Interest Rate Term Structure Modeling in the Presence of Missing Data

Kia Karelmo

September, 2010

Abstract

In this study, issuer specific term structures of interest rates are estimated in the presence of missing data. A smoothing spline estimation is applied on a set of bootstrapped yield notations for a number of Emerging Market issuers. Based on estimated term structures, in which a significant fraction of data is missing, univariate GARCH(1,1) filters are applied on single time series, after which a PCA based agorithm fills missing observations. Term structures are thereafter recursively reconstructed. In the center of the study lies the designed backfilling routine. Its application onto a number of synthetic data sets shows that its performance is satisfactory. Although, it is dependent on the fraction of missing values, the complexity of the underlying factor structure and the amount of noise in the data set. Applied onto the real set of data, the routine produces results of varying quality. Obtained estimates of missing term structures appear credible but their correctness remains to be assessed.

Keywords: Yield Curve Estimation, Backfilling, GARCH(1,1), Principal Component Analysis.

Acknowledgements

There are many who deserves a big thank you for helping me throughout the work on my thesis. First and foremost, my two supervisors for their solid support and guidance. Gianluca Marcoli at UBS for his inspiring ideas and experience within the topic and Filip Lindskog for his academic support and guidance, as my contact at the Royal Institute of Technology. Thank you also to Mark and Anja for your enthusiasm and practical suggestions which I would have struggled without.

I would also like to thank my family and my boyfriend for being such great sources of motivation and energy. And to all the rest of you who in any way have helped me on the way. You know who you are.

Kia Karelmo

Zürich, September 2010

Contents

Chapter 1

Introduction

1.1 Background

Financial risk modeling is an important task for banks and institutions active within the finance sector. Increasing pressure both from financial authorities and own management emphasizes the significance of sophisticated risk models, perhaps even more in the aftermath of the recent financial market turmoil. At this day, most financial institutions perform risk management through mathematical computations based upon historical observations. The availability of such data is unfortunately not always as good as one might wish for. Risk measures, such as Value at Risk (VaR) and Expected Tail Loss (ETL), must nevertheless be calculated and reported to regulators on a daily basis. Estimating one's risk profile and forecasting possible future losses based on scarce data is therefore a difficulty that many financial institutions are forced to confront.

In the field of financial risk management, risks are commonly divided into three categories: Market Risk, Credit Risk and Operational Risk. *Market risk* arises from movements in general market factors such as interest rates, commodity prices and currency exchange rates. Market Risk is undiversifiable and is also commonly referred to as Systematic Risk. *Credit risk* is defined as the risk carried by the lender that a debtor will not be able to repay her debt, but it can also be the risk of a change in credit rating, reflecting the market's perception of an issuer's creditworthiness. *Operational risk* is the risk of losses due to failure in internal processes (Hult and Lindskog, 2007).

Idiosyncratic credit risk (also referred to as *specific credit risk*) aims to explain that the price of a product bearing credit risk, does not necessarily follow general market movements. The value of a product can behave in a unique way and the concept of idiosyncratic risk captures these excess changes in the product's value (Frisch et al., 2002). A typical example of such product is a corporate bond, which also carries the risk of default of the issuer. One way of assessing the idiosyncratic risk of the product is to observe its yield to maturity and how it changes in comparison with a risk free benchmark. In focus is thus the credit spread between the different yields, a popular indicator of risk for fixed income products.

One common approach is to explain the credit spread by mapping various parts of it to different risk factors such as rating and currency. Products with similar risk factors are grouped together and risks are calculated based on these characteristics. However, this division does not provide an explanation of the entire spread. The part of the credit spread which cannot be generalized and thus remains unexplained corresponds to the idiosyncratic credit risk of the product.

Another option would be to observe the yield curve derived from fixed income products of a single issuer. In the issuer specific yield curve one can claim to find all information about how prices of the issuer's products behave. By subtracting a risk free benchmark curve (such as LIBOR or various swap rates) together with explained changes in the credit spread, idiosyncratic movements in the issuer yield curve can be obtained, if such representation is desired. In the the Basel Amendment for incorporating market risk, specific credit risk is generally said to arise from issuer related events.[1] Specific risk calculations on issuer level are for example applied in the standard model for calculation of capital adequacy suggested in the FINMA circular (FINMA, 2009).[2]

1.2 The Problematics of Missing Data

Practitioners in a wide range of industries are on a daily basis forced to confront the issue of missing data or data of poor quality. In *geosciences*, with *climate research* as one example, most data sets are obtained through historical observations. Within *DNA Microarray Technology* experimental data is used for investigations of genetic behaviour. *Demographic research* is dependent of proper documentation of human population. However, rough weather conditions, experimental failures and insufficient surveys are just some of the reasons which can cause such data sets to be incomplete (Kondrashov and Ghil, 2006; Zhang et al., 2008).

These are just a couple of examples where the problem of missing data is substantial, but the list can be made much longer. Nevertheless, statistical tools and computational procedures generally require full sets of data. Sometimes the only way to obtain such data sets, is through estimation and imputation of missing observations.

In the finance industry, the situation is at least as severe and one of the areas where the problem perhaps becomes most significant is risk management. Missing data might not be a large issue in the world of equity, where historical prices for a given share are fairly easy to find. That is however seldom the case for more exotic products, or even regular bonds issued by corporates or smaller emerging nations. Single missing price notations as well as continous gaps are common to find among historical price notations of traded products. Despite this shortcoming, regulatory requirements on risk calculations, such as the number of historical observations that risk measures must be based upon and the frequency of which these observations are updated, become more and more extensive (Basel Committee, 2005). Finding a way to deal with missing observations has thus never been more important.

[1] In the Basel amendment of 1996 (updated in 2005), specific risk is defined as the risk that the price of a product moves more or less than the general market together with the product's event risk which arises from irregular and infrequent events, generally connected to the issuer, such as the risk of default.

[2] The FINMA Circular (08/20) presents guidelines for calculation of capital adequacy for Swiss banks due to risks caused by interest rate and share price changes for positions held within the banks trading book, as suggested by the Swiss Financial Market Supervisory Authority (FINMA).

1.3 Aim of Thesis

The aim of this Master's Thesis is to design a procedure for estimation of issuer specific interest rate term structures with scarce availability of data. The procedure is initialized with a term structure estimation using historical yield notations, each for a number of bonds of a number of different issuers. Missing values, seen as missing observations in the term structure representation, will thereafter be treated with a backfilling algorithm. Based on the assumption that single time series of yield changes, at constant maturities, follow a GARCH(1,1) process, GARCH filters will be applied on single time series followed by an iterative value imputation based on Principal Component Analysis.

The main focus of this study is backfilling of missing observations, which is also where the mathematical emphasis will be placed. The backfilling algorithm introduced in this study will therefore be additionally validated through application on various synthetic data sets from which results are evaluated separately. Its performane is a crucial ingredient to fill missing data in the already estimated term structure dimension.

The designed procedure will be implemented in the context of Emerging Markets, where the availability and quality of data is low, as a typical example of a business for which banks calculate risk measures on a daily basis. To find an optimal representation for missing observations, based on available data, will be the main challenge of this thesis project. Generalizations will have to be weighted against loss of the information one has at hand.

1.4 Outline

The disposition of this paper is as follows. The first chapter gives an introduction to the topic and explains the aim of the study. Chapter 2 presents a literary review of methods and models applied within the practice today. The third chapter presents a theoretical framework which is found necessary for the chosen topic. Chapter 4 and 5 describe the applied method together with the data which has been used throughout this study. Chapter 6 consists of a further assessment of the backfilling routine and is followed by general results of the study in chapter 7. Chapter 8 contains a concluding discussion together with suggestions for further research.

Chapter 2

Literary Review

This chapter presents a review of existing literature where similar challenges, as those faced in this study, have been dealt with. Applied methods of the two main parts of the study; *yield curve estimation* and *backfilling of missing data*, are discussed together with their respective advantages and disadvantages.

2.1 Approaches to Yield Curve Estimation

The Term Structure of Interest Rates, or the *Yield Curve* (the two expressions will be used analogously throughout this document), is a powerful financial tool. It illustrates the relationship between the yield of a zero-coupon bond and its time to maturity. The term structure of interest rates is an essential ingredient within portfolio management, financial engineering and financial risk management and is frequently used for pricing of defaultable bonds and credit derivatives. Another example of its application is the calculation of Value at Risk (VaR) with means of Historical Simulation, where future scenarios are generated based on historical changes in interest rates and credit spreads (Houweling et al., 2001; Lin, 2002).

Due to its popularity, estimating the yield curve has become a common practice and there is a wide range of methods designed for the purpose. Whichever technique a practitioner chooses there are a number of specifically desired properties of a good yield curve estimation. As mentioned in Nawalka and Soto (2009), the following characteristics should be obtained:

- The estimated curve should give suitable fit of the data.
- Estimated zero and forward rates should remain positive over the estimation period.
- Countinous and smooth functions should be fitted to the discount function, zero yields and forward rates.
- The estimated curve should allow for asymptotic shapes.

The main challenge one must face when modeling the yield curve is caused by characteristics of the sample data used for the estimation. This holds regardless if one wishes to model the discount function, the spot curve or the forward curve, three analogous approaches (Lin, 2002). The term structure as such is seldom directly observable in terms of market notations. It must instead be estimated from available prices and yields of desired products. Commonly used as a mean of obtaining zero rates, is the yield to maturity for a coupon bearing bond (which in turn can be seen as the internal rate of return that gives the bond's present value as all cash flows are discounted).

Among the predecessors of yield curve estimation one can find polynomial splines by McCulloch (1971, 1975), exponential splines as in Vasicek and Fong (1982) and the equilibrium model created by Cox et al. (1985). During later years, models such as the Nelson Siegel (Nelson and Siegel, 1987), with the extension of Svensson (1994), together with different versions of B-splines (see for example Ioannides (2003) and Lin (2002)) are popular approaches.

The basic relations between bond prices, zero rates and forwars rates will not be presented in this section. It is assumed that the reader is familiar with these. If this is not the sace, or if one simply wants a reminder, it can be helpful to review Nawalka and Soto (2009) for a fundamental explanation.

2.1.1 Piecewise Spline Estimation

Using piecewise defined functions, *splines*, to obtain a complete representation of the yield curve is a well-tried method. One of the strengths of splines is that an *a priori* curvature of the term structure does not have to be imposed. This is fully determined by the data to which the splines adjust. The revolutionary example is given by McCulloch and his attempt to estimate the discount function with the help of polynomial splines (McCulloch, 1971), but the subsequent variations are many.

Polynomial Splines by McCulloch

The discount function $\delta(m)$, where m is the maturity denoted in years, tells us the present value of one unit of a given currency repayable in m years. The further ahead in the future a payment takes place, the lower is its value today. With an increasing m, $\delta(m)$ typically decreases exponentially towards zero. Since the value of one unit today is exactly one unit, $\delta(0) = 1$ by construction. If one achieves to properly estimate the discount function, calculating the corresponding yield to maturity, as well as the forward rate, for a given coupon bearing bond, becomes a straightforward task.

Perhaps not surprisignly, McCulloch's approach to modeling the yield curve builds upon such an estimation of the discount function. His idea is that the discount function can be estimated piecewise using continously defined and differentiable polynomials on subsections of the sample interval of times to maturity (McCulloch, 1971, 1975). McCulloch suggests representation (2.1), where the discount function is defined as a linear combination of a number of such functions $f()$:

$$\delta(m) = 1 + \sum_{j=1}^{k} a_j f_j(m) \tag{2.1}$$

If k functions are used, k subintervals must be defined and $k+1$ knots, that define the subintervals for the various functions, are required. McCulloch initially sets the functional forms as quadratic polynomials and obtains the estimates of the coefficients a_j by a weighted least squares regression (McCulloch, 1971). Instead of estimating an additional constant a_0 in the above equation, it can immediately be replaced by 1, as $\delta(0) = 1$ by definition.

Since the zero yield at maturity m can be expressed as the average forward rate up until this maturity, and by using the fact that the forward rate $\rho(m)$ is the factor of decay for the exponentially decreasing discount function $\delta(t)$, McCulloch derives an expression for the yield curve $\eta(m)$ as:

$$\eta(m) = -\frac{1}{m}\int_0^m \rho(x)dx = -\frac{1}{m}ln(\delta(m)) \tag{2.2}$$

In a later example, McCulloch makes adjustments to his model to incorporate tax effects. They can have a distorting effect on the curve, severely affecting its shape, if they are not taken into consideration (so McCulloch (1975)). In connection to these changes, he replaces the quadratic functinal forms by cubic polynomials with the motivation that they are more flexible, thus adapt faster to the behaviour of the discount function, and that they provide a smoother representation of the forward curve. However, any other p-degree polynomial, $p-1$ times differentiable, can be used if preferred. The modified model is presented in equation (2.3) where t denotes the estimated tax rate.

$$\eta(m) = -\frac{1}{m(1-t)}\int_0^m \rho(x)dx = -\frac{1}{m}ln(\delta(m)) \tag{2.3}$$

One uncertainty with McCulloch's model is related to the the number of connecting knots one wishes to use for the estimation. If they are too few, the adjustment of the curve to its actual shape becomes slow and a good fit can be difficult to obtain, especially for shorter maturities. If the knots are too many, there will instead be a risk of overfitting and possible outliers can have a large impact on the shape of the curve. When knots are unevenly distributed, the polynomial spline also tends to oscillate and the curve can generate values that are inconsistent with the behaviour of a discount function.

Exponential Splines by Vasicek and Fong

As already stated by McCulloch, the discount function is practically of exponentially decreasing shape. Vasicek and Fong (1982) uses this fact in their claim that the application of polynomials therefore is unsuitable for its estimation. They find that polynomials can be forced to appear exponential, if the number of knots are sufficient, but despite an arbitrarily good fit within sample, polynomials have undesirable asymptotic properties and should not be used for discount function estimation.

The authors find it more appropriate to use exponential splines which according to them fit the data well, give smoother forward rates and show desirable asymptotic characteristics for longer maturities. Their basic model for the price of a bond expressed in terms of the discount function $D()$, given below for the k:th bond among a total of n bonds, is:

$$P_k + A_k = D(T_k) + \sum_{j=1}^{L_k} C_k D(T_k - j + 1) - Q_k - W_k + \epsilon_k, \quad k = 1, 2, \ldots, n \tag{2.4}$$

where,

$D(\cdot)$ = is the discount funtion.
A_k = denotes accrued interest.
P_k = the price of the bond.
T_k = time to maturity measured in half years.
C_k = semiannual coupon rate.
Q_k = denotes tax effects.
W_k = denotes effects for callable bonds.
ϵ_k = residual term, for which $E[\epsilon_k] = 0$.

That the model is expressed in terms of the discount function rather than forward rates or yields is because bond prices are linear in the discount function which simplifies the estimation. When the discount function has been estimated, standard relationships make it easy to derive corresponding forward rates and yields. In order to further simplify computations, Vasicek and Fong impose the following transformation of the argument t:

$$t = -\frac{1}{\alpha}\log(1-x), \quad 0 \leq x < 1. \tag{2.5}$$

and define the function $G(x)$ such that:

$$D(t) = D\left(-\frac{1}{\alpha}\log(1-x)\right) \equiv G(x) \tag{2.6}$$

Since the discount function is exponential, its logarithm is basically represented by a straight line which can easily be estimated by splines. This is the purpose of imposing above transformation that results in the function $G(x)$, which displays desired logarithmic properties. The authors choose to define $G(x)$ as a cubic polynomial and a transformed version of equation (2.4) can be estimated using least squares regression. Without presenting a too high level of details, and with the help of imposing transformation (2.5), the discount function can be expressed in it's original parameter t in the following way:

$$D(t) = a_0 + a_1 exp(-\alpha t) + a_2 exp(-2\alpha t) + a_3 exp(-3\alpha t) \tag{2.7}$$

Vasicek's and Fong's model, which the authors denote *third order exponential splines*, gives an extensive representation of the term structure and leaves few things unsaid. However, exponential splines can give similar properties as polynomials, namely that the yield curve can display a sharp curvature towards longer maturities which is highly unlikely to occur in reality where an asymptotic behaviour generally can be seen (Nelson and Siegel, 1987).

B-Splines

A third example of functional forms of the spline family are the B-Splines.[1] B-Splines have also become popular among practitioners and examples can be seen where B-Splines are fitted to the discount function, the forward curve and the yield curve (Lin, 2002). In Houweling et al. (2001), B-Splines are even used to model the spread curve between a default-free government curve and a defaultable corporate curve, with satisfactory results.

[1] B is in this case short notation for *Basis*.

Lin (2002) presents an example of B-Spline application, as he uses the method to esitmate the yield curve of Taiwanese government bonds. With the notations of Lin each basis function of a B-spline interval can be expressed as:

$$g_s^p(t) = \sum_{i=1}^{s+p+1} \left[\left(\prod_{j-s, j\neq i}^{s+p+1} \frac{1}{T_j - T_i} \right) \right] [max(t - T_j, 0)]^p \tag{2.8}$$

$g_s^p(t)$ is the s:th p-order B-Spline function which is non-zero only if $t \in [T_s, T_{s+p+1}]$, $s = 1, 2, \ldots, m$, and where m is the number of subperiods between $t = 0$ and the longest maturity bond of the sample. This shows that a p-order spline only is defined in $p + 1$ subintervals. Furthermore, $p + m$ functions are needed, defined between a total of $2p + m + 1$ knots. An example of a set of basis functions can be seen in figure 2.1 below.

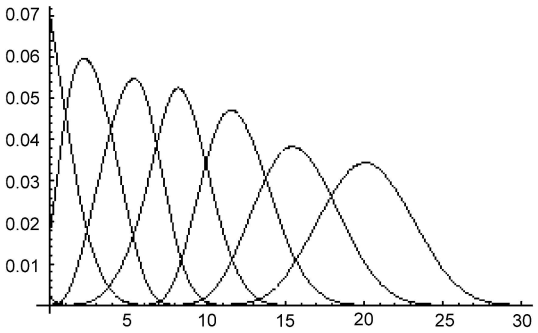

Figure 2.1: A set of basis functions defined on different intervals (Filipović, 2009).

In Houweling et al. (2001) the authors select a somewhat unusual setup as they attempt to jointly model a default-free government curve together with a defaultable corporate spread curve. It is believed by the authors that such setup results in smoother and more realistic spreads than if the two yield curves are modeled separately, where the spread curve is obtained by subtracting the government curve from the corporate yield curve.

Like McCulloch (1971, 1975) and Vasicek and Fong (1982), the authors place their focus on the discount function. Additionally, they categorize the sample bonds based on characteristics (such as credit rating and industry) and define the joint model, decomposed into the two curves, as:

$$D_1(t) = d(t)$$

$$D_c(t) = d(t) + s_c(t), \quad c = 2, 3, \ldots, C \tag{2.9}$$

In equation (2.9), $D_c(\cdot)$ denotes the discount curve of bond category c, $d(\cdot)$ is the government discount curve and $s_c(\cdot)$ is the discount spread curve for bond category c. When applying B-Splines, the representation of equation (2.9) is transformed into:

$$D_1(t) = \mathbf{g}_1(t)\beta_1$$

$$D_c(t) = \mathbf{g}_1(t)\beta_1 - \mathbf{g}_c(t)\beta_c, \quad c = 2, 3, \ldots, C \tag{2.10}$$

In representation (2.10), the authors choose to model the government yield curve by cubic B-Splines. The corporate spread curve however, is modeled by quadratic B-Splines. The motivation behind this choice, is that the spread curve generally displays less complicated shapes than the yield curve and is better represented by quadratic basis functions which reduces flexibility as well as degrees of freedom. The unknown spline weights $\beta_1, \beta_2, \ldots, \beta_C$ are estimated by restricted feasible generalized least squares.

One of the main disadvantages of the B-Spline methodology is the effect of the chosen knots on the result. Lin (2002) selects points *ad hoc* which could cause unrealistic curve shapes if placed inappropriately. Another drawback is the asymptotic behaviour of the method. Out of sample estimates cannot be obtained and due to this, despite satisfactory results from the application of B-Splines, Lin expresses concerns and mentions that an exponential spline representation as in Vasicek and Fong (1982), is a more appropriate choice.

2.1.2 The Nelson-Siegel-Svensson Models

Among the most popular methods for yield curve estimation, even applied at a number of central banks, are the Nelson-Siegel model and its extension by Svensson (Nelson and Siegel, 1987; Svensson, 1994). The desire of Charles Nelson and Andrew Siegel was to create a model that fits the whole family of basic shapes of the yield curve and not just creates a local approximation to observations. They also wanted a model that could be used for prediction beyond the sample range, something they found that many existing approaches failed to do.

The authors instead believed that the class of functions which can create the typical shapes of the yield curve are associated with solutions to differential equations. Their suggested model is based on the idea that the instantaneous forward rate of maturity m, here denoted $r(m)$, can be explained by such solution (equation 2.11); namely that to a second order differential equation.

$$r(m) = \beta_0 + \beta_1 \exp(-\frac{m}{\tau}) + \beta_2 \left[(\frac{m}{\tau}) \exp(-\frac{m}{\tau}) \right] \tag{2.11}$$

The yield to maturity $R(m)$ can further be derived from the relation between forward rate and yield:

$$R(m) = \frac{1}{m} \int_0^m r(x)dx \tag{2.12}$$

which finally gives the expression for yield to maturity:

$$R(m) = \beta_0 + (\beta_1 + \beta_2) \left[1 - \exp(-\frac{m}{\tau}) \right] \frac{m}{\tau} - \beta_2 \exp(-\frac{m}{\tau}) \tag{2.13}$$

In equations (2.11) and (2.13) above, τ denotes a time constant that determines the rate of which the regressor values (simplified as $m\exp(-m)$ and $\exp(-m)$) decay to 0. A small τ implies a fast decay and a very flexible adjustment of the curve whereas a large τ implies a slow decay which would fit curvature for longer maturities where sharp shifts in the curve are rare.

The coefficients β_0, β_1 and β_2 have relatively intuitive interpretations in the representation of the forward curve (equation (2.11)). β_0 is the long term component to which the curve converges as the maturity grows large, β_1 is the short term component that quickly becomes insignificant and β_2 is the medium term component that starts at zero, initially grows with the maturity m, and decreases again as the exponential term in the regressor takes over. β_2 thus creates a hump shape in the yield curve.

The creators claim that the suggested model gives a flexible and parsimonious representation of all shapes generally associated with the yield curve: monotonic, humped and s-shaped. In their application of the model on US Treasuries, they found that it could explain up to 96% of the variation in yields.

Lars Svensson (1994) wanted to further increase the flexibility and improve the fit of the model suggested by Nelson and Siegel, in his attempt to model and analyze Swedish forward rates. To achieve this he introduced a second hump shape, thus adding the two parameters τ_2 and β_3 to the original formula, which gives the representation of the instantaneous forward rate:

$$r(m) = \beta_0 + \beta_1 \exp(-\frac{m}{\tau_1}) + \beta_2 \left[(\frac{m}{\tau_1}) \exp(-\frac{m}{\tau_1}) \right] + \beta_3 \frac{m}{\tau_2} \exp(-\frac{m}{\tau_2}) \tag{2.14}$$

Furthermore, the zero yield curve is then given by:

$$R(m) = \beta_0 + \beta_1 \left(1 - exp(-\frac{m}{\tau_1}) \right) \left(-\frac{\tau_1}{m} \right) + \beta_2 \left[\left(1 - exp(-\frac{m}{\tau_1}) \right) \frac{\tau_1}{m} - \exp(\frac{m}{\tau_1}) \right]$$

$$+\beta_3 \left[\left(1 - exp(-\frac{m}{\tau_2}) \right) \frac{\tau_2}{m} - \exp(\frac{m}{\tau_2}) \right] \tag{2.15}$$

Coefficients are estimated with linear least squares by Nelson and Siegel and with Maximum Lieklihood by Svensson. Svensson however emphasizes that other estimation techniques also can be applied. An illustration of curves that can be obtained by the two estimation techniques can be seen below in figure 2.2 where Korean government zero yields are used as base for the estimation. The figure shows that both estimation techniques yield good results when data of high quality is used.

