Non-parametric Bayesian methods and deep learning for non-linear state-space models

Alireza Kahali
(2542764)

VU University Amsterdam
Department of Econometrics and Operations Research

Submitted in partial completion of the
Master of Science in Econometrics and Operations Research

March 7, 2017

Abstract

Gaussian Process based methods, neural networks, non-parametrics in Econometrics and signal processing techniques are reconciled to set-up methodology that is both computationally efficient as well as effective for semi-automatic modeling. In particular the methodology allows for non-linear state space model estimation without specifying the functional form of the equation of interest when there is a lack of large amounts of data. Moreover the functional form estimation is done in a, "see through" black-box Bayesian framework, thereby allowing for intuitive uncertainty quantification on the model level. Although the focus is on not so large data-sets, an approximation method is employed to allow for fast hierarchical Bayesian estimation and allow the model to scale efficiently with the sample size. A convolution covariance structure for the Gaussian Process prior is presented, through its spectral representation, allowing for computationally efficient and analytically tractable way of setting up multidimensional covariance structures of the prior. Furthermore to promote better frequentist coverage, the regularity of the prior sample paths are allowed to adapt to the data. In a rigorous simulation setting the performance of the sampling based learning algorithm as well as the ability to recover the true data generating distribution, in a finite sample setting, is verified. As an application two novel versions of the stochastic volatility model are introduced, namely the Reduced Rank Gaussian Process Stochastic Volatility (RR GP-SSM) model and the Deep RR GP-SSM. In-sample estimation results on index and stock data show that the proposed models recover stylized properties that are well accepted in the financial time series literature. Lastly forecasting performance on index and stock data is compared to classical volatility models, which are significantly outperformed by the RR GP-SSM and the deep RR GP-SSM.

Keywords: Non-parametrics, deep learning, state-space models, Bayesian learning, Gaussian Process, stochastic volatility

Non-parametric Bayesian methods and deep learning for non-linear state-space models

Alireza Kahali
(2542764)

VU University Amsterdam

Department of Econometrics and Operations Research

Thesis Committee:
Dr. Francisco Blasques
Dr. Charles Bos

Submitted in partial completion of the

Master of Science in Econometrics and Operations Research

March 7, 2017

Acknowledgements

Approximately three years ago, I knew little about mathematics, statistics, econometrics or computer science. At the moment of deciding to move from Business studies towards this field I was quite uncertain if I would be able to reach the Econometrics and Operations Research Master program, let alone me being able to complete it successfully. This thesis signifies the completion of the Master program and I want to sincerely thank Dr. Francisco Blasques for constantly challenging me to think about what is going on beneath the surface. After each one of our meetings I was left thinking: "alright, I guess I have not figured it out yet". Although my second supervisor, Dr. Charles Bos, came in later I am sincerely grateful to him for also giving me some important food for thought and granting me access to the server that allowed me to complete my empirical study. Finally I am grateful to my girlfriend for supporting me throughout these years, even-though I spent most of my time in the books and probably talked too much about my studies.

Abstract

Gaussian Process based methods, neural networks, non-parametrics in Econometrics and signal processing techniques are reconciled to set-up methodology that is both computationally efficient as well as effective for semi-automatic modeling. In particular the methodology allows for non-linear state space model estimation without specifying the functional form of the equation of interest when there is a lack of large amounts of data. Moreover the functional form estimation is done in a, "see through" black-box Bayesian framework, thereby allowing for intuitive uncertainty quantification on the model level. Although the focus is on not so large data-sets, an approximation method is employed to allow for fast hierarchical Bayesian estimation and allow the model to scale efficiently with the sample size. A convolution covariance structure for the Gaussian Process prior is presented, through its spectral representation, allowing for computationally efficient and analytically tractable way of setting up multidimensional covariance structures of the prior. Furthermore to promote better frequentist coverage, the regularity of the prior sample paths are allowed to adapt to the data. In a rigorous simulation setting the performance of the sampling based learning algorithm as well as the ability to recover the true data generating distribution, in a finite sample setting, is verified. As an application two novel versions of the stochastic volatility model are introduced, namely the Reduced Rank Gaussian Process Stochastic Volatility (RR GP-SSM) model and the Deep RR GP-SSM. In-sample estimation results on index and stock data show that the proposed models recover stylized properties that are well accepted in the financial time series literature. Lastly forecasting performance on index and stock data is compared to classical volatility models, which are significantly outperformed by the RR GP-SSM and the deep RR GP-SSM.

Keywords: Non-parametrics, deep learning, state-space models, Bayesian learning, Gaussian Process, stochastic volatility

Contents

List of Figures

List of Abbreviations

SSM	State-Space Model
GARCH	Generalized Autoregressive Conditional Heteroscedasticity
GJR-GARCH	Glosten, Jagannathan, and Runkle Generalized Autoregressive Conditional Heteroscedasticity
EGARCH	Exponential Generalized Autoregressive Conditional Heteroscedasticity
RKHS	Reproducing Kernel Hilbert Space
PSD	Positive Semi-Definite
MLE	Maximum Likelihood Estimator
GP	Gaussian Process
LTI	Linear time-invariant
SMC	Sequential Monte Carlo
PF	Particle Filter
PG	Particle Gibbs
PGAS	Particle Gibbs with Ancestral Sampling
KF	Kalman Filter
DGP	Data Generating Function
RMSE	Root Mean Square Error
ACF	Auto Correlation Function
IF	Inefficieny Factor
SE	Squared Exponential
SV	Stochastic Volatility
DGP	Data Generating Process
OU	Ornstein–Uhlenbeck

1 Introduction

The modeling of time series data in order to capture the dynamics of underlying data generating systems, of which we assume they exist, comes with various names and flavours depending on the field. In systems and control the process of capturing the dynamics by specifying a model and tuning its parameters is called systems identification (Walter and Pronzato 1997). In the econometrics or machine learning domain the process of systems identification is referred to as time series estimation, fitting, or learning (Durbin and Koopman 2012; Hamilton 1994; Barber et al. 2011). The areas in which this is applied are as wide as forecasting volatility, music analysis, robotics or genetic sequence analysis.

If we consider a black-box system with inputs and outputs as a model for the time series, we can let the observations $\{y_t\}_{t=1}^T$ be the output of the system. Now one way of modeling uncertainty is considering a distribution for $\{y_t\}_{t=1}^T$, for instance in the AR(p) model, nested in the framework of Box et al. (1994), the additive errors ϵ_t account for this. Thus we would have:

$$y_t \sim \mathbb{P}(y_t|y_{t-1}, \dots, y_{t-p}) \text{ for some p>1} \tag{1.1}$$

Suppose we have observations $\{y_t\}_{t=1}^{T-1}$ then this distributional assumption only affects the future output y_T but we do not account for uncertainty regarding the observations. It is unrealistic to assume noiseless measurements, where the classical example is noisy observation from sensors. Suppose there exists a certain set of information about the system that can determine its trajectory in the future referred to as its state. In State-Space modelling the states are modelled as latent variables with a certain distributional assumption. The hierarchical model where the states and inputs are higher up the chain as in equation 1.2 takes into account both measurement uncertainty as well as state uncertainty. Let $\mathbf{x}_t \in \mathbb{R}^d, \mathbf{u}_t \in \mathbb{R}^e, \mathbf{y}_t \in \mathbb{R}^n$, the general form of the state-space models we consider (Durbin and Koopman 2012) is:

$$\mathbf{y}_t \sim \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t), \qquad \mathbf{x}_{t+1} \sim \mathbb{P}_x(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t), \qquad \mathbf{x}_0 \sim \mathbb{P}_0(\mathbf{x}_0) \tag{1.2}$$

Where $\{\mathbf{y}_t\}_{t=1}^T$ are the observations, $\{\mathbf{x}_t\}_{t=1}^T$ the latent states , $\{\mathbf{u}_t\}_{t=1}^T$ the inputs and $\mathbf{x}_0$ the initialization. In smaller sample situations, we can restrict the general form in various

ways to downsize the model space as we wish by specifying the functional form, as well as some parametric family for the measurement- and the state-equation. Note that the states are latent. Furthermore when there is less data, given the increase in model possibilities, inference is of importance. Depending on the complexity of the parametric functional form, the probabilistic model and the estimation method in combination with the available amount of data, complications such as over-fitting can be a real danger lurking beneath the surface. Therefore it would be good idea to have a model with adaptive complexity according to the amount of available information. Non-parametric models, including Bayesian ones, aim to allow the complexity to automatically adapt to the modeling situation at hand. Furthermore in theory such non-parametric models can allow for functional form estimations in more general spaces, and although in theory the infinite parameter space could result in no misspecification, they can in practice at least lessen the specification error.

Often non-parametric methods and popular semi-parametric ones, such as deep neural networks are quite hungry for data, and as the amount of data decreases the problem of estimating an underlying data generating function from this finite data-set becomes more and more "ill-posed". To overcome this problem certain assumptions about this latent function are needed, but we wish to do this whilst remaining in the non-parametric setting. Non-parametrics in the Bayesian framework allow for higher level structural assumptions on these models, such as on the level of smoothness or periodicity, rather than on the functional form.

After a model has been specified, results strongly depend on the estimation techniques employed. Vanilla Maximum likelihood estimation for example is known for its proneness to over-fitting, especially in the presence of high flexibility. In the context of basis expansions corresponding to Gaussian Process SSM (with fixed hyper-prior parameters) Svensson and Schön (2016) compare various estimation methods amongst which vanilla and regularized (penalized) maximum likelihood. The former clearly performs poorly in comparison, especially in regions of the state space where there is little data available to excite the dynamics. In those regions there are many models possible, and MLE as well as regularized MLE only capture 1 (where the latter captures a much better suited one), whereas full Bayesian learning averages over all those possible models. We want a flexible model, in combination with an estimation/learning procedure, that aims to avoid over-fitting and propagates uncertainty about various regions of interest of the state-space forward in the form of a distribution rather than a point estimate. On the model level this would give us an estimated distribution over the function. And for decision making, we would also like to make function predictions with uncertainty quantification at locations of the state-space where there was little data to begin with. And finally given that we want these models to be useful for decision making we aim to make these black-box models "see-through", and aim for automatic modeling.

For these purposes Gaussian Process (GP) prior state-space models have been proposed (Frigola, Lindsten, Schön, and C. Rasmussen 2013; Svensson, Solin, et al. 2015), but although these models do allow for most of what we want, they come at a computational cost that

makes their implementation hard. To alleviate the computational burden many solutions have been proposed and after reviewing these we delve into one as the main approach that fits the objectives of this thesis. For this approximate model we find that as a by-product of the approximation approach we can easily implement convolution kernels over the state-dimensions through their spectral representations. We also find out that heuristically it is possible to let the regularity of the stochastic process prior sample paths be adaptable to the data, thereby promoting better frequentist coverage of the credible regions of the posterior we consider. For learning of the approximate model we employ a Blocked Gibbs scheme and find in a simulation setting that we can recover the distributions over the DGP parameters sufficiently well with a low amount of particles and Metropolis-within-Gibbs runs, and that a sample size of T=500 is sufficient. Of the two covariance function forms we use, the spectral mixture kernel (A. Wilson et al. 2014) and the Matern class one (Stein 1999), we find the former to not work so well in combination with the blocked Gibbs algorithm and the latter to give satisfactory results in terms of DGP recovery. We find depending on the DGP state transition function that relatively low expansion orders suffice and that when we have more basis functions than needed the posterior distribution around the unneeded weights are closely distributed around zero. From a frequentist point of view this relates to penalizing unneeded basis function weights commonly referred to as regularization.

