Non-stationary data-driven computational portfolio theory and algorithms

Doctoral Dissertation submitted to the
Faculty of Informatics of the Università della Svizzera italiana
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

presented by
Lars Putzig

under the supervision of
Prof. Illia Horenko

October 2014

Dissertation Committee

Prof. Rolf Krause	Università della Svizzera italiana
Prof. Igor Pivkin	Università della Svizzera italiana
Prof. Patrick Gagliardini	Università della Svizzera italiana
Prof. Alexander Schied	University of Mannheim, Germany

Dissertation accepted on 29 October 2014

Research Advisor	PhD Program Director
Prof. Illia Horenko	**Prof. Igor Pivkin, Prof. Stefan Wolf**

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Lars Putzig
Lugano, 29 October 2014

Abstract

The aim of the dissertation is the development of a data-driven portfolio optimization framework beyond standard assumptions. Investment decisions are either based on the opinion of a human expert, who evaluates information about companies, or on statistical models. The most famous methods based on statistics are the Markowitz portfolio model and utility maximization. All statistical methods assume certain knowledge over the underlying distribution of the returns, either by imposing Gaussianity, by expecting complete knowledge of the distribution or by inferring sufficiently good estimators of parameters. Yet in practice, all methods suffer from incomplete knowledge, small sample sizes and the problem that parameters might be varying over time.

A new, model-free approach to the portfolio optimization problem allowing for time-varying dynamics in the price processes is presented. The methods proposed in this work are designed to solve the problem with less a-priori assumptions than standard methods, like assumptions on the distribution of the price processes or assumptions on time-invariant statistical properties. The new approach introduces two new parameters and a method to chose these based on principles of information theory. An analysis of different approaches to incorporate additional information is performed before a straightforward approach to the out-of-sample application is introduced.

The structure of the numerical problem is obtained directly from the problem of portfolio optimization, resulting in a system of objective function and constraints known from non-stationary time series analysis. The incorporation of transaction costs allows to naturally obtain regularization that is normally included for numerical reasons.

The applicability and the numerical feasibility of the method are demonstrated in a low-dimensional example in-sample and in a high-dimensional example in- and out-of-sample in an environment with mixed transaction costs. The performance of both examples is measured and compared to standard methods, as the Markowitz approach and to methods based on techniques to analyse non-stationary data, like Hidden Markov Models.

Acknowledgements

I would like to express my deepest appreciation and thanks to my advisor Professor Illia Horenko, who was able to make time for discussion even when it was impossible. I would like to thank you for encouraging my research, guide me when needed and letting me roam free when not. Your advice has been beyond priceless. I would also like to thank my committee members, Professor Rolf Krause, Professor Igor Pivkin, Professor Patrick Gagliardini and Professor Alexander Schied for serving as my committee members. I would like to especially thank my colleagues and former colleagues, Susanne Gerber, Dimitri Igdalov, Olga Kaiser, Anna Marchenko and Philipp Metzner for long and sometime intense discussions. For being a distraction, when I needed one, I would like to thank my Lab-partners Dorian Krause and Johannes Steiner. When nothing worked, there was always a coffee break possible and nonsense to discuss.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all the sacrifices they made on my behalf. At the end I would like to express appreciation to my beloved girlfriend Katia Fiorucci, who tolerated my grumpiness, when nothing worked and kept me sane, when I questioned myself.

Contents

1 Introduction

In the wake of the globalization, and with communication channels allowing to access information worldwide within seconds, investors nowadays are confronted with an overwhelming amount of data. Not only do they face a large amount of information, e.g. in the form of news, industrial data, government data or employment statistics, but also a huge number of potential assets, including conventional investments like fixed interest bank-accounts and exchange traded stocks, a wide range of derivatives, Futures and other investment vehicles allowing to speculate on commodities, indices and other tradable or non-tradable underlying values. In fact, even virtual goods can be subject to investments, as can be seen e.g. in the case of bitcoins, a purely virtual currency. Yet not only is the investor faced with a vast number of assets, additionally each investment class comes with its own set of "rules": the typical investment in a stock consists of a one off payment and a potentially infinite holding period, while derivatives have expiry dates and Futures require a constant management of the generated cash flow. And not only do the rules depend on the asset class, but might also differ between single assets or even the same assets traded on different platforms, as holding policies or transaction costs might differ.

In this work, a unified framework is introduced, that is designed to support a potential investor in the decision making. The algorithms provided in this document are fashioned to choose assets to invest in and provide an advise on the amount to be invested for each asset. The method is applicable to many different asset classes, adaptable to different transaction cost structures typically encountered by investors and adjustable to the aim of the investor.

The problem of optimizing an investing strategy on a set of assets is not a recent one. Most of todays research is based on the work of Markowitz[61], who generalized the "expected returns - variance of returns" rule that was considered to be sound practical advice without theoretical background. As a result of this work, the positive effect of diversification is nowadays a widely accepted fact in the financial industry. Yet, the approach as it was formulated by Markowitz is based on a strong assumption: the knowledge (or firm believes) of mean and variance of future returns. To circumvent this knowledge, in the first approaches to weaken the assumption, stationarity of the parameters was assumed and the mean and variance were estimated from historical data. This corresponds to the assump-

tion of a log-normal distribution of the prices, as e.g. in the Black-Scholes-Model[12]. It was demonstrated[67], that the portfolio obtained when relying on stationarity assumptions tends to be an investment into the assets with highest uncertainty in the estimated parameters, leading to a large exposure of unaccounted risk. The proposed solution is based on a re-sampling of the data, thereby using the log-normality of the prices, causing an increase in the exposure to model risk due to the heavy reliance on the model.

Other models have been employed to overcome the stationarity assumption, e.g. in the GARCH-model[13], the variance of the log-normal distributed prices is a random, autoregressive process, that is additionally influenced by the observed prices. Yet also these models make use of the stationarity assumption, when their parameters have to be estimated.

More robust (in the sense of less exposure to model risk) approaches exist, e.g.[25], that are mainly based on assumptions about the parameters or on the distribution of the returns being limited to a certain range and then utilizing min-max-methods to find the strategy doing best in the worst case. But stationarity and ergodicity assumptions are still prominently used in these approaches. Another method is based on identifying the fair price of each asset and deciding whether or not to invest. This is typically done by a human expert, based on experience and data on the assets, a process known as fundamental analysis[1].

All the models mentioned can be categorized as stationary, parametric models as each model consists of a finite number of parameters that are assumed to be fixed. As an example for a non-parametric model, the Black-Scholes-Model[12] with local volatility[23] could be considered. For this approach, the variance of the log-normal distributed prices is not assumed to be constant, but to be defined by a function of time and current prices $\sigma(t, S(t))$. This function is then (point-wise) characterized by liquidly traded, observed options on the market, yet a complete formulation is (in the general case) not known. The known points of this function are more commonly used to calibrate stochastic volatility, e.g. in the Heston model[35]. In this case the model becomes stationary and parametric again, as the variance is modeled by fixed parameters.

A different approach is taken in the case of utility-based portfolio optimization. In this approach, the investors preferences are formulated in form of a utility function[79] $u(\cdot)$. This utility function is then maximized (e.g. in form of its expected value) to obtain the investment decisions fitting best to the investors preferences. To solve the maximization problem the distribution of the utility has to be known, which in turn depends on the distribution of the argument. This can be considered as a non-parametric model, and the distribution can be described, e.g. by its density function. In the classical case, the distribution is assumed to belong to a certain family (e.g. log-normal) and only the parameters are estimated, rendering the model parametric. The knowledge of the distribution is an assumption that can be weakened e.g. by employing min-max-optimizations over a set of possible distributions[24].

The stationarity assumption is widely used in the financial industry. Relying on this

assumption allows the investor to use long histories of observations to lower the estimation uncertainty of the underlying parameters. Yet it was demonstrated by the economy multiple times, that this assumption is not met by the real world economy[73]. Indeed, a common theory assumes that the economy goes through cycles of growth, followed by a recession.

A first attempt to take this into account was made by Hamilton[33], who used a Hidden-Markov-Model to allow for non-stationarity in an autoregressive integrated moving average model. This approach trades the global stationarity assumption for an assumption of local stationarity of the observations and a Markovianity assumption[5,6,43,44,62] for the hidden parameter, which is less restrictive.

Other approaches exist to overcome stationarity assumptions, e.g. by assuming the weaker local stationarity, as do local kernel methods[27,57]. Other approaches to analyze non-stationary data are based on probabilistic properties of a hidden process, e.g. Markovianity in a Hidden Markov Model[72]. Others again make use of probabilistic assumptions on the hidden process and the observed data, as it is the case for Gaussian Mixture Models[28,62]. Another approach is based on geometrical properties, as e.g. in the case of the $\mathbb{K}$-means approach[52,59].

Recently a more general method was developed[36–42], that allows to formulate a wide range of mixture models without imposing assumptions on the switching process. Where necessary, the switching process can be assumed to be an element of some function space and by using this additional information regularization of the switching process is possible. While it might seem counter-intuitive to replace assumptions on distributions by regularization, non-regularized solutions often suffer from over-fitting effects caused by the ill-posedness of the problem. The resulting optimization objective and the corresponding conditions form an infinite dimensional variational problem. To solve this, an approach to find numerical solutions for PDE's is employed: the solution is approximated by Finite Elements[14]. This allows to estimate the infinite dimensional solution by a finite dimensional representation using basis functions. The inclusion of the Finite Elements allows to apply some of the numerous techniques that exist to solve PDE's numerically.

Due to the simple and straight-forward inclusion of Finite Elements, this approach was dubbed the FEM-approach. The FEM-framework is formulated to be flexible enough to allow the handling of a broad range of problems while at the same time providing enough information to formulate concepts on an abstract level.

In this work, it will be demonstrated, that the portfolio optimization problem can be formulated in a way that directly results in the FEM-method. Apart from allowing the application of algorithms designed to handle this kind of problem, it will also be shown, that assumptions made for numerical or mathematical reasons in the FEM-framework are resulting naturally from the portfolio optimization problem and have a direct financial interpretation. This allows to find additional ways to treat obstacles encountered on the route to finding a solution.

A short introduction for the FEM-method and portfolio optimization methods is given in chapter 2 together with an example resulting from a direct application of both known methods. In chapter 3 a new model free portfolio optimization approach is introduced, that can be directly translated into the FEM-framework. The problem of choosing the meta-parameters in the FEM-framework is solved in chapter 4 in the general setting, before a specialized approach for the FEM-BV-Utility-method is introduced. In chapter 5, the problems of data assimilation and forecasting are tackled for the general FEM-approach and the special case of portfolio optimization. To show the applicability of the method in the case of high-dimensional data, in the final chapter Swiss market data are analyzed.

2 Non-stationary parameter identification and parameter-based market phase detection

In this chapter, a recently developed approach to analyze non-stationary time series is described in detail. This method is a generalization of well-known methods, e.g. Hidden Markov Models[72], Mixture Models[28] and kernel smoothing models[26,27], providing additional benefits as e.g. the lack of probabilistic a priori assumptions about the latent variables. In addition, the reader can find an introduction to portfolio optimization in this chapter. The advantages of applying the non-stationary approach to the portfolio optimization are then demonstrated in an example.

2.1 The FEM-approach

The FEM-approach was introduced and developed in[36–42], a complete introduction is also given in[66]. To formulate the basic idea of the FEM-approach, a number of objects need to be defined: let $x(t) \in \mathbb{R}^d$ and $u(t) \in \mathbb{R}^k$ be two d-dimensional resp. k-dimensional, observable processes. Further let

$$x(t) = \tilde{f}((x(\tau))_{\tau\in[t-m,t)}, u(t), \Theta(t), \varepsilon(t)) \tag{2.1}$$

be a *direct model* for $x(t)$ with *memory depth* m, *external influences* $u(t)$, *model parameters* $\Theta(t)$ and *noise* $\varepsilon(t)$. Additionally it is assumed, that $\Theta(t)$ and $\varepsilon(t)$ are not observable, although additional information on the noise might be available (e.g. white noise, red noise, grey noise). The *inverse problem* is finding the best fitting parameter series $(\Theta(t))_{t\in[0,T]}$ for a given observation $(x(t))_{t\in[0,T]}$ and $(u(t))_{t\in[0,T])}$, where best fitting is meant in the sense of *model distance*

$$\min_{\Theta(t)} \int_{t=m}^{T} \tilde{g}(x(t), (x(\tau))_{\tau\in[t-m,t)}, u(t), \Theta(t))\, dt \tag{2.2}$$

with *model distance function* $\tilde{g}(\cdot)$. This model distance function can be defined by geometrical distances,e.g.

$$\tilde{g}(x(t),(x(\tau))_{\tau\in[t-m,t)},u(t),\Theta(t)) = \left\| x(t) - E_{\varepsilon(t)}\left[\tilde{f}((x(\tau))_{\tau\in[t-m,t)},u(t),\Theta(t),\varepsilon(t))\right]\right\|, \tag{2.3}$$

by logarithmic likelihood, e.g.

$$\tilde{g}(x(t),(x(\tau))_{\tau\in[t-m,t)},u(t),\Theta(t)) = -\log\phi_\varepsilon(x(t);\tilde{f}((x(\tau))_{\tau\in[t-m,t)},u(t),\Theta(t),\varepsilon(t))), \tag{2.4}$$

where $\phi_\varepsilon(\cdot)$ is the density function induced by the noise, or other suitable measures.

Different model distance functions might lead to different optimal parameters, e.g. in the case of a linear model with multiplicative noise

$$\tilde{f}(u(t),\Theta,\varepsilon(t)) = (\alpha+\beta u(t))(1+C\varepsilon(t)), \quad \Theta=(\alpha,\beta,C),\varepsilon(t)\sim\mathcal{N}(0,1)i.i.d \tag{2.5}$$

a perfect observational series $x(t)=\tilde{f}(u(t),\Theta,0)$ and known parameter C, the Maximum Likelihood approach, thus maximizing the likelihood function, will lead to systematically false estimates of α and β, while e.g. the χ^2-minimization yields the right values[7].

Standard methods attempting to solve problem 2.2 do so by interpolating the parameters. A common example is the local kernel method[27,57], where the estimators are modified with a kernel function $W(\cdot)$ which is non-negative and satisfies the conditions

$$\int_{-\infty}^{\infty} W(s)\,ds = 1, \int_{-\infty}^{\infty} W^2(s)\,ds < \infty. \tag{2.6}$$

This function is used to a priori weight the observations. E.g. the estimator for the (time-varying) mean ($\theta(t)=\mu(t)$) of the time series $x(0),x(1),\ldots,x(T)$ is given[26,27] by

$$\mu(t) = \frac{1}{b}\sum_{j=0}^{T} x(j) \int_{j-\frac{1}{2}}^{j+\frac{1}{2}} W\left(\frac{t-s}{b}\right) ds. \tag{2.7}$$

Additionally to the fact that this is just a local estimator, the results of these methods are known to be either to noisy (for small smoothing parameter b) or to constant (for large smoothing parameter b)[66] and the estimators can only be applied locally. Choosing the right value for the smoothing parameter is a non-trivial problem.

Other standard methods assume the existence of a hidden (latent) variable that switches between different values and use this process to switch between parameters. Depending on the properties of this hidden variable, the according models are called, e.g. Hidden Markov Models[72], if the hidden variable is a homogenous Markov process. The distribution of the

observable variable is in this case governed by a (known) parametric distribution with (unknown) parameters depending on the hidden variable. E.g. a Hidden Markov Model with (one-dimensional) Gaussian emission and discrete observations would have the mathematical representation:

$$P[Z(t+1) = j | Z(t) = i] = P_{ij} \tag{2.8}$$

$$f(x(t); \mu(Z(t)), \sigma(Z(t))) = \frac{1}{\sqrt{2\pi}|\sigma(Z(t))} \exp\left(-\frac{(x-\mu(Z(t)))^2}{2\sigma^2(Z(t))}\right), \tag{2.9}$$

where $Z(\cdot)$ is the hidden state variable with transition matrix $P_{ij_{ij}}$ and $\mu(\cdot)$, $\sigma(\cdot)$ are the mean and standard deviation depending on the hidden state. A solution of this problem is typically found by formulating the log-likelihood of the observation and using the Expectation-Maximization-algorithm (EM-algorithm)[9,22] to find the parameters P, $\mu(\cdot)$ and $\sigma(\cdot)$.

The EM-algorithm is a standard approach to fit model parameters depending on a hidden variable to a sample of observed data, a short description is given in Algorithm 1.

Algorithm 1 EM-Algorithm

Require: Time series of observed data $X \in \mathbb{R}^{d \times T}$, model for the data depending on model parameters $\Theta(t)$, model for the hidden variable $Z(t)$

1: Choose an initial estimate for the parameters $\Theta^{(0)}$.

2: **repeat**

3: **Expectation step** Estimate the expected value of the log likelihood function with respect the hidden process conditioned on the observation and the parameter estimate $\Theta^{(i-1)}$

$$Q(\Theta|\Theta^{(i-1)}) = E_{Z|X,\Theta^{(i-1)}}\left[\log L(\Theta; X, Z)\right] \tag{2.10}$$

4: **Maximization step** Maximize the obtained quantity from the previous step with respect to the parameters Θ

$$\Theta^{(i)} = \arg\max_{\Theta} Q(\Theta|\Theta^{(i-1)}) \tag{2.11}$$

5: **until** Convergence

While this algorithms is guaranteed to improve the likelihood in each step, the convergence to the maximum likelihood estimator cannot be guaranteed in general. Depending on the model functions, only local optimality might be achieved.

Replacing the Markovianity assumption of the hidden variable by the assumption that $Z(t)$ are independent and identical distributed, the class of Mixture models is obtained. The

mathematical representation for e.g. a Gaussian Mixture Model[28,62] is given as

$$P[Z(t) = i] = \omega_i \tag{2.12}$$

$$f(x(t); \mu(Z(t)), \sigma(Z(t))) = \frac{1}{\sqrt{2\pi}|\sigma(Z(t))} \exp\left(-\frac{(x-\mu(Z(t)))^2}{2\sigma^2(Z(t))}\right), \tag{2.13}$$

where again $Z(\cdot)$ is the hidden variable, ω_i are the state probabilities of $Z(\cdot)$ and $\mu(\cdot)$ and $\sigma(\cdot)$ are the parameters of the Gaussian distribution depending on the hidden state. It should be noted, that in this case the density function of $x(t)$ could also be written unconditioned on $Z(t)$ as

$$f(x(t); \mu(1), \ldots, \mu(\mathbb{K}), \sigma(1), \ldots, \sigma(\mathbb{K})) = \sum_{i=1}^{\mathbb{K}} \omega_i \frac{1}{\sqrt{2\pi}\sigma(i)} \exp\left(-\frac{(x-\mu(i))^2}{2\sigma^2(i)}\right). \tag{2.14}$$

A solution of this problem can again be found using the EM-algorithm.

Other standard methods are not based on (assumed) probabilistic properties of the data, but on geometrical properties. An example is the $\mathbb{K}$-means algorithm[52,59], which attempts to identify $\mathbb{K}$ geometrical centers and assigns each data point to the closest (in the Euclidean norm) of these centers.

The underlying model for the $\mathbb{K}$-means algorithm is given by

$$Z(t) \in \{1, \ldots, \mathbb{K}\} \tag{2.15}$$

$$x(t) = \Theta(Z(t)) + \varepsilon(t),\ E[\varepsilon(t)] = 0, \text{independent}. \tag{2.16}$$

Thus each observation is generated by a random dislocation from a central point given by $\Theta(Z(t))$, where the state variable $Z(\cdot)$ is not observed. Without assuming a distribution for the latent variable $Z(\cdot)$, and without further knowledge of the distribution of the dislocations $\varepsilon(t)$, the $\mathbb{K}$-means algorithm is designed to exploit geometrical properties of the observations $x(\cdot)$ to estimate the center points $\Theta(1), \ldots, \Theta(\mathbb{K})$. To this end, the algorithm, as shown in Algorithm 2, divides the observation into sets (or clusters) and minimizes the sum of the squared distances of the data points to their respective centers

$$\min_{\gamma(\cdot),\Theta(\cdot)} \sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \gamma_i(t) \|x(t) - \Theta(i)\|_2^2 \tag{2.17}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \quad \forall t \in \mathcal{T}. \tag{2.18}$$

The parameters $\gamma_i(\cdot)$ are called (cluster) affiliation functions, they can be interpreted as indicator functions of the hidden variable $Z(\cdot)$ as

$$\gamma_i(t) = 1 \Leftrightarrow Z(t) = i. \tag{2.19}$$

Algorithm 2 $\mathbb{K}$-means-Algorithm

Require: Time series of observed data $x(\cdot) \in \mathbb{R}^d$

1: Choose an initial estimate for the parameters $\Theta^{(0)}$.

2: **repeat**

3: Estimate the affiliation of each observation $x(t)$ to the central points

$$\gamma_i^{(j)}(t) = \begin{cases} 1, & i = \operatorname{argmin}_k \|x(t) - \Theta^{(j-1)}(k)\|_2^2 \\ 0, & \text{else} \end{cases} \tag{2.20}$$

4: Estimate the central points

$$\Theta_i^{(j)} = \frac{\sum_{t \in \mathcal{T}} \gamma_i(t) x(t)}{\sum_{t \in \mathcal{T}} \gamma_i(t)} \tag{2.21}$$

5: **until** Convergence

The $\mathbb{K}$-means algorithm can be seen as an application of the EM-algorithm assuming the $\varepsilon(t)$ to be i.i.d. normal distributed and the hidden state $Z(t)$ to be i.i.d. uniform distributed on $\{1, \ldots, \mathbb{K}\}$.

In contrast to standard methods, the FEM-approach is not based on interpolating the parameters, but interpolating the model distance function[36–42], i.e. the problem is formulated via the *(total) model distance*, also called the *average clustering functional*

$$\tilde{L} = \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} \gamma_i(t) \tilde{g}(x(t), (x(\tau))_{\tau \in [t-m,t)}, u(t), \Theta^{(i)})\, dt. \tag{2.22}$$

The inverse problem then takes the form

$$\min_{\Theta, \gamma(\cdot)} \tilde{L}, \sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1 \ \forall t, \gamma_i(t) \geq 0 \ \forall i, t, \tag{2.23}$$

with affiliation functions $\gamma_i(\cdot)$. This method is a generalization over e.g. the $\mathbb{K}$-means approach. The later can be obtained by using the distance function $g(x(t), \Theta) = \|x_t - \Theta\|_2^2$. Unlike the local estimator obtained in the kernel smoothing approach, the above definition does not only take into account observations that are in close (temporal) vicinity, but all data with similar affiliation function.

With this approach, the model distance function is interpolated within a simplex, where $\mathbb{K}$ parameter-sets are used to describe the corners of this simplex.

The problem, as described in (2.23) is ill-posed in the sense of Hadamard[31], as the solution is e.g. not unique, i.e. swapping two parameter-sets $\Theta^{(i)}$ and $\Theta^{(j)}$ and the according

cluster affiliation functions $\gamma_i(\cdot)$ and $\gamma_j(\cdot)$ will lead to the same value of L thus rendering each minimum to exist in $\mathbb{K}!$ permutations. This problem can be overcome by prescribing a certain order, e.g.

$$\int_0^T \gamma_1(t)\,dt < \int_0^T \gamma_2(t)\,dt < \ldots < \int_0^T \gamma_{\mathbb{K}}(t)\,dt. \tag{2.24}$$

Moreover, a typical observation of the processes $x(t)$ and $u(t)$ does not contain the whole interval $[0,T]$, but only a discrete subset $\mathcal{T} := \{t_1,\ldots,t_n; 0 \le t_1 < t_2 < \cdots < t_n \le T\} \subset [0,T]$. The average clustering functional (as in equation (2.22)) then takes its discrete form

$$L = \sum_{i=1}^{\mathbb{K}} \sum_{t\in\mathcal{T}\cap[m;T]} \gamma_i(t) g(x(t), (x(\tau))_{\tau\in[t-m,t)\cap\mathcal{T}}, u(t), \Theta^{(i)}) \tag{2.25}$$

where $g(\cdot)$ might have to be adapted to the discrete set $\mathcal{T}$ as is demonstrated in example 2.1.2.

Example 2.1.1. Let $x(t) \in \mathbb{R}$ be a geometrical Brownian motion defined as

$$dx(t) = x(t)\,(\mu(t)\,dt + \sigma(t)\,dB(t)) \tag{2.26}$$

where $B(t)$ is a Brownian motion. A simple transformation yields

$$d\log(x(t)) = \left(\mu(t) - \frac{1}{2}\sigma(t)^2\right) dt + \sigma(t)\,dB(t). \tag{2.27}$$

Now let the observations be done equidistant with $\mathcal{T} = \{0,1,\ldots,n \le T\}$. Then the increments of the transformed process $\log(x(t))$ can be written as

$$\log(x(t+1)) - \log(x(t)) = \mu(t) - \frac{1}{2}\sigma(t)^2 + \sigma(t)\varepsilon(t),\ \varepsilon(t) \sim \mathcal{N}(0,1), t \in \mathcal{T}\setminus\{n\} \tag{2.28}$$

This allows to write the likelihood of n observation as

$$\begin{aligned} &L_n\left((x(t))_{t\in\mathcal{T}}, (\mu(t))_{t\in\mathcal{T}}, (\sigma(t))_{t\in\mathcal{T}}\right) \\ &\qquad = \prod_{t=0}^{n-1} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log\frac{x(t+1)}{x(t)} - \mu(t) + \frac{1}{2}\sigma^2(t)}{\sigma(t)}\right)^2\right). \end{aligned} \tag{2.29}$$

Changing to the log-likelihood yields in

$$\begin{aligned} &l_n\left((x(t))_{t\in\mathcal{T}}, (\mu(t))_{t\in\mathcal{T}}, (\sigma(t))_{t\in\mathcal{T}}\right) \\ &\qquad = \sum_{t=0}^{n-1}\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2\sigma(t)^2}\left(\log\frac{x(t+1)}{x(t)} - \mu(t) + \frac{1}{2}\sigma^2(t)\right)^2\right), \end{aligned} \tag{2.30}$$

which leads to a possible choice of $g(\cdot)$ as

$$g(x(t),x(t-1),\mu^{(i)},\sigma^{(i)}) = \frac{1}{2{\sigma^{(i)}}^2}\left(\log\frac{x(t)}{x(t-1)} - \mu^{(i)} + \frac{1}{2}{\sigma^{(i)}}^2\right)^2. \tag{2.31}$$

$\Diamond$

Example 2.1.2. Let $x(t) \in \{0,1,2,3\}$ be a process, where $x(0),x(1),\dots$ form a Markov chain with constant transition matrix $P = [P_{ij}]_{ij}$ and states $1,2,3$ and $x(t) = 0$ for all $t \notin \mathbb{N}_0$. The likelihood of an observation $(x(t))_{t=0,1,\dots,n}$ is

$$L_n\left((x(t))_{t=0,1,\dots,n},P\right) = \prod_{t=1}^{n} P_{x(t-1)x(t)}, \tag{2.32}$$

and the according log-likelihood can be written as

$$l_n\left((x(t))_{t=0,1,\dots,n},P\right) = \sum_{t=1}^{n} \log P_{x(t-1)x(t)}. \tag{2.33}$$

Depending on the choice of $\mathcal{T}$, e.g. $\mathcal{T} = \{0,1,\dots\}$ and $\mathcal{T}^*$ containing only even numbers $\mathcal{T}^* = \{0,2,\dots\}$, the model distance function, adapted to this observational time-frame can be chosen as

$$g(x(t),x(t-1),P) = -\log P_{x(t-1)x(t)}, \tag{2.34}$$

$$g^*(x(t),x(t-2),P) = -2\log P_{x(t-2)x(t)}. \tag{2.35}$$

The model distance functions are different as the transition probabilities are different depending on the length of the time steps. $\Diamond$

In the common case of discrete observations, the uniqueness condition on well-posed problems is violated again: the cluster affiliation functions $\gamma_i(\cdot)$ can exert any behavior between the observational points of the set $\mathcal{T}$. To avoid this, the functions $\gamma_i(\cdot)$ are approximated by Finite Elements[14], e.g. hat functions. Let $0 = \tau_1 < \tau_2 < \dots < \tau_N = T$ be a partition of the time axis, the function $\chi_j(t)$ can be defined as e.g.

$$\chi_j(t) = \begin{cases} 1, & t \in [\tau_j, \tau_{j+1}-1] \\ t+1-\tau_j, & t \in [\tau_j-1, \tau_j) \\ \tau_{j+1}-t, & t \in (\tau_{j+1}-1, \tau_{j+1}] \\ 0, & t \notin [\tau_j-1, \tau_{j+1}] \end{cases}, \tag{2.36}$$

where $\tau_0 < 0$ and $\tau_{N+1} > T$ are defined arbitrary. As a further condition, the finite element function have to be non-negative and sum up to 1 at all times:

$$\sum_{j=1}^{N} \chi_j(t) = 1, \quad \forall t \in \mathcal{T}, \qquad \chi_j(t) \geq 0, \quad \forall t \in \mathcal{T}, j = 1,\dots,N \tag{2.37}$$

This allows to approximate the function $\gamma_i(t)$ as sum of N finite elements

$$\gamma_i(t) \approx \sum_{j=1}^{N} \gamma_i^{(j)} \chi_j(t), \quad i = 1, \ldots, \mathbb{K} \tag{2.38}$$

and to define the problem of parameter identification as a finite-dimensional optimization problem (in $\gamma_i(\cdot)$), of the following form:

$$\min_{\Gamma,\Theta} \sum_{i=1}^{\mathbb{K}} \sum_{j=1}^{N} \gamma_i^{(j)} \int_m^T \chi_j(t) \tilde{g}(x(t), (x(\tau))_{\tau\in[t-m,t)}, u(t), \Theta^{(i)})\, dt, \tag{2.39}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i^{(j)} = 1, \quad \forall j = 1, \ldots, N, \tag{2.40}$$

$$\gamma_i^{(j)} \geq 0, \quad \forall t \in \mathcal{T}, j = 1, \ldots, N. \tag{2.41}$$

2.2 Regularization methods

Another standard assumption of time series analysis is the local stationarity of parameters. This assumption is implied by the time scale of parameter-changes versus the time scale of observations: in a typical setting, the parameters change at a slower pace than the observations are made, e.g. the local kernel methods are based on this idea.

