 99.9% of the variance, which means that the factor model built from these 3 factors is quite good.

For the second group of data ($\kappa = 2, \theta = 0.04, \sigma = 0.2, \rho = -0.4, S_0 = 1, V_0 = 0.04$), again the centering matrix satisfies the shape constraints. We interpolate the center (a 5×10 matrix) into a surface with the usual (unconstrained) 2-dimensional spline functions, which is illustrated in Figure 4.2.

It is easy to see that this surface satisfies the shape constraints, that is, increasing with respect to the maturity and convex with respect to the strike.

Again, the principal components might violate the shape constraints. For example, it is easy to see from Figure 4.5 that the at the money ($K = S_0 = 1$) part of

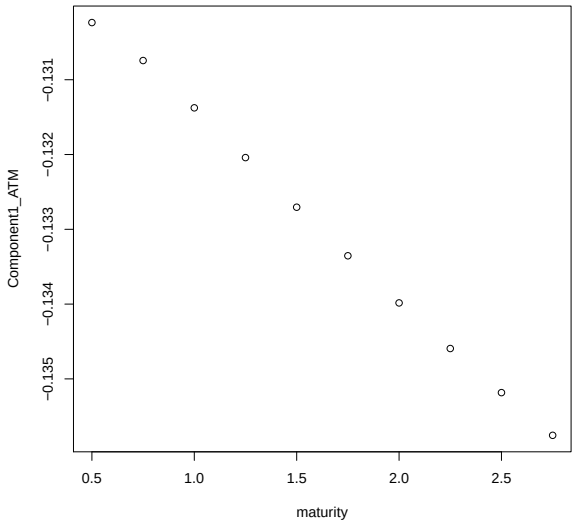

Figure 4.3: At the money ($K = S_0 = 1$) part of the first component in PCA. Group 1

the first component in PCA is decreasing with respect to T. Constrained spline functions are used for smoothing to avoid the violation of the shape constraints. The result is shown in Figure 4.6.

For the second component, Figure 4.9 shows that at the money part is increasing with respect to T, satisfying the constraint. Since the shape constraints are not violated, usual (unconstrained) spline functions can be applied for smoothing and the result is shown in Figure 4.10.

	Component 1	Component 2	Component 3
Standard deviation	0.4191405	0.038507003	0.007283338
Proportion of Variance	0.9912550	0.008366535	0.000299314
Cumulative Proportion	0.9912550	0.999621502	0.999920816

Table 4.4: Summary of the first 3 components in PCA. Group 2

The case for the third component is similar to the first one above. From Table 4.4, we find that the first 3 components can explain more than 99.9% of the variance, which means that the factor model built from these 3 factors is

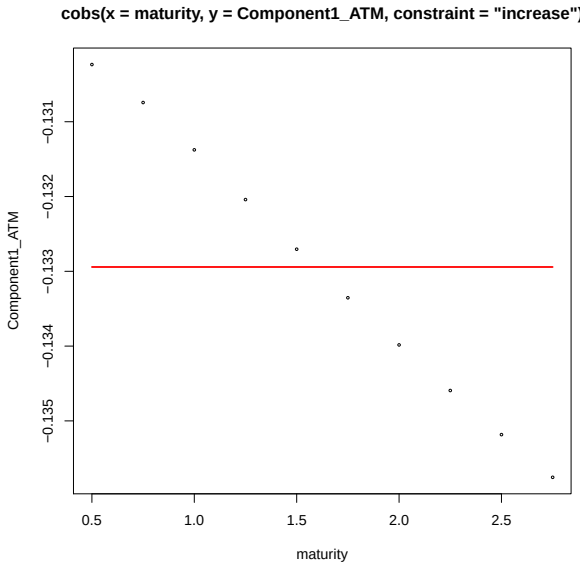

Figure 4.4: At the money ($K = S_0 = 1$) part of the first component in PCA, smoothed by constrained spline functions. Group 1

quite good.

Comparing the results for 2 groups of data, high similarity can be seen. This means that the pattern displayed in the centered Karhunen-Loéve procedure might not depend on the parameters in the model, but from some features of the model itself.

4.5.2 Uncentered Karhunen-Loéve Decomposition

We apply the uncentered Karhunen-Loéve expansion to the simulated data set described in Section 4.2.

For the first group of data ($\kappa = 2, \theta = 0.04, \sigma = 0.1, \rho = -0.25, S_0 = 1, V_0 = 0.04$), the first eigenmode matrix is shown in Table 4.5 and Table 4.6. Actually what the tables show are the negative of what R gives as the first eigenmode.

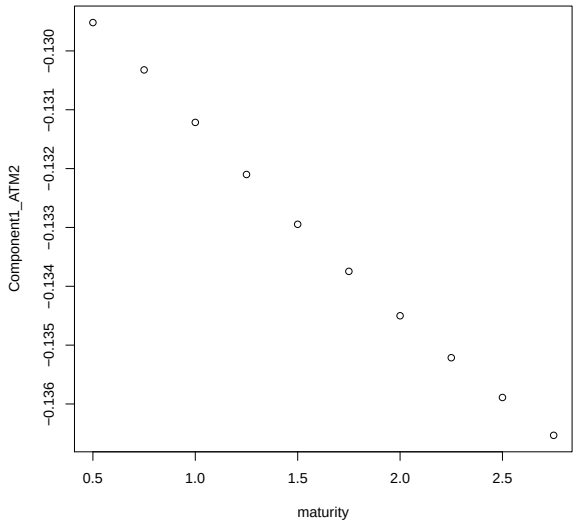

Figure 4.5: At the money ($K = S_0 = 1$) part of the first component in PCA. Group 2

We take the negative to make the all the entries positive.

