Mathematical Methods of Operations Research
https://doi.org/10.1007/s00186-022-00784-y

ORIGINAL ARTICLE

Signal-to-noise matrix and model reduction in continuous-time hidden Markov models

Elisabeth Leoff[1] · Leonie Ruderer[2] · Jörn Sass[3]

Received: 1 August 2021 / Revised: 18 February 2022 / Accepted: 19 April 2022

Abstract

Continuous-time regime-switching models are a very popular class of models for financial applications. In this work the so-called signal-to-noise matrix is introduced for hidden Markov models where the switching is driven by an unobservable Markov chain. Its relations to filtering, i.e. state estimation of the chain given the available observations, and portfolio optimization are investigated. A convergence result for the filter is derived: The filter converges to its invariant distribution if the eigenvalues of the signal-to-noise matrix converge to zero. This matrix is then also used to prove a mutual fund representation for regime-switching models and a corresponding market reduction which is consistent with filtering and portfolio optimization. Two canonical cases for the reduction are analyzed in more detail, the first based on the market regimes and the second depending on the eigenvalues. These considerations are presented both for observable and unobservable Markov chains. The results are illustrated by numerical simulations.

Keywords Hidden Markov model · Mutual fund · Portfolio optimization · Regime switching · Stochastic filtering

Mathematics Subject Classification Primary 91G15 · Secondary: 91G10 · 93E11

✉ Jörn Sass
sass@mathematik.uni-kl.de

Elisabeth Leoff
elisabeth.leoff@itwm.fraunhofer.de

[1] Department of Financial Mathematics, Fraunhofer Institute for Industrial Mathematics ITWM, Kaiserslautern, Germany

[2] Bad Kreuznach, Germany

[3] Department of Mathematics, Technische Universität Kaiserslautern, Kaiserslautern, Germany

Published online: 23 May 2022

1 Introduction

Regime-switching models are a very popular class of models in the field of mathematical finance. They describe return processes with time-changing drift or volatility parameters. Thus, they are a possible way to generalize the classical Black-Scholes lognormal stock price model by making the parameters dependent on a Markov chain with finitely many states. The switching parameters allow for flexible and realistic fits to observed market data. In the continuous-time model with switching volatility, the underlying Markov chain is observable (in theory) due to this stochastic volatility and no estimation (filtering) of it is needed. We call this model the Markov-switching model (MSM). In the model with constant volatility one has to filter for the underlying hidden Markov chain, i.e. to compute the conditional probability for the underlying state given the observed stock returns. Therefore, this model is called the hidden Markov model (HMM). The filtering problem was solved in the continuous-time model in Wonham (1965) and Elliott (1993). It was discretized and robustified in James et al. (1996) in the sense of Clark (1978), being consistent with filtering in discrete time as in Hamilton (1989). Portfolio optimization in the continuous-time HMM covering logarithmic and power utility was solved e.g. by Sass and Haussmann (2004) using Malliavin calculus and by Bäuerle and Rieder (2005) following an HJB approach. Also BSDE methods can be applied to portfolio optimization under partial information, see Papanicolaou (2019).

In portfolio optimization, estimating parameters or the true value of a system using "noisy" real-world observations can lead to poor portfolio performance already in the one-period model, when e.g. constructing a minimum-variance portfolio or applying mean-variance optimization. This is especially true for markets with a large number of assets (DeMiguel et al. 2007). There is an extensive literature available on how to approach such estimation problems, mainly in the one-period model. One may cluster according to the correlations and invest, in the spirit of DeMiguel et al. (2007), with equal weights in the representatives (e.g. Sass and Thös 2022). Or, as in Zhao et al. (2019), tackle such a problem by splitting the eigenvectors of the covariance matrix into well-estimated and poorly-estimated ones and use these to construct portfolios. Chen and Yuan (2016) restrict the investment in the mean-variance analysis on a subspace, using e.g. the leading eigenvectors of the covariance matrix to span this subspace. Avellaneda et al. (2021), Avellaneda and Lee (2010), Boyle (2014) also use a principal component analysis of the correlation matrix, leading to so-called eigenportfolios (see Remark 5.12 for how this relates to our setting).

Mutual funds are a well-known and famous concept from classical finance and financial mathematics which have been studied for a long time. Mutual fund separation theorems imply that it is optimal to trade in a portfolio that is a linear combination of single assets. The separation can e.g. be studied in the context of mean-variance analysis (Tobin 1958; Merton 1972) or using an expected utility setting (Cass and Stiglitz 1970; Schachermayer et al. 2009). The approach of Chamberlain (1988) uses martingale representation theory. General applicability of the Mutual Fund Theorem was discussed in Schachermayer et al. (2009) for classes of utility functions under some completeness condition. There, the mutual fund depends on time and there is only one fund of risky assets next to the riskfree bond.

In this work we bring both concepts, regime-switching models and mutual funds, together by introducing a reduced market model where one can trade in several mutual funds. In this reduced market we derive a result which provides a representation by as many mutual funds as we have market regimes. Each of these mutual funds would correspond to the classical mutual separation theorem, if we are in a model with static parameters corresponding to that regime. The reduced and the original model are connected by what we call the signal-to-noise matrix. It depends on the parameters of the HMM, namely it relates the states of the drift to the volatility. We prove that this matrix plays a central role in both portfolio optimization and filtering. Its name is motivated by the expression "signal-to-noise ratio" known from classical filtering theory and signal analysis. As an indicator of how much information about the signal can be present in the observation, the signal-to-noise matrix is also connected to the possible performance of the filter. The invariant distribution of the underlying Markov chain is always of significance for the filter as well, since it is approximately its expected value and also often chosen as the starting value.

We prove that for vanishing signal-to-noise matrix in terms of its eigenvalues, the filter indeed converges to this invariant distribution. The convergence result implies that in this case, the observations do not contain significant information about the signal anymore. We then derive a market reduction from the signal-to-noise matrix that depends on a condition on its eigenvalues. We establish the equivalence of the reduced market to the original model and the equivalence of the filters, depending on this condition. The funds are a linear combination of the original assets and do not change their composition over time. We also provide an explicit calculation for the composition of the funds. We show that portfolio optimization in the reduced market leads to an optimal wealth process that is identical to the optimal wealth in the original model. Furthermore, we consider two canonical cases for the reduced model: First the so-called "reduced regime representation model" (RRRM), where it turns out to be log-optimal to invest portions of wealth into the funds according to the filtered state probabilities. The second case is the "eigenvalue representation" (REVM) which is based on a principal component analysis of the signal-to-noise matrix. In this setting, the log-optimal strategy not only depends on the filter, but is also scaled by the eigenvalues of the signal-to-noise matrix. If the signal-to-noise matrix is singular, the market reduction looks differently. Then, we still arrive at an optimal terminal wealth that is identical to the original model, but the optimal strategy depends on the original observations. The theoretical results are underlined by a numerical evaluation of the optimization problem. The simulations suggest that the investor finds herself in the classical dilemma of risk versus gain. Again the signal-to-noise matrix assigning the funds is of great importance: controlling its values decides in which direction the portfolio is lead.

To summarize, in this work we prove a mutual fund representation and market reduction that is derived from the so-called signal-to-noise matrix. We investigate the relation of this matrix to filtering and portfolio optimization and prove convergence of the filter to the invariant distribution for vanishing eigenvalues. We present two canonical cases for the market reduction, the regime representation and the eigenvalue representation. For the latter case, a distinction is made depending on the number of non-zero eigenvalues of the signal-to-noise matrix. Furthermore, we present numerical simulations to illustrate our results. The main innovation is to base our analysis on the signal-to-noise matrix which is more informative in the filtering setting than the correlation matrix. This allows for new and more explicit results than focusing on the risk premium (cf. our discussion at the end of Sect. 4.1).

This work is organized as follows: In Sects. 2 and 3 we collect well-known results on filtering and portfolio optimization in regime-switching models. In Sect. 4 we formally introduce the signal-to-noise matrix and prove that a vanishing signal-to-noise matrix leads to convergence of the filter to the invariant distribution. In Sect. 5 we use a decomposition of the signal-to-noise matrix to obtain one of the main results of this work: A mutual fund theorem and corresponding model reduction. The effects of these model reductions on both portfolio optimization and filtering are investigated. As it turns out, the case of a singular signal-to-noise matrix has to be handled separately, which we do in Sect. 5.4. In Sect. 5.5 we then discuss corresponding results for the MSM and an HMM with non-constant volatility which can be seen as a model that lies between MSM and HMM. The conclusion in Sect. 6 summarizes our contributions.

In the following, we will use some abbreviations and acronyms. We summarize them here:

MSM	Markov-switching model
HMM	Hidden Markov model
RRRM	Reduced regime representation model
REVM	Reduced eigenvalue model
CRRA	Constant relative risk aversion
FB-HMM	Filter-based hidden Markov model

2 Financial market model and filtering

2.1 Regime-switching model

We consider a multivariate continuous-time regime switching model for asset returns, consisting of an n-dimensional observation process $R = (R_t)_{t\in[0,T]}$, which models the returns of n stocks,

$$R_t = \int_0^t \mu_s \,\mathrm{d}s + \int_0^t \sigma_s \,\mathrm{d}W_s. \tag{2.1}$$

While the n-dimensional Brownian motion $W = (W_t)_{t\in[0,T]}$ models the noise as usual, the regime changes, i.e. the switching of the drift and volatility parameters, are driven by a time-homogeneous, continuous-time Markov chain $Y = (Y_t)_{t\in[0,T)}$ which is independent of W. The Markov chain Y has d states and for convenience we use as states $\{e_1, \dots, e_d\}$, the unit vectors in $\mathbb{R}^d$. We write

$$\mu_t = BY_t \quad \text{and} \quad \sigma_t = \sigma(Y_t),$$

where $B \in \mathbb{R}^{n \times d}$ is any matrix while the matrices $\sigma(\mathrm{e}_k) \in \mathbb{R}^{n \times n}_{>0}$ are supposed to be non-singular. A popular intuition behind Y is that it reflects the current underlying state of the economy. To describe its dynamics we additionally need the rate matrix $Q \in \mathbb{R}^{d \times d}$ for which the negative of the diagonal element $-Q_{kk}$ provides the exponential rate for leaving state e_k, $k = 1, \ldots, d$, and the ratio $-Q_{kl}/Q_{kk}$ is the transition probability from e_k to e_l, $l \neq k$, if the chain jumps at all. We assume that the chain is irreducible. Therefore, under our condition of having finitely many states, a unique stationary distribution ν exists and is given by $\nu^\top Q = 0$, $\nu^\top \mathbf{1} = \mathbf{1}$.

The prices of the n stocks are then given by

$$dS_t = \mathrm{Diag}(S_t)\, \mathrm{d}R_t = \mathrm{Diag}(S_t)\, (BY_t\, \mathrm{d}t + \sigma(Y_t)\mathrm{d}W_t)$$

where $\mathrm{Diag}(y)$ denotes the diagonal matrix with diagonal y. Further, there is a money market account for which we assume interest rate $r = 0$ to keep notation simple. The results for filtering and for portfolio optimization below can be adapted to non-zero r easily.

2.2 MSM and HMM

In a suitable probability space $(\Omega, \mathcal{A}, P)$ for the model above we shall distinguish two filtrations. On the one hand we have $\mathcal{F} = (\mathcal{F}_t)_{t \in [0,T]}$, which is generated by Y and W, and augmented by the null sets. This corresponds to *full information*. For convenience we assume $\mathcal{A} = \mathcal{F}_T$. On the other hand, in real-world applications an investor typically can only rely on the observed stock prices or stock returns (this is equivalent in this model). Her information thus is given by $\mathcal{F}^R = (\mathcal{F}^R_t)_{t \in [0,T]}$, which is the filtration generated by R, again augmented by the null sets. We say that an investor with information $\mathcal{F}^R$ has *partial information*.

Let us denote $\Sigma(\mathrm{e}_i) = \sigma(\mathrm{e}_i)\sigma(e_i)^\top$. In the continuous-time regime switching model, $\Sigma(Y_t)$ can in theory be observed from the quadratic covariation of the stock returns and thus one can observe the jumps of the underlying chain if the matrices $\Sigma(\mathrm{e}_i)$, $i = 1, \ldots, d$ are pairwise different. Then—in theory—an investor with partial information has in fact full information, since Y_t can be obtained from $\Sigma(Y_t)$. For a full discussion and more details, see Elliott et al. (2008), Krishnamurthy et al. (2018). Note that in reality with discrete-time observations of the continuous process this would not be true. Nevertheless, in the theory we have to distinguish the following cases.

