June 28, 2020

A Bayesian Filtering Approach to Incorporate Views in Economic Scenario Generation

By Arvid Cederberg

Abstract

An economic scenario generator could be described as a tool used for simulating future paths of economies and financial markets. It should illuminate the dynamics of risk elements within the economy which drive financial variability, and usually includes models for variables such as sovereign interest rates, equity returns, credit spreads, exchange rates and inflation. Economists are increasingly requiring that their own views of the future market dynamics can be embedded in their economic scenarios, and this study proposes a Bayesian filtering approach to incorporate these views with models calibrated to historical data. With the vast amount of different processes one could choose to include in an ESG, providing a detailed yet completely general method is difficult. To limit the scope, the variables within the economic scenario generator are assumed to be modelled with (vector) autoregressive processes. However, it is shown that the method can be extended to allow for views on variables modelled with general first-order Markov chains, as well as memoryless linear state-space models such as the Dynamic Nelson-Siegel yield curve model. While the main focus will lie on unconditional views, where the imposed views are independent of previous observations and any other input parameters, the possibility of extending the model to allow for conditional views is also discussed.

Keywords: Economic Scenario Generator, Views, Outlook Correction, Bayesian Filtering, Vector Autoregression, VARX, Hidden Markov Model, Dynamic Nelson-Siegel Model

Acknowledgements

This master's thesis was written in collaboration with the Division of Mathematical Statistics at LTH and Kidbrooke Advisory. The writing process has been a major learning experience, and there are certain aspects of this thesis where support from my supervisors was essential. Therefore, I would like to express my deep gratitude to my research supervisors Erik Lindström and Edvard Sjögren for their patient guidance and useful critiques. I would also like to extend my sincere thanks to my friends and family for their support and encouragement.

Glossary

ESG An Economic Scenario Generator (ESG) is used for simulating future paths of financial markets. Its use ranges from outright forecasting to stress testing and calculation of capital requirements.

views A view (sometimes denoted outlook) is a belief about how the value of a variable, or a linear combination of variables, will develop in the future. These views are assumed to be based on expert knowledge.

baseline model A model without any incorporated views is denoted a baseline model. Similarly, a baseline estimate would be an estimate where no views are incorporated. A model is a part of an ESG.

external regressors When referring to external regressors in this study, it means external variables in relation to a model within an ESG. This does not necessarily mean that they are external inputs to the ESG, and in general they will have their own models.

Contents

Abstract **i**

Acknowledgements **iii**

Glossary **v**

1 Introduction **1**

1.1 Background 1

1.2 Aim and Scope 2

2 Theoretical Background **3**

2.1 Vector Autoregressive Process 3

2.2 Views and Recursive Bayesian Filtering 5

2.3 Filtration of Linear Gaussian Models 8

2.4 Views on Yield Curves and the Dynamic Nelson-Siegel Model 11

3 Method **15**

3.1 Model Construction 15

3.2 Estimating the Model 17

3.2.1 Unconstrained Maximum Likelihood Estimation 17

3.2.2 Constrained Maximum Likelihood Estimation 18

3.2.3 Volatility Clustering 20

3.2.4 Graphical Lasso 21

3.2.5 Parameter Inference 21

3.2.6 Selection of Hyperparameters 22

3.3 Defining Views 22

3.3.1 Setting the Covariance Matrix $\boldsymbol{\Omega}_t$ 23

3.3.2 Setting the Views $\boldsymbol{\psi}_t$ 25

3.3.2.1 Nelson-Siegel Approach 26

3.3.2.2 Gaussian Process Regression 27

3.3.2.3 Transformations of Views 28

3.4 Generating Scenarios 29

3.4.1 Views on Dependent Variables 29

3.4.2 Views on External Regressors 30

4 Example with Data **33**

4.1 Data 33

4.2 Estimating the ESG 35

4.3 Simulating Future 41

5 Discussion **45**
5.1 Loss of Conditionality . . . 45
5.2 Paths are Unattainable . . . 48
5.3 Contradicting Views . . . 49
5.4 Model Validation and Bias . . . 50

6 Conclusion **53**

References **55**

Appendices **57**

A Lag Operator **58**

B Matrix Operations **60**
B.1 Vectorization and Kronecker Products . . . 60
B.2 Matrix Derivatives . . . 61

C Multivariate Gaussian **62**
C.1 Conditional Distribution . . . 62
C.2 Matrix Normal Distribution . . . 62

1. Introduction

1.1 Background

An economic scenario generator can take many forms, and Pedersen et al. [2016] describes it as a software tool used for simulating future paths of economies and financial markets. It should illuminate the dynamics of risk elements within the economy which drive financial variability. Its use can range from tasks such as simulating the impact on Swedish equity from a change of the US Federal Reserve Bank's key rate to investigating how a portfolio allocation performs over time under different financial landscapes. What is important to model in an ESG will inevitably vary depending on the organization using it, but some common variables are sovereign interest rates, equity returns, credit spreads, exchange rates and inflation. Modelling other variables such as GDP, commodities, derivatives and mortgage-backed securities is also an option (Pedersen et al. [2016]).

Economists are increasingly requiring that their own outlooks of the future market dynamics can be embedded in their economic scenario generators. These views can be both short- and long-term and take into account factors not present in the historical data, such as current macroeconomic events likely to affect future trends. Embedding own views in statistical models is challenging, as textbooks usually only describe how these models are calibrated to historical data. The best way of incorporating views may depend on the models used in the ESG, the nature of the views and what is desired from the resulting scenario. One could consider direct moment targeting, where parameters are optimized to minimize deviance from the views and maximize the likelihood of the historical data being generated by the model. One could also consider ESGs only calibrated by user-provided parameters, which do not directly consider historical observations. The main problem with these approaches is the large number of parameters typically present in an ESG. Since analytical expressions for the time-varying moments generally are not tractable, brute-force optimization where full re-simulations are made at each optimization step would have to be used. This renders both direct moment targeting and individual calibration of parameters problematic for large and complex scenario generators.

This study proposes a method where the model parameters do not change when views are incorporated. Instead, two sources of information about the future distributions of the model variables are considered, where the forecasted distributions take into account both the model calibrated to the historical data and expert views. Thus, it is here assumed that the economic scenario generator is calibrated to optimally fit historical data as a basis point, upon which views are imposed in the forecasting procedure. There are two major drawbacks of the proposed method, whose implications and possible mitigations will be discussed. The fact that the views are specified unconditionally is an issue for processes with significant dependencies on previous lags. Furthermore, the method is difficult to apply on multi-level hierarchical models.

1.2 Aim and Scope

The aim is to find a practical method with good theoretical foundations for combining own outlooks with models calibrated to historical data in long-term economic scenario generation. These outlooks could in principle be of any variable present in the ESG, although there are some requirements on the assumed processes. Describing how to construct a complete economic scenario generator is beyond the scope of this study, although a simple example will be constructed to illustrate how the proposed method can be applied. With the vast amount of different processes one could choose to include in an ESG, providing a detailed yet completely general method is both difficult and questionable. The report will therefore proceed from a more general setting to a more detailed method. First, the idea of considering the views as observations in a hidden Markov model where the unconditional distribution is estimated by recursive Bayesian filtering will be described. This idea of recursive filtering can be applied to fairly general models. The special case where the transition of the state of the dependent variables in the baseline model is described by linear operators and Gaussian errors will then be considered, resulting in a slight modification of the well-known Kalman filter. A detailed method for incorporating views in ESGs based on (vector) autoregressive processes will then be proposed. While the main focus will lie on unconditional views, meaning that the imposed views are independent of previous observations and any other input parameters, the possibility of extending the model to allow for conditional views will also be discussed.

It will be shown that the vector autoregressive model with external regressors and parameter restrictions is a powerful tool when constructing ESGs. This, together with the fact that the recursive filtering approach is easily applicable, is the reason for the large focus on the VARX model in this study. However, there are certain economic variables which cannot directly be modelled with an autoregressive process. To illuminate this, and to show how the proposed method could be generalized, a briefer description of how views could be incorporated in the Dynamic Nelson-Siegel yield curve model is also given.

2. Theoretical Background

2.1 Vector Autoregressive Process

The vector autoregressive process can be used to capture linear interdependencies of multivariate time series. The model assumes that the conditional expectation is a linear function of past observations, where predictions are based on each variable's own lagged values and the lagged values of the other model variables. Consider modelling a multivariate time series $\{\mathbf{x}_t\}$ where $\mathbf{x}_t \in \mathbb{R}^n$. Using a linear dependence on the past p values in the prediction formula, the conditional expectation is given by

$$\mathbb{E}[\mathbf{x}_t|\mathcal{F}_{t-1}] = \mathbf{c} + \mathbf{A}_1\mathbf{x}_{t-1} + \cdots + \mathbf{A}_p\mathbf{x}_{t-p}$$

where $\mathbf{A}_i \in \mathbb{R}^{n\times n}$ for $i = 1, \ldots, p$. Denoting the forecast error, or innovation, $\mathbf{e}_t = \mathbf{x}_t - \hat{\mathbf{x}}_t$, the process is given by

$$\mathbf{x}_t = \mathbf{c} + \mathbf{A}_1\mathbf{x}_{t-1} + \cdots + \mathbf{A}_p\mathbf{x}_{t-p} + \mathbf{e}_t$$

Or in compact notation

$$\mathbf{x}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i\mathbf{x}_{t-i} + \mathbf{e}_t$$

If the innovations are serially independent, the above process is by definition a vector autoregressive process, denoted VAR(p) (Lütkepohl [2005]). It is often assumed that the innovations are normally distributed, i.e. $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$ where $\mathbf{\Sigma}_t \in \mathbb{R}^{n\times n}$. Similarly to the univariate autoregressive process, the vector autoregressive process is stable, or stationary, if all roots of the characteristic polynomial lie outside the unit circle (Lütkepohl [2005]). Thus, for stability, the condition

$$|\mathbf{I}_n - \sum_{i=1}^{p} \mathbf{A}_i z^i| \neq 0 \quad \text{for} \quad |z| \leq 1$$

must be fulfilled. A useful property is that every n-dimensional VAR(p) process has an np-dimensional VAR(1) representation. This representation is given by

$$\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{c}} + \mathbf{A}\tilde{\boldsymbol{x}}_{t-1} + \tilde{\boldsymbol{e}}_t \tag{2.1}$$

where (see Appendix B.1 for a definition of the vec-operator)

$$\begin{aligned}
\tilde{\boldsymbol{x}}_t &= \text{vec}(\mathbf{x}_t, \dots, \mathbf{x}_{t-p+1}) && (np \times 1) \\
\tilde{\boldsymbol{c}} &= \text{vec}(\mathbf{c}, \mathbf{0}, \dots, \mathbf{0}) && (np \times 1) \\
\mathbf{A} &= \begin{pmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I}_n & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_n & \mathbf{0} \end{pmatrix} && (np \times np) \\
\tilde{\boldsymbol{e}}_t &= \text{vec}(\mathbf{e}_t, \mathbf{0}, \dots, \mathbf{0}) && (np \times 1)
\end{aligned}$$

This means that any VAR(p) process can be converted to a first-order Markov process, meaning that the conditional density only depends on the previous observation. This property is a requirement for the proposed Bayesian filtering approach in the coming sections to be applicable, and thus of great importance in this study. In some cases, the data may show a preeminent moving average structure, i.e. some clear dependency on past innovations. In this case, a vector moving average (VMA) or a vector autoregressive moving average (VARMA) process may better explain the data generating process.

A VMA(q) process is given by

$$\mathbf{x}_t = \mathbf{e}_t + \mathbf{M}_1\mathbf{e}_{t-1} + \dots + \mathbf{M}_q\mathbf{e}_{t-p}$$

and a VARMA(p,q) process is given by

$$\mathbf{x}_t = \mathbf{c} + \mathbf{A}_1\mathbf{x}_{t-1} + \dots + \mathbf{A}_p\mathbf{x}_{t-p} + \mathbf{e}_t + \mathbf{M}_1\mathbf{e}_{t-1} + \dots + \mathbf{M}_q\mathbf{e}_{t-p}$$

or with the lag-operator (see Appendix A)

$$A(L)\mathbf{x}_t = \mathbf{c} + M(L)\mathbf{e}_t \tag{2.2}$$

where $A(L) = \mathbf{I}_n - \mathbf{A}_1 L - \dots - \mathbf{A}_p L^p$ and $M(L) = \mathbf{I}_n + \mathbf{M}_1 L + \dots + \mathbf{M}_q L^q$. An MA(q) process is invertible if

$$|\mathbf{I}_n + \sum_{i=1}^{q} \mathbf{M}_j z^j| \neq 0 \quad \text{for} \quad |z| \leq 1$$

and a VARMA(p,q) process is stable and invertible if its autoregressive part is stable and its moving average part is invertible (Lütkepohl [2005]). A VARMA process is not memoryless and is thus not fulfilling the Markovianity condition. However, any invertible VARMA(p,q) process has a VAR(∞) representation. Using the lag operator notation, the VAR(∞) representation is derived by multiplying both sides of (2.2) from the left by the inverse MA operator and matching coefficients (see Appendix A for more details)

$$M(L)^{-1}A(L)\mathbf{x}_t = M(L)^{-1}\mathbf{c} + \mathbf{e}_t$$

Naturally, an infinite order approximation is infeasible in practice, and the first p terms of the VAR(∞) representation would be used as an approximation. Relatively low lags is sometimes enough to replicate a the moving average part, as displayed with an example in the univariate case in Figure 2.1. Of course, the finite order VAR(p) approximation of the VARMA(p_0, q) process generated by this method is not the optimal estimator of order p, e.g. in terms of maximum likelihood, and given enough observations, fitting a VAR(p) process directly may be a better option. However, this would require a presample of length p, resulting in the use of fewer observations in the fitting process.

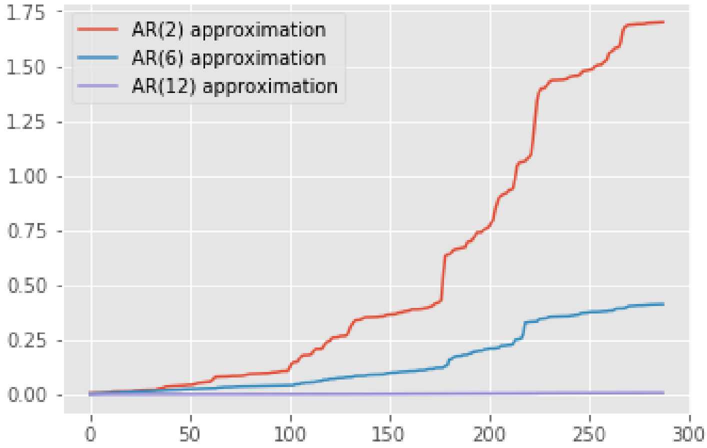


Figure 2.1: Finite order AR approximations of ARMA(1,2) process. Cumulative sum of squared deviance from ARMA(1,2) in-sample predictions.

Finally, it is noted that one may choose to include exogenous variables as predictors in the model, resulting in the vector autoregressive model with external regressors (VARX). The VARX(p,q) model is given by

$$\mathbf{x}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}_{t-i} + \sum_{j=0}^{q} \mathbf{B}_j \mathbf{u}_{t-j} + \mathbf{e}_t \tag{2.3}$$

where $\mathbf{u}_t \in \mathbb{R}^m$ and $\mathbf{B}_j \in \mathbb{R}^{n\times m}$. The term q is in this case the highest lag of the external regressors included, and not a moving average order. This process can be converted to a VARX(1,0) process given by

$$\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{c}} + \mathbf{A}\tilde{\boldsymbol{x}}_{t-1} + \mathbf{B}\tilde{\boldsymbol{u}}_t + \tilde{\boldsymbol{e}}_t \tag{2.4}$$

where

$$\tilde{\boldsymbol{u}}_t = \mathrm{vec}(\mathbf{u}_t, \dots, \mathbf{u}_{t-q}) \qquad (m(q+1)\times 1)$$

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_0 & \dots & \mathbf{B}_q \\ \mathbf{0} & \dots & \mathbf{0} \\ \vdots & & \vdots \\ \mathbf{0} & \dots & \mathbf{0} \end{pmatrix} \qquad (np \times m(q+1))$$

and $\tilde{\boldsymbol{x}}_t$, $\tilde{\boldsymbol{c}}$, $\mathbf{A}$ and $\tilde{\boldsymbol{e}}_t$ are defined as in (2.1). By definition, the variables $\mathbf{u}_t$ are strictly exogenous if all leads and lags are independent of all leads and lags of the error term $\mathbf{e}_t$. It is also assumed that $\mathbf{u}_t$ is stationary. A maximum likelihood procedure for estimating the VARX(p,q) model is covered in Section 3.2. For a more extensive review of the vector autoregressive model, the reader is referred to the book of Lütkepohl [2005].

2.2 Views and Recursive Bayesian Filtering

The Bayesian filtering approach covered in this section is to some degree inspired by the Black-Litterman model originally proposed by Black and Litterman [1992], as well as the time-dependent extension proposed by Steehouwer and van der Schans [2017]. However, there are some fundamental differences, and there is no need for the reader to be acquainted with the Black-Litterman model to understand the following theory. In contrast to the Black-Litterman model, this approach is not related to the CAPM framework. Instead, it is a proposed method for incorporating views in the estimation of the future distribution of discrete time Markov processes, and specifically the VARX process covered in the previous section. The idea is that the views are considered as noisy observations of the conditional expectation of the process at any given time. The fact that the views themselves are specified unconditionally – independent of the previous path – while being considered as observations of the *conditional* expectation is an issue for processes with a strong dependence on the previous observation. This could possibly be solved by defining the views conditional on the previous observation in a sampling approach, as discussed in Section 5.1.

In the following, a model already fitted to the historical data is considered, and the Bayesian filtering is applied when simulating the future distribution. Thus, the aim is to estimate the future distribution of the dependent variables, given some presample, a model estimated to the historical data and noisy observations of the conditional expectations over the forecast horizon. Denote the model fitted to the historical data the baseline model. This model can be any discrete-time first-order Markov chain of the form $\mathbf{x}_t = f(\mathbf{x}_{t-1}) + \mathbf{e}_t$, where $\mathbf{x}_t \in \mathbb{R}^n$ and $\mathbf{e}_t$ are independent and serially uncorrelated zero-mean random variables, possibly with time-varying covariance. The views are defined as a time-varying vector, specifying expert opinions of the expectations of one or more linear combinations of $\mathbf{x}_t$ at time t. The views are denoted $\boldsymbol{\psi}_t \in \mathbb{R}^d$, where d is the number of views. Furthermore, there is an uncertainty associated with the views, specified by an error term $\boldsymbol{\xi}_t$. This give rise to the model

$$\begin{aligned}
\boldsymbol{\psi}_t &= \mathbf{H}_t \boldsymbol{\mu}_t + \boldsymbol{\xi}_t \\
\boldsymbol{\mu}_t &= f(\mathbf{x}_{t-1}) \\
\mathbf{x}_t &= \boldsymbol{\mu}_t + \mathbf{e}_t
\end{aligned} \tag{2.5}$$

where $\mathbf{H}_t \in \mathbb{R}^{d\times n}$. It is assumed that $\boldsymbol{\xi}_t$ are serially uncorrelated zero-mean random variables independent of $\mathbf{e}_t$, and it is required that $\mathbf{H}_t$ has full rank. Note that this assumption will not exclude any views which are not contradicting or redundant. This set-up allows for expressing views such as

$$\begin{aligned}
\mathbf{H}_t &= \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\
\boldsymbol{\psi}_t &= (0.0001, 0.001)^\mathsf{T} \\
\mathbf{x}_t &\in \mathbb{R}^3
\end{aligned}$$

meaning that, at time t, the mean of the first variable is expected to be one basis point above the mean of second variable, while the mean of third variable is expected to be ten basis points. How to set $\boldsymbol{\psi}_t$, $\mathbf{H}_t$ and $\boldsymbol{\xi}_t$ will be discussed further in Section 3.3. The requirement that the baseline model is a Markov chain is at glance restrictive. However, in practice it is often possible to augment the state equation as

$$\tilde{\boldsymbol{x}}_t = \begin{pmatrix} \mathbf{x}_t \\ \mathbf{x}_{t-1} \\ \vdots \\ \mathbf{x}_{t-p+1} \end{pmatrix}$$

such that $\mathbf{x}_t = f(\mathbf{x}_{t-1}, \dots, \mathbf{x}_{t-p}) + \mathbf{e}_t$ can be written as $\tilde{\boldsymbol{x}}_t = f(\tilde{\boldsymbol{x}}_{t-1}) + \tilde{\boldsymbol{e}}_t$. See for example the VAR(1) representation of the VAR(p) model given in the previous section. The alteration of the matrix $\mathbf{H}_t$ needed in (2.5) is easily seen. Furthermore, any deterministic influence from other parameters is also allowed, meaning that one could have e.g. $\boldsymbol{\mu}_t = f(\mathbf{x}_{t-1}, \mathbf{u}_t, t)$ as a baseline model.

The model (2.5) has the form of the hidden Markov model given in Figure 2.2, and the recursive Bayesian filter computes the distribution $p(\mathbf{x}_t|\boldsymbol{\psi}_{1:t})$ given a prior distribution of $\mathbf{x}_0$, the model in (2.5) and some (noisy) observations $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_t$ of the conditional expectations $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_t$.

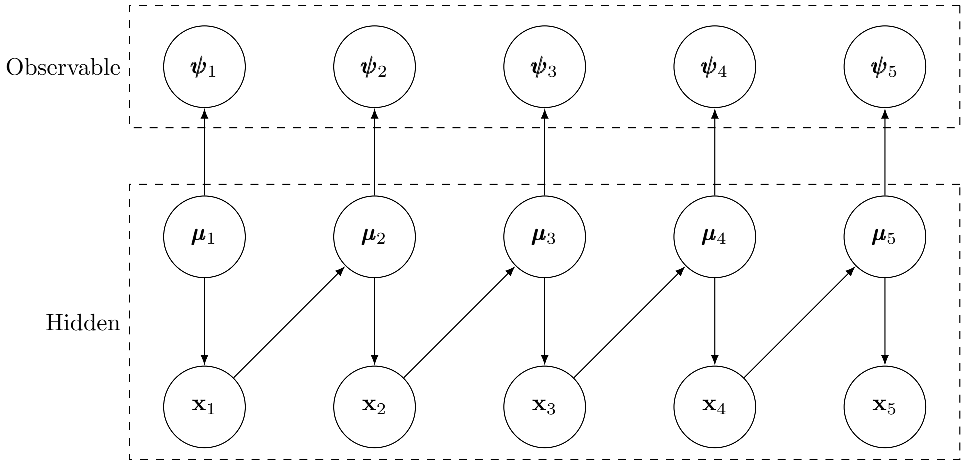


Figure 2.2: Hidden Markov Model

The general Bayesian filtering procedure can be outlined by

[1] Consider the filter density $p(\mathbf{x}_0)$ at time $t = 0$ as given

[2] At time t, compute the predictive density $p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\psi}_{1:t})$

[3] At time $t+1$, calculate $p(\boldsymbol{\psi}_{t+1}|\boldsymbol{\psi}_{1:t})$ and update filter density $p(\boldsymbol{\mu}_{t+1}|\boldsymbol{\psi}_{1:t+1})$

[4] Repeat steps 2 and 3

The densities are here expressed in terms of $\boldsymbol{\mu}_t$. One could of course add a layer and express the densities of $\mathbf{x}_t$, but since the values are always centered around the expectation (conditional on the path) with the distribution given by $\mathbf{e}_t$, it suffices to know the time-evolution of the distribution of $\boldsymbol{\mu}_t$ to sample from the distribution in practice. Furthermore, an analytical expression of the filter density of $\mathbf{x}_t$ can be derived in the Gaussian case (see Section 2.3). The predictive density $p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1})$ of $\boldsymbol{\mu}_t$, using the information from the observable views up to time $t-1$, can be derived from

$$\begin{aligned}
p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1}) &= \int p(\boldsymbol{\mu}_t|\mathbf{x}_{t-1},\boldsymbol{\psi}_{1:t-1})d\mathbf{x}_{t-1} \\
&= \int p(\boldsymbol{\mu}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\boldsymbol{\psi}_{1:t-1})d\mathbf{x}_{t-1} \\
&= \iint p(\boldsymbol{\mu}_t|\mathbf{x}_{t-1})p(\mathbf{x}_{t-1}|\boldsymbol{\mu}_{t-1})p(\boldsymbol{\mu}_{t-1}|\boldsymbol{\psi}_{1:t-1})d\boldsymbol{\mu}_{t-1}d\mathbf{x}_{t-1} \qquad (2.6)
\end{aligned}$$

and the filter density $p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t})$ of $\boldsymbol{\mu}_t$, using the information from the observable views up to time t, can be derived by using the fact that

$$\begin{aligned}
p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t}) &= \frac{p(\boldsymbol{\psi}_t|\boldsymbol{\mu}_t)p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1})}{p(\boldsymbol{\psi}_t|\boldsymbol{\psi}_{1:t-1})} \\
&= \frac{p(\boldsymbol{\psi}_t|\boldsymbol{\mu}_t)p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1})}{\int p(\boldsymbol{\psi}_t|\boldsymbol{\mu}_t)p(\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1})d\boldsymbol{\mu}_t} \qquad (2.7)
\end{aligned}$$

There is no closed form recursion in the general case. For non-linear f in (2.5) and/or non-Gaussian innovations, the reader is referred to the unscented Kalman filter proposed by Julier and Uhlmann [1997] and the Sequential Monte Carlo filter, or particle filter, described by e.g. Lopes and Tsay [2011]. Briefly, the UKF and the SMC filter are similar in the sense that they generate a set of points via known non-linear equations and combine the results to estimate the distribution of the state. However, the SMC filter generates points randomly, while the UKF generates points according to a certain algorithm. This means that the number of points needed, as well as the computational cost, is higher for the SMC filter, but also that the SMC filter will perform better given a large enough set of points, especially for non-Gaussian errors. The estimation error converges to zero as the number of points approaches infinity, which is not the case using the UKF. For Gaussian errors, the UKF computes a third order (Taylor expansion) accurate approximation of the first and second central moments, and for non-Gaussian errors the approximation is accurate to the second order (Julier and Uhlmann [1997]). If first order accuracy is enough, typically when the system is almost linear and the errors are Gaussian, an extended Kalman filter (Lee and Ricker [1994]) may yield comparable accuracy at a lower computational cost than both the UKF and the SMC filter. These methods are formulated for state-space models, and some alterations are needed due to the form of the model in (2.5).