As an application we set the GP-SSM up in a form similar to the log-normal discrete stochastic volatility model. These models are coined Reduced Rank Gaussian Process Stochastic Volatility (RR-GPSV) and the multilayer RR-GPSV. The key differences are that the transition equations are non-parametric now and we let the leverage effect be discovered by the functional form rather than through correlation between the state and measurement function errors. When we employ the proposed model in an in-sample context for index and stock data we find that the functional form estimation of the transition function is in line with certain stylized facts from the financial time series literature. In particular the estimated functional forms are able to capture volatility clustering and the leverage effect.

In a forecasting setting the presented models seem to outperform classical parametric volatility models on small and larger data-sets. The two-layer "deep" model of chapter 5 seems to perform similarly as the one-layer model in our forecasting experiments but is highly inefficient in terms of computation. It is likely that the current MCMC scheme for learning is not suited for a deeper model.

The contributions of this thesis are as follows. In terms of literature, an extensive (not exhaustive) review of the latest research is given, mostly conducted in the fields of engineering and robotics, and connected to topics in econometrics. In terms of methodology we present two novel econometric models: Reduced Rank GP-SV and Deep Reduced Rank GP-SV. Furthermore we expand the approximation method of Svensson, Solin, et al. (2015) to allow for a mean function to be employed and provide technical tools (proofs) for the effective implementation of a useful kernel. Furthermore we build a multidimensional

convolution covariance function that is computationally in efficient in multiple dimensions and rests on sound theory. In that sense we take into account an approximation method, a computationally efficient covariance structure, as well as careful considerations with regards to the frequentist coverage of the posterior regions we deem to be credible. Especially the last point is rarely discussed in Machine learning literature and certainly not in the context of the Gaussian Process State-space models. Often the Exponentiated-Quadratic/ Squared-exponential covariance function is employed, which has a high chance of leading to unreliable credible regions of the posterior in terms of frequentist coverage. By extension of the lack of popularity of GP based methods in the econometrics literature these are, to the best of my knowledge, not at all discussed there. We also aim to enrich the structure discovery procedure by employing the spectral mixture kernel of A. G. Wilson and Adams (2013) but this turns out to not be so efficient in combination with our sampling procedure for estimation. We also expose the models and inference algorithm to a rigorous simulation study, which I could not find in the context of GP-SSM's in the literature. Then finally, inspired by Mattos, Dai, et al. (2015), we set up a so called "deep" model for structure discovery and briefly experiment with it on data.

The following chapters of this thesis are structured as follows. In chapter 2 we give a review on Bayesian non-parametric methods, and ask ourselves the question of how these methods relate to other popular approaches in econometrics and machine learning (Linear Sieves, Neural Networks and kernel methods for example). We seek to answer how and why non-parametric Bayesian methods can be useful for a flexible modeling approach when little data is available. We dig into the theory of stationary covariance functions so that we can build a suitable one for the approximate model in chapter 3. In chapter 2 we also look at some asymptotic properties of non-parametric Bayesian models that are then used in chapter 3 to set up a heuristic for the adaptive regularity of the Gaussian Process prior. In chapter 3 we delve into the reduced rank model and build a convolution covariance structure for it. Furthermore in chapter 3 the complete sampling algorithm for finding the posterior over the reduced rank model is specified. In chapter 4 we discuss how and why we can make the reduced rank model "deeper", and in chapter 5 we discuss how to adapt these models for use in volatility estimation. In chapter 6 the models and sampling algorithm are evaluated in a simulation context and in chapter 7 these are employed in an empirical setting.

2

Review and theory of Bayesian nonparametrics and State-Space modeling

Contents

The aim of this chapter is to answer all the preliminary questions I have regarding the principles of the non-parametric Bayesian approach. While conducting the literature study, questions such as what is meant with a stochastic process as a prior, what is the role of the covariance function of that prior within the non-parametric Bayesian framework or what role do these priors on functions play in state-space models came up. Furthermore I was anxious to find out how these methods are connected to frequentist non-parametrics and popular non-probabilistic approaches in the machine learning community such as neural networks. These questions are answered in the subsequent sections of this chapter.

2.1 Probabilistic fundamentals

In this section we lay the groundwork for the probabilistic framework we employ when constructing suitable models for the applications of chapter 5. We only give an overview here and highlight only that which needs to be addressed in Bayesian Non-Parametrics and is used throughout the thesis.

2.1.1 Classical definitions

From a classical point of view, when we speak of a parametrized model we refer to the following:

Ch2. Definition 1. *A (frequentist) parametrized statistical model* $\mathbb{P}_\theta := \{p_\theta : \theta \in \Theta\}$ *on the measurable sample space* $(\mathcal{X}, \mathcal{B})$ *is a collection of probability measures* $p_\theta : \mathcal{B} \to [0,1]$ *parametrized by* θ *in the parameter space* Θ*. We assume here that the mapping from the parameter space* $\Theta \to \mathbb{P}_\theta : \theta \mapsto p_\theta$ *is bijective. We denote the collection of all probability measures on* $\mathcal{X}$ *with* $\mathcal{M}(\mathcal{X})$*.*

To denote the models as collections of densities we need to have that the model is dominated.

Ch2. Definition 2. *Let* p *and* μ *be two measures on the space* $(\mathcal{X}, \mathcal{B})$ *then we say that* p *is absolutely continuous with respect to* μ *if* $\mu(B) = 0 \Rightarrow p(B) = 0$*. We also say that* p *is dominated by* μ *and write* $p << \mu$*. And if* $\mathbb{P}_\theta$ *is a collection of probability measures on* $\mathcal{X}$ *we call it dominated and write* $\mathbb{P}_\theta << \nu$*, if there is a* σ*-finite measure on the space* $\mathcal{X}$ *such that* $p_\theta << \mu$ *for all* $p_\theta \in \mathbb{P}_\theta$

In order to get to a density we can start by constructing new measures from old ones as a form of point-wise re-weighting. For example if g is some non-negative measurable function then $p_\theta(B) = \int_B dp_\theta = \int_B g d\mu_x$ for this we need the following theorem.

Ch2. Theorem 1 (Radon-Nikodym)**.** *Let* p_θ *be a measure on* $(\mathcal{X}, \mathcal{B})$ *and* μ_x *a* σ*-finite measure on* $(\mathcal{X}, \mathcal{B})$ *such that* $p_\theta << \mu_X$*. Then there exists a real valued (so certainly a non-negative) function* $g : \mathcal{X} \to [0, \infty)$ *such that for any* $B \in \mathcal{B}$

$$p_\theta(B) = \int_B dp_\theta = \int_B g(x)\mu_x(dx)$$

We call g *the Radon-Nikodym derivative:*

$$\frac{dp_\theta}{d\mu_x}(x) = g(x)$$

In practice we usually have dominated models, for example the counting measure is a σ-finite measure on the discrete space and dominates all other measures in the space. Later on in the thesis when we say for instance that the spectral spectral measure has a density we mean that this Radon-Nikodym derivative exists. In the parametrized case we consider the model to be well-specified if it holds that the true data generating distribution of the data $p_0 : \mathcal{X} \to [0,1]$, of which we assume that $x \in \mathcal{X} \sim p_0$, is contained in the in the model $\mathbb{P}_\theta$. Then for an estimator such as the MLE asymptotic results such as consistency and some form of a central limit theorem have been established. When we speak of a non-parametric model we refer to dimension of the parameter space being infinite. Intuitively I like to imagine the

model using as many parameters as it needs from an infinite dimensional space in order to adjusts its complexity to the amount of available data. It does not mean that we necessarily have an infinite amount of parameters nor does it mean that we have no parameters. As an intuitive example suppose we are fitting a Gaussian distribution to data using maximum likelihood then we have, no matter the amount of available data, only 2 degrees of freedom (its location and scale). On the other hand suppose we employ a kernel density estimator and place a Gaussian on each data point. Then as the data grows we obtain an additional location and scale parameter, thus in the limit we have an infinite amount of parameters and need an infinite dimensional vector space for these parameters to live in.

2.1.2 Bayesian Framework

Often when presented with the Bayesian framework we encounter the Bayes equation:

$$\mathbb{P}(\theta|x) = \frac{\mathbb{P}(\theta, x)}{\mathbb{P}(x)} \tag{2.1}$$

Where $\mathbb{P}(\cdot)$ is a density and hence it is assumed that all these densities exist. Considering that in this thesis we are concerned with stochastic process priors and other infinite dimensional objects it is a good idea to start speaking of measures rather than densities, in fact in some situations it is possible that the equation as is stated above may not even be valid. On the probability space $(\Omega, \mathcal{A}, p)$ we let the random variable X be a function that induces a measure say μ_x on the sample space $(\mathcal{X}, \mathcal{B})$ and when it is realized we observe the mapping of the function $X(\omega) \in \mathcal{B}$ in $\mathcal{B}$. Suppose we also have the random variable $\theta : (\Omega, \mathcal{A}, p) \to (\Theta, \mathcal{C})$ on the space $(\Theta, \mathcal{C})$ with induced measure $\mu_\theta \in \mathcal{M}(\Theta)$ (Prior). Let the regular conditional measure $\mu_{X|\theta}$ be given by a Markov kernel κ from $(\Theta, \mathcal{C})$ into the sample space $(\mathcal{X}, \mathcal{B})$, that is:

- The map $c \mapsto \kappa(c, x)$ is $\mathcal{C}$- measurable for every $x \in \mathcal{B}$
- The map $x \mapsto \kappa(c, x)$ is a probability measure on $(\mathcal{X}, \mathcal{B})$ for every $c \in \Theta$

Note that to ensure that of this conditional is regular, we assume Θ to be Polish (its topology is metrizable, complete and separable). The Bayesian model $\mathbb{P}_\theta$ then is defined, through the choice of the prior, by the hierarchy:

$$\theta \sim \mu_\theta$$
$$X|\theta \sim \mu_{X|\theta}$$

We speak of a non-parametric Bayesian model when the space Θ is infinite, which is the case when the prior measure μ_θ is the law of a stochastic process whereof the sample paths reside in the function space Θ. Furthermore, on the product space $(\mathcal{X} \times \Theta)$ with the product σ-algebra $\mathcal{B} \times \mathcal{C} = \sigma(\mathcal{B} \times \mathcal{C})$, we have a well defined joint probability distribution $\mu_{X,\theta}$ and the marginal:

$$\mu_X = \int \mu_{X,\theta}(\theta, X)\mu_\theta(d\theta) \tag{2.4}$$

Essential to Bayesian inference is the other conditional distribution, namely the (regular) distribution $\mu_{\theta|X}$, which from a subjectivist point of view is the update of beliefs after observing additional data. In particular we are interested in the Markov kernel from $(\mathcal{X}, \mathcal{B})$ into $(\Theta, \mathcal{C})$, which exists under the (sufficient) assumption of the Polishness of the parameter space. This conditional distribution is referred to as the posterior. If the model $\mathbb{P}_\theta = \{p_\theta : \theta \in \Theta\}$ is dominated then in general it is possible to choose densities such that, relative to some σ-finite dominating measure, the maps $(\theta, x) \mapsto \mathbb{P}(\theta, x)$ are jointly measurable and hence obtain Bayes's formula. The existence of the formula means that the Posterior is dominated by the prior, and it can be written as a Radon-Nikodym derivative (see theorem below). Note for example that although the posterior for the Dirichlet process prior can be defined, because the model is typically not dominated, this posterior is usually not dictated by the Bayes formula (Hjort et al. 2010) (ch2.).