Definition 2.1 We call the model outlined in Sect. 2.1 *Markov switching model (MSM)*, if $\Sigma(\mathrm{e}_i) \neq \Sigma(\mathrm{e}_j)$ for all $i \neq j$. If $\Sigma(\mathrm{e}_1) = \cdots = \Sigma(e_d)$, then we call the model a *hidden Markov model (HMM)*.

MSM and HMM are the extreme cases. For settings for which only some of the $\Sigma(e_i)$ agree, the subsequent results can be adapted.

As discussed before, in the MSM an investor has full observation and thus she knows Y_t and hence $\mu_t = BY_t$ and $\Sigma_t = \Sigma(Y_t)$ at t. An investor with partial information has to estimate the underlying drift. The usual approach is to use the L^2-optimal estimate for μ_t at t. This quantity is called the *filter* and is defined by

$$\hat{\mu}_t := \mathrm{E}[\mu_t \mid \mathcal{F}_t^R] = B\,\mathrm{E}[Y_t \mid \mathcal{F}_t^R] = B\hat{Y}_t, \tag{2.2}$$

where $\hat{Y}_t := \mathrm{E}[Y_t \mid \mathcal{F}_t^R]$ is the well-known continuous-time Wonham filter for Y_t (Wonham 1965), see Sect. 2.3 below. Up to the beginning of Sect. 5.5 we shall concentrate on the HMM since filtering issues play an important role in our considerations.

2.3 Filtering in the HMM

We consider the HMM $\mathrm{d}R_t = \mu_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t$, $\mu_t = BY_t$, i.e. $\Sigma = \sigma\sigma^\top$ is constant and the filtering problem is non-trivial. To find the filter $\hat{Y}_t = \mathrm{E}[Y_t \mid \mathcal{F}_t^R]$ which yields $\hat{\mu}_t$ by (2.2), we can use a change of measure to $\tilde{P} \sim P$ with Radon-Nikodym derivative

$$\frac{\mathrm{d}\tilde{P}}{\mathrm{d}P} = Z_T, \quad \mathrm{d}Z_t = -Z_t Y_t^\top (\sigma^{-1}B)^\top\,\mathrm{d}W_t. \tag{2.3}$$

Under $\tilde{P}$, $\tilde{W}_t = \sigma^{-1}R_t$ is a Brownian motion independent of Y. $\tilde{P}$ is called *reference measure* in filtering and for interest rate 0 it corresponds to the *risk neutral* or *equivalent martingale measure* in finance, see e.g. Elliott (1993). By $\tilde{\mathrm{E}}$ we denote the expectation under $\tilde{P}$. This reference measure is used to introduce the unnormalized filter $\rho_t := \tilde{\mathrm{E}}[Z_T^{-1}Y_t \mid \mathcal{F}_t^R]$, which satisfies the *Zakai-equation* (Elliott 1993)

$$\mathrm{d}\rho_t = Q^\top \rho_t\,\mathrm{d}t + \mathrm{Diag}(\rho_t)B^\top\,\Sigma^{-1}\mathrm{d}R_t, \quad \rho_0 = \mathrm{E}[Y_0]. \tag{2.4}$$

The Zakai-equation is linear in ρ_t and driven by the observations. Using Bayes' formula for $\hat{Z}_t := \mathrm{E}[Z_t \mid \mathcal{F}_t^R]$ yields

$$\hat{Z}_t = \frac{1}{\tilde{\mathrm{E}}[Z_T^{-1} \mid \mathcal{F}_t^R]}.$$

By the definition of ρ_t this implies

$$\hat{Z}_t^{-1} = \tilde{\mathrm{E}}[Z_T^{-1} \mid \mathcal{F}_t^R] = \tilde{\mathrm{E}}[Z_T^{-1}\mathbf{1}^\top Y_t \mid \mathcal{F}_t^R] = \mathbf{1}^\top \rho_t \tag{2.5}$$

and thus by (2.4) and by $Q\mathbf{1} = 0$ we get

$$\mathrm{d}\hat{Z}_t^{-1} = \rho_t^\top B^\top \Sigma^{-1}\mathrm{d}R_t. \tag{2.6}$$

Using (2.5), Bayes' formula for conditional expectations, also called *Kallianpur-Striebel formula* in this context, reads

$$\hat{Y}_t = \frac{\rho_t}{\mathbf{1}^\top \rho_t}. \tag{2.7}$$

This implies that knowing ρ_t, the filter $\hat{Y}_t$ can be calculated directly. Thus, in filtering, one typically tries to compute ρ_t. By (2.4) and (2.6) and applying Itô's formula to (2.7) we get the *Kushner–Stratonovich equation*

$$\mathrm{d}\hat{Y}_t = Q^\top \hat{Y}_t \mathrm{d}t + \left(\mathrm{Diag}(\hat{Y}_t) - \hat{Y}_t \hat{Y}_t^\top\right)(\sigma^{-1}B)^\top \left(\sigma^{-1}\mathrm{d}R_t - \sigma^{-1}B\hat{Y}_t \mathrm{d}t\right). \tag{2.8}$$

3 Trading and portfolio optimization

Remember that we consider one money market account with interest rate 0 and n stocks with returns

$$\mathrm{d}R_t = \mu_t\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t, \quad \mu_t = BY_t,$$

where σ_t is switching with Y_t in the MSM, constant in the HMM. We may also allow for suitable $\mathcal{F}^R$-adapted volatility processes as we use them in Sect. 5.5. We set $\Sigma_t = \sigma_t\sigma_t^\top$. In all cases, the trading strategy of an investor can be described by her initial capital $x_0 > 0$ and the risky fraction process $\pi = (\pi_t)_{t\in[0,T]}$ if the wealth stays strictly positive (which is the case for the utility functions we will consider). The wealth process $(X_t^\pi)_{t\in[0,T]}$ then follows

$$\mathrm{d}X_t^\pi = X_t^\pi\,\pi_t^\top \mathrm{d}R_t, \quad X_0 = x_0. \tag{3.1}$$

So π_t^i denotes the fraction of wealth X_t invested in stock i. For given $x_0 > 0$ the admissible π are

$$\begin{aligned}\mathcal{A}(x_0) = \Big\{&\pi = (\pi_t)_{t\in[0,T]} \,:\, \pi \text{ progressively measurable, } \mathcal{F}^R\text{-adapted,}\\ &X_t^\pi > 0 \text{ for all } t \text{ a.s.}, \int_0^T (\|B^\top \pi_t\| + \|\sigma_t^\top \pi_t\|^2)\mathrm{d}t < \infty \text{ a.s.}\Big\},\end{aligned}$$

in particular an investor can at time t only use the information $\mathcal{F}_t^R$ obtained from observing the stock returns. However, for the MSM this is equivalent to having full information while for the HMM this is a case with strictly partial information, cf. the discussion in Sect. 2.2.

Note that (3.1) has the explicit solution

$$X_t^\pi = x_0 \,\exp\left\{\int_0^t \left(\pi_s^\top \mu_s - \frac{1}{2}\pi_s^\top \Sigma_s \pi_s\right)\mathrm{d}s + \int_0^t \pi_s^\top \sigma_s\,\mathrm{d}W_s\right\}, \quad t \in [0, T].$$

We evaluate the terminal wealth by a *utility function* $U : [0, \infty) \to \mathbb{R} \cup \{-\infty\}$ which is strictly increasing, strictly concave, twice continuously differentiable on $(0, \infty)$ with $\lim_{x\searrow 0} U'(x) = \infty$ and $\lim_{x\to\infty} U'(x) = 0$. We further denote by $I : (0, \infty) \to (0, \infty)$ the inverse of U'.

We will focus on power and logarithmic utility functions,

$$U_\alpha(x) = \frac{x^\alpha}{\alpha},\ \alpha < 1, \alpha \neq 1 \quad \text{and} \quad U_0(x) = \log(x)$$

which are utility functions with constant relative risk aversion (CRRA).

The problem of maximizing expected utility of terminal wealth then is: Maximize for $x_0 > 0$

$$\mathrm{E}[U(X_T^\pi)] \quad \text{over} \quad \pi \in \mathcal{A}_U(x_0), \tag{3.2}$$

where $\mathcal{A}_U(x_0) = \{\pi \in \mathcal{A}(x_0) : \mathrm{E}[U^-(X_T^\pi)] < \infty\}$ is the set of risky fraction processes admissible for U, with U^- its negative part. The problem (3.2) has been solved in Sass and Haussmann (2004) for the HMM with partial information. For general U, it is quite straightforward to show

$$X_T^* = I(\lambda \hat{Z}_T), \quad \text{where } \lambda > 0 \text{ is determined by} \quad \tilde{\mathrm{E}}[X_T^*] = x_0. \tag{3.3}$$

The difficulty lies then in finding the strategy as explicitly as possible. We cite the following result.

Theorem 3.1 *In the HMM for $U = U_\alpha$, $\alpha < 1$,*

$$\pi_t^* = \frac{1}{(1-\alpha)\mathrm{E}\left[\hat{Z}_{t,T}^{\frac{\alpha}{\alpha-1}} \mid \rho_t\right]} \Bigg\{ \Sigma^{-1} B \hat{Y}_t \mathrm{E}\left[\hat{Z}_{t,T}^{\frac{2\alpha-1}{\alpha-1}} \mid \rho_t\right] + (\sigma^\top)^{-1} \mathrm{E}\Big[\hat{Z}_{t,T}^{\frac{2\alpha-1}{\alpha-1}} \int_t^T (D_t \rho_{t,s}) B^\top \Sigma^{-1} \mathrm{d}R_s \;\Big|\; \rho_t\Big]\Bigg\},$$

where for $t \in [0, T]$ the Malliavin deriviative $D_t \rho_{t,s}$, $s \in [t, T]$, follows

$$D_t \rho_{t,s} = \sigma^{-1} B \mathrm{Diag}(\hat{Y}_t) + \int_t^s (D_t \rho_{t,u}) Q \,\mathrm{d}u + \int_t^s (D_t \rho_{t,u}) \mathrm{Diag}(B^\top \Sigma^{-1} \mathrm{d}R_u),$$

and where $\rho_{t,s} = \rho_s / \mathbf{1}^\top \rho_t$, $\hat{Z}_{t,s} = \hat{Z}_s / \hat{Z}_t$.

In particular $\pi_t^* = \Sigma^{-1} B \hat{Y}_t$ for $U = U_0 = \log$.

Proof This is a special case of Theorem 4.5 in Sass and Haussmann (2004) which uses the linearity of the Zakai equation (2.4) in order to show existence of the Malliavin derivative of X_T^*, cf. Corollaries 4.8, 4.9 and Proposition 4.10 in Sass and Haussmann (2004). □

Note that the result also states that ρ_t is a sufficient statistic to compute the conditional expectations. This is due to the fact that ρ_t satisfies the stochastic differential equation (2.4) which is driven by the observations and allows to derive the corresponding Markov property and thus to simplify the initial conditional expectation given $\mathcal{F}_t^R$. Practically it means that the strategy can be computed efficiently from the unnormalized filter ρ_t.

For the MSM, [Bäuerle and Rieder (2004), Theorems 2 and 3] provide optimal policies for CRRA utility functions:

Theorem 3.2 *In the MSM, for $U = U_\alpha$,*

$$\pi_t^* = \frac{1}{1-\alpha} \Sigma(Y_t)^{-1} B Y_t,$$

in particular $\pi_t^* = \Sigma(Y_t)^{-1} B Y_t$ *for* $U = \log$.

In the following we concentrate on the HMM since it involves the filtering problem, and discuss the MSM afterwards in Sect. 5.5 again.

4 Signal-to-noise matrix and convergence

The main idea for the model reduction in Sect. 5 is the observation that filter and portfolio optimization essentially depend on a lower-dimensional matrix which we introduce in Sect. 4.1 and whose influence we illustrate by a convergence result in Sect. 4.2.