	0.5	0.75	1.0	1.25	1.5
0.8	0.211972486	0.215167846	0.2186403	0.222224949	0.225837716
0.9	0.137477988	0.144092313	0.150306915	0.156155154	0.161682213
1.0	0.078901414	0.088132264	0.096285616	0.103665852	0.110455401
1.1	0.039769422	0.049199741	0.057589069	0.065217703	0.072258027
1.2	0.017688535	0.025217605	0.032347909	0.039092296	0.045487585

Table 4.5: First eigenmode matrix in uncentered Karhunen-Loéve (part): Strikes: 0.8, 0.9, 1.0, 1.1, 1.2; Maturities: 0.5, 0.75, 1.0, 1.25, 1.5. Group 1

As discussed in section 3.3, the first eigenmode satisfies the shape constraints. This can be justified from the above tables. We then interpolate it (a 5×10 matrix) into a surface with the usual (unconstrained) 2-dimensional spline functions, which is illustrated in Figure 4.11. Again it is easy to see that this surface satisfies the shape constraints, that is, increasing with respect to the maturity and convex with respect to the strike.

However, again the other eigenmodes might violate the shape constraints.

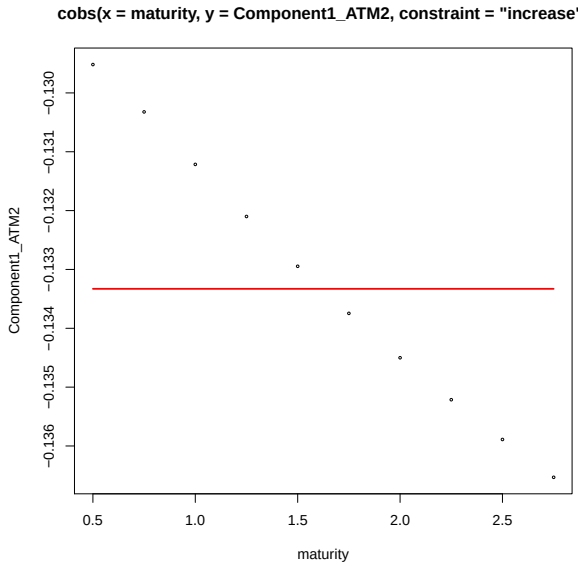

Figure 4.6: At the money ($K = S_0 = 1$) part of the first component in PCA, smoothed by constrained spline functions. Group 2

	1.75	2.0	2.25	2.5	2.75
0.8	0.229433052	0.232987026	0.236487125	0.239926299	0.243301727
0.9	0.166929304	0.171931157	0.176716703	0.181309828	0.185730819
1.0	0.116774923	0.122708476	0.128317522	0.133648748	0.138738553
1.1	0.078825305	0.08500143	0.090846913	0.096408116	0.101721439
1.2	0.051571812	0.057379763	0.062941428	0.068282539	0.073425047

Table 4.6: Second eigenmode matrix in uncentered Karhunen-Loéve (part): Strikes: 0.8, 0.9, 1.0, 1.1, 1.2; Maturities: 1.75, 2.0, 2.25, 2.5, 2.75. Group 1

For example, it is easy to see from Figure 4.12 that the at the money ($K = S_0 = 1$) part of the third component in uncentered Karhunen-Loéve decomposition is not increasing with respect to T.

Constrained spline functions are used for smoothing to avoid the violation of the shape constraints. The result is shown in Figure 4.13.

The second eigenmode is similar to the first and third components in PCA in Section 4.5.1.

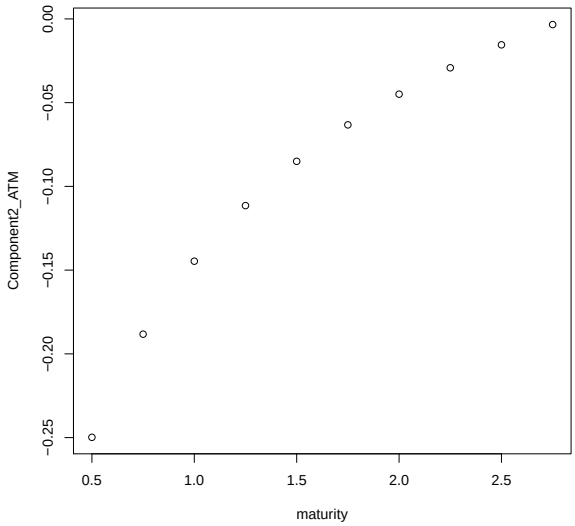

Figure 4.7: At the money ($K = S_0 = 1$) part of the second component in PCA. Group 1

For the second group of data ($\kappa = 2, \theta = 0.04, \sigma = 0.2, \rho = -0.4, S_0 = 1, V_0 = 0.04$), again the first eigenmode satisfies the shape constraints. We interpolate it (a 5×10 matrix) into a surface with the usual (unconstrained) 2-dimensional spline functions, which is illustrated in Figure 4.14. Again it is easy to see that this surface satisfies the shape constraints, that is, increasing with respect to the maturity and convex with respect to the strike.

The other eigenmodes might violate the shape constraints. For example, it is easy to see from Figure 4.15 that the at the money ($K = S_0 = 1$) part of the third component in uncentered Karhunen-Loéve decomposition is not increasing with respect to T. Constrained spline functions are used for smoothing to avoid the violation of the shape constraints. The result is shown in Figure 4.16. The second eigenmode is similar to the first and third components in PCA in Section 4.5.1.