Various studies show that the two models of Nelson, Siegel and Svensson are far better than others, something which was also found in a comparison of estimation techniques of the UK goverment curve by Ioannides (2003). Shortcomings of the models do however exist, for example when the sample data is of poor quality. Irregularly placed observations can force the models to create unrealistic humps and depending on the estimated parameters β_1 and β_2, the curve can sometimes take off to infinity as the maturity goes to zero.

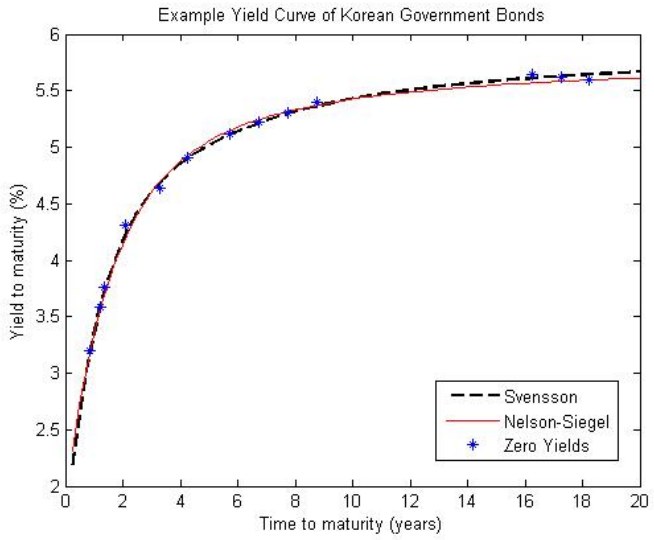

Figure 2.2: Example of yield curve estimation using the Nelson-Siegel and Svensson Models. Curves are estimated for maturities between 3 months and 20 years and are based on 13 zero yields of Korean goverment bonds on December 14th 2009. Both curves display a good fit to the underlying data, smooth curvature and desired asymptotic properties.

2.1.3 Estimating the Yield Curve with the means of PCA

Prinicpal Component Analysis (PCA) has become a popular estimation method when dealing with interest rates. It can be used to derive the driving components of the stochastic movements of the term structure and can be applied both on the forward curve and directly onto the yield curve (Filipović, 2009). The transformed data representation in terms of the Principal Components for example helps identifying key drivers of risk and thus facilitates portfolio risk management. For a detailed technical description of PCA the reader is referred to chapter 3 section 3.5, where the full procedure is explained.

The high colinearity that characterizes the term structure of interest rates is what makes it suitable for PCA. Dimension reduction, one of the main purposes of PCA, can be done successfully after which the variation in the term structure can be explained by just a few components. The typical PCA representation of the yield curve contains three components which explain a significant part of the total variation. These components are referred to as the *trend*, *tilt* and *curvature* components. For interest rate modeling in industrialized currencies, more than 95% of the variation in the data set is explained by these components, whereas smaller Emerging Market currencies might need a larger set of components, due to a higher level of illiquidity (Alexander, 2009).

2.2 Imputation of Missing Data

As discussed in the introductory chapter (section 1.2), missing values cause large problems for statistical analysis of univariate and multivariate data sets. Estimating these values with any method can lead to biased estimates of sample covariances, means and standard deviations as well as causing incorrect conclusions drawn from various statistical analyses (Stanimirova et al., 2007). However, excluding them can create just as much error and loss of prediction power (Heitjan, 1997). Many statistical analysis methods require complete sample data, which gives the practitioner no other choice but to fill gaps with various types of backfilling methods (Zhang et al., 2008).

Common approaches for filling unobserved or missing values are for example variations of single or multiple imputation together with the Expectation-Maximization algorithm (generally referred to as the EM-algorithm). Another possibility is to assume a distribution based on the observed data from which values are simulated. In this section, a number of methods found relevant for this study are discussed, many of which rely on the statistical tool Principal Component Analysis (PCA).

2.2.1 An Evolutionary Algorithm

In a somewhat creative approach introduced by Figueroa-Garcıa et al. (2008), univariate time series with missing data are filled with what the authors refer to as an iterative evolutionary algorithm where parallels can be drawn to the field of genetics. The algorithm is based on the minimization of a *fitness function* derived from a time series's autocorrelation function, mean and variance.

By using the largest and most recent complete subset l of the original time series x_t, the sample autocovariance function of the set is defined as:

$$\hat{\gamma}(h)^l = \sum_{t=n_1}^{n_2-|h|} \frac{(x_{t+|h|} - \bar{x})(x_t - \bar{x})}{n}, \quad (n_1 + n_2) < h < (n_2 - n_1), \quad n_1, n_2 \in T \tag{2.16}$$

In equation (2.16), t is the general time index of the time series x_t and (n_1, n_2) are the lower and upper bounds of an index vector ν containing all positions of missing values in x_t. By computing the mean and variance for x_t, based only on observed values and denoted $\bar{x^a}$ and $Var(x^a)$, the fitness function to be minimized can be expressed as:

$$\kappa = \sum_{h=1}^{H} [|\hat{\rho}(h)^l - \rho(h)|] + |\bar{x^a} - \bar{x}| + |Var(x^a) - Var(x)| \tag{2.17}$$

where $\hat{\rho}(h)^l$ is the autocorrelation function defined on the subset l given by standard relation with the autocovariance function.

The authors apply their model onto a set of weather time series and obtain satisfactory results. However, applied on a larger set of data, the correlation structure is ignored by only placing focus on univariate characteristics of the time series (within the scope of this study, interest rate time series can be mentioned as an appropriate example). Additionally, despite the authors' attempt to create an understandable algorithm by descriptive parallels, it becomes somewhat complicated for the reader to follow.

2.2.2 Sequential Local Least Squares (SLLS)

In DNA Microarray Technology, missing values in time series is a common occurence. Despite this fact, most statistical algorithms require 100% coverage. To solve the issue, Zhang et al. (2008) suggest a sequential local least squares estimator (SLLS) where k-neighbouring genes are used as estimation base.

The sample genes are initially separated into two groups: those that are complete and those that contain missing values. Thereafter the procedure starts with the gene with the lowest non-zero fraction of missing values, referred to as the the *target gene*. The target gene is regressed upon k neighbouring (complete) genes which are selected according to their similarities with the target gene. With g_i denoting the target gene, $\mathbf{g}_s$ the k neighbouring genes and $\mathbf{x}^T$ their coefficients, the regression can be expressed as:

$$g_i = x_1 g_{s_1} + x_2 g_{s_2} + \ldots + x_k g_{s_k} = \mathbf{x}^T \mathbf{g}_s \tag{2.18}$$

The target gene is thereafter redefined as the gene with second highest fraction of missing values and the procedure is repeated. The previous target gene has now been transferred to the group of genes containing complete data, and could thus be part of the regression base when values are imputed in the subsequent target gene. Hence, only the first target gene is certainly based upon originally full time series whereas following regressions also might include already imputed genes. The optimal number of neighbouring genes, k, is not defined as a constant but instead depends on the current target gene. The authors suggest an automatic parameter selection algorithm to determine the value of k.[2]

The procedure continues until all missing values have been filled. However, a lower boundary, that defines how high the rate of originally missing values can be for a gene to be used for recreation of values in other genes, is set by the authors. This threshold rate, r_0, is defined as:

$$r_0 = \frac{\sum_{i=1}^{m_2} c_i}{m_2 n} \tag{2.19}$$

where n is the total number of genes, m_2 is the number of genes with at least one missing value and c_i the number of missing values in gene i.

The presented algorithm is relatively intuitive for the user but it has one main drawback. It requires at least k complete genes for the initial target gene to be filled and can therefore not be used on data sets where all series have a least one missing observation.

[2]The selection algorithm is not presented in this section. The reader is instead referred to the original source of Zhang et al. (2008) for full details. However, the authors do emphasize the importance of the parameter and compare it with the number of Principal Components that are used in PCA related imputation methods.

2.2.3 PCA Based Backfilling Routines

A substantial number of backfilling routines are related to the statistical method Principal Component Analysis (PCA), of which a few relevant examples are presented below. For complete technical details of PCA the reader is referred to section 3.5 in chapter 3.

Singular Spectrum Analysis (SSA)

Kondrashov and Ghil (2006) present a data adaptive parametric method referred to as Singular Spectrum Analysis (SSA) in their attempt to backfill missing values in spatio-temporal data sets. The method is iterative and uses both spatial and temporal correlations. When applied on both simulated and real data the method produces good results, for single missing values as well as longer continous gaps.

The method is based on an eigenvalue decomposition of a lag-covariance matrix, C_x, obtained from the original data series $X_t : t = 1, \ldots, N$, and embedded in an $M \times M$ vector space. The eigenvectors with corresponding eigenvalues, denoted $\mathbf{E}_k$ and λ_k, explain the partial orthogonal variances of X_t where the sum of all eigenvalues gives its total variance. By projecting the time series onto each eigenvector (here also referred to as EOF:s for Empirical Orthogonal Function) the principal components are obtained, with the help of which the time series can be reconstructed through the relation:

$$R_\kappa(t) = \frac{1}{M_t} \sum_{k \in \kappa} \sum_{j=L_t}^{U_t} A_k(t - j + 1) E_k(j) \tag{2.20}$$

In the above equation κ defines the set of EOF:s that are used for reconstruction, M_t is a normalization factor and L_t and U_t denotes lower and upper bound, all of which are t-dependent. $A_k(\cdot)$ denotes the positions of the principal component $\mathbf{A}_k$.

After first centering the data set by its unbiased mean estimate and initially replacing missing values by zeros, the algorithm starts off by using the first EOF, $\mathbf{E}_1$, and reconstructing the data by applying relation (2.20). This gives the first reconstructed component ($\mathbf{R}_1$) of the data and non-zero values are now imputed in positions for originally missing values. The estimation is thereafter repeated, still using only $\mathbf{E}_1$. When convergence has been reached, the second EOF, $\mathbf{E}_2$, is added and missing values are again reconstructed until convergence using the two vectors $\mathbf{E}_1$ and $\mathbf{E}_2$. So the algorithm continues until convergence is reached as all the first κ EOF:s are used.

One of the weaknesses of the SSA algorithm presented by Kondrashov and Ghil (2006) is how to define the parameters κ and M. Optimal values differ between data sets and the authors find values for each set through cross-validation. The method does however seem to yield good results, even for longer continuous gaps in the original series.

EM-PCA and EM-SPCA

Despite the popularity of Principal Component Analysis, the method does have a number of shortcomings. It is highly sensitive to outliers and does not treat missing values sophisticatedly, two properties which are common in experimental data. Stanimirova et al. (2007) present a robust PCA, designed with the purpose of performing PCA on datasets where missing values and outliers occur, which also imputes values. They refer to the method as Expectation-Maximization Spherical PCA, or simply *EM-SPCA*. [3]

The data set $\mathbf{X}$ is initially centred around its robust centre, here defined as the L_1-median, and thereafter projected onto a subspace; a hypersphere with the radius 1. The projected version of a vector $\mathbf{x}_i$ (denoted $\mathbf{x}_i^p$) can be expressed as:

$$\mathbf{x}_i^p = \frac{\mathbf{x}_i - \mu_{L_1}(\mathbf{X})}{||\mathbf{x}_i - \mu_{L_1}(\mathbf{X})||} + \mu_{L_1}(\mathbf{X}) \tag{2.21}$$

where μ_{L_1} is the L_1-median centre and $||\cdot||$ denotes the Euclidean norm, here used as a weight factor. Classical PCA is then performed on the projected data, from which an estimate of $\mathbf{X}$ can be obtained with the help of robust loadings and factors, and missing values can be filled. The complete algorithm roughly follows the steps presented below.

After replacing originally missing values by the row and column means the algorithm is initiated with robust centering (as mentioned above). The data set is projected and an eigenvalue decomposition is performed on this projected set, from which values are reconstructed with obtained principal components. The authors suggest to use the s first principal components which explain about 80% of the variation in the data. The algorithm is then iterated until convergence is reached, which is measured with the help of the objective function defined for the k:th iteration as:

$$SS_k = \sum_p \sum_q (x_{p,q})^2, \quad p, q \in \textit{missing elements} \tag{2.22}$$

The algorithm is similar to "normal" EM-PCA where the procedure is identical but instead of robust loadings and factors, standard loadings and factors are used. The number of necessary iterations varies depending on the amount and positioning of missing values but convergence is generally reached faster if the correlation structure in the data is well defined. However, no matter the number of iterations, one of the strengths of the EM-algorithm is that it always converges to its optimal solution. In the case where outliers exist, the authors show that EM-SPCA outperforms standard EM-PCA.

Regression-based PCA

Another suggestion of PCA related backfilling procedures is presented by Grung and Manne (1998). They have selected a regression based method in which factors and factor loadings are obtained by a two-step regression. Although corresponding factors and loadings do not display the property of orthogonality, this can easily be obtained by one additional operation.

[3] The EM-algorithm will not be further explained in this study. For a specific explanation readers can for example turn to Moon (1996) found in the list of references.

Consider the $M \times N$ data matrix $\mathbf{Y}$ with some values Y_{ij} missing. Further consider the matrix $\mathbf{X}$ which contains the values from $\mathbf{Y}$ but where missing observations have been replaced by zeros. The relation between $\mathbf{Y}$ and $\mathbf{X}$ can be described with the help of a third matrix, namely the index matrix $\mathbf{C}$ for which $C_{ij} = 1$ where Y_{ij} is known and zero otherwise. The relation between all three matrices can be expressed by $X_{ij} = C_{ij}Y_{ij}$ and for the matrix $\mathbf{C}$, $C_{ij}^2 = C_{ij}$ also holds. With the help of this notation a set of regressions can be performed as below.

Define the matrices $\mathbf{A}^{(i)}$ and $\mathbf{B}^{(i)}$, where $A_{jk}^{(i)} = C_{ij}p_{jk}$ and $B_{ik}^{(j)} = t_{ik}C_{ij}$. The variables p_{jk} and t_{ik} are values of the factors $\mathbf{t}^{(i)}$ and the loadings $\mathbf{p}^{(j)}$. These values can be obtained by the following two regressions which are based on minimizing an objective function F (not presented here):

$$\mathbf{t}^{(i)} = \mathbf{x}^{(i)}\mathbf{A}^{(i)}(\mathbf{A}^{(i)^T}\mathbf{A}^{(i)})^{-1} \tag{2.23}$$

$$\mathbf{p}^{(j)^T} = (\mathbf{B}^{(j)^T}\mathbf{B}^{(j)})^{-1}\mathbf{B}^{(j)^T}\mathbf{x}^{(j)} \tag{2.24}$$

In the above equations, $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are the i:th row and the j:th column of the matrix $\mathbf{X}$ respectively. As previously mentioned, the factors and factor loadings obtained from above regression are not orthogonal. By estimating $\mathbf{Y}$ with $\hat{Y}_{ij} = \sum_k t_{ik}p_{jk}$, where missing values now have been filled with respect to the minimized objective function F, standard PCA can be performed on this estimate and orthogonality properties will be obtained.

The authors show that the method works well on data sets with fractions of missing values up to 25%, but that its performance is heavily dependent on their positions. If more than 25% of the values are missing, convergence which again is measured with the help of the objective function, takes a long time to reach.

Chapter 3

Theoretical Framework

In this chapter a theoretical framework, found necessary as foundation for this study, is presented. Relevant procedures for term structure estimation are described together with the GARCH(1,1) process, commonly used for financial time series modeling. The final part of the chapter covers the statistical tool Principal Component Analysis (PCA).

3.1 The Bootstrap Method

To obtain a good estimate of the term structure of interest rates might seem simple, if only one had a set of zero-coupon bonds at hand. Since zero yields are difficult to observe directly in the market they must be estimated and a common means for this is the Boostrap method. It provides a simple and straightforward approach and is based on iterative calculation of zero yields from coupon bonds with increasing maturiy. An illustrative example is given by Nawalka and Soto (2009) and is recited below.

Consider a set of N bonds that pay semiannual coupons where the shortest maturity bond reaches maturity within six months. Let $t_1, t_2, t_3, \ldots$ be the times for coupon payments, where t_1 denotes six months from today. The price of this bond, with continous compounding, is given by:

$$P(t_1) = \frac{C_{t_1} + F_{t_1}}{\exp(y(t_1)t_1)} \tag{3.1}$$

where C_{t_1} and F_{t_1} denote the semiannual coupon payment and the face value of the bond respectively. $y(t_1)$ is the annualized six month zero-coupon yield, corresponding to the first determined point on the yield curve, and which easily can be obtained with the help of equation (3.1).

When the six month zero yield has been obtained, the one year zero yield (where t_2 denotes one year) can be calculated from relation (3.2). $P(t_2)$ here denotes the price of a one year coupon bond, C_{t_2} the semiannual one year coupon paid at times t_1 and t_2, F_{t_2} the face value of the bond paid after one year and $y(t_2)$ the annualized one-year zero-coupon yield:

$$P(t_2) = \frac{C_{t_2}}{\exp(y(t_1)t_1)} + \frac{F_{t_2} + C_{t_2}}{\exp(y(t_2)t_2)} \tag{3.2}$$

Typically assumed when the Bootstrap method is applied, is that the zero curve is linear between determied points on the curve. Zero rates for maturities that lie in between these points can thus be obtained by simple linear interpolation. One other common characteristic is to define the yield curve as horizontal outside of maturities framed by determied points. This approach can in fact be seen as the simplest of examples on how to obtain a reptresentation of the yield curve. However, the only time when the Bootstrap method with linear interpolation can be of good use, is when one has a relatively large set of bonds evenly distributed over desired maturities (Hull, 2006).

3.2 Smoothing Splines

A set of functions, piecewise defined between a number of connecting knot points that together represent an arbitrary curve, are more commonly known as *splines* and have previously been disussed in chapter 2. The functional form of a spline can vary and a number of popular examples are found in section 2.1.1. Splines sometimes suffer from the problem of overfitting and can display undesired rapid oscillations or too high adjustment to outliers. To prevent such properties, the spline can be smoothed by introducing a penalty function, which penalizes excess variability of the fitted curve. A descriptive example by Fisher et al. (1995) is presented below.

Consider the example of the term structure of interest rates and let $h(\tau)$ denote an arbitrary spline function representing it, where τ denotes the time to maturity. Furthermore let $h(\tau)$ be related to the discount function $\delta(\tau)$ through a functional transformation $g(h(\cdot), \tau) \equiv \delta(\tau)$. The application of a smoothing spline will here penalize excess variability in the discount function so that a large number of knot points still can be used without causing overfitting or oscillation. The penalty function is defined as a constant multiplied by the integral (over all observed times $0, \ldots, T$) of the squared second derivative of the spline function and thus takes the following form:

$$\lambda \int_0^T h''(\tau)^2 d\tau \tag{3.3}$$

λ is sometimes referred to as the "smoothing factor" and decides to which extent the curve should adjust to the data as oppose to being smooth. $\lambda = 0$ implies no smoothing and corresponds to the application of a normal interpolating spline whereas with a growing value of λ the curve eventually converges to the least squares fit of a straight line.

With the penalty function defined as in (3.3) the optimal smoothing solution is found by simply minimizing the residual sum of squares plus the penalty function. In the case of term structure estimation, and with the help of the current prices p_i, coupon vectors c_i (one for each bond, containing all its coupon payments) and the functional transformation $g(\cdot)$, the expression to minimize over all bonds $i = 1, \ldots, n$ takes the form:[1]

$$\min_{h(\tau) \in H} \left[\sum_{i=1}^{n} p_i - c_i^T g(h(\cdot), \tau_i))^2 + \lambda \int_0^T h''(\tau)^2 d\tau \right] \tag{3.4}$$

[1] Furthermore, H denotes the subspace of all possible functions on R_+, twice differectiable with second derivatives that integrate to a finite value.

By smoothing the spline representation, previously mentioned shortcomings of standard splines can be reduced and a more stable curve estimate can be obtained. Due to the reduction in estimated parameters, the risk for poor curve fitting due to ad hoc parametrisation is also reduced. Though the example presented in this section is based on a term structure estimation, a smoothing spline can of course be applied on any other curve that one desires to estimate with the method. Figure 3.1 illustrates a smoothing spline fitted to a higher order polynomial with two different values of the smoothing factor.

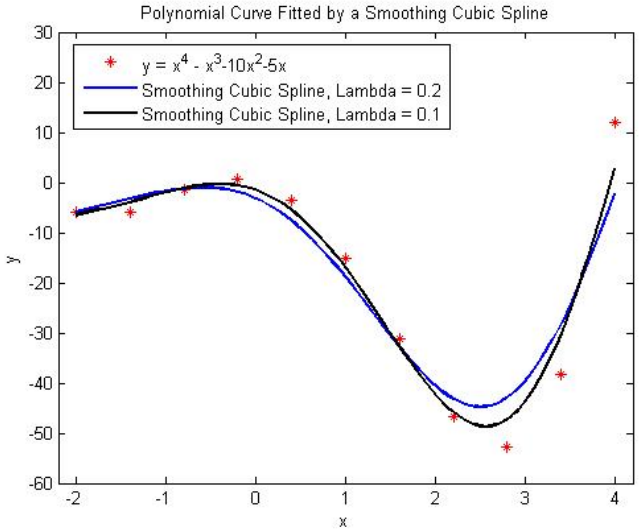

Figure 3.1: A smoothing cubic spline is fitted to the higher order polynomial $y = x^4 - x^3 - 10x^2 - 5x$ with the two different smoothing factors $\lambda_1 = 0.1$ and $\lambda_2 = 0.2$.

3.3 GARCH Models

Generalized Autoregressive Conditional Heteroskedasticity (GARCH) is a time series model introduced by Bollerslev (1986). Similar to the extension of AR processes to ARMA, the GARCH process is an extension of its predecessor, the ARCH process, first introduced by Engle (1982). The main purpose of the GARCH model is to predict conditional variances of a random variable. The more flexible lag structure and parsimonious data representation that GARCH permits for are found among its strengths (Bollerslev, 1986). Tsay (2005) also emphasizes how GARCH models are designed to capture volatility clustering of returns, a common phenomenon in high frequency financial data.

Within the econometric field, GARCH and ARCH models have become standard tools for volatility related questions, such as forecasting and analysis (Engle, 2001). Periods of high or low volatility have important implications for financial risk measurement as well as for pricing of financial products. Since volatility clustering effects is typically seen in short term data, GARCH is generally applied on daily or intraday data (Alexander, 2009). Additionally, GARCH displays a mean reverting property of volatility and is designed to deal with heteroskedastic error terms, another issue that normally arises when dealing with financial time series (Engle, 2001). A widespread number of applications of GARCH exist and there

are various extensions of the model.