4.1 Signal-to-noise matrix

In the fundamental results on filtering and portfolio optimization as presented in Sects. 2.3 and 3, for the n-dimensional HMM $\mathrm{d}R_t = BY_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t$, the dependency on σ and B is only via the *signal-to-noise matrix* A or its “root” Θ,

$$A := B^\top \Sigma^{-1} B = \Theta^\top \Theta, \quad \text{where} \quad \Theta := \sigma^{-1} B.$$

E.g., by (2.4) for the unnormalized filter

$$\mathrm{d}\rho_t = Q^\top \rho_t\,\mathrm{d}t + \mathrm{Diag}(\rho_t) B^\top\,\Sigma^{-1} \mathrm{d}R_t, \quad B^\top\,\Sigma^{-1}\mathrm{d}R_t = \Theta^\top \mathrm{d}\tilde{W}_t,$$

and for $U = \log$, $\pi_t^* = \Sigma^{-1} B \hat{Y}_t$, $\mathrm{d}X_t^* = X_t^* (\pi_t^*)^\top \mathrm{d}R_t = X_t^* \hat{Y}_t^\top \Theta^\top \mathrm{d}\tilde{W}_t$. By Theorem 3.1 we also have

$$\mathrm{E}[\log(X_T^*)] = \log(x_0) + \frac{1}{2}\mathrm{E}\left[\int_0^T \hat{Y}_t^\top A \hat{Y}_t\,\mathrm{d}t\right]. \tag{4.1}$$

But the signal-to-noise matrix A is $d \times d$-dimensional! This allows for a reduction of the model dimension in case $d < n$. Before we introduce this in Sect. 5, we first show in the following section that by decreasing the signal-to-noise matrix, we end up with a trivial filter which corresponds to having no information at all. This motivates our name for A and underlines the intuition that the relation between drift and volatility parameters is decisive for the performance of the filter. However, note that ΘY_t is the *market price of risk* or the *risk premium* and $Y_t^\top A Y_t$ (or their counterparts using the filter $\hat{Y}$ instead of Y) may be called the *risk premium function*. This quantity plays a prominent role in financial applications, e.g. in the analysis of robust continuous-time mean-variance problems, cf. Pham et al. (2022). However, since the filter $\hat{Y}$ depends on A, we can get more explicit structural results in our filtering setting by concentrating on A instead of the risk premium.

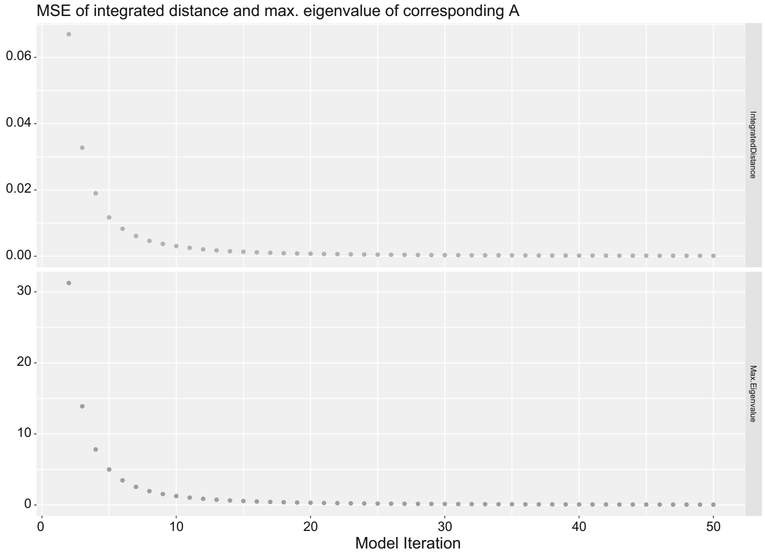

Fig. 1 Convergence to the stationary distribution for decreasing $\lambda_{max}(A)$

4.2 Convergence of the filter

As pointed out in Sect. 4.1, in (2.4) the dependence of the filter on σ and B is only indirectly through the matrix A. In the following we want to study the influence of changes in A on the behaviour of the filter. Intuitively, A describes a proportion between volatility and drift in the observations, and indirectly between volatility and Markov chain Y. This relation can also be seen as an indicator for how much information is present in the observation. Thus, in the following numerical example we consider a setting where the eigenvalues of A tend to zero and see how the average performance of the filter changes.

Example 4.1 We consider a sequence of 1-dimensional HMMs, where the eigenvalues of A decrease. For all iterations of the model, we choose the same rate matrix Q with invariant distribution ν and the same state matrix B as later on in Example 5.19.

The decrease in A is achieved by increasing the volatility, i.e. we choose a sequence of volatilities $\sigma_n = 0.2 \cdot n$. The expected squared distance between the filter and the invariant distribution integrated over time, $\mathrm{E}[\int_0^T ||\hat{Y}_t - \nu||^2 \mathrm{d}t]$, is plotted in Fig. 1.

We clearly see that for parameter choices where the largest eigenvalue of A is close to 0, the filter $\hat{Y}_t$ is close to the vector ν of the invariant distribution, where closeness to ν is not in a distribution sense but in $L^2(\mathbb{R}^d)$-distance. Since $\mathrm{E}[\hat{Y}_t] = \nu$ this means the filter does not contain much information about the true state of Y.

If the eigenvalues of A tend to 0, we can imagine that the volatility dominates the drifts, so the information about Y encoded in the returns is overlaid by too much noise. We have less information compared to a model where the ratio of drift to volatility is higher. Note that on the one hand, in the filtering equations we need σ to move away from the non-informative starting value of the invariant distribution in the first place to learn dynamically. On the other hand, as we see it here, “too much” σ compared to B means of course losing information.

We will formalize this intuition by proving that the distances between the filters and ν in a sequence of models converge to 0 if the eigenvalues of the corresponding A converge to 0. We first prove a stability result for SDEs using Doob’s martingale inequality and Gronwall’s lemma and then apply this result for the SDE of the normalized filter $\hat{Y}$. For the detailed proof see Appendix A.1.

Theorem 4.2 *Let X^n, X be d-dimensional processes bounded by* 1 *and satisfying*

$$\mathrm{d}X_t^n = a(X_t^n)\mathrm{d}t + b^n(X_t^n)\mathrm{d}W_t^n, \qquad X_0^n = x_0$$

and

$$\mathrm{d}X_t = a(X_t)\mathrm{d}t, \qquad X_0 = x_0$$

with W^n m-dimensional Brownian motions, $a : \mathbb{R}^d \mapsto \mathbb{R}^d$ Lipschitz-continuous and $b^n : \mathbb{R}^d \mapsto \mathbb{R}^{d\times m}$ bounded for X^n, i.e. $\|b^n(X_t^n)\|_{dm}^2 = \sum_{i,j} b_{ij}^n(X_t^n)^2$ is bounded. Further assume that b^n converges to 0 *along X^n in the sense that*

$$\|b^n(X^n)\|_{L^2([0,t])} = \left(\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s\right)^{\frac{1}{2}} \overset{(n\to\infty)}{\longrightarrow} 0$$

a.s. for all t. Then

$$\mathrm{E}\left[\sup_{s\le t} \|X_s^n - X_s\|^2\right] \overset{(n\to\infty)}{\longrightarrow} 0.$$

Now we consider a series of parameters σ_m, B_m giving rise to a series of HMMs R^m with respect to the same Markov chain and Brownian motion

$$\mathrm{d}R_t^m = B_m Y \mathrm{d}t + \sigma_m \mathrm{d}W_t.$$

Recall that the corresponding normalized filters $\hat{Y}^m$ are then given by

$$\mathrm{d}\hat{Y}_t^m = Q^T \mathrm{d}t + \sum_{i=1}^{n} (\mathrm{diag}(g_i^m)\hat{Y}_t^m - (g_i^m)^T \hat{Y}_t^m \hat{Y}_t^m \mathrm{d}(V_t^m)_i,$$

where g_i^m is the i^{th} column of $(\sigma_m^{-1} B_m)^T$ and V^m are the innovation processes. For all m we can define the matrix $A_m := \sigma_m^{-1} B_m (\sigma_m^{-1} B_m)^T$ of the HMM with parameters σ_m, B_m.

Our aim is to prove the convergence of $\hat{Y}_t^m$ for "too much" volatility. Using Theorem 4.2 we can show that this "too much" is governed by the behaviour of the eigenvalues of A_m. This is exactly what we have seen in Example 4.1. For the proof see Appendix A.2.

Theorem 4.3 *Consider the series of HMMs R^m as above and let $\lambda_m = (\lambda_{max}(A_m))^{1/2}$ be the sequence of square roots of the largest eigenvalue of A_m. Assume that $\lim_{m\to\infty} \lambda_m = 0$. Then we have for all t that*

$$\lim_{m\to\infty} \mathrm{E}\left[\sup_{s\le t} \left\| \hat{Y}_s^m - \nu \right\|^2\right] = 0,$$

that is for all $t > 0$ the sequence of normalized filters converges in L^2 to the probability vector of the invariant distribution.

5 Model reduction in the HMM

For a model reduction, we want to arrive at a d-dimensional model with the same performance and filter dynamics as the original n-dimensional model. To achieve this we should aim for a model with the same signal-to-noise matrix, as pointed out in Sect. 4.1. This observation inspires the definition of the model reduction that we discuss in the following.

5.1 HMM with non-singular signal-to-noise matrix

We consider the HMM

$$\mathrm{d}R_t = BY_t\,\mathrm{d}t + \sigma\,\mathrm{d}W_t, \quad \Sigma = \sigma\sigma^\top, \tag{5.1}$$

and first look in this section at the main case that the signal-to-noise matrix $A = B^\top \Sigma^{-1} B$ has full rank. This is the typical case if $n \ge d$, e.g. if we model a market with n risky assets for high n by an underlying Markov chain with d states corresponding to a few market regimes.

The idea is to find a d-dimensional return process

$$\mathrm{d}\breve{R}_t = CY_t\,\mathrm{d}t + \delta\,\mathrm{d}\breve{W}_t, \tag{5.2}$$

where $\breve{W}$ and $\delta^{-1}\breve{R}$ are d-dimensional Brownian motions under P and $\tilde{P}$, respectively. Let us denote by $\breve{\mathcal{F}} = (\breve{\mathcal{F}}_t)_{t\in[0,T]}$ the filtration generated by Y and $\breve{W}$, and by $\breve{\mathcal{F}}^R = (\breve{\mathcal{F}}_t^R)_{t\in[0,T]}$ the one generated by $\breve{R}$, both augmented by the null sets. Since the dimension of $\breve{R}$ is $d \le n$, we define:

Definition 5.1 We call a model with returns satisfying (5.2), where C and $D = \delta\delta^\top$ are non-singular matrices in $\mathbb{R}^{d\times d}$ with

$$C^\top D^{-1} C = A, \quad C^\top D^{-1}\mathrm{d}\breve{R}_t = B^\top \Sigma^{-1}\mathrm{d}R_t \tag{5.3}$$

a *reduced model* of the model with signal-to-noise matrix $A = B^\top \Sigma^{-1} B$.

Theorem 5.2 *(i) In a reduced model for C and δ, the Brownian motion $\breve{W}$ in (5.2) is given by*

$$\breve{W}_t = \delta^\top (C^{-1})^\top B^\top (\sigma^{-1})^\top W_t, \quad t \in [0, T].$$

Then, $\breve{\mathcal{F}}_t \subseteq \mathcal{F}_t$, $t \in [0, T]$.

(ii) The reference measure $\breve{\tilde{P}}$ in the reduced model has the same Radon–Nikodym derivative Z_T as $\tilde{P}$ and thus $\tilde{P}$ agrees on $\breve{\mathcal{F}}_T$ with $\breve{\tilde{P}}$.

(iii) The filter for Y in a reduced model is indistinguishable from the filter $\hat{Y}$ in the original model, in particular

$$\mathrm{E}[Y_t \mid \mathcal{F}_t^{\breve{R}}] = \hat{Y}_t \textit{ for all } t \in [0, T] \quad \textit{a.s.}.$$

The same is true for the unnormalized filter.