2.3 Filtration of Linear Gaussian Models

Again, consider the model given in (2.5), now with

$$\begin{aligned}\boldsymbol{\mu}_t &= \mathbf{c} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t \\ \mathbf{x}_t &= \boldsymbol{\mu}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_t)\end{aligned} \tag{2.8}$$

where $\mathbf{u}_t$ is modelled externally and seen as a non-random control vector. This does not mean that $\mathbf{u}_t$ cannot be sampled from a stochastic process, it just means that each filter recursion is done conditional on one path of $\{\mathbf{u}_t\}$. The process given in (2.4) is of this form, and the theory below will thus apply when considering views on variables modelled by a VARX(p,q) process with Gaussian innovations. Also assume that $\boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t)$. Due to the linearity of the prediction formula and the Gaussianity of $\mathbf{e}_t$ and $\boldsymbol{\xi}_t$, analytical expressions of the densities in (2.6) and (2.7) above can be derived. Denote $\mathbf{P}_t^x$ the (unconditional) covariance of $\mathbf{x}_t$ and $\mathbf{P}_t^\mu$ the covariance of $\boldsymbol{\mu}_t$. The indices $t|t-1$ and $t|t$ refer to the predictive and filter estimates respectively at time t. Assume an initial distribution $\mathbf{x}_{0|0} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{0|0}, \mathbf{P}_{0|0}^x)$. The predictive density of $\boldsymbol{\mu}_t$ at time t is given by

$$\begin{aligned}\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1} &\sim \boldsymbol{\mu}_t|\mathbf{c} + \mathbf{A}\hat{\boldsymbol{\mu}}_{t-1|t-1} + \mathbf{B}\mathbf{u}_t \\ &\overset{d}{=} \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{t|t-1}, \mathbf{P}_{t|t-1}^\mu\right)\end{aligned}$$

where

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{t|t-1} &= \mathbf{c} + \mathbf{A}\hat{\boldsymbol{\mu}}_{t-1|t-1} + \mathbf{B}\mathbf{u_t} \\ \mathbf{P}_{t|t-1}^\mu &= \mathbf{A}\mathbf{P}_{t-1|t-1}^x\mathbf{A}^\intercal\end{aligned}$$

and $\hat{\boldsymbol{\mu}}_{t-1|t-1}$ is the expectation of the filter density at time $t-1$. Furthermore, the joint distribution of $\boldsymbol{\mu}_t$ and $\boldsymbol{\psi}_t$ given the information at time $t-1$ (see Appendix C) is given by

$$\begin{pmatrix}\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1} \\ \boldsymbol{\psi}_t|\boldsymbol{\psi}_{1:t-1}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\hat{\boldsymbol{\mu}}_{t|t-1} \\ \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}\end{pmatrix}, \begin{pmatrix}\mathbf{P}_{t|t-1}^\mu & \mathbf{P}_{t|t-1}^\mu\mathbf{H}_t^\intercal \\ \mathbf{H}_t\mathbf{P}_{t|t-1}^\mu & \mathbf{H}_t\mathbf{P}_{t|t-1}^\mu\mathbf{H}_t^\intercal + \boldsymbol{\Omega}_\mathbf{t}\end{pmatrix}\right)$$

Again using Appendix C, this leads to the filter density of $\boldsymbol{\mu}_t$

$$\begin{aligned}\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t} &\sim \mathcal{N}\Big(\hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{P}_{t|t-1}^\mu\mathbf{H}_t^\intercal(\mathbf{H}_t\mathbf{P}_{t|t-1}^\mu\mathbf{H}_t^\intercal + \boldsymbol{\Omega}_\mathbf{t})^{-1}(\boldsymbol{\psi}_t - \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}), \\ &\qquad \mathbf{P}_{t|t-1}^\mu - \mathbf{P}_{t|t-1}^\mu\mathbf{H}_t^\intercal(\mathbf{H}_t\mathbf{P}_{t|t-1}^\mu\mathbf{H}_t^\intercal + \boldsymbol{\Omega}_\mathbf{t})^{-1}\mathbf{H}_t\mathbf{P}_{t|t-1}^\mu\Big) \\ &\overset{d}{=} \mathcal{N}\left(\hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\psi}_t - \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}), \mathbf{P}_{t|t-1}^\mu - \mathbf{K}_t\mathbf{H}_t\mathbf{P}_{t|t-1}^\mu\right)\end{aligned}$$

Since $\mathbf{x}_t$ is normally distributed around $\boldsymbol{\mu}_t$, the filter density of $\mathbf{x}_t$ is given by

$$\mathbf{x}_t|\boldsymbol{\psi}_{1:t} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{t|t}, \mathbf{P}^x_{t|t})$$

where

$$\begin{aligned}
\mathbf{K}_t &= \mathbf{P}^{\mu}_{t|t-1}\mathbf{H}_t^{\mathsf{T}}(\mathbf{H}_t\mathbf{P}^{\mu}_{t|t-1}\mathbf{H}_t^{\mathsf{T}} + \boldsymbol{\Omega}_{\mathbf{t}})^{-1} \\
\hat{\boldsymbol{\mu}}_{t|t} &= \hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\psi}_t - \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}) \\
\mathbf{P}^x_{t|t} &= \mathbf{P}^{\mu}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}^{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t
\end{aligned}$$

An apparent issue is that the covariance of the predictive density of the conditional expectation, $\mathbf{P}^{\mu}_{t|t-1}$, is not always invertible, for example if $\mathbf{A} = \mathbf{0}$. A somewhat crude, but practically feasible, solution is to add a small value to each diagonal element of $\mathbf{P}^{\mu}_{t|t-1}$. Since $\mathbf{H}_t$ has full rank by definition, and since $\mathbf{P}^{\mu}_{t|t-1}$ is positive semi definite, adding a small value to the diagonal will result in a feasible value of $\mathbf{K}_t$ for any valid covariance matrix $\boldsymbol{\Omega}_t$. This is a common approach, and it is essentially the same as adding a small noise to the conditional expectation in (2.8). Thus, the filtration procedure for the linear Gaussian process in (2.8) is done by the following steps:

[1] Initialize:

Set ϵ to a very small value, e.g. 10^{-10}. Let $\mathbf{x}_{0|0} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{0|0}, \mathbf{P}^x_{0|0})$. Here, the initial distribution is chosen as $\mathbf{P}^x_{0|0} = \mathbf{0}$ and $\hat{\boldsymbol{\mu}}_{0|0} = \mathbf{x}_0$ where $\mathbf{x}_0$ is the last known observation, since this gives the same initial predictive density as the baseline model.

[2] Predict:

$$\begin{aligned}
\hat{\boldsymbol{\mu}}_{t|t-1} &= \mathbf{c} + \mathbf{A}\hat{\boldsymbol{\mu}}_{t-1|t-1} + \mathbf{B}\mathbf{u}_{\mathbf{t}} \\
\mathbf{P}^{\mu}_{t|t-1} &= \mathbf{A}\mathbf{P}^x_{t-1|t-1}\mathbf{A}^{\mathsf{T}} + \epsilon\mathbf{I}_n
\end{aligned}$$

[3] Update:

$$\begin{aligned}
\mathbf{K}_t &= \mathbf{P}^{\mu}_{t|t-1}\mathbf{H}_t^{\mathsf{T}}(\mathbf{H}_t\mathbf{P}^{\mu}_{t|t-1}\mathbf{H}_t^{\mathsf{T}} + \boldsymbol{\Omega}_{\mathbf{t}})^{-1} \\
\hat{\boldsymbol{\mu}}_{t|t} &= \hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\psi}_t - \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}) \\
\mathbf{P}^x_{t|t} &= \mathbf{P}^{\mu}_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}^{\mu}_{t|t-1} + \boldsymbol{\Sigma}_t
\end{aligned}$$

This is just a slight modification of the well-known discrete time Kalman filter, originally derived by Kalman [1960], and perhaps, easier understood by reading the Bayesian derivation of Särkkä [2013]. The matrix $\mathbf{K}_t \in \mathbb{R}^{d\times n}$ is denoted the Kalman gain. To understand how the moments will be affected by the views, first consider the case where views of the expectations of every variable are considered, i.e. $\mathbf{H}_t = \mathbf{I}_n$. The Kalman gain is then given by $\mathbf{K}_t = \mathbf{P}^{\mu}_{t|t-1}(\mathbf{P}^{\mu}_{t|t-1} + \boldsymbol{\Omega}_{\mathbf{t}})^{-1}$, which makes it easier to realize that the ratio between the uncertainty of the model and the uncertainty in the views will impact how much the original process is affected. As the magnitude of the covariance matrix $\boldsymbol{\Omega}_{\mathbf{t}}$ approaches infinity, the Kalman gain will approach zero. This means that if the views are considered as completely uncertain, the views will be ignored, and the filter density will be the same as the predictive density. On the contrary, if the covariance matrix $\boldsymbol{\Omega}_t$ approaches zero, meaning that the views are considered as certain, the views will be fulfilled completely, i.e. the distribution of the specified linear combination will be given by $\mathbf{H}_t\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\psi}_t, \mathbf{H}_t\boldsymbol{\Sigma}_t\mathbf{H}_t^{\mathsf{T}})$. Thus, the magnitude of the change of the expectation at time t induced by the views is determined by $\boldsymbol{\Omega}_t$.

While the limitations and advantages of this approach will be discussed further in Chapter 5, it is important to understand what is happening to the posterior distribution. The views are adding information to the estimation procedure and the uncertainty of the conditional expectation will be lower than those implied by both the baseline model and the views (see Figure 2.3). Consider the case where views about the expected values of some individual assets are specified, i.e. not linear combinations of multiple variables, and where the residual covariance of the values around the conditional expectation is constant. In the limit $\boldsymbol{\Omega}_t \to \mathbf{0}$ for all t, the posterior covariance of the future values of the variables with views will approach the estimated residual covariance over the whole horizon, i.e. the values in e.g. 50 years are just as uncertain as the values tomorrow. Under the assumption that both the distribution from the baseline model and the distribution from the views are correct, meaning that the baseline model is correctly specified and that the views are observations of the true expectations with the supplied measurement error, the estimated filter density would be the true density of the future values. This is a far-fetched assumption however, and the issue of the loss of conditionality will be discussed in Section 5.1. For now, it is noted that this issue increases with both the magnitude of the elements in $\mathbf{A}$ and the amount of influence from the variables $\mathbf{u}_t$. When modelling processes which are noisy in comparison to the prediction power of previous lags and the external regressors, this issue is of less concern since the volatility is mainly driven by the uncertainty around the conditional expectation. Furthermore, the issue could possibly be handled at its core by formulating conditional views, as discussed in Section 5.1.

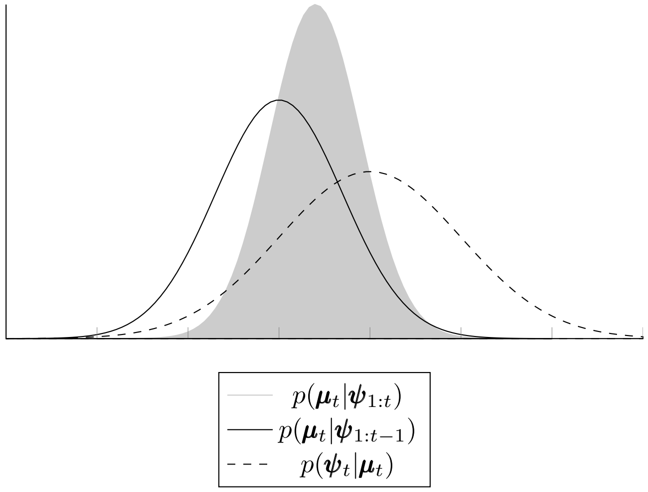


Figure 2.3: Illustration of how the filter density combines information from the views with information from the original process

2.4 Views on Yield Curves and the Dynamic Nelson-Siegel Model

It is clear that not all variables can be modelled directly (or after a simple transformation) with an autoregressive process, or any first-order Markov chain in general. To provide some guidance to the practitioner, a less detailed description of how to incorporate views in a yield curve model will be given. This is to illustrate the flexibility of the proposed method, and in the example with data in Chapter 4, a yield-curve model will not be included as the filtering procedure is best illustrated with a simpler example. Furthermore, the proposed methods for setting the level and uncertainty of the views are based on (vector) autoregressive processes, meaning that they will not directly apply to the state-space model described below. A classic yield curve model is the Nelson-Siegel curve, originally suggested by Nelson and Siegel [1987]. Let m denote the time to maturity and $y(m)$ the yield to maturity. The Nelson-Siegel curve is given by

$$y(m) = \beta_0 + (\beta_1 + \beta_2)\frac{1 - \exp(-m/\tau)}{m/\tau} - \beta_2 \exp(-m/\tau)$$

where the parameters β_0, β_1 and β_2 determines the shape of the curve. β_0 is simply a constant and does not require any explanation, but it is noted that this would correspond to the long-run level of interest rates. β_1 can be seen as a decay parameter, since the term

$$\frac{1 - \exp(-m/\tau)}{m/\tau}$$

starts at 1 and exponentially decays to zero as $m : 0^+ \to \infty$. The term β_2 could instead be seen as the size of a "hump", since the term

$$\frac{1 - \exp(-m/\tau)}{m/\tau} - \exp(-m/\tau)$$

starts at zero, increases fast and then decreases slowly to zero. τ controls the rate of exponential decay, and a larger value gives slower decay. Together, these terms have the ability to form many of the typical shapes seen in yield curves. Usually, one is interested in modelling the yield to a number of different maturities, which can be written as

$$\mathbf{y} = \begin{pmatrix} y(m_1) \\ y(m_2) \\ \vdots \\ y(m_N) \end{pmatrix} = \mathbf{G}\boldsymbol{\beta} \qquad (N \times 1)$$

where

$$\mathbf{G} = \begin{pmatrix} 1 & \frac{1-\exp(-m_1/\tau)}{m_1/\tau} & \frac{1-\exp(-m_1/\tau)}{m_1/\tau} - \exp(-m_1/\tau) \\ 1 & \frac{1-\exp(-m_2/\tau)}{m_2/\tau} & \frac{1-\exp(-m_2/\tau)}{m_2/\tau} - \exp(-m_2/\tau) \\ \vdots & \vdots & \vdots \\ 1 & \frac{1-\exp(-m_N/\tau)}{m_N/\tau} & \frac{1-\exp(-m_N/\tau)}{m_N/\tau} - \exp(-m_N/\tau) \end{pmatrix} \qquad (N \times 3)$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\mathsf{T} \qquad (3 \times 1)$$

This model can be extended to let the parameters vary with time, leading to the Dynamic Nelson-Siegel model (Koopman et al. [2010]). The choice of model for the parameters varies across the literature, but let us assume that they follow a vector autoregressive process of order 1 – the analogy for a VAR(p) model should be clear from Section 2.1. The model could then be represented as

$$\begin{aligned} \boldsymbol{\beta}_t &= \mathbf{c} + \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{e}_t \\ \mathbf{y}_t &= \mathbf{G}\boldsymbol{\beta}_t + \boldsymbol{\eta}_t \end{aligned} \tag{2.9}$$

where $\mathbf{e}_t$ and $\boldsymbol{\eta}_t$ are assumed to be independent, serially uncorrelated zero-centered Gaussian random variables with covariances $\boldsymbol{\Sigma}_e$ and $\boldsymbol{\Sigma}_\eta$ respectively. The estimation can then be done by using the (non-modified) Kalman filter (Särkkä [2013]), and maximize the likelihood conditional on the historical observations $\mathbf{y}_t$. Denoting $\hat{\mathbf{y}}_{t|t-1} = \mathbf{G}\hat{\boldsymbol{\beta}}_{t|t-1}$ the one-step prediction and $\mathbf{P}^y_{t|t-1}$ its covariance, the optimal parameters can be found by maximizing

$$\log\mathcal{L}(\theta|\mathbf{y}_0) = -\frac{1}{2}\sum_{t=1}^{T}\Big(\log|\mathbf{P}^y_{t|t-1}| + (\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})^\intercal[\mathbf{P}^y_{t|t-1}]^{-1}(\mathbf{y}_t - \hat{\mathbf{y}}_{t|t-1})\Big) + const$$

with the method of choice (Lindström et al. [2015]). Assuming that the parameter τ is fixed could simplify the optimization procedure. Thus, before any views are imposed, the baseline model in (2.9) is assumed to have been estimated. Consider imposing views on the conditional expectation of the yield to maturity, giving the model

$$\begin{aligned} \boldsymbol{\beta}_t &= \mathbf{c} + \mathbf{A}\boldsymbol{\beta}_{t-1} + \mathbf{e}_t & \mathbf{e}_t &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_e) \\ \boldsymbol{\mu}_t &= \mathbf{G}\boldsymbol{\beta}_t & & \\ \boldsymbol{\psi}_t &= \mathbf{H}_t\boldsymbol{\mu}_t + \boldsymbol{\xi}_t & \boldsymbol{\xi}_t &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t) \\ \mathbf{y}_t &= \boldsymbol{\mu}_t + \boldsymbol{\eta}_t & \boldsymbol{\eta}_t &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\eta) \end{aligned} \tag{2.10}$$

Denote $\mathbf{P}^\beta_t$, $\mathbf{P}^\mu_t$ and $\mathbf{P}^y_t$ the covariances of the variables $\boldsymbol{\beta}_t$, $\boldsymbol{\mu}_t$ and $\mathbf{y}_t$ respectively, and the indices $t|t-1$ and $t|t$ the predictive and filter estimates respectively. The predictive densities of $\boldsymbol{\beta}_t$ and $\boldsymbol{\mu}_t$ are given by

$$\begin{aligned} \boldsymbol{\beta}_t|\boldsymbol{\psi}_{1:t-1} &\sim \boldsymbol{\beta}_t|\mathbf{c} + \mathbf{A}\hat{\boldsymbol{\beta}}_{t-1|t-1} \\ &\overset{d}{=} \mathcal{N}(\hat{\boldsymbol{\beta}}_{t|t-1}, \mathbf{P}^\beta_{t|t-1}) \\ \boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t-1} &\sim \boldsymbol{\mu}_t|\mathbf{G}\hat{\boldsymbol{\beta}}_{t|t-1} \\ &\overset{d}{=} \mathcal{N}(\hat{\boldsymbol{\mu}}_{t|t-1}, \mathbf{P}^\mu_{t|t-1}) \end{aligned}$$

where

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{t|t-1} &= \mathbf{c} + \mathbf{A}\hat{\boldsymbol{\beta}}_{t-1|t-1} && (3 \times 1) \\ \mathbf{P}^\beta_{t|t-1} &= \mathbf{A}\mathbf{P}^\beta_{t-1|t-1}\mathbf{A}^\intercal + \boldsymbol{\Sigma}_e && (3 \times 3) \\ \hat{\boldsymbol{\mu}}_{t|t-1} &= \mathbf{G}\hat{\boldsymbol{\beta}}_{t|t-1} && (N \times 1) \\ \mathbf{P}^\mu_{t|t-1} &= \mathbf{G}\mathbf{P}^\beta_{t|t-1}\mathbf{G}^\intercal && (N \times N) \end{aligned}$$

The filter density of $\boldsymbol{\mu}_t$ can be calculated similarly as in Section 2.3 (following Appendix C.1) since $\boldsymbol{\psi}_t|\boldsymbol{\mu}_t \sim \mathcal{N}(\mathbf{H}_t\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t)$, yielding

$$\boldsymbol{\mu}_t|\boldsymbol{\psi}_{1:t} \sim \mathcal{N}\big(\hat{\boldsymbol{\mu}}_{t|t}, \mathbf{P}^\mu_{t|t}\big)$$

where

$$\begin{aligned} \hat{\boldsymbol{\mu}}_{t|t} &= \hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}^\mu_t(\boldsymbol{\psi}_t - \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}) && (N \times 1) \\ \mathbf{P}^\mu_{t|t} &= \mathbf{P}^\mu_{t|t-1} - \mathbf{K}^\mu_t\mathbf{H}_t\mathbf{P}^\mu_{t|t-1} && (N \times N) \\ \mathbf{K}^\mu_t &= \mathbf{P}^\mu_{t|t-1}\mathbf{H}^\intercal_t(\mathbf{H}_t\mathbf{P}^\mu_{t|t-1}\mathbf{H}^\intercal_t + \boldsymbol{\Omega}_\mathbf{t})^{-1} && (N \times d) \end{aligned}$$

Since $\mathbf{y}_t$ is normally distributed around $\boldsymbol{\mu}_t$, the filter density of $\mathbf{y}_t$ is given by

$$\mathbf{y}_t|\boldsymbol{\psi}_{1:t} \sim \mathcal{N}(\hat{\boldsymbol{\mu}}_{t|t}, \mathbf{P}^y_{t|t})$$

where

$$\mathbf{P}^y_{t|t} = \mathbf{P}^\mu_{t|t} + \boldsymbol{\Sigma}_\eta$$

What remains to specify to complete the recursion is the posterior of $\boldsymbol{\beta}_t$. Using the fact that $\boldsymbol{\psi}_t|\boldsymbol{\beta}_t \sim \mathcal{N}(\mathbf{H}_t\mathbf{G}\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t)$ together with Appendix C.1, the posterior of $\boldsymbol{\beta}_t$ is given by

$$\boldsymbol{\beta}_t|\boldsymbol{\psi}_{1:t} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}_{t|t}, \mathbf{P}^\beta_{t|t})$$

where

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{t|t} &= \hat{\boldsymbol{\beta}}_{t|t-1} + \mathbf{K}^\beta_t(\boldsymbol{\psi}_t - \mathbf{H}_t\mathbf{G}\hat{\boldsymbol{\beta}}_{t|t-1}) && (3 \times 1)\\
\mathbf{P}^\beta_{t|t} &= \mathbf{P}^\beta_{t|t-1} - \mathbf{K}^\beta_t\mathbf{H}_t\mathbf{G}\mathbf{P}^\beta_{t|t-1} && (3 \times 3)\\
\mathbf{K}^\beta_t &= \mathbf{P}^\beta_{t|t-1}\mathbf{G}^\intercal\mathbf{H}^\intercal_t(\mathbf{H}_t\mathbf{G}\mathbf{P}^\beta_{t|t-1}\mathbf{G}^\intercal\mathbf{H}^\intercal_t + \boldsymbol{\Omega}_t)^{-1} && (3 \times d)
\end{aligned}$$

The filter recursion is thus complete, which shows that the method may be generalized to more complex models than the VARX model, although the process must still be memoryless. Once again, some small value would be added to the diagonal of the predictive covariance of $\boldsymbol{\mu}_t$. While the proposed method of setting the covariance matrix $\boldsymbol{\Omega}_t$ in Section 3.3 would have to be altered, it could easily be extended by applying the same idea.