Ch2. Theorem 2 (Bayes)**.** *Suppose the model $\mathbb{P}_\theta = \{p_\theta : \theta \in \Theta\}$ is dominated dominated with with respect to some σ-finite measure, that is $\mathbb{P}_\theta << \nu$. Let the density of the conditional $\mu_{X|\theta}$ with respect to ν be given by $\mathbb{P}(x|\theta)$, then the posterior can be given as the Radon-Nikodym derivative:*

$$\frac{d\mu_{\theta|X}}{d\mu_\theta} = \frac{\mathbb{P}(x|\theta)}{\int \mathbb{P}(x|s)\mu_\theta(ds)} \tag{2.5}$$

This implies that $\mu_{X|\theta} << \mu_\theta$ almost surely with respect to the marginal of X.

2.2 Gaussian processes priors

Ch2. Definition 3 (Stochastic process)**.** *A collection $\mathbf{X} = \{X_t, t \in T\}$ of measurable maps indexed by the set T mapping values in the measurable space $(\mathcal{X}, \mathcal{B})$ is called a stochastic process (Random Field if T is multi-dimensional). The elements X_t map from a probability space $(\Omega, \mathcal{A}, p)$ to $(\mathcal{X}, \mathcal{B})$ where the latter is referred to as the state-space.*

As an example example in the context of housing prices the index t is time and the state space is $(\mathcal{X} = \mathbb{R}, \mathcal{B}(\mathbb{R}))$. For some $\omega \in \Omega$ the sample paths of $\mathbf{X}$ are the mappings $X_t(\omega) : T \to \mathcal{X}$ and hence elements of $\mathcal{X}^T = \mathbb{R}^T$. In the case of a Gaussian Process, depending on the definition of the covariance function, the sample paths can be elements of a desirable subset $F \subset \mathcal{X}^T$, where $(F, \mathcal{F})$ could for instance be the space of smooth continuous functions or maybe periodic ones. A forthright example is the case of $\mathbf{X}$ being the integrated Wiener Process where we have that by definition the sample paths are continuous (not smooth). The distribution of $\mathbf{X} : (\Omega, \mathcal{A}, p) \to (F, \mathcal{F})$ then is $\mathbb{P}(\mathbf{X} \in f)$ for some $f \in \mathcal{F}$. The stochastic process acts as a prior when for the model $\mathbb{P}_\theta$ we have $\theta = f, \Theta = F$.

Two of the most used non-parametric priors are the Dirichlet and the Gaussian process Hjort et al. (2010). Besides the fact that when employing the Dirichlet process prior the posterior is typically orthogonal to the prior (not dominated means no Bayes theorem,

but the posterior is often available), there has been a good number of recent research conducted on Gaussian processes priors in state-space models (Frigola, Lindsten, Schön, and Carl E Rasmussen 2014a; Mattos, Damianou, et al. 2016; Svensson, Solin, et al. 2015; Mattos, Dai, et al. 2015). It is therefore very interesting to delve into Gaussian Process based methods and find out how we can modify these and use them in some econometrics application, which in this thesis is stochastic volatility. Thus in thesis we focus solely on Gaussian Process Priors.

Gaussian Process

Ch2. Definition 4 (Gaussian Process)**.** *Let* $\mathbf{X}$ *be defined as in definition 3.* $\mathbf{X}$ *is called a Gaussian process if for any finite vector* $\{X_{t_j}\}_{t_j \in \mathcal{T}}, \mathcal{T} \subset T, |\mathcal{T}| < \infty$, $\{X_t\}_t$ *is Gaussian. We call this finite vector its finite dimensional distribution. Otherwise put:* $\sum_j c_j X_{t_j}$ *is Normally distributed for all* $c_j \in \mathbb{R}$, $j = 1 \dots n$. *We write* $\mathbf{X} \sim \mathcal{GP}(m, k)$, *where* m, k *are the mean and covariance functions respectively.*

Even more plainly put any finite vector of a Gaussian Process (finite dimensional distribution) has a normal Distribution. The mean function mapping from T given by $m(t) = \mathbb{E}[X_t]$ and the covariance function $k : T \times T \to \mathcal{V}$ (where $\mathcal{V}$ can be $\mathbb{R}$ for example) given by $k(t, s) = \mathbb{C}ov[X_t X_s]$ determine the finite dimensional distributions of the Gaussian process (as is the case for any Gaussian Distribution). In fact when we have m on T and k on $T \times T$ such that k is symmetric and positive definite then the existence of a GP characterized by its mean function m and its covariance function k is implied by Kolmogorov's extension theorem.

Ch2. Theorem 3 (Kolmogorov's extension theorem)**.** *For every finite subset* S *of an arbitrary set* T *let* p_S *be a probability distribution on* $\mathbb{R}^S$. *Then there exists a probability space* $(\Omega, \mathcal{A}, p)$ *and measurable maps* $X_t : \Omega \to \mathbb{R}$ *such that* $\{X_t : t \in S\} \sim p_S$ *for every finite set* S *if and only if for every pair* $S' \subset S$ *of finite subsets ,* $p_{S'}$ *is the marginal distribution of* p_S *on* $\mathbb{R}^{S'}$.

The other way around if we have that $\mathbf{X} \sim \mathcal{GP}(m, k)$ then we can see, by definition 4, that for k we have with $c_j \in \mathbb{R}$, $t_j \in \mathcal{T} \subset T$ and $|\mathcal{T}| < \infty$:

$$\sum_i \sum_j c_i c_j k(t_i, t_j) = \mathbb{V}ar\left[\sum_i c_i X_{t_i}\right] \geq 0 \tag{2.6}$$

Then building GP's can for example be accomplished by constructing valid covariance functions. Note that we can associate to each covariance function an operator in the form of an integral transform:

$$(T_{\mathcal{K}}\phi)(s) = \int_{\mathcal{X}} k(t, s)\phi(s)\mu_x(ds) = \hat{\phi}(t) \tag{2.7}$$

For some finite measure μ_x (possibly Lebesgue measure d_s on $\mathbb{R}$ or some probability measure that has a density $\mathbb{P}(s)d_s$) on the sample space $(\mathcal{X}, \mathcal{B})$ and ϕ in $L^2(\mu)$. In this sense

we often refer to the covariance function as the kernel. For this thesis we restrict our attention to stationary covariance functions, namely the Matern kernel and the spectral mixture kernel, which we discuss later on. When $m(t) = 0$ for all $t \in T$ we call the GP centered. Note that throughout the thesis we interchange the kernel mapping notation $k : T \times T \to \mathcal{V}$ with $k : \mathcal{X} \times \mathcal{X} \to \mathcal{V}$ given by $k(x_t, x_s)$, for x_t, x_s in the sample space $(\mathcal{X}, \mathcal{B})$, when it is more convenient.

The beauty now is that in this manner we can have priors on spaces of functions. Intuitively I interpret the GP prior as follows: if we were to employ a stochastic process prior we could let (each draw of) the prior function values be a sample path of a Gaussian process with a certain mean and covariance structure. Depending on this covariance structure encoded by k of $f : (\Omega, \mathcal{A}, p) \to (F, \mathcal{F})$ we in effect describe the prior space of functions $(F, \mathcal{F})$ that the sample paths find themselves in as well as a probability measure induced on that space. This probability measure places mass on those functions in the space that are in accordance with our prior beliefs.

2.3 Regression example

To elucidate the abstract concepts discussed in the previous sections we go through a regression example that is also referred to at times throughout the thesis. Suppose we have the univariate time series $\mathbf{y}_N = \{y_t\}_{t=1}^N$, and we wish to build a time-series model for it. Suppose we approach modeling the time series under the following prior assumption, let $t \in T \subset \mathbb{R}$:

$$y_t = f(t) + \epsilon_t \quad f \sim \mathcal{GP}(m(t), k(t,s)) \quad y_t | f(t) \sim \mathcal{NID}(f(t), \sigma_\epsilon^2) \tag{2.8}$$

Depending on the covariance function of GP this model nests the classical ARMA models (Turner 2012). Suppose we let $m(t) \equiv 0$, because we lack the necessary prior knowledge about the unknown function f, as can be seen later on this does not mean that the posterior mean is zero. We first describe the stochastic process prior, which in turn is determined by the covariance function k. As an example suppose we let the prior law be the Ornstein–Uhlenbeck (OU) process (intuitively a continuous version of a stationary AR(1)). Let W be a two-sided Brownian motion, σ, $\ell > 0$ and $r = s - t, s, t \in T$ then the OU process is given by:

$$X_t = \sigma \int_{-\infty}^{t} e^{-\frac{r}{\ell}} dW_s, t \in \mathbb{R} \tag{2.9}$$

As for the kernel specification note that for the Wiener integral $\int f dW$ we have that the linear map $L(f) = \int f dW$ is an isometry from the collection of functions in $L^2(\mathbb{R})$ into $L^2(p)$ with p the measure on $(\Omega, \mathcal{B}, p)$, that is for functions $f, g \in L^2(\mathbb{R})$ we have:

$$\mathbb{E}\left[\int f dW\right]\left[\int g dW\right] = \int_{\mathbb{R}} f(x) g(x) dx \tag{2.10}$$

Therefore we have that the covariance function is:

$$k(t,s) = \mathbb{E}[X_t X_s] = \sigma^2 \int_{\infty}^{\min\{t,s\}} e^{-\frac{t-u}{\ell}} e^{-\frac{s-u}{\ell}} du = \frac{\sigma^2 \ell}{2} \exp(-\frac{r}{\ell}) \tag{2.11}$$

We then have a prior $f(t) = X_t$, which can be viewed as a measure on the space $C(T)$ of continuous functions on T with their regularity depending on the sample paths of the OU process. Intuitively each time we roll the dice we get a sample path, and any combination of points on each path is jointly Gaussian. Suppose that we fix σ, ℓ, then under the prior for a finite collection of points $\mathbf{t} = \{t_i\}_{1=1}^N$ we have that:

$$\mathbf{f}|\mathbf{t} = \{f(t_i)\}_{i=1}^N|\mathbf{t} \sim \mathcal{N}\left(\{f(t_i)\}_{i=1}^N|\mathbf{0}, \begin{bmatrix} k(t_1,t_1) & \dots & k(t_1,t_N) \\ \vdots & \ddots & \vdots \\ k(t_N,t_1) & \dots & k(t_N,t_N) \end{bmatrix}\right) \tag{2.12}$$

We denote the invertible matrix of covariances at points (t_i, t_j) as $\mathbf{K}(\mathbf{t})$. Then by Bayes theorem:

$$\mathbb{P}(\mathbf{f}|\mathbf{y}_N, \mathbf{t}) = \frac{L_{\sigma_\epsilon^2}(\mathbf{y}_N|\mathbf{f})\mathbb{P}(\mathbf{f}|\mathbf{t})}{\mathbb{P}(\mathbf{y}_N|\mathbf{t})} \propto \mathcal{N}(\mathbf{y}_N|\mathbf{f}, \sigma_\epsilon^2 \mathbf{I}_N)\mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}(\mathbf{t})) \tag{2.13}$$