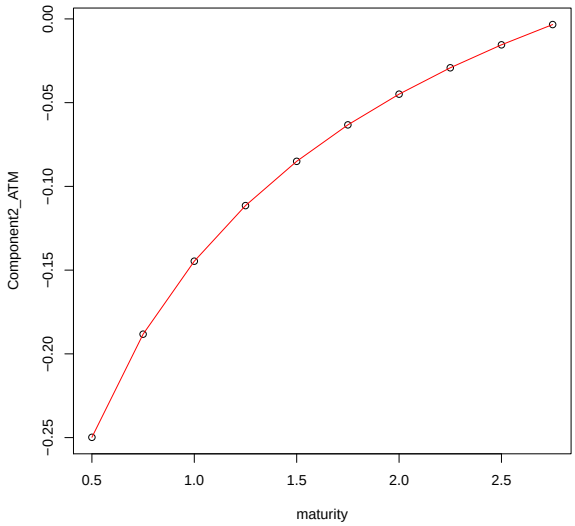

Figure 4.8: At the money ($K = S_0 = 1$) part of the secon component in PCA, smoothed by usual (unconstrained) spline functions. Group 1

Comparing the results for 2 groups of data, high similarity can be seen again. This means that the pattern displayed in the centered Karhunen-Loéve procedure might not depend on the parameters in the model, but from some features of the model itself.

From the results above, it can be seen that centered and uncentered Karhunen-Loéve decompositions might identify components and eigenmodes independent with the model parameters, but capturing some nature features of structure in the data set itself.

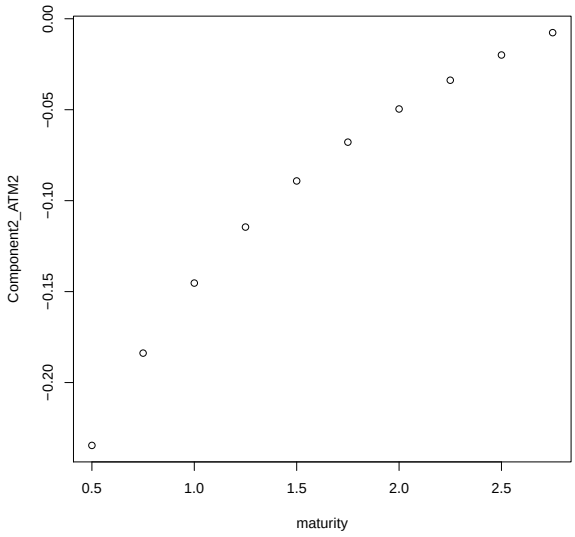

Figure 4.9: At the money ($K = S_0 = 1$) part of the second component in PCA. Group 2

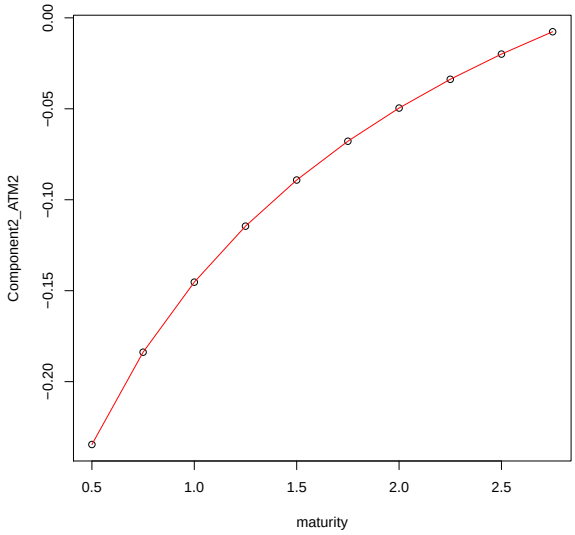

Figure 4.10: At the money ($K = S_0 = 1$) part of the secon component in PCA, smoothed by usual (unconstrained) spline functions. Group 2

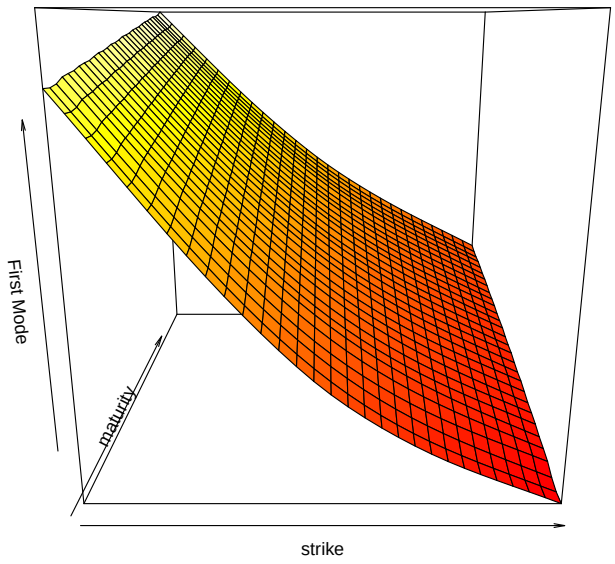

Figure 4.11: Surface interpolated from the first eigenmode in uncentered Karhunen-Loéve, with unconstrained spline functions. Group 1

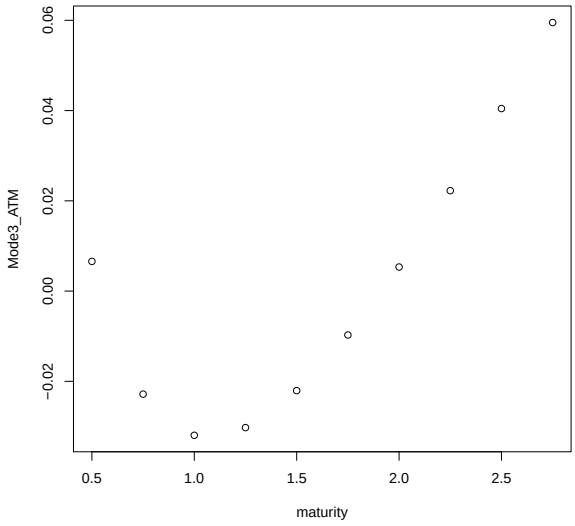

Figure 4.12: At the money ($K = S_0 = 1$) part of the third component in uncentered Karhunen-Loéve. Group 1