Other methods replace the local stationarity assumption with the assumption of sudden regime changes, i.e. the parameters can only be chosen from a finite set $\{\Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}\}$ where the values of these parameters are a priori unknown. A classical example for these methods is $\mathbb{K}$-means[52,59]. While this method works reasonable well if the parameters are distinct enough, even for low dimensional examples $\mathbb{K}$-means can fail to identify the true affiliations.

Example 2.2.1. [66]

For this example we consider a time series of two dimensional data $x(t) = (x_1(t), x_2(t))$ generated via a mixture model consisting of a time dependent convex combination of three (stationary) normal distributions,

$$x(t) \sim \sum_{i=1}^{3} \gamma_i(t) \mathcal{N}(\mu^{(i)}, \Sigma^{(i)}) \quad t = 1, \ldots, 6000, \tag{2.42}$$

where the coefficients $\Gamma(t) = (\gamma_1(t), \gamma_2(t), \gamma_3(t))$ are chosen such that the (hypothetically unknown) affiliation process changes twice, visiting all three states. For an illustration of $\Gamma(t)$ see the right panel of Fig. 2.1. From the scatter plot given in the left panel of Fig. 2.1

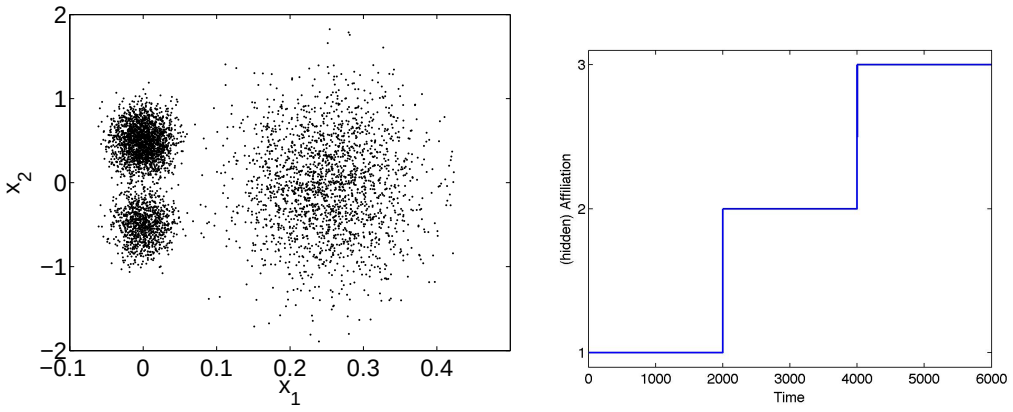

Figure 2.1. *Example 2.2.1[66] - Left: Scatter plot of the a time series generated via (2.42) where the parameters were chosen such that the data are concentrated in three geometric regions. Right: The graph of the affiliations used as a hidden process in parameter space for the generation of the time series shown in the left panel.*

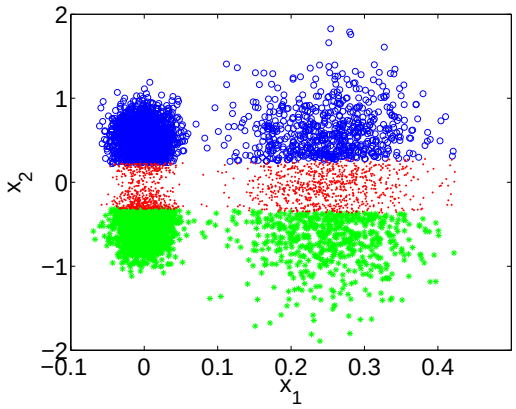

Figure 2.2. *Example 2.2.1[66] - Affiliations of the data points resulting from the classical $\mathbb{K}$-means algorithm. Due to the difference in scale of the dimension, the affiliations are not identified correctly.*

it can be seen, that the means and covariance matrices $(\mu^{(i)}, \Sigma^{(i)})$, $i = 1,2,3$ were chosen such that the data are mainly concentrated in three distinct areas. Applying the $\mathbb{K}$-means algorithm for $\mathbb{K} = 3$, leads to a classification (as illustrated in Fig. 2.2) that does not match these areas. This misclassification of the data points can be attributed to the different scales of the x_1 and x_2 components of the data. $\Diamond$

One way of combining both assumptions (local stationarity and sudden regime changes) is the concept of regularization. Thereby the changes in the affiliation function are penalized or limited to generate less transitions between different regimes. Different strategies for regularization exist, e.g. Tikhonov regularization, lasso regularization or bounding of the variation.

In the context of regularizing the affiliation function, Tikhonov regularization[76] describes the process of penalizing the objective function with the squared L2-norm of the first (weak) derivative. This limits the affiliation functions to elements of the H^1 space, as only in this case the penalization term is well-defined. The problem of equation (2.22) with constraints (2.23) becomes[37–40]

$$\tilde{L}_\varepsilon^2 = \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} \gamma_i(t)\tilde{g}(x(t),(x(\tau))_{\tau\in[t-m,t)},u(t),\Theta^{(i)})\,dt + \varepsilon^2 \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} (\partial_t \gamma_i(t))^2\,dt \tag{2.43}$$

$$\to \min_{\gamma,\Theta},$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1 \;\forall t, \gamma_i(t) \geq 0 \;\forall i,t. \tag{2.44}$$

The effects of this regularization are (i) that the resulting optimization problem becomes quadratic in γ and (ii) the derivatives of $\gamma(\cdot)$ are decreased when increasing ε.

The Lasso regularization[75] in turn penalizes the L1-norm of the first derivative, thereby implying the affiliation functions are from the Sobolev space $W^{1,1} \supset H^1$. With this regularization, the problem takes the form

$$\tilde{L}_\varepsilon^1 = \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} \gamma_i(t)\tilde{g}(x(t),(x(\tau))_{\tau\in[t-m,t)},u(t),\Theta^{(i)})\,dt + \varepsilon^2 \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} |\partial_t \gamma_i(t)|\,dt \tag{2.45}$$

$$\to \min_{\gamma,\Theta},$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1 \;\forall t, \gamma_i(t) \geq 0 \;\forall i,t. \tag{2.46}$$

This problem is convex, but neither linear nor quadratic. Yet by splitting the derivatives in positive and negative parts the problem can be transformed into a linear problem while doubling the dimensionality. An increase in ε leads to single derivatives of $\gamma(\cdot)$ going to

zero. The problem of positive and negative component being simultaneously non-zero is solved automatically, as the removal of a constant from both values decreases the penalty term without changing the (unpenalized) objective function, which means for the identification of the minimum, always a solution with at least one out of the two components being zero is chosen.

To see the effect of the regularization, the following consideration is useful. First a linear optimization problem with Tikhonov regularization is considered

$$c^\dagger x + \alpha \|x\|_2^2 \to \min_x . \tag{2.47}$$

The solution to this problem is given by

$$c + 2\alpha x^* = 0 \Rightarrow x^* = -\frac{1}{2\alpha} c. \tag{2.48}$$

Thus any change in α will result only in a re-scaling of the solution. While this does not hold in general, the fact that the components of the solution are decreased but stay non-zero does remain. To see the effect of the LASSO regularization, a quadratic problem is examined

$$x^\dagger H x^\dagger + c^\dagger x + \alpha \|x\|_1 \to \min_x . \tag{2.49}$$

The gradient of this problem can be written as

$$(H + H^\dagger) x + c + \alpha \operatorname{sgn}(x), \tag{2.50}$$

where $\operatorname{sgn}(\cdot)$ is the element-wise sign-function. The solution to the optimization problem has to satisfy the following first order condition

$$x^* = -(H + H^\dagger)^{-1}(c + \alpha \operatorname{sgn}(x^*)). \tag{2.51}$$

Assuming that the solution is known to contain positive elements only, this transforms into

$$x^* = -(H + H^\dagger)^{-1}(c + \alpha 1), \tag{2.52}$$

thus an increase in α results in elements of x^* being pushed to zero differently. The same holds for negative entries. In fact, the LASSO method is known to produce sparse solutions, thus single elements of the solution become zero[75].

A more general way to regularize is given by choosing the total variation as regularization functional. To this end, let $\mathcal{P}$ be the set of all partitions of the interval $[m,T]$, then the total variation is defined as

$$V_T^m(\gamma_i) = \sup_{P \in \mathcal{P}} \sum_{j=1}^{n_P} |\gamma_i(t_j^P) - \gamma_i(t_{j-1}^P)|,\ P = \{t_0^P, t_1^P, \ldots, t_{n_P}^P\}. \tag{2.53}$$

If $\gamma_i(\cdot)$ is differentiable and the derivative is Riemann-integrable, the total variation is

$$V_T^m(\gamma_i) = \int_m^T |\partial_t \gamma_i(t)|\, dt. \tag{2.54}$$

In the later case, a penalization of the objective functional $\tilde{L}$ would lead to the same optimization objective as the Lasso regularization. Yet the more general definition in Equation (2.53) together with the discretization implied by the observations allow for a threshold approach. As the (discrete) objective functional L (as in Equation (2.25)) does not depend on the function values of $\gamma_i(\cdot)$ on $[0,T] \setminus \mathcal{T}$, the function $\gamma_i(\cdot)$ can be chosen to minimize the total variation for a partition containing the elements of $\mathcal{T}$ only, as the same time the triangle inequality ensures[40–42]

$$V_T^m(\gamma_i) = \sup_{P \in \mathcal{P}(\mathcal{T})} \sum_{j=1}^{n_P} |\gamma_i(t_j^P) - \gamma_i(t_{j-1}^P)| = \sum_{j=2}^{n} |\gamma_i(t_j) - \gamma_i(t_{j-1})|,\ \mathcal{P}(\mathcal{T}) = \{P \in \mathcal{P}, \mathcal{T} \supset P\} \tag{2.55}$$

Depending on the regularization and the model used, the approach will be called FEM-*regularization-model*-method, e.g. by FEM-BV-Markov the FEM-method with Bounded Variation and a Markov model for the data is referred to. FEM-H1-$\mathbb{K}$-Means is the FEM-method with $H1$-regularization and a Euclidean distance function $g(\cdot)$. The algorithm used to solve the regularized problem is shown in Algorithm 3.

It should be noted, that for the unregularized case (i.e. $\varepsilon = 0$ or $C = \infty$) and $g(x(t), \Theta) = \|x(t) - \Theta\|_2^2$ this is exactly the $\mathbb{K}$-means algorithm shown as Algorithm 2.

Example 2.2.1 (cont.) To demonstrate the usefulness of the regularization, the example 2.2.1 is reconsidered. The time series was analyzed with the FEM-BV-approach which results from the simple model

$$x(t) = \Theta(t) + \varepsilon(t), \tag{2.64}$$

and the model distance function

$$g(x(t), \Theta) = \|x(t) - \Theta\|_2^2. \tag{2.65}$$

This corresponds to the Euclidean model distance used by the K-Means algorithm[52]. For $C = \infty$ the FEM-BV-Problem has a unique analytic solution (except for permutations of the hidden states) for Γ and for the model parameter Θ, which both are equal to the respective update formulas in the $\mathbb{K}$-means algorithm, as shown in Eq.s (2.20) and (2.21).

Solving the FEM-BV-$\mathbb{K}$-means problem repeatedly with the optimal combination $\mathbb{K} = 3, C = 2$ and a randomly chosen initial value for Γ, the best solution, in the sense of minimizing the objective functional, yields in the correct (up to a few isolated misfits) reproduction of the original convex coefficients $\Gamma(t)$. The classification of the data points is shown

Algorithm 3 FEM-algorithm

Require: Time series of observations $x(\cdot) \in \mathbb{R}^d$, number of parametersets $\mathbb{K}$, penalty factor ε or threshold level C

1: Choose $\gamma_i^{(0)}(t) \in [0,1]$ at random with for $t \in \mathcal{T}$ and $i = 1, \ldots, \mathbb{K}$.

2: **repeat**

3: Estimate parameters $(\Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})})^{(j)}$ by minimizing

$$(\Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})})^{(j)} = \underset{\Theta^{(1)},\ldots,\Theta^{(\mathbb{K})}}{\operatorname{argmin}} \sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \gamma_i^{(j-1)}(t) g(x_t, (x(\tau))_{\tau\in[t-m,t)\cap\mathcal{T}}, u(t), \Theta^{(i)}). \tag{2.56}$$

4: Estimate the affiliation functions $\gamma_i(\cdot)$ by solving one of the following minimization problems:

- H^1-regularized:

$$\gamma^{(j)}(\cdot) = \underset{\gamma(\cdot)}{\operatorname{argmin}} \quad \sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \gamma_i(t) g(x_t, (x(\tau))_{\tau\in[t-m,t)\cap\mathcal{T}}, u(t), \Theta^{(i)^{(j)}}) + \varepsilon \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} (\delta_t \gamma_i(t))^2 \, dt \tag{2.57}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \quad \gamma_i(t) \geq 0, \forall i, t \tag{2.58}$$

- $W^{1,1}$-regularized:

$$\gamma^{(j)}(\cdot) = \underset{\gamma(\cdot)}{\operatorname{argmin}} \quad \sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \gamma_i(t) g(x_t, (x(\tau))_{\tau\in[t-m,t)\cap\mathcal{T}}, u(t), \Theta^{(i)^{(j)}}) + \varepsilon \sum_{i=1}^{\mathbb{K}} \int_{t=m}^{T} |\delta_t \gamma_i(t)| \, dt \tag{2.59}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \quad \gamma_i(t) \geq 0, \forall i, t \tag{2.60}$$

- BV-regularized

$$\gamma^{(j)}(\cdot) \quad = \underset{\gamma(\cdot)}{\operatorname{argmin}} \sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \gamma_i(t) g(x_t, (x(\tau))_{\tau\in[t-m,t)\cap\mathcal{T}}, u(t), \Theta^{(i)^{(j)}}) \tag{2.61}$$

$$\sum_{i=1}^{\mathbb{K}} \|\gamma_i\|_{BV} \leq C \tag{2.62}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \quad \gamma_i(t) \geq 0, \forall i, t \tag{2.63}$$

5: **until** Convergence

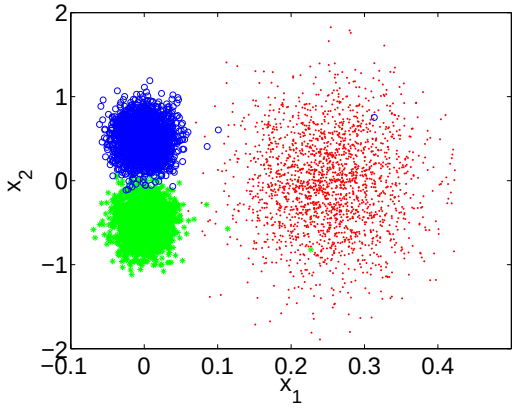

Figure 2.3. *Example 2.2.1*[66] *- The affiliations of the data points resulting from the FEM-$\mathbb{K}$-means method. Up to a few misclassifications, the method led to the right identification of the hidden process and therefore the right assignment of the data points to the geometrical centers whereas the unregularized method failed to do so (as can be seen in Figure 2.2).*

in Fig. 2.3. ◊

2.3 Portfolio optimization with known parameters

One of the major problems in mathematical finance is the portfolio optimization problem. From a practitioners viewpoint, it is of great interest to find an investment strategy (an optimal control) to maximize an objective function that depends on an underlying random process. The objective function can include e.g. the performance or some form of negative risk measure. The arguments of the objective function can range from a broad area of possible investments: financial instruments, like e.g. stocks, certificates and derivatives (in context of finance) or projects that can either be funded or delayed/canceled (e. g. in context of optimal project management), and more.

The Markowitz approach[61] to optimize a portfolio of risky assets is based on the proportion-of-wealth representation of investment strategies. Thereby the i-th component of the vector π is representing the proportion of the total wealth to be invested in the i-th asset. Using this representation and the data series of returns $R(t)$, the portfolio optimization problem can be written as

$$\pi^\dagger \operatorname{Cov}(R(t))\pi - \alpha\pi^\dagger E[R(t)] \to \min_\pi, 1^\dagger\pi = 1, \tag{2.66}$$

where α is chosen by the investor to balance risk (the quadratic term) and return (the linear part) to his preferences. Depending on whether the expectation and variance are con-

ditioned on the previous observations, static (unconditioned) and dynamic (conditioned) portfolio optimization are distinguished[54].

This portfolio model is based on the assumption of knowledge (or beliefs) of the mean and covariance matrix of the vector of asset returns. Markowitz suggests to use statistical techniques to obtain this knowledge, e.g. by using the empirical mean and covariance matrix from observations in the past. This and other statistical techniques are based on an underlying assumption of time-invariance of the mean and covariance matrix (see e.g.[49]).

As a fact known by researchers[11,23,35] and as was demonstrated by the economy multiple times, this assumption is wrong in the real world setting. One of the first approaches of handling the change in the distribution was given by Hamilton[33], showing that the application of Hidden Markov Models[72] can be used to account for changes in the financial environment. The identified periods of stationarity were called *Market phases*, implying a certain periodicity.

Current methods to overcome the stationarity assumption either include additional parameters which are in turn stationary (e.g.[55]) or assume further knowledge of the distribution (e.g. [34]).

Another aspect is, that whenever parameters have to be estimated from data, the size of the sample plays an important role. As demonstrated in[67], using standard estimators for the expectation and covariance will lead to portfolios maximizing the estimation error. Instead, as proposed in[68] one can use the estimators to resample the data and to generate a sample statistics of optimal portfolios, choosing the average (the empirical expectation) from the sample as investment choice. This reduces the exposure to the estimation error, yet it imposes the Gaussianity of the returns, leading to problems when the underlying processes are non-Gaussian.

A different approach to the problem of finding an optimal investment is based on the maximization of the expected utility. Let $u(\cdot)$ be a utility function in the sense of[79], and $V_\pi(t)$ be the wealth at time t resulting from the investment strategy π. Then e.g. the maximization of the expected terminal utility would be

$$E\left[u(V_\pi(T))\right] \to \max_\pi . \tag{2.67}$$

This approach can be seen as more general, yet it suffers from similar problems as the Markowitz approach: To estimate the mean of the utility function, the distribution of the returns has to be known a priori, inferring knowledge about the process which can be doubted. Moreover the parameters of this distribution have to be estimated, exposing this method to the same problem of sample size and stationarity assumptions.

A problem yet unmentioned is the occurrence of transaction costs. This problem is relevant for practitioners, as a direct application of e.g. the Markowitz approach will lead to a constant rebalancing of the portfolio, when applied to multiple time steps. This constant

trading will lead to a steady accumulation of transaction costs that might vanquish any profits made from the changes in investment. To incorporate transaction costs, different ways have been proposed: e.g. including them in the linear part of the objective function (as in[58]). Other approaches are based on assumptions on the underlying distribution of the returns (e.g. the Martingale approach shown in[19]).

2.4 Portfolio optimization with non-stationary parameter identification

A direct approach to utilize market phases in portfolio optimization is the usage of clustering methods on the parameters. The information obtained at this step is then used to construct the portfolios. E.g. the classical Black-Scholes model defines the price process as

$$dS(t) = \operatorname{diag}[S(t)]\,(\mu\,dt + \sigma\,dB(t))\,, \tag{2.68}$$

with constant parameters μ and σ. Assuming non-stationarity in the parameters, these constants are replaced by time-dependent versions

$$dS(t) = \operatorname{diag}[S(t)]\,(\mu(t)\,dt + \sigma(t)\,dB(t))\,. \tag{2.69}$$

A way to estimate the parameters in one dimension (based on observed, liquidly traded options) was shown in[23].

Discretizing the above stochastic differential equation to the observations $\mathcal{T} = \{t_1, \ldots, t_n\}$ will yield in the equation

$$\begin{aligned}\operatorname{diag}[S(t_i)]^{-1}\,(S(t_{i+1}) - S(t_i)) &= \mu(t_i)(t_{i+1} - t_i) + \sigma(t_i)(B(t_{i+1}) - B(t_i)) \\ &\sim \mathcal{N}(\mu(t_i)(t_{i+1} - t_i), (t_{i+1} - t_i)\sigma(t_i)\sigma^\dagger(t_i))\end{aligned} \tag{2.70}$$

with discretized time dependent parameters $\mu(t_i)$ and $\sigma(t_i)$. Alternatively to the parameter $\sigma(t_i)$, the covariance matrix $\Sigma(t_i) = \sigma(t_i)\sigma^\dagger(t_i)$ can be estimated.

Using this model for the asset returns, the expected return of the investment $\pi(t)$ is $\mu(t)^\dagger\pi(t)$ and the variance $\pi(t)^\dagger\Sigma(t)\pi(t)$. Yet, the estimation of the instantaneous parameters $\mu(t)$ and $\Sigma(t)$ is a problem. One approach to estimate these parameters is using a "moving average". This is a special case of the local kernel method[27,57] (see also Section 2.2). For the moving average approach, the kernel method $W(\cdot)$ is chosen as

$$W(s) = \begin{cases} 1, & s \in [-1, 0] \\ 0, & \text{else} \end{cases}, \tag{2.71}$$

and the estimation for a certain lookback-period τ on the dataset $(x(t))_{t\in\mathbb{R}}$ is obtained as

$$\mu(t) = \frac{1}{\tau}\int_{t-\tau}^{t} x(s)W\left(\frac{s-t}{\tau}\right)ds \tag{2.72}$$

$$\Sigma(t) = \frac{1}{\tau}\int_{t-\tau}^{t} (x(s)-\mu(t))^{\dagger}(x(s)-\mu(t))W\left(\frac{s-t}{\tau}\right)ds. \tag{2.73}$$

The justification of this approach is the assumption of local stationarity, thus the parameters $\mu(t)$ and $\Sigma(t)$ are (approximately) constant for the time frame $[t-\tau,t]$.

If this assumption is justified, than the question arises, which value to chose for τ, and whether τ depends on t as well. Thus instead of approximating the parameters with the assumptions on stationarity within a fixed interval, an assumption of constant parameters on the N intervals $[t_i,t_{i+1})$, $t_0 < t_1 < \ldots, t_N$ can be made. This can be seen as a more general version of the above idea: instead of assuming stationarity within a fixed time frame (independent of the time), the parameters are assumed to be constant on a subset of the whole time interval of variable length, that might be subject to randomness itself.

The effect of dropping the stationarity assumption will be demonstrated in the following by means of an example from [70]. Taking on the argument that the mean of daily returns can be better estimated than the covariance matrix, for the following example, the structure imposed to the mean is $\mu(t) = \sum_{j=0}^{\omega} \mu_j(t)t^j$. The optimization is performed based on the parameterization of the mean and the covariance matrix $\Sigma(t)$. Moreover, the covariance matrix and the coefficient of the mean are assumed to be constant on unknown intervalls $[t_i,t_{i+1})$, $0 = t_0 < t_1 < \ldots, t_N = T$ with a priori unknown $t_1,\ldots,t_{N-1}$ and N.

When working with financial data, one might end up with thousands of assets, as, for example, the Wilshire 5000 index, measuring the performance of most publicly traded companies available on the US market, consists of 3663 stocks (as of March 31st, 2014)[80]. With respect to this high dimensionality, one has to think about methods that are able to handle these data. The popular idea is to use principal component analysis (short: PCA, see [17,18,32,50,77]), where only a linear low-dimensional manifold of the data space is used to represent the original observation. The drawback here is the linearity of the manifold, as a globally optimal linear subspace might not exist. Instead it is assumed that these linear manifolds exist locally. To this end the principle directions of variance are identified (the largest eigenvectors of the covariance matrix) and a sufficient difference in this property of the covariance matrix, as indicated in [73], is seen as an indication of changed covariance. To further improve the estimation, different intervals are allowed to behave in the same way, thus the lower dimensional manifold is the same for multiple time intervals.

Let $Q^{(i)} \in \mathbb{R}^{d\times n}$ for $i = 1,\ldots,K$ and $n < d$ with ${Q^{(i)}}^{\dagger}Q^{(i)} = \mathtt{Id}$ be the projection matrix, and $(x(t) \in \mathbb{R}^d)_{t\in\mathcal{T}}$ be the time series of returns obtained from the observed price vector

$(S(t))_{t\in\mathcal{T}}$. Then we can define the distance function by

$$g(x(t),\theta_i) = \|(x(t)-\mu^{(i)}(t)) - Q^{(i)}Q^{(i)^\dagger}(x(t)-\mu^{(i)}(t))\|_2^2, \quad \theta^{(i)} = (\mu_0^{(i)},\ldots,\mu_\omega^{(i)},Q^{(i)}). \tag{2.74}$$

This distance function measures the reduction error resulting from the projection of the d-dimensional data on the i-th n-dimensional linear manifold. As x_t are the returns, $Q^{(i)}$ can be interpreted as the directions of maximal variance[45]. The covariance matrix of a sub-set $(x_t)_{t\in\mathcal{T}_i\subset\mathcal{T}}$ of the time series, where $\gamma_i(t) = 0 \forall t \notin \mathcal{T}_i$ for all i can be estimate by

$$\Sigma^{(i)} = \frac{1}{\sum_{t\in\mathcal{T}}\gamma_i(t)}\sum_{t\in\mathcal{T}}\gamma_i(t)(x(t)-\mu^{(i)}(t))(x(t)-\mu^{(i)}(t))^\dagger. \tag{2.75}$$

The return of an investment strategy $\pi(t)$ is the profit (or loss) within an investment period normalized by the value at the investment at the beginning of the investment period. This detail becomes important, when dealing with leveraged products as e.g. Futures. A Future is a contract between two parties to buy physical or virtual goods (e.g. commodities, indices) at a certain time in the Future (the settlement date) with a price agreed upon at the time of closing the contract. In contrast to a Zero-Strike Call Option, where the price is the current value of the underlying asset and is paid by the buyer immediately, a Future does not involve any cash-flow at the initial time between both parties. Yet the broker matching two interested parties typically demands a security payment from both parties, the so called Margin. After obtaining the contract, any change in the price is matched by the broker and the difference paid to or taken from the margin, thus changes in the price do affect the amount on the margin account instantly.

If an investor choses a typical broker, e.g.[3], he can obtain a Future contract for the DAX, which is standardized with a leverage of 25 Euro per point, the investor does have to provide a margin of 1800 Euro per contract. Then any increase or decrease of the DAX by one point will be directly reflected on the investors margin account by a gain (in case of an increase) or loss (in case of a decrease) of 25 Euro. If the DAX starts out at 9500 points at the time of purchasing one contract, the investor has to provide 1800 Euro on his margin account. If the DAX increases during the next hour by 10 points to 9510, the value on the margin account increases by 250 Euro to 2050 Euro. Assuming the DAX dropped 5 points when the investor decides to leave the position, the amount of 125 Euro is subtracted from the margin account, leaving the investor with 1925 Euro.

Due to the nature of this contract, closing a position, which would correspond to selling stocks from a portfolio, can be seen as obtaining a contract in the opposite direction, thus if the investor obtained a “Buy”-Future, the position can be closed by closing a contract for a “Sell”-Future of the same contract size. The cash-flow of both contracts will cancel out, leaving the investor with the profit or loss obtained till the time of closing the second contract. Nowadays, many Futures contracts are not physically settled, either because the

underlying asset can not be physically delivered (e.g. stock exchange indices) or the main purpose of the Future is speculative in nature. In this case, the broker will close all open positions at the settlement date.

At first glance, this contract type might seem counter intuitive. The investor could just obtain the underlying asset and generate the same (virtual) cash-flow, yet Futures hold their own advantages: (i) The underlying asset might not be tradable (e.g. the index of a stock exchange), or not exist yet (e.g. a company can sell the production of the next month, practically insuring itself against a drop in prices till the finishing of the production), (ii) the investment necessary is (typically) less than the "fair" price of the underlying asset, in the earlier example, the investor had to pay 1800 Euro for a contract which has a value of 237500 Euro. Practically providing the same cash-flow with less than a percent of the investment necessary (iii) entering a negative position by selling assets not in possession of the investor (known as short selling) is an integral part of the market, as all open positions over all market participants will always sum up to zero. Thus Futures can be used to practically perform short sales when they are not possible otherwise (e.g. for jurisdictional reasons), or if preferred by the investor.

It should not be left unmentioned that especially the second advantage can also prove to be a disadvantage: if the amount of money on the margin account drops under a certain level, typically a proportion of the money initially needed to open the positions, the broker will issue a "margin call", demanding additional security from the customer and ultimately, if no additional payment is made, close the position prematurely. Locking in a loss for the investor without the chance of a (potential) recovery. Here the leverage effect works to the disadvantage of the investor, while a classical investment into a stock limits the risk to the amount invested, if the company defaults and the stock price drops to zero, an investment into a Future will cause additional losses not covered by the initial margin.

The fact that selling in the market is easily done might seem like a big plus, but it presents a problem when combining it with classical portfolio optimization theory. If $\pi(t)$ represents the vector of investments in proportion of wealth at time t, a negative entry represent a short sale, yet a typical assumption is $1^\dagger \pi(t) \leq 1$, thus the investor has a total position covered by his wealth (or representing his wealth in case of equality). Any negative entries will allow for increased spending in the other assets, as the investor practically borrows money by short selling assets. In the context of Futures, that does not work: the margin payment does not depend on the "direction" of the trade, just on the absolute size. This can be reflected by splitting each Future into its long and short positions, creating a virtual second asset representing the short positions. This allows to handle both kind of trades by imposing non-negativity on the positions. An additional assumption that improves the result can be obtained by observing, that a long and short position at the same time (partially) cancel each other and provide no further advantage. If $\pi_i(t)$ is the total position in the i-th asset (a Future) at time t, and $\pi_i(t) = \pi_i^+(t) - \pi_i^-(t)$ is the decomposition into long

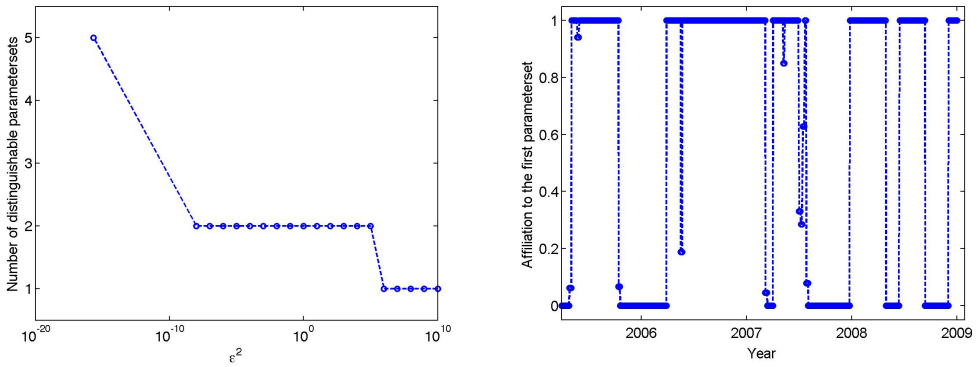

Figure 2.4. *Futures portfolio*[70] *- Left: Number of distinguishable market phases depending on* ε^2*. Right: Allocation of parameter-sets for the time series using* $\mathbb{K} = 2$ *and* $\varepsilon^2 = 1$*.*

and short positions, than the conditions $\pi_i^+(t) \geq 0, \pi_i^-(t) \geq 0$ are not sufficient to guarantee uniqueness. By adding the condition $\pi_i^+(t)\pi_i^-(t) = 0$ it is guaranteed, that at least one of the two positions is zero, providing an unique decomposition of the total position.