Note here that the family of possible posteriors is dominated. In this special case the posterior has an analytic expression that can easily be found, using the regression lemma as in the book of Durbin and Koopman (2012) (p77), to be:

$$\mathbb{P}(\mathbf{f}|\mathbf{y}_N, \mathbf{t}) \equiv \mathcal{N}(\mathbf{f}|\mathbf{K}(\mathbf{t})(\mathbf{K}(\mathbf{t}) + \sigma_\epsilon^2 \mathbf{I}_N)^{-1}\mathbf{y}_N, \mathbf{K}(\mathbf{t}) - \mathbf{K}(\mathbf{t})(\mathbf{K}(\mathbf{t}) + \sigma_\epsilon^2 \mathbf{I}_N)^{-1}\mathbf{K}(\mathbf{t})) \tag{2.14}$$

Given the joint distribution:

$$\begin{bmatrix} \mathbf{f}|\mathbf{t} \\ \mathbf{y}_N \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{t}) & \mathbf{K}(\mathbf{t}) \\ \mathbf{K}(\mathbf{t}) & \mathbf{K}(\mathbf{t}) + \sigma_\epsilon^2 \mathbf{I}_N \end{bmatrix}\right) \tag{2.15}$$

We may be interested in the distribution over the function value at some point t_* for which we have no data. Suppose we are interested in extrapolation at some point for some $* > N$ then we can find the predictive distribution of $f(t_*)$ to be:

$$\mathbb{P}(f(t_*)|\mathbf{y}_N, \mathbf{t}, t_*) = \int_{\mathbb{R}^N} \mathbb{P}(f(t_*), \mathbf{f}|\mathbf{y}_N, \mathbf{t}, t_*) d\mathbf{f} = \int_{\mathbb{R}^N} \mathbb{P}(f(t_*)|\mathbf{f}, \mathbf{t}, t_*)\mathbb{P}(\mathbf{f}|\mathbf{y}_N, \mathbf{t}) d\mathbf{f} \tag{2.16}$$

Which again, in this spacial case, can be found analytically to be:

$$\mathbb{P}(f(t_*)|\mathbf{y}_N, \mathbf{t}, t_*) \equiv \mathcal{N}(f(t_*)|\mathbf{K}(t_*, \mathbf{t})(\mathbf{K}(\mathbf{t}) + \sigma_\epsilon^2 \mathbf{I}_N)^{-1}\mathbf{y}_N, \tag{2.17a}$$

$$\mathbf{K}(t_*, t_*) - \mathbf{K}(t_*, \mathbf{t})(\mathbf{K}(\mathbf{t}) + \sigma_\epsilon^2 \mathbf{I}_N)^{-1}\mathbf{K}(\mathbf{t}, t_*)) \tag{2.17b}$$

Given the joint distribution:

$$\begin{bmatrix} f(t_*) \\ \mathbf{y}_N \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(t_*, t_*) & \mathbf{K}(t_*, \mathbf{t}) \\ \mathbf{K}(\mathbf{t}, t_*) & \mathbf{K}(\mathbf{t}) + \sigma_\epsilon^2 \mathbf{I}_N \end{bmatrix}\right) \tag{2.18}$$

Intuitively what we have done is obtain a posterior over functions that contain only those sample paths that are in accordance with the observations. Furthermore when we integrate

out the function values, we average over all the possible functions under the posterior, which is one view of why this approach is less prone to over-fitting.

In this example two important kernel parameters are present, namely the length-scale ℓ and signal variance σ^2. These parameters are often of importance in other GP kernels as well. The length-scale intuitively can be seen as controlling the amount of covariance between two points. Small ℓ leads to consecutive points being able to differ largely from each other, whereas a large ℓ results in more dependence and thus more smooth looking functions. Furthermore if we were to extrapolate points $f(t_*)$ for $* > N$ the length-scale influence how far away from the point t_N this is still reliable. The signal variance determines the amount of variation from the mean possible. These alongside the error variance in the regression can be learnt by placing priors on the hyper-parameters (in this case $(\ell, \sigma^2, \sigma_\epsilon^2)$) and integrating them out, referred to as hierarchical Bayes. Alternatively we can maximize the marginal likelihood (evidence) $\mathbb{P}(\mathbf{y}_N|\mathbf{t})$ as a function of the hyper-parameters, referred to as empirical Bayes/Maximum marginal likelihood/Type II MLE. In this example the empirical Bayes procedure is the most appealing one, because the marginal likelihood is analytically available.

Note that the mean of the posterior and the predictive distribution are determined by the data and the kernel, even though the mean function is identically zero everywhere. Furthermore note that, although in this special case these expressions are analytically available, the evaluation of the posterior and predictive distributions involve inverting a potentially large matrix. This is inversion is the most important hurdle for using this method, which has complexion $\mathcal{O}(N^3)$. Many solution to this complexity problem have been proposed such as structure exploiting methods (for Toeplitz and Kronecker structured covariance matrices), inducing point methods or variational methods. For example Hartikainen and Särkkä (2010) propose, for certain kernels, converting the model in this example to a linear Gaussian state-space model and then applying the Kalman Filter. In effect through an approximation of the covariance function, their method uses the Kalman filter to convert the batch inversion problem to an iterative one, which has $\mathcal{O}(N)$ complexity. However a drawback to the model in this example, for which their method is available, is that it is unable to capture non-linear dynamics, for that we would have to reformulate the model to:

$$y_{t+1} = f(y_t) + \epsilon_t \tag{2.19}$$

Or more generally:

$$y_{t+1} = f(y_t, \ldots, y_{t-p}) + \epsilon_t \tag{2.20}$$

And even this formulation has its drawbacks, particularly in terms of uncertainty quantification, as is described in the introduction. As a final note, as soon as the likelihood is no longer Gaussian intractability's of the posterior arise.

2.4 Connections to Neural Networks and Linear Sieves

Here we look at the connection between classical regression, Bayesian regression, and Neural networks in a simple way. The main question I had was how can we relate these methods to others in various fields, such as neural networks in machine learning or non parametric methods in econometrics such as sieve estimators. We keep the answers found and the notation brief. Let us for simplicity consider the univariate basis function expansion:

$$f_n(x_t) = \sum_{i=0}^{n} w_i \phi_i(x_t) \tag{2.21}$$

Where $\phi(x_t) : \mathcal{X} \to \mathcal{H}$ is often referred to as a feature map or a basis function that maps from the sample space to some $\mathbb{R}$-Hilbert space. For linearity, ϕ never depends on the weight vector $\mathbf{w}$. For example in the case of a polynomial basis (linear polynomial sieve in econometrics) we have that $\phi_i(x_t) = x_t^i$ and polynomials lie dense in the space of continuous functions defined on real compact intervals, but others are also possible and often more desirable. Fourier series, splines, wavelets are others that for example in the econometrics literature fall in the Hölder class of linear sieves (X. Chen 2007). In the neural networks literature a classical example is the 1 layer radial basis function (RBF) network where the ϕ_i often looks like $\phi_i(x_t) = \exp\left(-\frac{(x_t - c_i)^2}{\beta}\right)$. We consider the Bayesian (Gaussian) treatment where we impose the prior assumption of normality on the weights w_i, that is we let $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_n)$. By going through the basic steps exactly as with the GP regression example we would have that the this would imply some GP (Carl Edward Rasmussen 2006) (p84) with the kernel:

$$k(x_t, x_s) = \sigma^2 \sum_{i=1}^{n} \phi_i(x_t)\phi_i(x_s) \tag{2.22}$$

The question is what happens to the Gaussian Process when we let $n \to \infty$ and it turns out that this depends on the form of the basis function. MacKay (1998) showed that if the basis function is that of a RBF network then in the limit as $n \to \infty$ this corresponds to a GP with the Squared Exponential kernel:

$$k(x_t, x_s) = \exp\left(-\frac{(bx_t - bx_s)^2}{2\ell^2}\right) \tag{2.23}$$

Note that this result also holds for $\mathbf{x}_t \in \mathbb{R}^d$. As another example Wahba (1990) showed that if the basis function has the form of linear splines, then for $n \to \infty$ this corresponds to a GP with kernel:

$$k(x_t, x_s) = 1 + \alpha - 2\alpha|x_t - x_s| \tag{2.24}$$

Wahba (1990) gives similar results for cubic splines, but what is most important here is the connection between Sieves and Neural networks, and GP regression, which can be made through basis functions.

In short this means that certain kernels correspond to an infinite amount of basis functions, and of course this only means that an infinite amount of these are available and the model puts certain posterior mass on the weights to control the truncation as is appropriate according to the data that is available. From this point of view also we can look at GP based models as truly non-parametric.

2.5 Reproducing Kernel Hilbert Spaces and Regularization

There is another connection between the frequentist non-parametric methods and the Bayesian ones, which is in the way prior assumptions are incorporated in models. Finding the function f that drives the DGP from a finite set of data-points is not a well posed problem, and over-fitting is one of the known difficulties in these settings. Certain assumptions on the structure of f are needed, which in the GP case are imposed through the structure of the covariance function. In Bayesian non-parameterics the regularity of the prior sample paths, that is our prior assumptions on the well-behavedness of the DGP function, is of great importance (Knapik et al. 2011). In frequentist terms similar assumptions can be made through regularization. To answer the question of how these relate we closely follow Carl Edward Rasmussen (2006). Suppose we are interested in the minimizer of the functional:

$$F[f] = \frac{\lambda}{2}||f||^2_{\mathcal{H}} + L(\mathbf{y}_N, \mathbf{f}) \tag{2.25}$$

Where $L(\mathbf{y}_N, \mathbf{f})$ is some loss function, for example $L(\mathbf{y}_N, \mathbf{f}) = \frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_{t_i}) - y_{t_i})^2$. The first term contains the norm of f in a Reproducing Kernel Hilbert Space (RKHS) and is referred to as the regularizer.

Ch2. Definition 5 (Reproducing Kernel Hilbert Space)**.** *Let $\mathcal{H}$ be a Hilbert space of real functions f defined on an index set T. Then $\mathcal{H}$ is called a Reproducing Kernel Hilbert Space endowed with an inner product $< \cdot, \cdot >_{\mathcal{H}}$, and by definition the norm $|| \cdot ||_{\mathcal{H}} = \sqrt{< \cdot, \cdot >}_{\mathcal{H}}$, if there exists a function $K : T \times T \to \mathbb{R}$ such that:*

1 :) *for every $\mathbf{x}_t$, $k(\mathbf{x}_t, \mathbf{x}_s)$ as a function of $\mathbf{x}_s$ belongs to $\mathcal{H}$*

2 :) *k has the reproducing property: $f(\mathbf{x}_t) =< f(\cdot), k(\mathbf{x}_t, \cdot) >_{\mathcal{H}}$ for all $f \in \mathcal{H}$*

We consider the valid PSD kernel $k : T \times T \to \mathbb{R}$ with an eigenfunction[1] expansion $k(\mathbf{x}_t, \mathbf{y}_t) = \sum_{i=1}^{p} \lambda_i \phi_i(\mathbf{x}_t)\phi_i(\mathbf{x}_s)$ relative to a measure μ (see Merecer's Theorem below).