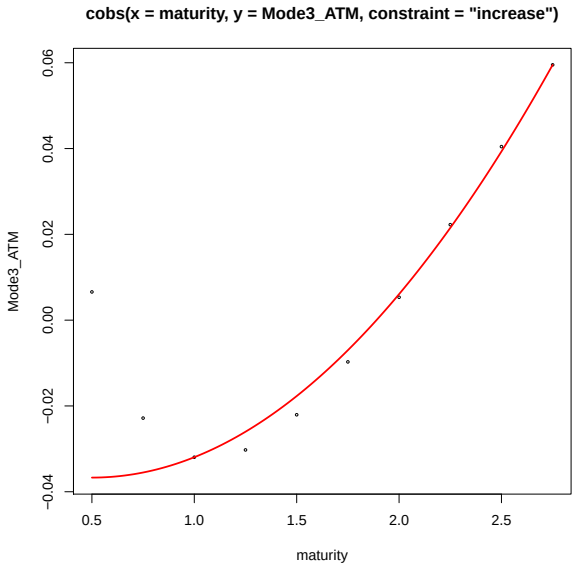

Figure 4.13: At the money ($K = S_0 = 1$) part of the third component in uncentered Karhunen-Loéve, smoothed by constrained spline functions. Group 1

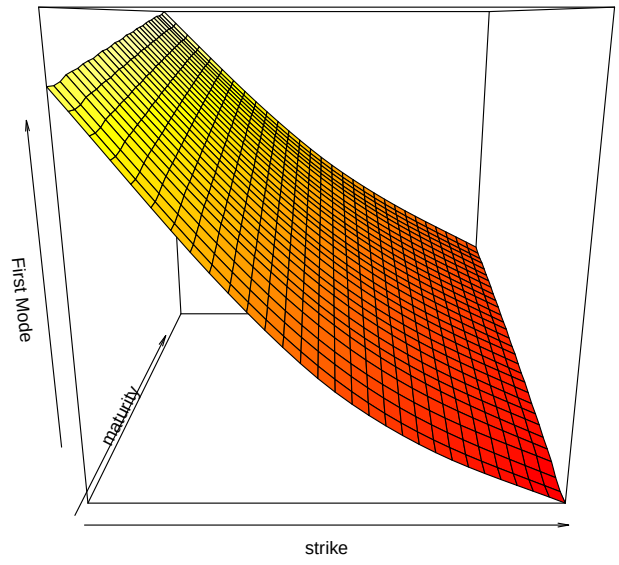

Figure 4.14: Surface interpolated from the first eigenmode in uncentered Karhunen-Loéve, with unconstrained spline functions. Group 2

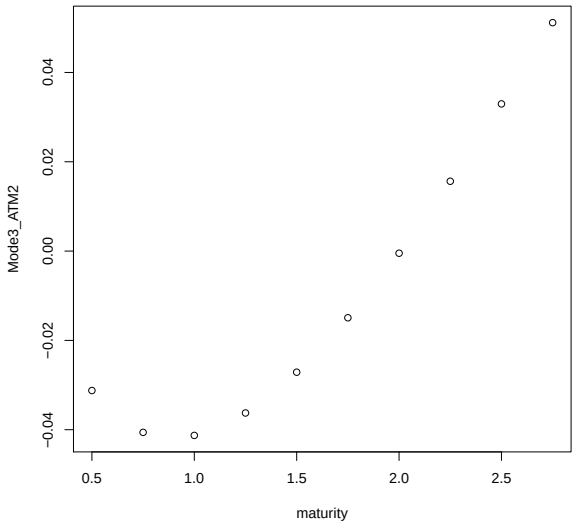

Figure 4.15: At the money ($K = S_0 = 1$) part of the third component in uncentered Karhunen-Loéve. Group 2

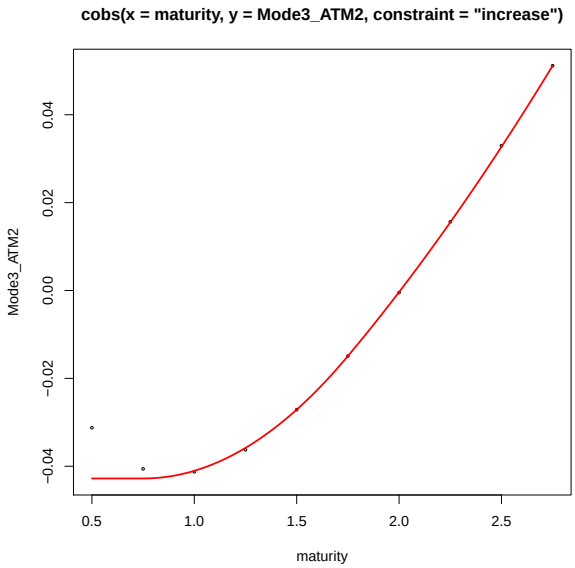

Figure 4.16: At the money ($K = S_0 = 1$) part of the third component in uncentered Karhunen-Loéve, smoothed by constrained spline functions. Group 2

CHAPTER 5

SUMMARY AND FUTURE RESEARCH

5.1 A Brief Summary

Exotic option pricing and hedging has long been the central topic in mathematical finance. The oversimplification and incapability of Black-Scholes model has led to a considerable literature on alternative option pricing models. The goal of this thesis is to move a step forward in the direction of market-based approach for exotic option pricing and hedging, which is a recently developed methodology in this area. We first build factor models for European call option price surface based on centered and uncentered Karhunen-Loéve decompositions, and then transform it to the Black-Scholes implied volatility. Since the spot volatility can be described by the asymptotic behavior of implied volatility, the models for implied volatility can give a description of the market dynamics and hence help pricing and hedging for exotic options.