As the initial investment necessary to obtain a Future does not depend on the price of the Future at that time, but on the margin, the return of the investment $\pi_i(t)$ can be written as

$$R_{i\pm}(t) = \frac{\pm\Delta S(t)}{M}, \tag{2.76}$$

with $i+$ being the dimension of π_i^+ and $i-$ the dimension of π_i^- respectively.

For the purpose of the following example, the prices of four wheat and four oil Futures of the years 2005-2008 are used. The Futures differ in the remaining time to maturity, the wheat Futures expire at the last trading day before the Saturday following the third Friday of March, July, September and December, the oil Futures are expiring monthly. The daily closing prices where obtained from[51] and[78]. The Futures are treated as "rolled over" contracts, which means: when one contract is expired, it is replaced by the next expiring contract, which in turn is replaced by the one with the next longer time to maturity, and so on, shifting the dimension corresponding to the Futures of one commodity by one. The contract with the longest time to maturity is replaced by a new one. Additionally for the purpose of this example, all assets share one margin account, containing the whole wealth of the investor.

Analyzing the whole dataset with H^1-regularization for different penalty parameters ε^2, different number of possible mean parameters, dimensionality of projection matrices and number of parameter-sets will lead to different results. Choosing out of the combinations is a non-trivial task that will be elaborated in Chapter 4. For this example, the number of mean parameters is chosen to be two (allowing for linear trends in the returns)

and the number of reduced dimensions is chosen to be 2 as well. One way to find good combinations of $\mathbb{K}$ and ε^2 is based on the number of statistically distinguishable clusters (see[38]). Thereby confidence intervals of the parameters are estimated by bootstrapping the relevant data points for each parameter and checking for overlap between these confidence intervals. If an overlap exists, the parameters are deemed statistically indistinguishable and the parameter $\mathbb{K}$ is considered too large. Out of the combinations $(\varepsilon^2, \mathbb{K})$, for each ε^2 the one with the largest $\mathbb{K}$ is chosen, the obtained values for $\mathbb{K}(\varepsilon^2)$ can be seen in the left panel of Figure 2.4. From this result the parameter ε^2 is chosen to be $\varepsilon^2 = 1$ and $\mathbb{K} = 2$. The resulting allocation of each datapoint to the parameters $\theta^{(1)}$ and $\theta^{(2)}$ can be seen in the right panel of Figure 2.4.

The interest of an investor typically does not lie with analyzing historical data, but finding portfolios to invest into today. To simulate this process, the following algorithm was run, starting with the first $T = 10$ observations, holding the parameters $\varepsilon^2 = 1$, the linear trend and the reduction to two dimension fixed:

1. Find the maximum number of distinguishable parameter-sets $\mathbb{K}$.

2. Estimate the affiliation functions $\gamma_i(\cdot)$ and the parametersets $\theta^{(1)}, \ldots, \theta^{(\mathbb{K})}$.

3. Chose the predicted parameterset $\hat{\theta} = \sum_{i=1}^{\mathbb{K}} \gamma_i(T)\theta^{(i)}$.

4. Estimate the instantaneous covariance matrix as linear combination of the full covariance matrices of the data affiliated with each of the parameter-sets, assuming the mean to be given by $\mu^{(i)}(t)$:

$$\hat{\Sigma} = \sum_{i=1}^{\mathbb{K}} \gamma_i(T) \frac{1}{\sum_{t=1}^{T} \gamma_i(t)} \sum_{t=1}^{T} \gamma_i(t) \left(R(t) - \mu_0^{(i)} - \mu_1^{(i)} t\right) \left(R(t) - \mu_0^{(i)} - \mu_1^{(i)} t\right)^{\dagger} \quad (2.77)$$

5. Find the minimum variance portfolio which exceeds a (daily) target return of 0.01415% by solving

$$\pi^*(T) = \underset{\pi}{\operatorname{argmin}}\, \pi^{\dagger} \hat{\Sigma} \pi, (\hat{\mu}_0 + \hat{\mu}_1 T)^{\dagger} \pi \geq 0.01415\%, 1^{\dagger} \pi = 1. \quad (2.78)$$

6. For each of the next 10 time steps estimate the rate of return as

$$R^{\pi}(t) = R(t)^{\dagger} \pi^*(T). \quad (2.79)$$

7. Repeat from step 1 with $T = T + 10$.

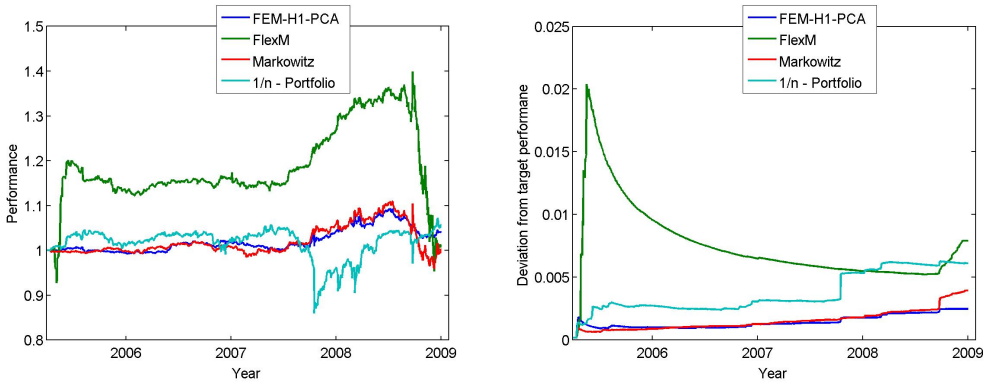

Figure 2.5. *Futures portfolio*[70] *- Comparison of the performance (left panel) and the deviation of the empirical return from the target return (right panel).*

The same is done for a standard Markowitz approach, assuming mean and covariance matrix are stationary and are estimated from the known history. Additionally a GARCH(1,1)[13] model in a multivariate form, is used to estimate the instantaneous covariance matrix. The parameters of this model are estimated with the FlexM-algorithm[55]. The last competing model is a $\frac{1}{n}$ portfolio, where $\frac{1}{8}$th of the total wealth is invested into each asset, buying the two Futures of each commodity with the longer time to maturity and selling the other two.

The performance of all four candidate methods can be seen in the left panel of Figure 2.5. It should be noted, that all four algorithms fall short of the targeted total return of 0.01431% per day. As the optimization was done based on the minimization of the variance from the target return rate, the deviation of the realized daily returns and the target return rate is of interest, that is the quantity

$$D(\tau) = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}(R^{\pi^*}(t) - C)^2}. \tag{2.80}$$

The evolution of this number can be seen in the right panel of Figure 2.5, where the FEM-H1-based portfolio does better than its competitors.

This example demonstrates, that a non-stationary approach might be superior to the usual stationary ansatz. Yet it still relies on a model, that has to be chosen and might be completely wrong. This problem will be addressed in the next chapter.

3 Non-stationary portfolio identification

In this chapter a new approach to the non-stationary portfolio optimization problem is introduced. This new ansatz is based on identifying different market phases by changes in the optimal allocation of assets, instead of relying on a model for observed financial data. Additionally, different kinds of transaction costs are taken into account and a numerical approach is developed, that can be related to the FEM-BV-algorithm. The usefulness of this approach is demonstrated in an example at the end of the chapter.

3.1 The Hybrid-Portfolio

In the literature, two different definitions of (financial) portfolios are known. The one used e.g. by Markowitz[61] is based on the proportion of the total wealth invested into each asset. If not existing already, a risk-free investment is added, the corresponding index is typically 0, to accentuate the special property of being risk-free. For the purpose of this work, a proportion of wealth portfolio will be denoted by the vector $\pi = [\pi_1, \ldots, \pi_d]^\dagger$. A typical assumption is $\pi_0 = 1 - \sum_{i \neq 0} \pi_i$, guaranteeing a complete investment and making the strategy independent of the actual amount of wealth. This allows to identify periods in the investment time frame, that are characterized by identical optimal relative asset allocation. Some advantages of this portfolio class are the independence from the total wealth and the simple structure of the Markowitz portfolio optimization problem

$$\pi^\dagger \mathrm{Cov}\left[R(t)\right] \pi - \alpha \pi^\dagger \mathbb{E}\left[R(t)\right] \to \min_\pi, \tag{3.1}$$

for some risk-affinity parameter α and asset returns $R(t)$.

The second portfolio class, more often used in derivative pricing than portfolio optimization, is the number-of-shares portfolio, here denoted by ζ. For this portfolio class, each component is defined by the number of shares owned by the investor. This kind of portfolio does not provide the scalability of the proportion of wealth type of portfolio, in the sense that is not normalized to wealth 1. Moreover, any normalization factor is only valid for a specific price vector. Yet it has its own advantages, i.e. the portfolio is independent of the development of the asset prices, as the number of shares does not change, while

the value of these shares and the proportion of total wealth do. Assuming a change in the number of shares only occurs in case of a trade (not taking into account stock splits etc.), this can be used as signal for the appearance of transaction costs. The relationship between these two classical portfolio approaches is given by

$$V\pi = \text{diag}[S]\zeta, \tag{3.2}$$

where V is the total wealth and S is the price vector.

As both kinds of portfolios have their own sets of advantages, a hybrid formulation is introduced here. This combines the independence from the current price vector of the proportion-of-wealth portfolio with the independence from the development of the asset prices. This formulation has links to both classical versions.

Definition 3.1.1. (Hybrid portfolio) Let $\xi \in \mathbb{R}^d$ with $\|\xi\|_1 = 1$, $S(t) \in \mathbb{R}^d$ a price vector, $V(t) \in \mathbb{R}^+$ the available wealth and $\Omega \subset \mathbb{R}^d$ the set of permitted investments in number of shares, then ξ is called a permitted hybrid portfolio if there is a $\beta \in \mathbb{R}^+$ such that $\beta\xi^\dagger S(t) = V(t)$ and $\beta\xi \in \Omega$. $\Diamond$

By formulating an investment in the hybrid portfolio form, the holding is split into a vector of proportions of the total holding (in number of shares, ξ) and a scaling factor (β).

The Hybrid portfolio (as defined above) corresponds to a scaled version of the classical number-of-shares portfolio, uniquely mapping any number-of-shares portfolio to a hybrid portfolio, except for the non-investment, $\zeta = 0$, which is mapped to $\beta\xi$ for any ξ and $\beta = 0$.

The definition leads to the following recursive condition on β:

$$\beta(0) = \frac{V(0)}{S(0)^\dagger \xi(0)}, \tag{3.3}$$

$$\beta(t) = \beta(t-1)\frac{S(t)^\dagger \xi(t-1)}{S(t)^\dagger \xi(t)}, t \geq 1. \tag{3.4}$$

This formulation has several advantages compared to the portfolio classes described above: The separation of the scaling parameter $\beta(t)$ from the actual distribution of the shares allows to identify periods of time that are characterized by holding a similar investment, i.e. the portfolio parameters $\xi(t)$ are the same, although they are separated by periods with other optimal investments. At the same time, for the duration of a market phase, neither of the parameters of the portfolio does change.

As already stated above, to distinguish between the different types of portfolios the following notation will be used: π is a proportion of wealth portfolio, ζ is a number of shares portfolio and ξ is a hybrid portfolio. All three classes are connected by the following relation:

$$V\pi = \text{diag}[S]\zeta = \beta\,\text{diag}[S]\xi, \tag{3.5}$$

where V is the total wealth and S is the price vector.

3.2 Utility of the Wealth process

Mathematically, the portfolio optimization problem can be described as an optimal control problem with stochastic (external) influences. Let $u((V(t))_{t\in\mathcal{T}}$ be the "utility" of the wealth process $V(\cdot)$, where $u(\cdot)$ is chosen according to an investors need or preferences. As the realization of the wealth process is a priori unknown, classical approaches aim e.g. at maximizing the expected value of the utility. The dynamics of an investment, depending on the price vector $S(\cdot)$ and the rate of return r can be formulated in all three portfolio definitions as:

$$dV^{\pi}(t) = V^{\pi}(t)\left(\pi(t)^{\dagger}\,\mathrm{diag}[S(t)]^{-1}\,dS(t) + (1-1^{\dagger}\pi(t))r\,dt\right), \tag{3.6}$$

$$dV^{\zeta}(t) = \zeta(t)^{\dagger}\,dS(t) + (V^{\zeta}(t) - \zeta(t)^{\dagger}S(t))r\,dt, \tag{3.7}$$

$$dV^{(\beta,\xi)}(t) = \beta(t)\xi(t)^{\dagger}\,dS(t) + (V^{(\beta,\xi)}(t) - \beta(t)\xi(t)^{\dagger}S(t))r\,dt. \tag{3.8}$$

Example 3.2.1. A classical approach to handle risk while maximizing profit is to maximize the expected logarithmic terminal wealth:

$$E\left[u(V^{\pi}(T))\right] \to \max_{\pi(\cdot)}, u(x) = \begin{cases}\log(x), & x>0\\ -\infty, & x\le 0\end{cases}, \tag{3.9}$$

if the wealth process is almost surely positive and finite with respect to the measure implied by the distribution of the price processes, the expectation takes a finite value that can be optimized. The classical example for this is the Kelly criterion for betting. Suppose $X \in \{0,1\}$ is a Bernoulli distributed random variable with $P[X=1]=p$, the rate of return $r=0$ and the price process $S(t):\{0,1\}\mapsto\mathbb{R}$ is defined as

$$S(0)=1, S(1)=QX \tag{3.10}$$

for some $1<Q\in\mathbb{R}$. Using the utility function defined in Eq. (3.9) and $\pi=\pi(0)$ the expectation takes the form

$$E\left[u(V^{\pi}(1))\right] = p\log\left(V(0)\left(\pi(Q-1)+1\right)\right) + (1-p)\left(V(0)\left(-\pi+1\right)\right). \tag{3.11}$$

With the constraints $\pi\in(-\frac{1}{Q-1},1)$ to satisfy the positivity constraint on $V^{\pi}(1)$. Maximizing Eq. (3.11) under these constraints results in

$$\pi = \frac{pQ-1}{Q-1} \tag{3.12}$$

which satisfies the constraints on π for $p\in(0,1)$. ◊

Example 3.2.2. As an alternative to the utility just depending on the terminal value, the utility can also depend on the time development of the wealth, e.g. in[63], the optimal investment and consumption strategies are found for the problem

$$\max E\left[\int_0^T e^{-\rho s}u(c(V(s),s))\,ds + e^{-\rho T}u(V(T))\right], \tag{3.13}$$

with $c(V(t),t)$ being the consumption at time t and discount factor ρ. Considering a risk free interest rate r, and an expected return of μ and a variance of σ^2 for the risky asset, the wealth process with investment $\pi(t)$ satisfies the dynamics

$$dV(t) = ((r+\pi(t)(\mu-r))V(t) - c(V(t),t))\,dt + V(t)\pi(t)\sigma\,dB(t). \tag{3.14}$$

The solution found in[63] is based on the use of the utility function $u(x) = \frac{x^{1-\alpha}}{1-\alpha}$ and non-negative consumption $c(\cdot) \geq 0$. Under these conditions, the solution is

$$\pi(t) = \frac{\mu - r}{\sigma^2\alpha} \tag{3.15}$$

$$c(V(t),t) = \begin{cases} \nu(1+(\nu\varepsilon-1)e^{-\nu(T-t)})^{-1}V(t), & T<\infty, \nu\neq 0 \\ (T-t+\varepsilon)^{-1}V(t), & T<\infty, \nu = 0 \\ \nu V(t), & T=\infty \end{cases} \tag{3.16}$$

$$\nu = \frac{\rho}{\alpha} - (1-\alpha)\left((\mu-\alpha)\frac{\pi(t)}{2\alpha} + \frac{r}{\alpha}\right), \tag{3.17}$$

for $0 \leq \varepsilon << 1$.. ◊

The main problem with these formulations is the dependence on an (unknown) distribution. Even if assumptions on the distribution are made (e.g. on the parametric family the distribution belongs to), parameters have to be estimated, typically from past observations. This creates the need for these parameters to be (sufficiently) constant in time or a model for the non-stationarity is needed, that comes with parameters itself.

In this work, a method based on empirical utility is proposed. Instead of tuning the parameters of a given model to fit an observed series of data and then using these parameters to find a portfolio that is optimal in the model, the aim is to find a set of portfolios that are optimal on the observed data and to utilize persistency to motivate the usage of these portfolios for future investments. The aim is to skip an intermediate step, where the parametric model is calibrated using the data. To this end, investments are considered as a series of one-period investments, e.g. using daily closing prices and assuming trades are permitted without constraints on the volume for the closing price. This allows to calculate the "empirical utility", that is the realized utility, if an investment strategy $\pi(t)$, $\zeta(t)$ or $(\beta(t),\xi(t))$ had been employed.

Example 3.2.3. Let $\mu_\tau^{(\beta,\xi)}$ be the unconditional mean return of the wealth process with investment strategy (β,ξ) and length of investment period τ, thus

$$\mu_\tau^{(\beta,\xi)} = E\left[\left.\frac{V_\tau^{(\beta,\xi)}(t+\tau)}{V_\tau^{(\beta,\xi)}(t)}\right| t \sim U[0,T-\tau]\right], \tag{3.18}$$

where $U[0,T-\tau]$ is the uniform distribution on $[0,T-\tau]$. Please note, that this does not imply that the returns of the wealth process have constant mean. Accordingly, the unconditional variance of the return of the wealth process shall be denoted by $\sigma_\tau^{(\beta,\xi)}$

$$\begin{aligned}\sigma_\tau^{(\beta,\xi)} &= \mathrm{Var}\left[\left.\frac{V_\tau^{(\beta,\xi)}(t+\tau)}{V_\tau^{(\beta,\xi)}(t)}\right| t \sim U[0,T-\tau]\right] \\ &= E\left[\left.\left(\frac{V_\tau^{(\beta,\xi)}(t+\tau)}{V_\tau^{(\beta,\xi)}(t)} - E\left[\left.\frac{V_\tau^{(\beta,\xi)}(t+\tau)}{V_\tau^{(\beta,\xi)}(t)}\right| t \sim U[0,T-\tau]\right]\right)^2\right| t \sim U[0,T-\tau]\right].\end{aligned} \tag{3.19}$$

Then the Markowitz problem with risk-affinity α for one period can be written as

$$\max_{(\beta,\xi)} \alpha\mu_\tau^{(\beta,\xi)} - \sigma_\tau^{(\beta,\xi)} \tag{3.20}$$

Moreover, the mean and variance can be estimated empirically from the N observation

$$\mu_\tau^{(\beta,\xi)} \approx \hat{\mu}_\tau^{(\beta,\xi)} = \frac{1}{N}\sum_{j=0}^{N-1} \frac{V_\tau^{(\beta,\xi)}((j+1)\tau)}{V_\tau^{(\beta,\xi)}(j\tau)}, \tag{3.21}$$

$$\sigma_\tau^{(\beta,\xi)} \approx \hat{\sigma}_\tau^{(\beta,\xi)} = \frac{1}{N}\sum_{j=0}^{N-1} \left(\frac{V_\tau^{(\beta,\xi)}((j+1)\tau)}{V_\tau^{(\beta,\xi)}(j\tau)} - \hat{\mu}_\tau^{(\beta,\xi)}\right)^2. \tag{3.22}$$

Now the wealth process can be replaced by the price process and the asset allocation to obtain

$$\hat{\mu}_\tau^{(\beta,\xi)} = \frac{1}{N}\sum_{j=0}^{N-1} \frac{S((j+1)\tau)^\dagger \xi(j\tau)}{S(j\tau)^\dagger \xi(j\tau)}, \tag{3.23}$$

$$\hat{\sigma}_\tau^{(\beta,\xi)} = \frac{1}{N}\sum_{j=0}^{N-1} \left(\frac{S((j+1)\tau)^\dagger \xi(j\tau)}{S(j\tau)^\dagger \xi(j\tau)} - \hat{\mu}_\tau^{(\beta,\xi)}\right)^2. \tag{3.24}$$

Thus by defining the empirical utility function as

$$u(V_\tau^{(\beta,\xi)}) = \alpha\hat{\mu}_\tau^{(\beta,\xi)} - \hat{\sigma}_\tau^{(\beta,\xi)} \tag{3.25}$$

the Markowitz portfolio problem is transformed into a problem with empirical utility, which is additionally independent of the scaling factor β. Moreover, the structure of the problem is similar to the inverse problem as in Eq. (2.2). $\Diamond$

By further assuming that the utility $u(\cdot)$ can be decomposed into a sum

$$u(V_\tau^{\beta,\xi}) = \sum_{j=0}^{N-1} \Delta u(V_\tau^{\beta,\xi(t_j)}, V_\tau^{\beta,\xi(t_{j+1})}, \hat{\Theta}), \tag{3.26}$$

with the aim of $\Delta u(\cdot)$ not depending on $(\beta(t),\xi(t))$ for $t \notin [j\tau,(j+1)\tau]$, with the dependence on the remaining investment decisions being expressed by the additional parameter $\hat{\Theta}$. E.g. for the logarithmic utility as in example 3.2.1, this can immediately be done by the decomposition

$$\begin{aligned} u(V_\tau^{(\beta,\xi)}) &= \log \frac{V_\tau^{\beta,\xi}(T)}{V_\tau^{(\beta,\xi)}(0)} = \sum_{j=0}^{N-1} \log \frac{V_\tau^{\beta,\xi}((j+1)\tau)}{V_\tau^{(\beta,\xi)}(j\tau)} \\ &= \sum_{j=0}^{N-1} \log \frac{S((j+1)\tau)^\dagger \xi(j\tau)}{S(j\tau)^\dagger \xi(j\tau)} = \sum_{j=0}^{N-1} \Delta u(V_\tau^{(\beta,\xi)}(j\tau), V_\tau^{(\beta,\xi)}((j+1)\tau)). \end{aligned} \tag{3.27}$$

In a case of the Markowitz portfolio problem, as in the example 3.2.3, the decomposition is not as fully independent as in the case of logarithmic utility. The expression of the mean in the variance calculation can not be completely separated. In the spirit of e.g. the Metropolis-Hastings-algorithm[65], the following decomposition is proposed:

$$\begin{aligned} \Delta u(V_\tau^{(\beta,\xi)}(j\tau), V_\tau^{(\beta,\xi)}((j+1)\tau), \hat{\mu}) &= \frac{\alpha}{N} \frac{V_\tau^{(\beta,\xi)}((j+1)\tau)}{V_\tau^{(\beta,\xi)}(j\tau)} - \frac{1}{N} \left(\frac{V_\tau^{(\beta,\xi)}((j+1)\tau)}{V_\tau^{(\beta,\xi)}(j\tau)} - \hat{\mu} \right)^2 \\ &= \frac{\alpha}{N} \frac{S((j+1)\tau)^\dagger \xi(j\tau)}{S(j\tau)^\dagger \xi(j\tau)} - \frac{1}{N} \left(\frac{S((j+1)\tau)^\dagger \xi(j\tau)}{S(j\tau)^\dagger \xi(j\tau)} - \hat{\mu} \right)^2, \end{aligned} \tag{3.28}$$

where the current estimate of the mean $\hat{\mu}$ is updated independently of the investment strategy $(\beta(t),\xi(t))$.

For the purpose of this work, the strategies are going to be further restricted. The set of permitted strategies is limited to the set $\mathcal{P}_\mathbb{K}$, defined as

$$\mathcal{P}_\mathbb{K} = \left\{ \xi : \mathcal{T} \mapsto \mathbb{R}^d \,\middle|\, \exists \gamma : \mathcal{T} \mapsto \{0,1\}^\mathbb{K}, \sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \xi(t) = \sum_{i=1}^{\mathbb{K}} \gamma_i(t) \xi^{(i)} \right\}. \tag{3.29}$$

This is the condition, that the shape parameter of the investment strategy can take only $\mathbb{K}$ different values. By introducing this assumption, the problem of finding the optimal asset allocation becomes the problem of identifying the optimal $\mathbb{K}, \xi^{(1)} \ldots . \xi^{(\mathbb{K})}$ and the affiliation function $\gamma(\cdot)$. As stated in Chapter 2, this problem is ill-posed. As a possible solution, regularization was introduced there, yet instead of artificially imposing regularization, the same can be achieved by introducing transaction costs. The transaction cost can be controlled by an upper bound, which is then in turn bounded from above to enforce regularization.

3.3 Transaction cost models

For the purpose of this work, three different transaction cost models are considered: fixed, volume and value based costs. Additionally, it will be demonstrated how to include each of these three models into the current setting. Most transaction costs in the real world can be categorized into these three classes, or as their combination.

3.3.1 Fixed transaction costs

Fixed transaction costs occur whenever a transaction is performed, but the cost does not depend on the volume (the number of assets traded) nor the value of the transaction, but on the presence of the assets that are traded. This is a typical case for small investors, who are charged a fixed fee per transaction e.g. because they do not reach the minimum transaction size to qualify for another type of transaction cost.

To formulate this type of transaction cost, let $\chi(\cdot)$ be the element-wise indicator function of $\mathbb{R}^d \setminus \{0\}$ for generic d, thus

$$\chi : \mathbb{R}^d \mapsto \mathbb{R}^d : \chi(x) = y \text{ with } y_i = \begin{cases} 0, & x_i = 0 \\ 1, & \text{else} \end{cases}. \tag{3.30}$$

Moreover let $c \in \mathbb{R}^d$ be the vector of per-transaction costs for each asset, then the cost inferred by the transaction at time t and period length τ is

$$c^\dagger \chi \left(\beta(t)\xi(t) - \beta(t-\tau)\xi(t-\tau)\right) = c^\dagger \chi \left(\beta(t)\sum_{i=1}^{\mathbb{K}} \gamma_i(t)\xi^{(i)} - \beta(t-\tau)\sum_{i=1}^{\mathbb{K}} \gamma_i(t-\tau)\xi^{(i)}\right) \tag{3.31}$$

$$= c^\dagger \chi \left(\sum_{i=1}^{\mathbb{K}} \left(\beta(t)\gamma_i(t) - \beta(t-\tau)\gamma_i(t-\tau)\right)\xi^{(i)}\right). \tag{3.32}$$

If $\gamma(t) = \gamma(t-\tau)$, the scaling parameter does not change ($\beta(t) = \beta(t-\tau)$), and thus the cost is zero. In the case $\gamma(t) \neq \gamma(t-\tau)$, the following upper bound holds:

$$c^\dagger \chi \left(\beta(t)\xi(t) - \beta(t-\tau)\xi(t-\tau)\right) \leq c^\dagger \sum_{i=1}^{\mathbb{K}} |\gamma_i(t) - \gamma_i(t-\tau)| \, \chi\left(\xi^{(i)}\right). \tag{3.33}$$

This bound corresponds to selling the whole current holding and buying the new one. This upper bound is twice the caused costs for each asset that has a non-zero position in the current and the new holding. If the two holdings in one asset are equal (but not necessarily

in the shape parameter ξ), no transaction would have been necessary, reducing the real cost for that asset to zero. The upper bound is sharp only for assets that appear in one of the two holdings only.

Additionally the upper bound is linear (almost everywhere) in the γ. This is used to efficiently solve the minimization problem in $\gamma(\cdot)$, as shown in Section 3.4.

3.3.2 Volume based transaction costs

In the case of volume based transaction costs, the fee depends on the number of shares (or options, futures, ...) traded. A common example for this type of transaction costs is the trading of futures, where the buy and sell prices offered by the broker are differing by a fixed gap, the so called spread. Unlike for example the stock exchange, where any transaction happens between two arbitrary market participants, futures are bought from or sold to the broker, who profits from the spread.

Volume based transaction costs can be formulated as follows: Let $c \in \mathbb{R}^d$ be the vector of cost per unit per asset, thus c_i is the cost for trading one unit of the ith asset. Then the cost of the transaction at time t can be formulated and bound from above as

$$\left\|\operatorname{diag}[c]\left(\beta(t)\xi(t)-\beta(t-\tau)\xi(t-\tau)\right)\right\|_1 \tag{3.34}$$

$$=\left\|\sum_{i=1}^{\mathbb{K}}\operatorname{diag}[c]\left(\beta(t)\gamma_i(t)\xi^{(i)}-\beta(t-\tau)\gamma_i(t-\tau)\xi^{(i)}\right)\right\|_1 \tag{3.35}$$

$$\leq\sum_{i=1}^{\mathbb{K}}\left\|\operatorname{diag}[c]\xi^{(i)}\right\|_1\left|\beta(t)\gamma_i(t)-\beta(t-\tau)\gamma_i(t-\tau)\right|. \tag{3.36}$$

This bound is modeling the transaction again as selling the whole investment at a given time and buying the new allocation, which infers to much transaction costs, as only the difference would have been needed to be bought or sold.