3.3.1 Symmetric Normal GARCH

GARCH(1,1), also denoted *Symmetric Normal GARCH*, is the simplest and most robust in the family of volatility models. The general representation of a time series X_t which follows a GARCH(1,1) process is (Penzer, 2007):

$$X_t = \sigma_t \epsilon_t$$

$$\sigma_t^2 = c + bX_{t-1}^2 + a\sigma_{t-1}^2, \quad F_{t-1} \tag{3.5}$$

In the above expressions, X_t is zero-mean distributed and σ_t^2 denotes its conditional variance at time t, given the information set F_{t-1}, containing all relevant information from the infinite past at time t. ϵ_t is a series of iid random variables where $E[\epsilon_t] = 0$ and $Var(\epsilon_t) = 1$. $[a, b, c]$ are the estimated GARCH coefficients defined such that $a, b \geq 0$, $c > 0$ and $a + b < 1$. In financial applications, X_t is sometimes said to be the market shock, generally given by the mean deviation of the return series (Alexander, 2009). Although, assuming that the returns themselves are zero-mean distributed and follow a GARCH(1,1) model is also a common approach. Furthermore, the GARCH parameters a, b and c are estimated by maximizing the log-likelihood function (equation 3.6), found to be a systematic approach to obtain the best possible fit (Engle, 2001):

$$-2lnL(\theta) = \sum_{i=1}^{T} \left(\ln\left(\sigma_t^2\right) + \left(\frac{\epsilon_t}{\sigma_t}\right)^2 \right) \tag{3.6}$$

Despite treatment of volatility clustering, the symmetric normal GARCH is not always an optimal model to describe financial time series. One reason for this is correlation clustering, which refers to increasing correlations between different assets when financial markets are more volatile. Nor does symmetric normal GARCH take the direction of returns into consideration and thus ignores the possibility that periods during which markets decline, have a larger impact on volatilities than upwards moving markets do, referred to as the "leverage effect", typically seen for equity and commodity markets and for which credible evidence exists (Alexander, 2009; Tsay, 2005).

To fully capture the characteristics of correlation clustering, a multivariate extension of the model is required. The extension to a multivariate GARCH model is however complicated. With an increasing number of time series, the number of parameters to be estimated increases fast and a growing covariance matrix makes positive definite estimation more and more difficult. Some models allow for low level multivariate GARCH estimation, such as the BEKK-representation but it remains that the application of multivariate GARCH is very difficult in practice.[2] To further reflect the leverage effect, an asymmetric GARCH model such as A-GARCH or GJR-GARCH would be a better choice (Alexander, 2009).

[2]The BEKK-representation is fully presented in Alexander (2009), page 165.

3.4 The Kalman Filter

The Kalman filter is a recursive process that provides a procedure for filtering of time-discrete linear data series (Welch and Bishop, 1995). It was invented by Rudolf E. Kalman during the 1960:s and has become a much appreciated tool. The Kalman filter separates the state variable x from a prediction error which is believed to be caused by the estimation measurement z. One of the strengths with the Kalman filter is that it can be used also in the presence of missing data; the procedure simply builds over such period in the data set.

Consider the discrete-time state variable $x \in R^n$ together with a measurement $z \in R^m$, described by the expressions:

$$x_k = Ax_{k-1} + Bu_{k-1} + w_{k-1} \tag{3.7}$$

$$z_k = Hx_k + v_k \tag{3.8}$$

where w_k and v_k are the process and the measurement noise for which $w \sim N(0, Q)$ and $v \sim N(0, R)$, where R and Q are the process and measurement covariance matrices. Furthermore, the matrix A defines the relation between the current and previous state variable (x_k and x_{k-1}) and the matrix H relates the state variable to the measurement z. (u denotes an optional control input related to the state variable x through the matrix B.)

The Kalman filtering equations for a series of the described character are then given by:

Time update equations:

$$\hat{x}_k^- = A\hat{x}_{k-1} + Bu_{k-1} \tag{3.9}$$

$$P_k^- = AP_{k-1}A^T + Q \tag{3.10}$$

Measurement update equations:

$$K_k = P_k^- H^T \left(HP_k^- H^T + R\right)^{-1} \tag{3.11}$$

$$\hat{x}_k = \hat{x}_k^- + K_k \left(z_k - H\hat{x}_k^-\right) \tag{3.12}$$

$$P_k = \left(I - K_k H\right) P_k^- \tag{3.13}$$

P_k^- and P_k above, are the *a priori* and *a posteriori* estimate error covariance and K_k is referred to as the Kalman gain.

The Kalman equations are thus divided into two types. The time updating equations create a prediction of the next state x_k by means of previously known information. When this has been done, the measurement equations are applied to correct the initial measure of the k:th state from the measurement error by first computing the Kalman gain and thereafter adding it to the *a priori* estimate of x_k to obtain its *a posteriori* estimate.

3.5 Principal Component Analysis

Principal Component Analysis (PCA) is a powerful and flexible statistical tool dating back to the beginning of the 20^{th} century. Its main purpose is to reduce the dimensionality of a large set of data while retaining as much variability as possible, mentioned among others by Jolliffe (2002) and Alexander (2009). The model is commonly mistaken for a factor model and even though similarities are many, Jolliffe claims that the main difference is the explicit underlying model of a factor model, which does not exist in PCA.

The fundamental idea of PCA is to transform a data set into a new set of variables called Principal Components. These variables are orthogonal and ordered such that the first component explains more of the varialbility within the original data than the second principal component, which in turn explains more than the third principal component and so on (the cumulative explained variability is generally high already after just a few components). Further analysis can thereafter be carried out directly on the principal components, something that significantly simplifies computations as the size of the data set has been reduced. As a pure statistical tool, the derived components do not need to have an interpretation but in the case where a lot of colinearity exists, interpretations can be found.

PCA has been applied in a large number of contexts and Jolliffe (2002) mentions an impressive number of 14 fields of application in the introducing chapter of his book. He also presents elaborated examples for spatio-temporal analysis, how PCA can explain the variability in human anatomy and how it can be used for demographic mapping of living arrangements among elderly. PCA has become a very important tool also in the field of finance as well as for data imputation, something which will be applied in this study.

3.5.1 Deriving the Principal Components

The Principal Component representation (described by Alexander (2009)) is defined as follows: Let $\mathbf{X}$ be a $T \times n$ matrix of random variables and let $\mathbf{V}$ be its corresponding $n \times n$ covariance matrix (or correlation matrix if preferred). Furthermore, let $\mathbf{W}$ be the $n \times n$ orthogonal matrix of $\mathbf{V}$. The $T \times n$ matrix $\mathbf{P}$, where columns that correspond to principal components (as an exact linear combination) of $\mathbf{X}$, is then given by the relation:

$$\mathbf{P} = \mathbf{XW} \tag{3.14}$$

or analogously,

$$\mathbf{X} = \mathbf{PW}^T \tag{3.15}$$

where $\mathbf{W}^T = \mathbf{W}^{-1}$ since $\mathbf{W}$ is an orthogonal matrix.

By selecting the first k columns of $\mathbf{P}$ and $\mathbf{W}$ and thus creating $\mathbf{P}^*$ and $\mathbf{W}^*$, an approximantion of $\mathbf{X}$ can be obtained through the relation:

$$\mathbf{X}^* = \mathbf{P}^*{\mathbf{W}^*}^T \tag{3.16}$$

where $\mathbf{P}^*$ hopefully explains as much as possible of the variation in the original data.

The principal components are thus retrieved through an eigenvalue decomposition of a covariance or correlation matrix of a set of observed variables. The i:th eigenvalue λ_i of $\mathbf{V}$ is obtained by taking the sum of squares of each element in the corresponding i:th principal component. Given that the total variation is explained by the sum of the eigenvalues of $\mathbf{V}$, one can easily derive an expression for the fraction of the variability which is explained by the first k components (Alexander, 2009):

$$\frac{\lambda_1 + \ldots + \lambda_k}{\lambda_1 + \ldots + \lambda_n} \tag{3.17}$$

Example

Consider the three randon variables (z_1, z_2, z_3) together defining the matrix Z. Observations are available of all three variables during a fixed time period $t = 1, 2, \ldots, 10$, as shown below:

$$Z \equiv (z_1, z_2, z_3) = \begin{pmatrix} -0.330 & -0.255 & -0.222 \\ -0.200 & -0.185 & -0.162 \\ 0.040 & 0.044 & 0.058 \\ 0.090 & 0.105 & 0.108 \\ 0.110 & 0.125 & 0.128 \\ 0.060 & 0.105 & 0.098 \\ 0.140 & 0.105 & 0.098 \\ 0.110 & 0.065 & 0.058 \\ -0.130 & -0.165 & -0.172 \\ 0.110 & 0.055 & 0.008 \end{pmatrix}$$

(z_1, z_2, z_3) are all zero-mean distributed by definition but if the data does not display this property it should be centered around its mean before PCA is performed. If the practitioner desires to reconstruct the original data from its transformed version, it is simple to add the originally estimated mean on top of the vectors after reconstruction which will result in the original set of observations.

The covariance matrix of Z is given by Σ:

$$\Sigma = \begin{pmatrix} 0.0261 & 0.0224 & 0.0201 \\ 0.0224 & 0.0205 & 0.0189 \\ 0.0201 & 0.0189 & 0.0177 \end{pmatrix}$$

Performing Principal Component Analysis on the covariance matrix Σ of Z and using notations as defined above gives us the matrix P of principal components:

$$P = \begin{pmatrix} -0.4707 & -0.0410 & 0.0039 \\ -0.3169 & -0.0001 & -0.0057 \\ 0.0813 & -0.0180 & -0.0074 \\ 0.1733 & -0.0274 & -0.0006 \\ 0.2078 & -0.0301 & -0.0011 \\ 0.1490 & -0.0432 & 0.0120 \\ 0.1998 & 0.0158 & -0.0063 \\ 0.1371 & 0.0287 & -0.0077 \\ -0.2663 & 0.0529 & -0.0003 \\ 0.1054 & 0.0625 & 0.0133 \end{pmatrix}$$

together with the orthogonal matrix W of Σ:

$$W = \begin{pmatrix} 0.6350 & 0.7377 & -0.2292 \\ 0.5697 & -0.2468 & 0.7839 \\ 0.5217 & -0.6284 & -0.5770 \end{pmatrix}$$

With this representation it is now a simple task to reconstruct the original data by executing the expression $\mathbf{X} = \mathbf{P}\mathbf{W}^T$ (equation 3.15) or with a reduced set of principal components through $\mathbf{X}^* = \mathbf{P}^*\mathbf{W}^{*^T}$ (equation 3.16).

3.5.2 PCA in Practice

The number of components one should use in equation (3.16) to estimate the matrix $\mathbf{X}^*$ depends on how much of the variability one wants to explain, as well as on the colinearity within the original data set. In highly colinear sets, already the first few principal components can give the desired level of explanation. In fact, if too many components are used, one might increase the risk of introducing noise into the equation (Alexander, 2009). What one must be aware of is that despite its strengths as a statistical tool, PCA is in general a non-robust method. Outliers can highly affect the principal components, something which is emphasized by Stanimirova et al. (2007).

To apply PCA on the correlation or the covariance matrix is yet another choice to be made by the practitioner. PCA performed on the correlation matrix is less dependent on the units in which the original data set is expressed and therefore produces more comparable results between different sets of data. The covariance matrix on the other hand, is highly unit dependet but gives results that are easier to interpret. Transforming between one or the other when the components are computed is however not an easy task since the eigenvectors of a covariance matrix has no simple relation to those of its corresponding correlation matrix. One should therefore carefully consider this choice at forehand (Jolliffe, 2002). Within the practice of finance, PCA is often used to model returns of portfolios and profit and loss cash flows but also for assessment of financial risk. The covariance (or correlation) matrix of equity returns or interest rate changes is then used as input.

Chapter 4

Method

This chapter presents the methods that have been applied throughout this study. Its aim is to provide the reader with enough information to understand how the work has been conducted and to be able to follow each step of the process.

4.1 Estimating the term structure

Historical yield notations are downloaded for a number of bonds for each of a number of different issuers. Let $I = 1, \ldots, k$ denote the various issuers. Yields are observed during the five year period 2005.05.03 - 2010.04.30. Each time series of yields is inspected and outliers are removed manually. Due to increasing volatility towards the end of a bond's lifetime, approximately the last month of notations before maturity is removed for all of the sample bonds that reaches expiry during the observed time window. As issuer specific term structures are desired, each issuer is now treated separately.

The term structure of interest rates for the issuer I, at a fixed day of the observed period, is given by the relation between the zero yield of each of the issuers's bonds and its current time to maturity. Since observed notations correspond to coupon bearing yields, these must first be transformed into zero yields for the representation of the yield curve to be correct. By placing the yield for each bond of the issuer I in relation to its time to maturity and thereafter applying the *Bootstrap metod* on the full set of bonds, yields are converted into zero yields and the representation of the yield curve is corrected.[1] Implementing the method in *Matlab* is done with the function *zbtyield()*, from which zero yields are returned for the same times to maturity which were used as input in the function.

The Bootstrap method is applied for each day of the observed period, always using all bonds of the issuer I, for which yield notations are available on that day. The procedure is finally repeated for the remaining issuers in the sample, resulting in a corresponding set of zero yields for the entire set of observed bonds.

[1] Each time the Bootstrap method is applied, it by construction done so on a full term structure. Higher maturity zero yields are constantly based upon already estimated zero yields for lower maturities of the same curve. For a descriptive example the reader is referred to section 3.1 where the method is explained.

With the aim to obtain yield time series at *constant maturities*, issuer specific yield curves are estimated at each day of the observed period, based on previously estimated zero rates. However, it was found that at least three zero yields are required, for the chosen estimation techniques to produce sound results. On days where this criteria is not fulfilled, the yield curve is not estimated and will further on be regarded as missing.[2]

Again consider the issuer I, for which a set of n bonds are observed. At a fixed day of the observed period, let y_i and T_i denote the yield and time to maturity of bond i, where $i = 1, \ldots, n$. Bonds are ordered such that $T_1 < \ldots < T_n$. Three different estimation techniques are applied and can with defined notations be described as follws:

1. In the first approach, bootstrapped zero yields are connected through simple *linear interpolation* between consecutive bonds, assuming constant forward rates. Thus, to obtain the unknown yield y for the known time to maturity T, where $T_{i-1} < T < T_i$ and generally $y_{i-1} < y < y_i$, relation (4.1) for linear interpolation is used:

$$y = \frac{y_i - y_{i-1}}{T_i - T_{i-1}}(T - T_{i-1}) \tag{4.1}$$

   For maturities outside of the sample maturity range, zero yields are assumed to be constant. The yield curve for concerned maturities is thus represented by a horizontal line.

2. The second estimation technique consists of a *cubic smoothing spline*, applied directly onto the estimated zero yields with the help of the *Matlab* function *csaps*(). The smoothing spline f is obtained by minimizing the sum of squares plus the penalty function (compare with section 3.2) of the expression:

$$\sum_{i=1}^{n} |y_i - f(T_i)|^2 + \lambda \int_0^T f''(T) dT \tag{4.2}$$

   Through visual inspection of estimated curves together with their fit to corresponding zero yields, the smoothing factor λ is set to the relatively high value 0.6 which gives a significant smoothing of the curve.[3] Such strong smoothing is found necessary for the curve not to display too high adjustment to the sometimes uneven positioning of yield notations, thus preventing an oscillating behaviour of the curve. Furthermore, a higher smoothing factor generally agrees with the smoothness typically characterizing a yield curve. In order to obtain estimates of maturities outside of the sample range, curves are extrapolated with a second order polynomial.

3. The third and most sophisticaed method is the *Svensson model*. Based on computed *clean prices*, *coupon rates*, *coupon frequencies* and *settlement dates*, curves are estimated to fit zero yields with the help of the *Matlab* function $fitsvensson()$, which uses non-linear least squares to find the optimal parameter values, $(\tau_1, \tau_2, \beta_0, \ldots, \beta_3)$. Due

[2]The motivation behind this choice is that the applied estimation techniques simply cannot produce reasonable curve estimates for less than three zero yields. The accurateness of estimation results can be questioned if less than three notations are used and one of the techniques can not even produce an estimate due to the model parametrization.

[3]The function *csaps()* requires the smoothing parameter $p = 1 - \lambda$ for which $p = 1$ results in an interpolating spline and $p = 0$ the least squares linear fit. The function is thus called with the value $p = 0.4$ equal to $\lambda = 0.6$.

to the extent of the model expression, its is here not presented again but the reader is referred to section 2.1.2 for complete representations.

Yield curves obtained from each technique are carefully reviewed, with qualities such as smoothness and closeness of fit with the underlying zero yields, at highest priority. Given the characteristics of the linear interpolation, smothness is essentially evaluated for the Svensson model and the smoothing spline. Additionally, basic properties of term structures such as long term convergence and non-negativity are verified.

The reason for which several techniques are evaluated is to identify one that best models the sample data. As the focus of this study rather is placed on backfilling, it was found appropriate to apply a number of established techniques and select the one that produces the best estimates. After careful review, the smoothing cubic spline is finally chosen for the remainder of this study. Despite being slightly less sophisticated than the Svensson model, curves estimated with the smoothing spline are found to best fulfill desired curve qualities without too much loss of already scarce information.

The yield curve estimation results in a transformed set of data. As oppose to originally observed yield notations for given bonds, time series now correspond to *zero yields at given maturities*, one set for each issuer. For more clarity in future computations, notations are hereby introduced.

Let $I = 1, \ldots, k$, once more denote the sample issuers and let $t \in \{s \in \mathbf{N}^* | s \leq 1304\}$ be the 5 year time period of daily observations. Furthermore let Y^I be the estimated term structure of issuer I. Denoting time to maturity m (also referred to as *time buckets*) and introducing matrix notations, Y^I can be expressed as $Y^I = (y^I_{t,m})$, where yield time series now are observed at constant maturities.

By the chosen construction of term structures, each Y^I consists of rows which are either *complete* or *entirely missing*. Additionally, to simplify computations, observations are placed in reverse order so that $t = 1$ corresponds to the most recent observation. The range of the discrete time buckets m differs between issuers, $m \in M^I = \{\ldots\}^I$, and depends on the longest maturity bond of each issuer. In most cases $M^I = \{3m, 6m, 1y, 1.5y, 2y, 3y, 4y, 5y, 10y, 15y, 20y\}$, but for a couple of issuers the longest maturity is instead $10y$.[4]

4.2 Backfilling of missing observations

As historical term structures have been estimated it is now time to place focus on filling the significant fraction of missing observations in the data set. With the aim to obtain time series that are stationary and iid, univariate GARCH(1,1) filters are applied on each of the time series. Prior to the filter application, the transformed data is once again reviewed for outliers. Based on visual inspection, parts of *or* entire term structures, are removed if they deviate too much from term structures of adjacent days. The reason for which this has to be done can simply be seen as shortcomings in the chosen term structure estimation technique, especially affecting the short and long ends of the curves. After this adjustment, term struc-

[4] Attepmting to estimate term structures of longer maturities than those of existing bonds is simply found meaningless since no data is available for such maturities and positions thereof cannot be held.

tures Y^I now also contain a number of rows where only a part of the observations are missing.

Reviewed term structures are placed horizontally after each other in a large matrix $\mathbf{Y} = [Y^1, Y^2, \ldots, Y^k]$. Daily yield changes are thereafter calculated for each time series in $\mathbf{Y}$ as: $x^I_{t,m} = y^I_{t,m} - y^I_{t+1,m}$, where t and $t+1$ are two consecutive days (recall that observations are placed in reverse time order). For the yield change to be calculated on a certain day one thus needs the current and previous day's yield notations. Since data at times is very scarce, the computation results in a matrix of daily yield changes with a higher amount of missing values than the matrix containing yield notations. The entire data matrix of yield changes can now be denoted $\mathbf{X} = [X^1, X^2, \ldots, X^k]$ and for which yield changes exist for the period $t = 1, \ldots, 1303$.

4.2.1 GARCH Estimation

The application of GARCH(1,1) filters on each univariate time series of yield changes is done with the belief that a correct filling is easier to obtain if time series are iid with mean zero and volatility one. To justify this choice, the existence of GARCH(1,1) properties must however first be established within them. As time series of yield changes are treated individually throughout this section, the simplified notation x_t will here be used, where $x_t := x^I_{t,m}$ for a fixed issuer I and maturity m.

Autocorrelations, conditional volatility functions and filtered yield changes (also referred to as GARCH residuals throughout the remainder of this study) are reviewed for a couple of selected time series. The *Matlab* function *garchfit()* is additionally applied to estimate GARCH(1,1) coefficients. For the assessment to be meaningful, the various steps must be performed on time series with higher quality data and it is further assumed that remaining time series behave similarly to those being analyzed. Results from the various tests are presented in Appendix A.

Once GARCH properties have been established, coefficients $\theta = (a, b, c)$, as defined in equation (4.3), are estimated for each x_t. Due to missing values, the estimation is executed with a method that allows for missing observations. By applying the Kalman filter equations, optimal GARCH coefficients together with conditional volatilities for observed data points are obtained according to the approach described by Penzer (2007), briefly recited below.[5]

$$x_t = \sigma_t \epsilon_t$$

$$\sigma_t^2 = c + b x_{t-1}^2 + a \sigma_{t-1}^2 \tag{4.3}$$

Assuming that each time series x_t of yield changes is zero mean distributed and follows a GARCH(1,1) process, a transformation is imposed on x_t. Define $u_t = x_t^2 - c/(1-a-b)$, let $\xi_t = \sigma_t^2 - c/(1-a-b)$ and reformulate the GARCH representation as $\xi_t = a\xi_{t-1} + bu_{t-1}$. An estimate of the sample conditional variance function is then given by the expression $\tilde{\sigma}_t^2 = \tilde{\xi}_t + \frac{c}{1-a-b}$. The optimal GARCH(1,1) coefficients $\theta = [a, b, c]$, finally used for estimating the conditional variance, are obtained through minimizing the expression below:

$$\hat{\theta} = arg \min_{\theta} \{ \sum_{t=\nu}^{n} \left[x_t^2 / \tilde{\sigma}_t^2(\theta | \mathbf{x}, F_0) + \log \tilde{\sigma}_t^2(\theta \| \mathbf{x}, F_0) \right] \} \tag{4.4}$$

[5] For a more detailed description of the estimation method the reader is asked to turn to Penzer (2007), found in the Bibliography.

When coefficients and conditional volatilities have been estimated for each of the constant maturity time series of yield changes in the sample, they are filtered by dividing them by their respective conditional volatilities. This gives a set of new time series of GARCH residuals, ϵ_t, with the same missing positions as corresponding yield changes x_t.

Let $\mathbf{E}$ denote the matrix of filtered time series, which is of the same dimension as the matrix $\mathbf{X}$. Thus, $\mathbf{E} = [E^1, E^2, \ldots, E^k]$, where E^I corresponds to filtered yield changes for issuer I, or with matrix notation $\mathbf{E} = (\epsilon^I_{t,m})$, where $t \in \{s \in \mathbf{N}^* | s \leq 1303\}$ and $m \in M^I$ as previously defined. GARCH(1,1) residuals are by definition iid distributed with $E[\epsilon_t] = 0$, $Var(\epsilon_t) = 1$ and are independent of the information set F_{t-1} given at time t, properties that hopefully can help ensure a sound estimation of missing positions in the following imputation algorithm. Filtered time series ϵ_t are finally reviewed visually and compared with their respective volatility functions σ_t and non-filtered time series x_t, before the filling algorithm is applied.