In Chapter 2, a rigorous mathematical framework for Karhunen-Loéve decomposition is given and a variant of it, the uncentered Karhunen-Loéve decomposition is developed. In Chapter 3, we discuss the static arbitrage constraints for call option price surface in details and proposes methods to incorporate these shape constraints into the centered and uncentered Karhunen-Loéve procedures. In Chapter 4, we apply the methodologies discussed in chapter 2 and chapter 3 to a simulated data set of call price surface samples. Some numerical analysis results is reported for samples simulated from Heston model with 2 different groups of parameters. Analysis and comparision for these results is carried out.

5.2 Some Future Research Directions

5.2.1 Real Data Issues

Real data is not dealt with in this thesis since there are many realistic issues which may seriously affect the results when considering real market observed data. These issues are very complicated and beyond the scope of this thesis. However, they are important for the implementation of the market-based approach in practice. We list several of them below for future research.

1. How to determine which call option is liquidly traded and which is not? The idea of market-based approach relies on the fact that call options are liquidly traded in the market and the prices of them can be directly quoted. However, in practice, one can find that only several call options, with particular strikes and maturities, have large trading volumes in the market, while other call options' trading volume are very small, even 0 in some cases. Hence it is important to propose realistic criteria for the liquidity of call options. A call option is liquidly traded or not should be determined according to these criteria. The proposition for these criteria relies on empirical research on the real data from call option market. Also economic analysis on the supply and demand for call options might be helpful.

2. Bid-ask spread (also known as buy-sell spread). It is known that there is a difference between the prices quoted (either by a single market maker or in a limit order book) for an immediate sale (ask) and an immediate purchase (bid) for call options (actually, for most kind of securities) in the market. A problem is what price we should use as the input in the market-based approach. The

use of any prices other than the ask price might cause problem when hedging if performed, since the call option needed might not be available in the market with that price. Similarly, using just the ask price might give overvaluation of the options. Also the size of the bid-ask spread in a security is one measure of the liquidity of the market and of the size of the transaction cost. So the bid-ask spread issue might be related to the liquidity issue. However, we do observe call options in the market with relatively large bid-ask spread and small trading volume. The real situation might be very complicated.

3. In the real data for call option prices, there is special structures. The sample is from a time series of daily prices of call options quoted from the market. The strike and maturities for liquidly traded call options usually changes every trading day, which means that the strikes and maturities in each sample are different. Special statistical estimators and procedures, especially for the surface interpolation / smoothing and covariance estimation, should be developed for this structure.

5.2.2 The General Structure of Market-Based Approach

There are many interesting topics in the general structures of the market-based approach. Several of them are listed here.

1. Can we go beyond the Brownian motion setup? For example, a few research has been done on adding jumps to the stochastic volatility models and local volatility model. Inspiring by this, we might try to add jumps in the market-based approach framework to generalize the Brownian motion assumption.

2. Market completeness. It is well-known that the market completeness can be related with the uniqueness of pricing measure (equivalent local martingale measure) by the second fundamental theorem in asset pricing. European call options are added as the primary securities in the market based approach and this enlarges the set of attainable claims. On the other hand, the adding of these call options also shrink the set of pricing measures. It is interesting to investigate the relations between these two sets and how much these call options improve the market completeness in the model.

3. Consistency problem. In HJM model, it is possible to determine all the possible finite factor models from a geometric point of view, i.e. the finite-dimensional realizations. Similarly, it would be interesting to ask if this can be done in the market-based approach. Actually, this problem is expected to be difficulty since the system of SDE is much more complicated for the call price surfaces than for the forward rate curves. However, this is no doubt an interesting problem in the market-based approach framework.

APPENDIX A

MATLAB CODE FOR DATA SIMULATION

This part gives all the Matlab code generating data set described in section 4.2.

10000 samples of pairs of the underlying price and variance are simulated in Heston model with parameters $\kappa = 2, \theta = 0.04, \sigma = 0.1, \rho = -0.25, S_0 = 1, V_0 = 0.04$. Then the call option prices are computed from these underlying prices and variances with the same parameters. The Matlab code is given below. For another group of data, with parameters $\kappa = 2, \theta = 0.04, \sigma = 0.2, \rho = -0.4, S_0 = 1, V_0 = 0.04$, the code is similar.