3. Method

3.1 Model Construction

As mentioned, the aim of this study is not to show how to construct a complete ESG. However, a simple example will be considered to highlight why the VARX model is of a convenient form, and to illustrate how views can be incorporated. For simplicity, only the US market is considered. For a multi-economy ESG, foreign exchange rates would have to be modelled to link the simulation results and aggregate portfolio outcomes across economies. The example ESG is sketched in Figure 3.1, where arrows indicate input type dependencies, i.e. not just noise correlation.

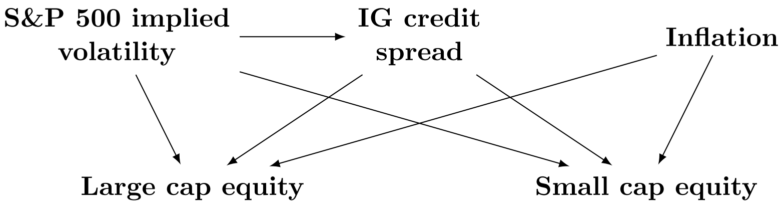


Figure 3.1: High-level sketch of a simple ESG

Now consider the vector autoregressive model with external regressors described in Section 2.1, given by

$$\mathbf{x}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}_{t-i} + \sum_{j=0}^{q} \mathbf{B}_j \mathbf{u}_{t-j} + \mathbf{e}_t \tag{3.1}$$

where $\mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)$ and where $\mathbf{x}_t$ has length n and $\mathbf{u}_t$ has length m. This model assumes multiple interacting dependent variables and a set of external regressors. This is a convenient process for constructing ESGs, since it provides a basis for controlling input type dependencies by imposing restrictions on the parameters $\mathbf{A}_i$ and $\mathbf{B}_j$, and since it allows for structural analysis (Lütkepohl [2005], Chapter 9). Furthermore, adding any deterministic (or Gaussian driven) influence to capture e.g. bias or a deterministic trend would not pose an issue when applying the filtering procedure described in Section 2.3. This makes it a flexible model. The external regressors, $\mathbf{u}_t$, would in this case be the S&P 500 implied volatility, the investment grade credit spread and inflation (modelled as consumer price index) after appropriate transformations. Furthermore, small cap equities are allowed to be dependent on large cap equities, but not the other way around, and small cap equities are not allowed to be dependent on each other. To clarify, no small cap equity is allowed to have a direct impact on another (large or small cap) equity, although noise correlation is allowed.

Adding some detail to the sketch in Figure 3.1, the dependency structure is depicted in Figure 3.2. This simple ESG can thus be modelled with a VARX(p,q) model with parameter restrictions, combined with univariate models for inflation, interest rate and volatility. To build more complex economic scenario generators, having one or more parallel VARX(p,q) model may not be sufficient, and one could also consider cascade type ESGs where dependent variables $\mathbf{x}_t$ are used as external regressors $\mathbf{u}_t$ at a lower level. However, incorporating views at a higher level would then be difficult, and to directly apply the proposed Bayesian filtering method, all variables with views must be endogenous (See Section 3.4.2). While this model is not necessarily appropriate for all possible variables in an ESG, the Bayesian filtering procedure for incorporating views can be applied to more general processes, although it may require a more computationally costly approach such as Sequential Monte Carlo filtering in some cases.

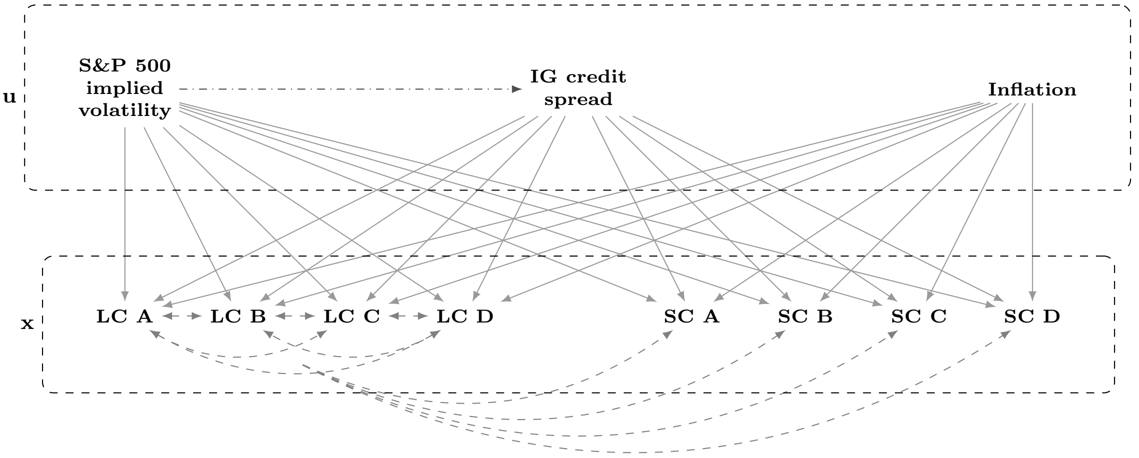


Figure 3.2: Detailed sketch of a simple ESG. Unbroken arrows are dependencies from external regressors and dashed arrows are dependencies from other dependent variables. The dash-dotted arrow is a dependency modelled externally.

In this study, the only source of uncertainty considered is the random error term $\mathbf{e}_t$. However, Hildebrand et al. [2019] states that when observing past observations, it only gives information about one previous state of the world, or one regime. With this point of view, the uncertainty should capture the fact that the data generating process may change in the future. Other sources of uncertainty to consider is the uncertainty from the parameter estimation and the uncertainty from the model choice. Quantifying the uncertainty arising from the possible discontinuation of the historical data generating process and the uncertainty due to the choice of model is difficult, and one may argue that keeping this in mind when analyzing prediction intervals is more sensible than trying to incorporate these uncertainties in the model. However, the bootstrapping procedure outlined in Section 3.2.5 can be applied to account for parameter uncertainty.

The models for volatility, credit spread and inflation used in the example ESG will be briefly described. The S&P 500 implied volatility is modelled with historical data from the VIX index. The process is modelled as an AR(p) process with mean reversion after a log-transformation for positivity

$$\begin{aligned}
\tilde{v}_t &= \log v_t \\
\theta &= \frac{1}{T}\sum_{t=1}^{T} \tilde{v}_t \\
\tilde{v}_t &= c(\tilde{v}_{t-1} - \theta) + a_1\tilde{v}_{t-1} + a_2\tilde{v}_{t-2} + \cdots + a_p\tilde{v}_{t-p} + e_t, \qquad e_t \sim N(0, \sigma^2)
\end{aligned}$$

This is of course the same as

$$\tilde{v}_t = c^* + a_1^*\tilde{v}_{t-1} + a_2\tilde{v}_{t-2} + \cdots + a_p\tilde{v}_{t-p} + e_t, \qquad e_t \sim N(0, \sigma^2)$$

with $c^* = -c\theta$ and $a_1^* = a_1 + c$, and the mean reversion coefficient is merely for interpretation. The investment grade credit spread is modelled directly, i.e. the risk-free rate and the investment grade yield are not modelled individually. The uncertainty is assumed to be driven by the S&P 500 implied volatility, and leverage from the volatility offset is also included as a predictor. Once again, an AR(p) model is chosen, and the model is thus given by

$$\mu = \frac{1}{T}\sum_{t=1}^{T} \log y_t$$
$$\tilde{y}_t = \log y_t - \mu$$
$$\tilde{y}_t = a_1\tilde{y}_{t-1} + a_2\tilde{y}_{t-2} + \cdots + a_p\tilde{y}_{t-p} + b(v_t - \bar{v}) + v_t e_t, \qquad e_t \sim N(0, \sigma^2)$$

The inflation is differenced before being modelled with an AR(p) model with constant volatility and no external predictors.

3.2 Estimating the Model

The estimation of the models for the external regressors will not covered here, but since they are all (univariate) autoregressive processes, the analogy is clear. The parameters to be estimated are thus the vector $\mathbf{c}$ and the matrices $\mathbf{A}_i$ and $\mathbf{B}_j$. First, the Maximum Likelihood estimates without parameter restrictions are derived. An iterative Maximum Likelihood procedure where parameter restrictions are included is then presented. In other sections in this study, the covariance is allowed to be time-varying, since the practitioner may want to include a stochastic model for the volatility. However, due to the difficulty of fitting a stochastic volatility process to high-dimensional VARX models, a constant covariance is assumed in the estimation procedure below. In the following, it is therefore assumed that the data is homoscedastic and that transformations have been applied to ensure (sufficient) stationarity of $\mathbf{x}_t$ and $\mathbf{u}_t$. For a volatility clustering model, where the same (historically observed) volatility process is assumed to drive all dependent variables, see Section 3.2.3.

3.2.1 Unconstrained Maximum Likelihood Estimation

To estimate the parameters in (3.1), the representation in (3.2) given an appropriate presample is considered.

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E} \tag{3.2}$$

where

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_T^\top \end{pmatrix} \qquad (T \times n)$$

$$\mathbf{Z} = \begin{pmatrix} 1 & \mathbf{x}_0^\top & \dots & \mathbf{x}_{1-p}^\top & \mathbf{u}_1^\top & \dots & \mathbf{u}_{1-q}^\top \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & \mathbf{x}_{T-1}^\top & \dots & \mathbf{x}_{T-p}^\top & \mathbf{u}_T^\top & \dots & \mathbf{u}_{T-q}^\top \end{pmatrix} \qquad (T \times k)$$

$$\boldsymbol{\Gamma} = (\mathbf{c}, \mathbf{A}_1, \dots, \mathbf{A}_p, \mathbf{B}_0, \dots, \mathbf{B}_q)^\top \qquad (k \times n)$$

$$\mathbf{E} = \begin{pmatrix} \mathbf{e}_1^\top \\ \vdots \\ \mathbf{e}_T^\top \end{pmatrix} \qquad (T \times n)$$

and $k = 1 + np + m(q + 1)$. Using the fact that $\text{vec}(\mathbf{E})$ is normally distributed with covariance $\mathbf{\Sigma} \otimes \mathbf{I}_T$, the likelihood can be expressed as (see Appendix C.2)

$$\begin{aligned}\mathcal{L} &= \frac{1}{(2\pi)^{nT/2}|\mathbf{\Sigma}|^{T/2}}\text{exp}\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Sigma}^{-1}\mathbf{E}^{\mathsf{T}}\mathbf{E}\right]\right) \\ &= (2\pi)^{-nT/2}|\mathbf{\Sigma}|^{-T/2}\text{exp}\left(-\frac{1}{2}\text{tr}\left[\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{Z\Gamma})^{\mathsf{T}}(\mathbf{X}-\mathbf{Z\Gamma})\right]\right) \qquad (3.3) \\ &\Rightarrow \\ \text{log}\mathcal{L} &= -\frac{nT}{2}\text{log}(2\pi) - \frac{T}{2}\text{log}|\mathbf{\Sigma}| - \frac{1}{2}\text{tr}\left[\mathbf{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X} - 2\mathbf{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Z\Gamma} + \mathbf{\Sigma}^{-1}\mathbf{\Gamma}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z\Gamma}\right]\end{aligned}$$

Using the rules given in Appendix B.2, the partial derivatives are calculated as

$$\begin{aligned}\frac{\partial\text{log}\mathcal{L}}{\partial\mathbf{\Gamma}} &= -\frac{1}{2}\left[-2\mathbf{\Sigma}^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Z} + 2\mathbf{\Sigma}^{-1}\mathbf{\Gamma}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\right]^{\mathsf{T}} \\ \frac{\partial\text{log}\mathcal{L}}{\partial\mathbf{\Sigma}} &= -\frac{T}{2}\mathbf{\Sigma}^{-\mathsf{T}} + \frac{1}{2}\mathbf{\Sigma}^{-\mathsf{T}}(\mathbf{X}-\mathbf{Z\Gamma})(\mathbf{X}-\mathbf{Z\Gamma})^{\mathsf{T}}\mathbf{\Sigma}^{-\mathsf{T}}\end{aligned}$$

This leads to the maximum likelihood estimates

$$\begin{aligned}\widehat{\mathbf{\Gamma}} &= (\mathbf{Z}^{\mathsf{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{X} \\ \widehat{\mathbf{\Sigma}} &= \frac{1}{T}(\mathbf{X}-\mathbf{Z}\widehat{\mathbf{\Gamma}})^{\mathsf{T}}(\mathbf{X}-\mathbf{Z}\widehat{\mathbf{\Gamma}})\end{aligned}$$

3.2.2 Constrained Maximum Likelihood Estimation

Due to the large number of parameters, $n(1 + np + m(q + 1))$, in comparison to the typical number of historical financial time series observations, an unconstrained Maximum Likelihood estimation is unlikely to perform well for higher dimensions n and m. One solution could be to use Lasso (Tibshirani [1996]) or some other technique to mitigate the curse of dimensionality. However, when constructing an ESG, the ability to manually specify which parameters to include is often desirable. When modelling the returns of an ETF based on a large index together with a few small cap stocks, it is reasonable to set restrictions to only allow the stocks to depend on the ETF and not the other way around. When using e.g. Lasso, this is no guarantee, and which parameters are estimated to be zero may vary with the estimation window. The purpose of an ESG is not only to predict future returns, but also to provide a deeper understanding of the market dynamics. It is therefore argued that user provided parameter restrictions, possibly with the help of e.g. Lasso, to sparsify the parameter matrices is the most flexible solution.

A constrained maximum likelihood estimation can be done by considering the vectorization of the parameter matrix $\mathbf{\Gamma}$. The constraint is formulated as

$$\boldsymbol{\alpha} = \text{vec}(\mathbf{\Gamma}) = \mathbf{R}\boldsymbol{\gamma}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^{kn\times 1}$, $\mathbf{R} \in \mathbb{R}^{kn\times\iota}$, $\boldsymbol{\gamma} \in \mathbb{R}^{\iota}$ and ι denotes the number of unrestricted parameters (Lütkepohl [2005]). The matrix $\mathbf{R}$ determines which parameters are set to zero by having one entry in each column set to 1 and all the other entries in that column set to 0. As an example, consider $n = 2$, $m = 1$, $p = 2$ and $q = 1$, and the desired parameters

$$\mathbf{\Gamma} = \begin{pmatrix} c(1) & A_1(1,1) & A_1(1,2) & A_2(1,1) & 0 & B_0(1) & 0 \\ c(2) & 0 & A_1(2,2) & 0 & 0 & B_0(2) & B_1(2)\end{pmatrix}^{\mathsf{T}}$$

The restriction would then be formulated as

$$\boldsymbol{\alpha} = \begin{pmatrix} c(1) \\ A_1(1,1) \\ A_1(1,2) \\ A_2(1,1) \\ 0 \\ B_0(1) \\ 0 \\ c(2) \\ 0 \\ A_1(2,2) \\ 0 \\ 0 \\ B_0(2) \\ B_1(2) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c(1) \\ A_1(1,1) \\ A_1(1,2) \\ A_2(1,1) \\ B_0(1) \\ c(2) \\ A_1(2,2) \\ B_0(2) \\ B_1(2) \end{pmatrix}$$

Vectorizing equation (3.2) and inserting the restriction gives (see Appendix B.1)

$$\begin{aligned} \text{vec}(\mathbf{X}) &= (\mathbf{I}_n \otimes \mathbf{Z})\text{vec}(\boldsymbol{\Gamma}) + \text{vec}(\mathbf{E}) \\ &= (\mathbf{I}_n \otimes \mathbf{Z})\mathbf{R}\boldsymbol{\gamma} + \text{vec}(\mathbf{E}) \end{aligned}$$

Again, using the fact that $\text{vec}(\mathbf{E}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}_T)$ the likelihood can be expressed as

$$\mathcal{L} = \frac{1}{(2\pi)^{nT/2}|\boldsymbol{\Sigma} \otimes \mathbf{I}_T|^{1/2}} \exp\left(-\frac{1}{2}\left[(\text{vec}(\mathbf{X}) - (\mathbf{I}_n \otimes \mathbf{Z})\mathbf{R}\boldsymbol{\gamma})^\mathsf{T} (\boldsymbol{\Sigma} \otimes \mathbf{I}_T)^{-1} (\text{vec}(\mathbf{X}) - (\mathbf{I}_n \otimes \mathbf{Z})\mathbf{R}\boldsymbol{\gamma})\right]\right)$$

Using the rules given in Appendix B.1, this gives a log-likelihood

$$\begin{aligned} \log \mathcal{L} &= -\frac{nT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\left[(\text{vec}(\mathbf{X}) - (\mathbf{I}_n \otimes \mathbf{Z})\,\mathbf{R}\boldsymbol{\gamma})^\mathsf{T} (\boldsymbol{\Sigma} \otimes \mathbf{I}_T)^{-1} (\text{vec}(\mathbf{X}) - (\mathbf{I}_n \otimes \mathbf{Z})\,\mathbf{R}\boldsymbol{\gamma})\right] \\ &= -\frac{nT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| \\ &\quad - \frac{1}{2}\left[\left(\text{vec}(\mathbf{X})^\mathsf{T}\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T\right) - \boldsymbol{\gamma}^\mathsf{T}\mathbf{R}^\mathsf{T}\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}^\mathsf{T}\right)\right)(\text{vec}(\mathbf{X}) - (\mathbf{I}_n \otimes \mathbf{Z})\,\mathbf{R}\boldsymbol{\gamma})\right] \\ &= -\frac{nT}{2}\log(2\pi) - \frac{T}{2}\log|\boldsymbol{\Sigma}| \\ &\quad - \frac{1}{2}\left[\text{vec}(\mathbf{X})^\mathsf{T}\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{I}_T\right)\text{vec}(\mathbf{X}) - 2\boldsymbol{\gamma}^\mathsf{T}\mathbf{R}^\mathsf{T}\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}^\mathsf{T}\right)\text{vec}(\mathbf{X}) + \boldsymbol{\gamma}^\mathsf{T}\mathbf{R}^\mathsf{T}\left(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}^\mathsf{T}\mathbf{Z}\right)\mathbf{R}\boldsymbol{\gamma}\right] \end{aligned}$$

The partial derivative with respect to the parameter vector is then given by

$$\frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\gamma}} = \mathbf{R}^\mathsf{T}(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}^\mathsf{T})\text{vec}(\mathbf{X}) - \mathbf{R}^\mathsf{T}(\boldsymbol{\Sigma}^{-1} \otimes \mathbf{Z}^\mathsf{T}\mathbf{Z})\mathbf{R}\boldsymbol{\gamma}$$

yielding the maximum likelihood estimate

$$\hat{\boldsymbol{\alpha}} = \mathbf{R}\hat{\boldsymbol{\gamma}} = \mathbf{R}[(\mathbf{R}^\mathsf{T}(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{Z}^\mathsf{T}\mathbf{Z})\mathbf{R})^{-1}\mathbf{R}^\mathsf{T}(\widehat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{Z}^\mathsf{T})\text{vec}(\mathbf{X})]$$

The maximum likelihood estimate of the covariance matrix is again given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}(\mathbf{X} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}})^\mathsf{T}(\mathbf{X} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}})$$

Unfortunately, there is no analytic solution since the estimation of the coefficients is commingled with the estimation of the covariance matrix. Thus, an iterative procedure has to be applied. However, each estimation of the parameter vector $\hat{\boldsymbol{\alpha}}$ can be done in reasonable time unless ι is very large ($>> 1000$). If encountering long run-times, it is likely due to sub-optimal matrix chain multiplications – how the products are parenthesized matters! It is suggested that the iteration is initialized by choosing the covariance estimate generated by the unconstrained maximum likelihood parameters. Since the problem is convex, the iteration should be stable and converge to an optimum.

3.2.3 Volatility Clustering

Consider the model

$$\mathbf{x}_t = \mathbf{c} + \sum_{i=1}^{p} \mathbf{A}_i \mathbf{x}_{t-i} + \sum_{j=0}^{q} \mathbf{B}_j \mathbf{u}_{t-j} + v_t \mathbf{e}_t, \quad \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{3.4}$$

where all dependent variables are assumed to be driven by a global volatility process $\{v_t\}$. The representation in (3.2) can then be written as

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{v} \circ \mathbf{E}$$

where

$$\mathbf{v} = (v_1, \dots, v_T)^\mathsf{T}$$

and $\circ$ denotes element wise multiplication of the rows. Denoting $\tilde{\mathbf{E}} = \mathbf{v} \circ \mathbf{E}$, the error distribution is given by $\text{vec}(\tilde{\mathbf{E}}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{V}))$, where $\mathbf{V} = \text{diag}(\mathbf{v} \circ \mathbf{v})$. First consider the unconstrained estimation. From Appendix C.2, the likelihood in (3.3) now evaluates to

$$\begin{aligned}\mathcal{L} &= (2\pi)^{-nT/2} |\boldsymbol{\Sigma}|^{-T/2} |\mathbf{V}|^{-n/2} \text{exp}\left(-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{E}}^\mathsf{T}\mathbf{V}^{-1}\tilde{\mathbf{E}}\right]\right) \\ &= (2\pi)^{-nT/2} |\boldsymbol{\Sigma}|^{-T/2} |\mathbf{V}|^{-n/2} \text{exp}\left(-\frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma})\right]\right)\end{aligned}$$

giving the log-likelihood

$$\begin{aligned}\text{log}\mathcal{L} = &-\frac{nT}{2}\text{log}(2\pi) - \frac{T}{2}\text{log}|\boldsymbol{\Sigma}| - \frac{n}{2}\text{log}|\mathbf{V}| \\ &- \frac{1}{2}\text{tr}\left[\boldsymbol{\Sigma}^{-1}\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{X} - 2\boldsymbol{\Sigma}^{-1}\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{Z}\boldsymbol{\Gamma} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{V}^{-1}\mathbf{Z}\boldsymbol{\Gamma}\right]\end{aligned}$$

and the partial derivatives are

$$\begin{aligned}\frac{\partial\text{log}\mathcal{L}}{\partial\boldsymbol{\Gamma}} &= -\frac{1}{2}\left[-2\boldsymbol{\Sigma}^{-1}\mathbf{X}^\mathsf{T}\mathbf{V}^{-1}\mathbf{Z} + 2\boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{V}^{-1}\mathbf{Z}\right]^\mathsf{T} \\ \frac{\partial\text{log}\mathcal{L}}{\partial\boldsymbol{\Sigma}} &= -\frac{T}{2}\boldsymbol{\Sigma}^{-\mathsf{T}} + \frac{1}{2}\boldsymbol{\Sigma}^{-\mathsf{T}}(\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma})\mathbf{V}^{-1}(\mathbf{X} - \mathbf{Z}\boldsymbol{\Gamma})^\mathsf{T}\boldsymbol{\Sigma}^{-\mathsf{T}}\end{aligned}$$

The maximum likelihood estimate of the parameters is thus the same as before, while the new maximum likelihood estimate of the covariance is given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{T}(\mathbf{X} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}})^\mathsf{T}\mathbf{V}^{-1}(\mathbf{X} - \mathbf{Z}\widehat{\boldsymbol{\Gamma}})$$

Some tedious calculations yield a similar result in the case with parameter constraints. Essentially, this means that if the innovations of all dependent variables are assumed to be driven by an historically observed global volatility process, the residuals can be standardized with the supplied volatility in the estimation procedure, and the new log-likelihood can be calculated by subtracting $\frac{n}{2}\text{log}|\mathbf{V}|$. The diagonal elements in $\mathbf{V}$ may be very small, leading to the determinant being evaluated as zero in practice. This can be overcome by applying LU-factorization, and some programming languages provide robust implementations for calculating log-determinants.