[1] In $\int k(\mathbf{x}_t, \mathbf{x}_s)\phi(\mathbf{x}_t)\mu(d\mathbf{x}_t) = \lambda\phi(\mathbf{x}_s)$, ϕ is the eigenfunction of k and λ its eigenvalue w.r.t. μ .

Ch2. Theorem 4 (Mercer's Thm.). *Let $(\mathcal{X}, \mathcal{B}, \mu_x)$ be a measure space with $\mu(\mathcal{X}) < \infty$ and $k \in L^\infty(\mathcal{X}^2)$ a kernel such that its corresponding operator $T_\mathcal{K} : L^2(\mathcal{X}) \to L^2(\mathcal{X})$ has the property:*

$$\int_{\mathcal{X}\times\mathcal{X}} k(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_t) f(\mathbf{x}_s) \mu_x(d\mathbf{x}_t) \mu_x(d\mathbf{x}_s) \geq 0 \text{ for all } f \in L^2(\mathcal{X})$$

and is symmetric. Then there exists a basis of the eigenfunctions $\phi_j \in L^2(\mathcal{X})$ of $T_\mathcal{K}$ associated with the eigenvalues $\lambda_j \geq 0$ with the property that $< \phi_k, \phi_l >_2^{\frac{1}{2}} = \begin{cases} 1 & k = l \\ 0 & k \neq l \end{cases}$ (orthonormal). Then:

$$\mathbf{1 :)}\ \sum_{j=0}^{\infty} |\phi_j| < \infty$$

$$\mathbf{2 :)}\ k(\mathbf{x}_t, \mathbf{x}_s) = \sum_{j=0}^{\infty} \lambda_j \phi_j(\mathbf{x}_t) \phi_j(\mathbf{x}_s)$$

With absolute and uniform convergence.

Suppose we have the Hilbert space consisting of linear combinations of the eigenfunctions, that is $f(\mathbf{x}_t) = \sum_{i=1}^p f_i \phi_i(\mathbf{x}_t)$ with $\sum_{i=1}^p \frac{f_i^2}{\lambda_i} < \infty$, endowed with the inner product

$$\sum_{i=1}^{p} \frac{f_i g_i}{\lambda_i} \tag{2.27}$$

Here it is important to note that $||f||_\mathcal{H}^2 =< f, f >_\mathcal{H} = \sum_{i=1}^p \frac{f_i^2}{\lambda_i}$ and the smoothness on the function is imposed by the fact that the coefficients $\{f_i\}_i$ must decay fast enough so that the norm is finite. In the case of of a basis expansion for example the $||f||_\mathcal{H}^2$ term penalises the coefficients f_i of $f(\mathbf{x}_t) = \sum_{i=1}^p f_i \phi_i(\mathbf{x}_t)$. For example when performing penalized maximum likelihood the regularizer $||f||_\mathcal{H}^2$ encodes the smoothness assumptions on the function and the loss function the likelihood fit.

By verifying the two conditions of the RKHS definition it can be shown that the Hilbert space comprising of linear combinations of the eigenfunctions is a RKHS of the kernel $k : T \times T \to \mathbb{R}$. Then by the Moore-Aronzajn theorem, which says that for every valid kernel there exists a unique RKHS (Aronszajn 1950), it follows that this Hilbert space is the one belonging to the kernel $k : T \times T \to \mathbb{R}$. By drawing the coefficients f_i from a Gaussian distribution, with its variance corresponding to the eigenvalues of the kernel, and letting $p \to \infty$ a Gaussian process prior can be generated of which its sample paths are not in the RKHS, but the posterior mean is (Carl Edward Rasmussen 2006). In fact if we consider Tikhonov regularization, that is $L(\mathbf{y}_N, \mathbf{f}) = \sum_{i=1}^N (< \{f_j\}_{j=1}^p, \{\phi_j(\mathbf{x}_{t_i})\}_{j=1}^p >_\mathcal{H} - y_{t_i})^2$ in Equation 2.25, for the model with Gaussian errors in Equation 2.8 then the minimizer of the functional in the RKHS is exactly the posterior mean in the regression section.

More generally, noting that from the regularization perspective we consider the solution $\hat{f} = \text{argmin}_f F[f]$, we can equivalently consider the maximization of:

$$\exp(-F[f]) = \exp\left(-\frac{\lambda}{2}||Pf||^2\right) \times \exp(-L(\mathbf{y}_N, \mathbf{f})) \tag{2.28}$$

Where P is constraint operator (such as the differential operator for example), and $||\cdot||$ the L^2 norm on the function space. Then the first term can be viewed as the GP prior and the second is proportional to the likelihood, thus the expression is proportional to the posterior in the example GP regression section. And maximizing this expression gives us the mode of the posterior (depending on the convexity of the likelihood possibly only a local maximum). This maximum of the posterior is referred to as the Maximum A Posteriori (MAP) solution. The optimization problem of course depends on whether for $L(\mathbf{y}_N, \mathbf{f})$ we can exploit either conjugacy or convexity, or if we are dealing with a multimodal search space.

Thus we can see that in both cases we are imposing structure on the function, but in a different way, in this case placing Gaussian priors an the expansion amounts to L^2 regularization. There are of course other possibilities such as placing Laplace priors on the weights which would coincide with L^1 regularization, but the Gaussian priors have a nice connection to GP's. In chapter 3 we approximate the GP based state-space model with a very specific stochastic basis-function expansion such that, given its kernel spectral representation, the expansion converges to a full GP state-space model for any kernel. This regularization perspective then becomes important.

2.6 Theory of stationary covariance functions

Stationary covariance functions have some desirable properties that are exploited throughout this thesis. In this section I give the needed definitions and theorems we use through the thesis, furthermore the two central kernels we use are introduced here.

Ch2. Definition 6 (Stationary Kernel)**.** *Let $\mathcal{X} \subset \mathbb{R}^d, T \subset \mathbb{R}^d$ and $\mathbf{x}_t, \mathbf{x}_s \in \mathcal{X}$ the kernel $k : T \times T \to \mathbb{R}$ such that $k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_k) \equiv k([\mathbf{x}_t - \mathbf{x}_s]; \boldsymbol{\theta}_k)$ is called (weakly) stationary.*

Stationarity of the covariance function implies that the distribution is invariant under rigid translations of the input, in symbolic terms: $\mathbb{E}(\mathbf{x}_t) = \mathbb{E}(\mathbf{x}_t + h)$ and $\mathbb{C}ov(\mathbf{x}_t + h, \mathbf{x}_s + h) = \mathbb{C}ov(\mathbf{x}_t, \mathbf{x}_s)$ for all $h \in \mathbb{Z}$.

Ch2. Definition 7 (Isotropy)**.** *Let $\mathcal{X} \subset \mathbb{R}^d, T \subset \mathbb{R}^d$ and $\mathbf{x}_t, \mathbf{x}_s \in \mathcal{X}$ the kernel $k : T \times T \to \mathbb{R}$ such that $k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_k) \equiv k(||\mathbf{x}_t - \mathbf{x}_s||; \boldsymbol{\theta})$ is called Isotropic. ($||\cdot||$ is the Euclidian norm).*

Isotropy implies invariance under rotations. For example suppose R is a rotation matrix then for an isotropic kernel we have $k(R\mathbf{x}_t, R\mathbf{x}_s; \boldsymbol{\theta}_k) = k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_k)$.

Ch2. Theorem 5. *(Bochner's theorem) A complex-valued function k on $\mathcal{X} \in \mathbb{R}^d$ is the covariance function of a weakly stationary mean square continuous complex-valued random process on $\mathbb{R}^D$ if and only if it can be represented as:*

$$k(\mathbf{r}) = (2\pi)^{-d} \int_{\mathbb{R}^{\mathbb{D}}} e^{i\omega'\mathbf{r}} \mu(d\omega), \qquad \mathbf{r} = \mathbf{x}_t - \mathbf{x}_s \tag{2.29}$$

for some positive σ-finite measure μ.

The transform in 2.29 is just a version the inverse Fourier-Stieltjes transform ($\omega = 2\pi x$). The measure μ is called the spectral measure and if it has a density $S(\omega)$ (i.e. if it is dominated) we refer to $S(\omega)$ as the spectral density corresponding to the kernel $k(\cdot)$. Note from Bochner's theorem the isometry between the linear span of the elements $\{X_t : t \in \mathbb{R}\}$ of the random process $\mathbf{X}$ on $(\Omega, \mathcal{B}, p)$ in $L^2(p)$ and closure of the linear span of functions $\{e^{i\omega' t} : t \in T \subset R^d\}$ in $L^2(\mu)$, which is called the spectral isometry. By the Wiener-Khintchine theorem when the spectral density corresponding to $k(\cdot)$ exists the kernel and the spectral density are Fourier duals of each-other (Carl Edward Rasmussen 2006). We then get the following transforms and inverse-transforms:

$$k(\mathbf{r}) = (2\pi)^{-d} \int S(\omega) e^{i\omega'\mathbf{r}} d\omega \tag{2.30}$$

$$S(\omega) = \int k(\mathbf{r}) e^{-i\omega'\mathbf{r}} d\mathbf{r} \tag{2.31}$$

In the isotropic case the spectral density $S(\omega)$ is also isotropic (invariant under rotations) and depends only on the length of ω (Adler 1981).

In this thesis, two kernels are implemented and the first one is the Matern class kernel (Stein 1999):

$$k^{mat}(\mathbf{r}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} r \mathbf{r}}{\ell} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}\mathbf{r}}{\ell} \right) \tag{2.32}$$

Where Γ is Gamma function, K_ν the modified Bessel function, $\nu > 0$ the degrees of freedom (discussed in more detail in ch.3), and ℓ the length-scale as discussed in the regression example. The spectrum of this kernel is discussed in more detail in chapter 3. Two things to note is that as $\nu \to \infty$ this becomes the squared-exponential/expatiated-quadratic kernel discussed before and for $\nu = \frac{1}{2}$ we obtain the OU covariance function from the Matern class. As such it is quite a flexible kernel. Especially in terms of asymptotic behavior in frequentist terms this flexibility is of importance as is elucidated later on.

Furthermore we employ the spectral mixture kernel of A. G. Wilson and Adams (2013). For the specific purposes of this thesis we are not interested in its kernel as much as its spectral density which is just a mixture of symmetrized Gaussians. The way I set up the

kernel and its corresponding spectral density in product spaces (see chapter 3) means we only need to consider the one-dimensional mixture spectral density:

$$\Phi^i(\omega|\mu_i, \sigma_i^2) = \mathcal{N}(\omega|\mu_i, \sigma_i^2) + \mathcal{N}(-\omega|\mu_i, \sigma_i^2) \tag{2.33a}$$

$$S^{mix}(\omega) = \sum_{i=1}^{q} a_i \Phi^i(\omega|\mu_i, \sigma_i^2) \tag{2.33b}$$

Where the mixture weights a_i are related to the signal variance of the data. If we we look at the covariance function given in A. G. Wilson and Adams (2013) corresponding to the spectral density given above, then we see that the variance $k(\mathbf{x}_t, \mathbf{x}_t)$ is the sum of the component weights. The inverse component variances $\frac{1}{\sigma_i^2}$ are the length-scales and the inverse locations $\frac{1}{\mu_i}$ are the periods of the components (note the periodic behavior here).