Proof (i) Solving $C^\top D^{-1} d\breve{R}_t = B^\top \Sigma^{-1} dR_t$ for $\breve{W}$ using (5.1), (5.2), we see that

$$\breve{W}_t := \delta^\top (C^{-1})^\top B^\top (\sigma^{-1})^\top W_t$$

provides the only possible candidate for $\breve{W}$. Then, by Lévy's characterization of Brownian motion, $\breve{W}$ is a Wiener process, since it is a continuous martingale with

$$\begin{aligned}
[\breve{W}, \breve{W}]_t &= \delta^\top (C^{-1})^\top B^\top (\sigma^{-1})^\top [W, W]_t (\delta^\top (C^{-1})^\top B^\top (\sigma^{-1})^\top)^\top \\
&= \delta^\top (C^{-1})^\top B^\top (\sigma^{-1})^\top I_d \, t \, \sigma^{-1} B C^{-1} \delta \\
&= \delta^\top (C^{-1})^\top A \sigma^{-1} B C^{-1} \delta \, t = I_d \, t.
\end{aligned}$$

Vice versa, for $\mathrm{d}R_t = BY_t \, \mathrm{d}t + \sigma \, \mathrm{d}W_t$ and $\mathrm{d}\breve{R}_t = CY_t \, \mathrm{d}t + \delta \, \mathrm{d}\breve{W}_t$ we then have

$$\begin{aligned}
B^\top \Sigma^{-1} \mathrm{d}R_t &= B^\top \Sigma^{-1} B Y_t \, \mathrm{d}t + B^\top \Sigma^{-1} \sigma \, \mathrm{d}W_t \\
&= A Y_t \, \mathrm{d}t + C^\top (\delta^\top)^{-1} \mathrm{d}\breve{W}_t \\
&= C^\top D^{-1} C Y_t \, \mathrm{d}t + C^\top D^{-1} \delta \, \mathrm{d}\breve{W}_t = C^\top D^{-1} \mathrm{d}\breve{R}_t.
\end{aligned}$$

So we have a reduced model in the sense of Definition 5.1. In this model, $\breve{\mathcal{F}}$ is generated by Y and $\breve{W}$, augmented by the null sets. Since $\breve{W}_s$ is a function of $\tilde{W}_s$ for all $s \leq t$ we have $\breve{\mathcal{F}}_t \subseteq \mathcal{F}_t$.

(ii) The reference measure $\breve{\tilde{P}}$ in the reduced model is given by, cf. (2.3),

$$\frac{\mathrm{d}\breve{\tilde{P}}}{\mathrm{d}P} = \breve{Z}_T, \quad \mathrm{d}\breve{Z}_t = -\breve{Z}_t Y_t^\top (\mathrm{d}^{-1} C)^\top \, \mathrm{d}\breve{W}_t.$$

Since $(\delta^{-1} C)^\top \breve{W}_t = (\sigma^{-1} B)^\top W_t$, we have by strong uniqueness or directly by the explicit representations of Z_T and $\breve{Z}_T$ that $Z_T = \breve{Z}_T$. Therefore, $\tilde{P}$ and $\breve{\tilde{P}}$ coincide on $\breve{\mathcal{F}}_T$.

(iii) By the Zakai equation (2.4), the unnormalized filter $\breve{\rho}$ in the reduced model satisfies $\breve{\rho}_0 = \mathrm{E}[Y_0]$ and

$$\begin{aligned} \mathrm{d}\breve{\rho}_t &= Q^\top \breve{\rho}_t \,\mathrm{d}t + \mathrm{Diag}(\breve{\rho}_t) C^\top D^{-1} \mathrm{d}\breve{R}_t \\ &= Q^\top \breve{\rho}_t \,\mathrm{d}t + \mathrm{Diag}(\breve{\rho}_t) B^\top \Sigma^{-1} \mathrm{d}R_t . \end{aligned}$$

Therefore, we have the same dynamics (under $\tilde{P}$ in the original model) and by strong uniqueness of this linear SDE we have that the continuous processes ρ and $\breve{\rho}$ are indistinguishable. Here we use that the corresponding reference measures $\tilde{P}$ and $\breve{\tilde{P}}$ are equivalent by (ii). By (2.7), also the corresponding normalized filters are indistinguishable. □

For C and δ as given in Definition 5.1 we now choose $\breve{W}$ as in Theorem 5.2 (i), i.e.,

$$\breve{W}_t := \delta^\top (C^{-1})^\top B^\top (\sigma^{-1})^\top W_t, \quad t \in [0, T],$$

and define $\breve{R}_t := \int_0^t C Y_s \,\mathrm{d}s + \delta \breve{W}_t$. Then the equivalence (5.3) holds pathwise.

Corollary 5.3 *We have $\mathcal{F}_t^{\breve{R}} \subseteq \mathcal{F}_t^R$ and $\mathcal{F}_t^{\breve{R}} = \mathcal{F}_t^{B^\top \Sigma^{-1} R}$ for $t \in [0, T]$. Further, the reference measures $\tilde{P}$ in the original model and $\breve{\tilde{P}}$ in the reduced model agree on $\mathcal{F}_T^{\breve{R}}$.*

Proof The inclusion $\mathcal{F}_t^{\breve{R}} \subseteq \mathcal{F}_t^R$ follows from $\breve{R}_s = D(C^\top)^{-1} B^\top \Sigma^{-1} R_s$, $s \leq t$. The equality of $\mathcal{F}_t^{\breve{R}}$ and $\mathcal{F}_t^{B^\top \Sigma^{-1} R}$ follows, since $D(C^\top)^{-1}$ is non-singular. The last statement is a consequence of Theorem 5.2 (ii) since $\mathcal{F}_T^{\breve{R}} \subseteq \breve{\mathcal{F}}_T$. □

Remark 5.4 In the case $d < n$ we clearly have strictly less information from observing $\breve{R}$ than from observing R: When observing $\breve{R}_t$ only, we can not distinguish between original returns R_t and $R_t + K_t$, where K_t lies in the kernel $\ker(B^\top \Sigma^{-1}) \subseteq \mathbb{R}^n$. This kernel is at least 1-dimensional and since A is assumed to have full rank in this section, it is in fact $(n - d)$-dimensional here. This is true for any choice of C and D according to Definition 5.1.

For example, in the simple case $n = 2$, $d = 1$, with diagonal σ and assuming $B_{21} \neq 0$, we would have that

$$\ker(B^\top \Sigma^{-1}) = \{(x, -B_{11}\sigma_{22}^2 x/(B_{21}\sigma_{11}^2))^\top \,:\, x \in \mathbb{R}\},$$

i.e. in the reduced model we cannot distinguish between original returns whose difference lies on the line $x \mapsto -B_{11}\sigma_{22}^2 x/(B_{21}\sigma_{11}^2)$.

Theorem 5.2 shows that interestingly the loss of information pointed out in Remark 5.4 does not affect the filter. Theorem 5.5 will show that this is also true for the optimal portfolio value.

By Theorem 5.2 and Corollary 5.3 we can from now on use the same notation for the filters $\hat{Y}$ and the unnormalized filters ρ in the original and the reduced models. We will also use the same notation $\tilde{P}$ for the reference measures, but have to keep in mind that these only agree on $\breve{\mathcal{F}}_T$ (and thus on $\mathcal{F}_T^{\breve{R}}$).

Theorem 5.5 *The optimal risky fraction process $\breve{\pi}^*$ for maximizing expected utility of terminal wealth for utility functions $U = U_\alpha$, $\alpha < 1$, leads to the same optimal wealth process as obtained in the original model, i.e.,*

$$\breve{X}_t^* = X_t^*, \quad t \in [0, T],$$

where $\breve{X}^*$ is the wealth process for $\breve{\pi}^*$ in the reduced model and X^* is the wealth process in the original model when following the optimal strategy π^* for maximizing expected utility of terminal wealth.

Proof Following the martingale approach, in the original model we have

$$X_T^* = I(\lambda \mathrm{E}[Z_T \mid \mathcal{F}_T^R]),$$

where $I = (U'_\alpha)^{-1}$ and $\lambda > 0$ is uniquely determined by $\mathrm{E}[Z_T X_T^*] = x_0$, cf. Sass and Haussmann (2004) and Theorem 3.1. Analogously, we obtain in the reduced model, since the Radon-Nikodym derivatives of the reference measures agree on $\mathcal{F}_t^{\breve{R}}$ by Theorem 5.2 (ii),

$$\breve{X}_T^* = I(\breve{\lambda} \mathrm{E}[Z_T \mid \mathcal{F}_T^{\breve{R}}]),$$

where $\breve{\lambda}$ is uniquely determined by $\mathrm{E}[Z_T \breve{X}_T^*] = x_0$.

By (2.5) we have $\mathrm{E}[Z_T \mid \mathcal{F}_T^R] = (\mathbf{1}^\top \rho_t)^{-1}$ and would get the analogous result in the reduced model, i.e., $\mathrm{E}[Z_T \mid \mathcal{F}_T^{\breve{R}}] = (\mathbf{1}^\top \breve{\rho}_t)^{-1}$. But by Theorem 5.2 (iii), ρ and $\breve{\rho}$ are the same, thus λ and $\breve{\lambda}$ are given uniquely by the same equation and hence agree. Therefore, $\breve{X}_T^* = X_T^*$.

This implies that also X_T^* is $\mathcal{F}_T^{\breve{R}}$-measurable and thus we get the same replicating strategies $\breve{\pi}^*$ and π^* by martingale representation. □

As outlined in the introduction there is a long history on mutual fund separation theorems. In continuous time, we could adapt Definition 2.4 for the more general model in Schachermayer et al. (2009) to our setting as follows: We say that the *mutual fund theorem holds* for a class of utility functions $\mathcal{U}$, if there exists a traded portfolio with values $M = (M_t)_{t \in [0,T]}$ and corresponding return process R^M such that for the optimal terminal wealth X_T^U under U there exists an $\mathcal{F}^R$-adapted, progressively measurable process η^U satisfying (for interest rate $r = 0$)

$$X_t^U = x_0 + \int_0^t \eta_s^U X_s^U \, \mathrm{d}R_s^M, \quad t \in [0, T]. \tag{5.4}$$

For example, in the Black-Scholes model, i.e. having constant parameters μ, σ in (2.1), $\Sigma = \sigma\sigma^\top$, we would get in the class of CRRA utility functions U_α the optimal $\pi_t^{U_\alpha} = \frac{1}{1-\alpha}\Sigma^{-1}\mu$ and

$$\mathrm{d}X_t^{U_\alpha} = X_t^{U_\alpha} (\pi_t^{U_\alpha})^\top \mathrm{d}R_t = \frac{1}{1-\alpha} X_t^{U_\alpha} \mu^\top \Sigma^{-1} \mathrm{d}R_t.$$

Therefore, the portfolio given by returns $\mathrm{d}R^M_t = \mu^\top \Sigma^{-1} \mathrm{d}R_t$ is the mutual fund, using $\eta^{U_\alpha}_t = \frac{1}{1-\alpha}$ here. So the mutual fund theorem holds in the Black-Scholes model in the class of CRRA functions. This was shown in Merton (1971) already. While we will refer to a representation like (5.4) in Remark 5.16 below, we want to point out here that in the following we think of building portfolios of mutual funds in the following sense: By (3.1) and Theorem 3.1 we have for logarithmic utility

$$\mathrm{d}X^*_t = X^*_t \hat{Y}^\top_t B^\top \Sigma^{-1} \mathrm{d}R_t = X^*_t \sum_{i=1}^{d} \hat{Y}^i_t \left(B^\top \Sigma^{-1} \mathrm{d}R_t\right)^i, \tag{5.5}$$

i.e. we can think of the optimal terminal wealth coming from investing in d mutual funds $\left(B^\top \Sigma^{-1} \mathrm{d}R_t\right)^i$. These are chosen with the weights $\hat{Y}^i_t$ which are the conditional probabilities for being in state i. So instead of using one mutual fund given by returns $\mathrm{d}R^M_t = \hat{Y}^\top_t B^\top \Sigma^{-1} \mathrm{d}R_t$, here we rather think of a representation as in (5.5) with d funds that have some correspondence to the states. In particular, in case of $\hat{Y} = \mathrm{e}_k$ this would boil down to investing in fund $b^\top \Sigma^{-1} R_t$, where b is row k of B. But the latter is the mutual fund in the Black-Scholes model with $\mu = b$. This way, we may think of (5.5) as a representations of mutual funds that would satisfy the mutual fund theorem in the degenerate cases that $\hat{Y}_t = \mathrm{e}_i$ for all t.

Our model reduction argument allows to introduce different decompositions into mutual funds $\breve{R}^1, \ldots, \breve{R}^d$ as discussed in the following remark. In Sect. 5.2 we introduce two canonical cases, one corresponding to (5.5) above.

Remark 5.6 (Interpretation of components of $\breve{R}$ as mutual funds) By (5.2) we can interpret the components of $\breve{R}$ as $d \le n$ asset returns which yield the same filters by Theorem 5.2 and lead to the same optimal portfolio value when building a portfolio only of these funds. Theorem 5.5 shows that it is sufficient to invest in risky assets $\breve{R}^1, \ldots, \breve{R}^d$. Note that these funds are given by

$$\breve{R}_t = D(C^\top)^{-1} B^\top \Sigma^{-1} R_t$$

which shows that the composition of fund i in terms of the original n assets is given by row i of $D(C^\top)^{-1} B^\top \Sigma^{-1}$. In particular, it is time-independent. This makes its interpretation more straightforward than a single mutual fund as in (5.4) with an (in our case) time-dependent composition.