Code for simulating underlying price and variance:

```
function [ HestonSample ] = HestonSim( kappa, theta, sigma, rho, S0, V0, t, N,
rep )
V = [V0*ones(rep, 1), zeros(rep, N)];
S = [S0*ones(rep,1),zeros(rep, N)];
NormRand1 = randn(rep, N);
NormRand2 = randn(rep, N);
dt = t / N;
for i = 1 : N
S(:, i+1) = S(:, i) + sqrt(V(:, i)) .* S(:, i) .* NormRand1(:, i) * sqrt(dt);
V(:, i+1) = V(:, i) + kappa * (theta-V(:, i)) * dt + sigma * sqrt(V(:,i)) .* (rho * Norm-
Rand1(:, i) + sqrt(1-rho^2)* NormRand2(:, i)) * sqrt(dt);
V(:, i+1) = abs(V(:, i+1));
end
```

```
HestonSample = [S(:, N+1), V(:, N+1)];
end
HestonSample = HestonSim(2, 0.04, 0.2, -0.4, 1, 0.04, 0.25, 1000, 10000);
```

Code for computing call prices:

```
function [ phi ] = phiHeston( kappa, theta, sigma, rho, u, T, S, V )
d = sqrt( (rho * sigma * 1i * u - kappa) ^ 2 + sigma ^ 2 * (1i * u + u ^ 2) );
g = (kappa - rho * sigma * 1i * u - d) / (kappa - rho * sigma * 1i * u + d);
C = (theta * kappa / sigma ^ 2) * ( (kappa - rho * sigma * 1i * u - d) * T - 2 *
log((1 - g * exp(-d * T)) / (1-g)) );
D = (kappa - rho * sigma * 1i * u -d) * (1 - exp(-d * T)) / (1 - g * exp(-d * T)) /
sigma ^ 2;
phi = exp(1i * u * log(S) + C + D * V);
end
```

```
function [ call_price ] = CallPriceHeston( kappa, theta, sigma, rho, K, T, S, V, al-
pha, umax )
function y = psifun( v )
z = exp(-1i * log(K) * v) * phiHeston( kappa, theta, sigma, rho, v-1i * (alpha+1),
T, S, V ) / (alpha ^ 2 + alpha - v ^ 2 + 1i * (2 * alpha+1) * v); y = real(z);
end
call_price = exp(-alpha * log(K)) * quadv(@psifun, 0, umax) ./ pi;
end
```

```
rep = 10000;
strike = [0.8 : 0.1 : 1.2];
maturity =[0.5 : 0.25 : 2.75];
for i = 1 : rep
for j = 1 : length(strike)
for k = 1 : length(maturity)
CallPriceSamp(j, k, i) = CallPriceHeston(2, 0.04, 0.1, -0.25, strike(j), maturity(k),
HestonSamp(i, 1), HestonSamp(i, 2), 1.5, 1000);
end
end
end
```

APPENDIX B

R CODE FOR CENTERED AND UNCENTERED KARHUNEN-LOÉVE ANALYSIS

This part gives all the R code for centered and uncentered Karhunen-Loéve analysis described in section 4.3.

Centered Karhunen-Loéve decomposition:

```
SimCallComp = princomp(SimCallPCA);
summary(SimCallComp);


Centermat = matrix(nrow = 5, ncol = 10);
Center = as.vector(SimCallComp $ center);
for (i in 1:10) {
Centermat[,i] = Center[(5 * (i-1)+1) : (5 * i)];
}
Comp1 = as.vector(SimCallComp $ loadings[,1]);
Comp2 = as.vector(SimCallComp $ loadings[,2]);
Comp3 = as.vector(SimCallComp $ loadings[,3]);
Comp1mat=matrix(nrow = 5, ncol = 10);
Comp2mat=matrix(nrow = 5, ncol = 10);
Comp3mat=matrix(nrow = 5, ncol = 10);
for (i in 1:10) {
Comp1mat[,i] = Comp1[(5 * (i-1)+1) : (5 * i)];
}
for (i in 1:10) {
Comp2mat[,i] = Comp2[(5 * (i-1)+1) : (5 * i)];
```

```
}
for (i in 1:10) {
Comp3mat[,i] = Comp3[(5 * (i-1)+1) : (5 * i)];
}
```

Uncentered Karhunen-Loéve decomposition:

```
SecMomSimCall = matrix(nrow = 50, ncol = 50);
for (i in 1:50) {
for (j in 1:50)
{
SecMomSimCall[i,j] = as.numeric(SimCallPCA[,i]) % * % as.numeric(SimCallPCA[,j]);
}
}
KLdecomp = eigen(SecMomSimCall);

Mode1 = KLdecomp $ vectors[,1];
Mode2 = KLdecomp $ vectors[,2];
Mode3 = KLdecomp $ vectors[,3];
Mode1mat = matrix(nrow = 5, ncol = 10);
for (i in 1:10) {
Mode1mat[,i] = Mode1[(5 * (i-1)+1) : (5 * i)]; }
Mode2mat = matrix(nrow = 5, ncol = 10);
for (i in 1:10) {
Mode2mat[,i] = Mode2[(5 * (i-1)+1) : (5 * i)]; }
Mode3mat = matrix(nrow = 5, ncol = 10);
```

```
for (i in 1:10) {
Mode3mat[,i] = Mode3[(5 * (i-1)+1) : (5 * i)];
}
```

APPENDIX C

R CODE FOR CONSTRAINED AND UNCONSTRAINED SPLINE SMOOTHING

This part gives all the R code for constrained and unconstrained spline smoothing described in section 4.4.

Unconstrained spline smoothing for the center in centered Karhunen-Loéve decomposition and the first eigenmode in uncentered Karhunen-Loéve decomposition:

```
y1 = as.vector(Centermat);
y2 = as.vector(-Mode1mat);
strike = rep(seq(.8,1.2,length=5),10);
maturity = rep(seq(.5,2.75,length=10),each = 5);
CenterSurf = gam(y1 s(strike, maturity));
CenterSurf2 = gam(y1 te(strike, maturity));
vis.gam(CenterSurf, main = ”Center in PCA”)
vis.gam(CenterSurf2, main = ”Center in PCA”)
```

Constrained spline smoothing for the first component in centered Karhunen-Loéve decomposition and the third eigenmode in uncentered Karhunen-Loéve decomposition:

```
Component1_ATM = Comp1mat[3,];
Mode3_ATM = -Mode3mat[3,];
maturity = seq(0.5,2.75, length = 10);
ConComp1_ATM < − cobs(maturity, Component1_ATM, constraint = ”in-
```

crease”);

ConMode1_ATM < − cobs(maturity, Mode3_ATM, constraint = ”increase”);

plot(ConComp1_ATM)

plot(ConMode1_ATM)

BIBLIOGRAPHY

[1] Y. Aït-Sahalia and J. Duarte. Nonparametric option pricing under shape restrictions. *Journal of Econometrics*, 116:9–47, 2003.