4.2.2 PCA-based Filling Algorithm

Similar to some of the procedures presented in chapter 2, the backfilling algorithm applied in this study is based upon a Principal Component Analysis decomposition, here applied on the matrix $\mathbf{E}$ of estimated GARCH(1,1) residuals. Before the procedure is initiated, a few preparatory steps must however be performed.

The number of optimal principal components k for filling missing values in $\mathbf{E}$ is defined and remains fixed throughout the algorithm. The selection of k is explained later on in this section and the reader is for now asked to assume that k is constant. Entirely empty rows of $\mathbf{E}$, and rows that contain fewer than k observations, are excluded from the matrix, reducing its dimension. On this reduced matrix, missing positions are identified and defined as a subspace Ω, such that $\Omega = (t, j)$ where $\epsilon_{t,j}$ (thus also $x_{t,j}$) are missing. The index j is here introduced as a general notation of the columns in the matrix $\mathbf{E}$, since the actual positions of missing values in the matrix now are of interest (as oppose to characteristics such as time to maturity and issuer). Important to note is also that t now is defined on a shorter interval, depending on the number of excluded rows of $\mathbf{E}$. Missing positions are finally replaced by zeros[6], resulting in the matrix $\mathbf{E}_0$, with slightly smaller dimension than $\mathbf{E}$ and which no longer contains missing values. The backfilling procedure can now be initiated with the matrix $\mathbf{E}_0$, but is described for the general matrix, $\mathbf{E}_l$, obtained in the l:th iteration, by below algorithm:

1. Each column of the input matrix $\mathbf{E}_l$ is standardized by subtracting its mean and dividing it by its standard deviation, resulting in the matrix $\hat{\mathbf{E}}_l$. Despite time series already being filtered, some sampling error is assumed to be present and this operation is done to ensure that column vectors are zero mean distributed and have standard deviation one.

2. Principal Component Analysis (PCA) is performed on the matrix $\hat{\mathbf{E}}_l$ such that Principal Components (PC:s) $\mathbf{P}_l$ and loadings $\mathbf{W}_l$ are obtained.

[6]Both zeros and row/column wise means are tried for this purpose but as no difference in performance can be detected, and since the input data consists of filtered time series with means very close to zero (assuming some sampling error), it is chosen to consistently use zeros for this purpose.

3. The first k PC:s are selected together with corresponding loadings, resulting in a subset of PC:s and loadings denoted $\mathbf{P}_l^*$ and $\mathbf{W}_l^*$. An estimate $\hat{\mathbf{E}}_l^*$ of the matrix $\hat{\mathbf{E}}_l$ is obtained as $\hat{\mathbf{E}}_l^* = \mathbf{P}_l^* \mathbf{W}_l^{*^T}$.

4. From the estimate $\hat{\mathbf{E}}_l^*$, values are taken *only* from positions initially identified as missing (i.e. $\forall(t,j) \in \Omega$) and imputed into corresponding positions of the matrix $\hat{\mathbf{E}}_l$, resulting in the new matrix $\mathbf{E}_{l+1}$. $\mathbf{E}_{l+1}$ is finally reblended with means and standard deviations estimated in step 1 (though remains denoted $\mathbf{E}_{l+1}$).

5. Step 1 to 4 are repeated until convergence of imputed values is reached. Convergence is measured as the maximum of absolute differences of each imputed position between two consecutive iterations of the procedure and is reached when this difference falls below the threshold value 10^{-5}:

$$\max_{(t,j)\in\Omega} |\epsilon_{t,j}^{l+1} - \epsilon_{t,j}^{l}| < 10^{-5} \tag{4.5}$$

The reason why certain rows are excluded from the initial matrix $\mathbf{E}$ prior to the backfilling procedure is initiated, is that they are found not to contain enough information for estimation of their missing values. These rows are instead filled through simulation from empirical factor distributions and placed back into their original positions (as in $\mathbf{E}$), when the backfilling procedure has reached convergence.[7] A new matrix $\tilde{\mathbf{E}}$, of the same dimension as $\mathbf{E}$, has now been generated where missing positions of $\mathbf{E}$ have been filled and originally observed values remain identical.

To ensure the validity of the backfilling procedure and to further reach understanding of its limitations it is initially applied on a number of synthetic data sets. The validation process, together with results thereof, is described in detail in chapter 6.

Application on Real Data

Once validity of the PCA-based backfilling procedure has been ensured, it is applied on the matrix $\mathbf{E}$ of estimated GARCH residuals. However, due to limitations in the data set, convergence is not easily reached. A number of subsets are therefore created to try to enhance the procedure's performance. For further improvement, it is also attempted to include additional risk factors in terms of extra time series, sought to be drivers of the concerned markets in the data set. The purpose of this is to include more information in the set with the hope that convergence easier can be reached, and that a larger amount of missing values can be filled. Additional time series are observed during the same time period as the chosen data.

The different tested approaches are:

1. *No decomposition.* The procedure is applied on the full matrix $\mathbf{E}$.

2. The data is divided according to the *regions* covered by the data. Each subset, corresponding to GARCH residual time series of issuers from one region, is filled separately. The reason for which this decomposition is applied is to further make use of the regional correlation structure, assumed to be significant within the data set.

[7] This procedure is not found to be of primary focus of this study and a description thereof can therefore be found in Appendix B.

3. Backfilling is performed on time series corresponding to *tenors between two and five years*. The decomposition thus contains all GARCH residual time series, $\epsilon^I_{t,m}$, for all issuers $I = 1, \ldots, k$, where $m \in M^I = \{2, 3, 4, 5\}$. The motivation behind this choice is to remove possible excess volatilities caused by the term structure estimation method, which might populate throughout the data set as the backfilling procedure is performed. The existence of excess volatilities is mainly expected in the long and short ends of estimated curves, which is why they are removed.

4. Backfilling is performed in *two steps*, where the matrix $\mathbf{E}$ initially is divided into subsets based on *region* and *tenors*, although still for a reduced number of tenors, $m \in M^I = \{2, 3, 4, 5\}$. For clarification this means that all time series with the same maturity m, belonging to issuers of the same region, correspond to one subset which is filled separately. The aim of this decomposition is to significantly reduce the complexity of the sets, upon which the backfilling procedure is applied. After this first run of the procedure, a second decomposition is created, where subsets instead correspond to *issuer specific time series*. In other words, the procedure is now applied on each E^I, $I = 1, \ldots, k$ (though still for maturities $m \in M^I = \{2, 3, 4, 5\}$). During this second run, additional values can be filled based on already imputed estimates.

5. The *two step application* described in item 4 is again used, with the difference that *additional risk factors* now are included in the first run of the procedure. Additional time series consist of one leading regional equity index and the currency exchange rate USD/local currency for concerned region. The same additional, region specific time series are thus placed together with all the tenor-dependent subsets of each region. The second run of the procedure is analogous to the one described in item 4.

The reason for which a single issuer decomposition is not mentioned among the items above, is that by construction of each term structure, the largest part of missing values corresponds to entire missing rows. Due to the condition that the number of observations on a row must exceed k for it to be filled by the means of PCA, most values would simply not be filled by the procedure but would instead have to be simulated (something which is desired to do for as few observations as possible). However, as the issuer specific decomposition is applied above, a large part of missing values have already been filled, and further estimates can be obtained based on already imputed values. Additionally, such representation is beneficial as factor simulation of remaning missing observations is applied (described in Appendix B), to retain the correlation structure between rows of observations, of one issuer.

Defining k

Defining the optimal number of principal components, k, is a common difficulty one comes accross when implementing a PCA-based backfilling routine (recall discussions in section 2.2.3). Nevertheless it's a difficulty that has to be dealt with. Cross validation is one common approach, defining a threshold of explained variance is another. However, the characteristics of the data used in this study, complicates the selection of k. Complete subsets for cross validation are difficult to find without being forced to impute initial estimates, and defining a threshold of explained variance, with such significant fraction of the data missing, might give incorrect implications.

The approach chosen in this study concerns the comparison of the input matrix ($\mathbf{E}$ or various subsets thereof) with a random matrix with the same basic properties. The underlying

idea is that those Principal Components of the input matrix that give a higher level of explanation than corresponding components of a randomly generated matrix, contain relevant information and should thus be kept. The selection is based upon eigenvalue decomposition and is described by the following steps:

1. A large number of matrices $\mathbf{Z}^p$, $p = 1, \ldots, N$, of the same dimension and with the same missing positions as the input matrix, are simulated. Each position of $\mathbf{Z}^p$ is $N(0, 1)$ distributed.

2. A pairwise Spearman's rank correlation matrix is computed for each simulated matrix.

3. Estimated rank correlation matrices are transformed into linear correlation matrices by applying relation (4.6) (as described by Hult and Lindskog (2002)). $\rho_{s,ij}$ here denotes *Spearman's rho* for the two column vectors $z_{t,i}$ and $z_{t,j}$ $(t = 1, \ldots, 1303)$, and ρ_{ij} is the corresponding linear correlation coefficient:

$$\rho_{ij} = 2\sin(\rho_{s,ij}\frac{\pi}{6}) \tag{4.6}$$

   The reason for which rank correlation is computed and thereafter transformed to linear correlation is simply to obtain a robust estimate of the correlation matrix. Rank correlation coefficients such as Spearman's rho and Kendall's tau are invariant under the marginal distributions of the data and the approach is appropriate for heavy tailed distributions, a characteristic often found in financial data.[8]

4. Eigenvalue decompositions are performed on the linear correlation matrices and their respective eigenvalues are sorted in descending order and divided by the sum of all eigenvalues of the corresponding matrix. This gives the fraction of the total variance explained by each eigenvalue (compare with discussions in section 3.5.1 regarding PCA).

5. Based on the N simulated matrices, an empirical distribution is created for each of the ordered and normalized eigenvalues. The 95% percentile is finally computed for the empirical distribution of each eigenvalue.

6. Step 2 to 4 are now performed on the input matrix $\mathbf{E}$, resulting in a set of ordered and normalized eigenvalues of its linear correlation matrix. k is thereafter defined as the number of eigenvalues which explain more than the corresponding 95% percentile.

Once k have been defined it remains constant throughout the iterative filling process on the selected data set.

[8] Futher motivation of the choice to use a rank correlation estimator is not described in this study. The reader is instead referred to Lindskog et al. (2003) for a more detailed description and justification of the approach.

4.3 Term Structures Revisited

Based on the complete matrix of GARCH residuals, $\tilde{\mathbf{E}}$, missing yield changes are now computed and filled into $\mathbf{X}$, leaving originally observed yield changes untouched. Each time series is treated separately and the computation is done using equation (4.3). For clarification, an estimate $\tilde{x}_{t,j}$ of the missing yield change $x_{t,j}$, is computed as:

$$\tilde{x}_{t,j} = \tilde{\sigma}_{t,j}\tilde{\epsilon}_{t,j}, \quad \tilde{\epsilon}_{t,j} \; \forall \, (t,j) \; \in \Omega \tag{4.7}$$

where $\tilde{\epsilon}_{t,j}$ has been filled by the designed backfilling procedure and $\tilde{\sigma}_{t,j}$ has been computed with the help of relation (4.3), using the conditional volatility $\tilde{\sigma}_{t-1,j}$, the yield change $x_{t-1,j}$ (both observed on previous day) and estimated GARCH(1,1) coefficients (a, b, c).

After this operation, the matrix $\mathbf{X}$ of yield changes is complete. Although such representation generally is of higher interest for many financial risk applications, it is also desirable to finalize the procedure by computing the actual yields. Through the simple computation, $y_{t,j} = y_{t-1,j} - x_{t-1,j}$, each constant maturity yield time series is recursively filled, and the matrix $\mathbf{Y}$ is completed (recall that yields and yield changes are placed in reverse order where the most recent observations are found on the first rows of the matrices $\mathbf{X}$ and $\mathbf{Y}$).

Chapter 5

Data

The real set of data, on which the entire procedure is applied, from term structure estimation to filling of missing values, consists of yield notations from eight different issuers. The issuers are taken from the two different Emerging Markets *South Korea* and *Mexico* and correspond to one government and three corporate issuers from each region. Data is downloaded for the five year period 2005.05.03 - 2010.04.30, mainly from Bloomberg but complementary sources, specific for the bank, are to some extent also used. For each of the eight issuers a set of bonds are used to estimate daily term structures. The number of bonds per issuer varies between 4 and 21 and the length of the historical time series differs beween issuers depending on the data availability among providers. Table 5.1 gives more detailed information about available data for each of the eight issuers.

Issuer Name	Bond Type	Industry	# Bonds	Available History (Years)
Industrial Bank of Korea	Corporate	Finance	18	≈ 4.5
Korea Exchange Bank	Corporate	Finance	21	≈ 4
Korea Telecom	Corporate	Telecom	19	5
Republic of Korea	Government	-	19	5
América Móvil	Corporate	Telecom	7	≈ 2
Petróleos Mexicanos	Corporate	Petrol	11	5
Telefonos de Mexico	Corporate	Telecom	4	≈ 2
United Mexican States	Government	-	19	5

Table 5.1: Description of the real data on which the full procedure is applied throughout this study.

As can be seen in table 5.1, the availability of data varies significantly among the selected issuers and a five year history can only be obtained for a few of them. Also worth to be noted is that despite a fairly long history of available data, illiquidity might cause a significant amount of gaps also during the observed time period, which is the case for a couple of issuers.

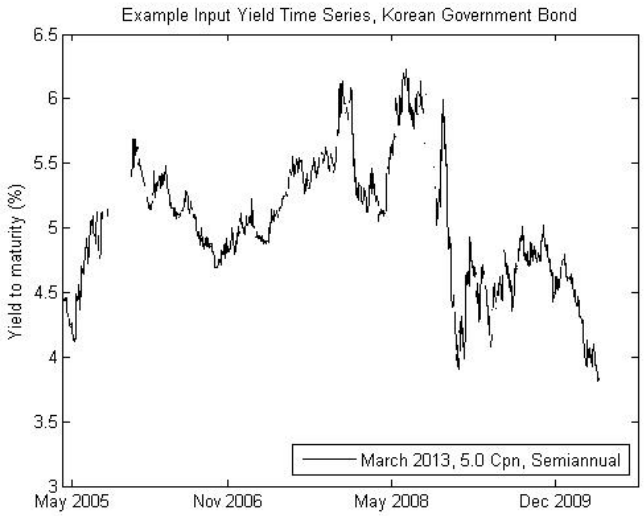

Figure 5.1: One of the many raw input yield time series is illustrated. This particular time series consists of coupon bearing yield notations for a Korean government bond with 5.0 semiannual coupon which matures in March 2013. Even for this time series, which is regarded as one of the better ones as it streches over the entire five year window, a number of gaps with missing values are clearly visible.

As a final illustration of the raw input data, an example time series (here taken from the set of Korean government bonds) is shown in figure 5.1. Even for this issuer, which is regarded as one of the better ones, missing values are clearly noticable in the time series.

Chapter 6

Model Validation

This chapter contains an evaluation of the iterative, PCA-based backfilling routine which later on is implemented with the purpose of filling missing observations in a term structure dimension. The application of the routine onto a number of different synthetic data sets, with different underlying factor structures and various fractions of missing data, is presented together with obtained results thereof.

6.1 Synthetic Data Sets

Two main data structures are created for the evaluation of the procedure. The first a more generic data set and the second somewhat more custom made, to resemble the true set of data onto which the procedure later on will be applied. The respective structures and the ways in which they are varied are further described below.

Generic Set

Let $\mathbf{X} = (x_{t,i})$, $i = 1, \ldots, 50$, be a matrix containing 50 simulated time series over the time period $t = 1, \ldots, 1000$. Furthermore let $\mathbf{Y} = (y_{t,j})$ denote the matrix of underlying factors where $j = 1, \ldots, k$ and $k \in \{5, 10\}$. Let error terms $\epsilon_{t,i}$ (not to be confused with GARCH residuals as described in chapter 4) of each simulated value $x_{t,i}$ be unique and let each $\epsilon_{t,i} \in N(0, 1)$. $x_{t,i}$, is related to the underlying factors $y_{t,j}$ by a set of unique coefficients $\alpha_{j,i}$ such that:

$$x_{t,i} = y_{t,1}\alpha_{1,i} + \ldots + y_{t,k}\alpha_{k,i} = \sum_{j=1}^{k} y_{t,j}\alpha_{j,i} \tag{6.1}$$

The underlying factors $y_{t,j}$ are simulated $N(0, 1)$ and the respective factor coefficients $\alpha_{j,i}$ are uniformly distributed over the interval $[0, 1]$.

The fraction of missing values of the data set is varied between 5%, 10%, 15%, 20%, 25%, 30%, 40% and 50%. Positions where observations are set as missing are randomly selected (discrete uniform) and true values of these positions (simulated from above factor structures) are saved for comparison with estimates imputed by the procedure. The number of Principal Components (k) used in the backfilling procedure is set to equal the number of underlying factors of the simulated structure.

To measure the performance of the backfilling procedure, the average sum of squared differences between true and imputed value over all imputed positions is calculated, as expressed in equation 6.2. The reason for this is that the amount of positions for which the sum is calculated is varied between 5% and 50%, and an average measure is necessary for comparison of the procedure's performance.

$$\frac{1}{|\Omega|} \sum_{x_{t,i} \in \Omega} (\tilde{x}_{t,i} - x_{t,i})^2 \tag{6.2}$$

In the above equation, $x_{t,i}$ denote true values which have been intentionally removed, $\tilde{x}_{t,i}$ correspond to the imputed values at the same position, Ω the set of missing positions in the observed data and $|\Omega|$ the number of positions in Ω.[1] Additionally, the sum of squared errors, the maximum absolute difference between true and estimated value among all imputed positions, and the number of iterations needed to reach convergence are reviewed. Results are presented later on in this chapter (section 6.2).

Customized Set

The second synthetic data set is created to resemble the true set of GARCH residuals onto which the procedure later on will be applied. It consists of 92 time series with 1303 observations each. Using notations as described for the generic data set above, $x_{t,i}$ and $y_{t,j}$, where $i = 1, \ldots, 92$ and $j = 1, \ldots, k$, are simulated time series and underlying factors respectively. In this exercise, it is however chosen to try a larger number of underlying structures and therefore $k \in \{1, 3, 5, 10, 15, 20\}$. To additionally investigate the influence of the error term, the exercise is also performed for two different sizes of errors: $N(0, 10^{-4})$ and $N(0, 10^{-2})$. The reason for this is that the amount of noise in daily financial data sometimes can be significant and the influence of the error term on the algorithm's performance is thus of interest.

To further resemble the true set of data, which is obtained from a number of eight underlying term structures, column vectors of the data matrix $\mathbf{X}$ are groupwise assigned different variances and small means. Even though the procedure itself is designed to standardize vectors of the input data at each iteration, this characteristic is imposed since it is also found in the true set of data.

[1] $|\Omega|$ is referred to as the *cardinal* of Ω and defines the number of positions in the set Ω.

Missing positions are now set to be identical to those in the matrix of filtered GARCH residuals. Missing values are thus not randomly selected and their character is more concentrated to larger horizontal and vertical gaps instead of randomly spread single positions. The total fraction of missing positions thus corresponds to 33.7% of the data, but the fraction varies among subsections since the availability of data differs between various issuers. k is again given by the number of underlying factors and the model performance is measured as previously described for the generic data set.

6.2 Results from Model Validation

Below, results obtained from the application of the backfilling procedure onto the two synthetic data sets are presented, starting with the generic set thereafter followed by the customized set, both with various imposed characteristics as explained in section 6.1.

6.2.1 Generic Data Set

Table 6.1 presents results obtained for the generic data set with a 5 factor underlying structure.

# Missing (%)	Max Abs Diff	Sum Square Diff	Avg Sum Square Diff (10^{-3})	# Iterations
5%	0.0439	0.3042	0.1261	17
10%	0.0464	0.6132	0.1226	17
15%	0.0455	0.9368	0.1249	29
20%	0.0476	1.2574	0.1257	34
25%	0.0523	1.6075	0.1286	52
30%	0.0508	1.9467	0.1298	56
40%	0.0491	2.7188	0.1359	67
50%	0.0559	3.6019	0.1441	178

Table 6.1: Results from generic data set with underlying 5 factor structure. Columns in the table correspond to the fraction of missing values in the data set (1), the maximum absolute difference between imputed and original values (2), the sum (3) and average sum (4) of squared differences over all imputed positions and the number of iterations needed to reach convergence (5).

As can be seen from the results in table 6.1, the algorithm produces sound results as the distances between imputed values and true values in general are small. Convergence of the algorithm is easily reached with a slight increase of necessary iterations, proportional to the increasing amount of missing values. The average sum square difference also displays a slight increase as the fraction of missing observations grow large whereas it remains on a constant level for lower fractions. The maximum absolute difference between real and imputed values remains relatively constant for all scenarios. A larger number of iterations as well as a larger difference between true and imputed values with a growing number of missing positions is expected as the amount of information in the set decreases, but results here show that the precision of the model remains high.

Figures 6.1 and 6.2 illustrate some of the results obtained with the underlying five factor structure. The same vector $x_{t,i}$ of the data set ($t = 1, \ldots, 1000$ and i fixed) where 10% and 30% of the positions have been removed, is shown together with imputed missing positions

and the difference between imputed and true values over *all* positions of the vector. For positions where observations have not been removed, this difference is simply represented by a zero.

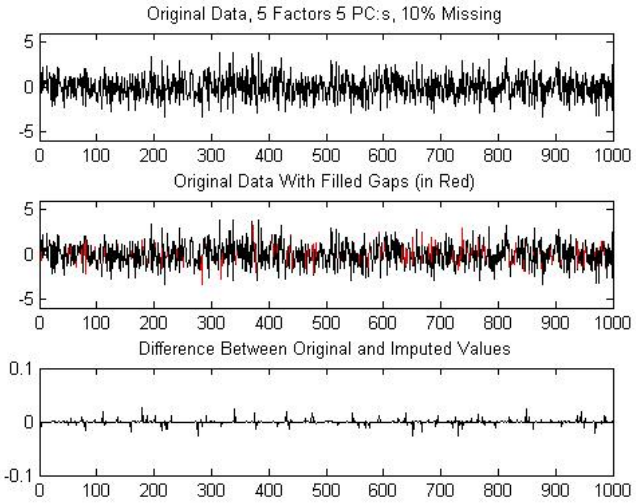

Figure 6.1: Example vector from a generic 5 factor structure where missing values have been imposed on 10% of the positions in the data set and filled with the designed algorithm. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

# Missing (%)	Max Abs Diff	Sum Square Diff	Avg Sum Square Diff (10^{-3})	# Iterations
5%	0.0397	0.3287	0.1315	22
10%	0.0454	0.6699	0.1340	29
15%	0.0425	1.0134	0.1351	43
20%	0.0422	1.3888	0.1389	60
25%	0.0447	1.7764	0.1421	84
30%	0.0503	2.2126	0.1475	106
40%	0.0698	3.2322	0.1616	148
50%	0.0886	4.6693	0.1868	639

Table 6.2: Results from generic data set with underlying 10 factor structure. Columns in the table correspond to the fraction of missing values in the data set (1), the maximum absolute difference between imputed and original values (2), the sum (3) and average sum (4) of squared differences over all imputed positions and the number of iterations needed to reach convergence (5).