3.2.4 Graphical Lasso

The Graphical Lasso estimator is a sparse penalized maximum likelihood estimator for the precision matrix (inverse covariance matrix) of a multivariate elliptical distribution. When the sample size is relatively small in comparison to the number of features, the non-penalized Maximum Likelihood estimator of the covariance matrix is likely inducing spurious dependencies, and Graphical Lasso solves this issue by penalizing the sum of absolute values of the precision matrix. The Graphical Lasso estimator of the precision matrix is given by

$$\arg\max_{\mathbf{\Theta}} \left(\log|\mathbf{\Theta}| - \mathrm{tr}(\mathbf{S}\mathbf{\Theta}) - \rho||\mathbf{\Theta}||_1\right) \tag{3.5}$$

where $\mathbf{\Theta}$ is positive semi definite and $\mathbf{S}$ is the empirical covariance matrix (Friedman et al. [2008]). In this case the empirical covariance matrix (of the residuals) is the Maximum Likelihood estimate. The objective function in (3.5) is the (penalized) Gaussian log-likelihood of the data, partially maximized with respect to the parameters. To see why this is the case, take the logarithm of (3.3) and remove the constant, yielding

$$-\frac{T}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\mathrm{tr}\left[\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{Z}\mathbf{\Gamma})^{\mathsf{T}}(\mathbf{X}-\mathbf{Z}\mathbf{\Gamma})\right] \propto \log|\mathbf{\Sigma}^{-1}| - \mathrm{tr}\left[\mathbf{\Sigma}^{-1}\frac{1}{T}(\mathbf{X}-\mathbf{Z}\mathbf{\Gamma})^{\mathsf{T}}(\mathbf{X}-\mathbf{Z}\mathbf{\Gamma})\right]$$

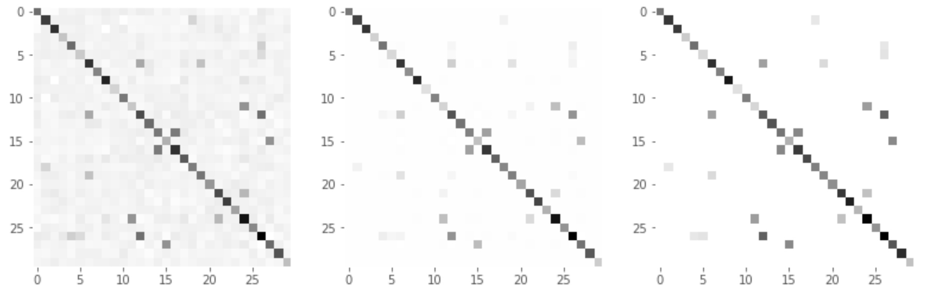
Partially maximizing with respect to $\mathbf{\Gamma}$, and denoting

$$\mathbf{S} = \widehat{\mathbf{\Sigma}} = \frac{1}{T}(\mathbf{X}-\mathbf{Z}\widehat{\mathbf{\Gamma}})^{\mathsf{T}}(\mathbf{X}-\mathbf{Z}\widehat{\mathbf{\Gamma}})$$

and $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$, this simplifies to

$$\log|\mathbf{\Theta}| - \mathrm{tr}(\mathbf{S}\mathbf{\Theta})$$

The penalizing parameter ρ determines the sparsity of the resulting estimate. Figure 3.3 illustrates the result of applying Graphical Lasso with $\rho = 0.1$ to the sample covariance estimated from 300 observations of a 30-dimensional multivariate normal distribution.



(a) Empirical covariance **(b)** GLasso estimate **(c)** True covariance

Figure 3.3: Illustration of Graphical Lasso estimation of the covariance matrix

3.2.5 Parameter Inference

Under some certain conditions, the parameters are asymptotically normally distributed (Lütkepohl [2005]). However, for high dimensional data with relatively few observations, a bootstrapping procedure may be more reliable, especially if the normality assumption is violated. While multiple bootstrapping methods for time-series have been applied in the literature, the method proposed below is based on the one outlined by Lütkepohl [2005], Appendix D.3. Consider the parameter estimates $\widehat{\mathbf{\Gamma}}$ and the residuals $\hat{\mathbf{e}}_{1:T}$ obtained from one of the estimation procedures above. Also consider some quantity of interest $\hat{y} = y(\widehat{\mathbf{\Gamma}})$. The usual residual bootstrap procedure then proceeds as:

[1] Estimate the parameters $\widehat{\boldsymbol{\Gamma}}$ given $\mathbf{x}_{1:T}$ and an appropriate presample, where $\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_T$ are the resulting residuals

[2] Compute the centered residuals $\hat{\mathbf{e}}_1 - \bar{\mathbf{e}}, \ldots, \hat{\mathbf{e}}_T - \bar{\mathbf{e}}$, where $\bar{\mathbf{e}} = \sum \frac{1}{T}\hat{\mathbf{e}}_t$. Obtain bootstrap residuals $\mathbf{e}^*_{1:T}$ by randomly drawing with replacement from the centered residuals

[3] Calculate a bootstrap time series as $\mathbf{x}^*_t = \hat{\mathbf{c}} + \sum_{i=1}^{p} \hat{\mathbf{A}}_i \mathbf{x}^*_{t-i} + \sum_{j=0}^{q} \hat{\mathbf{B}}_j \mathbf{u}_{t-j} + \mathbf{e}^*_t$, initialized by the same presample as in the estimation procedure.

[4] Re-estimate the parameters given the bootstrap time series to obtain $\boldsymbol{\Gamma}^*$

[5] Calculate a bootstrap version, y^*, of the quantity of interest

[6] Repeat [2]-[5] to obtain a sufficiently large sample $y^*_{1:N}$

Of course, the procedure above is only valid for serially uncorrelated residuals. For parameter confidence intervals, the quantities of interest are simply $\hat{y}_{ij} = \widehat{\boldsymbol{\Gamma}}_{ij}$.

3.2.6 Selection of Hyperparameters

Possible hyper parameters in the estimation procedures above are p, q, $\mathbf{R}$ and ρ. How to set the Graphical Lasso penalization parameter ρ will not be discussed, but the Python library scikit-learn provides an implementation of Graphical Lasso with cross-validation (scikit-learn developers [2019]).

If there are no parameter restrictions, p and q can easily be selected according to e.g. AIC. When including parameter restrictions however, restrictions must be set for each combination of p and q. For certain parameters, there may be clear restrictions according to economic theory. However, there may also be a large number of parameters which are allowed to vary in theory, but where imposing restrictions may yield better results.

One approach is to

[1] Specify some requirements on the restrictions, e.g. always include intercept or always set certain restrictions to zero. Economic theory should be taken into consideration.

[2] For each combination of p and q, find the restriction matrix $\mathbf{R}$ which gives the lowest AIC among matrices fulfilling the requirements in [1].

If $\mathbf{R}$ is large, brute force estimation in [2] where each combination of possible restrictions is tested is infeasible. Instead, one could use Lasso, which optimizes the parameters by minimizing the sum of squared residuals with a penalty on the absolute value of the coefficients (Tibshirani [1996]). This will result in a number of zero coefficients, the number of which depends on a provided penalty parameter. A specified number of values of the penalty parameter can then be tested for each combination of p and q by combining the result with [1] and calculating AIC. With this approach, there is no guarantee of finding the optimal (in terms of AIC) parameter restriction matrix. However, the algorithm can give an initial restriction matrix with a relatively short run-time. The parameters could then further be tested for significance with the bootstrap procedure outlined in Section 3.2.5, or the restriction matrix further optimized by e.g. random perturbations.

3.3 Defining Views

Following the method proposed in Section 2.2, the user should provide $\boldsymbol{\psi}_t \in \mathbb{R}^d$, $\mathbf{H}_t \in \mathbb{R}^{d \times n}$ and the distribution of $\boldsymbol{\xi}_t \in \mathbb{R}^d$. For simplicity, the linear process given in (2.8) will be considered – it was shown in Section 2.1 that the model given in (3.1) can be represented similarly. In the most general case, i.e. where the modelled process may be non-linear and/or non-Gaussian, the following methods would have to be extended, although the same underlying reason could be applied. The same goes for the Dynamic Nelson-Siegel model. Thus, the system is assumed to be given by

$$\begin{aligned} \boldsymbol{\psi}_t &= \mathbf{H}_t\boldsymbol{\mu}_t + \boldsymbol{\xi}_t && \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t) \\ \boldsymbol{\mu}_t &= \mathbf{c} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t && \\ \mathbf{x}_t &= \boldsymbol{\mu}_t + \mathbf{e}_t && \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t) \end{aligned} \tag{3.6}$$

Providing $\boldsymbol{\psi}_t$ and $\mathbf{H}_t$ is theoretically straightforward (how to set $\boldsymbol{\psi}_t$ in practice is discussed in Section 3.3.2). Consider an example where the log returns of 5 investable assets are modelled, i.e. $\mathbf{x}_t \in \mathbb{R}^5$. The inputs

$$\begin{aligned} \boldsymbol{\psi}_t &= (0.0004, 0.0003, 0.00015)^\mathsf{T} \\ \mathbf{H}_t &= \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \end{aligned}$$

would then imply views that the conditional expectation of the log-returns of the first, third and fourth asset at time t are expected to be 4, 3 and 1.5 basis points respectively, while

$$\begin{aligned} \boldsymbol{\psi}_t &= (0.0001, 0.00005)^\mathsf{T} \\ \mathbf{H}_t &= \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 \end{pmatrix} \end{aligned}$$

would imply views that the conditional expectation of the log-return of the first asset is one basis point higher than that of the second asset and that the log-return of the third asset is 0.5 basis points higher than that of the fifth asset at time t.

3.3.1 Setting the Covariance Matrix $\boldsymbol{\Omega}_t$

The covariance matrix is more difficult to define. The interpretation of $\boldsymbol{\Omega}_t$ is that it defines the covariance of the distribution of $\boldsymbol{\psi}_t$ around the *true conditional* expectation of the specified linear combination of variables. While this provides flexibility for someone with a strong opinion about the future covariance of the views, a reasonable assumption could be that it relates to the covariance of the forecast of the conditional expectation indicated by the baseline model. The forecast of the conditional expectation means the prediction made today about the expectation at time t given an observation at $t-1$. It will become clearer below, but the covariance of the forecast of the conditional expectation would be the forecast error covariance of the value minus the noise covariance. In the following proposed value of $\boldsymbol{\Omega}_t$, it is assumed that

[1] $\boldsymbol{\Omega}_t$ is a function of the covariance of the forecast of the conditional expectation indicated by the baseline model

[2] The trust in the views in relation to the trust in the model is a deterministic function of the number of steps into the future, possibly constant

The effects of macroeconomic shocks, such as the one induced by Covid-19 in 2020, will have a limited effect on the long-term behavior of the process (decades ahead), and one may argue that the process fitted to the historical data will provide the best estimate far into the future. This motivates why the relative trust in the views may vary over the forecast horizon in assumption [2]. For long-term views however, one may exclude dependence on the number of steps into the future, h, since the trust in the views in comparison to the trust in the baseline model would likely be similar 40 years into the future as 50 years into the future.

Denote $\text{Cov}[\boldsymbol{\mu}_{t=t_0+h}] = \tilde{\boldsymbol{\Sigma}}_{t=t_0+h}$ the forecast error covariance of the conditional expectation at time t assuming no views, where t is h steps into the future from today. To clarify what this means, assume that today's values of the dependent variables are $\mathbf{x}_0$, i.e. $t_0 = 0$. The conditional expectation of $\mathbf{x}_1$ is given by

$$\boldsymbol{\mu}_1 = \mathbf{c} + \mathbf{A}\mathbf{x}_0 + \mathbf{B}\mathbf{u}_1$$

and is thus deterministic, i.e. $\tilde{\boldsymbol{\Sigma}}_{t=1} = \mathbf{0}$. However, $\mathbf{x}_1$ is uncertain with covariance $\boldsymbol{\Sigma}_1$. Thus, the covariance of the forecasted conditional expectation at time $t = 2$ is given by

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_{t=2} &= \text{Cov}[\mathbf{c} + \mathbf{A}\mathbf{x}_1 + \mathbf{B}\mathbf{u}_2] \\ &= \mathbf{A}\text{Cov}[\mathbf{x}_1]\mathbf{A}^\intercal \\ &= \mathbf{A}\boldsymbol{\Sigma}_1\mathbf{A}^\intercal\end{aligned}$$

Recall Section 2.3, where a very small value was added to the diagonal elements of the covariance of the predictive density. The same must be done in this case, and the reason why will become apparent below. Thus, the forecast error covariance of the conditional expectation can be calculated recursively as

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_{t=t_0+h} &= \mathbf{A}\text{Cov}[\mathbf{x}_{t=t_0+h-1}]\mathbf{A}^\intercal + \epsilon\mathbf{I}_n \\ \text{Cov}[\mathbf{x}_{t=t_0+h}] &= \begin{cases}\mathbf{0} & \text{if } h = 0 \\ \mathbf{A}\text{Cov}[\mathbf{x}_{t=t_0+h-1}]\mathbf{A}^\intercal + \boldsymbol{\Sigma}_t & \text{if } h > 0\end{cases}\end{aligned}$$

The small value ϵ must be the same value as used in the filtering procedure for consistency. Now that the uncertainty of the forecasted conditional expectation implied by the baseline model has been specified, it is proposed that $\boldsymbol{\Omega}_t$ is set to

$$\boldsymbol{\Omega}_t = \begin{cases}\tau(h)\mathbf{H}_t\tilde{\boldsymbol{\Sigma}}_{t=t_0+h}\mathbf{H}_t^\intercal & \text{if } t \in I_\psi \\ \infty\mathbf{I}_d, & \text{otherwise}\end{cases} \tag{3.7}$$

where I_ψ is the interval, belonging to the forecast horizon, on which the views are defined. This means that the covariance $\boldsymbol{\Omega}_t$ is set to infinity, resulting in no Kalman gain, outside of the interval where the views are set. There is now an interpretation of $\boldsymbol{\Omega}_t$ – it is the baseline model covariance of the forecasted conditional expectation of the linear combination of variables where views are specified, multiplied by a scalar determined by the function $\tau(h)$. This reduces the non-trivial task of choosing an appropriate covariance matrix of the views to the choice of a scalar-valued function $\tau(h)$ describing the relative uncertainty of the views to the uncertainty of the baseline model over a specified interval belonging to the forecast horizon. To further understand how the estimated density will evolve given this specification of $\boldsymbol{\Omega}_t$, recall the filtering procedure in Section 2.3, where

$$\begin{aligned}\boldsymbol{\mu}_{t|t-1} &= \mathbf{c} + \mathbf{A}\hat{\boldsymbol{\mu}}_{t-1|t-1} + \mathbf{B}\mathbf{u_t} \\ \mathbf{P}^\mu_{t|t-1} &= \mathbf{A}\mathbf{P}^x_{t-1|t-1}\mathbf{A}^\intercal + \epsilon\mathbf{I}_n \\ \mathbf{K}_t &= \mathbf{P}^\mu_{t|t-1}\mathbf{H}_t^\intercal(\mathbf{H}_t\mathbf{P}^\mu_{t|t-1}\mathbf{H}_t^\intercal + \boldsymbol{\Omega}_\mathbf{t})^{-1} \\ \hat{\boldsymbol{\mu}}_{t|t} &= \hat{\boldsymbol{\mu}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\psi}_t - \mathbf{H}_t\hat{\boldsymbol{\mu}}_{t|t-1}) \\ \mathbf{P}^x_{t|t} &= \mathbf{P}^\mu_{t|t-1} - \mathbf{K}_t\mathbf{H}_t\mathbf{P}^\mu_{t|t-1} + \boldsymbol{\Sigma}_t\end{aligned}$$

Thus, the Kalman gain at time $t \in I_\psi$ given the matrix in (3.7) is

$$\mathbf{K}_t = \mathbf{P}^\mu_{t|t-1}\mathbf{H}_t^\intercal(\mathbf{H}_t\mathbf{P}^\mu_{t|t-1}\mathbf{H}_t^\intercal + \tau(h)\mathbf{H}_t\tilde{\boldsymbol{\Sigma}}_{t=t_0+h}\mathbf{H}_t^\intercal)^{-1}$$

and zero for $t \notin I_\psi$. The initial covariance of $\mathbf{x}_{t_0}$ is set to $\mathbf{P}^x_{t_0|t_0} = \text{Cov}[\mathbf{x}_{t_0}] = \mathbf{0}$. Thus

$$\begin{aligned}\mathbf{K}_{t_0+1} &= \epsilon \mathbf{I}_n \mathbf{H}^{\mathsf{T}}_{t_0+1}(\mathbf{H}_{t_0+1}\epsilon\mathbf{I}_n\mathbf{H}^{\mathsf{T}}_{t_0+1} + \tau(h)\mathbf{H}_{t_0+1}\epsilon\mathbf{I}_n\mathbf{H}^{\mathsf{T}}_{t_0+1})^{-1} \\ &= \frac{1}{(1+\tau(1))}\mathbf{H}^{\mathsf{T}}_{t_0+1}(\mathbf{H}_{t_0+1}\mathbf{H}^{\mathsf{T}}_{t_0+1})^{-1}\end{aligned}$$

This means that the Kalman gain at time $t_0 + 1$ is inversely proportional to $1 + \tau(1)$. It is also noted that since $\tilde{\mathbf{\Sigma}}_{t=t_0+h} \geq \mathbf{P}^\mu_{t|t-1}$ for all $t > t_0$, the Kalman gain would reduce over the forecast horizon until reaching a steady-state if $\tau(h)$ was constant. This is due to the fact that the uncertainty in the predictive density is lower than the uncertainty indicated by the model (or the views) alone. Deriving the Kalman gain at a given time for general matrices $\mathbf{A}$ and $\tau(\cdot)$ is difficult. For relatively small magnitudes of the elements in $\mathbf{A}$ however, the Kalman gain will always be approximately inversely proportional to $(1+\tau(h))$, or specifically $\mathbf{H}_t\mathbf{K}_t(1+\tau(h)) \lesssim 1$ where $\lesssim$ would mean less than and almost equal. The larger the magnitude of the autoregressive coefficients, the smaller the Kalman gain. For a more extensive description of the Kalman filter and its steady-state, the reader is referred to the work of Crassidis and Junkins [2012].

This gives two interpretations of $\tau(h) \in [0, \infty)$

[1] $\tau(h)$ is the relative uncertainty of the views and the baseline model prediction of the conditional expectation, where $\tau(h) = 1$ would indicate similar uncertainties at time $t = t_0 + h$

[2] The Kalman gain at time $t = t_0 + h$ is approximately proportional to $\frac{1}{\tau(h)+1}$

In Chapter 5, the prospect of having historical data of views is discussed. In a situation where $\tau(h)$ is tuned, or where biases are estimated, it would be appropriate to choose a simple function e.g. $\tau(h) = \tau_0$, $\tau(h) = \tau_0 h$ or $\tau(h) = \tau_0 \log(h)$. For a more in-depth discussion about the case where historical views are available, the reader is referred to Section 5.4.

3.3.2 Setting the Views $\boldsymbol{\psi}_t$

In this section, some practical considerations when setting the views $\boldsymbol{\psi}_t$ throughout the forecast horizon are discussed. Some informal requirements on the views are proposed. These are based on reason, and do not relate to the stability of the filter procedure. First, consider what properties would be expected from a forecast generated by the model with incorporated views. One may expect that the resulting forecast is reasonably close to the baseline model prediction. Secondly, one would expect that the behavior of the forecast is not changing dramatically from one time step to another. Considering that the model includes external regressors with generic models, there is no analytically tractable expression for the forecasted mean of the baseline model in the general case. However, it is assumed that routines are implemented to calculate the approximate h-step prediction interval from the baseline model.

In principle, there are no restrictions on the general form of the views, and the user could provide a sine wave if that is the belief. Since the attributes of what could be seen as a ”reasonable” view will vary depending on which variable is modelled, the curvature of the views is not discussed, and it should be decided by the user according to economic theory. However, two informal restrictions are proposed. Let $t = t_0$ denote today and denote N the length of the forecast horizon $I_N = \{t_0+1, t_0+2, \dots, t_0+N\}$. Let $I_\psi = \{t_0+a, \dots, t_0+b\} \subseteq I_N$ be the interval on which the views are given. It is required that views are set at each time step within I_ψ, i.e. intervals such as $I_\psi = \{t_0+1, t_0+2, t_0+4, \dots\}$ are not allowed. Consider a baseline model prediction region $I_{\boldsymbol{\mu}_t,\alpha}$ at a specified significance level, α, centered at $\bar{\boldsymbol{\mu}}_t$. Note that $\boldsymbol{\mu}_t$ is the conditional expectation given some path until that point, and the uncertainty of $\boldsymbol{\mu}_t$ far into the future may be approximately the same as the uncertainty of $\mathbf{x}_t$ for some models, whereas $\bar{\boldsymbol{\mu}}_t$ is the mean of the conditional expectation at a given time calculated from a large number of simulated paths, i.e. the unconditional mean. The proposed restrictions are then

$$\boldsymbol{\psi}_t \in I_{\boldsymbol{\mu}_t,\alpha} \tag{3.8}$$

$$\begin{cases} ||\boldsymbol{\psi}_{t_0+a} - \mathbf{H}_{t_0+a}\bar{\boldsymbol{\mu}}_{t_0+a}||_2 < \epsilon_2 & \forall a \\ ||\boldsymbol{\psi}_{t_0+b} - \mathbf{H}_{t_0+b}\bar{\boldsymbol{\mu}}_{t_0+b}||_2 < \epsilon_2 & \text{if} \quad b < N \end{cases} \tag{3.9}$$

where ϵ_2 is some small value in relation to the magnitude of the values of $\{\boldsymbol{\mu}_t\}$. Setting the values of ϵ_2 and α is left to the implementer, and the rationale behind the restrictions is instead discussed. Restriction (3.8) is simply saying that the view of the expected value at a given time is reasonably close to the forecast from the baseline model. If a view lies outside, say, the 90% quantile of the baseline model prediction, one may consider reevaluating the model choice. In some cases, this restriction may have to be loosened. Consider a random walk where the conditional expectation is always zero. Restriction (3.9) means that if the views are set on an interval which does not end at the last time step, the transition should be smooth. Likewise, the transition into the region where views are imposed should be smooth, whether or not the interval starts at the next time step or further into the forecast horizon. These restrictions alone are obviously no guarantee for "reasonable" views, and the user needs to take into account variable specific theory and assumptions.

Two approaches upon which a user interface could be based are discussed. Neither of them by themselves guarantee that the restrictions above are fulfilled, meaning that some alterations to make sure that the bound and end point conditions are satisfied would have to be implemented. Furthermore, these approaches are assuming that the views are specified unconditionally, i.e. independent of the path until that point.