The justification for this specific covariance function comes from the fact that mixtures of Gaussians lie dense in the set of all distribution functions (Kostantinos 2000). Therefore by using a finite Mixture of Gaussians, given enough mixture components, any spectral distribution can be approximated arbitrarily well. By the Fourier duality of stationary covariance functions then, so can any covariance function be approximated. A. G. Wilson and Adams (2013) show how for the number of components $q \leq 10$ covariances such as the periodic kernel or the rational quadratic one (Carl Edward Rasmussen 2006) are well approximated. If we recall that the space of functions that can be generated as well as the prior probability measure induced on that space are determined by the kernel, then we in effect allow for a lot generality in terms of our prior beliefs by using the Gaussian mixture.

2.7 Asymptotics

Given what I have learnt from the advanced econometrics classes, consistency results are not to be taken lightly, therefore a similar question arises in terms of Bayesian inference, which is still a very young study as compared to asymptotics in classical statistics. Usually in a parametric Bayesian setting we say that the model is well specified if the true underlying distribution lies in the support of the prior, and under mild conditions the posterior often concentrates its mass around this true model. In the infinite dimensional case the situation is more difficult, and although there was no time to study the theory and literature regarding the contraction rates of the posterior for Gaussian Process prior models (see for example (A. W. van der Vaart and J. H. van Zanten 2008)), we do want to understand how the choice of the GP prior influences the credible regions we use to quantify uncertainty. Arguably the coverage of the credible regions are more important than contraction rates of the posterior. These regions are often the area between two quantiles of the posterior. For example in our application section we use up to 3 standard deviations from the mean of the posterior as the regions that might contain some true model of which we assume it exists. Thus we are interested in coverage from a frequentist point of view as the data goes to infinity.

In the literature there is evidence that assuming a data generating process (DGP), for non-parametric Bayesian models, regions of the posterior mass at some α-level taken as credible sets do not necessarily contain the DGP with a probability at least as large as α (Freedman 1999; Johnstone 2010). Knapik et al. (2011) conclude that these credible regions typically show good coverage in frequentist terms when the prior process has a lower degree of regularity than the true DGP. On the other hand if this is not the case then the coverage can in fact be very bad, thus it would be odd to just assume that when we employ Bayesian non-parametric models we are good to go in terms of uncertainty quantification.

Therefore at least heuristically it is a good idea to choose the Prior such that its regularity can be controlled, and it would be even better if the regularity could be adapted to the data. In that sense the most used Gaussian Process prior in machine learning based on the squared exponential kernel, with a spectral density that has finite moments of any order and hence has the process has infinitely differentiable sample paths is out of the question, unless we have information on the regularity of the DGP that justifies this choice. Of course we don't know the regularity of the true DGP.

This regularity of the prior is determined by the tails of the spectral density in Fourier space and we explore this further in chapter 3 where, at least for the Matern prior process, we show how we can use the parameters of the spectral density corresponding to the convolution kernel to let it be adaptable to the data.

When we show that the convolution kernel we employ has the ability to learn the regularity of the prior from data, that is in the case the approximate model converges to the full GP-model. And the approximation brings along its own source of error. Note also that the context in which we employ the GP priors, state-space models, is different than the ones often encountered in the literature which is the signal-in-white-noise model (Knapik et al. 2011; Szabó et al. 2015), but the regularity is a good starting point from which to look at the problem. For the Matern process with the convolution kernel we learn the regularity parameter with hierarchical Bayesian learning, that is we place priors on the parameters of the spectral density and proceed to learn the posterior in a Bayesian way. The empirical Bayes procedure, again in the context of the signal-in-white-noise model, has been shown to have the posterior contract, at a near optimal rate, around the true DGP for a range of true regularities. Of course this does not imply good coverage of the credible sets in the empirical Bayes setting. For the hierarchical Bayes procedure I was unable to find such a result, but I think its still a good idea to let the regularity be adaptable to the data in this procedure. Szabó et al. (2015) show that there are always true models that are not within the credible region of the posterior asymptotically, but these are few and when taken out the credible regions have good frequentist coverage.

So in this thesis the best we have is but a heuristic, but one that sits on sounds theory and for future reference we can delve into the asymptotic behavior of the models we propose. In the simulation analysis of chapter 6 the finite sample results are promising.

2.8 GP's in state-space modeling

Given that we have a non-parametric and probabilistic way of describing functions, the next logical step is to implement GP's in a state-space model. However in a state-space model the regressors are no longer directly available and we are in what is referred to as a reinforcement learning setting (because of the feedback that can be introduced into the system). The question then is how do GP priors fit in state-space modelling and how are these models estimated. Furthermore we analyse the prior GP-SSM distribution.

State Space Model with a GP in the transition equation One way of setting up a non-parametric SSM is to proceed in a Bayesian manner and set up a prior model for it with a GP prior over the state function, and let the measurement equation be parametric.

$$f(\mathbf{x}_t) \sim \mathcal{GP}(m(\mathbf{x}_t; \boldsymbol{\theta}_f), k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f)) \tag{2.34a}$$

$$f(\mathbf{x}_t) := \mathbf{f}_{t+1} \qquad \mathbf{x}_0 \sim \mathbb{P}_0(\mathbf{x}_0) \tag{2.34b}$$

$$\mathbf{x}_{t+1}|\mathbf{f}_{t+1} \sim \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{f}_{t+1}, \mathbf{Q}) \qquad \mathbf{x}_{t+1} = f(\mathbf{x}_t) + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\eta}_t|0, \mathbf{Q}) \tag{2.34c}$$

$$\mathbf{y}_t|\mathbf{x}_t \sim \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t) \tag{2.34d}$$

Note that in this model the structure of f is time invariant. The chain structure of the inputs and outputs of the latent function f of the states $\mathbf{x}_t \to f(\mathbf{x}_t) \to \mathbf{x}_{t+1}$ much like a recurrent neural network makes estimation of this model difficult. We can either let $\mathbf{x}_t \in \mathbb{R}^d$ be a vector of independent states, which means that we have a diagonal $k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f) \in \mathbb{R}^{d\times d}$ where $k_l(\mathbf{x}_{i,t}, \mathbf{x}_{j,s}; \boldsymbol{\theta}_f) = 0 \in \mathbb{R}$. Alternatively we can allow correlations between the states in various ways, for example by allowing for separate kernels on the outputs of the GP like Rakitsch et al. (2013). As is clarified in chapter 3, covariance between the states is allowed for through the basis function expansion which incorporates the spectral density of the kernel.

If we lack the necessary prior knowledge we can let $m(\mathbf{x}_t; \boldsymbol{\theta}_f) = 0 \in \mathbb{R}^d$, which is the approach we take in this thesis. On the other hand if we have more prior knowledge then this can be Incorporated into the functional form specification, we can for example let the state equation be of the form $f + g$, where g is user specified and f has a GP prior. As such, if there is already a decent parametric model in place we can replace the parts of the functional form we are most uncertain about with a GP based function. This insight reconciles the data-driven philosophy in machine learning literature with the model-driven one in econometrics. There is no need to throw away expert knowledge when employing machine learning or non-parametric methods and there is no need to throw away undiscovered insights from the data when employing a more model-driven approach. In this thesis the goal is to verify the performance of the non-parametric approach as such I impose as little structure as possible, but this is certainly not needed.

The simple but infinite Bayesian network implied by prior model in the equations in 2.34 allows us to generate the samples $\mathbf{x}_{0:T}|\mathbf{f}_{1:T}$ and $\mathbf{y}_{1:T}|\mathbf{x}_{0:T}$ from it in a chain like manner and we can get the joint distribution as follows:

$$\mathbb{P}(\mathbf{f}_{1:T,}\mathbf{x}_{0:T},\mathbf{y}_{1:t}) = \mathbb{P}_0(\mathbf{x}_0)\prod_{t=1}^{T}\mathbb{P}(\mathbf{f}_t|\mathbf{f}_{1:t-1},\mathbf{x}_{0:t-1})\mathbb{P}(\mathbf{x}_t|\mathbf{f}_t)\mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t) \tag{2.35}$$

Again it is good to note that two things are omitted here in order to simplify the notation: explicitly conditioning on $\boldsymbol{\theta}_f$ as well as on the inputs. For the inputs we can imagine $\mathbf{x}_t$ to be defined as $(\mathbf{x}_t, \mathbf{u}_t)$. Note also that samples from this model are prior samples, that is state draws from this model are not conditioned on any measurements $\mathbf{y}_{1:t}$ we might have.

Let $m(\mathbf{x}_t;\boldsymbol{\theta}) = 0$, as in the regression example we can obtain the conditional distribution (Frigola, Y. Chen, et al. 2014b) on the function values at time t given previous information (again this follows directly from regression lemma in Durbin and Koopman (2012), where $y = \mathbf{f}_{1:t-1}$) :

$$\mathbf{f}_t|\mathbf{f}_{1:t-1},\mathbf{x}_{0:t-1} \sim \mathcal{N}(\mathbf{f}_t|\mathbf{K}_{t-1,0:t-2}\mathbf{K}^{-1}_{0:t-2,0:t-2}\mathbf{f}_{1:t-1}, \mathbf{K}_{0:t-1,0:t-1} - \mathbf{K}_{t-1,0:t-2}\mathbf{K}^{-1}_{0:t-2,0:t-2}\mathbf{K}'_{t-1,0:t-2}) \tag{2.36}$$

In this section, for the sake of understanding the GP-SSM, we adopt the notation of Frigola, Y. Chen, et al. (2014b) where

$$\mathbf{K}_{i:j,k:l} := \begin{bmatrix} k(x_i,x_k;\boldsymbol{\theta}_f) & \cdots & k(x_i,x_l;\boldsymbol{\theta}_f) \\ \vdots & \ddots & \vdots \\ k(x_j,x_k;\boldsymbol{\theta}_f) & \cdots & k(x_j,x_l;\boldsymbol{\theta}_f) \end{bmatrix} \tag{2.37}$$

If we look at the first couple of iterations of drawing functions and states following equation 2.36.

$$\mathbf{x}_0 \sim \mathbb{P}_0(\mathbf{x}_0) \tag{2.38a}$$

$$\mathbf{f}(\mathbf{x}_0) \equiv \mathbf{f}_1|\mathbf{x}_0 \sim \mathcal{N}(\mathbf{f}_t|0, k(\mathbf{x}_0,\mathbf{x}_0;\boldsymbol{\theta}_f)) \tag{2.38b}$$

$$\mathbf{x}_1|\mathbf{f}_1 \sim \mathcal{N}(\mathbf{x}_1|\mathbf{f}_1,\mathbf{Q}) \tag{2.38c}$$

$$\mathbf{f}(\mathbf{x}_1) \equiv \mathbf{f}_2|\mathbf{f}_1,\mathbf{x}_{0:1} \sim (\mathbf{f}_2|A\mathbf{f}_1, k(\mathbf{x}_1,\mathbf{x}_1;\boldsymbol{\theta}_f) - AB) \tag{2.38d}$$

Where A, B are as in equation 2.36. We can write these as follows:

$$\mathbf{f}_1|\mathbf{x}_0 = 0 + \epsilon_{f_1} \qquad \epsilon_{f_1} \sim \mathcal{N}(\epsilon_{f_1}|0, k(\mathbf{x}_0,\mathbf{x}_0;\boldsymbol{\theta}_f)) \tag{2.39a}$$

$$\mathbf{f}_2|\mathbf{f}_1,\mathbf{x}_{0:1} = A\mathbf{f}_1 + \epsilon_{f_2} \qquad \epsilon_{f_2} \sim \mathcal{N}(\epsilon_{f_2}|0, k(\mathbf{x}_1,\mathbf{x}_1;\boldsymbol{\theta}_f) - AB) \tag{2.39b}$$