For $U = \log$, we get from Theorem 3.1, replacing R, W, B, Σ, by $\breve{R}, \breve{W}, C, D$ the optimal risky fraction

$$\breve{\pi}^*_t = D^{-1} C \hat{Y}_t.$$

Since in the original model the optimal risky fraction is $\hat{\pi}^*_t = \Sigma^{-1} B \hat{Y}_t$ and since $B^\top \hat{\pi}^*_t = A \hat{Y}_t = C^\top \breve{\pi}^*_t$, we also have the representation

$$\breve{\pi}^*_t = (C^\top)^{-1} B^\top \hat{\pi}^*_t$$

which allows to compute the optimal fund investment directly from the optimal portfolio in the original model. Note that this does not work in the opposite direction, since B is not a square matrix for $d < n$. Clearly, several models can lead to the same reduced model but not vice versa.

Finally note that we have as many risky mutual funds as we have market regimes (states of the Markov chain), where the remainder is put in the riskfree asset. In particular, in case of $d = 1$ we obtain the classical mutual fund theorem in the Black–Scholes model.

5.2 Canonical cases of reduced models in the HMM

There are two canonical cases for choosing C and δ (or $D = \delta\delta^\top$) in (5.2). By its definition, according to (5.3) any choice has to satisfy

$$C^\top D^{-1} C = A = B^\top \Sigma^{-1} B, \tag{5.6}$$

$$\text{and } C^\top D^{-1} \mathrm{d}\breve{R}_t = B^\top \Sigma^{-1} \mathrm{d}R_t. \tag{5.7}$$

Definition 5.7 The reduced model is a *reduced regime representation model (RRRM)*, if the matrices C, D in Definition 5.1 are

$$C = A \quad \text{and} \quad D = A.$$

This RRRM yields returns

$$\mathrm{d}\breve{R}_t = B^\top \Sigma^{-1} \mathrm{d}R_t.$$

For $U = \log$ the optimal strategy $\breve{\pi}$ in the lower-dimensional model is

$$\breve{\pi}_t = D^{-1} C \hat{Y}_t = \hat{Y}_t.$$

Note that indeed (as proved in Theorem 5.5) the corresponding optimal wealth $\breve{X}$ is pathwise the same as in the original n-dimensional model, since

$$\mathrm{d}\breve{X}_t^* / \breve{X}_t^* = \breve{\pi}_t^\top \mathrm{d}\breve{R}_t = \hat{Y}_t^\top B^\top \Sigma^{-1} \mathrm{d}R_t = \mathrm{d}X_t^* / X_t^*.$$

The interpretation of the RRRM is according to Remark 5.6 that we invest in d funds, where fund k has returns evolving as $\breve{R}_t^k$ and a proportion $\hat{Y}_t^k = P(Y_t = \mathrm{e}_k \mid \mathcal{F}_t^R)$ of our money is invested in fund k. So in the degenerate case $\hat{Y}_t = \mathrm{e}_k$ we would invest everything in fund k. This extreme case cannot happen for $t > 0$, due to the dynamics of $\hat{Y}$. However, we see that the component $\breve{R}^k$ corresponds to the mutual fund k that would be optimal if we knew with certainty we were in state e_k.

The more involved case is the second one which is based on a principal component analysis of the signal-to-noise matrix A.

Definition 5.8 The reduced model is a *reduced eigenvalue model (REVM)*, if we choose C and D in Definition 5.1 as follows: Since A is non-singular and positive definite, we can decompose A as follows:

$$A = V\Lambda V^{\top}, \quad \text{where} \quad \Lambda = \text{Diag}(\lambda_1, \ldots, \lambda_d), \quad V = (v_1 \ldots v_d) \tag{5.8}$$

and V is orthogonal, $\lambda_1 \geq \cdots \geq \lambda_d > 0$, $Av_k = \lambda_k v_k$. So λ_k is the kth eigenvalue of A and v_k a corresponding eigenvector. Then we choose $C = V^{\top}$ and $D = \Lambda^{-1}$.

In the REVM we get returns

$$\mathrm{d}\breve{R}_t = \Lambda^{-1}V^{\top}B^{\top}\Sigma^{-1}\mathrm{d}R_t \quad \text{and} \quad \breve{\pi}_t = \Lambda V^{\top}\hat{Y}_t \quad \text{for} \quad U = \log. \tag{5.9}$$

Now the interpretation of the mutual funds represented by $\breve{R}$ is that we would invest optimally in the kth fund only if $\hat{Y}_t = v_k$. But the filter does not stay constant. In general, we invest $\breve{\pi}_t^i = \lambda_i v_i^{\top}\hat{Y}_t$ in $\breve{R}_t^i$, $i = 1, \ldots, d$. In the following, we will consider this setting in more detail.

5.3 Mutual funds in the REVM

While the RRRM has a straightforward interpretation, the REVM is more sophisticated. In the following, we shall therefore have a more detailed look at this decomposition. The results in the next section provide a more fundamental interpretation on the structure of the mutual funds in the REVM.

Example 5.9 (REVM for $n = 4$, $d = 3$) In this example we consider an HMM where the chain has 3 states. For the continuous-time model, returns from real-world applications are still sampled in discrete time. For discretizing the filters, as presented in Sect. 2.3, we use a robust scheme as introduced in James et al. (1996) (see also Sass and Haussmann (2004) for a multivariate version). We use the following rate matrix Q of the chain, which yields invariant distribution ν,

$$Q = \begin{pmatrix} -7 & 4 & 3 \\ 2 & -4 & 2 \\ 3 & 5 & -8 \end{pmatrix}, \quad \nu = \begin{pmatrix} 0.247 \\ 0.528 \\ 0.225 \end{pmatrix}.$$

For a 4-dim. return process we consider further

$$B = \begin{pmatrix} 1.2 & 0.1 & -1.0 \\ 1.0 & 0.1 & -0.8 \\ 0.5 & 0.01 & -0.4 \\ 0.6 & -0.01 & -0.3 \end{pmatrix}, \quad \sigma = \begin{pmatrix} 0.2 & 0.12 & 0.1 & 0.11 \\ 0.08 & 0.21 & 0.1 & 0.12 \\ 0.1 & 0.11 & 0.22 & 0.09 \\ 0.11 & 0.09 & 0.1 & 0.2 \end{pmatrix}.$$

Note that the average values for the drift are $B\nu = (0.12, 0.12, 0.04, 0.08)^{\top}$ which lie in the range of market data, as does the volatility matrix σ. The values for Q are chosen

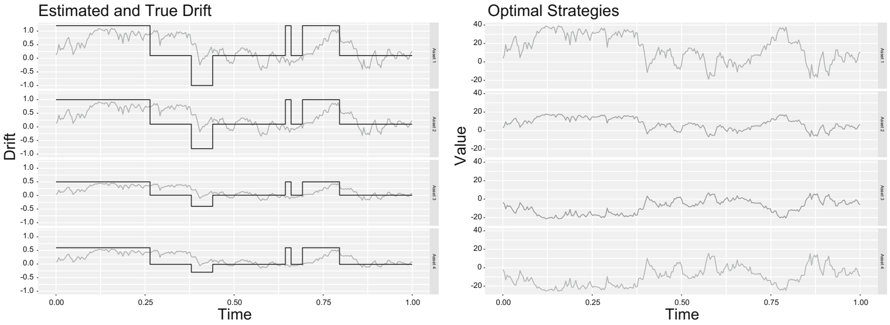

Fig. 2 Filters and strategy in full model in Example 5.9

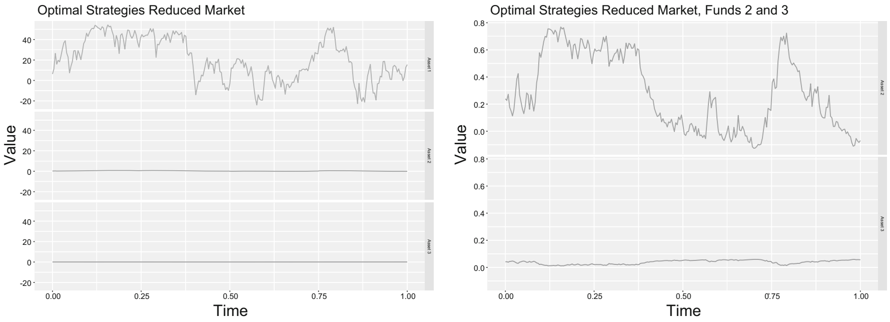

Fig. 3 Strategy in reduced model (left hand: all three components, right hand: zoom in to second and third component) in Example 5.9

such that we see a suitable number of jumps in the graphs, but Q has no influence on the signal-to-noise matrix and thus is irrelevant for the subsequent results.

An example path of the filters and the log-optimal strategy in the full model is given in Fig. 2. In Fig. 3 we see the optimal strategies in the reduced model. It can be seen that investment in the first fund fluctuates much more than in the other funds, contrary to the strategy in the full model. There, the wealth invested in the single assets fluctuates for all assets.

Proposition 5.10 *Consider the returns $\breve{R}$ in the REVM.*

(i) We have $\mathrm{E}[\breve{R}^i_t] = \nu^\top v_i\, t$,

$$\operatorname{Var}(\breve{R}^i_t) = \lambda_i^{-1}\, t + 2 \sum_{k,l=1}^{d} (v_i^k - \nu^\top v_i)(v_i^l - \nu^\top v_i)\nu_k \int_0^t \int_0^s \left(\mathrm{e}^{Q(s-u)}\right)_{kl} \mathrm{d}u\, \mathrm{d}s,$$

and $\breve{R}^1_t, \dots, \breve{R}^d_t$ are independent conditionally on Y.

(ii) For the optimal strategy $\breve{\pi} = \breve{\pi}^*$ for $U = \log$ we have $\breve{\pi}^*_t = \Lambda V^\top \hat{Y}_t$ and thus

$$\breve{\pi}^i_t \in [-\lambda_i, \lambda_i], \quad i = 1, \dots, d.$$

(iii) For $U = \log$ the optimal value is

$$\mathrm{E}[\log(\breve{X}_T^*)] = \log(x_0) + \frac{1}{2}\sum_{i=1}^{d} \lambda_i v_i^\top \left(\int_0^T \mathrm{E}\left[\hat{Y}_t \hat{Y}_t^\top\right] \mathrm{d}t\right) v_i.$$

Proof In the reduced eigenvalue model we have

$$\mathrm{d}\breve{R}_t = V^\top Y_t \,\mathrm{d}t + \Lambda^{-1/2}\,\mathrm{d}\breve{W}_t$$

for the diagonal matrix Λ of the eigenvalues $\lambda_k > 0$ and for the matrix V of eigenvectors $V = (v_1 \dots v_d)$. Therefore

$$\mathrm{d}\breve{R}_t^k = v_k^\top Y_t \,dt + \lambda_k^{-1/2}\mathrm{d}\breve{W}_t^k.$$

Using the independence of Y and $\breve{W}$, which follows from the independence of Y and W, and that $\mathrm{E}[Y_t] = \nu$ by starting with the invariant distribution ν, the claims in (i) can be derived straightforwardly (cf. Elliott et al. 2008). For (ii) note that by (5.9) we have $\breve{\pi}_t = \Lambda V^\top \hat{Y}_t$ and thus $\breve{\pi}_t^k = \lambda_k v_k^\top \hat{Y}_t$. This yields

$$|\breve{\pi}_t^k| = \lambda_k |v_k^\top \hat{Y}_t| \le \lambda_k \sqrt{\|v_k\|^2 \|\hat{Y}_t\|^2} \le \lambda_k \tag{5.10}$$

since $\|v_k\| = 1$ and $\|\hat{Y}_t\| \le 1$.