3.3.2.1 Nelson-Siegel Approach

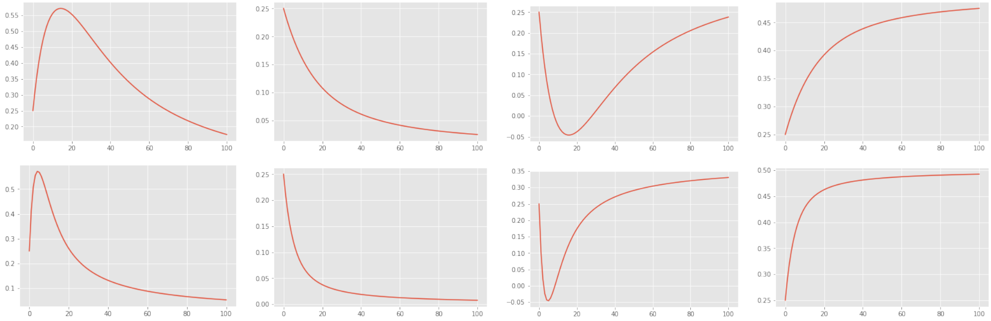
Often the views may either be beliefs about the long-term mean, where the view could be expressed as a gradual change from the long-term mean indicated by the baseline model to the new mean, or beliefs about a short-term shock. This could for example be a shorter period of increased volatility and decreased mean returns, in other words a short-term shock to the market. In this case a view of the expected return could be described by some sort of short-term negative impulse, after which the expectation gradually returns to its long-term value.

One proposed method for constructing views of these types is the Nelson-Siegel curve, already introduced in Section 2.4.

The curve would now be given by

$$y(h) = \beta_0 + (\beta_1 + \beta_2)\frac{1 - \exp(-h/\tau)}{h/\tau} - \beta_2 \exp(-h/\tau)$$

where h again denotes the number of time steps into the forecast horizon (maturity in the original yield-curve model). Given some data points, this curve can be fitted by e.g. OLS. If there is no data, the Nelson-Siegel curve can still be used as a tool for expressing the views as a smooth function of h. As previously seen, the three parameters are describing a constant, an exponential decay and a "hump". This provides a framework for constructing views about both long-term mean and short-term shocks, as illustrated in Figure 3.4. In a practical application, the user may be asked to either use a slider for each parameter, or to enter, say, 50 points, upon which the least squares estimate of the curve is returned.



$[\beta_0, \beta_1, \beta_2] =$ $[0.0, 0.25, 1.5]$

$[\beta_0, \beta_1, \beta_2] =$ $[0.0, 0.25, 0.0]$

$[\beta_0, \beta_1, \beta_2] =$ $[0.37, -0.12, -1.2]$

$[\beta_0, \beta_1, \beta_2] =$ $[0.5, -0.25, 0.0]$

Figure 3.4: A few realizations of Nelson-Siegel curves. Top: $\tau = 10$, Bottom: $\tau = 3$.

3.3.2.2 Gaussian Process Regression

In some cases, the views may be more general than what could be described by a Nelson-Siegel curve. An example could be views about business cycles. A suggested practical approach for dealing with more general views is Gaussian process regression. By definition, a Gaussian process "is a collection of random variables, any finite number of which have (consistent) joint Gaussian distributions" (Rasmussen [2003]), which means that a GP describes a distribution over functions and can be seen as an extension of the multivariate Gaussian distribution to infinite dimensionality. A GP is completely described by a mean function and a covariance function, given by

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\left[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))\right] \end{aligned}$$

and denoted $f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$. A common choice is $m(x) = 0$ and $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2l}|\mathbf{x} - \mathbf{x}'|^2\right)$ where l is a length-scale factor. This is denoted the squared-exponential covariance function, squared-exponential kernel or Radial basis function kernel. In this case, the GP would be univariate and defined over time, and thus given by $f_i(t) = \psi_{i,t} \sim \mathcal{GP}\left(m_i(t), k_i(t, t+\tau)\right)$ where $i = 1, \ldots, d$. Furthermore, if the observations are considered noisy, the squared-exponential covariance function can be extended as

$$\mathrm{Cov}(\psi_{i,t}, \psi_{i,t+\tau}) = k(t, t+\tau) = \sigma_f^2 \exp\left(-\frac{1}{2l}|\tau|^2\right) + \sigma_n^2 \delta_\tau \tag{3.10}$$

where δ_τ is the Kronecker delta function which is 1 if and only if $\tau = 0$ and zero otherwise. A detailed description of Gaussian processes is beyond the scope of this study, but the general idea is that given a set of known function outputs $f(t)$, unknown values $f^* = f(t^*)$ can be inferred from the joint probability

$$\begin{pmatrix} f \\ f^* \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu^* \end{pmatrix}, \begin{pmatrix} k(t,t) & k(t,t^*) \\ k(t^*,t) & k(t^*,t^*) \end{pmatrix}\right)$$

The covariance function, or kernel, defines the similarity measure between points, in this case the auto-covariance. There are many types of covariance functions with different characteristics. Furthermore, new covariance functions can be constructed by summing and multiplying kernels. This means that the GP can capture characteristics such as linear trends, quadratic trends, seasonality and randomness by combining kernel functions, making it a very flexible tool. In (3.10), a White noise kernel, $\sigma_n^2 \delta_\tau$, is added to the squared-exponential kernel (Rasmussen [2003]). Figure 3.5 shows a Gaussian process fitted to 100 data points.

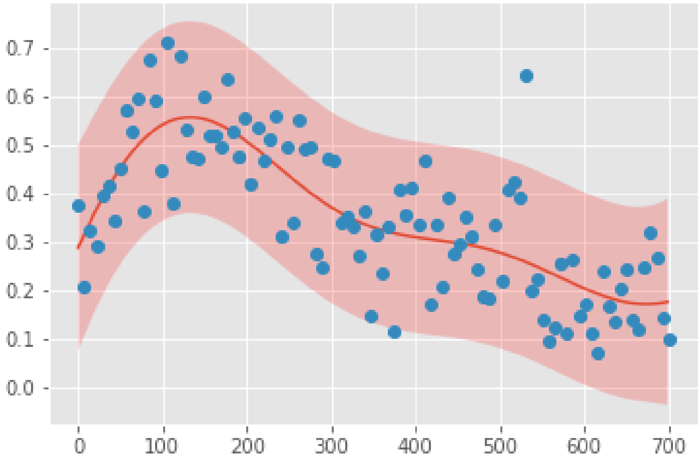


Figure 3.5: Gaussian process regression with $m(x) = 0$ and the kernel function given in (3.10) with $l = 100$, $\sigma_f = 1$ and $\sigma_n = 0.15$. Estimate and 95% prediction interval.

In a practical implementation, the user could interactively place points and see how the distribution of $\{\psi_{i,t}\}$ evolves over the forecast horizon. The covariance matrix $\mathbf{\Omega}_t$ could then either be calculated as in (3.7) or by constructing a diagonal matrix from the variances estimated by the GP regression for each individual view. This approach could also be used to include views from external sources with lower granularity, e.g. yearly outlooks of GDP produced by IMF.

3.3.2.3 Transformations of Views

There are mainly two common transformations of time series data, namely differentiation and the log-transformation. Regarding differentiation, the distribution is still Gaussian. This means that if one has views $\{\boldsymbol{\psi}_t^*\}$ of the expectation of variables which are differenced in the baseline model, the input views can simply be calculated as $\boldsymbol{\psi}_t = \boldsymbol{\psi}_t^* - \boldsymbol{\psi}_{t-1}^*$. The log-transformation is a bit more intricate. If

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$$

then $\mathbf{y} = \exp(\mathbf{x})$ has a log-normal distribution with expectation

$$\mathbb{E}[\mathbf{y}]_i = \exp(\mu_i + \frac{1}{2}\Sigma_{ii})$$

First consider a univariate time series, $\{y_t\}$, where the log-level is modelled as $x_t = \log(y) \sim \mathcal{N}(\mu_t, \sigma_t^2)$. Further assume that views ψ_t^* are formulated as the expected value of y_t. Since $\mathbb{E}[y_t] = \mathbb{E}[\exp(x_t)] = \exp(\mu + \frac{1}{2}\sigma^2)$, the views should then be transformed as $\psi_t = \log(\psi_t^*) - \frac{1}{2}\sigma^2$. In the multivariate case, the views would be transformed as

$$\boldsymbol{\psi}_t = \log(\boldsymbol{\psi}_t^*) - \frac{1}{2}\text{diag}(\mathbf{H}_t\mathbf{\Sigma}_t\mathbf{H}_t^\intercal)$$

since the views are specified as the expected value of a linear combination of the conditional expectation, $\mathbf{H}_t\boldsymbol{\mu}_t$, where $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_t, \mathbf{\Sigma}_t)$.

Another common way of expressing views is annualized returns over different horizons, e.g. five, ten and twenty years. Here, one method would be to calculate the implied mean prices at the end of each horizon as

$$\mathbf{p}_{t_0+h} = \mathbf{p}_{t_0}(1 + \mathbf{y}_{t_0+h})^h$$

where $\mathbf{y}_{t_0+h}$ denotes the h-year annualized return and $\mathbf{p}_t$ the price at time t. These values can then be interpolated with a method of choice, after which views of log-returns can be constructed. The transformed views would of course in this case indicate a stronger assumption than the original views, and these types of transformations should therefore be handled carefully.

3.4 Generating Scenarios

This section gives a walk-through of how the filtering procedure given in Section 2.3 can be applied to the model given in (3.1). In cases where the external regressors are also variables of interest themselves, as in the example ESG in Section 3.1, it may be relevant to impose views on them as well as the dependent variables. However, the proposed method would have to be extended to allow for views on external regressors, and the possibilities and difficulties of this are discussed in Section 3.4.2.

3.4.1 Views on Dependent Variables

Consider the VARX(1,0) representation given in (2.4) of the VARX(p,q) model in (3.1). By constructing the matrix $\mathbf{\Lambda}_t = \big(\mathbf{H}_t, \mathbf{0}, \ldots, \mathbf{0}\big) \in \mathbb{R}^{d\times np}$, the system

$$\begin{aligned}
\boldsymbol{\psi}_t &= \mathbf{H}_t\boldsymbol{\mu}_t + \boldsymbol{\xi}_t & \boldsymbol{\xi}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_t)\\
\boldsymbol{\mu}_t &= \mathbf{c} + \sum_{i=1}^{p}\mathbf{A}_i\mathbf{x}_{t-i} + \sum_{j=0}^{q}\mathbf{B}_j\mathbf{u}_{t-j} & &\\
\mathbf{x}_t &= \boldsymbol{\mu}_t + \mathbf{e}_t & \mathbf{e}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t)
\end{aligned}$$

has the representation

$$\begin{aligned}
\boldsymbol{\psi}_t &= \mathbf{\Lambda}_t\tilde{\boldsymbol{\mu}}_t + \boldsymbol{\xi}_t & \boldsymbol{\xi}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_t)\\
\tilde{\boldsymbol{\mu}}_t &= \tilde{\boldsymbol{c}} + \mathbf{A}\tilde{\boldsymbol{x}}_{t-1} + \mathbf{B}\tilde{\boldsymbol{u}}_t & &\\
\tilde{\boldsymbol{x}}_t &= \tilde{\boldsymbol{\mu}}_t + \tilde{\boldsymbol{e}}_t & \tilde{\boldsymbol{e}}_t &\sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(1,0,\ldots,0)\otimes\mathbf{\Sigma}_t)
\end{aligned}$$

where $\tilde{\boldsymbol{\mu}}_t = \mathrm{vec}(\boldsymbol{\mu}_t, \boldsymbol{\mu}_{t-1}, \ldots, \boldsymbol{\mu}_{t-p+1})$ and everything else is defined as in (2.4). This dynamic system of equations can be treated with a modified Kalman filter as seen in Section 2.3. Denoting $\mathbf{Q}_t$ the $np \times np$ covariance matrix $\mathrm{diag}(1,0,\ldots,0)\otimes\mathbf{\Sigma}_t$, the filter recursion is thus given by

$$\begin{aligned}
\hat{\tilde{\boldsymbol{\mu}}}_{t|t-1} &= \tilde{\boldsymbol{c}} + \mathbf{A}\hat{\tilde{\boldsymbol{\mu}}}_{t-1|t-1} + \mathbf{B}\tilde{\boldsymbol{u}}_t\\
\mathbf{P}^{\mu}_{t|t-1} &= \mathbf{A}\mathbf{P}^{x}_{t-1|t-1}\mathbf{A}^{\mathsf{T}} + \epsilon\mathbf{I}_{np}\\
\mathbf{K}_t &= \mathbf{P}^{\mu}_{t|t-1}\mathbf{\Lambda}_t^{\mathsf{T}}(\mathbf{\Lambda}_t\mathbf{P}^{\mu}_{t|t-1}\mathbf{\Lambda}_t^{\mathsf{T}} + \mathbf{\Omega_t})^{-1}\\
\hat{\tilde{\boldsymbol{\mu}}}_{t|t} &= \hat{\tilde{\boldsymbol{\mu}}}_{t|t-1} + \mathbf{K}_t(\boldsymbol{\psi}_t - \mathbf{\Lambda}_t\hat{\tilde{\boldsymbol{\mu}}}_{t|t-1})\\
\mathbf{P}^{x}_{t|t} &= \mathbf{P}^{\mu}_{t|t-1} - \mathbf{K}_t\mathbf{\Lambda}_t\mathbf{P}^{\mu}_{t|t-1} + \mathbf{Q}_t
\end{aligned}$$

and, to facilitate implementation, the dimensions of all vectors and matrices in the filter recursion are given in Table 3.1.

What remains to specify is how to set $\mathbf{\Omega}_t$ according to (3.7) in this case. This can now done with the recursion

$$\begin{aligned}
\tilde{\mathbf{\Sigma}}_{t=t_0+h} &= \left[\mathbf{A}\mathrm{Cov}\left[\tilde{\boldsymbol{x}}_{t=t_0+h-1}\right]\mathbf{A}^{\mathsf{T}}\right]_{1:n,1:n} + \epsilon\mathbf{I}_n\\
\mathrm{Cov}[\tilde{\boldsymbol{x}}_{t=t_0+h}] &= \begin{cases}\mathbf{0} & \text{if } h = 0\\ \mathbf{A}\mathrm{Cov}[\tilde{\boldsymbol{x}}_{t=t_0+h-1}]\mathbf{A}^{\mathsf{T}} + \mathbf{Q}_t & \text{if } h > 0\end{cases}
\end{aligned}$$

Object(s)	Dimensions
$\hat{\tilde{\boldsymbol{\mu}}}_{t\|t-1}$, $\hat{\tilde{\boldsymbol{\mu}}}_{t\|t}$, $\tilde{\boldsymbol{c}}$	$np \times 1$
$\tilde{\boldsymbol{u}}_t$	$m(q+1) \times 1$
$\mathbf{A}$	$np \times np$
$\mathbf{B}$	$np \times m(q+1)$
$\mathbf{P}^{\mu}_{t\|t-1}$, $\mathbf{P}^{x}_{t\|t}$, $\mathbf{Q}_t$	$np \times np$
$\mathbf{K}_t$	$np \times d$
$\boldsymbol{\psi}_t$	$d \times 1$
$\boldsymbol{\Lambda}_t$	$d \times np$

Table 3.1: Dimensions of all objects in the filter recursion

where $\mathbf{M}_{1:n,1:n}$ would denote the $n \times n$ matrix consisting of the first n rows and columns of $\mathbf{M}$. This is due to the fact that $\text{Cov}[\mathbf{x}_t] = \text{Cov}\,[\tilde{\boldsymbol{x}}_{t,1:n}]$.

3.4.2 Views on External Regressors

The external regressors are modelled externally and if there are no views of $\mathbf{u}_t$, these could simply be sampled from their respective processes and the filtering procedure for the dependent variables outlined in the previous section would be done conditional on each sample. Incorporating views when estimating the future *distribution* of the external regressors does not pose an issue. If $\mathbf{u}_t$ are simulated from one or multiple autoregressive processes with Gaussian innovations, the same filtering approach as outlined in Section 2.3 can be applied. If a process is non-linear or non-Gaussian, the methods mentioned in Section 2.2 could be applied. All external regressors are in this case modelled as AR(p) processes. The same way as a VAR(p) process can be represented as a VAR(1) process, the AR(p) model can be reformulated as an VAR(1) model. The representation is given by

$$
\begin{aligned}
&u_t = \gamma + \alpha_1 u_{t-1} + \cdots + \alpha_p u_{t-p} + \eta_t \\
&\iff \\
&\tilde{u}_t = \tilde{\gamma} + \tilde{\alpha}\tilde{u}_{t-1} + \tilde{\eta}_t
\end{aligned}
$$

where u_t is one of the external regressors and

$$
\begin{aligned}
\tilde{u}_t &= (u_t, \ldots, u_{t-p+1})^{\mathsf{T}} && (p \times 1) \\
\tilde{\gamma} &= (\gamma, 0, \ldots, 0)^{\mathsf{T}} && (p \times 1) \\
\tilde{\alpha} &= \begin{pmatrix} \alpha_1 & \alpha_2 & \ldots & \alpha_{p-1} & \alpha_p \\ 1 & 0 & \ldots & 0 & 0 \\ 0 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \ldots & 1 & 0 \end{pmatrix} && (p \times p) \\
\tilde{\eta}_t &= (\eta_t, 0, \ldots, 0)^{\mathsf{T}} && (p \times 1)
\end{aligned}
$$

Thus, the approach outlined in Section 2.3 could be applied when estimating the future distribution of these variables given views. The issue is that while the distribution of the values at any given time is given by the filter density, it is not possible to sample paths given the views. It may be tempting to sample from the process (assuming no views on the dependent variables) as

$$
\tilde{\boldsymbol{x}}_t = \tilde{\boldsymbol{c}} + \mathbf{A}\tilde{\boldsymbol{x}}_{t-1} + \mathbf{B}\tilde{\boldsymbol{u}}_{t|t} + \tilde{\boldsymbol{e}}_t
$$

where $\tilde{\boldsymbol{u}}_{t|t} \in \mathbb{R}^{m(q+1)\times 1}$ is sampled from the posterior density of $\tilde{\boldsymbol{u}}_t$ given views, and the variables are defined as in the previous section. This is however not a valid approach. Assume for simplicity that the external regressors are all modelled as AR(1) processes, or possibly together as a VAR(1) process. The system

$$\begin{aligned}
\tilde{\boldsymbol{u}}_t &= \tilde{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{u}}_{t-1} + \tilde{\boldsymbol{\eta}}_t \\
\tilde{\boldsymbol{x}}_t &= \tilde{\boldsymbol{c}} + \mathbf{A}\tilde{\boldsymbol{x}}_{t-1} + \mathbf{B}\tilde{\boldsymbol{u}}_t + \tilde{\boldsymbol{e}}_t
\end{aligned}$$

where

$$\mathbf{u}_t = \boldsymbol{\gamma} + \boldsymbol{\alpha}\mathbf{u}_{t-1} + \boldsymbol{\eta}_t \qquad (m \times 1)$$

and

$$\begin{aligned}
\tilde{\boldsymbol{\gamma}} &= \text{vec}(\boldsymbol{\gamma}, \mathbf{0}, \dots, \mathbf{0}) && (m(q+1) \times 1) \\
\tilde{\boldsymbol{\alpha}} &= \begin{pmatrix} \boldsymbol{\alpha} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_m & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{I}_m & \mathbf{0} \end{pmatrix} && (m(q+1) \times m(q+1)) \\
\tilde{\boldsymbol{\eta}}_t &= \text{vec}(\boldsymbol{\eta}_t, \mathbf{0}, \dots, \mathbf{0}) && (m(q+1) \times 1)
\end{aligned}$$

would then have to be considered from the start, upon deriving the filter recursion. When considering views on both the dependent variables and the external regressors, the filter recursion would be difficult to derive. The simplest approach to set views on the external regressors would of course be to include these variables as dependent variables (making them internal). This means that instead of modelling the VARX(p,q) model, the VAR(p) model

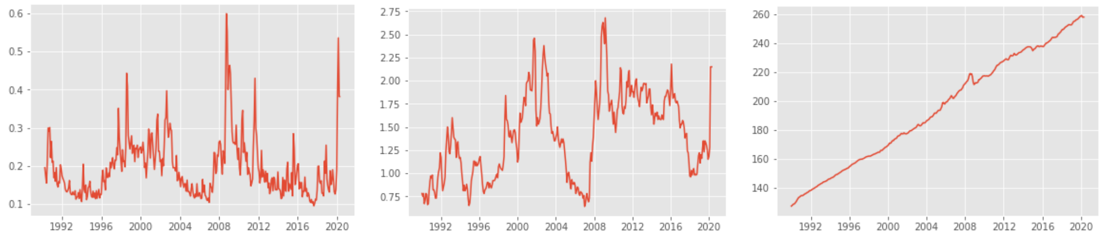
$$\begin{pmatrix} \mathbf{x}_t \\ \mathbf{u}_t \end{pmatrix} = \begin{pmatrix} \mathbf{c} \\ \boldsymbol{\gamma} \end{pmatrix} + \sum_{i=1}^{p^*} \begin{pmatrix} \mathbf{A}_i & \mathbf{B}_i \\ \mathbf{0} & \boldsymbol{\alpha}_i \end{pmatrix} \begin{pmatrix} \mathbf{x}_{t-i} \\ \mathbf{u}_{t-i} \end{pmatrix} + \begin{pmatrix} \mathbf{e}_t \\ \boldsymbol{\eta}_t \end{pmatrix} \qquad ((n+m) \times 1)$$

would be assumed, where p^* is the maximum lag used in any individual process. Whether joint estimation is used or not is a matter of preference, but the parameters would of course have to be re-estimated. Naturally, this model would generally result in less influence from the (previously) external regressors, since the simultaneous values typically would have the most predictive power. Therefore, it is not a perfect solution to the issue.

4. Example with Data

4.1 Data

The data consists of 362 monthly observations from 1990-03-01 to 2020-04-17. The S&P 500 implied volatility is modelled with observations from the VIX index [Yahoo Finance] and the credit spread is modelled with observations from Moody's seasoned Aaa corporate bond yield relative to yield on 10-Year treasury constant maturity [Federal Reserve Bank of St. Luis]. The inflation is modelled with data from the consumer price index for all urban consumers [Federal Reserve Bank of St. Luis].



(a) Volatility (VIX) **(b)** Aaa credit spread **(c)** Inflation (CPI)

Figure 4.1: Historical observations of volatility, credit spread and inflation

The equity data consists of 12 large cap stocks from S&P 500 and 20 small cap stocks from S&P SmallCap 600 [Yahoo Finance]. The log-returns are for simplicity equally weighted (at any given time), to create two indices. These are not very representative as general small/large cap indices due to both the small number of stocks and the weighting method, and could instead be viewed as two dynamically reweighted portfolios. In a global economy ESG, one could instead choose to model a similar number of major global stock indices together with foreign exchange rates. Figure 4.2 shows the historical levels of the two indices and Table 4.1 shows historical means and standard deviations of the log-returns of all constituents, where the column *Code* shows the trading name of the stock (e.g. DIS = Walt Disney Company). The indices have actually performed very similarly in terms of aggregate return over the whole period. The log-returns of the two indices also have similar volatilities, 4.6% and 4.7% respectively, even though the average volatility over the horizon is significantly larger for the small cap index. This could be explained by the large cap equities being more correlated, while the returns of the small cap equities equalize each other to a larger extent. Again, a larger number of stocks would be needed to accurately describe the general market movements in the two segments.