Where the epsilons are conditionally independent. Then the moments with respect to the joint distribution are given by:

$$\mathbb{E}_{\mathbf{f}_{1:2}}[\mathbf{f}_2|\mathbf{f}_1, \mathbf{x}_{0:1}] = A\mathbb{E}_{\mathbf{f}_{1:2}}[\epsilon_{f_1}] + \mathbb{E}_{\mathbf{f}_{1:2}}[\epsilon_{f_2}] = 0 \tag{2.40a}$$

$$\mathbb{V}ar_{\mathbf{f}_{1:2}}[\mathbf{f}_2|\mathbf{f}_1, \mathbf{x}_{0:1}] = A\mathbb{V}ar_{\mathbf{f}_{1:2}}[\epsilon_{f_1}]A' + \mathbb{V}ar_{\mathbf{f}_{1:2}}[\epsilon_{f_2}] \tag{2.40b}$$

$$= k(\mathbf{x}_1, \mathbf{x}_0; \boldsymbol{\theta}_f)k(\mathbf{x}_0, \mathbf{x}_0; \boldsymbol{\theta}_f)^{-1}k(\mathbf{x}_0, \mathbf{x}_0; \boldsymbol{\theta}_f)k(\mathbf{x}_0, \mathbf{x}_0; \boldsymbol{\theta}_f)^{-1}k(\mathbf{x}_1, \mathbf{x}_0; \boldsymbol{\theta}_f) \tag{2.40c}$$

$$+ k(\mathbf{x}_1, \mathbf{x}_1; \boldsymbol{\theta}_f) - k(\mathbf{x}_1, \mathbf{x}_0; \boldsymbol{\theta}_f)k(\mathbf{x}_0, \mathbf{x}_0; \boldsymbol{\theta}_f)^{-1}k(\mathbf{x}_1, \mathbf{x}_0; \boldsymbol{\theta}_f) \tag{2.40d}$$

$$= k(\mathbf{x}_1, \mathbf{x}_1; \boldsymbol{\theta}_f) \tag{2.40e}$$

$$\mathbb{C}ov_{\mathbf{f}_{1:2}}[\mathbf{f}_2|\mathbf{f}_1, \mathbf{x}_{0:1}, \mathbf{f}_1|\mathbf{x}_0] = A\mathbb{C}ov_{\mathbf{f}_{1:2}}[\epsilon_{f_1}, \epsilon_{f_1}] + \underbrace{\mathbb{C}ov_{\mathbf{f}_{1:2}}[\epsilon_{f_2}, \epsilon_{f_1}]}_{=0} \tag{2.40f}$$

$$= k(\mathbf{x}_1, \mathbf{x}_0; \boldsymbol{\theta}_f)k(\mathbf{x}_0, \mathbf{x}_0; \boldsymbol{\theta}_f)^{-1}k(\mathbf{x}_0, \mathbf{x}_0; \boldsymbol{\theta}_f) \tag{2.40g}$$

$$= k(\mathbf{x}_1, \mathbf{x}_0; \boldsymbol{\theta}_f) \tag{2.40h}$$

Following this logic we can write down the distribution of the conditional likelihood:

$$\widetilde{\mathbb{P}}(\mathbf{f}_{1:T}|\mathbf{x}_{0:T}) := \prod_{t=1}^{T} \mathbb{P}(\mathbf{f}_t|\mathbf{f}_{1:t-1}, \mathbf{x}_{0:t-1}) \equiv \mathcal{N}(\mathbf{f}_{1:T}|\mathbf{0}, \mathbf{K}_{0:T-1,0:T-1}) \tag{2.41}$$

Where we introduce the notation $\widetilde{\mathbb{P}}(\mathbf{f}_{1:T}|\mathbf{x}_{0:T})$ to emphasize that $\widetilde{\mathbb{P}}(\mathbf{f}_{1:T}|\mathbf{x}_{0:T}) \neq \mathbb{P}(\mathbf{f}_{1:T}|\mathbf{x}_{0:T})$, it is just the conditional distribution taken from the joint density in 2.35. Then the prior joint distribution of the model in equation 2.34 is:

$$\mathbb{P}(\mathbf{f}_{1:T}, \mathbf{x}_{0:T}, \mathbf{y}_{1:t}) \equiv \mathbb{P}_0(\mathbf{x}_0)\mathcal{N}(\mathbf{f}_{1:T}|\mathbf{0}, \mathbf{K}_{0:T-1,0:T-1})\mathcal{N}(\mathbf{x}_{1:T}|\mathbf{f}_{1:T}, \mathbb{I}_T \otimes \mathbf{Q}) \prod_{t=1}^{T} \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t) \tag{2.42}$$

More general GP State Space Model We can also place a GP prior on the measurement function:

$$f(\mathbf{x}_t) \sim \mathcal{GP}(m(\mathbf{x}_t; \boldsymbol{\theta}_f), k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f)) \tag{2.43a}$$

$$f(\mathbf{x}_t) := \mathbf{f}_{t+1} \qquad \mathbf{x}_0 \sim \mathbb{P}_0(\mathbf{x}_0) \tag{2.43b}$$

$$\mathbf{x}_{t+1}|\mathbf{f}_{t+1} \sim \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{f}_{t+1}, \mathbf{Q}) \qquad \mathbf{x}_{t+1} = f(\mathbf{x}_t) + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim \mathcal{N}(\boldsymbol{\eta}_t|0, \mathbf{Q}) \tag{2.43c}$$

$$g(\mathbf{x}_t) \sim \mathcal{GP}(m(\mathbf{x}_t; \boldsymbol{\theta}_g), k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_g) \tag{2.43d}$$

$$\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{y}_t|g(\mathbf{x}_t), \mathbf{H}) \qquad \mathbf{y}_t = g(\mathbf{x}_t) + \boldsymbol{\epsilon}_t \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{\epsilon}_t|0, \mathbf{H}) \tag{2.43e}$$

It can be shown that the model in Equation 2.34 and this model are equivalent by redefining the states. Having said that, for the purposes of this thesis it is more appealing to parametrically define the measurement function because usually we have prior information about this function and furthermore in that case we are not burdened with Gaussianity in the measurement equation.

2.8.1 Various inference/learning approaches

Learning boils down to estimating the function values, the hyper-parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \mathbf{Q}, \boldsymbol{\theta}_y)$ and the hidden states. Two common ways of learning are, as in the regression example, empirical Bayes and hierarchical Bayes (also referred to as full Bayesian learning). However in the GP-SSM setting, due to the interdependence between state draws and and function draws, intractability's arise. For example suppose we would want to maximize the marginal likelihood $\mathbb{P}(\mathbf{y}_t|\boldsymbol{\theta}_f)$ with the hyper-parameters as the arguments, then this would involve integrating out the latent function values as well as the states. The latent function values $\mathbf{f}_{1:T}$ can be integrated out and the likelihood $\mathbb{P}(\mathbf{x}_{0:T}) = \int_{\mathcal{X}} \mathbb{P}(\mathbf{x}_{0:T}, \mathbf{f}_{1:T}) d\mathbf{f}_{1:T}$ is analytically available (J. Wang et al. 2005) but it is not Gaussian. The non-Gaussianity of this distribution follows, because the conditional likelihood given in 2.41 includes a covariance matrix obtained by evaluating the kernel of f at locations $\mathbf{x}_{0:T-1}$, and this results in us not being able to analytically integrate the states out to obtain the marginal likelihood. Then we would have to rely on approximation methods. The main ingredients in the hierarchical Bayesian estimation/learning/identification are the posterior distribution over the function values, the hyper-parameters $\boldsymbol{\theta} = (\boldsymbol{\theta}_f, \mathbf{Q}, \boldsymbol{\theta}_y)$ and the hidden states. Although obtaining the posterior over the parameters through integration is analytically intractable for the GP-SSM, given the analytical availability of the distribution over the states with the latent function values integrated out, a sampling algorithm can be used to obtain samples from the smoothing distribution.

Suppose we know $\boldsymbol{\theta}$ and we have the smoothing distribution $\mathbb{P}(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}; \boldsymbol{\theta})$ if we wish to find the distribution on the latent function $f(\cdot)$ at some input point $\mathbf{x}_{\tau-1}$ we can find it in a similar fashion as we did in the case of the regression example (Frigola, Lindsten, Schön, and C. Rasmussen 2013):

$$\mathbb{P}(f(\mathbf{x}_\tau)|\mathbf{x}_\tau, \mathbf{y}_{1:T}, \boldsymbol{\theta}) = \int_{\mathcal{X}^{T+1}} \mathbb{P}(f(\mathbf{x}_\tau)|\mathbf{x}_\tau, \mathbf{x}_{0:T}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}, \boldsymbol{\theta}) d\mathbf{x}_{0:T} \tag{2.44}$$

If we are interested in one-step ahead prediction for example then $\tau = T$. However in practice the smoothing distribution is not readily available and furthermore the hyper-parameters need to be inferred from the data. In Bayesian learning then we would like to draw samples from the joint posterior over the hyper-parameters and the states in order to approximate the integral:

$$\mathbb{P}(f(\mathbf{x}_\tau)|\mathbf{x}_\tau, \mathbf{y}_{1:T}, \boldsymbol{\theta}) = \int \mathbb{P}(f(\mathbf{x}_\tau)|\mathbf{x}_\tau, \mathbf{x}_{0:T}, \boldsymbol{\theta}) \mathbb{P}(\mathbf{x}_{0:T}, \boldsymbol{\theta}|\mathbf{y}_{1:T}) d\mathbf{x}_{0:T} d\boldsymbol{\theta} \tag{2.45}$$

The distribution of $\mathbb{P}(f(\mathbf{x}_\tau)|\mathbf{x}_\tau, \mathbf{x}_{0:T}\boldsymbol{\theta})$ can be easily found through standard GP prediction on $\mathbf{x}_{1:T} = f(\mathbf{x}_{0:T-1}) + \mathbb{I}_T \otimes \mathbf{Q}$ as in the regression example. To draw samples from the joint posterior Frigola, Lindsten, Schön, and C. Rasmussen (2013) employ a particle Gibbs with ancestor sampling (PGAS) algorithm which consists of a blocked Gibbs scheme and a conditional particle filter with ancestor sampling (CPF-AS) (Lindsten et al. 2014) made possible by the analytically available non-Gaussian density $\mathbb{P}(\mathbf{x}_{0:T}) = \int_{\mathcal{X}} \mathbb{P}(\mathbf{x}_{0:T}, \mathbf{f}_{1:T}) d\mathbf{f}_{1:T}$