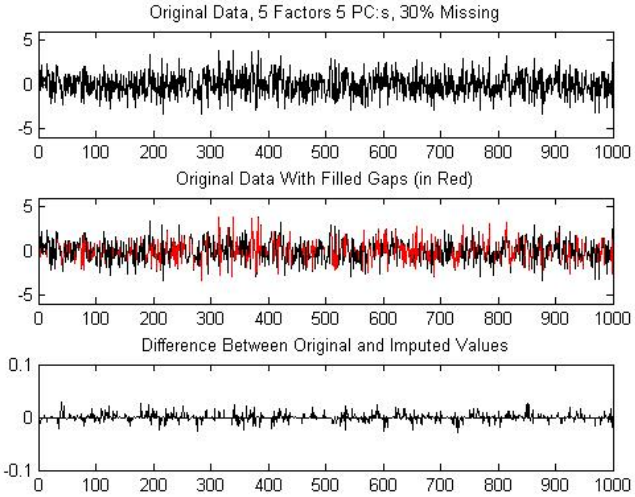

Figure 6.2: Example vector from a generic 5 factor structure where missing values have been imposed on 30% of the positions in the data set and filled with the designed algorithm. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

Results in table 6.2, obtained as the number of underlying factors is increased from 5 to 10, roughly display the same characteristics as for the previous less complex structure. Average sum square differences are still fairly small as are the maximum absolute differences, although these measures show larger increases than in previous example. The most significant difference is seen in the number of iterations needed for the algorithm to reach convergence. In general convergence is still easily reached but for the set which contains 50% of missing values, a much larger number of iterations is needed than for previous factor structure, as well as compared with lower fractions of missing values for the same underlying structure. These facts indicate that the factor complexity has a notable impact on convergence of the model, as well as some impact also on other performance measures. The loss of precision with increasing fraction of missing positions is here more significant than for the less complex factor structure.

Again a couple of sample plots illustrate some of the results obatined with the 10 factor structure. An example vector is shown in figures 6.3 and 6.4, with 10% and 30% of missing values respectively.

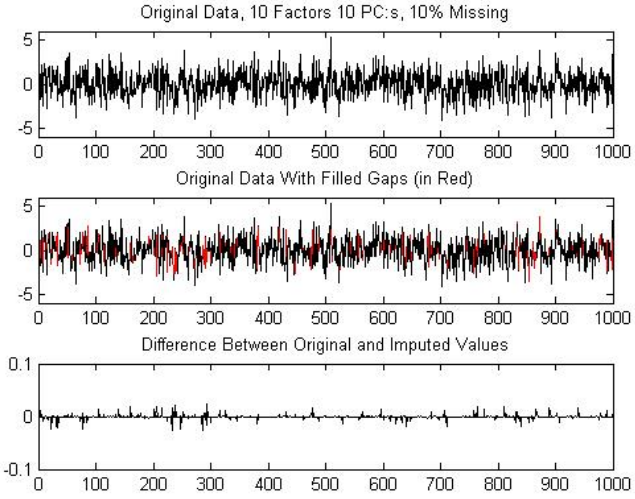

Figure 6.3: Example vector from a generic 10 factor structure where missing values have been imposed on 10% of the positions in the data set and filled with the designed algorithm. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

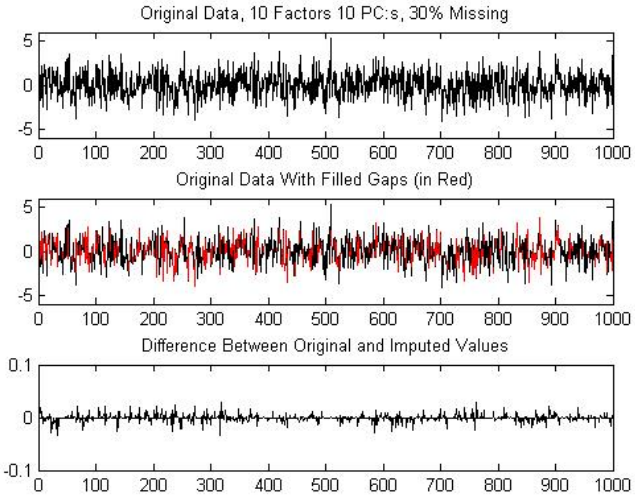

Figure 6.4: Example vector from a generic 10 factor structure where missing values have been imposed on 30% of the positions in the data set and filled with the designed algorithm. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

6.2.2 Customized Data Set

As the smaller N(0,10^{-4}) distributed error terms are applied for the customized data set whith constant fraction and positioning of missing values and where the number of underlying factors are varied, results presented in table 6.3 are obtained.

# Factors	Max Abs Diff	Sum Square Diff	Avg Sum Square Diff (10^{-6})	# Iterations
1	0.00228	0.0031	0.0758	79
3	0.00262	0.0027	0.0660	117
5	0.00566	0.0044	0.1080	225
10	0.03017	0.0145	0.3594	2079
15	0.00954	0.0195	0.4830	4727
20	0.01791	0.0436	1.0786	9740

Table 6.3: Results from customized data set with N(0,10^{-4}) distributed error terms, constant fraction of missing values (33.7%) and various underlying factor structures. Columns in the table correspond to the number of underlying factors (1), the maximum absolute difference between imputed and original values (2), the sum (3) and average sum (4) of squared differences between true and imputed value over all imputed positions and finally the number of iterations needed to reach convergence of the algorithm (5).

One of the most evident differences compared with the generic data set are the number of necessary iterations to reach convergence. Even for a less complex factor structure with 10 underlying factors, more than 2'000 iterations are needed to reach convergence whereas less than 150 iterations would have been necessary for the corresponding fraction of missing values of the generic data set. Although, certain conclusions cannot be made since also other characteristics differ, such as the dimension of the data set, this could be an indication that the positioning of missing values has an impact on the time needed to reach convergence. Such indications have been observed before, among others by Grung and Manne (1998).

As expected with a small error term, other performance measures display good results in each scenario. The average sum square differences are as small as of the order 10^{-6} (for the simplest structure) and grow with an increasing number of factors, in line with expectations. The maximum absolute difference between true and imputed values also increases with the number of underlying factors, though with one exception; the underlying 10 factor structure. That this particular data set yields the biggest absolute distance is somewhat surprising but not worrying since average sum square difference instead displays expected behaviour where the largest value is found for the largest number of factors. A number of example illustrations are found in figures 6.5 - 6.8 where two different vectors with various characteristics of missing values are shown.

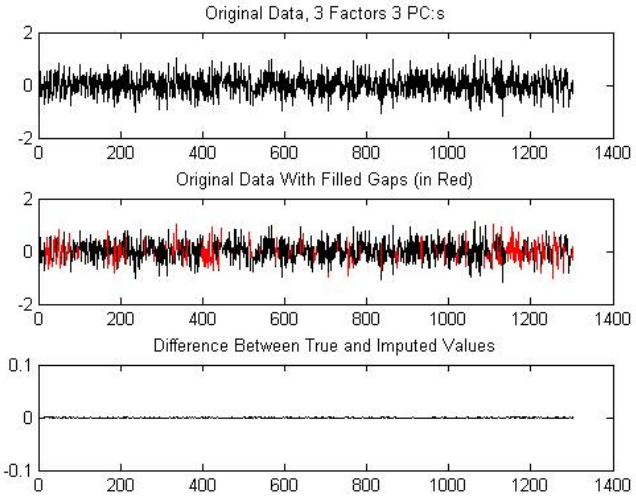

Figure 6.5: Example vector from customized data set based on a 3 factor structure with N(0,10^{-4}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is farily even. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

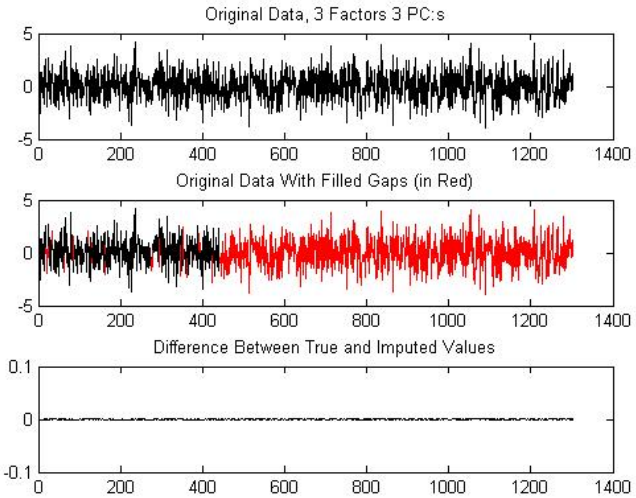

Figure 6.6: Example vector from customized data set based on a 3 factor structure with N(0,10^{-4}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is characterized by a long consecutive gap. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

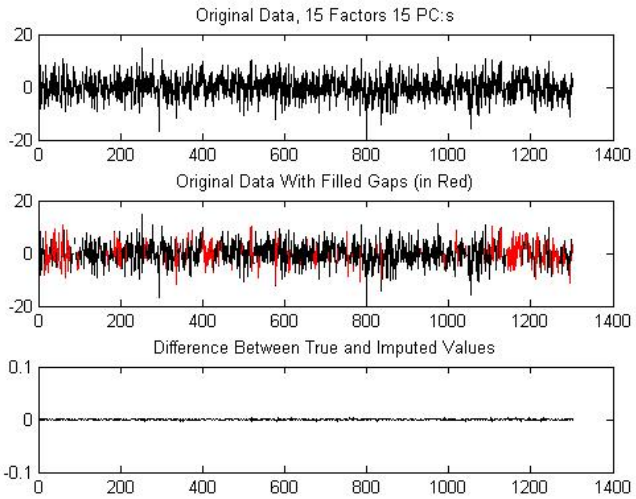

Figure 6.7: Example vector from customized data set based on a 15 factor structure with N(0,10^{-4}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is farily even. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

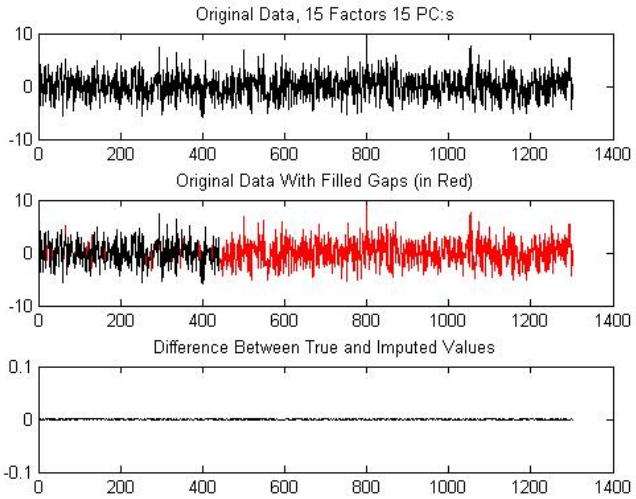

Figure 6.8: Example vector from customized data set based on a 15 factor structure with N(0,10^{-4}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is characterized by a long consecutive gap. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

As error terms instead are set to be somewhat larger, namely N(0,10^{-2}) distributed, the algorithm instead yields results presented in table 6.4.

# Factors	Max Abs Diff	Sum Square Diff	Avg Sum Square Diff (10^{-3})	#Iterations
1	0.17737	16.155	0.3997	80
3	0.25506	26.251	0.6495	116
5	0.55895	43.224	1.1000	225
10	2.90030	140.869	3.5000	2144
15	0.96018	190.651	4.7000	4431
20	4.20830	478.129	11.8000	9651

Table 6.4: Results from customized data set with N(0,10^{-2}) distributed error terms, constant fraction of missing values (33.7%) and various underlying factor structures. Columns in the table correspond to the number of underlying factors (1), the maximum absolute difference among imputed and original values (2), the sum (3) and average sum (4) of squared differences over all imputed positions and finally the number of iterations needed to reach convergence of the algorithm (5).

When reviewing the results in table 6.4, one can state that the performance measures behave similar to those obtained with the smaller error term. The maximum absolute difference, the sum of squared differences and the average sum of squared differences typically increase with an increasing number of factors. Although with the same exception of the 10 factor structure where the largest absolute difference between true and imputed value can be seen. However, the sizes of the measures are here notably larger. Convergence is again reached with similar numbers of iterations as for previous approach and results are satisfactory at every attempt. The behaviour of the number of needed iterations imply that the error term does in fact not affect the time needed to reach convergence but only the correctness of imputed values.

A number of illustrative examples are shown in figures 6.9 - 6.12. Note the different scales for the illustrated differences between imputed and true positions, as well as the average sum squared differences presented in table 6.4, compared to the factor structure containing smaller error terms.

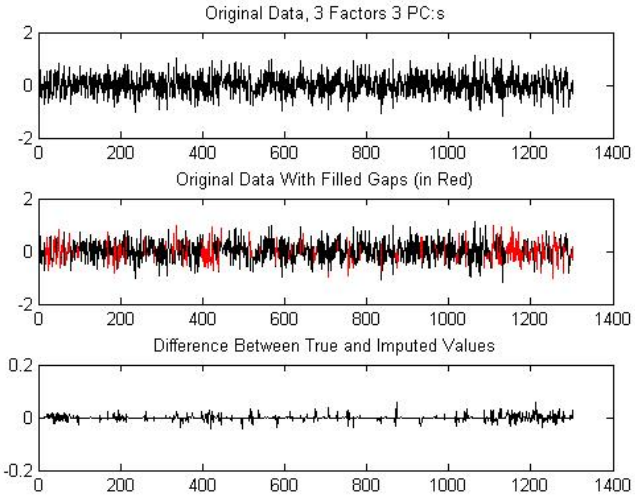

Figure 6.9: Example vector from customized data set based on a 3 factor structure with N(0,10^{-2}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is farily even. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

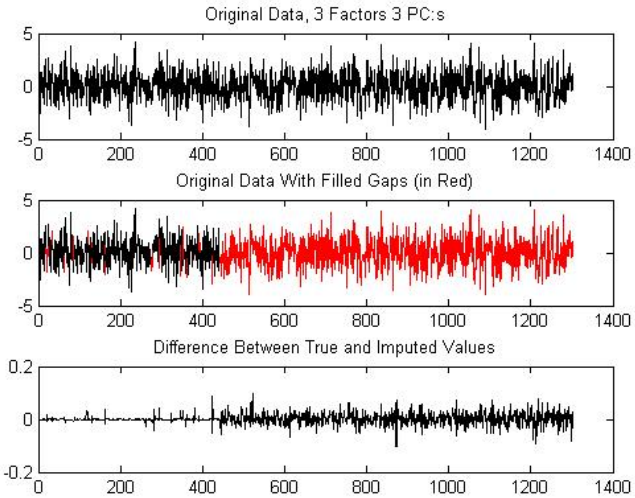

Figure 6.10: Example vector from customized data set based on a 3 factor structure with N(0,10^{-2}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is characterized by a long consecutive gap. Filled Values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

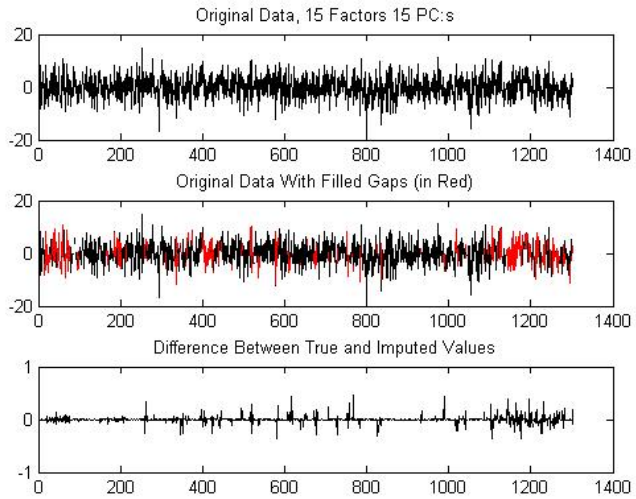

Figure 6.11: Example vector from customized data set based on a 15 factor structure with N(0,10^{-2}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is farily even. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

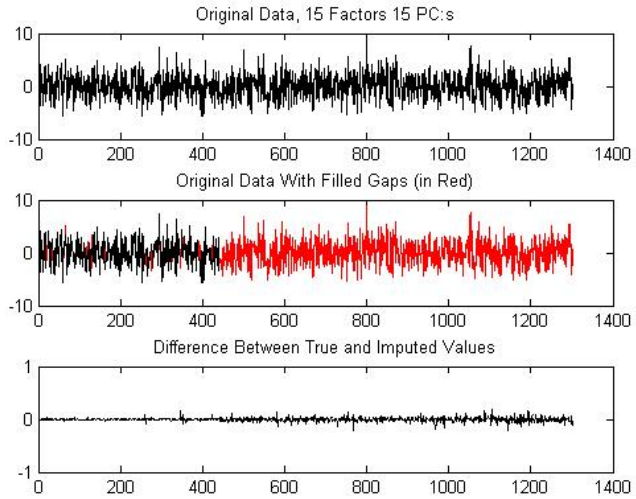

Figure 6.12: Example vector from customized data set based on a 15 factor structure with N(0,10^{-2}) distributed errors and where missing values have been imposed on 33.7% of the positions in accordance with the original data set. In this example, the distribution thereof is characterized by a long consecutive gap. Filled values are illustrated in red on the centered plot of the figure and the differences between imputed and original values are shown in the bottommost plot of the figure.

Overall validation results imply that the model produces good estimates when it is applied on data sets with various underlying structures and varying fractions of missing data. Convergence is certainly reached slower when the number of factors as well as the fraction of missing values increase. Nevertheless, the backfilling procedure produces sound estimates which in fact lie very close to true values in all scenarios.

From the presented validation tests, implications are also found that the error term has an impact on the performance of the model. As the size of the error terms of the customized data set increases, imputed values differ more from true values than with smaller error terms. The number of iterations remain roughly the same, but sum squared differences increase significantly, something which might be important to bear in mind as the algorithm is applied on a true set of daily financial data which generally contains a substantial part of noise.

Finally judging from the larger number of iterations seen for the customized data set, indications are found that the positioning of missing values might have an impact on the time needed for the routine to converge. Although imputed estimates are satisfactory, larger horizontal and vertical gaps appear to notably slow down the routine, something which also has been expressed in previous studies.

Chapter 7

Results

This chapter presents results from the various components of the study, applied on a true set of data with the aim to estimate historical yield curves. Estimated term structures are illustrated and assessed. Thereafter follows results obtained from the backfilling procedure, which are of special importance in this study. Last but not least, results from the reconstruction of term structures with the help of imputed estimated values and GARCH(1,1) parametrizations are presented. At the very end of the chapter, a short summary of gathered results is presented to provide the reader with an overview of the most important findings.

7.1 Estimated Yield Curves

Figures 7.1, 7.2 and 7.3 illustrate zero yields obtained from the Bootstrap method applied onto various sets of coupon bearing yield notations.

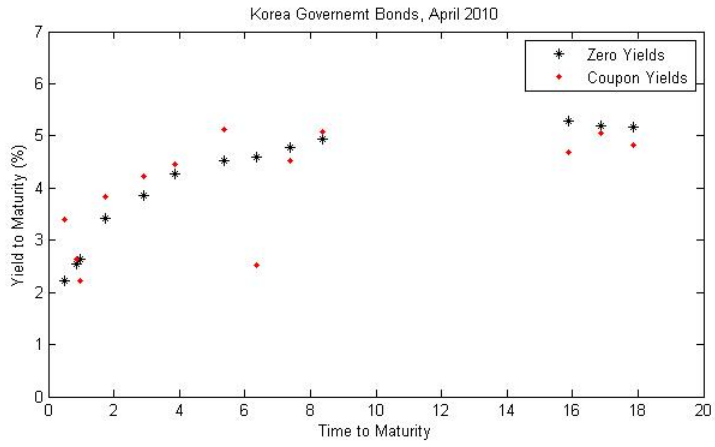

Figure 7.1: Bootstrapped zero yields are illustrated together with corresponding coupon bearing yield notations, here for a set of Korean government bonds with various coupon and time to maturity, observed in April 2010. Zero yields display a farily nice curvature of an upward sloping curve with a slight dip towards longer maturities.

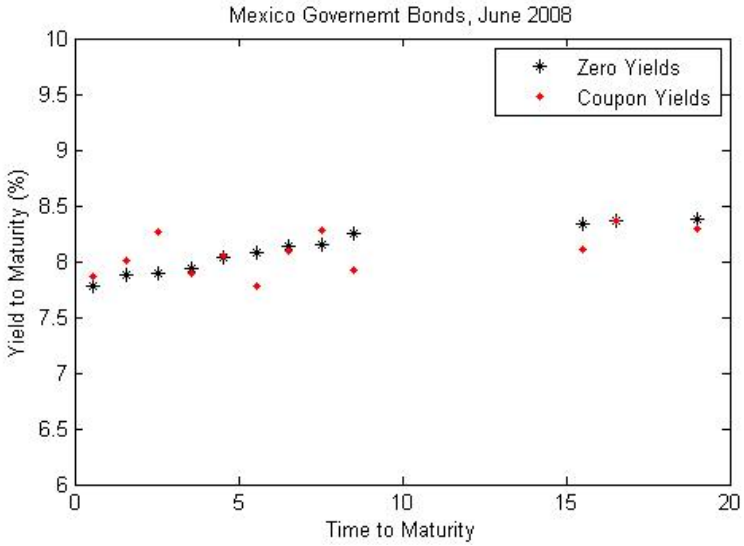

Figure 7.2: Bootstrapped zero yields are illustrated together with corresponding coupon bearing yield notations, here for a set of Mexican government bonds with various coupon and time to maturity, observed in June 2008. Zero yields imply a relatively flat term structure during the observed period.

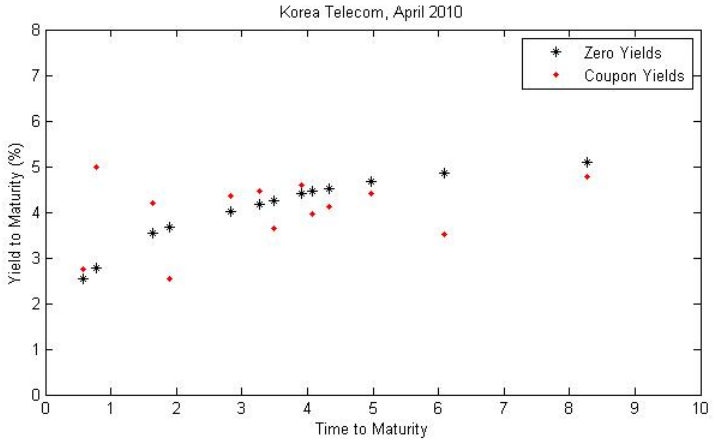

Figure 7.3: Bootstrapped zero yields are illustrated together with corresponding coupon bearing yield notations, here for a set of corporate bonds issued by Korea Telecom, with various coupon and time to maturity, observed in April 2010. Similar to figure 7.1 above, estimated zero yields create a smooth and nice curvature of one of the typical shapes of a yield curve.