[2] Y. Aït-Sahalia, Y. Wang, and F. Yared. Do option markets correctly price the probabilities of movement of the underlying asset? *Journal of Econometrics*, 102:67–110, 2001.

[3] C. Alexander. Principles of the skew. *Risk*, January:S29–S32, 2001.

[4] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis (2nd Edition)*. John Wiley and Sons, 1971.

[5] Louis Bachelier. *Théorie de la spéculation*. Gabay, 1995.

[6] G. Bakshi, C. Cao, and Z. Chen. Empirical performance of alternative option pricing models. *Journal of Finance*, LII, no. 5:2003–2049, 1997.

[7] G. Bakshi, C. Cao, and Z. Chen. Do call prices and the underlying stock always move in the same direction? *Review of Financial Studies*, 13:549–584, 2000.

[8] A. Basilevsky. *Statistical Factor Analysis and Related Methods*. John Wiley and Sons, 1994.

[9] N. H. Bingham and R. Kiesel. *Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives*. Springer, 2004.

[10] F. Black and M. Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economy*, pages 637–654, 1973.

[11] T. P. Bollerslev. A conditional heteroscedastic time series model for security prices and rates of return data. *Review of Economics and Statistics*, 69:542–547, 1987.

[12] D. T. Breeden and R. H. Litzenberger. Prices of state-contingent claims implicit in option prices. *Journal of Business*, 51(4):621–651, 1978.

[13] C. L. Brooks, M. Karplus, and B. M. Pettitt. *Proteins: A Theoretical Perspective of Dynamics, Structure and Thermodynamics*. Wiley-Interscience, 1988.

[14] B. Brunner and R. Hafner. Arbitrage-free estimation of the risk-neutral density from the implied volatility smile. *Journal of Computational Finance*, 7:75–106, 2003.

[15] R. Buff. *Uncertain Volatility Models-Theory and Application*. Springer, 2002.

[16] R. Carmona and S. Nadtochiy. Local volatility dynamic models. *Finance and Stochastics*, 13:1–48, 2009.

[17] P. Carr, H. Géman, D. B. Madan, and M. Yor. Stochastic volatility for lévy processes. *Mathematical Finance*, 13:345–382, 2003.

[18] P. Carr and D. B. Madan. Option valuation using the fast fourier transform. *Journal of Computational Finance*, 2:61–73, 1998.

[19] L. Clewlow and X. Xu. A review of option pricing with stochastic volatility. *Working Paper*, 1992.

[20] R. Cont and J. da Fonseca. Dynamics of implied volatility surfaces. *Quantitative Finance*, 2:45–60, 2002.

[21] M. G. Cox and D. G. Hobson. Local martingales, bubbles and option prices. *Finance and Stochastics*, 9:477–492, 2005.

[22] S. R. Das and R. K. Sundaram. Of smiles and smirks: A term-structure perspective. *Journal of Financial and Quantitative Analysis*, 34:211–240, 1999.

[23] F. Delbaen and W. Schachermayer. A general version of the fundamental theorem of asset pricing. *Mathematische Annalen*, 300:463–520, 1994.

[24] E. Derman and I. Kani. Riding on a simle. *Risk*, 7(2):32–39, 1994.

[25] J. L. Doob. Generalized sweeping-out and probability. *Journal of Functional Analysis*, 2:207–225, 1968.

[26] B. Dumas, J. Fleming, and R. E. Whaley. Implied volatility functions: Empirical tests. *Journal of Finance*, 8:2059–2106, 1998.

[27] B. Dupire. Pricing with a smile. *Risk*, 7(1):18–20, 1994.

[28] V. Durrleman. From implied to spot volatilities. *Working Paper*, 2005.

[29] R. F. Engle and G. GonzalezRivera. Semiparametric ARCH models. *Journal of Business and Economics Statistics*, 9:345–359, 1991.

[30] P. Hagan et al. Managing smile risk. *Wilmott Magazine*, 1:84–108, 2002.

[31] M. R. Fengler, W. K. Härdle, and E. Mammen. A semiparametric factor model for implied volatility surface dynamics. *Journal of Financial Econometrics*, 5(2):189–218, 2007.

[32] M. R. Fengler, W. K. Härdle, and P. Schmidt. Common factors governing vdax movements and the maximum loss. *Journal of Financial Markets and Portfolio Management*, 16(1):16–29, 2002.

[33] M. R. Fengler, W. K. Härdle, and C. Villa. The dynamics of implied volatilities: A common principal component approach. *Review of Derivatives Research*, 6:179–202, 2003.

[34] J. R. Franks and E. J. Schwartz. The stochastic behaviour of market variance implied in the prices of index options. *The Economic Journal*, 101:1460–1475, 1991.

[35] K. R. French, G. W. Schwert, and R. F. Stambaugh. Expected stock returns and volatility. *Journal of Financial Economics*, 19:3–29, 1987.

[36] R. G. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer-Verlag, 1991.

[37] I. Gyöngy. Mimicking the one-dimensional marginal distributions of processes having an ito differential. *Probability Theory and Related Fields*, 71(4):501–516, 1986.