(iii) By Theorem 5.5 we have $\breve{X}_T^* = X_T^*$ and thus by (4.1)

$$\mathrm{E}[\log(\breve{X}_T^*)] = \log(x_0) + \frac{1}{2}\mathrm{E}\left[\int_0^T \hat{Y}_t^\top A \hat{Y}_t \mathrm{d}t\right].$$

Computing

$$\hat{Y}_t^\top A \hat{Y}_t = \hat{Y}_t^\top V \Lambda V^\top \hat{Y}_t = \sum_{i=1}^{d} \lambda_i \left(v_i^\top \hat{Y}_t\right)^2$$

and applying the Tonelli theorem yields the result. □

Using Proposition 5.10 (iii), one can compute directly (using e.g. the Cauchy-Schwarz inequality)

Corollary 5.11 *The optimal wealth under $U = \log$ satisfies*

$$\log(x_0) + \frac{1}{2}T\sum_{i=1}^{d} \nu_i A_{ii} \ge \mathrm{E}[\log(\breve{X}_T^*)] \ge \log(x_0) + \frac{1}{2}T\,\nu^\top A \nu.$$

The bounds in Corollary 5.11 have the following interpretation. Remember that $\breve{X}_T^*$ corresponds to investing according to the optimal $\breve{\pi}_t^* = \Lambda V^\top \hat{Y}_t$ under partial information in the reduced model. The upper bound corresponds to the optimal value for an investor with full information who uses $\breve{\pi}_t^{\text{full}} = \Lambda V^\top Y_t$ in the reduced model, and the lower bound to an investor which uses no further information and just invests according to the strategy based on the mean of the trend, i.e. $\breve{\pi}_t^{\text{no-info}} = \Lambda V^\top \nu$.

In the original model, we would obtain the same optimal value by Theorem 5.5 using the optimal strategy $\pi_t^* = (\sigma\sigma^\top)^{-1} B\hat{Y}_t$ for partial information and the same bounds corresponding to full information by $\pi_t^{\text{full}} = (\sigma\sigma^\top)^{-1} B\hat{Y}_t$ and $\pi_t^{\text{no-info}} = (\sigma\sigma^\top)^{-1} B\nu$ for assuming constant parameters.

Remark 5.12 In case of the REVM, we can now be more specific on the structure of the d mutual funds than in Remark 5.6. Remember that $\lambda_1 \geq \cdots \geq \lambda_d$ are the eigenvalues of the signal-to-noise matrix A. The funds are ordered according to the sizes of the eigenvalues. Proposition 5.10(i) shows that the variances are decreasing of order t (the second term is of order t^2). By part (ii) this leads to a possibly higher investment in fund i than in fund j if $i < j$, since the invested fraction in fund i is bounded by λ_i. So a fund with lower index may lead to a higher investment and thus is more attractive. From the estimate (5.10) we see that this happens in particular if $\|\hat{Y}_t\|$ is close to 1, which is the case if and only if $\hat{Y}_t \approx \mathrm{e}_k$ for some k, i.e. if the filter is quite informative. Because the funds are conditionally independent by part (i), for diversification we should still invest in all funds. This will always happen, since $\hat{Y}_t = \mathrm{e}_k$ is not possible due to the dynamics of the filters.
These relations also imply that using an approximate strategy by investing only $\pi_t^1 = \breve{\pi}_t^1$ and $\pi_t^i = 0$ for $i = 2, \ldots, d$, i.e. investing in fund 1 only, we can get a quite good approximation to the optimal value as long as $\lambda_1 \gg \lambda_2$. This is rather the typical case since the eigenvalues result from a principal component analysis providing the most influential investment direction.
The approach in the REVM is similar to the idea of eigenportfolios. While we decompose the signal-to-noise matrix, these are based on a decomposition of the correlation matrix and choosing the portfolio corresponding to the principal eigenvector, see e.g. Avellaneda et al. (2021), Avellaneda and Lee (2010), Boyle (2014), Chen and Yuan (2016) and the references therein.

Example 5.13 Let us illustrate the preceding remark. In the setting of Example 5.9, the signal-to-noise matrix is

$$A = \begin{pmatrix} 42.13709 & 6.211840 & -40.350279 \\ 6.21184 & 1.128064 & -6.501476 \\ -40.35028 & -6.501476 & 40.879834 \end{pmatrix},$$

with eigenvalues $\lambda_1 = 82.85$, $\lambda_2 = 1.22$, $\lambda_3 = 0.07$.

So λ_1 is clearly dominant, which leads to the strong investment into the first fund as seen in Fig. 3. λ_2 is much smaller than λ_1, but still of higher order than λ_3, thus we see more investment into the second fund compared to the third.

Let us point out one relation to the convergence result Theorem 4.3. Note first that trivially, when the largest eigenvalue λ_1 of the signal-to-noise matrix A converges to

0 we have by (4.1) that the optimal expected utility converges to $\log(x_0)$, i.e., there is approximately no gain from investing in the stocks. However, based on the results of this section, we can utilize Theorem 4.3 for a more subtle argument on the relation of the funds in the reduced model as we outline in the following remark.

Remark 5.14 Let us consider a sequence of models with signal-to-noise matrices $A^{(n)} = \frac{1}{n}A = B^\top \Sigma^{-1} B$. Then the eigenvalues satisfy $\lambda_i^{(n)} = \frac{1}{n}\lambda_i$ but the eigenvectors remain unchanged, i.e., $\Lambda^n = \frac{1}{n}\Lambda$, $V^{(n)} = V$ in the decomposition (5.8) of $A^{(n)}$.

As discussed in Remark 5.12 the REVM decomposition may be used to invest only in the first k, $k < d$, portfolios of the decomposition. Analogously to Proposition 5.10 (ii) this would yield an expected utility of

$$\log(x_0) + \frac{1}{2n}\sum_{i=1}^{k} \lambda_i v_i^\top \left(\int_0^T \mathrm{E}\left[\hat{Y}_t^{(n)} (\hat{Y}_t^{(n)})^\top \right] \mathrm{d}t \right) v_i, \tag{5.11}$$

where $\hat{Y}^{(n)}$ are the filters computed in the model with $A^{(n)}$ and where we used $\lambda_i^{(n)} = \frac{1}{n}\lambda_i$. We can compare this with the performance of the portfolio not taking the information into account, i.e. using the strategy $\breve{\pi}_t^{\text{no-info}} = \Lambda V^\top \nu$ as discussed after Corollary 5.11, leading to expected utility

$$\log(x_0) + \frac{T}{2}\,\nu^\top A^{(n)} \nu = \log(x_0) + \frac{T}{2n}\,\nu^\top A \nu = \log(x_0) + \frac{T}{2n}\sum_{i=1}^{d} \lambda_i v_i^\top \nu \nu^\top v_i \tag{5.12}$$

as stated in that corollary. Example 5.13 shows that often the performance in (5.11) for $k < d$ can be expected to be better than in (5.12). Formally, by taking derivatives in (5.11) and (5.12) we see that this is true if

$$\sum_{i=1}^{k} \lambda_i v_i^\top \mathrm{E}\left[\hat{Y}_T^{(n)} (\hat{Y}_T^{(n)})^\top \right] v_i > \sum_{i=1}^{d} \lambda_i v_i^\top \nu \nu^\top v_i$$

which is equivalent to

$$\sum_{i=1}^{k} \lambda_i v_i^\top \left(\mathrm{E}\left[\hat{Y}_T^{(n)} (\hat{Y}_T^{(n)})^\top \right] - \nu\nu^\top \right) v_i > \sum_{j=k+1}^{d} \lambda_j v_j^\top \nu \nu^\top v_j. \tag{5.13}$$

Now the left-hand side converges to 0 due to the L^2-convergence in Theorem 4.3 while the right-hand side is strictly positive and constant in n. This means that, if the eigenvalues become too low, then a portfolio of a strict subset of the mutual funds can no longer dominate the constant portfolio which takes no information via filtering into account.

5.4 Singular signal-to-noise matrix in HMM

Let A now be singular. A typical example is when $d > n$, i.e., we have a Markov chain with more states than risky assets. This already occurs when we have only one risky asset and consider $d \geq 2$ market regimes.

The RRRM as in Definition 5.7 does not work in this case since it requires A to be non-singular for computing its inverse. Also using the REVM directly as in Definition 5.8 does not work since it uses that the diagonal matrix Λ of the eigenvalues is non-singular – but now at least one eigenvalue is 0.

Utilizing the same idea, we now try to reduce the model to a dimension corresponding to the number of strictly positive eigenvalues. More precisely, having p strictly positive eigenvalues, $1 \leq p < d$, we can order the eigenvalues of A as follows (A is at least positive semi-definite)

$$\lambda_1 \geq \ldots \geq \lambda_p > \lambda_{p+1} = \ldots = \lambda_d = 0 \quad \text{for some} \quad p < d,$$

i.e., we assume that A has rank $1 \leq p < d$. We denote as in Definition 5.8

$$\Lambda = \mathrm{Diag}(\lambda_1, \ldots, \lambda_d), \quad V = (v_1 \ldots v_d),$$

where v_i is a normalized eigenvector for λ_i and V is orthogonal. Then, we have $A = V\Lambda V^\top$ as in Definition 5.8 again, but we can not proceed as we did in Theorem 5.2. Λ is singular, so we can not define $\delta\delta^\top$ by Λ^{-1} in order to introduce $\breve{W}$ as we did in that theorem.

Instead, reducing the dimension even further to p, we set

$$\Lambda^{(p)} := \mathrm{Diag}(\lambda_1, \ldots, \lambda_p) \in \mathbb{R}^{p\times p}, \quad V^{(p)} = (v_1 \ldots v_p) \in \mathbb{R}^{d\times p}.$$

Then we have

$$A = V\Lambda V^\top = V^{(p)}\Lambda^{(p)}(V^{(p)})^\top.$$

Now we can define a Brownian motion

$$\breve{W}_t^{(p)} = \left(\Lambda^{(p)}\right)^{-1/2} \left(V^{(p)}\right)^\top B^\top \left(\sigma^\top\right)^{-1} W_t, \quad t \in [0, T].$$

The same arguments as in the proof of Theorem 5.2 show that $\breve{W}^{(p)}$ is a p-dimensional Brownian motion. For the p-dimensional model with returns

$$\mathrm{d}\breve{R}_t^{(p)} = (V^{(p)})^\top Y_t \mathrm{d}t + (\Lambda^{(p)})^{-1/2}\mathrm{d}\breve{W}_t^{(p)}$$

we then have

$$B^\top \Sigma^{-1} dR_t = V^{(p)}\Lambda^{(p)} d\breve{R}_t^{(p)} + (I_d - V^{(p)}(V^{(p)})^\top)B^\top(\sigma^\top)^{-1} dW_t. \tag{5.14}$$

Theorem 5.15 *For maximizing power or logarithmic utility we get the same optimal terminal wealth as in the original model from investing in only p mutual funds with returns $R^{(p)}$, where the optimal strategy for $U = \log$ is*

$$\breve{\pi}_t^{(p)} = \Lambda^{(p)} \left(V^{(p)}\right)^\top \hat{Y}_t.$$

But to compute the filters and thus the optimal strategy $\breve{\pi}^{(p)}$ we need the observation from all d assets on the right-hand side of (5.14)*. In particular, $\breve{\pi}^{(p)}$ is in general not $\mathcal{F}^{R^{(p)}}$-adapted.*

Proof Adapting the argument in the proof of Theorem 5.2, by (5.14) we get the same filters when we use the whole information from the right-hand side. Then we can also use the arguments in the proof of Theorem 5.5 to conclude that the optimal wealth processes agree. □

Note that the second term in (5.14) really adds information. We can see this by rewriting (5.14) as

$$\begin{aligned} B^\top \Sigma^{-1} dR_t = & \; AY_t \mathrm{d}t + V^{(p)}(V^{(p)})^\top B^\top (\sigma^\top)^{-1} \mathrm{d}W_t \\ & + (I_d - V^{(p)}(V^{(p)})^\top) B^\top (\sigma^\top)^{-1} \mathrm{d}W_t \end{aligned}$$

and noting that $V^{(p)}(V^{(p)})^\top$ and $I_d - V^{(p)}(V^{(p)})^\top$ are orthogonal.

5.5 Model reduction in MSM and filter-based HMM

Consider the MSM as we introduced it in Sect. 2.2,

$$\mathrm{d}R_t = BY_t \,\mathrm{d}t + \sigma(Y_t)\mathrm{d}W_t,$$

where $\sigma(e_1), \ldots, \sigma(\mathrm{e}_d)$ are pairwise different. As discussed there, Y_t then is observable from the returns, i.e. it is $\mathcal{F}_t^R$-measurable, and thus we know the current parameters $B\mathrm{e}_k, \sigma(\mathrm{e}_k)$ if $Y_t = \mathrm{e}_k$, cf. Krishnamurthy et al. (2018). For optimization problem (3.2) with $U = U_\alpha$, the optimal strategies are of the form $\pi_t^* = \frac{1}{1-\alpha} \Sigma(Y_t)^{-1} BY_t$, see Theorem 3.2.