As can be seen, estimated zero yields display a more even pattern than initial coupon bearing yields. This implies that the Bootstrap method produces sound zero notations which can be used as a base for the continued yield curve estimation. However, with a decreasing number of points, the method tends to produce less stable estimates where yields sometimes diverge from an imagined curve. This feature will be dealt with later on as estimated yield curves will be closely obsereved and outliers removed from the sample.

The application of the three different yield curve estimation techniques onto obtained zero yields results in curves of widely varying quality. A number of example curves are illustrated in figures 7.4, 7.5 and 7.6, where results from all three techniques are shown togeher with underlying zero yields.

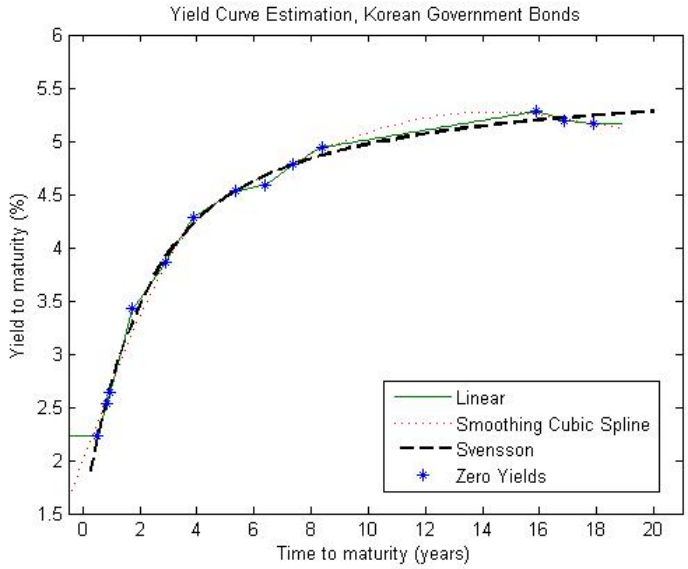

Figure 7.4: Three different techniques are applied onto a set of zero yields to obtain estimates of the yield curve: linear interpolation between estimated zero yields, a smoothing cubic spline and the Svensson model. A somewhat naive estimate is obtained through linear interpolation whereas other techniques produce sensible curve estimates.

Linear interpolation between estimated zero yields, where yields of maturities outside of the sample data are treated as constant, always produces results but the quality thereof is generally very poor. The method works fairly well when a large set of evenly distributed zero yields are available. Results are however less satisfactory when observations are few an unevenly distributed. Even for a set of yields as in figure 7.5, where notations are frequent and a decent curve shape can be sensed, an uneven shape of the interpolated curve is evident.

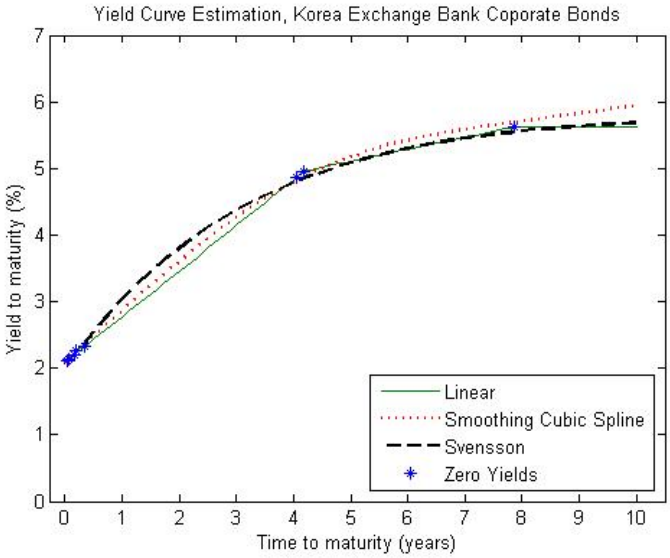

Figure 7.5: Three different techniques are applied onto a set of zero yields to obtain sound estimates of the yield curve: linear interpolation between estimated zero yields, a smoothing cubic spline and the Svensson model. Good estimates are produced by the Svensson model and the smoothing spline whereas linear interpolation appears to underestimate the curve, entirely without smoothness.

Depsite not being the most sophisticated method for yield curve estimation, the *smoothing cubic spline* with smoothing factor 0.6 generally produces results of good quality. Good sample data gives estimates almost in the range of what the Svensson model produces and during periods, and for issuers, where data is of lower quality, the smoothing spline also generates reasonable estimates in almost all of the cases. Figure 7.7 illustrates this fact where a smooth curvature and asymptotic properties are obtained with the smoothing spline whereas a Svensson estimate could not even be obtained due to restrictions in the data.

For a good set of zero yields, the *Svensson model* clearly produces the best results among the three techniques. Smoothness of the curve as well as fit to the data are well fulfilled which can be seen in figures 7.4 and 7.5. In many of the cases though, the results obtained with the Svensson model are disappointing. As the quality of the data used in this study does not seem to agree with what is necessary for the model, a large part of the estimated yield curves display uneven curvature, poor fit to zero yields and even yields that are negative or that goes to infinity as the time to maturity decreases towards zero. One poor example of the capability of the Svensson model is illustrated in figure 7.6 where a corporate yield curve for Petróleos Mexicanos is estimated. Two sharp hump shapes can be seen together with a yield decreasing toward minus infinity. Figure 7.7 shows another example, though in this case the Svensson model cannot even create a curve estimate as too few yield notations are observed.

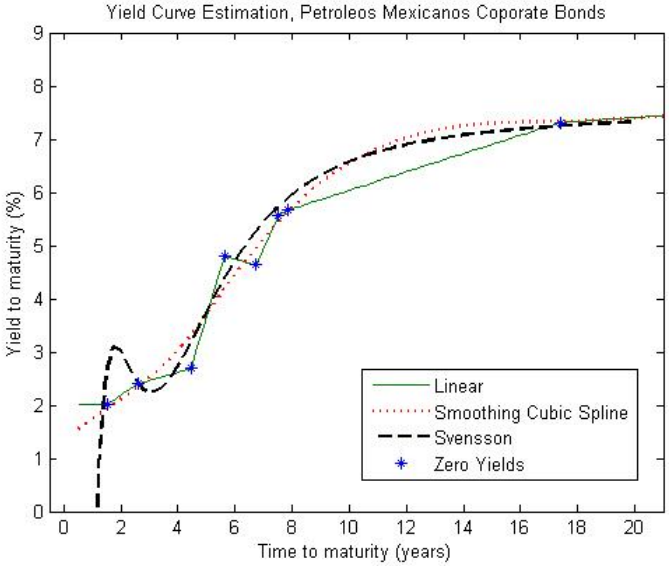

Figure 7.6: Three different estimation techniques are applied onto a set of zero yields to obtain estimates of the yield curve: linear interpolation between estimated zero yields, a smoothing cubic spline and the Svensson model. The figure illustrates an example where very poor estimates of zero yields have been obtained. The Svensson model here produces an overfitted curve, as does the linear interpolation which by construction intersects each point. The smoothing spline however produces a smooth curve where only a slight kink can be detected toward the short end of the curve.

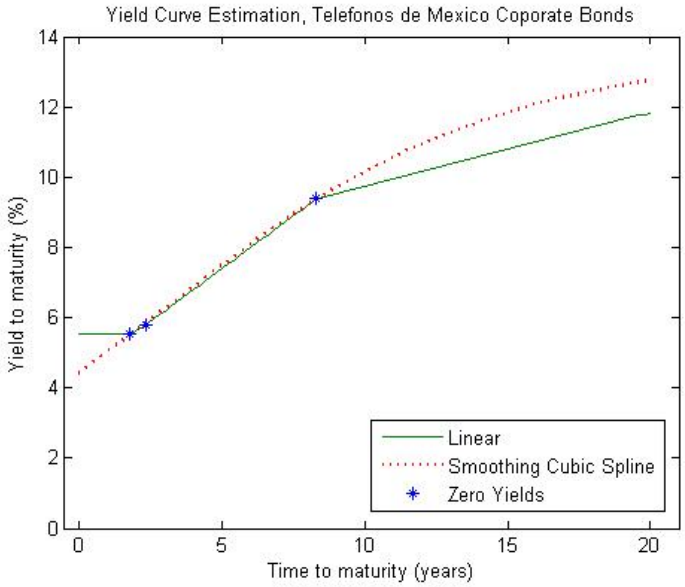

Figure 7.7: Example yield curve of the corporate issuer Telefonos de Mexico. This figure illustrates how the smoothing spline can produce reasonable estimates of the yield curve at times when an estimate by the Svensson model cannot even be obtained. The simple characteristics of linear interpolation are once again evident.

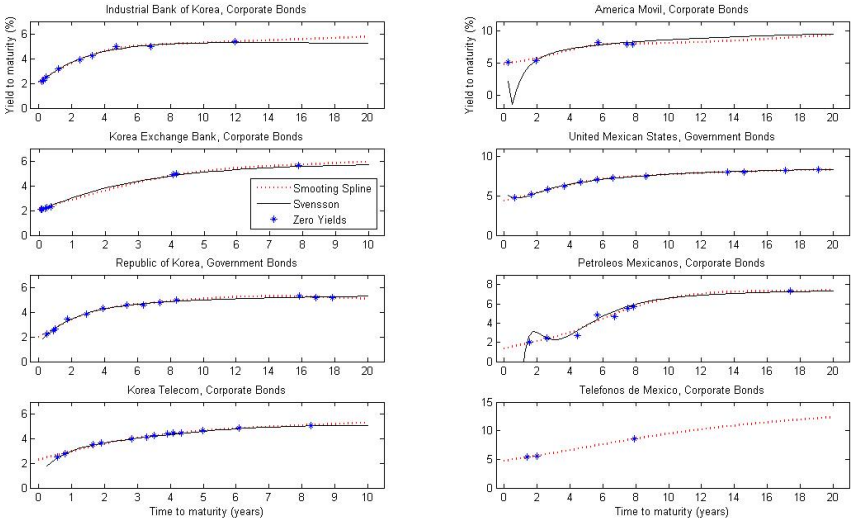

Figure 7.8: Term structures estimated on the 30th of April 2010 for all of the eight issuers covered in the sample are illustrated. It is clear that even for the most recent period, where data quality in general is good for all issuers, the Svensson model sometimes creates unreliable results whereas the smoohing spline yields reasonable estimates for all issuers. For *Telefonons de Mexico* an estimate could not even be obtained with the Svensson model due to the limited amount of data points.

Judging from obtained curve estimates, the choice of which technique to apply is not a difficult one. The smoothing cubic spline outstands the simplicity of the linear interpolation and produces significantly more stable estimates than the Svensson model. The Svensson model would surely be the best choice for consistently good sets of zero yields but the quality of the data used in this study is too poor for the technique to be suitable. The smoothing spline is found to be a good compromise between model sophistication and loss of already scarce data. Figure 7.8 shows estimates obtained with the smoothing cubic spline and the Svensson model for all issuers on the most recent day of the sample. The exception is Telefonos de Mexico where the Svensson model could not produce any estimate due to data restrictions.

The condition of at least three valid yield notations to attempt an estimate of the yield curve resulted in a somewhat reduced set of data. However, as the quality thereof varies between issuers, some sets remain fairly intact whereas others are more affected. A smaller fraction of full term structures as well as single notations from the estimated yield curves must also be removed due to unsatisfactory results, which even further decreases the number of remanining yield notations. With the smoothing spline as selected estimation method, and where all identified outliers have been removed, the respective histories of estimated term structures show the characteristics presented in table 7.1. A total of 92 constant maturity yield time series, all with varying fractions of missing values, have now been obtained.

A couple of examples of how estimated yields develop over time, for a reduced set of maturities, are shown in figures 7.9 and 7.10. Periods of steeper as well as flatter yield curves can be identified as vertical distances between time series vary. Generally, longer term series lie above shorter term series, implying the most common scenario of an upwards sloping curve.

Issuer Name	Observed Dates	Tenors	# Time Series	# Missing (%)
Industrial Bank of Korea	01/08/2005-30/04/2010	3M-20Y	12	22.76%
Korea Exchange Bank	29/03/2006-30/04/2010	3M-10Y	10	37.80%
Korea Telecom	04/05/2005-30/04/2010	3M-10Y	10	20.46%
Republic of Korea	03/05/2005-30/04/2010	3M-20Y	12	6.90%
Total Korea	-	-	*44*	*21.33%*
América Móvil	25/08/2008-30/04/2010	3M-20Y	12	68.60%
Petróleos Mexicanos	03/05/2005-30/04/2010	3M-20Y	12	2.07%
Telefonos de Mexico	25/08/2008-30/04/2010	3M-20Y	12	68.31%
United Mexican States	03/05/2005-30/04/2010	3M-20Y	12	2.88%
Total Mexico	-	-	*48*	*35.46%*
Total All	-	-	**92**	**28.70%**

Table 7.1: Static information of resulting sets of term structures for each of the selected issuers (1). Columns in the table correspond to the dates between which term structures are estimated (2), for which tenors the term structure has been estimated (3), the number of raw time series per issuer (4) and the fraction of missing values in the data set after the estimation has taken place (5) (where a complete 5 year history corresponds to a full set of data).

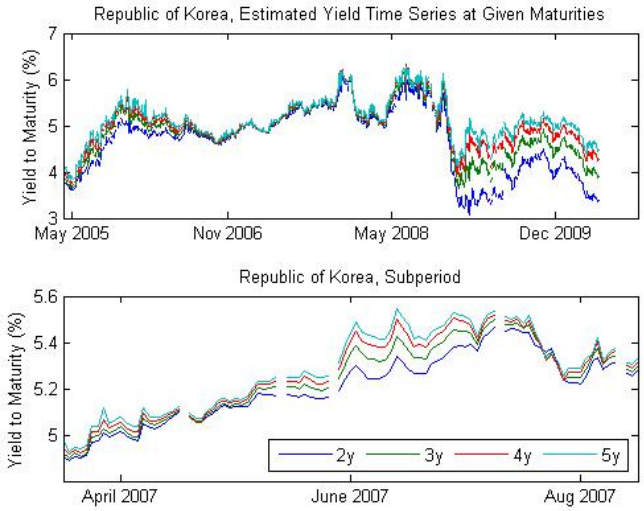

Figure 7.9: Estimated time series corresponding to maturities 2y to 5y are illustrated for the Korean government bond curve. In the lower part of the figure, a subperiod is shown where a fraction of the observations are missing.

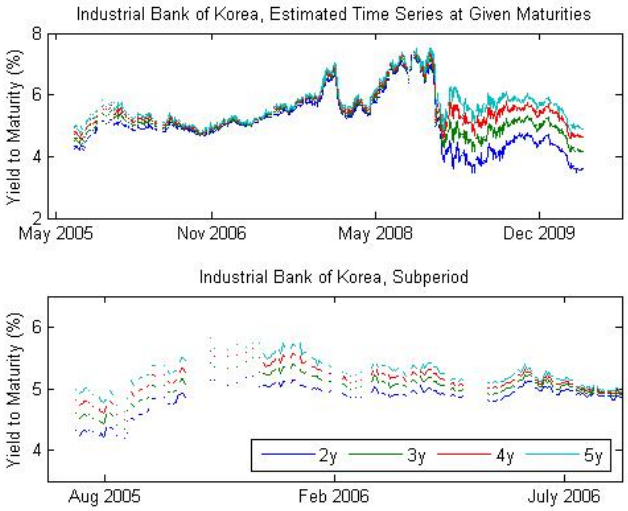

Figure 7.10: Estimated time series corresponding to maturities 2y to 5y are illustrated for the corporate bond curve of Industrial Bank of Korea. The lower part of the figure shows a subperiod which contains a notable amount of missing positions.

Issuer Name	# Missing (%)	Issuer Name	# Missing (%)
Industrial Bank of Korea	32.66%	América Móvil	70.17%
Korea Exchange Bank	49.14%	Petróleos Mexicanos	3.60%
Korea Telecom	29.27%	Telefonos de Mexico	69.88%
Republic of Korea	12.38%	United Mexican States	4.44%
Total Korea	*30.10%*	*Total Mexico*	*37.02%*
Total All	**33.71%**		

Table 7.2: Resulting set of yield changes for each of the eight issuers computed from the set of obtained issuer term structures. It is clearly notable that the fraction of missing values significantly increases as yield changes are computed. A complete 5 year history corresponds to a full set of data.

As yield changes are computed from estimated term structures, an increase in missing positions can be observed (as previously discussed in section 4.2). Presented in table 7.2 are the fractions of missing values in the set of yield changes, displayed per issuer, in total per region and in total over the entire data set. The additional loss of information, as yield changes are computed from yields, is especially visible among a couple of the Korean issuers. The fraction of missing observations displayed in table 7.2 will remain through the GARCH filter and are thus the same positions which later on will be filled by the iterative backfilling procedure.

7.2 Backfilling Performance

Below follows a brief presentation of the GARCH(1,1) filter application onto constant maturity time series of yield changes. This is followed by results obtained by the PCA-based backfilling procedure, applied onto the various described sets and subsets of GARCH residuals.

7.2.1 GARCH(1,1) Filtering

When GARCH properties have been established (from which results are presented in Appendix A), all time series of yield changes are filtered with the procedure described by Penzer (2007) and recited in section 4.2.1. A few examples of filtered GARCH residual time series of varying length is illustrated in figure 7.11 and it is upon time series of this character that the backfilling procedure will be applied. Further assessing the total of 92 filtered time series, it is found that their *average* mean and volatility (based only on available observations for each time series) are -0.036 and 1.023 respectively. The standard deviation of these measures are 0.049 for the mean and 0.034 for the volatility. Thus, some sampling error is present but on average time series are approximately (0,1) distributed.

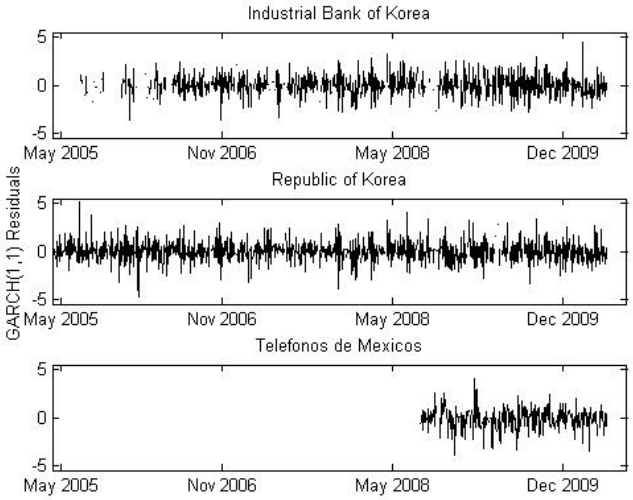

Figure 7.11: Three examples of filtered GARCH residuals are illustrated, all of varying quality. The length of the time series as well as the frequency of notations differs among all of the three issuers. Periods of higher and lower quality can be seen with the naked eye.

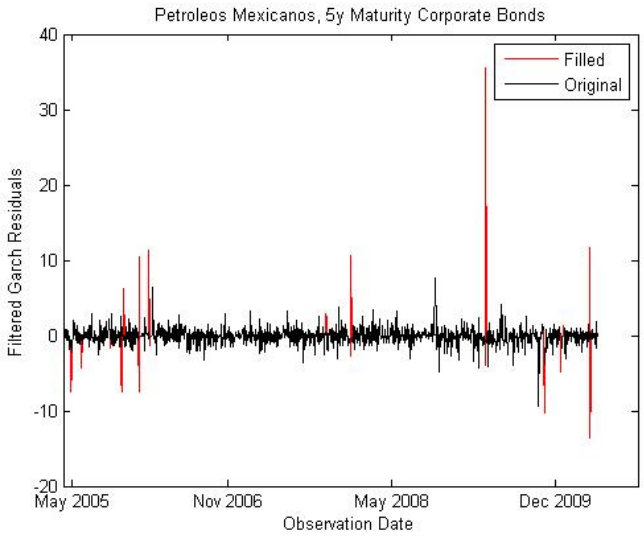

Figure 7.12: GARCH residuals for the corporate 5y maturity time series of Petróleos Mexicanos. After 1000 iterations convergence of the procedure has not yet been reached and imputed values are significantly larger than original values of the same time series.

7.2.2 Filling of Time Series

No Decomposition

Applying the backfilling procedure directly onto the entire set of data is in some regards a beneficial scenario. In such approach, no correlations would be neglected and no extra time would need to be spent on decomposing the data into smaller subsets. However, when applying the backfilling procedure onto the 92 estimated GARCH residual time series, results are not satisfactory. 12 optimal Principal Components are estimated but as iterations are initiated, convergence is not reached. Imputed estimates rapidly exceeds observed values and show no sign of slowing down.

The backfilling procedure is initiated once more on the set, but is now stopped after 1000 iterations. Time series are at this point reviewed and it is evident that imputed values are largely overestimated. The smallest and largest imputed values are -43.7 and 51.1, both of far larger size than what is reasonable for a set of (0,1) distributed GARCH residuals. As the procedure is allowed to continue for even longer time, imputed values grow even larger. An example is illustrated in figure 7.12 where imputed values are clearly overestimated.

Regional Decomposition

The regional decomposition results in two different subsets corresponding to South Korean and Mexican issuers. The Korean subset contains of 44 time series for which 4 optimal PC:s are estimated and the Mexican subset contains of remaining 48 time series for which the larger number of 8 PC:s are obtained. The algorithm does however not reach convergence for any of the two subsets. Again letting the algorithm iterate 1000 times for each of the subsets, similar results are obtained as for the previous scenario. Many of the imputed values are far larger than observed ones, reaching up to -42.5 and 43.8 for Korea, and -38.3 and 20.3 for Mexico. The convergence factor has at this stage reached 0.023 but is in fact

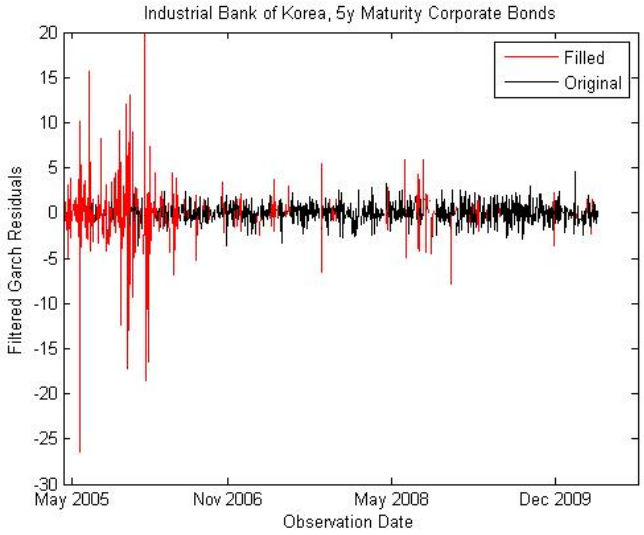

Figure 7.13: GARCH residuals for the corporate 5y maturity time series of Industrial Bank of Korea. After 1000 iterations convergence of the procedure has not yet been reached and imputed values are significantly larger than original observations of the same time series.

increasing again as the procedure is stopped. An example time series is again illustrated in figure 7.13, this time for the corporate issuer Industrial Bank of Korea.

Two to five Year Tenors

When visually reviewing the constant maturity time series of yield changes, it can be confirmed that tenors on the long and short ends of the curves in general display a more volatile behaviour than tenors at the midmost part of the curves. Removing these excess volatile time series, in order to reduce the risk of error population in the PCA-based backfilling routine, is thus justified. Reducing the number of tenors for each of the issuers to those between two and five years, results in one set of 32 time series, 4 belonging to each issuer, and with 4 optimal PC:s estimated for the entire set.