The bound can be further simplified by defining $\beta_{\max}\geq\max_{t\in\mathcal{T}}\beta(t)$ and estimating

$$\sum_{i=1}^{\mathbb{K}}\left\|\operatorname{diag}[c]\xi^{(i)}\right\|_1\left|\beta(t)\gamma_i(t)-\beta(t-\tau)\gamma_i(t-\tau)\right| \tag{3.37}$$

$$\leq\sum_{i=1}^{\mathbb{K}}\left\|\operatorname{diag}[c]\xi^{(i)}\right\|_1\max\{\beta(t),\beta(t-\tau)\}\left|\gamma_i(t)-\gamma_i(t-\tau)\right| \tag{3.38}$$

$$\leq\beta_{\max}\sum_{i=1}^{\mathbb{K}}\left\|\operatorname{diag}[c]\xi^{(i)}\right\|_1\left|\gamma_i(t)-\gamma_i(t-\tau)\right|. \tag{3.39}$$

This bound, while not sharp, has the advantage of linearity almost everywhere with respect to $\gamma_i(\cdot)$, as in the case of fixed transaction costs.

3.3.3 Value based transaction costs

The case of value based transaction costs is very common for investors. Here the cost depends on the value of a transaction, as is typical for tax, exchange fees and banking fees for larger transactions. This kind of transaction cost is similar to the one in the previous subsection, with one additional factor:

$$\|\mathrm{diag}[c]\,\mathrm{diag}[S(t)]\,(\beta(t)\xi(t)-\beta(t-\tau)\xi(t-\tau))\|_1 \tag{3.40}$$

$$=\left\|\sum_{i=1}^{\mathbb{K}}\mathrm{diag}[c]\,\mathrm{diag}[S(t)]\left(\beta(t)\gamma_i(t)\xi^{(i)}-\beta(t-\tau)\gamma_i(t-\tau)\xi^{(i)}\right)\right\|_1 \tag{3.41}$$

$$\leq\sum_{i=1}^{\mathbb{K}}\left\|\mathrm{diag}[c]\,\mathrm{diag}[S(t)]\xi^{(i)}\right\|_1|\beta(t)\gamma_i(t)-\beta(t-\tau)\gamma_i(t-\tau)|\,. \tag{3.42}$$

Subsequently, also the upper bound to model the transaction costs in this example can be interpreted as a complete liquidation followed by buying the new allocation. And again this bound is very crude, yet it can be further estimated as

$$\sum_{i=1}^{\mathbb{K}}\left\|\mathrm{diag}[c]\,\mathrm{diag}[S(t)]\xi^{(i)}\right\|_1|\beta(t)\gamma_i(t)-\beta(t-\tau)\gamma_i(t-\tau)| \tag{3.43}$$

$$\leq\sum_{i=1}^{\mathbb{K}}\left\|\mathrm{diag}[c]\,\mathrm{diag}[S(t)]\xi^{(i)}\right\|_1\max\{\beta(t),\beta(t-\tau)\}\,|\gamma_i(t)-\gamma_i(t-\tau)| \tag{3.44}$$

$$\leq\beta_{\max}\sum_{i=1}^{\mathbb{K}}\left\|\mathrm{diag}[c]\,\mathrm{diag}[S(t)]\xi^{(i)}\right\|_1|\gamma_i(t)-\gamma_i(t-\tau)|\,, \tag{3.45}$$

with $\beta_{\max}\geq\max_{t\in\mathcal{T}}\beta(t)$. Again linearity in $\gamma_i(\cdot)$ is achieved almost everywhere at the cost of the sharpness of the bound.

3.4 Analysis of Historical Data for fixed meta-parameters

The process of analyzing historical data is going to be demonstrated in this section. In this first step the optimal asset allocations for the past are found. For the purpose of this section, it will be assumed that the number of individual portfolios $\mathbb{K}$ and the maximum accumulated cost C are known. If this is not the case, the optimal combination of $\mathbb{K}$ and C has to be identified. Different approaches to this problem are investigated in Chapter 4.

Let $\mathcal{T}$ be the set of observation times, $S(t)\in\mathbb{R}^d, t\in\mathcal{T}$ the price vector at the observation times and $u(\cdot)$ the utility function that has to be maximized with decomposition $\Delta u(\cdot)$. Moreover, let $c(\cdot)$ be the cost function. The portfolio optimization problem with limited

cost level C can then be formulated as

$$\max_{(\beta(t),\xi(t))_{t\in\mathcal{T}}} u(V^{(\beta,\xi)}), \sum_{t_j\in\mathcal{T}\setminus\{T\}} c(\beta(t_j),\xi(t_j),\beta(t_{j+1}),\xi(t_{j+1}),S(t_{j+1})) \leq C. \tag{3.46}$$

As stated in Section 3.2, by limiting the set of permitted strategies to

$$\mathcal{P}_{\mathbb{K}} = \left\{ \xi : \mathcal{T} \mapsto \mathbb{R}^d \middle| \exists \gamma : \mathcal{T} \mapsto \{0,1\}^{\mathbb{K}}, \sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \xi(t) = \sum_{i=1}^{\mathbb{K}} \gamma_i(t)\xi^{(i)} \right\}, \tag{3.47}$$

the problem is transformed to the FEM-BV-Utility problem:

$$\min_{\gamma_i(\cdot),\xi^{(1)},\ldots,\xi^{(\mathbb{K})},\hat{\Theta}} -\sum_{i=1}^{\mathbb{K}} \sum_{t_j\in\mathcal{T}\setminus\{T\}} \gamma_i(t_j)\Delta u(V^{(\beta(t_j),\xi(t_j))}, V^{(\beta(t_{j+1}),\xi(t_{j+1}))}, \hat{\Theta}) \tag{3.48}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \forall t \in \mathcal{T} \tag{3.49}$$

$$\gamma_i(t) \in \{0,1\}, \forall i = 1,\ldots,\mathbb{K}, t \in \mathcal{T} \tag{3.50}$$

$$\sum_{t_j\in\mathcal{T}\setminus\{T\}} c\left(\beta(t_j), \sum_{i=1}^{\mathbb{K}} \gamma_i(t_j)\xi^{(i)}, \beta(t_{j+1}), \sum_{i=1}^{\mathbb{K}} \gamma_i(t_{j+1})\xi^{(i)}, S(t_{j+1})\right) \leq C. \tag{3.51}$$

Please note that the function $c(\cdot)$ in inequality (3.51) is almost linear, if fixed transaction costs or the crude bounds from Equation (3.39) or (3.45) are used. Indeed this can be interpreted as a weighted version of the constraints of the FEM-BV-method. To solve this kind of problem, additional slip variables $\omega_i(\cdot)$ are defined. Let $\varphi_i(t)$ be the weights of the variational bounds, thus

$$c\left(\beta(t_j), \sum_{i=1}^{\mathbb{K}} \gamma_i(t_j)\xi^{(i)}, \beta(t_{j+1}), \sum_{i=1}^{\mathbb{K}} \gamma_i(t_{j+1})\xi^{(i)}, S(t_{j+1})\right) = \sum_{i=1}^{\mathbb{K}} \varphi_i(t)|\gamma_i(t_{j+1}) - \gamma_i(t_j)| \tag{3.52}$$

then the problem can be reformulated as

$$\min_{\gamma_i(\cdot),\xi^{(1)},\ldots,\xi^{(\mathbb{K})},\hat{\Theta},\omega_i(\cdot)} -\sum_{i=1}^{\mathbb{K}} \sum_{t_j\in\mathcal{T}\setminus\{T\}} \gamma_i(t_j)\Delta u(V^{(\beta(t_j),\xi(t_j))}, V^{(\beta(t_{j+1}),\xi(t_{j+1}))}, \hat{\Theta}) \tag{3.53}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \forall t \in \mathcal{T} \tag{3.54}$$

$$\gamma_i(t) \in \{0,1\}, \forall i = 1,\ldots,\mathbb{K}, t \in \mathcal{T} \tag{3.55}$$

$$\varphi_i(t)(\gamma_i(t_{j+1}) - \gamma_i(t_j)) \leq \omega_i(t_j), \forall i = 1,\ldots,\mathbb{K}, t_j \in \mathcal{T}\setminus\{T\} \tag{3.56}$$

$$-\varphi_i(t)(\gamma_i(t_{j+1}) - \gamma_i(t_j)) \leq \omega_i(t_j), \forall i = 1,\ldots,\mathbb{K}, t_j \in \mathcal{T}\setminus\{T\} \tag{3.57}$$

$$\sum_{i=1}^{\mathbb{K}} \sum_{t_j\in\mathcal{T}\setminus\{T\}} \omega_i(t_j) \leq C. \tag{3.58}$$

To solve this problem, the Algorithm 4 is used.

Algorithm 4 Data based Portfolio Optimization Algorithm

Require: Time series of asset prices $S(\cdot) \in \mathbb{R}^d$, number of portfolios $\mathbb{K}$, maximum transaction cost level C, function of cost-weights $\varphi(\cdot)$

1: Choose $\mathbb{K}-1$ random transition points $0 < \tau_1 < \ldots, \tau_{\mathbb{K}-1} < T$ points and generate $\Gamma^{(0)}$ with

$$\gamma_i^{(0)}(t) = \begin{cases} 1, & t \in [\tau_{i-1}, \tau_i) \\ 0, & \text{else} \end{cases}, \tau_0 = 0, \tau_{\mathbb{K}} = T. \tag{3.59}$$

2: Choose $\mathbb{K}$ random vectors $\xi^{(1)},\ldots,\xi^{(\mathbb{K})} \in \mathbb{R}^{d+1}$ with $\|\xi^{(\cdot)}\|_1 = 1$,
3: **while** Total Cost > C **do**
4: Set smallest entry in $[\xi_1,\ldots,\xi_{\mathbb{K}}]$ that is not the cash component to zero and renormalize.
5: **end while**
6: Call Optimization loop (Algorithm 5)

3.5 Example for German market data

As an example, the FEM-BV-Utility is compared to the (textbook) Markowitz-approach and an ansatz based on the idea, that the asset-returns can be modeled by a Gaussian Mixture Model[9,62]. Additionally a GARCH-approach[13], as formulated in its multivariate form[55], and a buy-and-hold investment with equal investment into each asset are considered. A Hidden Markov Model with Gaussian emission[33] is included as well and is shown

Algorithm 5 Portfolio Optimization Algorithm - Optimization loop (part of 4)

repeat
  for k=1 to $\mathbb{K}$ **do**
    Estimate $\xi^{(k)}$ for fixed Γ by finding an improvement over the last estimate as an approximate solution to the optimization problem

$$\min_{\xi^{(k)}} -\sum_{t_j \in \mathcal{T}\setminus\{T\}} \gamma_k(t_j)\Delta u(V^{(\beta(t_j),\xi(t_j))}, V^{(\beta(t_{j+1}),\xi(t_{j+1}))}, \hat{\Theta}) \tag{3.60}$$

$$\gamma_i(t) \in \{0,1\}, \forall i = 1,\ldots,\mathbb{K}, t \in \mathcal{T} \tag{3.61}$$

$$\varphi_i(t)(\gamma_i(t_{j+1}) - \gamma_i(t_j)) \leq \omega_i(t_j), \forall i = 1,\ldots,\mathbb{K}, t_j \in \mathcal{T}\setminus\{T\} \tag{3.62}$$

$$-\varphi_i(t)(\gamma_i(t_{j+1}) - \gamma_i(t_j)) \leq \omega_i(t_j), \forall i = 1,\ldots,\mathbb{K}, t_j \in \mathcal{T}\setminus\{T\} \tag{3.63}$$

$$\sum_{i=1}^{\mathbb{K}} \sum_{t_j \in \mathcal{T}\setminus\{T\}} \omega_i(t_j) \leq C. \tag{3.64}$$

  end for
  Set $L = \emptyset$ the set of locked-out solutions in Γ.
  repeat
    Estimate the current weights $\bar{\varphi}(t) = \varphi(t)$ and compute Γ for fixed $\xi^{(1)},\ldots,\xi^{(\mathbb{K})}$ by solving the integer problem

$$\min_{\gamma_i(\cdot),\omega_i(\cdot)} -\sum_{i=1}^{\mathbb{K}} \sum_{t_j \in \mathcal{T}\setminus\{T\}} \gamma_i(t_j)\Delta u(V^{(\beta(t_j),\xi(t_j))}, V^{(\beta(t_{j+1}),\xi(t_{j+1}))}, \hat{\Theta}) \tag{3.65}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \forall t \in \mathcal{T} \tag{3.66}$$

$$\gamma_i(t) \in \{0,1\}, \forall i = 1,\ldots,\mathbb{K}, t \in \mathcal{T} \tag{3.67}$$

$$\bar{\varphi}_i(t)(\gamma_i(t_{j+1}) - \gamma_i(t_j)) \leq \omega_i(t_j), \forall i = 1,\ldots,\mathbb{K}, t_j \in \mathcal{T}\setminus\{T\} \tag{3.68}$$

$$-\bar{\varphi}_i(t)(\gamma_i(t_{j+1}) - \gamma_i(t_j)) \leq \omega_i(t_j), \forall i = 1,\ldots,\mathbb{K}, t_j \in \mathcal{T}\setminus\{T\} \tag{3.69}$$

$$\sum_{i=1}^{\mathbb{K}} \sum_{t_j \in \mathcal{T}\setminus\{T\}} \omega_i(t_j) \leq C, \tag{3.70}$$

$$\gamma(\cdot) \notin L. \tag{3.71}$$

    if Total Cost > C **then**
      (7) Block the solution that was just found by $L = L \cup \{\Gamma\}$
    end if
  until Total Cost <= C
until Convergence

in the following as a "Hamilton"-portfolio. To complete the comparisons, a simple 60-days Moving Average (or "Rolling Window") approach is considered. This is a special version of the local kernel method as in Eq. (2.6), with kernel function defined as

$$W(s) = \begin{cases} \frac{1}{60}, & s \in [-60, 0) \\ 0, & \text{else} \end{cases}. \tag{3.72}$$

This results in the estimation of the Markowitz parameters by using only the last 60 observations. The Markowitz-approach and the buy-and-hold investment have to be classified as static investment strategies, as new observations do not influence the investment in any form. For all other approaches, the optimal portfolio might change with newly available information and thus the strategies are dynamic.

The methods are compared in-sample and first without transaction costs, before a fixed transaction cost is introduced.

The data are German market data: One asset is the German Stock index (DAX) in the realization as a total return index (as opposed to a price index). As a second asset a German government issued bond is considered, that is stripped of its yearly interest-payments. This allows to trade both assets without taking into account possible payouts during the time-span of the trading.

The data used for the analysis are 04-Oct-1999 to 14-Jun-2013, yielding in 3495 daily closing prices. Out of this data set, the first 3000 time steps are used for the In-sample analysis, the remaining 495 time steps are held "hidden" from the models to allow a later study of the Out-of-sample performance. Additionally a "no-short-selling" constraint was used. This constraint reflects the typical situation for a private investor and reduces the risk of default or subsequent capital contributions.

For the numerical solving of the problem, the SCIP[2] mixed integer programming solver, the SoPlex[81] Linear Programming solver and an SQP solver[69] were used in the context of the Algorithm 4 described above.

3.5.1 Analysis without transaction costs

For this example no trading costs are considered, therefore the setting is as proposed in[61].

For the FEM-BV-Utility-method, the number of local models ($\mathbb{K}$) is chosen as $\mathbb{K} = 2$. The Gaussian Mixture Model analysis of the returns leads to a superior (in the sense of information theory, see also Chapter 4) result for $\mathbb{K} = 2$ when compared to the stationary case ($\mathbb{K} = 1$) or a larger number of Gaussians. Additionally time dependent means of the form $\mu_i(t) = \mu_i^{(0)} + \mu_i^{(1)} t$ were considered, but have been inferior. For the Hidden Markov Model with Gaussian emissions, the number of states of the hidden variable was estimated to be 2.

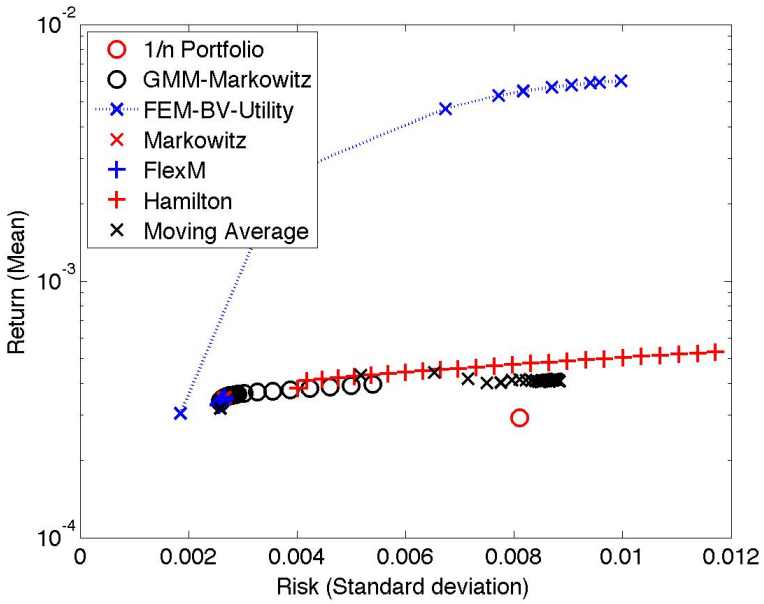

Figure 3.1. *Simple Example, No Transaction costs - In-sample comparison of different portfolio optimization algorithms. The FEM-BV-Utility-approach shows a significant better risk-return rate than the Gaussian Mixture Model-based Markowitz approach, the standard Markowitz approach, the multivariate GARCH model (FlexM), the* $\frac{1}{n}$*-Portfolio, the Hamilton approach or the Moving Average (MA) based model.*

The optimal portfolios for different risk-aversion parameters α are estimated and the mean and the standard deviation of the daily (in-sample) returns are calculated and plotted. The according results can be seen in Figure 3.1.

To take into account both statistical properties, the Sharpe-Ratio, as introduced in[74] is used. The Sharpe-Ratio R is defined as

$$R = \frac{\mu - r}{\sigma}, \tag{3.73}$$

$$\mu = E\left[\frac{V_t}{V_{t-1}} - 1\right], \tag{3.74}$$

$$\sigma = \sqrt{E\left[\left(\frac{V_t}{V_{t-1}} - 1 - \mu\right)^2\right]}. \tag{3.75}$$

As the Sharpe-Ratio is given versus a reference interest rate r, this parameter is chosen such that the ECB's target-inflation rate of 2% per year is matched. The Sharpe-Ratios of the FEM-BV-Utility-algorithm for different risk-aversion parameters α can be seen in Figure

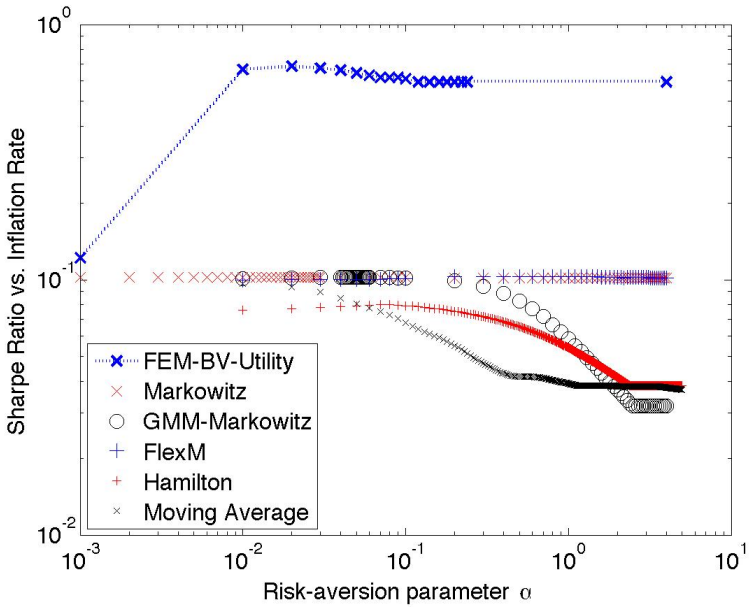

Figure 3.2. *Simple Example, No Transaction costs - In-sample Sharpe-Ratio of the optimized portfolio depending on the risk-aversion parameter α for considered portfolio optimization methods.*

3.2. The optimal values for the risk-affinity parameter α and the resulting Sharpe-Ratio are shown in table 3.1. Additionally the (annual) portfolio turnover (ATO) is shown. This value is calculated as:

$$\text{ATO} = \sum_t \|\pi_t - \pi_{t-}\|_1 \frac{365}{\text{Length of observation in days}}, \tag{3.76}$$

as e.g. in [15,21], where π_t is the (chosen) investment in proportion of wealth after possible adjustments at time t, and π_{t-} is the investment before the adjustment at time t.

It should be noted, that the Sharpe-Ratio of the FEM-BV-Utility-method is higher, due to the much better (in-sample) performance of the FEM-BV-Utility-method.

3.5.2 Analysis with fixed transaction costs

In this subsection the performance of the FEM-BV-Utility-algorithm is compared to the same competitors as in the last subsection, but this time fixed transaction costs are considered.

To compare the different algorithms, the whole dataset (3000 observations) is used to estimate the parameters required by the different methods. Those parameters are then used

Method	FEM-BV-Utility	Markowitz	GMM	FlexM	$\frac{1}{n}$-Portfolio	Hamilton	MA
α	0.02	0.013	0.051	3.2	-	0.07	0.01
R	0.6855	0.1022	0.1024	0.1019	0.0286	0.08	0.0965
ATO	248.3727	0.2426	0.2492	0.1531	0.0003	1.5771	3.5815
#Tran.	2022	2998	2998	2997	0	2998	2996

Table 3.1. *Optimal (maximizing Sharpe-Ratio) risk-aversion parameters obtained in-sample. The obtained Sharpe-Ratio for the FEM-BV-Utility method is significantly higher, indicating a higher return and/or less risk.The annualized portfolio turnover (ATO) of the FEM-BV-Utility is much higher than the ones of the standard methods, as the method makes use of the free transactions.*

to calculate the portfolio at time t. For the Markowitz algorithm, the FlexM based multivariate GARCH algorithm, the Gaussian Mixture Model-Markowitz methods, the Hidden markov Model and the 60-days moving average approach, the candidate portfolio at time t is chosen by solving the minimization problem

$$\pi^\dagger \Sigma(t) \pi^\dagger - \alpha \left(\mu(t)^\dagger \pi - \frac{2c}{V(t)} \right) \to \min_\pi, 1^\dagger \pi = 1, \pi_i \geq 0. \tag{3.77}$$

Where c is the cost per transaction and $\mu(t)$ and $\Sigma(t)$ are the current estimates of the mean and covariance matrix according to the model. The result is then compared to

$$\pi_{t-}{}^\dagger \Sigma(t) \pi_{t-}{}^\dagger - \alpha \mu(t)^\dagger \pi_{t-}, \tag{3.78}$$

where the portfolio π_{t-} is the portfolio obtained from the holding at the previous time-step, modified by the "mechanical" changes, thus

$$\pi_{t-} = \frac{\operatorname{diag}[S(t)] \operatorname{diag}[S(t-1)]^{-1} \pi_{t-1}}{S(t)^\dagger \operatorname{diag}[S(t-1)]^{-1} \pi_{t-1}}. \tag{3.79}$$

This means that the costs are subtracted from the next periods return, making small adjustments of the portfolio undesirable.

The portfolio remains unchanged if the result of Eq. (3.78) is smaller than the result of the optimization as in Eq. (3.77). This method is similar to the optimization proposed in[58], with the additional benefit of minimizing the variance.

For the FEM-BV-Utility method, the stationary case $\mathbb{K} = 1$ and the non-stationary case with two portfolios $\mathbb{K} = 2$ are considered. The maximum cost level C and the risk-aversion parameter α were chosen to maximize the Sharpe-Ratio for all competitors.

The analysis results in a three dimensional efficiency frontier, which is shown for the FEM-BV-Utility method with $\mathbb{K} = 1$ and $\mathbb{K} = 2$ in Figure 3.3. This figure implies, that

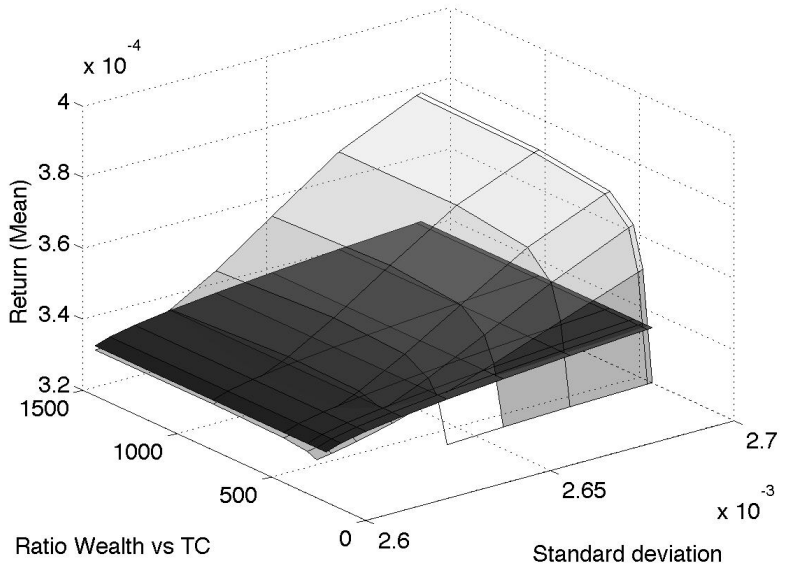

Figure 3.3. *Simple Example, Fixed Transaction Costs - The efficiency frontier depending on the ratio of initial wealth to cost per transaction (TC). The solid surface corresponds to the FEM-BV-Utility method for $K = 1$, while the transparent surface represents the case $K = 2$. For high transaction costs (relative to the initial wealth), or to achieve better performance with small risk, the case $K = 1$ is giving better results. This corresponds to buying a portfolio once and holding it for the rest of the time-frame, as opposed to the case of $\mathbb{K} = 2$, where additional trades are performed.*

for a high cost per transaction to initial wealth ratio, the stationary case $\mathbb{K} = 1$, which corresponds to a Buy-and-Hold investment, is favorable over the strategy which switches between two different portfolios.

The influence of the transaction costs on all portfolio optimization methods is graphically demonstrated in Figure 6.4, showing only the best value for each case, with respect to α and C. As was observed also in the three dimensional plot in Figure 3.3, for high transaction costs, the stationary case is superior to the non-stationary case of the FEM-BV-Utility method. Moreover in all cases, the FEM-BV-Utility method shows better in-sample performance.

This analysis leaves us with the following open questions:

- How to choose the parameters $\mathbb{K}$ and C?
- Can the result of the analysis be used for future investment decisions?

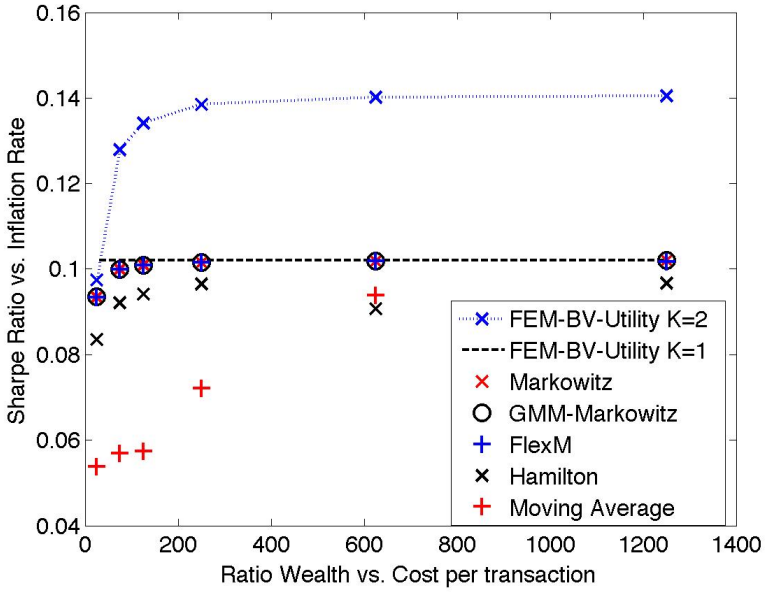

Figure 3.4. *Simple Example, Fixed Transaction Costs - Sharpe-Ratio vs. the ECB target inflation rate for different ratios of starting wealth vs. (fixed) cost per transaction. The risk aversion parameter α is chosen to maximize this measure. For high transaction costs (covering 25 transactions, left-most point) the parameter choice $K = 1$ in the FEM-BV-Utility method produces better results, this corresponds to buying a portfolio once and holding it for the rest of the time-frame.*

- How can new observations be incorporated efficiently in the analysis?

The answer to the first question is given in the following chapter, while the later two questions are answered in Chapter 5

4 Model discrimination

In this chapter, the problem of identifying the parameters for the FEM-BV-Method is examined, i.e. finding the "best" combination of the number of parametersets $\mathbb{K}$ and the regularity bound C. To this end, first a way to measure the information content of a time series is introduced. This method can be directly applied to the data in case of probabilistic models and binary affiliation functions as will be shown in the first section. In the second section, this criterion is expanded to the case of non-probabilistic and non-parametric models, which also allows to evade the binary cluster affiliations and thus makes it applicable also in the case of e.g. H^1-regularization. In the last section of this chapter, a more pragmatic approach is derived directly from the portfolio optimization problem.

4.1 Model selection for probabilistic models and binary affiliation functions: Akaike's information criterion

Using the model distance directly to measure the optimality of a model in terms of its parameters $\mathbb{K}$ and C is bound to fail. An increase in the number of possible parameter sets $\mathbb{K}$ or in the permitted transitions between them corresponds to a relaxation of the constraints and thus to a decrease in the model distance. In particular, if the number of possible parameter sets and the number of transitions is increased to the number of observations, then the model degenerates to describing each data point individually. This effect is known as over-fitting.

One way of avoiding over-fitting, while simultaneously allowing enough freedom to identify good parameters is the separation of a training set and a test set, a method called cross-validation[53] and typically used for training models with external influences, e.g. artificial neural networks[10]. While the number of free parameters is to small, an increase will improve the parameter fit on the training data as well as on the test set. When the number of free parameters grows further, at a certain point over-fitting occurs. From this point on, increasing the number of free parameters will still improve the fit on the training data, but not on the test set. In most cases, the fit on the test data gets worse, as the model

gets adjusted to effects that are merrily random effects than results of underlying dynamics, and thus becomes ineffective outside the training set. This effect is also known as "curve fitting", as the model is adjusted to a specific observation (plotted as a curve) and not to the underlying dynamics.