Equity	Code	Mean Return (%)	Standard deviation (%)	Market Capitalization 2020 ($B)
LC1	BAC	0.2	11.2	194.2
LC2	DIS	0.7	7.3	186.1
LC3	HD	1.3	7.4	243.5
LC4	INT	1.1	10.4	252.7
LC5	JNJ	0.8	5.5	394.4
LC6	JPM	0.7	9.5	275.4
LC7	KO	0.6	5.8	193.7
LC8	MRK	0.5	7.1	197.0
LC9	MSF	1.5	8.7	1392
LC10	PFE	0.7	6.4	215.4
LC11	PG	0.7	5.7	285.4
LC12	T	0.2	6.3	207.6
Average	-	**0.77**	**7.6**	**336.5**
SC1	MLAB	1.6	10.4	1.0
SC2	LNN	0.8	10.0	0.9
SC3	PDCE	0.9	17.1	1.2
SC4	AWR	0.8	6.2	2.8
SC5	AXE	0.6	10.5	3.2
SC6	B	0.6	8.8	1.8
SC7	PLXS	1.1	14.7	1.7
SC8	CBU	0.8	7.1	3.1
SC9	CVBF	0.6	8.5	2.6
SC10	CWT	0.6	5.9	2.2
SC11	GBCI	0.9	7.8	3.5
SC12	INDB	0.6	11.0	2.1
SC13	JJSF	0.9	8.7	2.3
SC14	KWR	0.6	10.5	2.5
SC15	MDC	1.1	13.8	1.7
SC16	SFNC	0.7	9.0	1.8
SC17	EE	0.2	9.6	2.8
SC18	UNF	0.7	8.6	3.0
SC19	WDFC	0.7	6.5	2.3
SC20	COKE	0.6	8.0	2.2
Average	-	**0.78**	**9.7**	**2.2**

Table 4.1: Mean and standard deviation of historical log returns of each modelled stock.

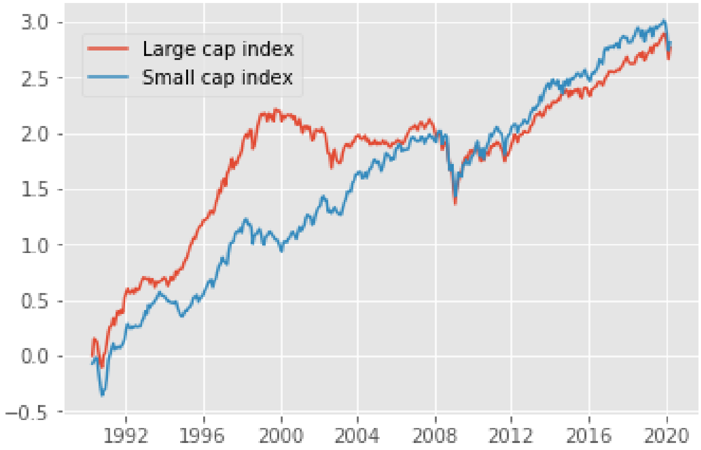
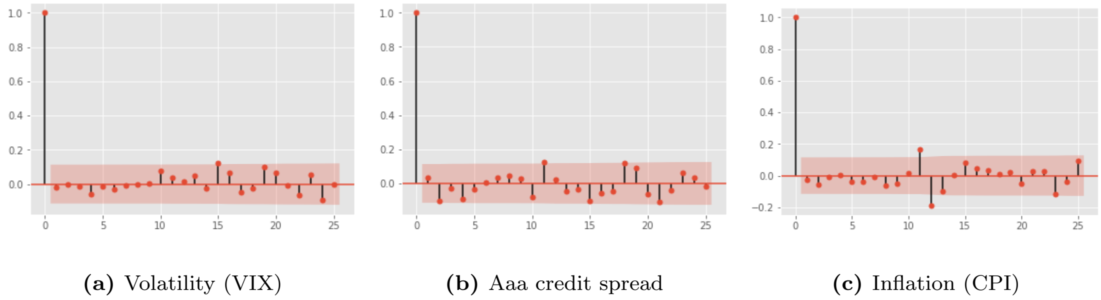


Figure 4.2: Historical observations over the period 1990-03-02 - 2020-04-17 of the log-level of constructed large cap and small cap indices, both starting at 1 in 1990-03-01.

4.2 Estimating the ESG

A training subset of the data is constructed from the first 300 observations, and major outliers are removed (only for SC17, or EE, in this case). Selection according to AIC yields an AR(4) model for the volatility and an AR(2) model (with volatility leverage) for the credit spread. The inflation data, however, shows a prominent moving average structure, and an AR(10) approximation of the ARMA(1,2) model giving the minimum AIC is calculated. The autocorrelations of the resulting residuals are shown in Figure 4.3. There seems to be some dependency on the previous year left in the residuals from the inflation model, but including a lag of 12 months did not lower the resulting AIC. Confidence intervals of the parameters are given in Table 4.2. The parameters in the inflation model are referring to the coefficients in the ARMA(1,2) model, and all parameters for the other two models are described in Section 3.1. The volatility model shows a large dependence on the previous observation, and the negative sign of c indicates mean reversion. The credit spread model also indicates large dependence on the previous observation, as well as a positive dependence on volatility, which is in line with what one would expect.



(a) Volatility (VIX)

(b) Aaa credit spread

(c) Inflation (CPI)

Figure 4.3: Autocorrelation of residuals

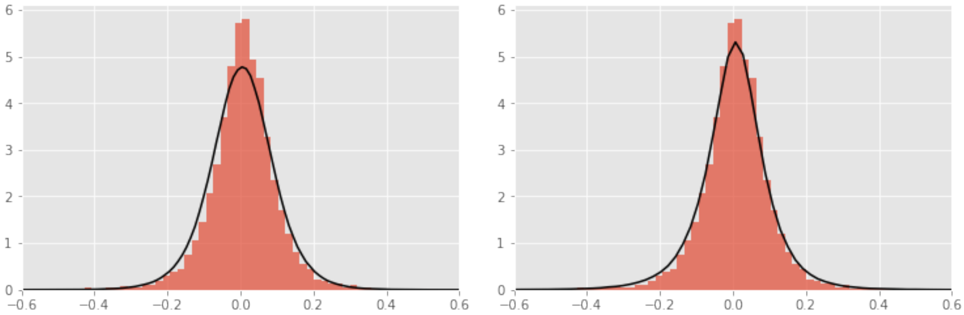
Volatility model		
Parameter	$[\alpha/2$	$1-\alpha/2]$
θ	-1.8400	-1.5276
c	-0.1874	-0.0679
a_1	0.6691	0.8606
a_2	-0.0934	0.1405
a_3	-0.1024	0.1298
a_4	-0.0179	0.1731
σ	0.1514	0.1799

Credit spread model		
Parameter	$[\alpha/2$	$1-\alpha/2]$
μ	0.0836	0.4065
a_1	1.0215	1.1834
a_2	-0.2456	0.0833
b	0.0535	0.3810
σ	0.3917	0.4744

Inflation model		
Parameter	$[\alpha/2$	$1-\alpha/2]$
intercept	0.3860	0.7965
a_1	-0.9189	-0.3859
b_1	0.9759	1.4664
b_2	0.3714	0.6216
σ	0.2003	0.2353

Table 4.2: Confidence intervals of fitted parameters at significance $\alpha = 0.05$

The VARX(p,q) model with parameter restrictions is first selected in terms of lowest AIC according to the method proposed in Section 3.2.6, where the volatility and inflation are both differenced when used as external regressors. Furthermore, no influence from small cap stocks on other stocks is allowed. Including volatility clustering from the VIX index is both reducing AIC and giving better residual structure (more like white noise) for most of the constituents. A comparison of the distribution of the log-returns from 2 000 simulated paths with and without volatility clustering is displayed in Figure 4.4. The AIC search yields $p = 1$ and $q = 0$, with parameter restrictions such that there are 32 free parameters in $\mathbf{c}$, 272 free parameters in $\mathbf{A}_1$ (of which 16 are on the diagonal) and 85 free parameters in $\mathbf{B}_0$, meaning that 859 of 1248 parameters are set to zero. Since this could still be considered as an overparameterized model, only significant parameters are included. After bootstrapping parameter confidence intervals with 5 000 bootstrap time series according to Section 3.2.5 and imposing restrictions on any insignificant parameters, there are 111 free parameters left. Graphical Lasso is applied to the covariance matrix with $\rho = 0.05$. A summary of the parameter restrictions is given in Table 4.3. As one might expect, the simultaneous values of the (differenced) VIX index seem to have most explanatory power. Some suspicions of multicollinearity may be raised, since VIX is derived from an index containing the large cap equities, but analysis shows that the dependence is not large enough to be of any concern. Furthermore, the average $\mathbf{B}_0$ coefficient for VIX is -0.68 for the large cap equities and -0.66 for the small cap equities, suggesting that the explanatory power of the volatility processes in not only due to the fact that it is partly derived from the large cap equities. The impact of differenced S&P 500 volatility on small cap equities is also large in the training data, although less prominent since the magnitude of the returns are larger in the small cap segment. Figure 4.5 shows the structure of the standardized residuals of both indices. The assumption is that the standardized errors have a multivariate normal distribution, which implies that any weighted sum should be normally distributed. It seems plausible that the weighted sums are normally distributed, although there are a few major outliers. This is however no guarantee for the multivariate normality to hold, since that is a stronger assumption. The estimated covariance matrices of the standardized residuals before and after Graphical Lasso are displayed in Figure 4.6.



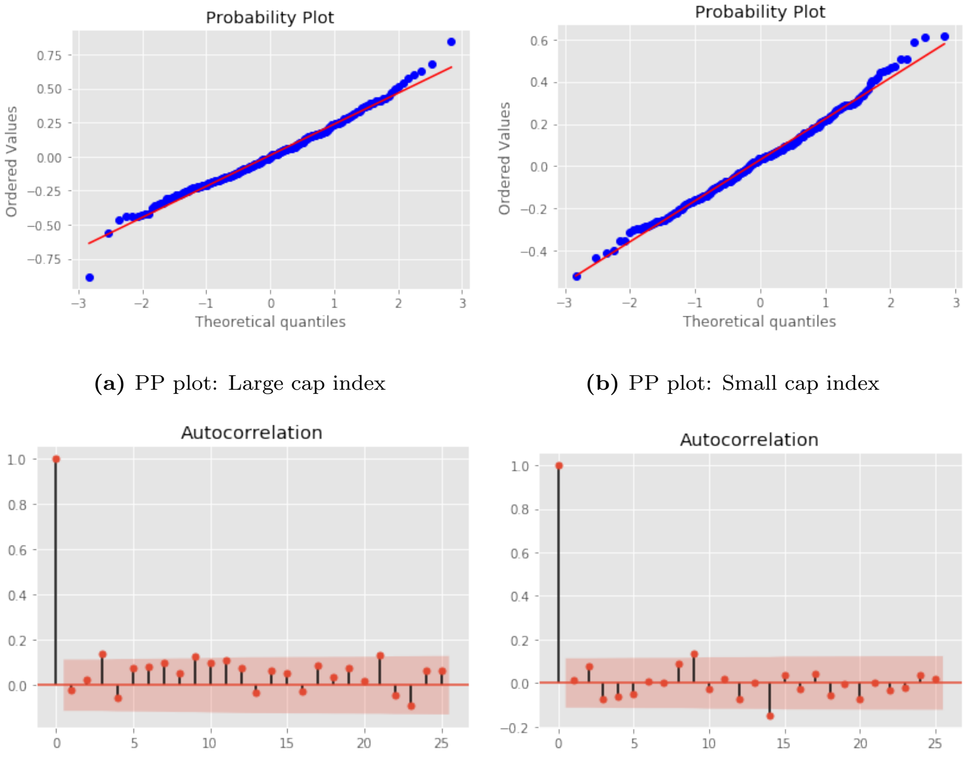
(a) No volatility clustering

(b) Volatility clustering

Figure 4.4: Distribution of all log-returns. Histogram of training data and distribution fitted to 2 000 simulated paths of same length and with same presample.

Object	Possible parameters	Free parameters
$\mathbf{c}$	**32**	**17**
$\mathbf{A}_1$	**1024**	**58**
diag($\mathbf{A}_1$)	32	9
$\mathbf{B}_0$	**96**	**36**
(VIX)	32	32
(Aaa)	32	3
(CPI)	32	1
$\mathbf{\Gamma}$	**1248**	**111**

Table 4.3: Summary of parameter restrictions.



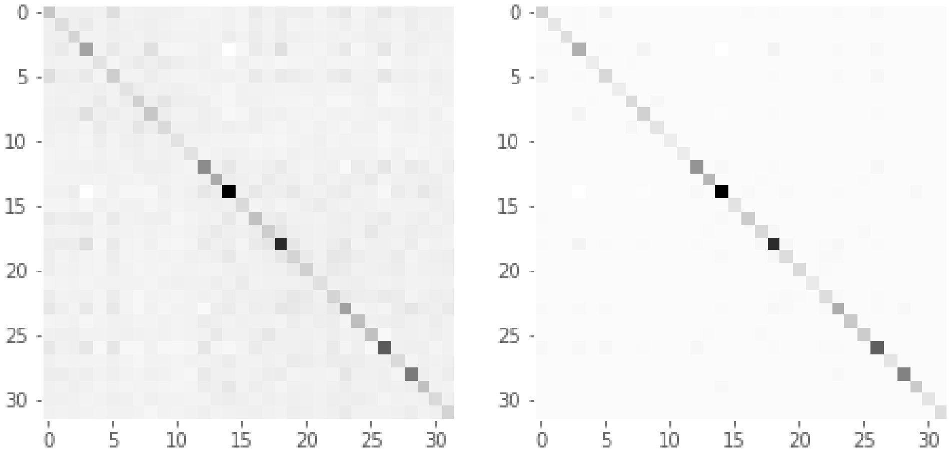
(a) PP plot: Large cap index

(b) PP plot: Small cap index

(c) ACF plot: Large cap index

(d) ACF plot: Small cap index

Figure 4.5: Probability-probability plots and autocorrelation plots for the standardized log-returns of both indices.

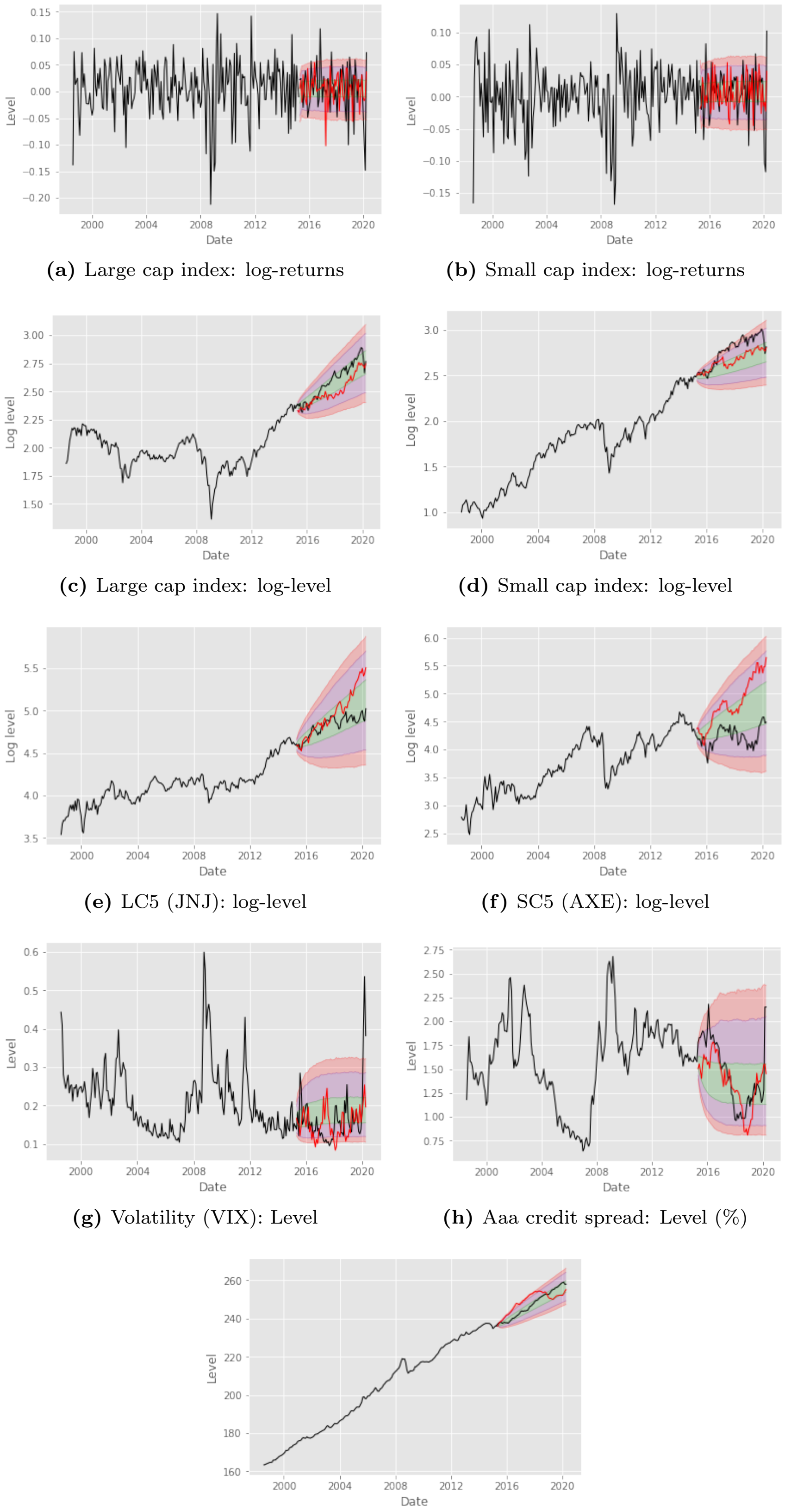


(a) Empirical estimate

(b) Graphical Lasso estimate

Figure 4.6: Covariance of standardized residuals

Figure 4.7 shows prediction intervals at a few different significance levels of the two indices, two of the individual stocks and the external regressors over the tuning period. One simulated path, the true historical values over the tuning horizon and historical observations over the training period are also displayed. To test the forecast accuracy of the model, a rolling window could have been used. This could also have been used to examine whether the assumption that the parameters are time-invariant holds. The main focus is however to illustrate how views can be incorporated, and nothing about the behavior of the simulated paths or prediction bounds (over the tuning horizon) suggests that the model is severely misspecified.



(a) Large cap index: log-returns

(b) Small cap index: log-returns

(c) Large cap index: log-level

(d) Small cap index: log-level

(e) LC5 (JNJ): log-level

(f) SC5 (AXE): log-level

(g) Volatility (VIX): Level

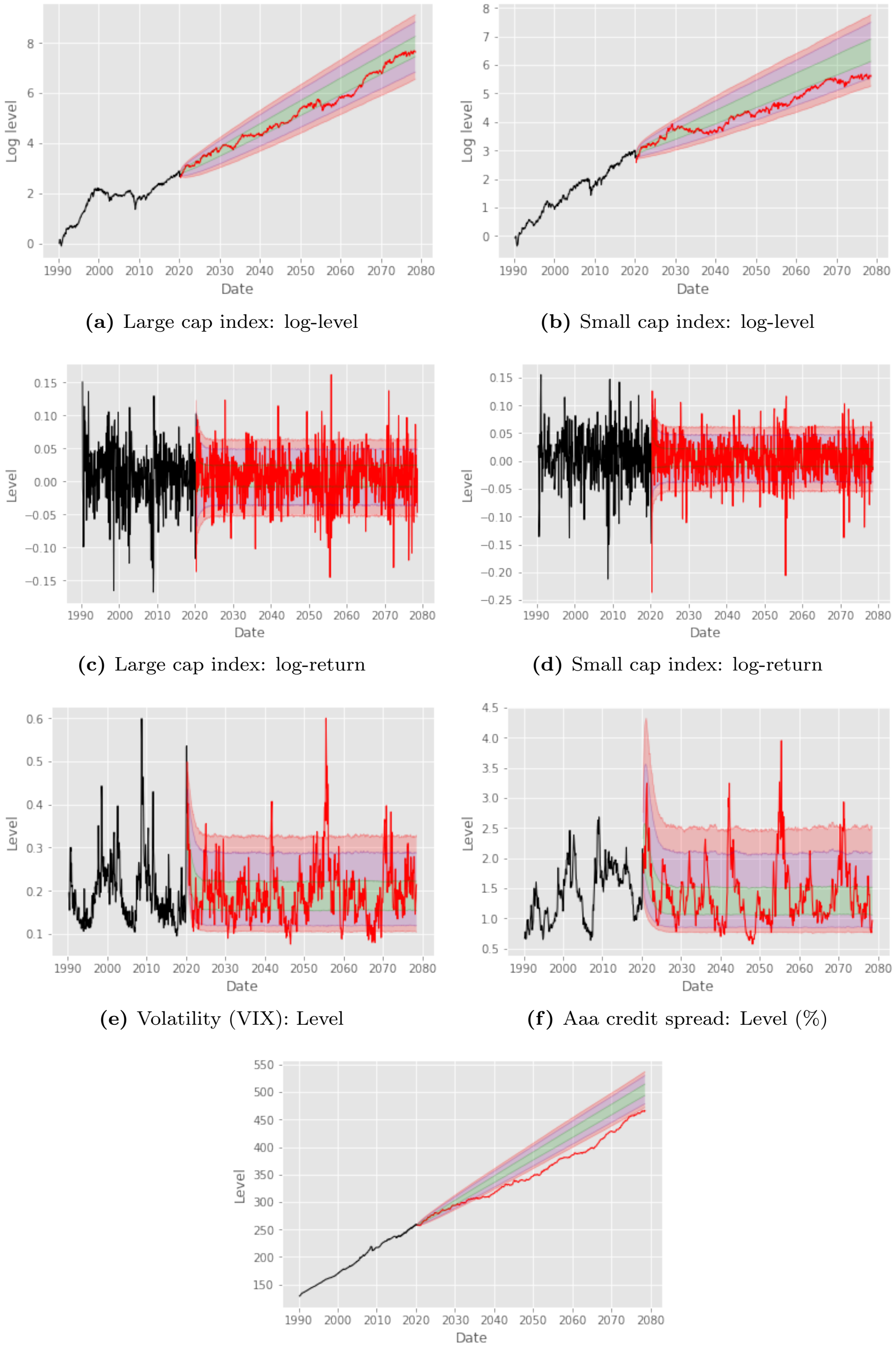
(h) Aaa credit spread: Level (%)

(i) Inflation (CPI): Index

Figure 4.7: Historical observations over whole period are displayed in black. Simulated paths over forecast horizon are displayed in red. Shaded areas are prediction intervals at significance $\alpha = 0.3$, $\alpha = 0.1$ and $\alpha = 0.05$ respectively.

4.3 Simulating Future

Using the ESG fitted to the training period, the aim is now to simulate the future development of the modelled variables. The horizon is chosen to 700 months, i.e. almost 60 years. Before any views are imposed, the baseline model estimate of the future is analyzed. Figure 4.8 shows prediction intervals of the levels of the two indices as well as the external regressors over the horizon. One simulated path over the horizon and all historical values are also displayed. The prediction of the small cap index seems too low judging from the historical trend. Furthermore, the prediction intervals of the returns of both indices could be considered a bit narrow, especially for the small cap index. Reviewing the model should normally be the starting point. Should intercepts be included for all variables due to the long-term nature of the forecast? Is there more or less comovement in the data than what is induced by the external regressors? Is the Graphical Lasso penalization removing too much of the noise correlation, resulting in more equalization of returns and smaller prediction intervals of the indices? However, for illustrative purposes, the predictions will instead be altered by imposing views.



(a) Large cap index: log-level

(b) Small cap index: log-level

(c) Large cap index: log-return

(d) Small cap index: log-return

(e) Volatility (VIX): Level

(f) Aaa credit spread: Level (%)

(g) Inflation (CPI): Index

Figure 4.8: Historical observations are displayed in black. Simulated paths over forecast horizon are displayed in red. Shaded areas are prediction intervals at significance $\alpha = 0.3$, $\alpha = 0.1$ and $\alpha = 0.05$ respectively.

There are thus two beliefs that could be considered, namely higher *weighted sum of* the returns of the small cap equities and slightly higher standard deviation of returns in general. To alter the standard deviation of returns, views about the volatility process would have to be imposed. The question is however whether or not the actual views are about the volatility process, or about the magnitude of the covariance matrix. The volatility process is in this case also affecting the credit spread prediction due to the model configuration, and in an ESG, the volatility process is not just a driver of other variables, but also an important variable to study in itself. This is one of the limitations of the proposed method – it does not provide a framework for imposing views about neither the functional relationship between variables, nor the noise covariances. It does not make sense to impose views about the distribution of one variable only to change the distribution of another! Furthermore, the filter recursion would not be analytically tractable when imposing views on the volatility process. As seen in Section 3.4, the method is only easily applicable when considering views on the dependent variables. Therefore, only views about the expectation of the small cap index will be imposed.