Unfortunately, convergence remains unreached also for this decomposition and imputed positions are again clearly too large. Similarly to previous approaches the convergence factor shows no sign of decreasing to zero and after 1000 iterations, values in the range of -31.3 to 35.0 are imputed in formerly missing positions. (As results are similar to those of previous approach it is chosen not to include any illustration for current subset.)

Regional and Tenor-wise Decomposition - Two-Step Filling

As the first decomposition according to *regions* and *tenors* is applied, 8 subsets of 4 time series each are obtained. Complexity is thus significantly reduced prior to the application of the backfilling procedure. This fact is also reflected in a low number of optimal PC:s for the different subsets, all for which convergence is reached within just a few seconds. The total fraction of missing values in the set decreases from 33.71% to 15.66%.

Issuer Name	PC:s	Itera-tions	# Missing Pre Fill (%)	#Missing After 1st Run (%)	# Missing After 2nd Run (%)
Industrial Bank of Korea	1	18	31.33%	22.14%	21.72%
Korea Exchange Bank	1	19	48.83%	22.22%	21.76%
Korea Telecom	1	19	30.47%	22.05%	21.58%
Republic of Korea	1	18	11.86%	9.79%	9.32%
Total Korea	-	-	*30.62%*	*19.05%*	*18.60%*
América Móvil	1	42	69.59%	21.45%	5.24%
Petróleos Mexicanos	1	18	3.40%	2.94%	2.40%
Telefonos de Mexico	2	138	69.84%	21.55%	5.66%
United Mexican States	1	10	3.49%	3.13%	2.99%
Total Mexico	-	-	*36.58%*	*12.26%*	*4.07%*
Total All	-	-	**33.71%**	**15.66%**	**11.33%**

Table 7.3: Filling results from the two-step decomposition of the original data set. The number of optimal PC:s (2) and necessary iterations (3) are shown as the second run of the procedure is performed on each of the issuer sets of time series in the sample. Furthermore, the fraction of missing values is presented before (4), in between (5) and after (6) the filling has been performed twice, here illustrated per issuer which was found to be the most relevant representation.

As the second decomposition is performed, where each *issuer* is treated separately, 8 subsets consisting of 4 time series each are again created. Table 7.3 presents the number of principal components and the number of iterations needed to reach convergence solely for this second decomposition (as the issuer term structure is the representation of interest). The fraction of missing values is however presented for the various issuers prior to filling, after the first run and finally after the second run of the backfilling procedure.

Clearly notable in table 7.3 is the significant reduction in missing values among Mexican issuers. The total fraction of missing observations here decreases from 36.58% to 4.07% after applying the procedure twice, whereas for Korea the fraction decreases from 30.62% to 18.60%. Although the quality of the imputed values is the most crucial aspect, it is in this study also desired to fill a large amount of missing values.

Figure 7.14 illustrates an example time series of GARCH residuals, where imputed values (combined from the first and second run) are displayed in red. Judging from the size of imputed values they seem to be reasonable estimates. However, as the figures basically illustrate noise and since true values for imputed positions are unknown, further analysis of their quality is difficult to make in this dimension.

Apart from the fear that correlations between time series are not optimally used as they are shuffled around, the main drawback of this choice of decomposition is the relatively low number of imputed values. This fact becomes clear when inspecting figure 7.14 where large gaps still are visible. Fewer time series are here treated simultaneously by the backfilling routine and depending on the number of optimal Principal Components, the fraction of the data which remains unfilled is quite significant. Imagine for example that two Principal Components are found optimal for filling of a data set consisting of 4 time series. This means that all rows containing 1 or 2 observations will be excluded from the filling procedure. Since 4 observations here corresponds to a complete row, only rows with excatly 3 observations will be filled by the routine.

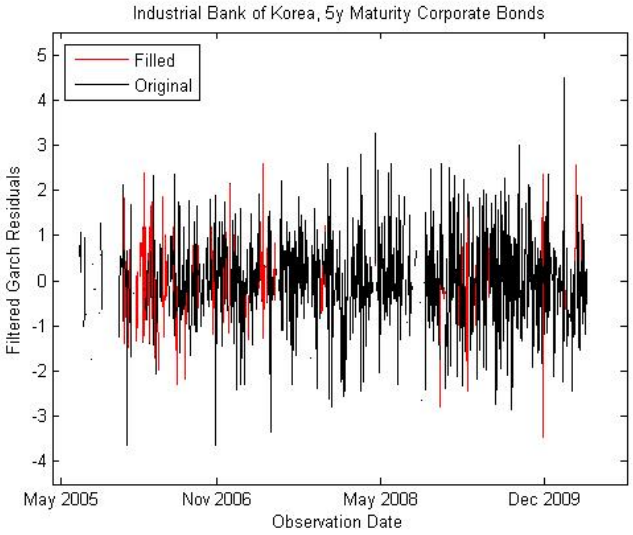

Figure 7.14: The figure illustrates the 5y maturity time series of the corporate issuer Industrial Bank of Korea, after application of the backfilling procedure. Displayed in red are all imputed values after the first and second sun of the procedure. No additional time series have been used during the filling procedure.

Including Additional Risk Factors - Two-step Filling

To increase the fraction of positions being filled by the algorithm, additional time series are placed in subsets created by the the first decomposition. As the two regions covered by the data correspond to South Korea and Mexico the additional time series are: *USD/KRW* - Currency exchange rate between US Dollar and Korean Won, *Kospi 200 Index* - Capitalization weighted index based on 200 Korean stocks corresponding to 93% of the Korea Stock Exchange market value, *USD/MXN* - Currency exchange rate between US Dollar and Mexican Peso and the *Mexbol Index* - Capitalization weighted index of leading stocks traded on the Mexican Stock Exchange.

With additional risk factors taken into account, the first run of the procedure is now performed on 8 different subsets each consisting of 6 time series (instead of 4 as in previous scenario). As a result of adding more information to the data set, the number of optimal PC:s in this first run of the procedure generally increases from 1 to 2. Since more information is added to the data set it is however found that this increase is reasonable. Convergence is once again easily reached for all subsets and the total fraction of missing values decreases from 33.71% to astonishing 5.76%.

As the second decomposition is applied, additional time series are removed and 8 subsets of 4 time series each are again created. Results are presented in table 7.4 and a clear improvement can be seen. The number of missing values which are filled by the procedure have now significantly increased and in the total set, only 5.53% of the observations remain missing after the filling procedure as oppose to 11.33% when no additional risk factors were used.

Issuer Name	PC:s	Itera-tions	# Missing Pre Fill (%)	#Missing After 1st Run (%)	# Missing After 2nd Run (%)
Industrial Bank of Korea	1	17	31.33%	9.69%	9.27%
Korea Exchange Bank	1	18	48.83%	9.69%	9.27%
Korea Telecom	1	18	30.47%	9.69%	9.27%
Republic of Korea	1	18	11.86%	9.54%	9.11%
Total Korea	-	-	*30.62%*	*9.65%*	*9.23%*
América Móvil	2	0	69.59%	2.40%	2.40%
Petróleos Mexicanos	1	17	3.40%	0.61%	0.50%
Telefonos de Mexico	2	0	69.84%	2.40%	2.40%
United Mexican States	2	0	3.49%	2.05%	2.05%
Total Mexico	-	-	*36.58%*	*1.87%*	*1.84%*
Total All	-	-	**33.71%**	**5.76%**	**5.53%**

Table 7.4: Filling results from the two step decomposition of the original data set, with additional risk factors included. The number of optimal PC:s (2) and necessary iterations (3) are shown as the second run of the procedure is performed on each of the issuer sets of time series in the sample. Furthermore, the fraction of missing values is presented before (4), in between (5) and after (6) the filling has been performed twice, here illustrated per issuer, which was found to be the most relevant representation.

Something that might appear strange in table 7.4, are the "zero" iterations noted for some of the issuers. This simply implies that the restriction on the number of observations needed for a day to be included in the filling procedure (in this case 3, since k equals 2), is not fulfilled for any of the days that still contain missing values. All values that are possible to fill with the means of PCA are thus done so during the first run of the procedure. The time series illustrated in figure 7.15 is the same time series as shown in figure 7.14, though this time with estimates obtained with the help of additional information.

Comparing figures 7.14 and 7.15, it appears as if positions which are successfully estimated in both scenarios are identical. This would in turn imply that adding additional time series will in fact not distort imputed values but will simply help to fill a larger number of missing positions. To illustrate this further, an example is shown in figure 7.16 where the same two sections of a time series is filled without and with additional information. The differences between imputed values are seemingly small but with the help of additional information, more values can be filled.

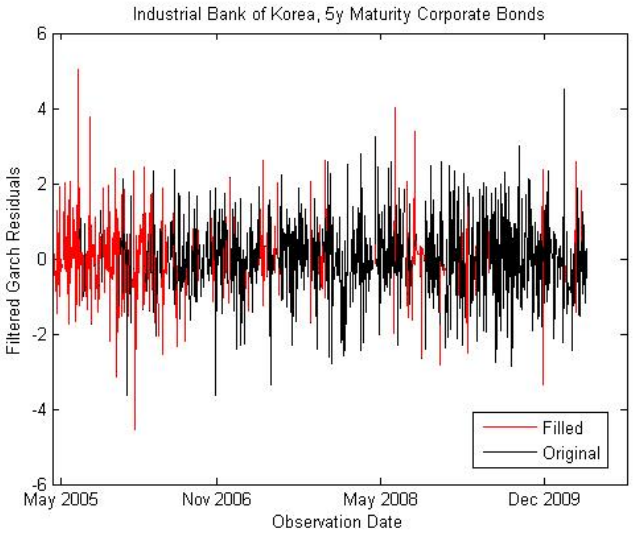

Figure 7.15: The figure illustrates the 5y maturity time series of the corporate issuer Industrial Bank of Korea, after application of the backfilling procedure. Imputed values from the first and second run of the procedure are displayed in red. Additional time series have been used to increase the fraction of filled values.

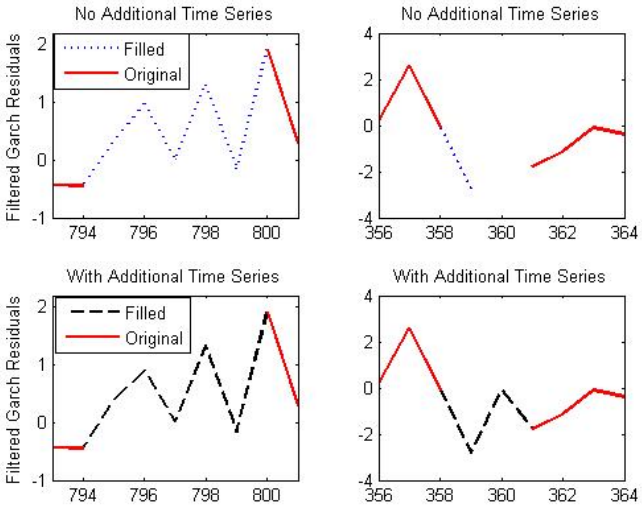

Figure 7.16: Gaps in one of the original time series have here been filled with and without the presence of additional risk factors. Differences between imputed values (comparing the upper and lower parts of the figure) are seemingly small, but the amount of positions which are imputed is higher as additional time series are included.

Region	Issuer Name	Avg. Sum Sq. Diff.
Korea	Industrial Bank of Korea	0.00396
	Korea Exchange Bank	0.00356
	Korea Telecom	0.00552
	Republic of Korea	0.00360
Mexico	América Móvil	0.28930
	Petróleos Mexicanos	0.65125
	Telefonos de Mexico	0.29166
	United Mexican States	0.36266

Table 7.5: Average sum square differences between unitely imputed values as the filling procedure has been performed without and with additional time series. Results show that differences in general are small for Korean issuers but significantly larger for Mexican issuers, something which is believed to be related to the much larger fraction of originally missing values among Mexican issuers.

Unfortunately, this is not always the scenario as some of the issuers illustrate larger differences between the two approaches than what can be seen here. To further assess this matter, the average sum square differences between values imputed without and with additional risk factors, are assessed for each issuer and presented in table 7.5. Korean issuers in general show very small differences between the two approaches whereas differences are more significant for Mexican issuers. This is rather believed to be related to the fraction of initially missing values in the data set being filled. This fraction is notably higher among a couple of Mexican issuers and since a regional (and tenor-wise) decomposition is applied this will influence time series of the same tenor, for other issuers of the same region. A higher level of originally missing values makes the difference of including additional time series even more distinct as the relative amount of new information will be very high, something which might steer imputed values more than desired.

7.3 Term Structure Reconstruction

Figures 7.17 and 7.18 illustrate reconstructed histories of term structures for the Republic of Korea and Industrial Bank of Korea, from a historical yield perspective. Both figures show reasonable yield notations on positions that were previously missing. Illustrated time series are generated without the presence of additional risk factors, meaning that a larger fraction of missing observations finally had to be filled through simulation from factor distributions (see Appendix B). To assess the credibility of estimated missing values, compare with figures 7.9 and 7.10 on pages 59 and 60.

Figure 7.19 instead shows an example of the impact of including additional risk factors in the backfilling algorithm, on the final yields. Time series corresponding to the 2y to 5y maturity yields for the corporate issuer Industrial Bank of Korea are again illustrated over the observed period. Comparing the lower part of the plot with that of figure 7.18, a slight difference in the shape of the time series can be detected, mainly in the leftmost half of the plots.

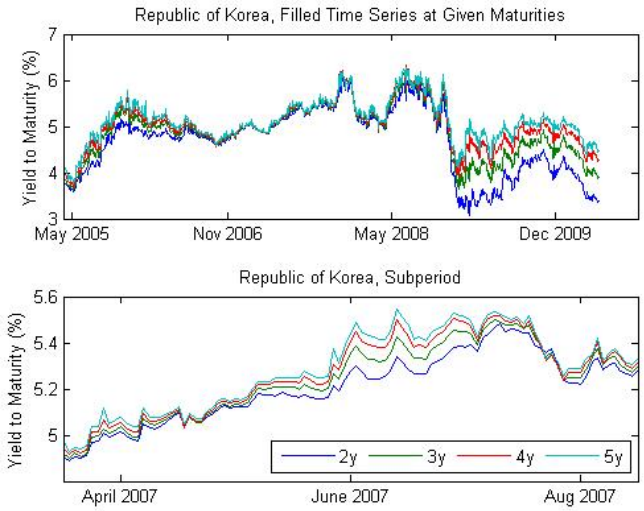

Figure 7.17: Estimated time series corresponding to maturities 2y to 5y, where missing values have been imputed with the suggested filling algorithm, here illustrated for the Korean government bond curve. Values are imputed without additional risk factors. In the lower part of the figure, a subperiod is shown where missing values, as shown in figure 7.9, have been filled.

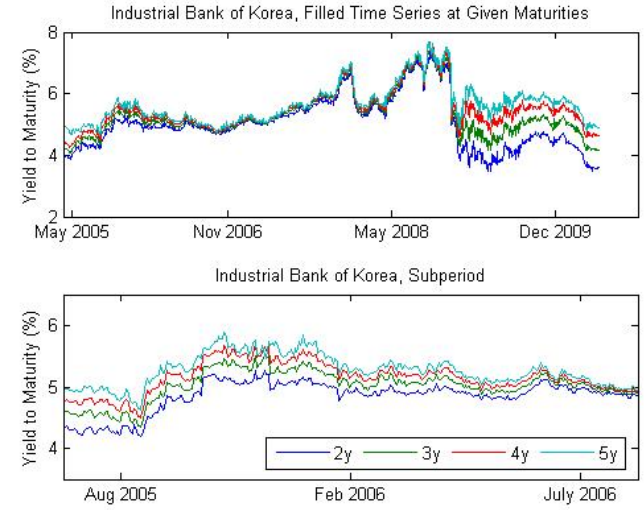

Figure 7.18: Estimated time series corresponding to maturities 2y to 5y, where missing values have been imputed with the suggested filling algorithm, here illustrated for the corporate yield curve of Industrial Bank of Korea. Values are imputed witout additional risk factors. In the lower part of the figure, a subperiod is shown where missing values, as shown in figure 7.10, have been filled.

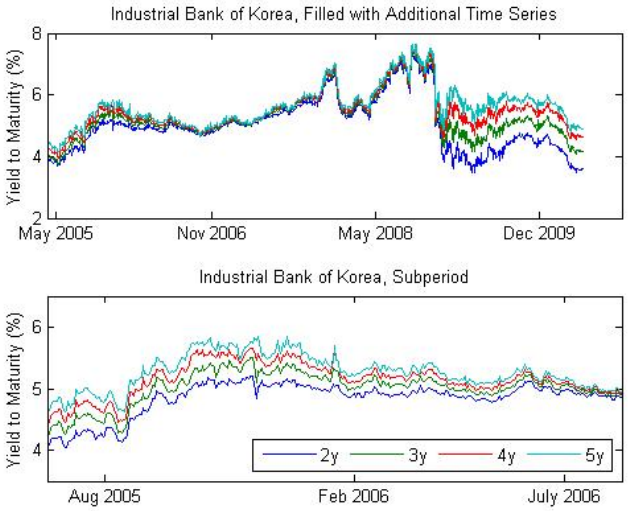

Figure 7.19: Estimated time series corresponding to maturities 2y to 5y, where missing values have been imputed with the suggested filling algorithm, again illustrated for the corporate yield curve of Industrial Bank of Korea. Additional time series have been used and a slight difference can be seen through visually comparison with figure 7.18.

If one looks closely, a dip can be observed for the 5y maturity time series in the lower part of figure 7.18. Approximately 1/3 into the subperiod (corresponding to $\approx$ November 2005) the 5y maturity yield in fact decreases to a level below the 4y and even the 3y yield. Lower yields for longer maturities do sometimes occur, but generally on the long end of curves covering a broader range of maturities. As the figure illustrates yields only for a limited number of fairly short maturities, it is considered unlikely to see the yield of the longest maturity dip below those of shorter maturities, in turn implying that the imputed value here might be incorrect. Such dip cannot be seen in figure 7.19, which in turn speaks in favour of the scenario where additional risk factors are included.

In contrast to examples illustrated so far, where results appear good, a couple of term structures are less successfully filled. To give an example, a part of the filled history of the Mexican corporate issuer América Móvil is illustrated in figure 7.20. América Móvil is the issuer with the highest fraction of originally missing values as more than 70% of its yield changes were unobserved. As historical yield notations were available during the period 25/08/2008-30/04/2010, approximately the left half of figure 7.20 has been entirely constructed with estimates from the PCA-based filling routine. This turned out to be too much for the filling algorithm. Even though imputed GARCH residuals appear reasonable (though this dimension practically consists of pure noise and therefore is difficult to correctly assess), it becomes clear when the term structure dimension is reconstructed, that filled values don't make much sense.

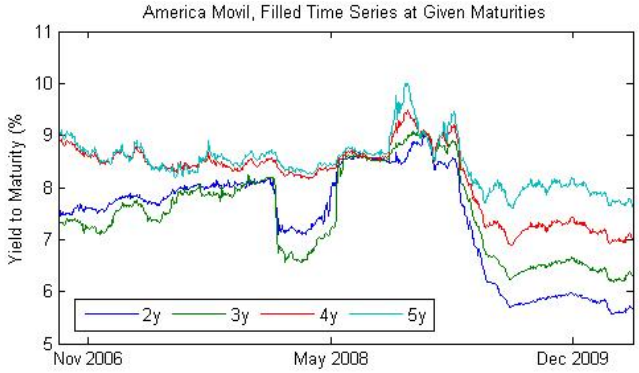

Figure 7.20: Illustration of a less successfully filled history of term structures, in this case for the corporate issuer América Móvil. The issuer corresponds the one with the highest fraction of originally missing values where roughly only two years of observations were available.

As a final evaluation of the resulting yield estimates, it remains to review them from a term structure perspective. Figures 7.21 to 7.23 show a few examples of where parts of, or entire term structures, estimated per issuer, have been imputed. Here illustrated together with observed term structures of adjacent days. The first thing one notes is the shape of the imputed yield curves, which correspond well with originally observed curves. The distance to these curves is also reasonable and imply that daily changes of yields at constant maturities are in range of what can be expected. Unfortunately it is difficult to further discuss the correctness of the imputed values. Sine they are unknown there is no "true" value for such positions which makes visual inspection one of the strongest tools to assess the correctness of curves and time series.

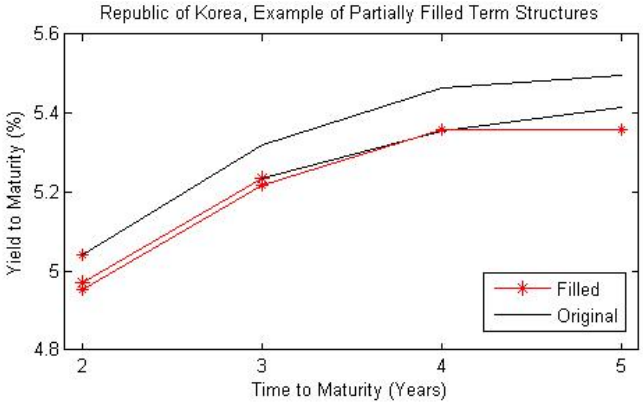

Figure 7.21: The figure illustrates partially and entirely filled yield curves together with originally observed ones. Also in this example simulated curves are aligned with observed curves from adjacent days and seemingly display a credible shape.

Especially interesting is figure 7.23, where the same two yield curves have been filled without and with additional time series respectively. The difference among imputed GARCH residuals is in general very small for the issuer and here it can be seen that also the final yield curve representations seem very close if not identical.

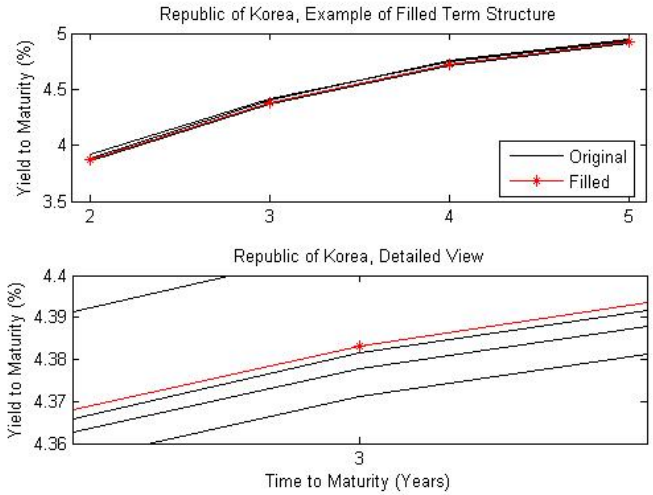

Figure 7.22: Imputed yield curve for Korean government bonds is illustrated (in red) among already observed yield curves (in black) from adjacent days of the history. The imputed yield curves display a credible behaviour and correspond to the shape of observed curves.