Another way of selecting the number of parameter sets $\mathbb{K}$ was used in[70]. This method is based on the observation, that increasing the number of parameter sets will lead to an increase in the uncertainty of the parameters, as each parameter set is estimated with less data. In this method, a large number of possible parameter sets $\mathbb{K}$ is chosen and the optimal parameters are estimated. Then for each parameter a confidence region is identified by means of bootstrapping over the data set used to estimated this parameter. If these confidence regions overlap for at least two sets of parameters, these sets are considered statistically indistinguishable and the total number of sets is reduced, until this does no longer occur. This method, while rather simple, has the disadvantage, that the persistency parameter C can not be estimated in the same way.

A different method, that can be used to estimate both parameters simultaneously is based on the principle of Occam's razor: The optimal model should describe the data sufficiently good while exhibiting the least possible number of free parameters. This principle has been given a more formal depiction by various information measures. One of the most prominent ones is the Akaike information criterion (AIC)[4].

The derivation of the AIC is based on the assumption, that the observed data were sampled from an unknown distribution f and aims at finding the best parameter Θ^* for the model $h(\cdot,\Theta)$. Best parameter is hereby to be understood as minimizing the Kullback-Leibler divergence[29] between the unknown distribution f and the model $h(\cdot,\Theta^*)$. It should be remarked, that the AIC cannot be used to compare models based on different data sets.

The AIC is formally given by

$$AIC(M) = -2\log(\mathcal{L}(M)) + 2|M|, \tag{4.1}$$

where $\mathcal{L}(M)$ is the Likelihood of the Model M with number of free parameters $|M|$. Obviously, out of a set of models with the same number of free parameters, based on this criterion, the model which maximizes the likelihood is taken. On the other hand, out of a set of models with the same likelihood of making a certain observation, the model with the smallest number of free parameters is chosen. If a set of models contains models with different number of free parameters and which result in different likelihoods of observation, the model is selected which balances these two values. A detailed analysis of the criterion can be found in[16], which also includes the further information on the correction for small samples, as developed in[46].

In the FEM-BV-framework the AIC can be directly applied in the case of binary affiliation functions $\gamma_i(t) \in \{0,1\}$ and probabilistic model distance functions without memory

or dependence, as for example

$$g(x,\Theta) = -\log P[x|\Theta], g(x,\Theta) = -\log f(x;\Theta), \tag{4.2}$$

where $f(\cdot;\Theta)$ is a probability density function with parameters Θ. In this setting, the log-likelihood of the observation is

$$\log \mathcal{L}(\Theta^{(1)},\dots,\Theta^{(\mathbb{K})}) = -\sum_{i=1}^{\mathbb{K}} \sum_{t\in\mathcal{T}} \gamma_i(t) g(x(t),\Theta^{(i)}), \tag{4.3}$$

which is the negative of the target functional that is minimized in the FEM-BV-framework. It should be stressed again, that the independence of the observation is the crucial assumption allowing to directly apply the result of[4].

In this case the result of the optimization can be directly used as an input in the Akaike information criterion. The other necessary input is the number of free parameters. This number is given by

$$|M(K,C)| \leq \mathbb{K}|\Theta| + \frac{C}{2} + \left(\frac{C}{2}+1\right), \tag{4.4}$$

where $|\Theta|$ is the number of free parameters in each $\Theta^{(i)}$. Additionally, the $\gamma_i(\cdot)$'s have to be characterized. Their complete description is given by the times of regime changes $0 < t_1 < \cdots < t_{\frac{C}{2}} < T$ and the index of the parameter set that is chosen for each time interval, starting from the choice made at time 0. Using $\frac{C}{2}$ instead of the number of transitions that are obtained during the optimization does only increase the Akaike Information Criterion, thereby favoring a solution with a C that is closer to the sharp upper bound of the BV-norm.

This works similarly in the case of memory effects, e.g. if $(x(t) \in \{1,\dots,s\})_{t=0,1,\dots,T}$ is modeled by a Markov chain with s states and time-invariant transition matrices $P^{(1)},\dots,P^{(\mathbb{K})}$, a possible model distance function is

$$g(x(t),x(t-1),P) = -\log P_{x(t-1)x(t)}. \tag{4.5}$$

Even though the single observations are not independent of each other, the log-likelihood of the whole observation is given by

$$\log \mathcal{L}(P^{(1)},\dots,P^{(\mathbb{K})}) = -\sum_{i=1}^{\mathbb{K}} \sum_{t=1}^{T} \gamma_i(t) g(x(t),x(t-1),P^{(i)}) + \log P[x(0)], \tag{4.6}$$

again the target functional of the FEM-BV-method does coincide with the log-likelihood except for the influence of the initial distribution. As the above model does not estimate the initial distribution (a task that would prove futile in the case of a single observation), this part of the sum can be ignored. In this case, the number of free parameters per parameter set can be estimated as follows: all entries in the transition matrix $P^{(i)}$ except for one per

row are free parameters, as each row has to sum to 1 and thus one element per row is completely described by the $s-1$ others. Thus the Akaike information criterion takes the form

$$AIC(\mathbb{K},C) = 2\sum_{i=1}^{\mathbb{K}}\sum_{t=1}^{T}\gamma_i(t)g(x(t),x(t-1),P^{(i)}) + 2(\mathbb{K}s(s-1)+C+1). \tag{4.7}$$

4.2 Model selection in the general setting: A modified information criterion

In the case were neither the model distance function nor the model function can be interpreted as a parametric probabilistic model, e.g. if the model is

$$x(t) = \Theta(t) + \varepsilon(t), E[\varepsilon(t)] = 0, \varepsilon(\cdot) \text{ independent}, \tag{4.8}$$

and the model distance is chosen as

$$g(x,\Theta) = \|x-\Theta\|_2^2, \tag{4.9}$$

a solution might still be found. In the above example, the unregularized solution can be obtained by the K-Means algorithm, or, if regularization is required, by the FEM-BV-K-Means-method. Yet, the knowledge about the noise process is not sufficient to allow for a probabilistic interpretation of the data or some function of it. To avoid this issue, the distribution of the scalar time series of model distances affiliated to each parameter set can be used to derive a probabilistic interpretation. The key idea of this approach is to find a distribution that fits this time series sufficiently well and use the likelihood based on that distribution to detect the optimal model in the FEM-BV-approach[66].

To this end, let $\text{supp}(\gamma_i) = \{t \in \mathcal{T} : \gamma_i(t) = 1\}$ be the support of $\gamma_i(\cdot)$. Please note that $\gamma_i(t)$ is still considered to be binary, thus the classical non-zero definition can be replaced by $\gamma_i(t) = 1$. Furthermore, let the time series of model distances for each parameter set on the according support be distributed according to a parametric probability density function $\rho_i(\cdot;\Lambda_i)$. If the time series of model distances are assumed to be independent, the likelihood function takes the form

$$\mathcal{L}(\mathbb{K},C) = \prod_t \left(\sum_{i=1}^{\mathbb{K}} \gamma_i(t)\rho_i\left(g(x(t),\Theta^{(i)});\Lambda_i\right)\right). \tag{4.10}$$

This likelihood can now be directly used in the original formulation of AIC to obtain

$$AIC(\mathbb{K},C) = -2\sum_t\sum_{i=1}^{\mathbb{K}}\gamma_i(t)\log\rho_i\left(g(x(t),\Theta^{(i)});\Lambda_i\right) + 2\left(\mathbb{K}|\Theta| + C + 1 + \sum_{i=1}^{\mathbb{K}}|\Lambda_i|\right). \tag{4.11}$$

This leaves the problem of finding the density functions $\rho_i(\cdot;\Lambda)$, that are supposed to represent distributions of the model distances for each individual parameter set. One option would be to simply assume a certain class of parametric density function, e.g. the class of Gaussians. Then the parameters Λ_i can be calculated for each parameter set using a maximum likelihood approach. However, assuming a fixed class of parametric density functions might be too restrictive and thereby can lead to the selection of wrong models.

Instead of relying on a priori assumptions, a better approach is to make use of statistical properties that can be obtained from the model distances themselves, e.g. the first k non-central moments. Given this information, the distribution function which contains the least additional information while providing the statistical properties of the sample can be chosen. According to[47,48,60], this distribution is characterized by maximizing the entropy as a measure of uncertainty. To this end, let $\eta_j, j = 1,\dots,k$ be the empirical estimates of the first k non-central moments, $\eta_0 = 1$ the normalization constraint and Ω be the support of $\rho(\cdot)$. The maximum entropy distribution is then given as the solution to the variational problem

$$\mathcal{H}(\rho) = -\int_\Omega \rho(x)\log(\rho(x))\,dx \to \max_\rho \tag{4.12}$$

subject to (4.13)

$$\eta_j = \int_\Omega x^j \rho(x)\,dx, \quad j = 0,\dots,k. \tag{4.14}$$

This problem can be solved by applying the Euler-Lagrange equation, known from the calculus of variations. This yields in

$$\rho^*(x) = \exp\left(\sum_{j=0}^{k} \lambda_j x^j - 1\right), \tag{4.15}$$

$$\eta_j = \int_\Omega \rho^*(x) x^j\,dx, \quad j = 0,\dots,k, \tag{4.16}$$

where the set of equalities (4.16) is enforcing the first k non-central moments of the distribution to fit the k first empirical non-central moments and the function $\rho^*(\cdot)$ to be normalized on Ω. The additional constraint $\rho(x) \geq 0, \forall x \in \Omega$ is automatically fulfilled and does not need to be considered in the optimization.

In the case $k = 1$ and $\Omega = [0,\infty)$ the Maximum entropy distribution is given by the exponential distribution, similar, in the case $k = 2$ and $\Omega = \mathbb{R}$, the $\rho^*(\cdot)$ is the density function of a normal distribution. The parameters can here be estimated in the usual way. Yet in more general cases, an analytic solution is often unknown. In these cases, the Lagrange multiplies have to be computed numerically. A detailed analysis of this problem is shown e.g. in[82].

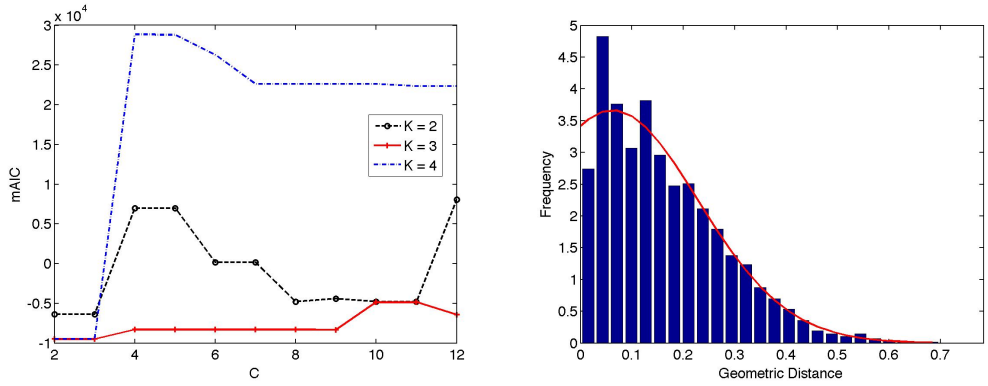

Figure 4.1. *Example 2.2.1[66] - Left: Values of the AIC for fixed $\mathbb{K}$ as a function of C using the Maximum Entropy approach. Right: Histogram of the geometrical distance of the data in the right geometrical cluster to the cluster center together with the graph of the fitted distribution based on the maximum entropy approach with $k = 3$ (red line).*

This approach allows to utilize the model distances and cluster affiliations obtained for each parameter set to estimate the AIC without imposing additional a priori assumptions on the distribution of the model distances, obtaining values for each combination $(\mathbb{K}, C)$. The number of non-central moments k to be considered can be chosen in line with the AIC approach by treating the non-central moments as k additional parameters and including the choice of k in the choice of the optimal model.

It should not be left unremarked, that this approach relies on the assumption that the model distances are independent and identical distributed for each parameter set. However, this has to be seen in the context, that standard methods in Bayesian time series analysis (e.g. Gaussian Mixture Models or Hidden Markov Models) assume, besides of the i.i.d. assumption, that the parametric multivariate distribution of the data is known, an assumption that is not necessary in the approach presented here.

Example 2.2.1 (cont.)

To demonstrate the usefulness of this approach, the example from Chapter 2 is revised. The analysis of the data set was repeated for different choices of $\mathbb{K}$ and C and the resulting model distances where fitted with the Maximum entropy approach introduced in this section. In the left panel of Figure 4.1 the values of the AIC for different pairs $(\mathbb{K}, C)$ are shown. The number of moments was chosen between 1 and 10 to minimize the AIC value. The right panel of Figure 4.1 shows a typical histogram of the geometrical distances to the center and the obtained maximum entropy distribution for $k = 3$ fitted non-central moments, the choice minimizing the AIC value. $\Diamond$

Example 4.2.1. [66] In example 2.2.1, the model distance function was based on the geometrical distance of the time series to a number of centroids, i.e. mean values. In this example a dataset with constant mean but time-varying variance is considered. Let $x(t) \in \mathbb{R}^2$ be time series of two dimensional data and $\{x(t)\}_{t\in\{1,\dots,10000\}}$ be a series of 10000 observations. The data are generated by

$$x(t) \sim \gamma_1(t)\mathcal{N}(0,\Sigma_1) + \gamma_2(t)\mathcal{N}(0,\Sigma_2). \tag{4.17}$$

The affiliation functions $\Gamma(t) = (\gamma_1(t), \gamma_2(t))$ with $\gamma_2(t) = 1 - \gamma_1(t)$ are chosen to take only values in $\{0,1\}$. The covariance matrices Σ_1 and Σ_2 are

$$\Sigma_1 = \begin{bmatrix} 4 & 0 \\ 0 & 0.25 \end{bmatrix}, \Sigma_2(\omega) = \begin{bmatrix} \cos\omega & \sin\omega \\ -\sin\omega & \cos\omega \end{bmatrix} \Sigma_1 \begin{bmatrix} \cos\omega & -\sin\omega \\ \sin\omega & \cos\omega \end{bmatrix}, \tag{4.18}$$

with $\omega = 15$ degrees. The generated data are shown in Figure 4.2. As can be seen, the two clouds formed by the data are overlapping for the majority of the points. Since the mean of both normal distributions is the origin of the ordinates, K-means is bound to fail to recover the affiliation functions used to generate the data. As an alternative, the identification by means of the eigenvectors of the covariance matrices will be used. This approach is known as Principle Component Analysis[50]. The basic idea is to use a lower dimensional dataset, that is projected on the dimension of the observed data by means of distinct projection matrices $Q_1, \dots, Q_{\mathbb{K}}$, where they are subject to noise

$$x(t) = Q_i \bar{x}(t) + \varepsilon(t). \tag{4.19}$$

The projection matrices have the additional property ${Q_i}^\dagger Q_i = I$. To identify the different projections, the model distance function

$$g(x(t), Q_i) = \|x(t) - Q_i {Q_i}^\dagger x(t)\|_2^2 \tag{4.20}$$

is used. The affiliation function used to generate the data can be seen in the right panel in Figure 4.3.

To give a reference to the result of the FEM-BV-PCA-approach, a Gaussian Mixture Model is also fitted to the data set. This is a classical and widely accepted method for unsupervised clustering, with the additional benefit of being designed for cases like this, where the data is the weighted sum of Gaussian processes. A two-dimensional stationary mixture model of two Gaussian distribution was trained (fitted) to the data using the Expectation-Maximization algorithm[22]. Even so this case seems to be a home game for the Gaussian Mixture model, the estimated covariance matrices are nowhere close to the covariance matrices used to generate the data. The reason for this can be seen when the associated most likely values of the hidden process are visualized (as done for the beginning of the time

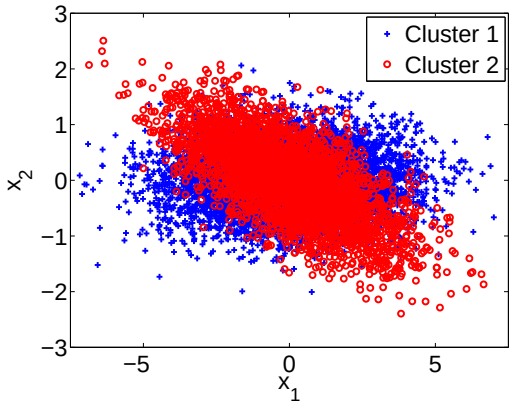

Figure 4.2. *Example 4.2.1[66] - Scatter-plot of a time series generated via the mixture model in (4.17) consisting of a time dependent convex combination of two (stationary) normal distributions with mean zero and covariance matrices given in (4.18) and a rotation angle* $\omega = 15$ *degrees.*

series in the left panel of Fig. 4.3). The close proximity of the two Gaussian distribution leads to a highly oscillatory result rather than a persistent path. In turn, most of the data points are not affiliated with the distribution they were originally sampled from, leading to incorrect estimation of the parameters. This behavior of the hidden process is caused by the strong stationarity assumption included in the Gaussian Mixture Model, i.e. the parameters of the distribution and the affiliation weights are assumed to be time-independent.

When using the FEM-BV-PCA-approach, the persistence of the affiliation functions is taken into account. Additionally the FEM-BV-PCA-algorithm was not granted the knowledge of the right parameters $\mathbb{K} = 2$ and $C = 8$, but instead the sets of possible parameters $\mathbb{K} \in \{1,2,3\}$ and $C \in \{2,4,6,8,10,14,20\}$. The according AIC values are shown in Figure 4.4, leading to a recovery of the original parameters. The resulting path is shown in the right panel of Figure 4.3, where it can be seen that the obtained affiliation matches with the one used to generate the data. ◊

The most restrictive assumption, that is still remaining, is the assumption that the affiliation functions $\gamma_i(\cdot)$ are taking only binary values at the observation times. Indeed, when utilizing the FEM-H1-framework the typical outcome is an affiliation function that does not satisfy this condition. This case needs some interpretation, as a non-binary affiliation function does not automatically indicate, that the optimal parameter can be interpreted as the convex combination of the $\mathbb{K}$ parameter sets. This is due to the interpolation on the level of model distances instead of parameters.

For this difference to be seen, the case of a probabilistic model with non-binary affiliation function is considered. Especially, let the model distance function $g(x(t),\Theta)$ be

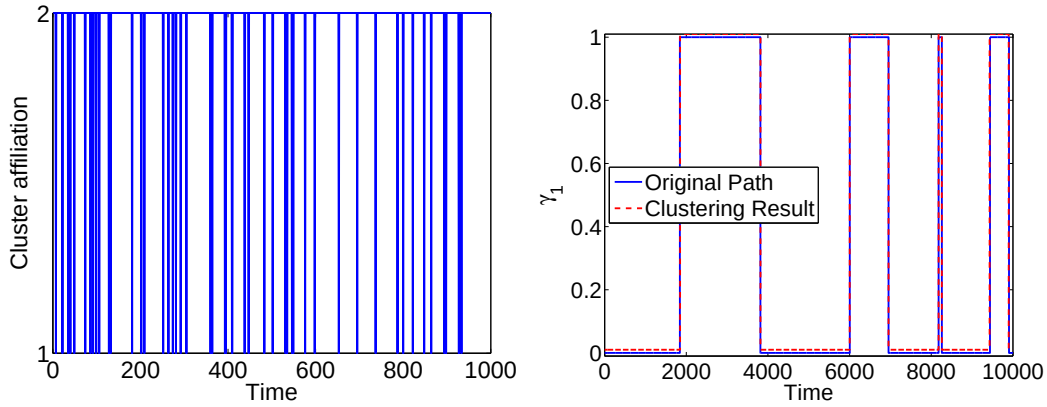

Figure 4.3. *Example 4.2.1[66]- Left: Part of the affiliation function for* $(1 \leq t \leq 1000)$ *obtained from fitting a two-dimensional stationary mixture model of two Gaussian distributions (via the GMM-method) on the data shown in Fig. 4.2. Right: The prescribed affiliation function* $\gamma_1(t)$ *(solid line) completely coincides with one obtained from the FEM-BV-PCA-analysis (red dashed line).*

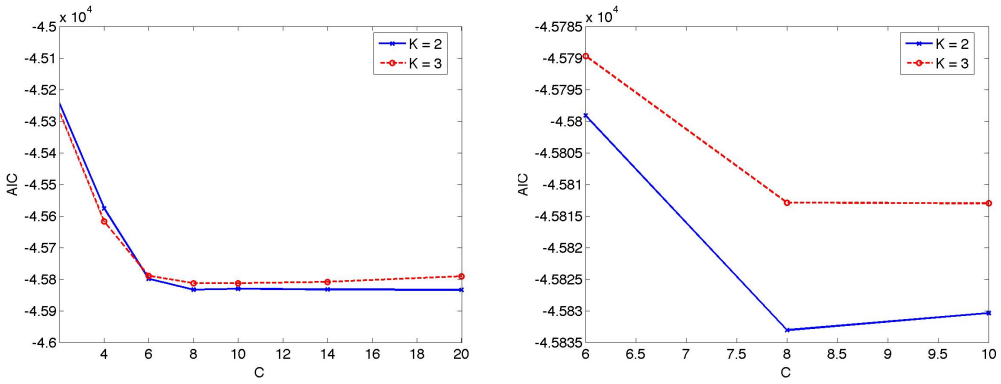

Figure 4.4. *Example 4.2.1[66] - Left: AIC for different number of parameter sets and number of possible transitions. The values for* $K = 1$ *are outside the plotted area above the other two plots. Right: Detailed view on the optimal point.*

chosen as the logarithmic density of the observation, i.e. let $f(x,\Theta)$ be the probability density function. Then the model distance function shall be chosen as

$$g(x,\Theta) = -\log(f(x;\Theta)). \tag{4.21}$$

Now let $\Theta^{(1)},\dots,\Theta^{(\mathbb{K})}$ be the parameter sets and $\gamma_1(\cdot),\dots,\gamma_{\mathbb{K}}(\cdot)$ be the affiliation functions, then the minimization of the (unregularized) model distance

$$\sum_t \sum_{i=1}^{\mathbb{K}} \gamma_i(t) g(x(t);\Theta^{(i)}) \to \min_{\gamma(\cdot),\Theta}, \tag{4.22}$$

can be interpreted as follows: Let $Z(\cdot) \in \{1,\dots,\mathbb{K}\}$ be a hidden, random process with independent realizations, governing the dynamics of the observed random variable $X(t) \sim f(\cdot,\Theta(t))$ where $\Theta(\cdot)$ is chosen out of the set $\{\Theta^{(1)},\dots,\Theta^{(\mathbb{K})}\}$ according to the hidden variable, thus

$$X(t) \sim f(\cdot,\Theta^{(Z(t))}), \forall t \in \mathcal{T}. \tag{4.23}$$

Further, let the density function $f(\cdot)$ be known. By inserting $g(x(t),\Theta^{(i)}) = -\log f(x(t),\Theta^{(i)})$ and $\gamma_i(t) = P[Z(t)=i]$, the model distance in Eq. (4.22) takes the form

$$\sum_t \sum_{i=1}^{\mathbb{K}} P[Z(t)=i] \log f(x(t),\Theta^{(i)}) = \sum_t E_{Z(t)}\left[\log f(x(t),\Theta^{(Z(t))})\right] \to \max. \tag{4.24}$$

Since the logarithm is a concave function, from Jensen's inequality follows

$$\sum_t E_{Z(t)}\left[\log f(x(t),\Theta^{(Z(t))})\right] \le \sum_t \log E_{Z(t)}\left[f(x(t),\Theta^{(Z(t))})\right], \tag{4.25}$$

thus the minimization of the model distance corresponds to a maximization of a lower bound of the log-likelihood of the observation. It might not be obvious why the lower bound is maximized instead of the log-likelihood directly as in e.g. the EM-algorithm. Yet the EM-algorithm relies on assumptions on the distribution of $Z(\cdot)$, e.g. the $Z(\cdot)$ are i.i.d. as in the case of a Gaussian Mixture Model or, in the case of a Hidden Markov Model, the $Z(\cdot)$ form a Markov chain. Assuming that each of the $Z(\cdot)$ has its own distribution leads to a series of Dirac distributions, i.e. the $Z(\cdot)$ is deterministic for each time step and the $\gamma_i(\cdot)$ are binary, in which case the lower bound is sharp. This problem is avoided in the FEM-framework by regularizing the hidden process, an assumption less restrictive than knowledge about the dynamics of the hidden process.

Now the same interpretation can be applied in the non-probabilistic case. Let $Z(\cdot) \in \{1,\dots\mathbb{K}\}$ be a hidden random process with independent realizations and $\gamma_i(t)$ the probabilities of $Z(t)$ being i. Moreover, let $\rho_i(\cdot)$ be the density function of $g(x(t),\Theta^{(i)})|_{Z(t)=i}$.

The density of the model distances $\rho(\cdot)$ can then be interpreted as a mixture model of the densities $\rho_i(\cdot)$

$$\rho(\cdot) = \sum_{i=1}^{\mathbb{K}} \gamma_i(\cdot)\rho_i(\cdot). \tag{4.26}$$

Assuming the hidden process is known $Z(t) = z(t)$ and the density function $\rho_i(\cdot)$ depends on the parameters $\Lambda^{(i)}$, then the likelihood function of the parameters $\Lambda = (\Lambda^{(1)}, \ldots, \Lambda^{(\mathbb{K})})$ can be written as

$$L(\Lambda; x(t), z(t)) = \gamma_{z(t)}(t)\rho_{z(t)}(g(x(t), \Theta^{(z(t))}), \Lambda^{(z(t))}) \tag{4.27}$$

or, for the whole observation $(x(t))_{t\in\mathcal{T}}$

$$L(\Lambda; x, z) = \prod_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \mathbb{I}_i(z(t))\gamma_i(t)\rho_i(g(x(t), \Theta^{(i)}), \Lambda^{(i)}), \tag{4.28}$$

inserting the exponential family, the following structure is obtained

$$L(\Lambda; x, z) = \exp\left(\sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \mathbb{I}_i(z(t)) \left(\log(\gamma_i(t)) + \sum_{j=0}^{k} \lambda_j^{(i)} (g(x(t), \Theta^{(i)}))^j - 1\right)\right). \tag{4.29}$$

In the next step, the expectation of the logarithm of $L(\Lambda; x, z)$ with respect to hidden variable is taken:

$$\begin{aligned} Q(\Lambda) &= E_z\left[\log L(\Lambda; x, z)\right] && (4.30)\\ &= \sum_{t\in\mathcal{T}} \sum_{i=1}^{\mathbb{K}} \gamma_i(t) \left(\log(\gamma_i(t)) + \sum_{j=0}^{k} \lambda_j^{(i)} (g(x(t), \Theta^{(i)}))^j - 1\right) && (4.31) \end{aligned}$$

This function can now be maximized with respect to the parameters Λ, with the constraint that each $\rho_i(\cdot, \Lambda^{(i)})$ is a probability density, thus

$$\int_{-\infty}^{\infty} \exp\left(\sum_{j=0}^{k} \lambda_j^{(i)} x^j - 1\right) dx = 1. \tag{4.32}$$

This condition can be included by explicitly solving the above equation for $\lambda_0^{(i)}$

$$\lambda_0^{(i)} = 1 - \log\left(\int_{-\infty}^{\infty} \exp\left(\sum_{j=1}^{k} \lambda_j^{(i)} x^j\right) dx\right), \tag{4.33}$$

and inserting this into $Q(\Lambda)$:

$$Q(\Lambda) = \sum_{t\in\mathcal{T}}\sum_{i=1}^{\mathbb{K}}\gamma_i(t)\left(\log(\gamma_i i(t)) + \sum_{j=1}^{k}\lambda_j^{(i)}(g(x(t),\Theta^{(i)}))^j - \log\left(\int_{-\infty}^{\infty}\exp\left(\sum_{j=1}^{k}\lambda_j^{(i)}x^j\right)dx\right)\right). \tag{4.34}$$

The maximization of $Q(\Lambda)$ yields

$$\frac{\partial Q(\Lambda)}{\partial\lambda_j^{(i)}} = \sum_{t\in\mathcal{T}}\gamma_i(t)\left((g(x(t),\Theta^{(i)}))^j - \left(\int_{-\infty}^{\infty}\exp\left(\sum_{j=1}^{k}\lambda_j^{(i)}x^j\right)dx\right)^{-1}\int_{-\infty}^{\infty}x^j\exp\left(\sum_{j=1}^{k}\lambda_j^{(i)}x^j\right)dx\right)$$

$$\frac{\sum_{t\in\mathcal{T}}\gamma_i(t)(g(x(t),\Theta^{(i)}))^j}{\sum_{t\in\mathcal{T}}\gamma_i(t)} = \int_{-\infty}^{\infty}x^j\exp\left(\sum_{j=0}^{k}\lambda_j^{(i)}x^j - 1\right)dx. \tag{4.35}$$

Thus as in the case of binary $\gamma_i(\cdot)$, the parameters $\lambda_j^{(i)}$ can be obtained by using the empirical non-central moments. In the case of non-binary $\gamma_i(\cdot)$, the estimators have to be weighted by the $\gamma_i(\cdot)$.

4.3 Model selection for the FEM-BV-Utility-method: A more pragmatic approach

In the FEM-BV-Utility method, of course the approach introduced in the previous section is still valid. And indeed, in the absence of transaction costs, an increase in $\mathbb{K}$ will (typically) lead to an improvement in the target functional, as the parameters can be chosen optimal for this observation. In this case, the information content of the model should be used to avoid over- or under-fitting.

With transaction costs present, the picture changes. An increase in $\mathbb{K}$ will either increase the number of transactions, as the new possibilities are used, or the additional portfolio remains unused, thus $\exists i \in \{1,\ldots,\mathbb{K}\} : \sum_{t\in\mathcal{T}}\gamma_i(t) = 0$. In the later case, it is obvious that the number of base portfolios can be reduced by at least one without changing the result, which, in the spirit of Occam's razor, should be done.

If the number of transactions changes, so does the total cost of all transactions. Thus, changing the maximum allowed cost level C will directly influence the number of transactions and thus the total cost.