First, the belief about the expectation of the (equally weighted) log-returns of the small cap equities over the horizon needs to be specified. In this case, let us start by formulating a belief about the log-level of the index. For simplicity, the baseline model estimate is going to be merged with views that the expectation of the log-index is going to follow a linear trend fitted to the historical observations. Thus, the views about the log-level of the index, together with the implied log-return, are given in Figure 4.9. The matrix specifying the linear combination of variables on which the views are set is in this case just the small cap weight matrix, i.e.

$$\begin{aligned}\mathbf{H}_t &= \left(0, \ldots, 0, \frac{1}{20}, \ldots, \frac{1}{20}\right)\\ &\in \mathbb{R}^{1\times 32}\end{aligned}$$

As seen in Section 3.3.2, the views should be transformed to account for the fact that the log-returns are modelled. Denote ψ_t^* the views about the log-level of the index. The views, ψ_t, which should be imposed in the filtering procedure are then given by

$$\begin{aligned}\psi_t &= \psi_t^* - \psi_{t-1}^* - \frac{1}{2}\text{diag}(\mathbf{H}_t\widehat{\mathbf{\Sigma}}_t\mathbf{H}_t^\top)\\ \widehat{\mathbf{\Sigma}}_t &= v_t\widehat{\mathbf{\Sigma}}\end{aligned}$$

where v_t is the (simulated) value of the volatility process and $\widehat{\mathbf{\Sigma}}$ is the estimated covariance matrix of the standardized errors. In line with the proposed restrictions in 3.3.2, the interval I_ψ is set to $[t_0+10, \ldots, t_0+700]$ to ensure a smooth transition. For the covariance matrix, $\tau(h) = 1$ for $h = 1, \ldots, 700$ is chosen.

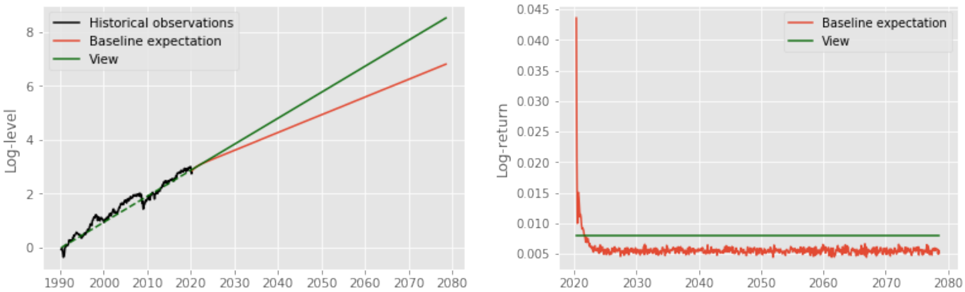


Figure 4.9: Left: Historical observations, together with the baseline model expectation and the views (log-level). Right: Baseline model expectation and views (log-return).

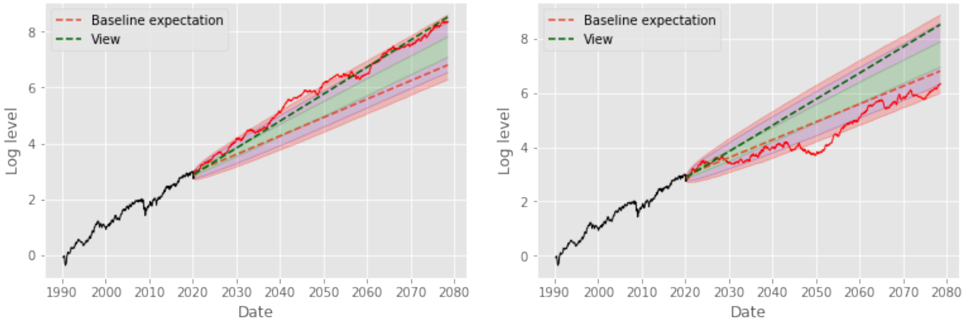
Since the log-returns are heavily dependent on the volatility process, it would be false to say that the process is close to being described by a random walk (conditional on observed values of the external regressors). Thus, specifying views unconditionally will reduce the uncertainty of the forecast (see Section 5.1). Therefore, views conditional on $\mathbf{u}_t$ are also considered. Denote ψ_t^k the views and $\mathbf{u}_t^k$ the values of the external regressors of path k. With the unconditional views $\psi_t = 0.008038$ for all t, the conditional views are set to

$$\psi_t^k = 0.008038 + \mathbf{H}_t\mathbf{B}_0\mathbf{u}_t^k - \delta$$

where δ is the average impact on the log-return from the external regressors in the baseline prediction, i.e.

$$\delta = \frac{1}{N}\sum_{k=1}^{N}\mathbf{H}_t\mathbf{B}_0\mathbf{u}_{0,t}^k$$

where $\mathbf{u}_{0,t}^k$ are the values of the external regressors in one of the simulated paths from the baseline model and N is the number of paths. This would indicate beliefs that the conditional expectation will have the same dependence on the external regressors as in the baseline model, but where the unconditional expectation is higher. The adjusted predictions with unconditional and conditional views respectively are shown in Figure 4.10. It is clear that the conditional views result in wider prediction intervals than the unconditional views.



(a) Unconditional views: log-level

(b) Conditional views: log-level

Figure 4.10: Predictions of the small cap index with unconditional and conditional views respectively. Historical observations are displayed in black. Simulated paths over forecast horizon are displayed in red. Shaded areas are prediction intervals at significance $\alpha = 0.3$, $\alpha = 0.1$ and $\alpha = 0.05$ respectively. Note that the paths are sampled from the filter density.

5. Discussion

There are four main issues with the suggested approach. The possibility of having contradicting views can be removed by imposing relevant restrictions on the views, while model validation and bias correction may be applied as a sufficient amount of data on views becomes available. The other two issues are more difficult to handle. As seen in Section 2.3, the views add information in the estimation procedure, resulting in a posterior covariance lower than what is implied by the baseline model if $\tau(t) < \infty$ for any t. If $\tau(t)$ is relatively small, the width of the prediction intervals could reduce significantly depending on the amount of memory in the original process. In certain settings this is desirable. If the assumption that the error terms ξ_t and $\mathbf{e}_t$ are independent, and both the view-distribution and the baseline model are correctly specified, the posterior density would be an accurate estimation of the future development of the variables. Consider for example a robot receiving noisy inputs about its position from both a sensor and a satellite with (presumably) independent noise. However, making these assumptions is in this case far-fetched, since the distribution of the views is just a theoretical construction. Furthermore, the fact that the method can be used to estimate the posterior distribution, but not for generating paths, is an issue for hierarchical models.

5.1 Loss of Conditionality

The first main issue stems from the fact that the views are not specified conditional on previous values, while the baseline model typically is. Furthermore, the baseline model will usually influence the views. Consider an example where an economist looks at the baseline model prediction over the forecast horizon and considers it to be reasonable in terms of uncertainty, but with a slightly low expectation. The belief is therefore that the *unconditional* expectation is higher than what was originally estimated, and the economist imposes this view with similar uncertainty as implied by the baseline model (i.e. $\tau(h) = 1$). Should the resulting estimate of the unconditional expectation then be more certain than then original prediction? This is of course only an issue if there is a significant dependence on previous lags and/or external regressors, meaning that there is some non-negligible uncertainty of the conditional expectation in the baseline model. As an extreme case where this is no issue, consider a random walk. The conditional expectation would then always be zero, meaning that the uncertainty is only driven by the error term $\mathbf{e}_t$, and that imposing views of the type mentioned above will yield similar widths of the prediction intervals as the original forecast.

Since the views are not specified conditional on previous values, they are independent of the path until that point. Consider setting the uncertainty of the views according to (3.7). For an autoregressive process with a large dependence on previous values, i.e. where the magnitudes of the AR-coefficients are large, this would mean that the h-step forecast error covariance generally would be lower than what is implied by the baseline model. It would also give a resulting expectation at time t closer to the views than the baseline estimate, even though $\tau(h) = 1$ for all h. First consider the case where there is no dependence on previous observations. In this case, one would always have $\mathbf{H}_t\mathbf{K}_t(1 + \tau(h)) = \mathbf{I}_d$, and given $\tau(h) = 1$ for all $t = t_0 + h$, the prediction would lie in between the baseline estimate and the views.

This will approximately be the case when imposing views on processes which are mostly described by noise, e.g. when modelling log-returns without external regressors. However, this is not the case when there is a larger dependence on previous lags, such as when modelling log-volatility without differencing. Although setting the parameter $\tau(t) = 1$ implies that the uncertainty of the views is the same as that of the unconditional prediction of the baseline model, the fact that the views are unconditional on previous values will result in the views dominating the prediction. This poses an issue since the expert's beliefs in practice will be about the *unconditional* expectation at a given time – the path until that point is unknown. As an illustration, consider the (univariate) AR(2)-process given by

$$\begin{aligned} x_t &= -0.18 + 0.8x_{t-1} + 0.1x_{t-2} + e_t, \quad e_t \sim \mathcal{N}(0, 0.15^2) \\ x_0 &= 0.3 \end{aligned}$$

This process is stationary, although one of the roots of the characteristic polynomial is close to the unit circle, and the expectation is relatively heavily dependent on previous observations. Setting $\tau(t) = 1$ and $\psi_t = 0.1$ for all t, the resulting prediction after incorporating the views, together with the baseline prediction, is given in Figure 5.1. Notice that, even though the variance of the distribution around the conditional expectation is the same, the forecast error covariance is reduced when applying the views. The resulting expectation is also closer to the views than the baseline model estimate. The evolution of the Kalman gain is also displayed. It starts at $1/(1 + \tau) = 0.5$ and decreases until reaching its steady state. This is in line with what is expected. When setting the covariance of the views according to (3.7), larger magnitudes of the AR parameters give a smaller Kalman gain.

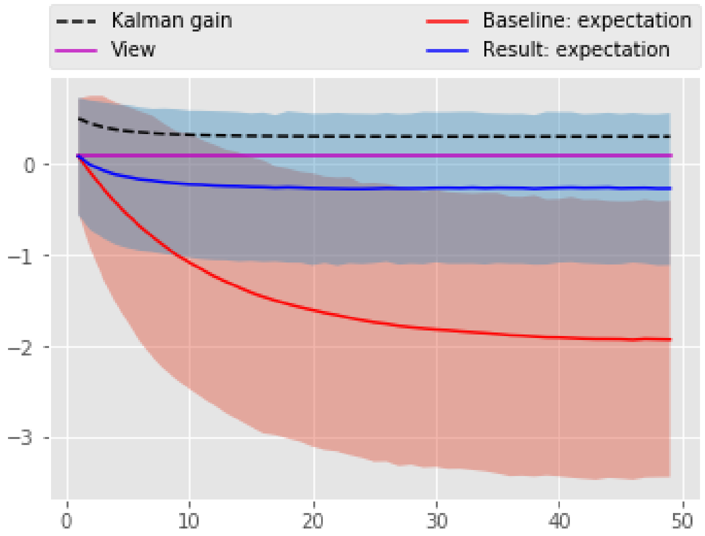


Figure 5.1: Illustration of the loss of lag dependence when incorporating views. Shaded areas are within 95% empirical quantiles.

There is no easy solution to this issue in the general case. One option is to calibrate the function $\tau(h)$ – in this case the function $\tau(h) = 1 + 0.18h$ would result in a predicted mean approximately in between the view and the baseline prediction over the whole forecast horizon. This would however mean that the value of the covariance matrix of the views is less interpretable. If possible, modelling the differenced time series would reduce the loss of lag dependence. However, when modelling e.g. log-volatility, the mean reversion property would disappear when modelling the differenced series. A possible option would be to penalize deviance from the mean, and model e.g. the process

$$
\begin{aligned}
&\boldsymbol{\mu}_t = \mathbf{A}\mathbf{x}_{t-1} - c(\mathbf{y}_t - \mathbf{m})\\
&\mathbf{x}_t = \boldsymbol{\mu}_t + \mathbf{e}_t\\
&\mathbf{y}_t = \mathbf{y}_0 + \sum_{i=0}^{t} \mathbf{x}_i
\end{aligned}
$$

where $\{\mathbf{y}_t\}$ is the logarithm of the modelled time series, $\{\mathbf{x}_t\}$ is the log-return and $\mathbf{m}$ is a global mean to which the logged time series is expected to revert. However, this would complicate both the estimation and the filtering procedure. Furthermore, it may not be possible to specify the model such that it behaves similarly as a model which assumes an autoregressive process for $\{\mathbf{y}_t\}$. Since the process is assumed to be mean reverting, the views of the differenced process would also have to approach, or be centered around, zero – setting views about a long-term mean would be impossible.

A final option which may be worth to investigate further would be to set conditional views. When modelling processes which sometimes drift far from the global mean (e.g. the volatility and the credit spread processes modelled Chapter 4), providing views conditional on previous values would be more reasonable. Naturally, our views about next month's volatility depend on the current value, while it would not be as far-fetched to say that our views about the expected values of next month's equity returns are independent of the current values. Even in a case where log-returns are modelled, setting conditional views may sometimes be a better option. If some external regressor has a large impact on the predictions in the baseline model, formulating views based on the observed value of that variable could be appropriate.

Specifying views conditional on the external regressors $\mathbf{u}_t$ does in principle not pose an issue – the value of $\mathbf{u}_t$ is considered as known at each filter recursion. Consider a linear Gaussian system and a similar assumption as in Section 4.3, where the views are specified as the expectation with the influence from the external regressors excluded. This is exactly the same as considering the system

$$
\begin{aligned}
&\boldsymbol{\psi}_t = \mathbf{H}_t\boldsymbol{\mu}_t - \mathbf{H}_t\mathbf{B}\mathbf{u}_t + \boldsymbol{\xi}_t \qquad && \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t)\\
&\boldsymbol{\mu}_t = \mathbf{c} + \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t\\
&\mathbf{x}_t = \boldsymbol{\mu}_t + \mathbf{e}_t && \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)
\end{aligned}
$$

where $\boldsymbol{\psi}_t$ are the views about the expectation discarding influence from the external regressors. Since the values of $\mathbf{u}_t$ are considered as deterministic at each filter recursion, this could be rewritten as

$$
\begin{aligned}
&\boldsymbol{\psi}_t = \boldsymbol{\alpha}_t + \mathbf{H}_t\boldsymbol{\mu}_t + \boldsymbol{\xi}_t \qquad && \boldsymbol{\xi}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_t)\\
&\boldsymbol{\mu}_t = \boldsymbol{\beta}_t + \mathbf{A}\mathbf{x}_{t-1}\\
&\mathbf{x}_t = \boldsymbol{\mu}_t + \mathbf{e}_t && \mathbf{e}_t \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_t)
\end{aligned}
$$

where $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ are both deterministic, and the filter recursion is thus easily found. If the regressors have their own stochastic models, which is generally the case unless they would be known beforehand (e.g. weekdays), the realized views $\boldsymbol{\psi}_t$ would of course also be stochastic. A drawback of this is that the actual moments of the prediction would be more difficult to derive. However, for general processes of $\mathbf{u}_t$, specifying the moments of the forecast may be difficult anyway. Conditioning the views on past observations is less trivial – the paths of the dependent variables are not known. An option that may be considered is sampling the views according to some belief about the process. Consider the setting where the expected value of the baseline prediction is believed to be too low. The views of conditional expectations could then be constructed iteratively by generating paths (of $\boldsymbol{\mu}_t$) from the baseline model and adding the believed shift in expectation to each generated path $\boldsymbol{\psi}_t$. Again, this would mean that the moments of the forecasted distribution are not easily tractable, and that simulation is needed. Furthermore, the interpretation of $\boldsymbol{\Omega}_t$ discussed in Section 3.3.1 would not hold. Assume that the belief is that the process will behave exactly the same as the baseline model, meaning that for each simulated path k, the views are set as

$$\boldsymbol{\psi}_t^k = \boldsymbol{\mu}_t^k \qquad (\mathbf{H}_t = \mathbf{I}_n)$$

where

$$\begin{aligned}\boldsymbol{\mu}_t^k &= \mathbf{c} + \mathbf{A}\mathbf{x}_{t-1}^k + \mathbf{B}\mathbf{u}_t^k \\ \mathbf{x}_t^k &= \boldsymbol{\mu}_t^k + \mathbf{e}_t, \quad \mathbf{e}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_t)\end{aligned}$$

The filter recursion would then be done for each sampled path of the views, upon studying the empirical distribution (sampled from the posterior at each iteration). If the covariance $\boldsymbol{\Omega}_t$ is always zero, the forecasted distribution would be exactly the same as for the baseline model. This would obviously be the case when the uncertainty of the views approaches infinity as well. The question is what happens in between, and maybe more importantly, what would be the interpretation?

Thus, further research would have to be made to give a clear interpretation for how the uncertainty of the views would affect the resulting moments in a sampling approach. Furthermore, requiring the user to specify some functional dependence on previous values may not always be a practical solution. However, it would allow for more flexibility, and for processes which sometimes drift far from the unconditional mean before returning (while still stationary), it would give a clearer interpretation.

5.2 Paths are Unattainable

The Bayesian filtering approach is a way of estimating the distribution over the forecast interval. This means that at each point in time, the expectation and uncertainty are known (or the approximate density for general processes). However, sampling from the filter density would *not* yield a plausible future path. Although we are typically only interested in the moments, and in turn an expectation and a prediction region, studying the generated paths could be considered as one step in validating the model. If a simulated path looks unrealistic given the historical observations, it would indicate that the model may be misspecified. However, by validating the baseline model, and by fully understanding the filter recursion, interpreting the results should not be a concern. A bigger issue is the fact that economic scenario generators often are specified hierarchically. This means that paths simulated from some variables may be used as predictors for other variables. As seen in Section 3.4.2, this makes it difficult to impose views on variables which are not lowest in hierarchy. Depending on the complexity, it may be an option to specify the filter recursion for the whole ESG (although maybe not analytically). In the general case however, this may be difficult.

There is a simple, although imperfect, solution. This is to only allow for lagged dependence between the variables. For the VARX(p,q) model, this would mean that any (previously) external regressor where there may be views is added to the dependent variables, i.e. modelled as internal. This would of course mean that dependence on the simultaneous value cannot be modelled, which would typically result in less dependence. On the other hand, when making long-term forecasts, there is another crude but sometimes practical solution: estimate a VARX(p,q) model and construct a VAR(p) model (or VAR(q) if $q > p$) where the parameters of the previously external variables are shifted. Letting the estimated parameters $\mathbf{B}_0$ be the corresponding parameters in $\mathbf{A}_1$ and $\mathbf{B}_1$ be the corresponding parameters in $\mathbf{A}_2$ etc. and adjusting the covariance matrix would not result in a noticeable difference in the resulting forecast far into the future. As an example, one could consider respecifying the system

$$\begin{aligned}x_t &= ax_{t-1} + e_t^x \\ y_t &= by_{t-1} + \beta x_t + e_t^y \\ z_t &= cz_{t-1} + \gamma y_t + e_t^z\end{aligned} \tag{5.1}$$

as

$$\begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} = \begin{pmatrix} a & 0 & 0 \\ \beta & b & 0 \\ 0 & \gamma & c \end{pmatrix} \begin{pmatrix} x_{t-1} \\ y_{t-1} \\ z_{t-1} \end{pmatrix} + \mathbf{e}_t \tag{5.2}$$

where

$$\mathbf{e}_t \sim \mathcal{N}\left(0, \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_y^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix}\right)$$

without re-estimating the parameters, such that the three-level hierarchy is approximated by a single level. The predictions for y_t and z_t would of course have to be shifted one and two time-steps respectively assuming the same presample. Although the distributions would converge asymptotically, it is impossible to specify the presample such that the (5.1) equals (5.2) in the short-term for all three variables. Assuming monthly intervals, the resulting distributions may differ significantly multiple years ahead. Thus, the resulting distributions would have to be studied thoroughly before making predictions with the re-specified system, and this cannot be regarded as a valid solution in general unless short-term predictions are considered as irrelevant.

5.3 Contradicting Views

The third issue is the possibility of contradicting views. Since it is required that $\mathbf{H}_t$ has full rank, no views which are theoretically impossible are allowed. However, if the difference between the views $\boldsymbol{\psi}_t$ and the expectation of the predictive density is large in comparison to the uncertainty of the predictive density and the uncertainty of the views, the expectation of the filter estimate would be far into the tails of both distributions (See Figure 5.2). This would suggest that either the distribution according to the baseline model or according to the views (or both) is misspecified, and that the uncertainty of the filter estimate should be large due to the model uncertainty. However, the posterior estimate of the covariance is independent of the level of $\boldsymbol{\psi}_t$. To some extent, the proposed method of setting $\boldsymbol{\Omega}_t$ solves this. If the uncertainty of the views is large, the distribution around $\boldsymbol{\psi}_t$ is wide. On the other hand, if the uncertainty is small, the predictive density will converge to the views fast. Furthermore, by applying the informal restrictions on the views proposed in Section 3.3.2, the initial densities would always be consistent for small enough ϵ_2. Those restrictions also make sure that the views and the baseline model estimate of the *unconditional* expectation are reasonably close.

Another possibility in a complete economic scenario generator is that contradicting views on different processes are imposed. Consider as an extreme case modelling a EUR-based ETF on the S&P 500 index (e.g. VUSA.AS) and a USD-based ETF on the same index (e.g. SPY) in the same economic scenario generator. Imposing views that the expected return of the EUR-based ETF will be higher than the expected return of the USD-based ETF, together with views that the currency exchange rate EUR/USD will decrease over the period, would of course not be reasonable. However, with the model specification in (2.5), there are no views which are infeasible.

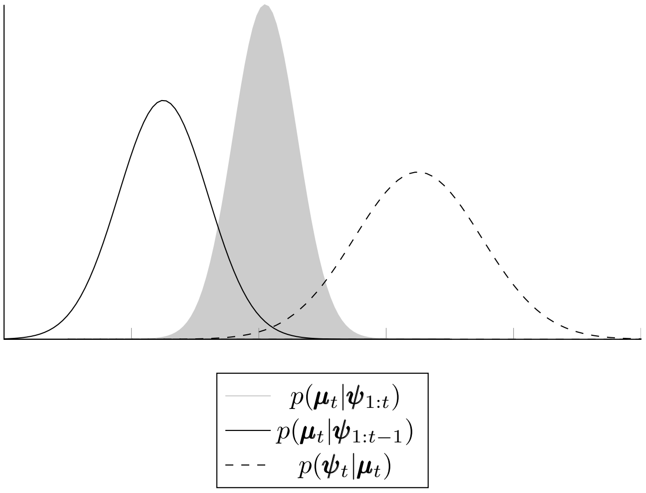


Figure 5.2: Large difference between the imposed view and the expectation of predictive density in comparison to the assumed uncertainties.

5.4 Model Validation and Bias

The final issue is partly due to the lack of data of historical views and partly due to the bias of the practitioner. Everyone has biases, and beliefs about the future development of assets and macroeconomic factors may depend on the individual's experiences and, in some cases, even mood. As staded by Parker [2019], "all kinds of day-to-day activities are primarily driven by behavioral patterns. These same behavioral patterns can also influence investing actions." There are two main categories of investment biases. Cognitive biases involve making decisions based on established concepts that may be inaccurate, while emotional biases typically occur impulsively based on personal feelings at the time a decision is made. These biases can to some extent be mitigated by understanding and identifying them.