The signal-to-noise matrix then depends on time via Y_t,

$$A(Y_t) = B^\top \Sigma^{-1}(Y_t) B \quad \text{for} \quad \Sigma(Y_t) = \sigma(Y_t)(\sigma(Y_t))^\top.$$

We can apply the same decompositions as in the HMM, but now depending on Y. Since Y is observable, this can be calculated based on the available information. We discuss the details in the following two remarks.

Remark 5.16 (Mutual funds and RRRM in the MSM) In analogy to the RRRM in Definition 5.7 we can introduce a d-dimensional reduced MSM by

$$\mathrm{d}\breve{R}_t = A(Y_t)Y_t \mathrm{d}t + \delta(Y_t)\mathrm{d}\breve{W}_t = B^\top \Sigma(Y_t)^{-1} \mathrm{d}R_t,$$

where $\delta(Y_t)$ is some square root of $A(Y_t)$ and $\breve{W}_t = \delta(Y_t)^{-1}B^\top(\sigma(Y_t)^{-1})^\top$. Applying Theorem 3.2 to the reduced MSM, we have

$$\breve{\pi}_t = (\delta(Y_t)\delta(Y_t)^\top)^{-1}A(Y_t)Y_t = Y_t.$$

Therefore, just as in the RRRM for the HMM, fund $\breve{R}_t^k$ will be chosen for $Y_t = e_k$. While we speak here of d funds, note that the MSM allows for the two-fund separation in the sense of Schachermayer et al. (2009, Definition 2.4) in the class of CRRA utility functions, cf. (5.4) above. Indeed, according to Theorem 3.2

$$\pi_t^* = \frac{1}{1-\alpha}\Sigma^{-1}(Y_t)BY_t, \quad \mathrm{d}X_t^* = X_t^*\pi_t^*\mathrm{d}R_t = \frac{1}{1-\alpha}X_t^*Y_t^\top B^\top\Sigma^{-1}(Y_t)\mathrm{d}R_t,$$

where the latter part $Y_t^\top B^\top\Sigma^{-1}(Y_t)\mathrm{d}R_t$ would correspond to the risky portfolio in which the fraction $\eta = \frac{1}{1-\alpha}$ of the current wealth would be invested.

Remark 5.17 (REVM in the MSM) In the MSM, in analogy to the REVM in Definition 5.8, we can also introduce for each value e_k of Y_t, $k = 1, \ldots, d$, an eigenvalue decomposition

$$A(\mathrm{e}_k) = V(\mathrm{e}_k)\,\Lambda(\mathrm{e}_k)V(\mathrm{e}_k)^\top,$$

where

$$\Lambda(\mathrm{e}_k) = \mathrm{Diag}(\lambda_1^{(k)}, \ldots, \lambda_d^{(k)}), \quad V(\mathrm{e}_k) = (v_1^{(k)} \ldots v_d^{(k)})$$

and $V(\mathrm{e}_k)$ is orthogonal, $\lambda_1^{(k)} \geq \cdots \geq \lambda_d^{(k)} > 0$, $Av_i^{(k)} = \lambda_i^{(k)}v_i^{(k)}$. Then we have

$$\mathrm{d}\breve{R}_t = \Lambda(Y_t)^{-1}V(Y_t)^\top B^\top\Sigma^{-1}\mathrm{d}R_t = V(Y_t)^\top Y_t\mathrm{d}t + \Lambda(Y_t)^{-1/2}\mathrm{d}\breve{W}_t$$

for a suitable Brownian motion. By Theorem 3.2 we then have optimal

$$\breve{\pi}_t = \Lambda(Y_t)V(Y_t)^\top\, Y_t \tag{5.15}$$

for logarithmic utility. But in fact, in the MSM we can simplify (5.15) considerably since Y_t is a unit vector and $\Lambda(\mathrm{e}_k)$ is diagonal, yielding

$$\breve{\pi}_t = M\,Y_t, \quad \text{where} \quad M_{ik} = \lambda_i^{(k)}(v_k^{(k)})_i.$$

In terms of mutual funds, we can look at this result as setting up, for each state of Y_t, d funds ordered according to the eigenvalues of $A(Y_t)$. Then M_{ik} provides the fraction to be invested in fund i if the chain is in state $Y_t = \mathrm{e}_k$, of the d funds for this state.

So in both cases we have a decomposition independent of time t, choosing the corresponding funds based on the observable state of Y_t.

Typically a continuous-time MSM has better econometric properties than the HMM, e.g. it allows for volatility clustering. The reason for considering a continuous-time MSM (or HMM) is that we obtain more explicit results than in discrete time, e.g. like in Theorems 3.1 and 3.2, where corresponding discrete-time results are not explicit. But a continuous-time MSM may be a poor approximation for a discrete-time MSM in the sense that in continuous time no filtering problem exists. But in the discrete-time MSM the underlying Markov chain cannot be observed and its states have to be estimated by the corresponding filters. Therefore, solving e.g. portfolio optimization problems in the continuous-time MSM as in Theorem 3.2 provides a poor approximation for the discrete-time model. This is not the case for the HMM, where the discretization of continuous-time filters yields the filters which are optimal in the corresponding discrete-time HMM, see James et al. (1996).

This motivates to introduce a *HMM with non-constant volatility* as approximation for the MSM,

$$\mathrm{d}R_t = BY_t\mathrm{d}t + \sigma_t\,\mathrm{d}W_t, \quad \sigma_t = f(\hat{Y}_t).$$

This yields consistent approximations since the filtering problem is non-trivial. Filters can be computed similar as in Sect. 2.3. A choice of f which satisfies $f(\mathrm{e}_k) = \sigma(\mathrm{e}_k)$ for $k = 1, \dots, d$ is possible, cf. Krishnamurthy et al. (2018) for details in the one-dimensional case.

The best model in an MSE-sense is

$$f(\hat{Y}_t) = \sum_{k=1}^{d} \sigma(\mathrm{e}_k)\hat{Y}_t^k. \tag{5.16}$$

Using this parametrization, we speak of the *filter-based hidden Markov model (FB-HMM)*.

Remark 5.18 (Optimization and mutual funds in the FB-HMM) Portfolio optimization also works for the FB-HMM (5.16), cf. Haussmann and Sass (2004), where filtering would have to be addressed as in Krishnamurthy et al. (2018). A reduced model can be introduced, but due to the relation (5.16) we have a signal-to-noise matrix $A_t = B^\top(f(\hat{Y}_t)f(\hat{Y}_t)^\top)^{-1}B$ depending on the filter $\hat{Y}_t$. Thus, in the reduced model the composition of the mutual funds would then also depend on time and state via the filter. This could still be used to define an RRRM similar as in Definition 5.7, yielding optimal $\breve{\pi}_t = \hat{Y}_t$, but, as pointed out above, the composition of the funds would depend on the filter value via

$$\mathrm{d}\breve{R}_t = B^\top(f(\hat{Y}_t)f(\hat{Y}_t)^\top)^{-1}\mathrm{d}R_t.$$

For the FB-HMM also other decompositions of A_t are reasonable, but they would also be filter-dependent.

In summary, we have seen that for the HMM we have a very good interpretation of the reduced models in terms of d funds which have a time-independent composition,

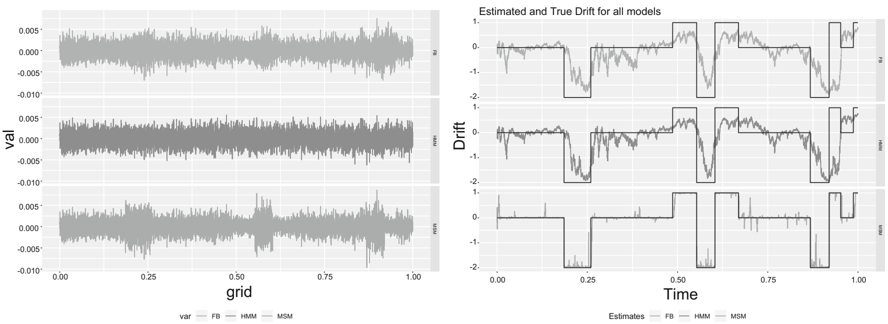

Fig. 4 HMM, FB-HMM and MSM: Returns and Filters in Example 5.19

while for the MSM this composition depends on the state of Y_t and for the FB-HMM on the filter $\hat{Y}_t$. We shall close this section with a comparison of the three models which makes in particular evident why there is no reasonable other choice for the MSM than the one in Remark 5.16.

Example 5.19 (Filters in HMM, discrete-time MSM and FB-HMM) For all three models, HMM, FB-HMM, MSM, we use $d = 3$ states and the rate matrix Q from Example 5.9. But we look at $n = 1$ only, using as state matrix for the drift $B = (1, 0, -2)$. The vector of volatilities for the MSM and the FB-HMM are $\sigma^{MSM}(\mathrm{e}_1) = 0.1, \sigma^{MSM}(\mathrm{e}_2) = 0.15, \sigma^{MSM}(\mathrm{e}_3) = 0.25$, and $\sigma^{FB}(\mathrm{e}_k) = \sigma^{MSM}(\mathrm{e}_k)$, while as constant volatility in the HMM we use the average over the invariant distribution, i.e. $\sigma^{HMM} = \sum_{k=1}^{3} \sigma^{MSM}(\mathrm{e}_k)\nu_k = 0.16$.

Also in the continuous-time model the returns are sampled in discrete time, as is typical for real-world applications. For filtering in the MSM in discrete time we refer to Elliott et al. (1995), for the HMM to Sect. 2.3 and for filters in the FB-HMM see Krishnamurthy et al. (2018).

The returns, the drift and the filter for the drift, $B\hat{Y}_t$, are plotted in Fig. 4 for one simulation of Y and W. We see clearly, that the filter in the MSM provides the true state quite exactly as we expected since in continuous time the chain is observable. Therefore, the only reasonable decomposition into funds is the regime parametrization as discussed in Remark 5.16 since by $\breve{\pi}_t = Y_t$ it leads to choosing the best portfolio in the current state. One also sees that the FB-HMM lies regarding filtering and volatility clustering between both models and just provides a good compromise between the more realistic filtering in the HMM and the more realistic econometric properties of the MSM.

6 Conclusion

In the context of hidden Markov models we showed that the signal-to-noise matrix plays a prominent role for portfolio optimization as well as for filtering. The convergence result in Chapter 4 gives an exact formulation of the intuition that we can retrieve less information on the underlying chain from observing the stock prices when the signal (drift) is small compared to the noise (volatility). This is shown by proving that for decreasing eigenvalues of the signal-to-noise matrix, the filters converge uniformly in L^2 to the invariant distribution of the chain. Since the latter is the distribution of the chain, we gain no additional information in the limit.

The important role of the signal-to-noise matrix, which is of dimension d, and of its d eigenvalues then motivated us to reduce the dimension (if $d \leq n$) of the model by decomposing the signal-to-noise matrix and setting up a d-dimensional model based on this decomposition in (5.3). The returns in the reduced model can be seen as d mutual funds. We proved that portfolio optimization and filtering in the reduced model yields pathwise the same optimal wealth and filter processes.

Two special cases were introduced, using in the RRRM a decomposition which yields the optimal portfolios in the single states as funds, and in the REVM an eigenvalue decomposition which leads to funds which contribute according to the corresponding eigenvalues more or less to the optimal portfolio.

To complete the survey we looked at the case of a singular signal-to-noise matrix (e.g. when $d > n$) and at model reduction in related models as the MSM and the filter-based HMM.

Our analysis showed that in the standard case of a non-singular signal-to-noise matrix in the HMM, while there is less information in the reduced model from observing the funds only, the filters and the optimization based on this observation still yield the same results. For future research it would be interesting to analyze further if this reduction helps to identify relevant model parameters better. Also the effect of including expert opinions, see e.g. Frey et al. (2012), on the model reduction would be of interest.

Funding Open Access funding enabled and organized by Projekt DEAL.