In contrast to the standard FEM-method, where the regularization is added to provide mathematical and numerical benefits, in the FEM-BV-Utility-method, the regularization does directly result from the problem formulation. Moreover, the limited quantity can be directly interpreted and related to the optimized quantity. In fact, if the investor chooses a maximum level of transaction costs of C, the "true" initial wealth needs to be the amount used in the trading $V(0)$ and the amount reserved for the transaction costs C, resulting in the initial exposure $\bar{V}(0) = V(0) + C$. By replacing the initial wealth $V(0)$ by $\bar{V}(0)$ in the target functional, at least the result of the first trading period is influenced, allowing to choose the combination of $(\mathbb{K}, C)$ that results in optimizing the target function.

Example 4.3.1. Let $S(0) = [1,1]^\dagger$ and $S(1) = [1,Q]^\dagger$ with $Q > 1$ and an initial target investment of $V(0)$. Additionally, the transaction costs are value based with a 1% fee on the second component. The utility function is the logarithm of the relative profit, thus

$$u(V(0),V(1)) = \log\left(\frac{V(1)}{V(0)}\right) = \log V(1) - \log V(0). \tag{4.36}$$

Additionally, no short-selling is allowed. The $\beta(t)$ is in this case the constant $\beta = V(0)$, as $\beta(0) = \frac{V(0)}{S(0)^\dagger \xi(0)}$ and no further investments are considered. The matrix with $S(0)$ on the diagonal is the identity matrix. The optimal portfolio shall be $\xi = [\xi_1, \xi_2]^\dagger$. Thus the trading cost constraints become

$$C \geq \beta_{\max} \| c^\dagger \operatorname{diag}[S(0)] \xi \|_1 = V(0) \left| \frac{1}{100} \xi_2 \right| = \frac{1}{100} V(0) \xi_2. \tag{4.37}$$

This leads to the optimization problem

$$\max_{\xi} \log(\xi_1 + Q\xi_2) - \log(\xi_1 + \xi_2), \xi_1 + \xi_2 = 1, 0 \leq \xi_2 \leq 1, \xi_2 \leq 100 \frac{C}{V(0)}. \tag{4.38}$$

As $Q > 1$, the solution to this problem is the maximal permitted ξ_2 with

$$\xi_2^* = \min\left\{100\frac{C}{V(0)}, 1\right\}, \xi_1^* = 1 - \xi_2^*. \tag{4.39}$$

Replacing the initial value $V(0)$ (without changing the influence on later values, e.g. $V(1)$) in the utility function by the initial exposure $\bar{V}(0)$ results in the optimized utility function

$$u(\xi^*, C) = \begin{cases} \log(V(0)Q) - \log(V(0) + C), & C \geq \frac{V(0)}{100} \\ \log\left(V(0)\left(\left(1 - 100\frac{C}{V(0)}\right) + 100Q\frac{C}{V(0)}\right)\right) - \log(V(0) + C), & C < \frac{V(0)}{100} \end{cases} \tag{4.40}$$

$$= \begin{cases} \log\left(\frac{V(0)Q}{V(0)+C}\right), & C \geq \frac{V(0)}{100} \\ \log\left(\frac{V(0)+100C(Q-1)}{V(0)+C}\right), & C < \frac{V(0)}{100} \end{cases}. \tag{4.41}$$

Maximizing this function for $C \geq 0$ leads to the solution

$$C^* = \begin{cases} 0, & Q < \frac{101}{100} \\ \frac{V(0)}{100}, & Q \geq \frac{101}{100} \end{cases}. \tag{4.42}$$

Thus C is chosen just large enough to allow a whole transfer of the money to the profitable security, if and only if the profit is enough to cover the cost of the transaction. Else the C is chosen as 0 to preserve the initial wealth in the free asset. This is the result one would expect intuitively. ◊

In general no closed form solution for the optimal $\xi(\cdot)$ is known, thus an analytical solution to the general problem can not as easily be found as in the above example.

Yet the solution can be approximated by solving the problem for a set of combinations $(\mathbb{K}, C)$ and then either choosing the best combination out of the set or using a suitable method to propose a new set of candidates. E.g. one could employ a Simulated Annealing[30] approach by choosing an initial combination $(\mathbb{K}, C)^{(0)}$, calculating the utility of this combination (called $u^{(0)}$), selecting an initial acceptance parameter η and then iterating (with iterator i) over the following steps:

1. Create a new candidate $(\bar{\mathbb{K}}, \bar{C})$ by randomly modifying the last solution $(\mathbb{K}, C)^{(i-1)}$.
2. Solve the problem of finding $\xi^{(1)}, \ldots, \xi^{(\bar{\mathbb{K}})}$ and $\gamma(\cdot)$.
3. Estimate the quality of the combination by replacing the initial wealth in the utility function by the initial exposure $V(0) + \bar{C}$, called $\bar{u}^{(i)}$.
4. Accept the new solution if $\bar{u}^{(i)} > u^{(i-1)}$ or with a probability decreasing with $u^{(i-1)} - \bar{u}^{(i)}$ and η.

   If the solution is accepted, the results are taken for the next iteration, thus $(\mathbb{K}, C)^{(i)} = (\bar{\mathbb{K}}, \bar{C})$ and $u^{(i)} = \bar{u}^{(i)}$, else the solution is discarded and the result of the last iteration is reused, thus $(\mathbb{K}, C)^{(i)} = (\mathbb{K}, C)^{(i-1)}$ and $u^{(i)} = u^{(i-1)}$.
5. Increase the iterator i, increase η if necessary, terminate if no solution was accepted for a suitable number of steps, else return to step 1.

The random choice of $(\bar{\mathbb{K}}, \bar{C})$ should be suitable, as $\mathbb{K}$ should be an integer value and $\bar{C}$ should be positive. The algorithm is converging to a local maximum in the worst case. Convergence to the global maximum depends on the increase in η[8].

Example 4.3.2. In this example, the data from subsection 3.5.2 are reconsidered in the case of an initial investment of 3000 currency units and fixed transaction costs of 20 currency units per asset. Only the case $\mathbb{K} = 2$ is considered in this example. Please note, that in this

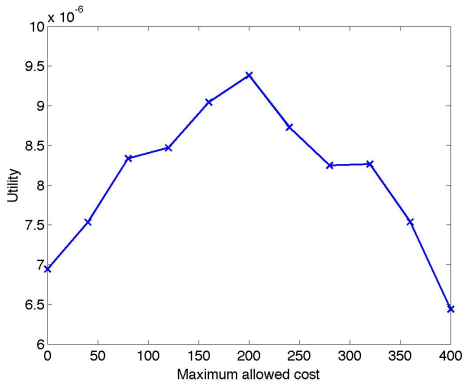

Figure 4.5. *Simple Example, Fixed Transaction Costs - Utility of the wealth process, including the maximum accumulated transaction costs as an immediate loss in the first period.*

case the parameter C should only take even integer values, as any transaction in one asset has to be followed by a transaction in the other asset, to avoid the need for the investor to add or remove money from the portfolio in whole. Additionally, if the maximal cost level C drops below 40, no transactions are allowed, as not enough money to cover the transactions is provided. This causes the solution to revert to the case of $\mathbb{K} = 1$, thus a simple Buy-and-Hold strategy.

The utility obtained by modifying the wealth process to include the maximal allowed transaction cost C as a loss in the first investment period can be seen in Figure 4.5. The observed behavior is as one would expect: while allowing for transaction at first improves the result, as the number of transactions (and thus the cost of these transactions) is further increased a maximum is reached, before the accumulating costs start to occupy to much of the performance and thus the obtained utility starts decreasing. $\Diamond$

5 Out-of-sample Analysis: Data Assimilation and prediction

In this chapter, two problems of practical interest are considered: (i) the problem of data assimilation, that is the inclusion of newly observed data into a model with parameters fitted to an earlier observation, and (ii) the problem of forecasting data for yet unobserved times.

Suppose that a solution of the inverse problem is known for some data set $(x(t))_{t\in\mathcal{T}}$ and model $f(\cdot,\Theta(t))$, thus an estimation $\Theta^*(t)$ is known for $t\in\mathcal{T}$. Now a number of new observations $(x(t))_{t\in\mathcal{T}^*}$ is made. The problem that will be closely examined in this chapter is how to expand the known solution $\Theta^*(t)|_{t\in\mathcal{T}}$ to the new observations, i.e. obtaining $\Theta^*(t)|_{t\in\mathcal{T}^*}$. To compare the different methods that are going to be introduced in this chapter, the following example will be used:

Example 5.0.3. Let $(x(t))_{t=1,\dots,10000}$ be a series of observation with $x(t)\in\{1,2\}$ simulated from a Markov chain with transition matrix $P(t)$:

$$P(t)=\begin{bmatrix}0.7 & 0.3\\ 0.3 & 0.7\end{bmatrix}, t\in\bigcup_{k=1}^{5}[2000(k-1)+1,1000(2k-1)], \tag{5.1}$$

$$P(t)=\begin{bmatrix}0.7 & 0.3\\ 0.7 & 0.3\end{bmatrix}, t\in\bigcup_{k=1}^{6}[1000(2k-1)+1,2000(k-1)], \tag{5.2}$$

The hidden process switches every 1000 time steps between a Markov chain with a 70% probability to be equal to the last observation, and independent Bernoulli distributed random variables. This is a perfect setting for the FEM-BV-Markov-method, as the construction of the data is exactly following the underlying model. To compare different assimilation methods, the first 5000 time steps are analyzed first, then the different assimilation method will be tested in the following scenarios:

1. Additional observation of 1000 time steps,
2. Additional observation of 2500 time steps,

New data points	abs. diff. in aff.	rel. diff. in aff.	cpu time (sec.)
1000	9	0.90%	257.47
2500	36	1.44%	271.76
5000	43	0.86%	311.52

Table 5.1. *Data Assimilation - Reanalysis: Quality of fit and computational cost to identify the model parameters and the affiliation function. If additional $\mathbb{K}$ and C are to be tested, the computational cost increases accordingly. Shown are the absolute difference in the affiliation functions in observations (abs.diff.in aff.), the relative difference (rel.diff.in aff.) and the cpu time in seconds.*

3. Additional observation of 5000 time steps.

The result in the "perfect" case would be to observe transitions every 1000 time steps. Additionally, for each of the transition points $5001,\ldots,9001$, the first observation that confirms the transition is reported. ◊

5.1 Complete Reanalysis

The first method is very straightforward: All information that are obtained in the earlier analysis are scrapped and the whole process of analyzing the data, thus finding the parameters $\mathbb{K}$ and C, finding the affiliation functions $\gamma(\cdot)$ and the optimal parametersets $\Theta_1,\ldots,\Theta_{\mathbb{K}}$, is repeated. A slight modification of this methods uses the previously obtained parametersets $\Theta_1,\ldots,\Theta_{\mathbb{K}}$ as initial values.

This method has the advantage, that the new information is treated equally to the previously known data. Changes in the affiliation function are not limited to the new sample, new parameter-sets are easily identified and the number of permitted transitions C can be adapted to the new situation.

The drawback of this approach is the computational cost, the whole analysis has to be repeated including the estimation of the parameters $\mathbb{K}$ and C. All information that was obtained previously is dismissed but, especially if the additional data set is small compared to the original data set, very likely recovered from the data. This process could be potentially avoided by recycling the known information.

Example 5.0.3 (cont.)

The complete reanalysis results in a very good result, although it comes at a high computational cost. The obtained affiliation functions for the second and third scenario are shown in Figure 5.1. In Table 5.1 the number of wrongly assigned data points and the computational cost are shown. The data points that are assigned to wrong parameter sets

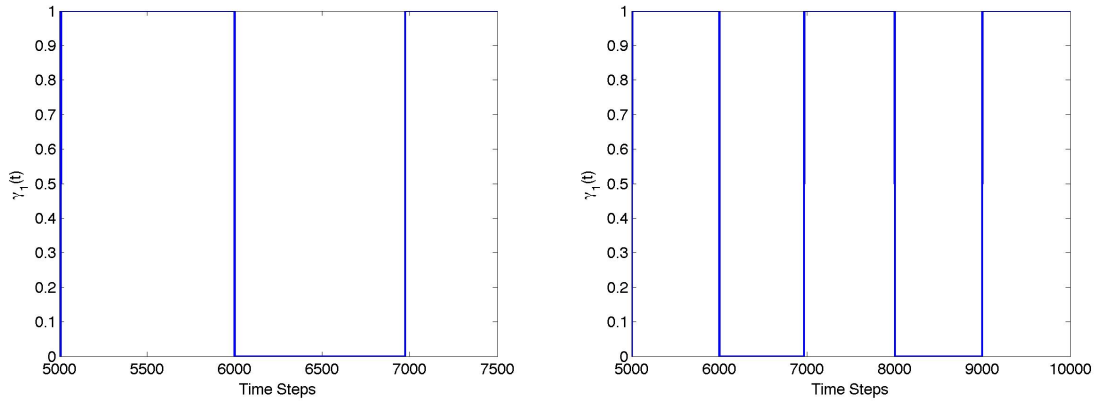

Figure 5.1. *Data Assimilation - Reanalysis: The obtained path for the 2500 time steps (left) and 5000 time steps (right) are shown. All parameters are chosen optimally: the model parameters and affiliation functions minimize the sum of the model distances, under regularity conditions, and the number of parameter sets $\mathbb{K}$ and the regularity parameter C minimize the AIC.*

Time steps before new transition	Observations needed
5000	31
6000	15
7000	14
8000	21
9000	09

Table 5.2. *Data Assimilation - Reanalysis: Number of observations before the transition is recognized*

are few, this number could be further reduced if the parameters chosen to generate the time series could be easier distinguished by the generated data. The number of observations needed to confirm a transition is shown in Table 5.2, on average 18 observations are needed. This is due to the fact, that the original parameters in both parameter set are chosen very close to each other. If they were less equal, the number of necessary observations would be smaller. ◊

New data points	abs. diff. in aff.	rel. diff. in aff.	cpu time (sec.)
1000	472	47.20%	0.01
2500	997	39.88%	0.16
5000	2065	41.30%	0.42

Table 5.3. *Data Assimilation - Minimization of the model distance: Quality of fit and computational cost to identify the affiliation function. While this method is much faster than the complete reanalysis, the results are not very reliable.*

5.2 Minimizing the model distance function

This method was first proposed in[41]. It is based on the assumption that the parameters obtained on the original observation are sufficiently optimal on the extended time series as well, thus the influence of the new data on the parameter estimation is sufficiently small. In this case, the affiliation function can be chosen to minimize the average clustering functional on the new observations

$$\sum_{t\in\mathcal{T}^*}\sum_{i=1}^{\mathbb{K}}\gamma_i(t)g(x(t),\Theta^{(i)}) \to \min_{\gamma(\cdot)}. \tag{5.3}$$

The solution to this problem can be found analytically, the optimal affiliation functions $\gamma_i^*(\cdot)$ are then

$$\gamma_i^*(t) = \begin{cases} 1, & i = \arg\min_j g(x(t),\Theta^{(j)}) \\ 0, & \text{else} \end{cases}, t \in \mathcal{T}^*. \tag{5.4}$$

Since an analytical solution is known, the computational cost of this scheme is reducing to the computational cost of evaluating the model distance function, which makes this scheme very easy to be applied.

On the downside, the updating procedure does not take into account the regularization that was part of the original problem. Additionally, the new information does have no influence on the affiliation functions for the original observation time or on the parameters. Moreover, a previously unobserved parameter set $\Theta^{(\mathbb{K}+1)}$ can not be discovered with this method.

Example 5.0.3 (cont.)

As can be seen from Table 5.3, choosing the affiliation functions to minimize the model distance is a very fast method to assimilate data, yet the fact that both parameters are close leads to a high number of wrongly assigned points. Another reason for this behavior is the fact that the regularization is ignored on the assimilated time frame. A plot of the affiliation function $\gamma_1(\cdot)$ is not shown for this example, as the affiliation function oscillates heavily

Time steps before new transition	Observations needed
5000	3
6000	1
7000	1
8000	1
9000	1

Table 5.4. *Data Assimilation - Minimization of the model distance: Number of observations before the transition is recognized*

between 0 and 1, rendering the plot useless. In contrast to the complete reanalysis, the minimization of the model distance does detect the changes (as shown in Table 5.4) very fast (on average 1.4 observations are needed). Yet, this is not due to an improved detection of change points, but caused by the lack of regularization. In fact, the number of (wrongly) detected change points in the whole test set is very large, leading to a large number of wrongly affiliated points. ◊

This method is suitable for cases where the data assigned to different parameter sets are sufficiently separated, thus in cases where the regularization of the affiliation functions is not necessary. E.g. in the case of a K-Means model with a clear geometric separation of the data.

5.3 Bayesian inference

To address the negative points in the previous scheme, a different approach was developed. This approach is based on the assumption, that the data $x(\cdot)$ are modeled by a probabilistic model and the distribution of the data is completely described by the parameter sets $\Theta^{(1)},\dots,\Theta^{(\mathbb{K})}$ and the previous observations. Furthermore a model for the hidden affiliation process $Z(t) \in \{1,\dots,\mathbb{K}\}$ is needed. This model should come with some memory, to support the regularization on this level. An example for a model fitting these criteria is the stationary Markov chain, thus if the distribution of $Z(t)$ (denoted by $f_Z(t) \in \mathbb{R}^{\mathbb{K}}$) is known, the distribution of the next time step is $P^\dagger f_Z(t)$, for some transition matrix P and equidistant time steps. The matrix P can be estimated from the result of the analysis of the original observation on the time frame $\mathcal{T}$.

First the case of a single additional observation is considered, as was introduced in[20], let $\mathcal{T} = \{1,\dots,T\}$ and $\mathcal{T}^* = \{T+1\}$. If the hidden process $Z(\cdot)$ was known on $\mathcal{T}$, this information and the actual observation $x(T+1)$ could be used to obtain the posterior dis-

tribution for $Z(T+1)$:

$$\mathbb{P}[Z(T+1)=i|(x(t))_{t\in\mathcal{T}\cup\mathcal{T}^*};Z(T)] \tag{5.5}$$
$$=\frac{\mathbb{P}[Z(T+1)=i;x(T+1)|(x(t))_{t\in\mathcal{T}}];Z(T)}{\mathbb{P}[x(T+1)|(x(t))_{t\in\mathcal{T}};Z(T)]} \tag{5.6}$$
$$=\frac{\mathbb{P}[x(T+1)|Z(T+1)=i;(x(t))_{t\in\mathcal{T}};Z(T)]\mathbb{P}[Z(T+1)=i|(x(t))_{t\in\mathcal{T}};Z(T)]}{\mathbb{P}[x(T+1)|(x(t))_{t\in\mathcal{T}};Z(T)]} \tag{5.7}$$
$$=\frac{\mathbb{P}[x(T+1)|Z(T+1)=i;(x(t))_{t\in\mathcal{T}};Z(T)]\mathbb{P}[Z(T+1)=i|(x(t))_{t\in\mathcal{T}};Z(T)]}{\sum_{k=1}^{\mathbb{K}}\mathbb{P}[x(T+1)|Z(T+1)=k;(x(t))_{t\in\mathcal{T}};Z(T)]\mathbb{P}[Z(T+1)=k|(x(t))_{t\in\mathcal{T}};Z(T)]}. \tag{5.8}$$

Moreover, if only a distribution of $Z(T)$ is known, then

$$\mathbb{P}[Z(T+1)=i|(x(t))_{t\in\mathcal{T}\cup\mathcal{T}^*}] \tag{5.9}$$
$$=\sum_{j=1}^{\mathbb{K}}\mathbb{P}[Z(T+1)=i|(x(t))_{t\in\mathcal{T}\cup\mathcal{T}^*};Z(T)=j]P[Z(T)=j|(x(t))_{t\in\mathcal{T}\cup\mathcal{T}^*}] \tag{5.10}$$
$$=\sum_{j=1}^{\mathbb{K}}\frac{\mathbb{P}[x(T+1)|Z(T+1)=i;(x(t))_{t\in\mathcal{T}};Z(T)=j]\mathbb{P}[Z(T+1)=i|(x(t))_{t\in\mathcal{T}};Z(T)=j]}{\sum_{k=1}^{\mathbb{K}}\mathbb{P}[x(T+1)|Z(T+1)=k;(x(t))_{t\in\mathcal{T}};Z(T)=j]\mathbb{P}[Z(T+1)=k|(x(t))_{t\in\mathcal{T}};Z(T)=j]} \tag{5.11}$$
$$P[Z(T)=j|(x(t))_{t\in\mathcal{T}\cup\mathcal{T}^*}]. \tag{5.12}$$

One possible approach to incorporate new observations into the existing solution would be to assume the observed affiliation functions to describe the hidden process, thus in the case of binary affiliation functions, if $\gamma_i(t)=1$ then $Z(t)=i$, in case of non-binary affiliation functions

$$\mathbb{P}[Z(t)=i]=\gamma_i(t). \tag{5.13}$$

This approximation can then be used to calculate the posterior distributions for each newly observed time step using only the information about the previous data. This procedure is known as the forward-algorithm.

For multiple time steps, the later information should be taken into account as well. One possibility is to use the Baum-Welch algorithm[71] in a modified form. The original algorithm simultaneously optimizes for the transition matrix of the hidden states, the observation probabilities of the observed data, conditioned on the hidden states and the initial distribution. In the setting that was introduced here, the transition matrix of the hidden states and the observation probabilities of the observations made are assumed to be obtained by the FEM-algorithm, the missing parameters are the hidden states at the new observations $\mathcal{T}^*$. To this end, the following algorithm is proposed:

1. Calculate the *forward probabilities*

$$f_i(T) = \gamma_i(T), \quad i = 1,\dots,\mathbb{K} \tag{5.14}$$

$$f_i(t) = \mathbb{P}[x(t)|Z(t) = i, x(t-1)] \sum_{j=1}^{\mathbb{K}} f_j(t-1)\mathbb{P}[Z(t+1) = i|Z(t) = j,], \tag{5.15}$$

$$i = 1,\dots,\mathbb{K}, t \in \mathcal{T}^*. \tag{5.16}$$

2. Calculate the *backward probabilities*

$$b_i(T^*) = 1, \quad i = 1,\dots,\mathbb{K} \tag{5.17}$$

$$b_i(t) = \sum_{j=1}^{\mathbb{K}} b_j(t+1)\mathbb{P}[x(t+1)|Z(t+1) = j, x(t)]\mathbb{P}[Z(t+1) = j|Z(t) = i] \tag{5.18}$$

$$i = 1,\dots,\mathbb{K}, t \in \mathcal{T}^* \setminus \{T^*\}. \tag{5.19}$$

where $T^* = \max \mathcal{T}^*$.

3. Estimate the path as

$$\hat{\gamma}_i(t) = \frac{f_i(t)b_i(t)}{\sum_{j=1}^{\mathbb{K}} f_j(t)b_j(t)}, \quad i = 1,\dots,\mathbb{K}, t \in \mathcal{T}^* \tag{5.20}$$

Please note that unlike in the classical Baum-Welch-algorithm[71] , no iterations are needed, as the estimation does not depend on the latest estimate.

This method does attempt to preserve the regularization by using the regularized solution to fit a Markov model. This model is then used to provide forward and backward information which allows for a close approximation of the hidden variable.

A drawback of this method is the fact, that only probabilities of the hidden process are obtained, no strict assignments are made. Another disadvantage is the fact, that the regularization is only approximated by the memory effects of the model of the hidden process. And lastly, the parameters have to be valid for the additional data as well, thus completely new parameter sets $\Theta^{(\mathbb{K}+1)}$ can not be discovered.

Example 5.0.3 (cont.)

Assimilating the data with the Bayesian inference is a much faster method than the complete reanalysis, while providing only slightly worse results, as can be seen in Table 5.5. The obtained affiliation functions are visualized in Figure 5.2. Another fact, that is visible from the Figure 5.2 is, that the assignment far from transition points is almost deterministic, in fact the smallest obtained value of $\gamma_1(\cdot)$ is $1.3 \cdot 10^{-5}$, the largest is $1 - 5.2 \cdot 10^{-7}$.

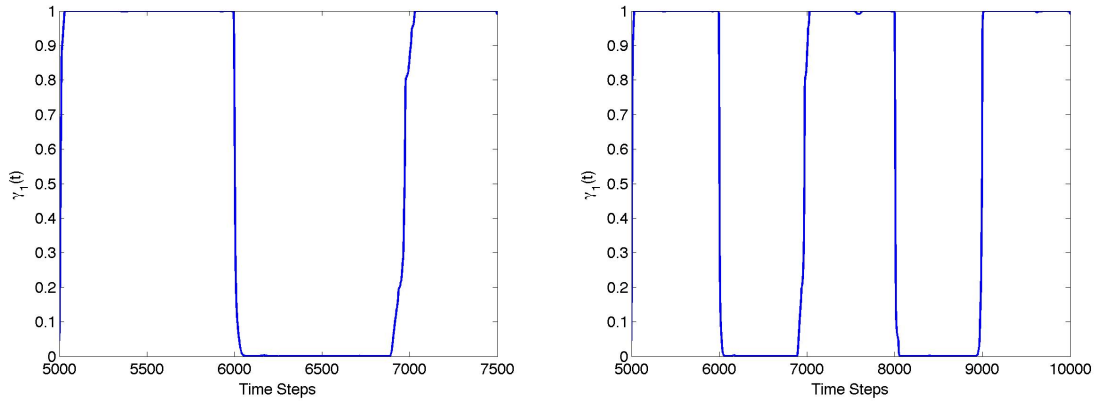

Figure 5.2. *Data Assimilation - Bayesian inference: The obtained path for the 2500 time steps (left) and 5000 time steps (right) are shown. The affiliation functions $\gamma_i(\cdot)$ are estimated using the forward-backward algorithm.*

New data points	abs. diff. in aff.	rel. diff. in aff.	cpu time (sec.)
1000	8.5643	0.86%	0.06
2500	52.7971	2.11%	0.17
5000	71.8021	1.44%	0.29

Table 5.5. *Data Assimilation - Bayesian inference: Quality of fit and computational cost to identify the model parameters and the affiliation function. This method is similarly fast as the minimization of the model distance while providing better results.*

In this case, a change point is considered detected, when the probability of the previous observed parameter set is no longer the largest, in this example if the probability drops below 50%. The number of observations needed to achieve that is shown in the Table 5.6. The average number of observations needed is 26.8, which is slightly higher than the in the case of a complete reanalysis. ◊

The need for a probabilistic model for the data and the hidden process limit the use of this approach. The fuzzy assignment is an additional drawback. Yet the small computational cost and the implicit regularization make this approach useful in the cases that it applies to.

Time steps before new transition	Observations needed
5000	31
6000	39
7000	17
8000	37
9000	10

Table 5.6. *Data Assimilation - Bayesian inference: Number of observations before the transition is recognized*

5.4 Preserving the regularization: Supervised Learning

The last method can be seen as a compromise between the complete reanalysis (as in Section 5.1) and the minimization of the model distance (Section 5.2). The idea is to use the FEM-BV-framework with fixed parameters $\Theta^{(1)},\dots,\Theta^{(\mathbb{K})}$, thus only the affiliation functions $\gamma_i(\cdot)$ are subject to optimization. Additionally, the range of the meta parameter C is strongly limited, leading to a smaller set of problems then the complete reanalysis, with the added benefit that all optimization problem are linear and can be solved rather quickly.

The mathematical formulation of this problem is:

$$\min_{\gamma_1(\cdot),\dots,\gamma_{\mathbb{K}}(\cdot)} \sum_{t\in\mathcal{T}\cup\mathcal{T}^*} g(x(t),(x(\tau))_{\tau\in[t-m,t)},\Theta^{(i)}) \tag{5.21}$$

$$\sum_{i=1}^{\mathbb{K}} \gamma_i(t) = 1, \quad t\in\mathcal{T}\cup\mathcal{T}^* \tag{5.22}$$

$$\gamma_i(t) \geq 0, \quad i=1,\dots,\mathbb{K}, t\in\mathcal{T}\cup\mathcal{T}^* \tag{5.23}$$

$$\sum_{i=1}^{\mathbb{K}} V_{T^*}^m \leq C^* \tag{5.24}$$

$$C^* \in \left[C, \left\lceil \frac{T^*}{T} C \right\rceil\right], T^* = \max \mathcal{T}^*, \tag{5.25}$$

where $\lceil x \rceil$ is the ceiling function, thus the smallest integer not less than x. The choice of the range of C^* here is motivated by the fact, that the number of transitions between two parameter sets is assumed to be growing linearly with the number of observations, yet this might not be suitable in general and a problem-driven estimation of the range is recommended.

Advantages of this approach are the decrease in computational cost, when compared to a complete reanalysis and the fact that the regularization is retained without the need of (explicit) Markovianity assumptions, in contrast to the (unregularized) minimization of

the model distance. Another fact to be pointed out is the ability to reassign data from the original set $\mathcal{T}$ to a different parameter set. This can be useful if, for example, the last few observations would have belonged to a different parameter set, yet their number did not warrant the introduction of an additional transition. New observations can confirm the existence of a transition, requiring a change in the affiliation function for the last points of $\mathcal{T}$, or contradict the additional transition, thereby supporting the decision that was based on the regularization.

A disadvantages of this approach is a somewhat higher computational cost, since the optimization of the affiliation functions has to be performed and can not be circumvented by an analytic solution as in the case of minimizing the model distance. And unlike the complete reanalysis, a change in the number of parameter sets $\mathbb{K}$ can not be detected.

Example 5.0.3 (cont.)