[38] R. Hafner. *Stochastic Implied Volatility*. Springer, 2004.

[39] M. Harrison and S. Pliska. Martingales and stochastic integrals in the theory of continuous trading. *Stochastic Processes and Their Applications*, 11:215–260, 1981.

[40] X. He and P. Ng. COBS: Qualitatively constrained smoothing via linear programming. *Computational Statistics*, 14:315–337, 1999.

[41] D. Heath, R. Jarrow, and A. Morton. Bond pricing and the term structure of

interest rates: A new methodology for contingent claims valuation. *Econometrica*, 60:77–105, 1992.

[42] S. Heston. A closed form solution for options with stochastic volatility with applications to bond and currency options. *Review of Financial Studies*, 6(2):327–343, 1993.

[43] R. Heynen. An empirical investigation of observed smile patterns. *Review of Futures Markets*, 13:317–353, 1993.

[44] R. Heynen, A. Kemma, and T. Vorst. Analysis of the term structure of implied volatilities. *Journal of Financial and Quantitative Analysis*, 29:31–56, 1994.

[45] J. Hull and A. White. The pricing of options on asset with stochastic volatilities. *Journal of Finance*, 42(2):281–300, 1987.

[46] J. Jacod and P. Protter. Risk neutral compatibility with option prices. *Working Paper*, 2006.

[47] N. Kahalé. An arbitrage-free interpolation of volatilities. *RISK*, 17(5):102–106, 2004.

[48] I. Karatzas and S. Shreve. *Methods of Mathematical Finance*. Springer, 1998.

[49] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Annales Academi Scientiarum Fennic*, 37, 1946.

[50] H. G. Kellerer. Markov-komposition und eine anwendung auf martingale. *Mathematische Annalen*, 198:99–122, 1972.

[51] O. Ledoit, P. Santa-Clara, and S. Yan. Relative pricing of options with stochastic volatility. *Working Paper*, 2002.

[52] A. L. Lewis. *Option Valuation under Stochastic Volatility*. Finance Press, 2000.

[53] A. S. Lewis, D. R. Luke, and J. Malick. Local linear convergence for alternating and averaged nonconvex projections. *Foundations of Computational Mathematics*, 9(4):485513, 2009.

[54] A. S. Lewis and J. Malick. Alternating projections on manifolds. *Mathematical of Operations Research*, 33(1):216234, 2008.

[55] M. M. Loéve. *Probability Theory*. VanNostrand, 1955.

[56] E. N. Lorenz. *Empirical Orthogonal Functions and Statistical Weather Prediction*. Statistical Forecasting Project, Department of Meteorology, MIT, 1956.

[57] J. L. Lumley. The structure of inhomogeneous turbulent flows. *In Atmospheric Turbulence and Radio Propagation*, pages 166–178, 1967.

[58] T. J. Lyons. Derivatives as tradable assets. *RISK 10th Anniversary Global Summit*, 1997.

[59] R. C. Merton. Theory of rational option pricing. *Bell Journal of Economics and Management Science*, 4 no. 1:141–183, 1973.

[60] P. Ng and M. Maechler. A fast and efficient implementation of qualitatively constrained quantile smoothing splines. *Statistical Modelling*, 7(4):315–328, 2007.

[61] S. Pal and P. Protter. Strict local martingales, bubbles, and no early exercise. *Working Paper*, 2007.

[62] D. Psychoyios, G. Skiadopoulos, and P. Alexakis. A review of stochastic volatility processes: Properties and implications. *Journal of Risk Finance*, 4(3):43–60, 2003.

[63] R. Rebonato. *Volatility and Correlation*. John Wiley, 1999.

[64] W. Rudin. *Functional Analysis (2nd Edition)*. McGraw-Hill, 1991.

[65] P. Samuelson. Rational theory of warrant pricing. *Industrial Management Review*, 6(2):13–32, 1965.

[66] P. J. Schönbucher. A market model for stochastic implied volatility. *Phil. Trans. of the Royal Soc., Series A*, 357:2071–2092, 1999.

[67] M. Schweizer and J. Wissel. Arbitrage-free market models for option prices: The multi-strike case. *Finance and Stochastics*, 12:469–505, 2008.

[68] M. Schweizer and J. Wissel. Term structures of implied volatilities: Absence of arbitrage and existence results. *Mathematical Finance*, 18(1):77–114, 2008.

[69] L. Scott. Option pricing when the variance changes randomly: Theory, estimation and an application. *Journal of Financial and Quantitative Analysis*, 22:419–438, 1987.

[70] G. Skiadopoulos. Volatility smile-consistent option models: A survey. *International Journal of Theoretical and Applied Finance*, 4(3):403–438, 2001.

[71] G. Skiadopoulos, S. Hodges, and L. Clewlow. The dynamics of the S&P 500 implied volatility surface. *Review of Derivatives Research*, 3:263–282, 2000.

[72] E. M. Stein and J. C. Stein. Stock price distributions with stochastic volatility: an analytic approach. *Review of Financial Studies*, 4:727–752, 1991.

[73] V. Strassen. The existence of probability measures with given marginals. *Annals of Mathematical Statistics*, 36:423–439, 1965.

[74] J. Wiggins. Option values under stochastic volatility. *Journal of Financial Economics*, 19:351–372, 1987.

[75] J. Wissel. Arbitrage-free market models for option prices. *Working Paper*, 2007.

[76] K. Yosida. *Functional Analysis (6th Edition)*. Springer-Verlag, 1980.

[77] Y. Zhu and M. Avellaneda. An e-arch model for the term structure of implied volatility of fx options. *Applied Mathematical Finance*, 4:81–100, 1997.