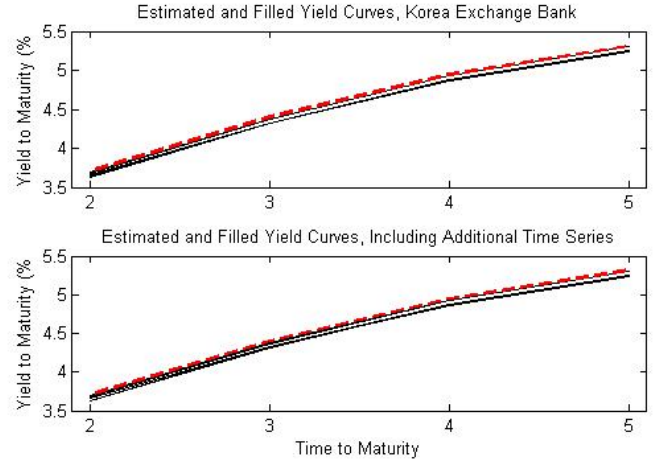

Figure 7.23: Examples of imputed yield curves are illustrated for the corporate issuer Korea Exchange Bank. Above plot shows estimates produced without additional risk factors and below plot shows the same yield curves estimated in the presence of additional information. Fortunately, no significant difference can be noted, implying that the inclusion of additional time series can be helpful.

Summarized Results

Zero yields are estimated with the Boostrap method based on coupon bearing yield notations of a number of different issuers. The method generally produces reasonable zero yields but as the number of observed bonds decreases and their positioning relative to each other becomes more uneven, the quality of estimated yields decreases.

Three different techniques are thereafter applied onto zero yields for estimation of issuer specific yield curves. It is found that the cubic smoothing spline with a smoothing factor $\lambda = 0.6$ gives the best fit to the data and produces consistent results. Yield curves are thereafter estimated for all issuers and a set of 10-12 time series of historical yields at constant maturity (y_t) is obtained for each issuer. With a simple computation, corresponding time series of yield changes (x_t) are created.

Assuming they follow univariate GARCH(1,1) processes, each time series x_t is filtered to obtain GARCH residuals (ϵ_t). PCA-based backfilling is performed on these residuals and depending on the set of data the procedure is applied on, the performance of the procedure varies. It is quickly realized that a decomposition of the real data used in this study is necessary for the procedure to reach convergence. After trying a number of different approaches, a two-step filling based on two different decompositions is found to be the most successful. Attempting to include additional risk factors in the backfilling procedure generally gives positive results. For Korean issuers the effect seems solely positive as a larger amount of positions can be filled. However, estimates produced for Mexican issuers clearly differs from those obtained without additional time series, although something rather believed to depend on the higher fraction of missing values among Mexican issuers.

Recursive reconstruction of yield changes and thereafter term structures generally produces reasonable results. However, time series with a higher fraction of originally missing values display less successful outcome. Assessing the credibility of obtained results is mainly done visually, other approaches are difficult since missing observations truly are unknown and meaningful numerical comparisons are difficult to make.

Chapter 8

Conclusions and Discussion

In this Master's Thesis a procedure has been designed for estimation of issuer specific term structures with scarce availability of data. The aim has been to obtain complete histories of yield curves based upon which financial risk measures can be computed. To reach this objective, the main focus has been placed upon filling of missing values, for which a specific algorithm has been developed.

As term structures were estimated based on bootstrapped zero yields, it very early became clear that the most sophisticated method does not always produce the best result. The appropriate choice of model is highly dependent on the characteristics and quality of the data and should carefully be considered. The smoothing cubic spline which was chosen in this study, turned out to be the best technique to obtain reasonable estimates without too much loss of data. Data which was scarce already from the beginning. However, unstable estimates at curves' ends made it possible to only use the midmost part of the estimated term structures to obtain sound results in the subsequent backfilling procedure.

To avoid estimation errors, which here turned out to affect the performace of the latter applied filling procedure, corresponding CDS notations or other suitable proxies could instead be used. Such measures can however be difficult to find for smaller issuers, where finding bond notations already is a challenge. The term structure estimation technique is also a part of the study with room for improvements as this aspect was somewhat de-emphasized in favour of the backfilling routine.

The designed backfilling algorithm is proven to yield successful results for a number of synthetic data sets of which underlying factor structures are known. The fraction of missing values, the number of underlying factors and the amount of noise in the data set are all shown to have an impact on the model's performance. Applied on a true set of data, convergence was however found difficult to reach. This could in fact be related to the amount of noise which can be high among daily financial data. Even more so due to additional errors assumed to be introduced during the term structure estimation. It is also commonly known that Principal Component Analysis is a non-robust method, and the data set might have contained too much noise for the filling routine to correclty reach convergence. As an alternative to a more profound yield curve estimation to minimize errors, robust PCA could be used to reduce the impact of noise and possible outliers throughout the backfilling algorithm.

The higher the fraction of missing values, the larger is the risk that these values are not correctly estimated and one should carefully consider if it is justified to apply the procedure on data sets of too low quality. On sets with a lower fraction and a more sparse structure of missing values, it has been shown in this study that satisfactory results are obtainable, but where does one draw the line? Is it possible to define a measure of the character of missing observations and set a threshold for when values still can be filled with the suggested procedure? In this study, term structures are successfully estimated as long as the fraction of missing values is kept on a reasonable level. The actual validity of the estimates is however still to be assessed. A good way of evaluating filled term structures has not been thought of, something left to further investigate.

Also worth to highligt is the question of how many Principal Components to use in the imputation algorithm. The approach presented in this study was found meanigful for the chosen data, but is nevertheless only one among many different possibilities. Difficulties to reach convergence could possibly be related to this question and alternative ways of defining the number of components remains to be tried. As dimensionality of the data was reduced through decomposition, the optimal number of components might have been easier to detect, leading to reasonable results and convergence, but unfortunately also to the neglection of important correlations as time series are separated.

As a final word, the strongest lesson learned in this study is related to the data. Regardless of what one wishes to do, one must be aware that every step of the process will be highly dependent on the chosen data set.

Bibliography

C. Alexander. Market Risk Analysis, Volume II: Practical Financial Econometrics, 2009.

Basel Committee. Amendment to the Capital Accord to Incorporate market risks, 2005.

T. Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

J.C. Cox, J.E. Ingersoll Jr, and S.A. Ross. A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, 53(2):385–407, 1985.

R. Engle. GARCH 101: The use of ARCH/GARCH models in applied econometrics. *Journal of Economic Perspectives*, 15(4):157–168, 2001.

R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica: Journal of the Econometric Society*, 50(4): 987–1007, 1982.

J.C. Figueroa-Garcıa, D. Kalenatic, and C. Lopez-Bello. Missing Data Imputation in Time Series by Evolutionary Algorithms. In *Advanced intelligent computing theories and applications: with aspects of artificial intelligence: 4th International Conference on Intelligent Computing, ICIC 2008, Shanghai, China, September 15-18, 2008: proceedings*, page 275. Springer-Verlag New York Inc, 2008.

D. Filipović. *Term-Structure Models: A Graduate Course*. Springer Verlag, 2009.

Swiss Financial Market Supervisory Authority FINMA. FINMA Circular 08/20 - Market Risks Banks, 2009.

M. Fisher, D. Nychka, and D. Zervos. Fitting the term structure of interest rates with smoothing splines. *Finance and Economics Discussion Series*, 1, 1995.

C. Frisch, W. Knöchlein, G. in Härdle, T. Kleinow, and G. Stahl. *Applied quantitative finance: theory and computational tools*. Springer Verlag, 2002.

B. Grung and R. Manne. Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 42(1-2):125–139, 1998.

D.F. Heitjan. Annotation: what can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4):548, 1997.

P. Houweling, J. Hoek, and F. Kleibergen. The joint estimation of term structures and credit spreads. *Journal of Empirical Finance*, 8(3):297–323, 2001.

J.C. Hull. *Options, futures, and other derivatives*. Pearson Education India, 2006.

H. Hult and F. Lindskog. Multivariate extremes, aggregation and dependence in elliptical distributions. *Advances in Applied probability*, 34(3):587–608, 2002.

H. Hult and F. Lindskog. Mathematical Modeling and Statistical Methods for Risk Management - Lecture Notes. Technical report, Matematisk Statistik KTH, 2007.

M. Ioannides. A comparison of yield curve estimation techniques using UK data. *Journal of Banking & Finance*, 27(1):1–26, 2003.

I.T. Jolliffe. *Principal component analysis*. Springer verlag, 2002.

D. Kondrashov and M. Ghil. Spatio-temporal filling of missing points in geophysical data sets. *Nonlinear Processes in Geophysics*, 13(2):151, 2006.

B.H. Lin. Fitting term structure of interest rates using B-splines: the case of Taiwanese Government bonds. *Applied Financial Economics*, 12(1):57–75, 2002.

F. Lindskog, A. Mcneil, and U. Schmock. Kendall's tau for elliptical distributions. *Credit Risk–measurement, evaluation and management, Bol, Nakhaeizade et al., eds. Physica–Verlag, Heidelberg*, pages 149–156, 2003.

J.H. McCulloch. Measuring the term structure of interest rates. *Journal of Business*, 44(1): 19–31, 1971.

J.H. McCulloch. The tax-adjusted yield curve. *Journal of finance*, 30(3):811–830, 1975.

T.K. Moon. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6):47–60, 1996.

S. Nawalka and G. Soto. Term Structure Estimation, 2009.

C.R. Nelson and A.F. Siegel. Parsimonious modeling of yield curves. *Journal of business*, pages 473–489, 1987.

J. Penzer. Inference for GARCH and Related Models with Incomplete Data. *Department of Statistics*, 2007.

I. Stanimirova, M. Daszykowski, and B. Walczak. Dealing with missing values and outliers in principal component analysis. *Talanta*, 72(1):172–178, 2007.

L.E.O. Svensson. Estimating and interpreting forward interest rates: Sweden 1992-1994. *NBER Working paper*, 1994.

R.S. Tsay. *Analysis of financial time series*. Wiley-Interscience, 2005.

O.A. Vasicek and H.G. Fong. Term structure modeling using exponential splines. *Journal of Finance*, 37(2):339–348, 1982.

G. Welch and G. Bishop. An introduction to the Kalman filter. *University of North Carolina at Chapel Hill, Chapel Hill, NC*, 1995.

X. Zhang, X. Song, H. Wang, and H. Zhang. Sequential local least squares imputation estimating missing value of microarray data. *Computers in Biology and Medicine*, 38(10): 1112–1120, 2008.

Appendix A

Ensuring GARCH(1,1) Properties

As mentioned in section 4.2.1, it must be shown that GARCH(1,1) properties exist among estimated yield changes to justify the GARCH(1,1) assumption and filter application. Ensuring the existence of such properties is the purpose of this appendix.

Figures A.1 and A.2 illustrate the 5y maturity time series together with daily yield changes for the two government curves in the sample. It is easy to see in both figures that yield changes show periods of higher and lower volatilities. Volatility clustering thus clearly exist, which suggests that GARCH(1,1) characteristics might be present in the data.

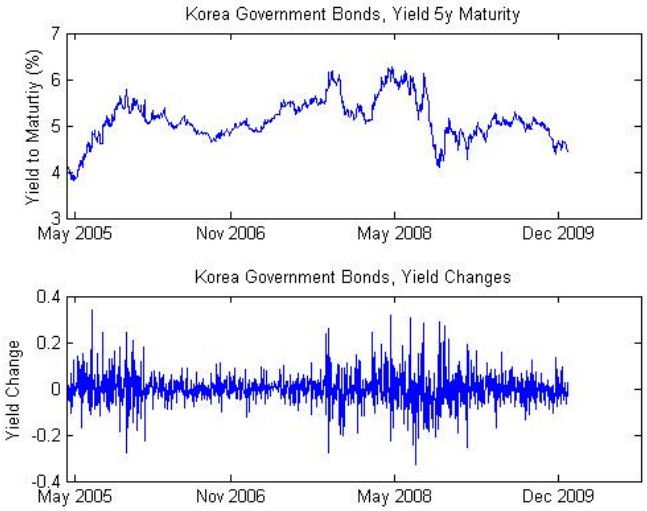

Figure A.1: The 5y maturity yield time series of the Korean government bond curve is illustrated together with its daily changes. Periods of different volatilities are easily detected among yield changes.

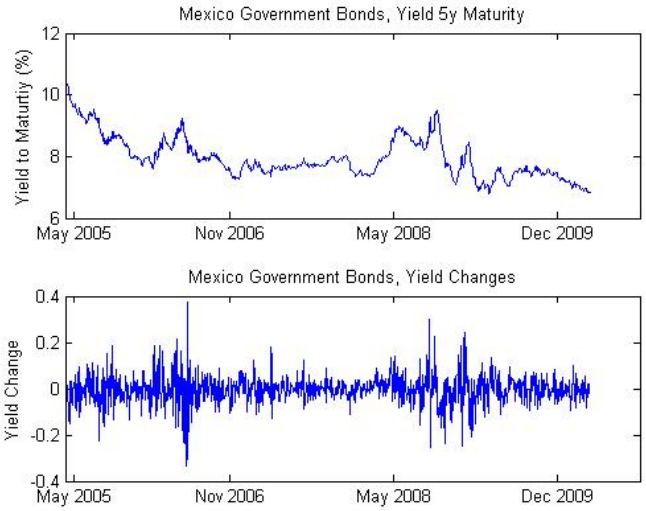

Figure A.2: The 5y maturity yield time series of the Mexican government curve is illustrated together with corresponding daily yield changes. Volatility clusters are evident also in this example.

Further reviewed are autocorrelations, partial autocorrelations and autocorrelations between squared yield changes, here illustrated for the Korean 5y government time series in figures A.3 - A.5. Resulting plots suggest that no autocorrelations exist between observed yield changes, whereas they are clearly noticable among squared changes, thus indicating the existence of a volatility process. Similar implications are given as the function *garchfit()* is applied, where GARCH(1,1) processes successfully are fitted to each of the time series. Coefficients (a, b, c) (as defined in equation (3.5) of section 3.3.1) and their respective standard errors are presented in table A.1.

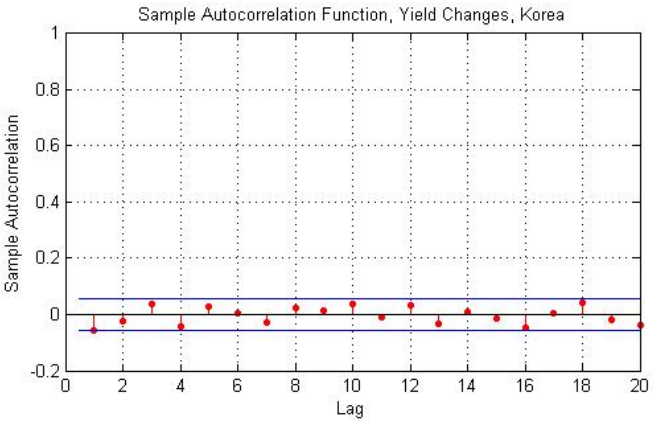

Figure A.3: Sample autocorrelation function illustrated for the 5y maturity time series of yield changes of the Korean government curve. No significant autocorrelations can be detected for the various presented lags.

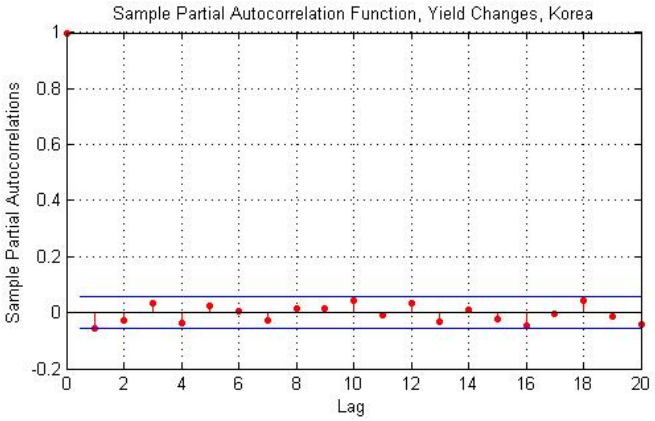

Figure A.4: Sample partial autocorrelation function illustrated for the Korean government 5y maturity time series of yield changes. No significant partial autocorrelations can be detected for the various lags.

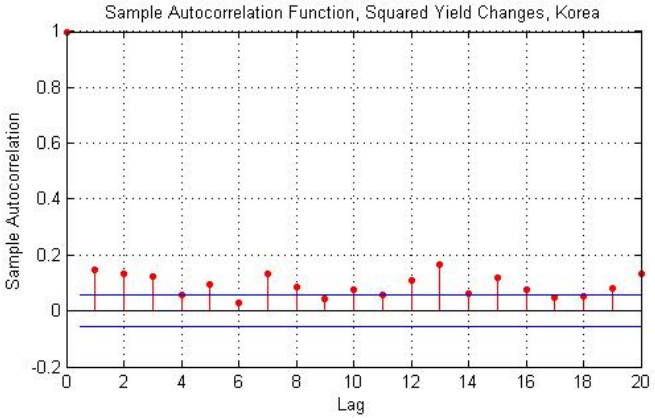

Figure A.5: Sample autocorrelation function of squared yield changes for the 5y maturity Korean government time series. Though no correlations could be detected among yield changes, squared changes clearly display correlations between observations of different lags.

Issuer Name	**c** (10^{-3})	SE_c (10^{-3})	**b**	SE_b	**a**	SE_a
Republic of Korea	0.0170	0.0066	0.0478	0.0051	0.9496	0.0049
United Mexican States	0.0942	0.0190	0.1695	0.0183	0.8103	0.0176

Table A.1: GARCH(1,1) coefficients with respective standard errors (SE_*) for the two example time series, selected to ensure the existence of GARCH(1,1) properties.

Conditional volatilities together with filtered GARCH residuals are finally illustrated in figures A.6 and A.7. Filtered GARCH residuals display a significantly more stable behaviour than time series containing yield changes, although some outliers do appear in the respective example series which can be assumed to be a result of sampling error.

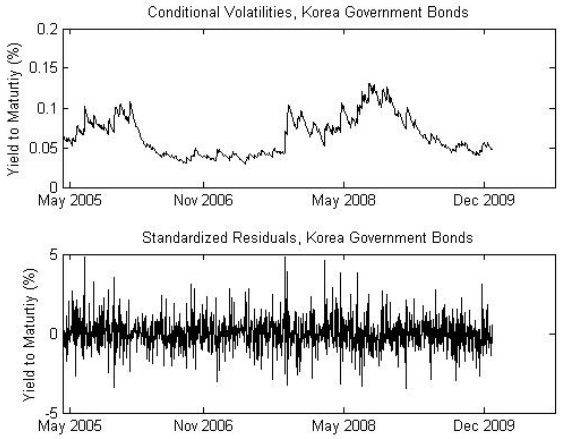

Figure A.6: Conditional volatilities are illustrated together with filtered yield changes of the 5y maturity time series of Korean government bonds. The lower plot illustrates a much more even behaviour where clusters of volatilities noted in figure A.1 have been removed.

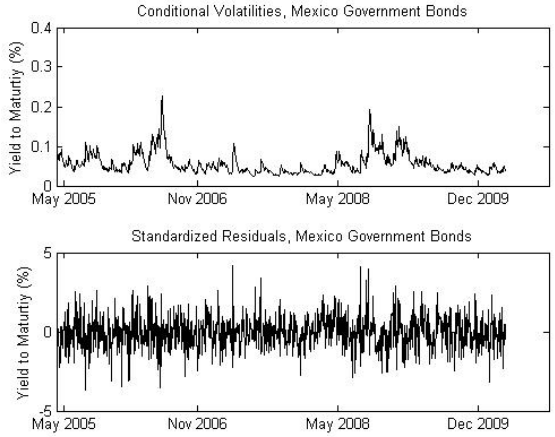

Figure A.7: Conditional volatilities are illustrated together with filtered yield changes of the 5y maturity time series of Mexican government bonds. In the lower plot of filtered GARCH residuals, a much more even behaviour is now seen than compared to corresponding yield changes in figure A.2

Appendix B

Simulation of Remaining Missing Values

Due to the constraint set in the filling routine, where the number observations for a fixed t must be larger than the number of optimal PC:s for that given row to be included, a fraction of values will remain missing as the procedure reaches convergence. The size of this fraction in turn depends on the character of the input data. Optimally, no positions would remain missing after the procedure has been applied but those that still do must nevertheless be filled. This is done by random draws from empirical distributions of Principal Components (PC:s) and factor residuals.

Example

Consider the matrix $\mathbf{X}$ consisting of $i = 1, \ldots, n$ time series observed during the period $t = 1, \ldots, T$. With matrix notation, $\mathbf{X} = (x_{t,i})$, where $x_{\cdot,i}$ further denotes the i:th column vector of $\mathbf{X}$, i.e. the i:th time series. Now let $\mathbf{X}$ be defined by the two underlying factors (f_1, f_2) (observed during the same period) such that each $x_{t,i}$ is given by factor observations $(f_{t,1}, f_{t,2})$ and corresponding set of factor coefficients $\alpha_{1,i}$ and $\alpha_{2,i}$ through the following relation:

$$x_{t,i} = f_{t,1}\alpha_{1,i} + f_{t,2}\alpha_{2,i} + \varphi_{t,i} \tag{B.1}$$

$\varphi_{t,i}$, which corresponds to a residual term of the factor model, can be "backed out" from the model by rewriting the relation such that:

$$\varphi_{t,i} = x_{t,i} - f_{t,1}\alpha_{1,i} - f_{t,2}\alpha_{2,i} \tag{B.2}$$

In this study, the factors f correspond to the set of Principal Components that are obtained as the iterative PCA-based filling procedure has reached convergence and coefficients are the corresponding loadings of the same PCA decomposition. If two PC:s are found optimal for the filling of a certain data set it means that these two components correspond to factors in above representations. After this final filling step, the matrix of GARCH residuals is complete, $\tilde{\mathbf{E}}$ has been created.

To fill entire *rows* of missing values, random draws are made from the Principal Component vectors and values are created by multiplying corresponding loadings with these components. Assuming that the filled part of the matrix contains "true" values, PC vectors are at convergence seen as their respective empirical distribution. If thus two PC:s are found relevant for reconstruction of the data, one random couple of these two components are drawn independent of each other, and independently with replacement for each row where missing values still exist. This gives us a set of factor pairs (f_1, f_2), one for each t where values still need to be filled.

Additionally, a residual term $\varphi_{t,i}$ is added on top of each simulated value, obtained by subtracting the part explained by relevant PC:s from the values of $\mathbf{X}$ at convergence. Independent random draws are made column wise from corresponding residual distribution $\varphi_{\cdot,i}$ for each of the positions that are simulated in corresponding time series $x_{\cdot,i}$.

For *partially empty rows*, which do not contain enough information to be kept in the filling algorithm, but that still have at least one "true" observation, the above procedure is somewhat modified. Instead of just one random draw per row to fill, $N = 1000$ draws are made, both from PC:s and residuals, and values of $\mathbf{X}$ are reconstructed for these N different scenarios. The selected scenario is the one that minimizes the sum of squared differences between that (or those) value(s) that are originally observed and that (or those) that are generated at corresponding positions of each of the N scenarios. Consider for example the row vector $x_{t,\cdot} = (NaN, NaN, NaN, 0.73, -0, 49)$, where three missing positions remain to be simulated. Furthermore let the row vector $x^*_{t,\cdot} = (x^*_{t,1}, x^*_{t,2}, x^*_{t,3}, x^*_{t,4}, x^*_{t,5},)$ be simulated values from one of the N random draws. The squared difference to be minimized over all the N simulations is then given by $(0.73 - x^*_{t,4})^2 + (-0.49 - x^*_{t,5})^2$.