A Appendix: proofs of convergence results

A.1 Proof of Theorem 4.2

For $t > 0$

$$\begin{aligned}\|X_t^n - X_t\| &= \|\int_0^t a(X_s^n) - a(X_s)\mathrm{d}s + \int_0^t b^n(X_s^n)\mathrm{d}W_s^n\| \\ &\leq \|\int_0^t a(X_s^n) - a(X_s)\mathrm{d}s\| + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\| \\ &\leq \int_0^t \|a(X_s^n) - a(X_s)\|\,\mathrm{d}s + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\| \\ &=: M_t^n.\end{aligned}$$

Since b^n is bounded $\int_0^t b^n(X_s^n)\mathrm{d}W_s^n$ is a true martingale and we can show that M^n is a submartingale: For $t > u$

$$\begin{aligned}\mathrm{E}\big[M_t^n|\mathcal{F}_u\big] &= \int_0^u \|a(X_s^n) - a(X_s)\|\,\mathrm{d}s + \mathrm{E}\big[\int_u^t \|a(X_s^n) - a(X_s)\|\,\mathrm{d}s|\mathcal{F}_u\big] \\ &\quad + \mathrm{E}\big[\|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\||\mathcal{F}_u\big] \\ &\geq \int_0^u \|a(X_s^n) - a(X_s)\|\,\mathrm{d}s + \mathrm{E}\big[\|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\||\mathcal{F}_u\big] \\ &\geq \int_0^u \|a(X_s^n) - a(X_s)\|\,\mathrm{d}s + \|\mathrm{E}\big[\int_0^t b^n(X_s^n)\mathrm{d}W_s^n|\mathcal{F}_u\big]\| \\ &= \int_0^u \|a(X_s^n) - a(X_s)\|\,\mathrm{d}s + \|\int_0^u b^n(X_s^n)\mathrm{d}W_s^n\| = M_u^n.\end{aligned}$$

By Doob's inequality for M^n and $p = 2$ it follows that

$$\mathrm{E}[\sup_{s\leq t}\|X_s^n - X_s\|^2] \leq \mathrm{E}[(\sup_{s\leq t}\|M_s^n\|)^2] \leq 4\mathrm{E}[(M_t^n)^2].$$

Let L be the Lipschitz constant of a, then

$$\begin{aligned}
\mathrm{E}[(M_t^n)^2] &= \mathrm{E}\big[\big(\int_0^t \|a(X_s^n) - a(X_s)\| \,\mathrm{d}s + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|\big)^2\big] \\
&\leq 4\mathrm{E}\big[\big(\int_0^t \|a(X_s^n) - a(X_s)\| \,\mathrm{d}s\big)^2 + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|^2\big] \\
&\leq 4\mathrm{E}\big[\int_0^t \|a(X_s^n) - a(X_s)\|^2 \,\mathrm{d}s + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|^2\big] \\
&\leq 4\mathrm{E}\big[\int_0^t L^2 \|X_s^n - X_s\|^2 \,\mathrm{d}s + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|^2\big] \\
&\leq 4\mathrm{E}\big[\int_0^t L^2 \sup_{u\leq s} \|X_u^n - X_u\|^2 \,\mathrm{d}s + \|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|^2\big] \\
&= 4\int_0^t L^2\mathrm{E}[\sup_{u\leq s} \|X_u^n - X_u\|^2]\mathrm{d}s + 4\mathrm{E}\big[\|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|^2\big].
\end{aligned}$$

Using the Ito-isometry componentwise we see that

$$\mathrm{E}\big[\|\int_0^t b^n(X_s^n)\mathrm{d}W_s^n\|^2\big] = \mathrm{E}\big[\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s\big]$$

which implies

$$\mathrm{E}[\sup_{s\leq t} \|X_s^n - X_s\|^2] \leq 16\int_0^t L^2\mathrm{E}[\sup_{u\leq s} \|X_u^n - X_u\|^2]\mathrm{d}s + 16\mathrm{E}\big[\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s\big].$$

Since $\mathrm{E}\big[\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s\big]$ is non-decresaing in t we can apply Gronwall's inequality to see that

$$\mathrm{E}[\sup_{s\leq t} \|X_s^n - X_s\|^2] \leq \exp\big(16L^2 t\big)16\mathrm{E}\big[\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s\big].$$

Now $\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s$ converges to 0 a.s. and $b^n(X_s^n)$ is bounded, so by dominated convergence we can conclude that

$$\lim_{n\to\infty} \mathrm{E}\big[\int_0^t \|b^n(X_s^n)\|_{dm}^2 \,\mathrm{d}s\big] = 0$$

and thus

$$\lim_{n\to\infty} \mathrm{E}[\sup_{s\leq t} \|X_s^n - X_s\|^2] = 0.$$

A.2 Appendix: Proof of Theorem 4.3

We want to apply Theorem 4.2 for

$$X_t^m := \hat{Y}_t^m \qquad \text{and} \qquad X_t := \nu$$

satisfying $\mathrm{d}X_t = Q^T \nu \mathrm{d}t = Q^T X_t \mathrm{d}t$. Lipschitz-continuity of the drift is fulfilled since it is linear and what is left to check are the assumptions on the functions b^m.

Set $P_m := \sigma_m^{-1} B_m \in \mathbb{R}^{n\times d}$, so $(g_i^m)^T = (p_{i1}^m, \ldots, p_{id}^m)$, i.e. $(g_i^m)_j = p_{ij}^m$. Then $\|P_m\| = \lambda_m$ with $\|\cdot\|$ the norm on $\mathbb{R}^{d\times n}$ induced by the euclidean norms on $\mathbb{R}^d$ and $\mathbb{R}^n$. By assumption $\lim_{m\to\infty} \lambda_m = 0$, thus also

$$\lim_{m\to\infty} \max_{i,j} \left| p_{ij}^m \right|^2 = 0$$

since all matrix norms are equivalent.

Consider for $i = 1, \ldots, d$ the i^{th} component of the diffusion part

$$\begin{aligned}
&\sum_{j=1}^{n} \big((g_j^m)_i (X_t^m)_i - \sum_{k=1}^{d} (g_j^m)_k (X_t^m)_k (X_t^m)_i\big) \mathrm{d}(V_t^m)_j \\
&\quad = \sum_{j=1}^{n} \big(p_{ji}^m (X_t^m)_i - \sum_{k=1}^{d} p_{jk}^m (X_t^m)_k (X_t^m)_i\big) \mathrm{d}(V_t^m)_j \\
&\quad = \sum_{j=1}^{n} b_{ij}(X_t^m) \mathrm{d}(V_t^m)_j = \big(b^m(X_t^m)\mathrm{d}V_t^m\big)_i
\end{aligned}$$

with

$$b_{ij}^m(x) := p_{ji}^m x_i - \sum_{k=1}^{d} p_{jk}^m x_k x_i$$

for $x \in \mathbb{R}^d$. Using the matrix valued function b^m we can reformulate the SDE as

$$\mathrm{d}X_t^m = Q^T X_t^m \mathrm{d}t + b^m(X_t^m)\mathrm{d}V_t^m.$$

Now for $x \in \mathbb{R}^d$ with $\|x\| \leq 1$

$$\begin{aligned}
\left\|b^m(x)\right\|_{dn}^2 &= \sum_{i=1}^d \sum_{j=1}^n b_{ij}^m(x)^2 = \sum_{i=1}^d \sum_{j=1}^n \left(p_{ji}^m x_i - \sum_{k=1}^d p_{jk}^m x_k x_i\right)^2 \\
&\leq 4 \sum_{i=1}^d \sum_{j=1}^n (p_{ji}^m x_i)^2 + (- \sum_{k=1}^d p_{jk}^m x_k x_i)^2 \\
&= 4 \sum_{i=1}^d \sum_{j=1}^n (p_{ji}^m x_i)^2 + \left(\sum_{k=1}^d p_{jk}^m x_k\right)^2 x_i^2 \\
&\leq 4 \sum_{i=1}^d \sum_{j=1}^n (p_{ji}^m)^2 + \left(\sum_{k=1}^d p_{jk}^m x_k\right)^2 \\
&\overset{\text{CSI}}{\leq} 4 \sum_{i=1}^d \sum_{j=1}^n (p_{ji}^m)^2 + \left(\sum_{k=1}^d p_{jk}^m\right)^2 \|x\|^2 \\
&\leq 4 \sum_{i=1}^d \sum_{j=1}^n (p_{ji}^m)^2 + \left(\sum_{k=1}^d p_{jk}^m\right)^2.
\end{aligned}$$

X_t^m is the filter and satisfies $\left\|X_t^m\right\| \leq 1$, so b^m is bounded for X_t^m. Also due to $\lim_{m\to\infty} \max_{i,j} \left|p_{ij}^m\right|^2 = 0$ it follows that

$$\left\|b^m(X_t^m)\right\|_{dn}^2 \leq c_m$$

for some sequence c_m converging to 0. Thus

$$\left\|b^m(X^m)\right\|_{L^2([0,t])}^2 = \int_0^t \left\|b^m(X_s^m)\right\|_{dn}^2 \,\mathrm{d}s \leq \int_0^t c_m \mathrm{d}s = t c_m \overset{m\to\infty}{\longrightarrow} 0$$

and we can apply Theorem 4.2 to conclude that

$$\lim_{m\to\infty} \mathrm{E}\big[\sup_{s\leq t} \left\|\hat{Y}_s^m - \nu\right\|^2\big] = 0.$$

References

Avellaneda M, Lee J-H (2010) Statistical arbitrage in the US equities market. Quant Finance 10(7):761–782

Avellaneda M, Healy B, Papanicolaou A, Papanicolaou G (2021) Principal eigenportfolios for U.S. equities. Preprint: ssrn.com/abstract=3738769

Bäuerle N, Rieder U (2004) Portfolio optimization with Markov-modulated stock prices and interest rates. IEEE Trans Autom Control 29(3):442–447

Bäuerle N, Rieder U (2005) Portfolio optimization with unobservable Markov-modulated drift process. J Appl Probab 42(2):362–378

Boyle P (2014) Positive weights on the efficient frontier. N Am Actuar J 18(4):462–477

Cass D, Stiglitz JE (1970) The structure of investor preferences and asset returns, and separability in portfolio allocation: a contribution to the pure theory of mutual funds. J Econ Theory 2(2):122–160

Chamberlain G (1988) Asset pricing in multiperiod securities markets. Econ J Econ Soc, 1283–1300

Chen J, Yuan M (2016) Efficient portfolio selection in a large market. J Financ Econom 14(3):496–524

Clark JMC (1978) The design of robust approximations to the stochastic differential equations of nonlinear filtering. Commun Syst Random Process Theory 25:721–734

DeMiguel V, Garlappi L, Uppal R (2007) Optimal versus Naive diversification: How inefficient is the 1/N portfolio strategy? Rev Financ Stud 22(5):1915–1953

Elliott RJ (1993) New finite-dimensional filters and smoothers for noisily observed Markov chains. IEEE Trans Inf Theory 39(1):265–271

Elliott RJ, Aggoun L, Moore JB (1995) Hidden Markov models. Springer, New York

Elliott RJ, Krishnamurthy V, Sass J (2008) Moment based regression algorithms for drift and volatility estimation in continuous-time Markov switching models. Econom J 11:244–270

Frey R, Gabih A, Wunderlich R (2012) Portfolio optimization under partial information with expert opinions. J Theor Appl Finance 15(1):1–18

Hamilton JD (1989) A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57(2):357–384

Haussmann UG, Sass J (2004) Optimal terminal wealth under partial information for HMM stock returns. Contemp Math 351:171–185

James MR, Krishnamurthy V, Le Gland F (1996) Time discretization of continuous-time filters and smoothers for HMM parameter estimation. IEEE Trans Inf Theory 42:593–605

Krishnamurthy V, Leoff E, Sass J (2018) Filterbased stochastic volatility in continuous-time hidden Markov models. Econom Stat 6:1–21

Merton RC (1971) Optimum consumption and portfolio rules in a continuous-time model. J Econ Theory 3:373–413

Merton RC (1972) An analytic derivation of the efficient portfolio frontier. J Financ Quant Anal 7(4):1851–1872

Papanicolaou A (2019) Backward SDEs for control with partial information. Math Financ 29(1):208–248

Pham H, Wei X, Zhou C (2022) Portfolio diversification and model uncertainty: a robust dynamic mean-variance approach. Math Financ 32(1):349–404

Sass J, Haussmann UG (2004) Optimizing the terminal wealth under partial information: the drift process as a continuous time Markov chain. Finance Stochast 8(4):553–577

Sass J, Thös A-K (2022) Risk reduction and portfolio optimization using clustering methods. Econom Stat (to appear)

Schachermayer W, Sirbu M, Tafflin E (2009) In which financial markets do mutual fund theorems hold true. Finance Stochast 13:49–77

Tobin J (1958) Liquidity preference as behavior towards risk. Rev Econ Stud, 65–86

Wonham WM (1965) Some applications of stochastic differential equations to optimal nonlinear filtering. J Soc Ind Appl Math Ser A Control 2(3):347–369

Zhao L, Chakrabarti D, Muthuraman K (2019) Portfolio construction by mitigating error amplification: the bounded-noise portfolio. Oper Res 67(4):965–983