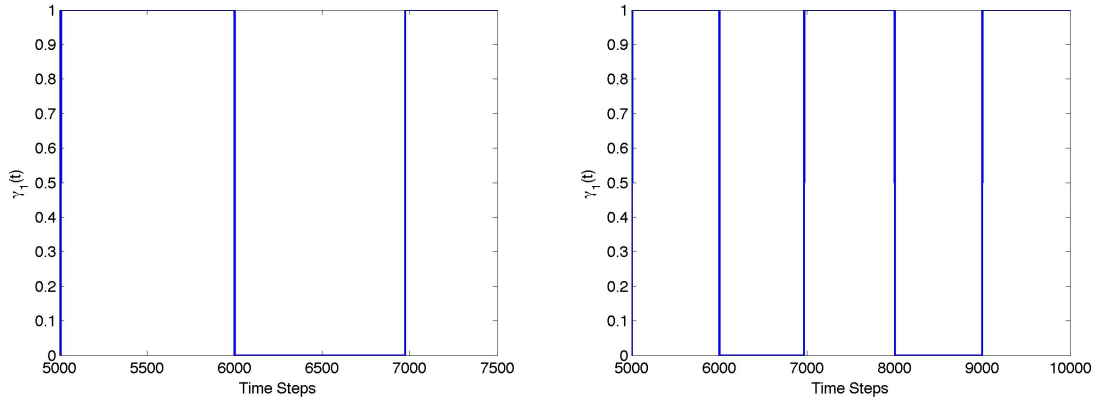

Figure 5.3. *Data Assimilation - Supervised Learning: The obtained path for the 2500 time steps (left) and 5000 time steps (right) are shown. The affiliation functions $\gamma_i(\cdot)$ are minimizing the weighted sum of the model distance functions, subject to regularity constraints.*

As can be seen from Table 5.7, the Supervised Learning approach is only slightly slower than the Bayesian Inference of the minimization of the model distances, yet it provides a result that is identical to the complete reanalysis. In fact, the path shown in Figure 5.3 are identical to the ones obtained by a complete reanalysis. The number of additional observations needed to detect a change point, as can be seen in Table 5.8 is identical to the case of a complete reanalysis. Summarizing, this supervised approach based on solving the linear optimization problem (5.21)-(5.25) is a good candidate to replace the complete reanalysis for short term assimilations. $\Diamond$

New data points	abs. diff. in aff.	rel. diff. in aff.	cpu time (sec.)
1000	9	0.90%	0.66
2500	36	1.44%	0.71
5000	43	0.86%	0.80

Table 5.7. *Data Assimilation - Supervised Learning: Quality of fit and computational cost to identify the model parameters and the affiliation function. This method is still much faster then the complete reanalysis, yet slower then the minimization of the model distance or the Bayesian inference. Yet the results are identical to the complete reanalysis.*

Time steps before new transition	Observations needed
5000	31
6000	15
7000	14
8000	21
9000	09

Table 5.8. *Data Assimilation - Supervised Learning: Number of observations before the transition is recognized*

The maintaining of the regularization and the smaller computational cost when compared to a complete reanalysis and the general applicability make this approach very useful. The lack of ability to identify additional parameter sets and to update the previously estimated parameters $\Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}$ justify a mixed strategy: For sufficiently short additional data, the supervised approach is used, while a complete reanalysis is performed when the newly accumulated data reach a certain threshold.

5.5 Forecasting

In this section, the problem of forecasting is discussed. The classical methodology to do this, under a given model, is to estimate the (stationary) parameters of the model using the observations and then use these to predict the next time step, e.g. as the expected value, the most likely observation or just a random realization. In the FEM-framework, this is not as simple, as also the parameters in the next time step are unknown. Thus the problem of forecasting in the FEM-framework can be divided into two subproblems: The prediction of the hidden process and the prediction of the data.

E.g. let $f(\Theta(t), \varepsilon)$ be a model with stochastic influence ε for the data and assume the model for the distribution of the hidden process $Z(\cdot)$ to be known. As prediction, the expected value of the model given the information available is then

$$\hat{x}(T+1) = E_{\varepsilon, Z(T+1)}\left[f(\Theta(T+1), \varepsilon) | (Z(t))_{t \in \mathcal{T}}\right] \tag{5.26}$$

$$= \sum_{i=1}^{\mathbb{K}} \mathbb{P}[Z(T+1) = i | (Z(t))_{t \in \mathcal{T}}] E_{\varepsilon}\left[f(\Theta^{(i)}, \varepsilon)\right]. \tag{5.27}$$

If the hidden states and the data cannot be separately estimated, e.g. if the distribution of the hidden state depends on previous observations and multiple time steps are to be forecasted, methods like Monte Carlo can be employed. Thereby a sample of independent realizations of the hidden and the observed process for all time steps subject to the forecasting is generated. The expectation of the time steps can then be approximated by the average over the sample.

Obtaining a data point from a model with known parameters and known hidden states is the well-studied forward-problem and will not be further discussed. Especially in the case of the FEM-BV-Utility-method and in contrast to standard methods of portfolio theory, forecasting the actual observed data (the prices of the assets) is not the aim of the method, and the lack of a forward model makes it a futile task. Instead the emphasis in this chapter is put on modeling and forecasting the hidden process, as this will directly allow to make investment decisions in the FEM-BV-Utility-context.

5.5.1 Markov approach

One of the most straightforward ideas in this context would be to model the hidden process as a Markov process. The advantage of this idea is that the regularization translates into small off-diagonal entries of the transition matrix. Under the assumption of stationarity of the distribution of the hidden process and equidistant time steps, a closed form for the estimator of the transition matrix is known to be

$$P_{ij} = \frac{N_{ij}}{\sum_{k=1}^{\mathbb{K}} N_{ik}}, \tag{5.28}$$

where N_{ij} is the number of observed transitions of the hidden process from state i to state j. The model can be adapted for cases with more memory. To this end, let $S = \{1, \ldots, \mathbb{K}\}$ be the state space of the hidden process $Z(t)$ which should be modeled with memory depth m. Assuming that $Z(\cdot)$ has indeed a memory depth of m, and the distribution of $Z(t)$ given the previous m observations does not depend on t, guarantees, that the process

$$\bar{Z}(t) = \begin{bmatrix} Z(t) \\ Z(t-1) \\ \vdots \\ Z(t-m+1) \end{bmatrix} \in S^m \tag{5.29}$$

has a memory depth of 1 and can be modeled as a Markov chain. The drawback of this approach is a curse of dimension. The state space of $\bar{Z}(\cdot)$ contains $\mathbb{K}^m$ elements, thus the (stationary) transition matrix will contain $\mathbb{K}^m(\mathbb{K}^m - 1)$ free parameters. Hence the number of free parameters for the model of the hidden state grows exponentially with the memory depth. The choice of the most appropriate memory depth m can be made e.g. using AIC.

5.5.2 Semi-Markov approach

As an alternative to the Markov approach, a less restrictive version can be chosen by means of the Semi-Markov approach. In this case, the hidden process is modeled by a Markov chain $X : \mathbb{N} \to S$ and a timing process $Y : \mathbb{N} \to \mathbb{R}$. The timing process is defining the time of the next transition, and the Markov chain $X(\cdot)$ the actual transition. E.g. if the timing process is exponentially distributed with mean a

$$Y(n) - Y(n-1) \sim \exp(a), \quad Y(0) = 0 \tag{5.30}$$

$$Z(t) = X(n), \quad n = \arg\max_i \{Y(i) \leq t\}, \tag{5.31}$$

the resulting process $Z(t)$ is a continuous time Markov process. A typical assumption made to get a unique definition is, that the Markov chain $X(\cdot)$ has a transition matrix with diagonal entries zero, thus at each transition time, the state has to change.

One advantage of this model is the fact, that the observation times, and thus the times at which the hidden process is estimated, do not have to be equidistant. Yet this additional flexibility comes at the price of having to model the transition times as well.

5.5.3 One step predictions

An important observation can be made when looking at a typical estimation of the hidden process resulting from the FEM-method: a transition at the end of the observational series is not identified if not enough data support the transition. This is a result of the regularization,

that smoothes out short periods in a state in favor of a more persistent hidden process. If the analysis would be repeated with a single additional data point, holding the affiliation functions constant on the previous observations, the affiliation functions for the additional observation would be identical to the previous step, if the additional observation was not lying "to far off". To find a more mathematical description of "to far off", this will be investigated in more detail here.

Let $\mathcal{L}((x(t))_{t\in\mathcal{T}}, \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}, C)$ be the likelihood of the parameters $\Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}$ and the affiliation functions $\gamma_i(\cdot)$ given the observation $(x(t))_{t\in\mathcal{T}}$ (as in Chapter 4). Additionally let the choice of the parameters $\mathbb{K}$ and C be optimal in the sense of AIC, thus

$$(\mathbb{K}, C) = \arg\min_{(\mathbb{K}^*, C^*)} \left(-2\log\mathcal{L}((x(t))_{t\in\mathcal{T}}, \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}^*, C^*) + 2(\mathbb{K}^*|\Theta| + C^* + 1)\right). \tag{5.32}$$

The likelihood of the observation $(x(t))_{t\in\mathcal{T}^*}$ where $T^* > T$, $\mathcal{T}^* = \mathcal{T} \cup \{T^*\}$ can then be written as

$$\begin{aligned}&\mathcal{L}((x(t))_{t\in\mathcal{T}^*}, \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}, C)\\ &\quad = \mathcal{L}(x(T^*), \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}, C|(x(t))_{t\in\mathcal{T}})\mathcal{L}((x(t))_{t\in\mathcal{T}}, \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}, C).\end{aligned} \tag{5.33}$$

This allows to write the AIC of the new time series as

$$\begin{aligned}&-2\log\mathcal{L}((x(t))_{t\in\mathcal{T}^*}, \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}^*, C^*) + 2(\mathbb{K}^*|\Theta| + C^* + 1)\\ &\qquad = -2\log\mathcal{L}(x(T^*), \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}^*, C^*|(x(t))_{t\in\mathcal{T}})\\ &\qquad -2\log\mathcal{L}((x(t))_{t\in\mathcal{T}}, \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}^*, C^*) + 2(\mathbb{K}^*|\Theta| + C^* + 1).\end{aligned} \tag{5.34}$$

Then the AIC of the parameters $(\mathbb{K}, C)$, chosen optimal as in Eq. (5.32) can be written as

$$AIC^*(\mathbb{K}, C) = -2\log\mathcal{L}(x(T^*), \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}, C|(x(t))_{t\in\mathcal{T}}) + AIC(\mathbb{K}, C), \tag{5.35}$$

while the AIC for an additional transition in the time interval (T, T^*) can be written as

$$\begin{aligned}&AIC^*(\mathbb{K}, C^+)\\ &\quad = -2\log\mathcal{L}(x(T^*), \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma^+, \mathbb{K}, C^+|(x(t))_{t\in\mathcal{T}}) + AIC(\mathbb{K}, C) + 2(C^+ - C).\end{aligned} \tag{5.36}$$

Thus for the additional transition to be accepted by the criterion, the value of the AIC in the setting of C^+ has to be smaller (or equal) to the value of the AIC without the additional transition, hence

$$\begin{aligned}0 \leq\ &AIC^*(\mathbb{K}, C) - AIC^*(\mathbb{K}, C^+)\\ &= -2\log\mathcal{L}(x(T^*), \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma, \mathbb{K}, C|(x(t))_{t\in\mathcal{T}})\\ &\quad + 2\log\mathcal{L}(x(T^*), \Theta^{(1)}, \ldots, \Theta^{(\mathbb{K})}, \Gamma^+, \mathbb{K}, C^+|(x(t))_{t\in\mathcal{T}}) - 2(C^+ - C),\end{aligned} \tag{5.37}$$

which can be directly reformulated into

$$\begin{aligned}\mathcal{L}(x(T^*),\Theta^{(1)},\ldots,\Theta^{(\mathbb{K})},\Gamma^+,\mathbb{K},C^+|(x(t))_{t\in\mathcal{T}}) \\ \geq \mathcal{L}(x(T^*),\Theta^{(1)},\ldots,\Theta^{(\mathbb{K})},\Gamma,\mathbb{K},C|(x(t))_{t\in\mathcal{T}})\exp(C^+-C)\end{aligned} \quad (5.38)$$

In the case of the standard, unweighted FEM-BV-method, the term $(C^+ - C)$ is equal to 2, thus the observation $x(T^*)$ has to be at least $\exp(2)$-times more likely under a different parameter set to justify an additional transition.

Using the argument, that a single observation is very unlikely to cause a transition if the remaining affiliation functions are held constant, a sensible one-step forecast for the affiliation function is to hold it constant.

6 Numerical results

In this chapter, the numerical feasibility of the introduced FEM-BV-Utility-method will be demonstrated with the help of two examples. First the approach will be shown in whole by analyzing a 31-dimensional financial time series of Swiss Market data. In this example, all steps will be taken in the order typically encountered by a practitioner. The second part of this chapter consists of the out-of-sample analysis of the data introduced in section 3.5, as the previous steps have already been taken in the chapters 3 and 4.

6.1 Complete Analysis of Swiss Market Data

In this example the approach is tested by building a portfolio in a real world setting. To this end, 30 different assets of the Swiss market are taken. The set of assets contains 14 stocks, 6 bonds (4 industry, 1 federal government, 1 cantonal owned company), 6 currencies, 3 commodities and the index *SMI*. An interest and risk free account is added, allowing to simulate the complete withdrawal of the investment.

All prices are adjusted according to the algorithm used by Yahoo Finance `https://help.yahoo.com/kb/finance/historical-prices-sln2311.html?impressions=true`, that was developed by the Center for Research in Security Prices. This allows to embed payments made during the observation time in the price process.

The cost structure in this example is a mixture of the three types of transaction cost introduced in Section 3.3. For the stocks and bonds, a fixed fee of 20 CHF is charged per transaction. For the currencies, the fee is 0.015% of the exchange money. For the index a fee of 1.50 CHF per contract is charged. Finally the commodities are traded with a fee of 0.2% of the value of the transaction. Changes on the cash account do not cause any fees.

The data are available for the 20-12-2010 to 01-12-2013, which includes the majority of the euro crisis and the resulting capital flight from the eurozone to the "safe haven" Switzerland, severely affecting the exchange rates and the Swiss economy. For the purpose of this example, the data is divided into a training and a test set. The data from 20-12-2010 to 01-05-2013 are used to train the parameters and the data of the remaining 7 month are then used to perform an out-of-sample analysis. This setting is in line with the situation an

investor is typically faced, a high dimensional data series is available for a relatively short period of time.

Two different situations were analyzed, the first consisted of a hypothetical investment of 100′000 CHF, the second of a much smaller investment of 5′000 CHF. The utility is chosen as the Markowitz-utility from Example 3.2.3. The risk-affinity parameter is set to $\alpha = 0.1$, meaning that an increase of the mean by 1 is only accepted if the variance is increase by less than 0.1.

6.1.1 The competing models

A well-known problem for the selection of portfolios in high-dimensional settings is the sensitivity of the sample covariance matrix to the data. E.g. in case of 10 years of monthly observations of 1000 stocks, the number of parameters in the covariance matrix (taking into account the symmetry) is 500′500, yet only 120′000 observations are made. As a result of this imbalance, the sample covariance matrix is typically singular, even if the "true" covariance matrix is known to be well-conditioned, rendering the result of the Markowitz-portfolio optimization problem exposed to the full estimation error. Indeed, recent studies[21] show, that in cases like this, the $\frac{1}{N}$-portfolio, where the portfolio is build by investing the same proportion of wealth into each asset, is surprisingly efficient. One way to reduce the exposure of the investment choice to the estimation error is the incorporation of additional information. In addition, the mean is not estimated, as a sufficiently long history to distinguish the mean from 0 is typically not given[64]. This has the additional benefit of avoiding extreme allocations due to estimation errors of the mean. The resulting strategy is based on minimizing the covariance of the investment without considering the returns.

To assess the usefulness of the FEM-BV-Utility-approach, it will be compared with two standard methods, that are known to do well in high-dimensional settings.

1. A $\frac{1}{n}$-portfolio, where the investor is just investing the same proportion of its wealth into each asset.

2. A minimum variance portfolio as a special case of the Markowitz problem with mean estimator $\hat{\mu} = 0$ and a well-conditioned estimator for the covariance matrix[56].

As neither model considers the return of the portfolio an unmanaged investment is chosen, i.e. the portfolio can be selected only at the initial time and stays constant for the whole time frame.

Other approaches that where considered where the 60-day Moving Average ansatz, as a special version of the Local Kernel Method with kernel function as in 3.72, and the model consisting of a Hidden Markov process and Gaussian outputs proposed by Hamilton[33]. The first approach did result in singular covariance matrices, due to the window size. A

mixed approach, using a well-conditioned estimator for the covariance matrix[56] resulted in covariance matrices dominated by the regularization. The Hidden Markov approach suffered from similar problems. The high dimensionality of the data caused a large amount of Gaussian distributions with almost singular covariance matrices to be identified. A regularization approach solved the singularity, but also caused the model with a single hidden state (thus a single Gaussian) to be considered the superior model by the AIC[4].

6.1.2 Estimation of the parameters $\mathbb{K}$ and C

The first difficulty is to find the optimal parameters $\mathbb{K}$ and C in the FEM-BV-Utility. This is done in line with chapter 4, especially using the approach of Section 4.3.

To this end, different combinations of $\mathbb{K}$ and C are used for the optimization and the utility of the result of the analysis of the test set for each pair $(\mathbb{K}, C)$ is compared.

As the maximum allowed transaction cost C increases, the constraints on the problem get relaxed and the utility increases before reaching a plateau. Taking into account the cost for the transactions as well, the situation changes. As can be seen in the Figure 6.1, increasing the maximum allowed transaction cost C will not necessarily increase the utility, but causes the plateau to be decreasing. In the case of an investment of $100'000$ CHF, this allows to estimate the global optimal parameters as $\mathbb{K} = 2$ and $C = 1800$. The optimal cost level for the case $\mathbb{K} = 3$ is $C = 1400$. The fact that the optimal cost level for the case of additional portfolios is smaller might seem counter intuitive at first sight, yet the higher number of base portfolios leads to a higher specialization. But the payout of this additional effort does not outweigh the inferred cost, and therefore the case $\mathbb{K} = 2$ is taken as the reference solution.

The lower panel of Figure 6.1 shows the empirical utility including the transaction costs for the case of an initial investment of $5'000$ CHF. As can be seen here, not only does the total transaction cost that is tolerated for the cases $\mathbb{K} = 2$ and $\mathbb{K} = 3$ decrease, but the solution without transactions $\mathbb{K} = 1$ is preferred over a solution with transactions.

6.1.3 In-sample analysis

As the parameters $\mathbb{K}$ and C are known, the in-sample performance can be measured.

As can be seen in the upper panel of Figure 6.2, in the case of an initial investment of $100'000$ CHF, the choice of the parameters $\mathbb{K} = 2$, together with the optimized total cost leads to a higher return than the according investments for $\mathbb{K} = 1$ or $\mathbb{K} = 3$. The numbers can be seen in Table 6.1. Comparing the results to the two standard approaches shows superior performance for the rejected cases $\mathbb{K} = 1$ and $\mathbb{K} = 3$ over the standard approaches as well. A comparison in terms of Sharpe-Ratio versus a risk-free investment into just the cash account indicates an advantage for the FEM-BV-Utility method with $\mathbb{K} = 2$ over its

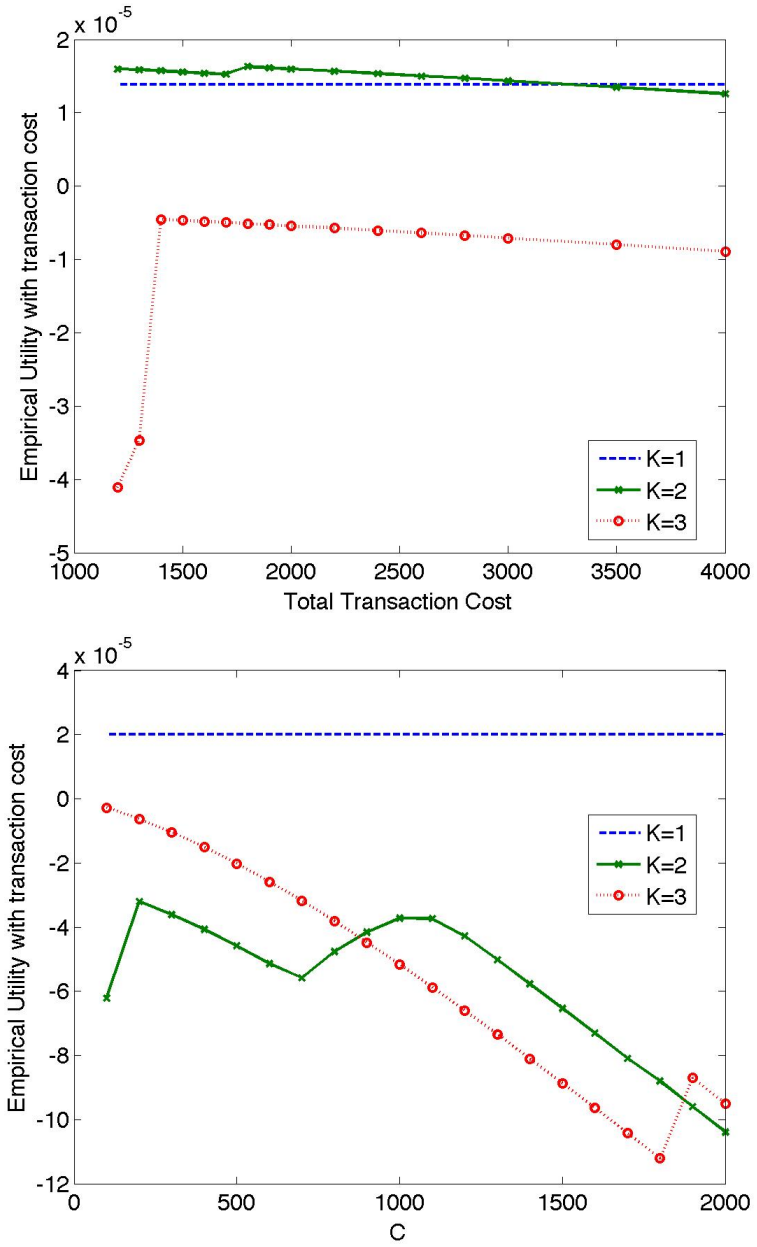

Figure 6.1. *High-dimensional example - Utility for different combinations of* $\mathbb{K}$ *and* C *for the case of a* 100′000 *CHF investment (upper) and* 5′000 *CHF investment (lower), considering the transaction costs in the utility as in Section 4.3.*

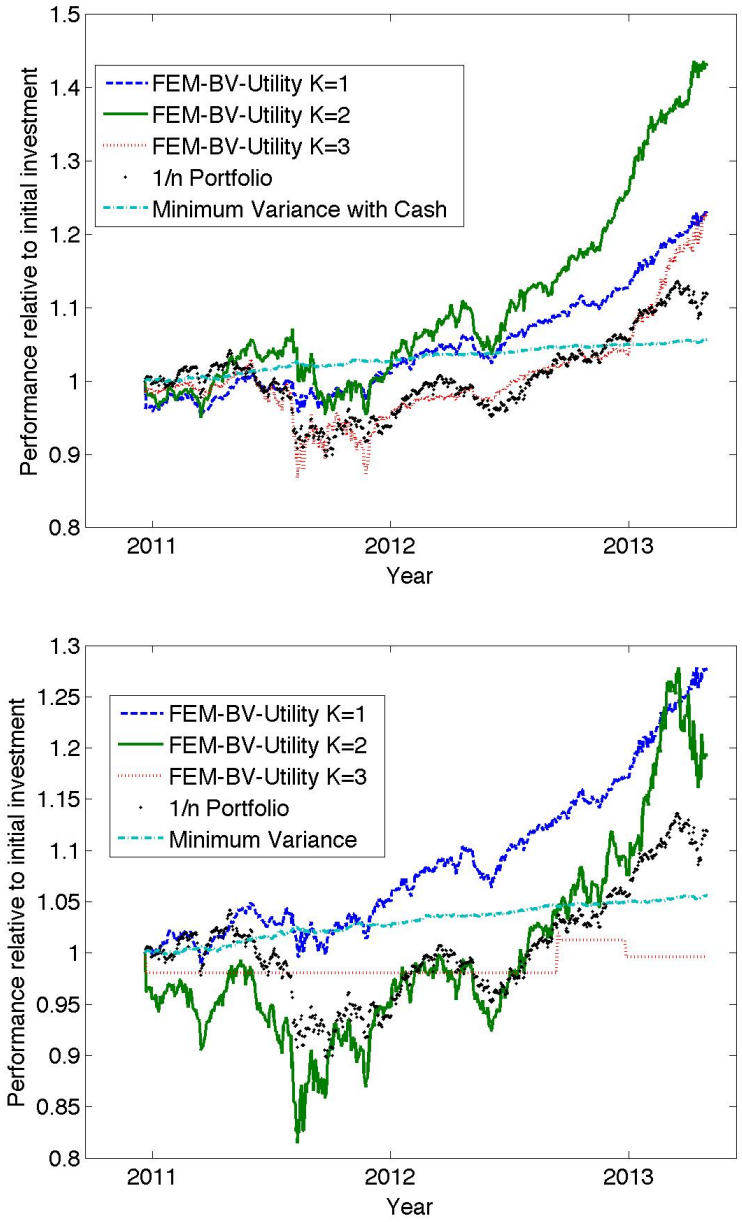

Figure 6.2. *High-dimensional example - In-sample performance of the FEM-BV-Utility approach and two standard methods for the case of an initial investment of* 100′000 *CHF (upper) and* 5′000 *CHF (lower).*

	$\mathbb{K}=1$	$\mathbb{K}=2$	$\mathbb{K}=3$	$\frac{1}{n}$	MinVar
Total Return	22.89%	43.19%	22.61%	11.87%	5.59
Avg. Daily Return	0.0244%	0.0430%	0.0251%	0.0141%	0.0063%
StD (Daily)	0.0033	0.0052	0.0054	0.0047	0.0006
Sharpe-Ratio (risk-free)	0.0739	0.0827	0.0465	0.0300	0.1050
Sharpe-Ratio (Inflation)	0.0585	0.0727	0.0362	0.0184	0.0148
Transactions	0	4	3	0	0

Table 6.1. *High-dimensional example - In-sample comparison of the FEM-BV-Utility approach and the two standard methods for an initial investment of* $100'000$ *CHF.*

	$\mathbb{K}=1$	$\mathbb{K}=2$	$\mathbb{K}=3$	$\frac{1}{n}$	MinVar
Total Return	27.70%	19.26%	−0.40%	11.87%	5.59%
Avg. Daily Return	0.0288%	0.0232%	−0.0003%	0.0141%	0.0063%
StD (Daily)	0.0030	0.0074	0.0016	0.0047	0.0006
Sharpe-Ratio (risk-free)	0.0971	0.0312	−0.0022	0.0300	0.1050
Sharpe-Ratio (Inflation)	0.0704	0.0206	−0.0524	0.0184	0.0148
Transactions	0	1	4	0	0

Table 6.2. *High-dimensional example - In-sample comparison of the FEM-BV-Utility approach and the two standard methods for an initial investment of* $5'000$ *CHF.*

competitors as well, only the minimum variance approach does better, which is mainly due to the low standard deviation of the returns.

Using the risk free rate as benchmark in the Sharpe-Ratio is questionable, as the investor should not be interested in maintaining the face value of the investment, but the value of the money in comparison to the prices of buyable goods.

If the Sharpe-Ratio is calculated versus the threshold inflation rate of the Swiss National Bank (2% per year), the picture changes, the FEM-BV-Utility method provides superior results in the Sharpe-Ratio to the standard methods even in the non-optimal cases $\mathbb{K} = 1$ and $\mathbb{K} = 3$.

In the case of an initial investment of $5'000$ CHF, the reference solution of the FEM-BV-Utility-method ($\mathbb{K} = 1$) outperforms the standard methods as well, as can be seen in the lower panel of Figure 6.2. The solution for the case $\mathbb{K} = 2$ is also resulting in a better performance than the standard methods, but this comes with a higher variance. The behavior of the solution for the case $\mathbb{K} = 3$ can be explained when looking at the obtained portfolios: this strategy consists of only one portfolio with investments in the risky assets, but two portfolios with risk-less investments only. Thus in this setting, money is invested only for a short period of time before it is withdrawn back to the cash-account. As can be seen in Table 6.2, not only is the performance of the FEM-BV-Utility-method in the cases $\mathbb{K} = 1$ and $\mathbb{K} = 2$ better than the performance of the standard-methods, but also the Sharpe-Ratio when calculated against the threshold inflation rate is better. This does not hold, when the Sharpe-Ratio is calculated against the risk-free asset, here the Minimum Variance approach takes a narrow lead.

6.1.4 Out-of-sample analysis

The out-of-sample analysis is done by performing successive 1-day predictions and assimilation of the data for the 7 months (between May 2013 and December 2013) not included in the analysis. The additional time steps might justify additional transactions. To cover this case, the maximum allowed cost is increased linearly with the new data.

It is considered to be an open secret[21] that in high-dimensional settings the $\frac{1}{n}$-Portfolio is hard to beat. In fact a wide range of methods, that do great in the in-sample case, fall of the cliff in out-of-sample comparisons.