When there is no historical data on views, identifying and accounting for these biases is difficult. However, in a situation where there are some historical observations of views, biases can be incorporated in the model. Furthermore, one could also choose to include biases when there are no historical views available by dynamically updating the bias coefficients over time as data on views accumulate. Including bias correction, the model (2.5) could look like

$$\begin{aligned}
\boldsymbol{\psi}_t &= \mathbf{H}\boldsymbol{\mu}_t + \boldsymbol{\xi}_t + \mathbf{b}_t^{\psi} \\
\boldsymbol{\mu}_t &= f(\mathbf{x}_{t-1}) + \mathbf{b}_t^{\mu} \\
\mathbf{x}_t &= \boldsymbol{\mu}_t + \mathbf{e}_t
\end{aligned} \tag{5.3}$$

Note that the baseline model is biased as well. One could say that it has a cognitive bias in assuming that the data generating process will not change. It is assumed that $\mathbf{H}_t = \mathbf{H}$, i.e. that the views are imposed on the same linear combination of variables over time. Since one of the issues with the model without bias correction is that similar uncertainties of the views and the baseline model prediction results in less uncertainty than what is historically observed, it is proposed that both the bias of the baseline model estimate and the bias of the views are random. Letting $\mathbf{b}_t^{\mu} \sim \mathcal{N}(\mathbf{m}^{\mu}, \boldsymbol{\Theta}^{\mu})$ and $\mathbf{b}_t^{\psi} \sim \mathcal{N}(\mathbf{m}^{\psi}, \boldsymbol{\Theta}^{\psi})$, the filtering recursion is easily extended when f is linear and $\mathbf{e}_t$ and $\boldsymbol{\xi}_t$ are Gaussian. How to estimate the bias, and how to dynamically update the model over time, is however non-trivial. One could use the following chain of events

[1] Fit the model f to the window $t \in [t_0, t_1]$.

[2] Impose views and forecast on the window $[t_1, t_2]$ with no bias.

[3] Estimate bias

[4] Fit the model f to the window $t \in [t_0, t_2]$.

[5] Impose views and forecast on the window $[t_2, t_3]$ given bias

[6] Re-estimate bias

[7] Fit the model f to the window $t \in [t_0, t_3]$.

[8] ...

The model may of course be used to make predictions decades ahead, but a forecast would have to be made more often to update the bias. Naturally, $\tau(h)$ used to set $\mathbf{\Omega}_t$ could be estimated as well, although this would complicate things further. If $\tau(h)$ is not estimated (or constant), it should always be set by the same person according to the same rationale (with the same bias) over time. Consider the linear Gaussian case. The model in (5.3) could then be written as

$$\begin{aligned} \boldsymbol{\psi}_t &= \mathbf{H}\boldsymbol{\mu}_t + \boldsymbol{\xi}_t^*, & \boldsymbol{\xi}_t^* &\sim \mathcal{N}(\mathbf{m}_t^{\psi}, \mathbf{\Omega}_t + \mathbf{\Theta}_t^{\psi}) \\ \boldsymbol{\mu}_t &= \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{u}_t + \mathbf{b}_t^{\mu} & \mathbf{b}_t^{\mu} &\sim \mathcal{N}(\mathbf{m}_t^{\mu}, \mathbf{\Theta}_t^{\mu}) \\ \mathbf{x}_t &= \boldsymbol{\mu}_t + \mathbf{e}_t & \mathbf{e}_t &\sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_t) \end{aligned} \tag{5.4}$$

The new information on which the biases should be estimated are the observations of $\mathbf{x}_t$. Unlike typical filtering procedures, where the process of some hidden variable is estimated, the objective is to maximize the likelihood of the process in (5.4) having generated the new observations. The problem is that he Bayesian filtering procedure generates (long-term) forecasts of the moments, and not paths. One approach could be to maximize the likelihood of the posterior, but this would require data on views at each time step and it would not be able to catch long-term biases, since the biases would be estimated to correct one-step predictions. Another approach could be to manually tune the biases. This would mean reformulating the views of both the expectation and the covariance given the observed data, and correct both the views and the baseline model estimate by tuning the bias parameters. This approach may however result in the introduction of new biases, and the corrections should be based on careful statistical analysis of the observed data. A third option may be to formulate the process such that e.g. Gibbs sampling can be applied, where some initial priors are assumed, and the biases are sampled conditional on the previously observed biases and observations. To be able to include biases, some practical considerations would have to be made. Letting $\mathbf{H}_t$ be time-varying may not be an option and letting $\tau(h)$ be constant could also increase the interpretability in the case where biases are included. In a full ESG, possibly with more than one (vector) autoregressive process, different individuals may set views on different variables. If not, each user should use its own bias corrections, in which case the biases will most likely differ between individuals.

Considering views from different investors on the same linear combination of variables may also be interesting. In Section 2.2, the matrix $\mathbf{H}_t$ is said to be required to have full rank. This would mean that having multiple views on the same linear combination of variables is not an option. Using the approach to set $\mathbf{\Omega}_t$ outlined in Section 3.3.1, specifying a linear combination $\mathbf{H}_t$ without full rank would result in a singular matrix in the expression of the Kalman gain in the filter recursion. However, when formulating views of the same linear combination of variables by different investors, the approach to setting the covariance of the views would clearly have to be changed. Instead, the covariance of the views should explain both the relative trust in the views by each investor and the relative trust of the views and the baseline model. Since the recursion would be tractable also for rank-deficient matrices $\mathbf{H}_t$ as long as $\mathbf{\Omega}_t$ is positive definite, this opens up some possibilities. One approach could be to assume that $\mathbf{\Omega}_t$ is diagonal, where the magnitude of the diagonal elements may be time-varying and dependent on the financial landscape. One investor may for example be more reliable in volatile markets, while another investor performs better in more stable financial landscapes. Another way of letting investors specify views on the same combinations of variables could be a two-step approach, where the level of the views $\boldsymbol{\psi}_t$ is specified by some weighted sum of investor views, where the weights may be time-varying.

6. Conclusion

The proposed method has some clear advantages in comparison to targeting the moments when estimating the parameters. Due to the form of the process in (2.5), the only requirement is that any variable in the ESG upon which views are to be imposed can be modelled as a first-order Markov chain and is lowest in hierarchy. This means that the distribution of the current value is completely determined by the previous value, and that the current value of the variable is not a predictor for another variable. Furthermore, the filter recursion has a closed form for linear Gaussian models, including the vector autoregressive model and the Dynamic Nelson-Siegel (state-space) model. Although the VARX(p,q) process can only model linear dependencies, the possibilities of including external regressors, deterministic trends and Gaussian driven influences makes it a fairly flexible model. For more general models, the computational costs of applying a Sequential Monte Carlo based filtering would still be substantially lower than direct moment targeting with re-simulation at each optimization step. However, the method is of course not directly comparable to directly targeting the moments when estimating the model. While the moments can be derived for each input with the proposed method, the second moment cannot be targeted. On the other hand, the results could be considered as more interpretable since the model parameters are still the Maximum Likelihood estimates.

As discussed, there are four main issues with the proposed method. The issue of contradicting views can to a large extent be solved by imposing relevant restrictions, possibly dependent on which variable is modelled. While the proposed informal restrictions could be seen as a starting point, one may also have to consider relationships between variables modelled in different process within the ESG – which are not present in the same VARX(p,q) model. Regarding the issue of bias and model validation, introducing dynamically updated (random) biases could illuminate the tendencies to over- or underestimate certain parameters by individuals, as well as correct for this when producing forecasts. While a suggestion of how a bias correcting model may look is proposed in Section 5.4, practical considerations of how to estimate or specify the biases and dynamically update the model would have to be taken. Estimating the biases to account for long-term forecasts is difficult. Furthermore, some restrictions may have to be considered, such as always imposing views on the same linear combinations of variables.

The most problematic issues are the loss of conditionality, where the baseline model may be heavily dependent on previous observations as well as external regressors, and the difficulty of applying the method on hierarchical models. The incorporation of the views in the process (2.5) is formulated such that the views are seen as noisy observations of the *conditional expectation* of the specified linear combination of the dependent variables at a given time point. The actual views will however generally be *unconditional.* An expert may think that the credit spread is expected to be on a certain level with some uncertainty at a given time in the future. However, since he cannot observe the path until that point, manually specifying conditional views is impossible. While specifying conditional views with functional relationships could be a possibility with a sampling approach to setting the views, further investigation of the resulting moments would have to be made to form clear interpretations for how to set the level and the uncertainty of the views. Furthermore, specifying *unconditional* views over the whole forecast horizon is difficult in itself, and requiring the user to enter functional inputs will not always be an option. However, formulating the views as any function of the external regressors is easily done, since the filter recursion is done conditional on one path of $\{\mathbf{u}_t\}$.

The fact that the method cannot be used to sample paths means that the dynamics of all variables with views must be accounted for in the filter recursion. With multiple hierarchical levels, and possibly views on the volatility process, this gets complicated fast even in the case with linear dependencies and Gaussian errors. In the general case, it may be impossible to even formulate the likelihoods. Whether the complexity is justified in these cases is questionable, since the results may be difficult to interpret. However, for linear models without views on the volatility process, there is the possibility of modelling the previously external regressors as internal, and only allow for lagged dependence. In a setting where long-term forecasts are of interest, one may even choose to rewrite the VARX(p,q) model as a VAR(p) model without re-estimating the parameters, since the asymptotic distribution would be the same.

The proposed method may be seen as blueprint at this stage, and some suggestions for further research are therefore given. Investigating the prospect of setting the views conditionally, and how *unconditional* views about future moments can be transformed into conditional views for specific values of the VARX(p,q) parameters. How can the conditional expectations be filtered such that the views about unconditional moments do not interfere with the conditional behavior of the process? In a setting where some historical views are available, or where views are stored over time, introducing biases to the model would provide useful information and likely yield better results. Investigating how these biases can be estimated, and how the model can be updated when data becomes available, would therefore be another topic for research. Finally, extending the method such that views can be incorporated on all variables in models with multiple hierarchical levels and possibly also volatility processes would make the method more directly applicable to general ESGs. However, constructing a plug-and-play approach for complex economic scenario generators with general processes would not be possible analytically.

6. References

[1] Black, F. and Litterman, R. “Global Portfolio Optimization”. In: *Financial Analysts Journal* vol. 48.5 (1992), pp. 28–43.

[2] Brandt Petersen, K. and Syskind Pedersen, M. *The Matrix Cookbook.* Nov. 2012. URL: `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf` (visited on 03/11/2020).

[3] Crassidis, J. L. and Junkins, J. L. *Optimal Estimation of Dynamic Systems.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Taylor & Francis Group, LLC, 2012. Chap. 3.3.4.

[4] Ding, S. and Cook, R. D. “Dimension Folding PCA and PFC for Matrix-Valued Predictors”. In: *Statistica Sinica* vol. 24.1 (2014), pp. 463–492.

[5] Friedman, J., Hastie, T., and Tibshirani, R. “Sparse Inverse Covariance Estimation with the Graphical Lasso”. In: *Biostatistics* vol. 9.3 (2008), pp. 432–441.

[6] Hildebrand, P. et al. *Understanding Uncertainty.* Apr. 2019. URL: `https://www.blackrock.com/corporate/literature/whitepaper/bii-portfolio-perspectives-april-2019.pdf` (visited on 04/08/2020).

[7] Julier, S. J. and Uhlmann, J. K. “New extension of the Kalman filter to nonlinear systems”. In: *Signal Processing, Sensor Fusion, and Target Recognition VI.* Vol. 3068. International Society for Optics and Photonics. SPIE, 1997, pp. 182–193.

[8] Kalman, R. E. “A New Approach to Linear Filtering and Prediction Problems”. In: *Journal of Basic Engineering* vol. 82 (1960), pp. 35–45.

[9] Koopman, S. J., Mallee, M. I. P., and Van der Wel, M. “Analyzing the Term Structure of Interest Rates Using the Dynamic Nelson–Siegel Model With Time-Varying Parameters”. In: *Journal of Business & Economic Statistics* vol. 28.3 (2010), pp. 329–343.

[10] Lee, J. H. and Ricker, N. L. “Extended Kalman Filter Based Nonlinear Model Predictive Control”. In: *Industrial & Engineering Chemistry Research* vol. 33.6 (1994), pp. 1530–1541.

[11] Lindström, E., Madsen, H., and Nygaard Nielsen, J. *Statistics for Finance.* Chapman and Hall/CRC Press, 2015, pp. 302–305.

[12] Lopes, H. F. and Tsay, R. S. “Particle Filters and Bayesian Inference in Financial Econometrics”. In: *Journal of Forecasting* vol. 30.1 (2011), pp. 168–209.

[13] Lütkepohl, H. *New Introduction to Multiple Time Series Analysis.* Springer Science & Business Media, 2005.

[14] Nelson, C. R. and Siegel, A. F. “Parsimonious Modeling of Yield Curves”. In: *The Journal of Business* vol. 60.4 (Oct. 1987), pp. 473–489.

[15] Parker, T. *Cognitive vs. Emotional Investing Bias: What’s the Difference?* July 2019. URL: `https://www.investopedia.com/articles/investing/051613/behavioral-bias-cognitive-vs-emotional-bias-investing.asp` (visited on 05/24/2020).

[16] Pedersen, H. et al. “Economic Scenario Generators: A Practical Guide”. In: *The Society of Actuaries* (July 2016).

[17] Rasmussen, C. E. “Gaussian Processes in Machine Learning”. In: *Summer School on Machine Learning.* Springer. 2003, pp. 63–71.

[18] Särkkä, S. *Bayesian Filtering and Smoothing*. Vol. 3. Cambridge University Press, 2013, pp. 56–62.

[19] scikit-learn developers. *scikit-learn user guide*. Release 0.20.4. July 2019. URL: `https://scikit-learn.org/0.20/_downloads/scikit-learn-docs.pdf` (visited on 05/10/2020).

[20] Steehouwer, H. and van der Schans, M. "Time-Dependent Black-Litterman". In: *Journal of Asset Management* vol. 18 (Sept. 2017), pp. 371–387.

[21] Tibshirani, R. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* vol. 58.1 (1996), pp. 267–288.

Appendices

A. Lag Operator

The lag operator (L), sometimes denoted the backshift operator (B), is a time series operation which retrieves the previous observation. For a given (multivariate) time series $\{\mathbf{x}_t\}$ and a constant (vector) $\mathbf{c}$, the following equalities hold

$$\begin{aligned} L\mathbf{x}_t &= \mathbf{x}_{t-1} \\ L^{-1}\mathbf{x}_{t-1} &= \mathbf{x}_t \\ L^k\mathbf{x}_t &= \mathbf{x}_{t-k} \\ L\mathbf{c} &= \mathbf{c} \end{aligned}$$

Thus, the vector autoregressive moving average model in section 2.1 can be written as

$$A(L)\mathbf{x}_t = \mathbf{c} + M(L)\mathbf{e}_t$$

where $A(L) = \mathbf{I}_n - \mathbf{A}_1 L - \cdots - \mathbf{A}_p L^p$ and $M(L) = \mathbf{I}_n + \mathbf{M}_1 L + \cdots + \mathbf{M}_q L^q$. Therefore, the VAR($\infty$) representation given by

$$\mathbf{x}_t = \boldsymbol{\pi}_0 + \sum_{i=1}^{\infty} \mathbf{\Pi}_i \mathbf{x}_{t-i} + \mathbf{e}_t$$

can be calculated by (Lütkepohl [2005]) comparing coefficients in

$$\begin{gathered} \mathbf{x}_t - \sum_{i=1}^{\infty} \mathbf{\Pi}_i \mathbf{x}_{t-i} = M(L)^{-1}A(L)\mathbf{x}_t = M(L)^{-1}\mathbf{c} + \mathbf{e}_t \\ \Longrightarrow \\ \mathbf{I}_n - \sum_{i=1}^{\infty} \mathbf{\Pi}_i L^i = M(L)^{-1}A(L) \end{gathered}$$

The intercept is given by

$$\boldsymbol{\pi}_0 = M(L)^{-1}\mathbf{c}$$

Finding $M(L)^{-1}$ is however non-trivial, and it involves matching coefficients in

$$M(L)M(L)^{-1} = (\mathbf{I}_n + \mathbf{M}_1 L + \cdots + \mathbf{M}_q L^q)(\mathbf{M}_0^- + \mathbf{M}_1^- L + \mathbf{M}_2^- L^2 + \dots) = \mathbf{I}_n$$

where a large number of lags in $M(L)^{-1}$ may be needed depending on the desired tolerance level. Taking the simple example of a univariate MA(1) lag polynomial, the inverse is calculated as

$$\begin{aligned} M(L)M(L)^{-1} &= (1 + bL)(b_0^- + b_1^- L + b_2^- L^2 \dots) \\ &= b_0^- + (b_1^- + bb_0^-) + (b_2^- + bb_1^-)L^2 \dots \\ &= 1 \end{aligned}$$

giving

$$(b_0^-, b_1^-, b_2^-, \dots) = (1, -b, b^2, \dots)$$

B. Matrix Operations

B.1 Vectorization and Kronecker Products

Definition (Kronecker product)**.** The Kronecker product of two matrices $\mathbf{A} \in \mathbb{R}^{n \times m}$ and $\mathbf{B} \in \mathbb{R}^{p \times q}$ is denoted $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{np \times mq}$ and given by

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \dots & a_{1m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\mathbf{B} & \dots & a_{nm}\mathbf{B} \end{pmatrix}$$

The Kronecker product has the following properties (Brandt Petersen and Syskind Pedersen [2012])

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = \mathbf{AC} \otimes \mathbf{BD} \tag{B.1}$$

$$(\mathbf{A} \otimes \mathbf{B})^{-1} = (\mathbf{A}^{-1} \otimes \mathbf{B}^{-1}) \tag{B.2}$$

$$|\mathbf{A} \otimes \mathbf{B}| = |\mathbf{A}|^{\mathrm{rank}(\mathbf{B})} |\mathbf{B}|^{\mathrm{rank}(\mathbf{A})} \tag{B.3}$$

$$(\mathbf{A} \otimes \mathbf{B})^\intercal = \mathbf{A}^\intercal \otimes \mathbf{B}^\intercal \tag{B.4}$$

Definition (Vectorization)**.** The vectorization of a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ is denoted $\mathrm{vec}(\mathbf{A}) \in \mathbb{R}^{nm}$ and given by

$$\mathrm{vec}(\mathbf{A}) = \begin{pmatrix} a_{11} \\ \vdots \\ a_{n1} \\ \vdots \\ a_{1m} \\ \vdots \\ a_{nm} \end{pmatrix}$$

The vec operator has the following properties (Brandt Petersen and Syskind Pedersen [2012])

$$\mathrm{vec}(\mathbf{AXB}) = (\mathbf{B}^\intercal \otimes \mathbf{A})\mathrm{vec}(\mathbf{X}) \tag{B.5}$$

$$\mathrm{tr}(\mathbf{A}^\intercal \mathbf{B}) = \mathrm{vec}(\mathbf{A})^\intercal \mathrm{vec}(\mathbf{B}) \tag{B.6}$$

$$\mathrm{vec}(\mathbf{A}) + \mathrm{vec}(\mathbf{B}) = \mathrm{vec}(\mathbf{A} + \mathbf{B}) \tag{B.7}$$

B.2 Matrix Derivatives

All of the following derivative rules are given by (Brandt Petersen and Syskind Pedersen [2012]).

If $F(\mathbf{X})$ is differentiable function of each of the variables of $\mathbf{X}$, then

$$\frac{\partial \mathrm{tr}(F(\mathbf{X}))}{\partial \mathbf{X}} = f(\mathbf{X})^{\mathsf{T}}$$

where $f(\cdot)$ is the scalar derivative of $F(\cdot)$. Other derivative rules used in this paper are

$$\begin{aligned}\frac{\partial |\mathbf{X}|}{\partial \mathbf{X}} &= |\mathbf{X}|\mathbf{X}^{-\mathsf{T}} \\ \frac{\partial \mathbf{Y}^{-1}}{\partial x} &= -\mathbf{Y}^{-1}\frac{\partial \mathbf{Y}}{\partial x}\mathbf{Y}^{-1}\end{aligned}$$

C. Multivariate Gaussian

C.1 Conditional Distribution

Let $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$ be two multivariate Gaussian random variables with expectations $\boldsymbol{\mu_x}$ and $\boldsymbol{\mu_y}$ respectively. Then, the random variable $\mathbf{z} = (x_1, \ldots, x_n, y_1, \ldots, y_m)$ is distributed as

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu_x} \\ \boldsymbol{\mu_y} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{xx}} & \boldsymbol{\Sigma}_{\mathbf{xy}} \\ \boldsymbol{\Sigma}_{\mathbf{yx}} & \boldsymbol{\Sigma}_{\mathbf{yy}} \end{pmatrix} \right) \tag{C.1}$$

Further consider the case where $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{y}|\mathbf{x} \sim \mathcal{N}(\mathbf{Ax}, \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}})$. Formula (C.1) then gives

$$\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A}\boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\mathbf{A}^\top \\ \mathbf{A}\boldsymbol{\Sigma} & \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}} \end{pmatrix} \right)$$

The conditional distributions of $\mathbf{x}$ given $\mathbf{y}$ is then given by (Brandt Petersen and Syskind Pedersen [2012])

$$\begin{aligned} \mathbf{x}|\mathbf{y} &\sim \mathcal{N}\big(\boldsymbol{\mu_x} + \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}{}^{-1}(\mathbf{y} - \boldsymbol{\mu_y}), \boldsymbol{\Sigma}_{\mathbf{xx}} - \boldsymbol{\Sigma}_{\mathbf{xy}}\boldsymbol{\Sigma}_{\mathbf{yy}}^{-1}\boldsymbol{\Sigma}_{\mathbf{yx}}\big) \\ &\overset{d}{=} \mathcal{N}\big(\boldsymbol{\mu} + \boldsymbol{\Sigma}\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}})^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\mu}), \boldsymbol{\Sigma} - \boldsymbol{\Sigma}\mathbf{A}^\top(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top + \boldsymbol{\Sigma}_{\mathbf{y}|\mathbf{x}})^{-1}\mathbf{A}\boldsymbol{\Sigma}\big) \end{aligned}$$

C.2 Matrix Normal Distribution

Definition (Matrix normal distribution)**.** Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ be a random matrix such that $\text{vec}(\mathbf{X}) \sim \mathcal{N}\left(\text{vec}(\mathbf{M}), \boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}\right)$ where $\boldsymbol{\Omega} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n\times n}$ are positive definite. Then, the matrix $\mathbf{X}$ has the probability density (Ding and Cook [2014])

$$f_{\mathbf{X}}(\mathbf{X}) = (2\pi)^{-\frac{nm}{2}}|\boldsymbol{\Omega}|^{-n/2}|\boldsymbol{\Sigma}|^{-m/2}\text{exp}\left(-\frac{1}{2}\text{tr}\left[\boldsymbol{\Omega}^{-1}\left(\mathbf{X} - \mathbf{M}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{M}\right)\right]\right)$$

This is called a matrix normal distribution and is often denoted $\mathcal{MN}_{n\times m}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Omega})$

Proof. Using B.2, B.7, B.5 and B.6 in the written order, the exponent can be rewritten as

$$\begin{aligned} -\frac{1}{2}\left(\text{vec}(\mathbf{X}) - \text{vec}(\mathbf{M})\right)^\top\left(\boldsymbol{\Omega} \otimes \boldsymbol{\Sigma}\right)^{-1}\left(\text{vec}(\mathbf{X}) - \text{vec}(\mathbf{M})\right) &= -\frac{1}{2}\text{vec}(\mathbf{X} - \mathbf{M})^\top\left(\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Sigma}^{-1}\right)\text{vec}(\mathbf{X} - \mathbf{M}) \\ &= -\frac{1}{2}\text{vec}(\mathbf{X} - \mathbf{M})^\top\text{vec}\left(\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{M}\right)\boldsymbol{\Omega}^{-1}\right) \\ &= -\frac{1}{2}\text{tr}\left[\boldsymbol{\Omega}^{-1}\left(\mathbf{X} - \mathbf{M}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{M}\right)\right] \end{aligned}$$

Using B.3, one has

$$|\mathbf{\Omega} \otimes \mathbf{\Sigma}|^{-1/2} = |\mathbf{\Omega}|^{-n/2}|\mathbf{\Sigma}|^{-m/2}$$

□