The out-of-sample performance for an initial investment of $100'000$ CHF can be seen in the upper panel Figure 6.3, showing a slight advantage of the FEM-BV-Utility method. In fact, as can be seen in Table 6.3, the case $\mathbb{K} = 3$ finishes with a slightly higher total return, when compared to the case $\mathbb{K} = 2$. For both parameters, the performance is better then the performance of the standard methods, while the case $\mathbb{K} = 1$ finishes in the middle of the field. When using the Sharpe-Ratio against the risk-free cash account as the criterion for the comparison, the $\frac{1}{n}$-portfolio is leading the field, before the FEM-BV-Utility method

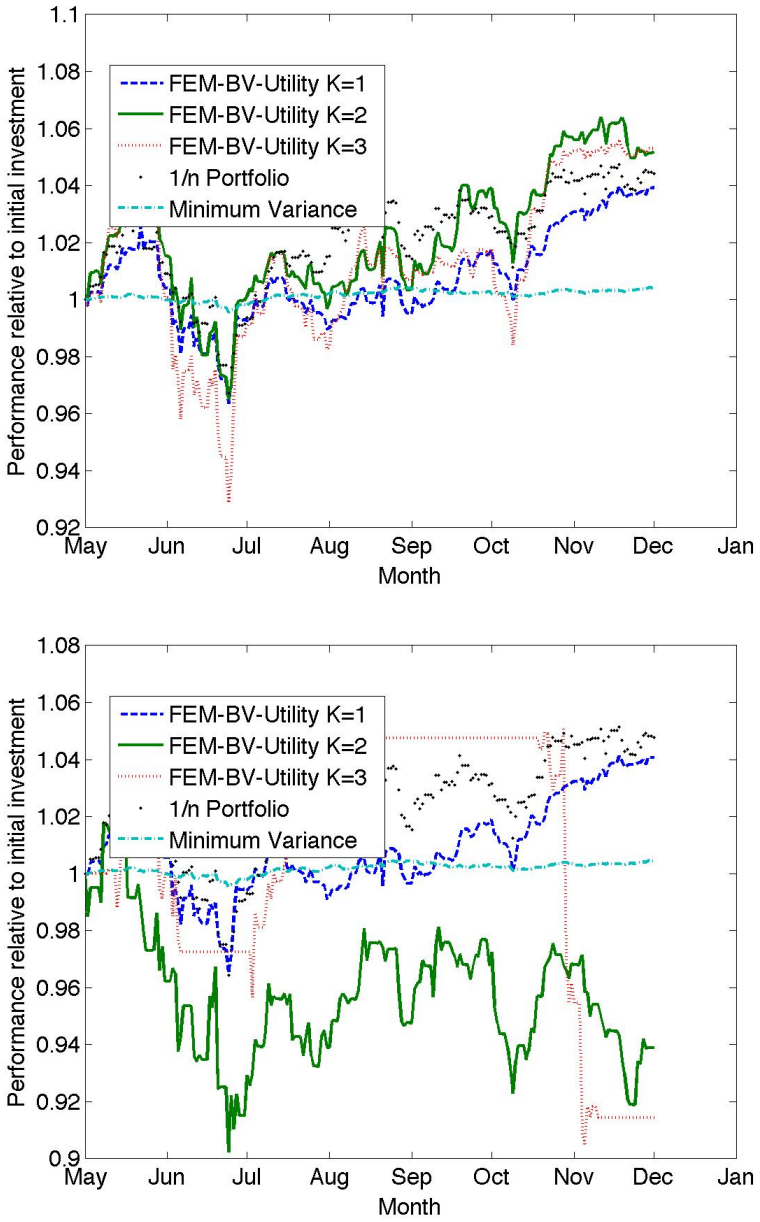

Figure 6.3. *High-dimensional example - Out-of-sample performance of the FEM-BV-Utility approach and two standard methods for an initial investment of* 100′000 *CHF (upper) and* 5′000 *CHF (lower).*

	$\mathbb{K}=1$	$\mathbb{K}=2$	$\mathbb{K}=3$	$\frac{1}{n}$	MinVar
Total Return	3.92%	5.17%	5.30%	4.42%	0.40%
Avg. Daily Return	0.0186%	0.0246%	0.0262%	0.0209%	0.0019%
StD (Daily)	0.0035	0.0047	0.0064	0.0036	0.0004
Sharpe-Ratio (risk-free)	0.0534	0.0525	0.0409	0.0572	0.0426
Sharpe-Ratio (Inflation)	0.0378	0.0409	0.0324	0.0424	−0.0799
Transactions	0	2	3	0	0

Table 6.3. *High-dimensional example - Out-of-sample comparison of the FEM-BV-Utility approach and the two standard methods for an initial investment of* $100'000$ *CHF.*

	$\mathbb{K}=1$	$\mathbb{K}=2$	$\mathbb{K}=3$	$\frac{1}{n}$	MinVar
Total Return	4.07%	−6.11%	−8.56%	4.79%	0.44%
Avg. Daily Return	0.0193%	−0.0259%	−0.0383%	0.0226%	0.0020%
StD (Daily)	0.0035	0.0084	0.0083	0.0039	0.0005
Sharpe-Ratio (risk-free)	0.0556	−0.0309	−0.0464	0.0573	0.0426
Sharpe-Ratio (Inflation)	0.0327	−0.0404	−0.0560	0.0373	−0.1226
Transactions	0	0	0	0	0

Table 6.4. *High-dimensional example - Out-of-sample comparison of the FEM-BV-Utility approach and the two standard methods for an investment of* $5'000$ *CHF.*

with $\mathbb{K} = 1$ and $\mathbb{K} = 2$, that are in turn superior to the minimum variance approach. If the benchmark is changed to the target inflation rate of 2% again, the advantage of the $\frac{1}{n}$-portfolio is reduced to a narrow lead, while the Sharpe-Ratio of the Minimum Variance portfolio becomes negative, indicating under performance.

For the case of an initial investment of 5′000 CHF, the reference solution of the FEM-BV-Utility-method ($\mathbb{K} = 1$) is outperformed by the $\frac{1}{n}$-portfolio, as can be seen in the lower panel of Figure 6.3. When comparing the numbers in Table 6.4, it can be seen, that the $\frac{1}{n}$-portfolio does also show a slight advantage in terms of the Sharpe-Ratio, in the case of the risk-free reference return as well as versus the inflation rate.

6.2 Out-of-sample analysis of German Market Data

As already mentioned at the beginning of the chapter, in this example the out-of-sample analysis of the data used in Section 3.5 and example 4.3.2 is performed.

Please be reminded, that 495 datapoints of the dataset have not been included in the in-sample-analysis (Section 3.5). These additional observations are compared assuming fixed transaction costs as in Subsection 3.5.2. The same set of algorithms is used as competitors as in the in-sample-analysis.

For the FEM-BV-Utility method, the optimal cost level C has to be chosen according to the ratio of initial wealth to transaction costs (per transaction), this is done in-sample. For the out-of-sample analysis, the ratio $\frac{C}{t}$ is kept constant (where t is the current length of the known data-series). Practically, this means that the investor, having decided how much transaction costs to accept for a given amount of time, scales this linearly with longer investment time frames.

The risk-aversion parameter α was chosen to maximize the Sharpe-Ratio in the in-sample example under the consideration of the transaction costs, as was the maximum allowed accumulated transaction costs C.

As the Sharpe-Ratio on the training set for the FEM-BV-Utility method with parameter $K = 1$ is superior to the Sharpe-Ratio for the parameter $K = 2$ for the case of high transaction costs in comparison to the initial wealth (see Figure 6.4 left), it is advisable to apply the investment strategy obtained with the former without further transactions (Buy&Hold). Indeed, this strategy is outperforming the case with transactions and the standard methods in this setting. And it remains to deliver a better or equal performance than the standard methods (except for FlexM in one case) with lower transaction costs (see Figure 6.4 right), even so it is outperformed by the strategy obtained by $K = 2$, as the better performance on the training set indicated (see Figure 6.4 left).

For the following calculations, a ratio of initial wealth to cost per transaction of 250 was used. This corresponds e.g. to an initial amount of wealth of 5000 euro and a cost per

transaction of 20 euro. The simple $\frac{1}{n}$-Portfolio was not updated during the process, meaning half the initial wealth was invested in the bond, the other half in the index and then the portfolio was not touched anymore. The result of the five algorithms can be seen in Figure 6.5. The FEM-BV-Utility-method is not only performing better also out-of-sample, the Sharpe-Ratio of the result is also superior to the Sharpe-Ratios of the other six algorithms which can be seen in Table 6.5. This indicates that the better return of the FEM-BV-Utility-method is also better balanced with the risk when compared to the other algorithms. The fact that the advantage in the Out-of-sample case is not as large as in the In-sample scenario can be explained by the problem of detecting changes in $\Gamma(\cdot)$, as a change point will need additional confirmation by more observations before it results in a transaction. Additionally, the optimal portfolios are estimated based on the training set, without taking into consideration the new information obtained during the out-of-sample period.

Method	FEM-BV-Utility	Markowitz	GMM	FlexM	$\frac{1}{n}$-Portfolio	Hamilton	MA
R	0.0665	0.0560	0.0554	0.0705	0.0407	0.0332	-0.0029
ATO	8.8006	0.5192	0.5192	1.8850	0.5192	0.5192	2.5960
#Tran.	8	0	0	4	0	0	2

Table 6.5. *Sharpe-Ratios of the obtained out-of-sample returns. The result of the FEM-BV-Utility method is still better than the results of the standard approaches, also the advantage is less distinct. The reason for this behavior is very likely the uncertainty of knowing* Γ *out-of-sample, as optimal change points might not be detected until confirmation by more observations is given. As the optimal portfolios are estimated based on the training set, new information obtained during the out-of-sample period are not taken into consideration.The standard methods (except for the FlexM and the Moving Average approach) do not readjust the holding during the out-of-sample period, while the FEM-BV-Utility method by doing so achieves a better performance.*

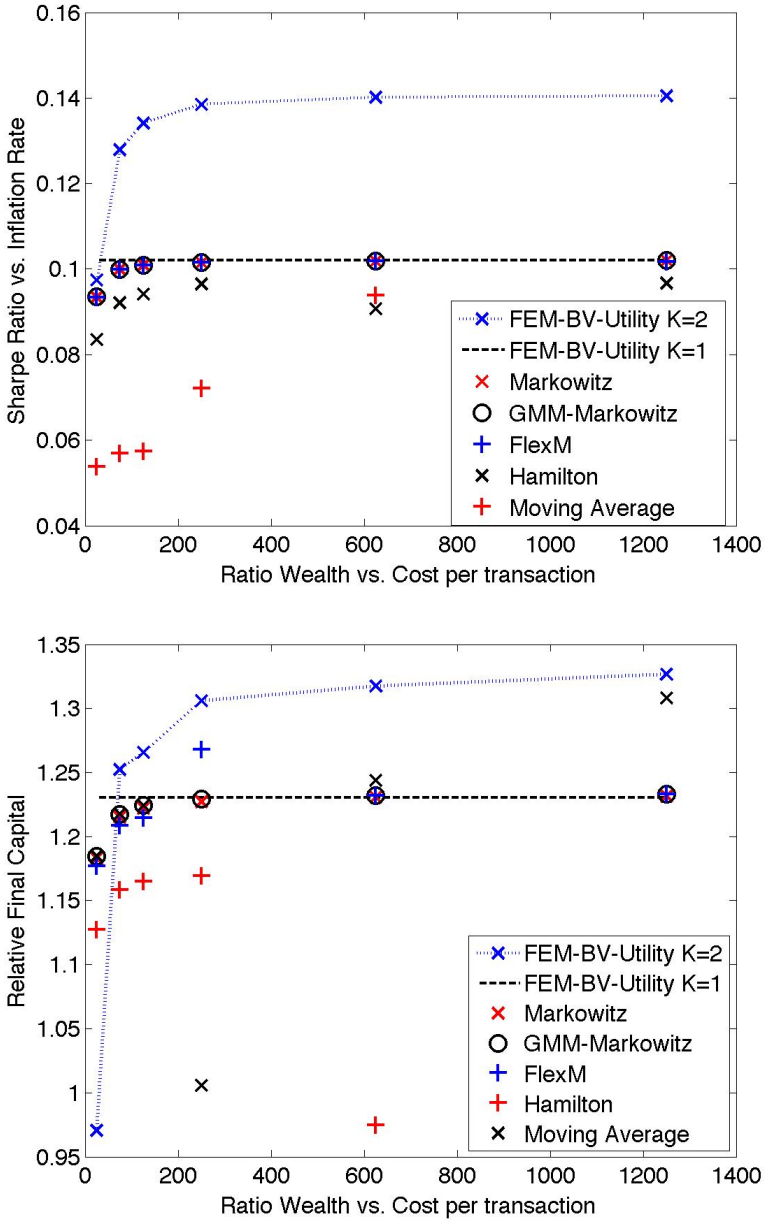

Figure 6.4. *Left: Sharpe-Ratio vs. the ECB target inflation rate for different ratios of starting wealth vs. (fixed) cost per transaction. The risk aversion parameter α is chosen to maximize this measure. For high transaction costs (covering* 25 *transactions, left-most point) the parameter K in the FEM-BV-Utility method should be set to* 1*, thus the algorithm reverts to buying a portfolio once and holding it for the rest of the time-frame. Right: Relative final capital of a hypothetical investment during the (non-trained) last* 495 *observations of the dataset, using the α obtained by maximizing the Sharpe-Ratio at the training set.*

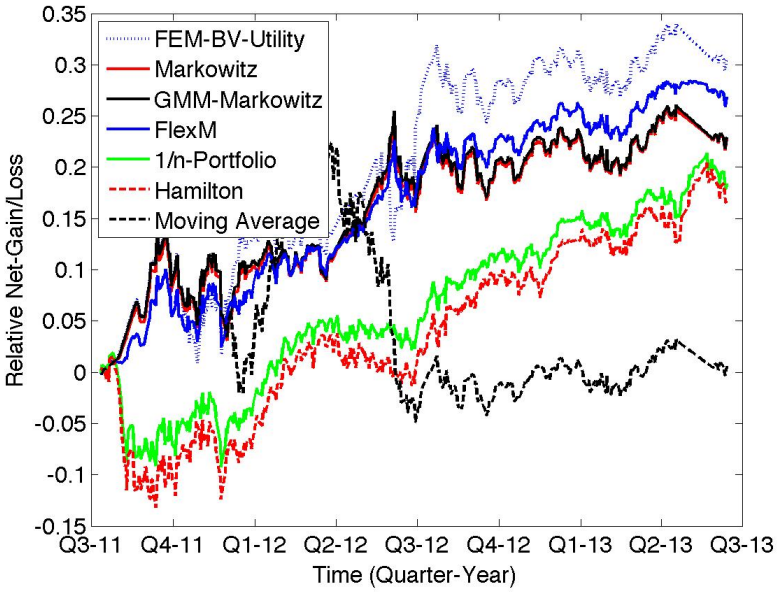

Figure 6.5. *Out-of-sample performance of different portfolio optimization strategies. Plotted are the relative Net-Gains, hence the gain/loss at a certain time divided by the initial wealth (initial wealth to cost per transaction ratio is* 250*). Each method's parameters were optimized over the training set (not included in the plot).*

7 Conclusion

A model-free approach to data-driven portfolio allocation was introduced that is based on local stationarity assumptions of the investment strategy. These assumptions are supported by the incorporation of transaction costs in the obtained numerical problem. Different ways of handling the identification of parameters in the approach are introduced for the in-sample-case and for the event of new observations. The numerical feasibility of the proposed algorithms was demonstrated for a low-dimensional example in Section 3.5 and for a high-dimensional example in Chapter 6.

The problem of portfolio optimization is a wide field of research. Ever since the famous introduction of Markowitz's portfolio optimization approach[61], approaches have been explored to widen the possible preferences of an investor, e.g. by introducing utility functions[24,79], or to relax the assumptions of perfect knowledge of model parameters. Today it is a well known fact in the literature that diversification can significantly reduce risk. Known approaches are based on an underlying model, where the estimation of the parameters of such a model infers risk from a different angle, as the identified parameters are perturbed. While different concepts exist, that try to handle this additional source of risk, e.g. by resampling the data to avoid an over-exposure to the estimation error[67], the problem still relies heavily on the validity of the used model. Another source of error is the assumption of stationary parameters, either of the prices, as e.g. in the Black-Scholes-Model[12], or of underlying processes driving the prices, e.g. in the GARCH-model[13]. While these assumptions can be considered valid for short time-frames, long-term observations[73] suggest that parameters can be subject to sudden, significant changes, motivating a relaxation of the stationarity assumption towards local stationarity.

Different approaches exist to analyse series of non-stationary observations. Some are purely local, e.g. the local kernel method[27,57], others rely on assumed probabilistic properties of a hidden process, e.g. Hidden Markov Models[72], and the observed process, as e.g. Gaussian Mixture Models[28,62]. Others again avoid probabilistic assumptions by considering geometrical properties, e.g. $\mathbb{K}$-means[52,59], or rely on dynamical systems to identify non-stationary features, e.g. artificial neural networks[10]. Recently a method was developed that does not rely on a priori probabilistic assumptions on the hidden process or station-

arity of parameters. The FEM-framework has been successfully used in a wide range of problems[36–42,66,70], where standard methods fail or provide inaccurate results.

In this work, a data-driven approach to the portfolio optimization problem was developed, that does not rely on the validity of a specific model for the observation data, nor on the knowledge of some (instantaneous) parameters. Instead, based on the effect of transaction costs, local stationarity of the investment strategy is assumed. The new approach allows to consider different preferences of the investor to be used as optimization target by making use of utility functions, while avoiding assumptions on distributions of price processes. Instead the identification of investment periods can be considered as the identification of a hidden process with unknown dynamics. The lack of a model for the price process render standard approaches for analyzing non-stationary time series useless, yet the structure of the problem proves to be similar to the FEM-BV-method. This allows to make use of a range of algorithms developed in this environment in the context of the portfolio optimization problem.

The resulting method identifies the (in hindsight) optimal investment strategy including transaction costs for an observed time series (in-sample). Different ways were shown to incorporate new observations into a known solution and to expand the strategy out-of-sample.

The empirical analysis showed that discarding the assumption of an a priori known model improved the performance of the obtained investment strategies in-sample as well as out-of-sample, when compared to standard approaches.

While the empirical results are just snapshots and are not sufficient to consider the proposed methodology superior to model-based approaches, they warrant a certain skepticism when dealing with models in a real world setting. Whether or not this is a problem of parameter-uncertainty or of the reliance on the validity of the model is a question that cannot be fully answered, yet.

In the current version of the framework, the transaction costs are represented by crude upper bounds that aid to preserve the linear structure of the subproblem of identifying the affiliation functions. While this allows to solve this sub-problem efficiently using mixed-integer solvers, it also ignores part of the permitted solution space, i.e. solutions that satisfy the constraints for the true cost, but do not satisfy the constraints for the upper bound. Another eventual limitation of the presented approach in the current form is the reliance on standard solvers for the subproblem of identifying the shape parameter of the portfolio. In the cases considered in this work, the objective function is given by a sum of ratios, a problem where no efficient solving strategy exists.

When the performance of the FEM-BV-Utility-method is compared to other approaches, it should be considered, that the optimization is currently done numerically. This does not only result in increased computational cost, but the obtained solution might only be a local optimum, a fact that might explain the poor out-of-sample performance in the high-

dimensional example.

From the author's point of view, the crudeness of the upper bound and the numerical optimization using standard solvers are important working points for future research. A different approach to the first point would allow to further improve the solution, thereby adding to its efficiency. A specialized solver or an analytical solution can significantly decrease the computational cost of the approach and might further improve the performance in high-dimensional settings.

Bibliography

[1] Abarbanell, J. S. and Bushee, B. J. 1997. Fundamental analysis, future earnings, and stock prices, *Journal of Accounting Research* **35**(1).

[2] Achterberg, T. 2009. Scip: Solving constraint integer programs, *Mathematical Programming Computation* **1**(1): 1–41. `http://mpc.zib.de/index.php/MPC/article/view/4`.

[3] *ActivTrades Online Broker Webpage, as of April* 2014. `http://www.activtrades.co.uk/index.aspx?page=cfds_indices_margins`.

[4] Akaike, H. 1974. A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6): 716–723.

[5] Baum, L. E. 1972. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes, *in* O. Shisha (ed.), *Inequalities III: Proceedings of the Third Symposium on Inequalities*, Academic Press, University of California, Los Angeles, pp. 1–8.

[6] Baum, L. and Petrie, T. 1966. Statistical inference for probabilistic functions of finite state Markov chains, *The Annals of Mathematical Statistics* **37**(6): 1554–1563.

[7] Berkson, J. 1980. Minimum chi-square, not maximum likelihood!, *The Annals of Statistics* **8**(3): pp. 457–487.
URL: *http://www.jstor.org/stable/2240587*

[8] Bertsimas, D. and Tsitsiklis, J. 1993. Simulated annealing, *Statistical Science* **8**(1): pp. 10–15.
URL: *http://www.jstor.org/stable/2246034*

[9] Bilmes, J. 1998. *A Gentle Tutorial of the EM Algorithm and its Applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report*, International Computer Science Institute, Berkeley.

[10] Bishop, C. M. 1996. *Neural Networks for Pattern Recognition*, 1 edn, Oxford University Press, USA.
URL: *http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0198538642*

[11] Black, F. 1976. Studies of stock price volatility changes, *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economics Statistics Section*, pp. 177–181.

[12] Black, F. and Scholes, M. 1973. The pricing of options and corporate liabilities, *The journal of political economy* pp. 637–654.

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

[14] Braess, D. 2001. *Finite elements: Theory, fast solvers, and applications in solid mechanics*, Cambridge University Press.

[15] Brandt, M. W., Santa-Clara, P. and Valkanov, R. 2009. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns, *Review of Financial Studies* **22**(9): 3411–3447.

[16] Burnham, K. P. and Anderson, D. R. 2002. *Model selection and multi-model inference: a practical information-theoretic approach*, Springer.

[17] Christiansen, B. 2005. The shortcomings of nonlinear principal component analysis in identifying circulation regimes, *J. Climate* **18**: 4814–4823.

[18] Crommelin, D. and Majda, A. 2004. Strategies for model reduction: Comparing different optimal bases, *J. Atmos. Sci.* **61**(17): 2206–2217.

[19] Cvitanić, J. and Karatzas, I. 1996. Hedging and portfolio optimization under transaction costs: A martingale approach, *Mathematical finance* **6**(2): 133–165.

[20] de Wijes, J., Putzig, L. and Horenko, I. 2014. Discrete non-homogenous and non-stationary logistic and markov regression models for spatio-temporal data with unresolved external influences, *accepted for publication in Communications in Applied Mathematics and Computational Science* .

[21] DeMiguel, V., Garlappi, L. and Uppal, R. 2009. Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy?, *Review of Financial Studies* **22**(5): 1915–1953.

[22] Dempster, A. P., Laird, N. M. and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1): 1–38.

[23] Dupire, B. 1994. Pricing with a smile, *Risk* pp. 18–20.

[24] Föllmer, H. and Schied, A. 2004. *Stochastic Finance, An Introduction in Discrete Time*, Studies in Mathematics 27, 2nd edn, de Gruyter.

[25] Garlappi, L., Uppal, R. and Wang, T. 2007. Portfolio selection with parameter and model uncertainty: A multi-prior approach, *Review of Financial Studies* **20**(1): 41–81.

[26] Gasser, T. and Müller, H. G. 1979. Kernel estimation of regression functions, *in* T. Gasser and M. Rosenblatt (eds), *Smoothing Techniques for Curve Estimation*, Vol. 757 of *Lecture Notes in Mathematics*, Springer Berlin / Heidelberg, pp. 23–68.

[27] Gasser, T. and Müller, H.-G. 1984. Estimating regression functions and their derivatives by the kernel method, *Scandinavian J. of Stat.* **11**(3): 171–185.

[28] Gelman, A., Carlin, J., Stern, H. and Rubin, D. 2004. *Bayesian Data Analysis*, Chapman and Hall.

[29] Gibbs, A. L. and Su, F. E. 2002. On Choosing and Bounding Probability Metrics, *International Statistical Review* **70**(3): 419–435.

[30] Goffe, W. L., Ferrier, G. D. and Rogers, J. 1994. Global optimization of statistical functions with simulated annealing, *Journal of Econometrics* **60**(1): 65–99.

[31] Hadamard, J. 1902. Sur les problèmes aux dérivées partielles et leur signification physique, *Princeton University Bulletin* **13**: 49–52.

[32] Haider, S., Parkinson, G. and Neidle, S. 2008. Molecular dynamics and principal components analysis of human telomeric quadruplex multimers, *Biophys. J.* **95**(1): 296–311.

[33] Hamilton, J. 1989. A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* **57**(2): 357–384.

[34] Haussmann, U. G. and Sass, J. 2004. Optimal terminal wealth under partial information for hmm stock returns, *Contemporary Mathematics* **351**: 171–186.

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

[36] Horenko, I. 2008. On simultaneous data-based dimension reduction and hidden phase identification, *J. Atmos. Sci.* **65**(6): 1941–1954.

[37] Horenko, I. 2009. On robust estimation of low-frequency variability trends in discrete Markovian sequences of Atmospherical Circulation Patterns, *J. Atmos. Sci.* **66**(7): 2059–2072.

[38] Horenko, I. 2010a. Finite element approach to clustering of multidimensional time series, *J. of Sci. Comp.* **32**(1): 62–83.

[39] Horenko, I. 2010b. On clustering of non-stationary meteorological time series, *Dyn. of Atm. and Oc.* **49**(2-3): 164–187.

[40] Horenko, I. 2010c. On identification of non-stationary factor models and its application to atmospherical data analysis, *J. Atmos. Sci.* **67**(5): 1559–1574.

[41] Horenko, I. 2011a. Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling, *J. Atmos. Sci.* **68**(7): 1493–1506.

[42] Horenko, I. 2011b. On analysis of nonstationary categorical data time series: dynamical dimension reduction, model selection and applications to computational sociology, *Multi. Mod. Sim.* **9**(4): 1700–1726.

[43] Horenko, I., Dittmer, E., Fischer, A. and Schütte, C. 2006. Automated model reduction for complex systems exhibiting metastability, *Mult. Mod. Sim.* **5**(3): 802–827.

[44] Horenko, I., Dittmer, E. and Schütte, C. 2005. Reduced stochastic models for complex molecular systems, *Comp. Vis. Sci.* **9**(2): 89–102.

[45] Horenko, I., Klein, R., Dolaptchiev, S. and Schütte, C. 2008. Automated generation of reduced stochastic weather models I: Simultaneous dimension and model reduction for time series analysis, *Mult. Mod. Sim.* **6**(4): 1125–1145.

[46] Hurvich, C. M. and Tsai, C.-L. 1989. Regression and time series model selection in small samples, *Biometrika* **76**(2): 297–307.

[47] Jaynes, E. 1957a. Information Theory and Statistical Mechanics, *Physical Review* **106**(4): 620–630.

[48] Jaynes, E. 1957b. Informational Theory and Statistical Mechanics II, *Physical Review* **108**(2): 171–190.

[49] Jobson, J. and Korkie, B. 1980. Estimation for markowitz efficient portfolios, *Journal of the American Statistical Association* pp. 544–554.

[50] Jolliffe, I. 2002. *Principal Component Analysis*, Springer.

[51] *Kansas City Board of Trade* n.d.. `http://www.kcbt.com/historical_data.asp`.

[52] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R. and Wu, A. Y. 2002. An efficient k-means clustering algorithm: Analysis and implementation, *IEEE Trans. Pattern Analysis and Machine Intelligence* **24**: 881–892.

[53] Kohavi, R. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection, Morgan Kaufmann, pp. 1137–1143.

[54] Kroner, K. F. and Sultan, J. 1993. Time-varying distributions and dynamic hedging with foreign currency futures, *Journal of Financial and Quantitative Analysis* **28**(04): 535–551.

[55] Ledoit, O., Santa-Clara, P. and Wolf, M. 2003. Flexible multivariate garch modeling with an application to international stock markets, *Review of Economics and Statistics* **85**(3): 735–747.

[56] Ledoit, O. and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices, *Journal of multivariate analysis* **88**(2): 365–411.

[57] Loader, C. 1999. *Local Regression and Likelihood*, Springer, New York.

[58] Lobo, M. S., Fazel, M. and Boyd, S. 2007. Portfolio optimization with linear and fixed transaction costs, *Annals of Operations Research* **152**(1): 341–365.

[59] MacQueen, J. B. 1967. Some Methods for Classification and Analysis of Multivariate Observations, *in* L. M. L. Cam and J. Neyman (eds), *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, pp. 281–297.

[60] Majda, A. and Wang, X. 2006. *Nonlinear Dynamics and Statistical Theories for Basic Geophysical Flows*, Cambridge University Press.

[61] Markowitz, H. 1952. Portfolio selection, *The journal of finance* **7**(1): 77–91.

[62] McLachlan, G. and Peel, D. 2004. *Finite mixture models*, Wiley. com.

[63] Merton, R. C. 1969. Lifetime portfolio selection under uncertainty: The continuous-time case, *The Review of Economics and Statistics* **51**(3): pp. 247–257.
URL: *http://www.jstor.org/stable/1926560*

[64] Merton, R. C. 1980. On estimating the expected return on the market: An exploratory investigation, *Journal of Financial Economics* **8**(4): 323–361.

[65] Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. 1953. Equation of state calculations by fast computing machines, *The journal of chemical physics* **21**(6): 1087–1092.

[66] Metzner, P., Putzig, L. and Horenko, I. 2012. Analysis of persistent nonstationary time series and applications, *Communications in Applied Mathematics and Computational Science* **7**(2): 175–229.

[67] Michaud, R. O. 1989. The markowitz optimization enigma: is 'optimized' optimal?, *Financial Analysts Journal* pp. 31–42.

[68] Michaud, R. O. and Michaud, R. 1999. Portfolio optimization by means of resampled efficient frontiers. US Patent 6,003,018.

[69] Powell, M. J. 1978. A fast algorithm for nonlinearly constrained optimization calculations, *Numerical analysis*, Springer, pp. 144–157.

[70] Putzig, L., Becherer, D. and Horenko, I. 2010. Optimial Allocation of a Futures Portfolio Utilizing Numerical Market Phase Detection, *SIAM J. Finan. Math.* **1**(1): 752–779.

[71] Rabiner, L. and Juang, B.-H. 1986. An introduction to hidden markov models, *ASSP Magazine, IEEE* **3**(1): 4–16.

[72] Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2): 257–286.

[73] Ragunathan, V. and Mitchell, H. 1997. *Modelling the time-varying correlations between national stock market returns*, Department of Economics and Finance, Faculty of Business, RMIT.

[74] Sharpe, W. F. 1998. The sharpe ratio, *Streetwise-The best of the Journal of Portfolio Management, Princeton University Press Princeton, New Jersy* pp. 169–185.

[75] Tibshirani, R. 1996. Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

[76] Tikhonov, A. 1943. On the stability of inverse problems, *Dokl. Akad. Nauk SSSR* **39**(5): 195–198.

[77] Tsay, R. 2005. *Analysis of financial time series*, Wiley Series in Probability and Statistics, Wiley-Interscience.

[78] *U.S. Department of Energy* n.d.. `http://tonto.eia.doe.gov/dnav/pet/pet_pri_fut_s1_d.htm`.

[79] von Neumann, J. and Morgenstern, O. 1947. *Theory of games and economic behavior*, 2nd edn, Princeton University Press.

[80] *Wilshire Webpage as of April* 2014. `http://web.wilshire.com/Indexes/Broad/Wilshire5000/Characteristics.html`.

[81] Wunderling, R. 1996. *Paralleler und objektorientierter Simplex-Algorithmus*, PhD thesis, Technische Universität Berlin. `http://www.zib.de/Publications/abstracts/TR-96-09/`.

[82] Zellner, A. and Highfield, R. A. 1988. Calculation of maximum entropy distributions and approximation of marginal posterior distributions, *Journal of Econometrics* **37**(2): 195 – 209.