(J. Wang et al. 2005). When the latent function values have been integrated out the prior model on the states becomes non-Markovian but can still be represented as a product of Gaussians $\mathbb{P}(\mathbf{x}_t|\mathbf{x}_{0:t-1})$ which are obtained through the standard GP predictive distribution given $\mathbf{x}_{0:t-1}$. These analytically available densities make the sampling scheme possible (for details see Frigola, Lindsten, Schön, and C. Rasmussen (2013)). Integrating out the latent function values can be interpreted as marginalizing the uncertainty over the latent function values and obtaining a prior over the states, which are derived from the latent function that carries the assumptions on the transition function (dependence structure, smoothness, stationarity etc.). This is also visible from the non-Gaussian prior distribution $\mathbb{P}(\mathbf{x}_{0:T}) = \mathbb{P}(\mathbf{x}_0) \int \mathcal{N}(\mathbf{x}_{1:T}|\mathbf{f}_{1:T}, \mathbb{I}_T \otimes \mathbf{Q})\mathcal{N}(\mathbf{f}_{1:T}|\mathbf{m}_{0:T-1}, \mathbf{K}_{0:T-1})d\mathbf{f}_{1:T}$ which after integration keeps the information obtained from $\mathbf{m}_{0:T-1}, \mathbf{K}_{0:T-1}$.The hyper-parameters $\boldsymbol{\theta}$ given the data and the state are sampled in an easier fashion where the parameters for the states ($\boldsymbol{\theta}_f$) and the likelihood ($\boldsymbol{\theta}_\mathbf{y}$) are assumed to allow for independent sampling. The more challenging task of sampling the hyper-parameters of the states are tackled using slice sampling (R. M. Neal 2003; Agarwal and Gelfand 2005). Finally the blocked Gibbs scheme alternates between sampling from the smoothing distributing $\mathbb{P}(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}; \boldsymbol{\theta})$ and that of the hyper-parameter distribution $\mathbb{P}(\boldsymbol{\theta}|\mathbf{y}_{1:T}, \mathbf{x}_{0:T})$ resulting in samples from the joint posterior. Suppose we have K samples $\{(\mathbf{x}^k_{0:T}, \boldsymbol{\theta}^k) \sim \mathbb{P}(\boldsymbol{\theta}^k, \mathbf{x}^k_{0:T}|\mathbf{y}_{1:T})\}_{k=1}^K$ then these can be used for prediction:

$$\mathbb{P}(f(\mathbf{x}_{\tau-1})|\mathbf{x}_{\tau-1}, \mathbf{y}_{1:T}, \boldsymbol{\theta}) \approx \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}(f(\mathbf{x}_{\tau-1})|\mathbf{x}_{\tau-1}, \mathbf{x}^k_{0:T}, \boldsymbol{\theta}^k) \tag{2.46}$$

Initial experimentation with this method resulted in long computation times, we thus continued to study why this is the case and what methods there are to circumvent this obstacle. This approach is the most accurate when it comes to inference and it would be ideal if we had infinite amount of computing and memory capabilities, which in practice is limited. For each batch of observations $\mathbf{y}_{1:T}, \mathbf{y}_{1:T+1}, \ldots, \mathbf{y}_{1:T+\tau}$, N particles and M iterations, the vanilla implementation of the Blocked sampling algorithm has complexity $\mathcal{O}(NMT^4)$ (Frigola, Lindsten, Schön, and C. Rasmussen 2013; Y. Wu et al. 2014) due to the covariance matrix inversion. It can be brought down to $\mathcal{O}(NMT^3)$ by reusing Cholesky factors (rank-one update) from the previous $T \times T$ matrix Frigola, Lindsten, Schön, and C. Rasmussen (2013).

Y. Wu et al. (2014) approximate the smoothing distribution using particles while simultaneously learning $\boldsymbol{\theta}$ using an extension to the Regularized Auxiliary Particle Filter (RAPCF) (Liu and Mike West 2001) to overcome the non-Markovian structure of $\mathbb{P}(\mathbf{x}_{0:T}) = \int \mathbb{P}(\mathbf{x}_{0:T}, \mathbf{f}_{1:T})d\mathbf{f}_{1:T}$. Implementing the algorithm naively would result in $\mathcal{O}(NT^4)$ in the online/filtering case and can brought down to $\mathcal{O}(NT^3)$ by rank-one updating at each step time step. Although the algorithm is faster it has certain limitation as discussed in Y. Wu et al. (2014), however when the PGAS and RAPCF are compared to each other in an empirical study no significant differences are found between the two methods. Thus they conclude that the RAPCF gives faster inference at negligible cost in terms of predictive accuracy.

In this thesis we take a different approach namely by approximating the GP-SSM such that given a complexity control parameter on the model m we can reduce computation time, and this approach is discussed in the next section.

3

Methods for approximate inference

Contents

As mentioned in the regression example employing a GP means having to invert large matrices for inference which can become very computationally cumbersome ($\mathcal{O}(T^3)$ not including the complexity of the estimation algorithm), which for non-sparse matrices constrains the GP framework to situations where we have $T \lesssim 10000$. If we don't have labeled data to estimate the distribution over the GP function on, as is the case in the SSM framework, the computational complexity can become even worse. There are vast amounts of methods designed to speed-up inference/learning/identification of GP prior based methods. In this chapter we quickly review current literature and decide upon a reduced rank approximation method, after which in the first section we delve into the method of Svensson, Solin, et al. (2015) and seek to find if its generalizable. Just having an approximation framework does not make the method implementable, because in the GP framework we must always choose a covariance function and in section 3.1.2 we answer design questions and build two suitable covariance functions. In section 3.1.3 we answer how the covariance function we build in the previous section relates to the regularity of the prior (see section 2.7), and how we can allow for adaptive regularity. In section 3.2 we discuss the implementation of the sampling method for Bayesian inference.

3.1 Hilbert space approximation of SSM

One class of methods that aims to speed up GP inference is that of the Kronecker methods, which are especially useful for multidimensional problems. These methods exploit the structure of product kernels (see 3.1.2), and allow the complexity to drop to $\mathcal{O}(m^2T)$ (see Rakitsch et al. (2013) for details). However given that the aim of this thesis is not necessarily a fast method for high dimensional state-spaces this method is not particularly appealing. Another option is to employ many GP based model models on data subsets of size m, rendering the computational complexity proportional to $\mathcal{O}(m^3)$ (Deisenroth and J. W. Ng 2015). These methods are coined product of experts methods and are known to reduce the ability of the model to pick up on complex structures. Then there are those that are members of the family of Nyström methods. Often referred to as inducing points methods or sparse approximations, the idea is to choose a set of inducing points $\mathbf{x}_{t_u}$ with an associated covariance matrix (the kernel evaluated at these points). Then the eigendecomposition of this covariance matrix $\mathbf{K}(\mathbf{t}_u) \in \mathbb{R}^{m\times m}$ can be scaled to match that of the covariance matrix corresponding to the whole data-set $\mathbf{K}(\mathbf{t})$. This corresponds to the Nyström approximation of the eigen-values and -vectors of the $\mathbf{K}(\mathbf{t})$ (Carl Edward Rasmussen 2006):

$$\hat{\lambda}_j = \frac{1}{m}\lambda_j^u \quad \hat{\phi}_j(\mathbf{x}_t) = \frac{\sqrt{m}}{\lambda_j^u}k(\mathbf{x}_t, \mathbf{x}_{t_u})\mathbf{u}_j \tag{3.1}$$

Where the $\hat{\lambda}_j, \hat{\phi}_j(\mathbf{x}_t)$ are the approximated eigenvalues and eigenfunctions respectively and $\lambda_j^u, \mathbf{u}_j$ the eigenvalues and eigenvectors of $\mathbf{K}(\mathbf{t}_u)$ respectively. Thereby obtaining the approximate covariance matrix of the prior:

$$\mathbf{K}(\mathbf{t}, \mathbf{t}_u)\mathbf{K}(\mathbf{t}_u, \mathbf{t}_u)^{-1}\mathbf{K}(\mathbf{t}_u, \mathbf{t}) \tag{3.2}$$

Then the Woodburry identity can be used for example to speed up inference, for a review see Snelson and Ghahramani (2005) and Quiñonero-Candela and Carl Edward Rasmussen (2005). Additionally Titsias (2009) introduces uncertainty at the inducing points $\mathbf{x}_{t_u}$ via a variational distribution. The variational approach has been implemented in the SSM's by for example Frigola, Y. Chen, et al. (2014b), but in their results it is visible that a large amount of data is needed for the method to perform significantly well.

Solin and Särkkä (2014) take an alternative approach and approximate the Karhunen–Loeve eigenbasis, and because no variational approximations are employed the approximation can be estimated via hierarchical Bayes. We employ sampling methods and obtain a joint posterior that includes the latent function values. In the paper of Svensson, Solin, et al. (2015) they consider the general GP-SSM given in equation 2.43 with $m(\mathbf{x}_t; \boldsymbol{\theta}_f) = 0$. This is the approach we employ in this thesis, because we find in the literature that variational sparse methods often require a large amount of data (Frigola, Y. Chen, et al. 2014b). However, in contrast to the method we employ in this thesis, variational based methods scale very well

to high dimensional problems, but that is not the main target of this thesis. Additionally on a personal note the explicit connection between GP's and basis function expansion in the method is highly educative and allows for a probabilistic view on basis expansion methods. The approach detailed below boils down to using the spectral properties of stationary kernels discussed in section 2.6 (Solin and Särkkä 2014) in combination with methods used for approximating differential operators for partial differential equations (Showalter 2010) to give a computationally efficient approximation of the operator $T_{\mathcal{K}}$ of equation 2.7. As such for this method a necessary assumption is the that the kernel is stationary (6).

3.1.1 The approach

We give an outline of the approach of Solin and Särkkä (2014) including details that fill in the blanks in the paper. The question I aim to answer here is whether or not we can generalise this method to allow for more general covariance structures such in the work of Kom Samo and S. Roberts (2015). Within the time-frame available for this thesis (to design a suitable covariance structure, implement, analyse the methods performance, experiment with various hyper-priors, etc.) I could not easily find a way to generalize the method. I also aim to find whether or not more structure can imposed on the function through defining the mean function. In the analysis below I do find this is possible, but after careful consideration I decided not implement a mean function with the reason being as in section 5.2, which is that we want to see whether the models we propose can extract structural properties, known as stylized facts in the literature, from the data. The details below did result in insights that allow us to answer the question of how to construct a useful kernel for the method.

As in equation 2.7 of section 2.6 we can view the covariance function as an operator. When the covariance function is stationary(def 6) we have:

$$T_\alpha T_{\mathcal{K}}\phi(\mathbf{x}_s) = T_\alpha\hat{\phi}(\mathbf{x}_t) = \int_{\mathcal{X}} k(\mathbf{x}_{t+\alpha},\mathbf{x}_s)\phi(\mathbf{x}_s)\mu_x(d\mathbf{x}_s) \tag{3.3a}$$

$$= \int_{\mathcal{X}} \underbrace{k(\mathbf{x}_t,\mathbf{x}_s)\phi(\mathbf{x}_s)\mu_x(d\mathbf{x}_s)}_{\text{By stationarity}} = \hat{\phi}(\mathbf{x}_t) \tag{3.3b}$$

$$T_{\mathcal{K}} T_\alpha\phi(\mathbf{x}_s) = T_{\mathcal{K}}\phi(\mathbf{x}_{s+\alpha}) = \int_{\mathcal{X}} k(\mathbf{x}_t,\mathbf{x}_{s+\alpha})\phi(\mathbf{x}_{s+\alpha})\mu_x(d\mathbf{x}_s) \tag{3.3c}$$

$$= \int_{\mathcal{X}} k(\mathbf{x}_t,\mathbf{x}_s)\phi(\mathbf{x}_{s+\alpha})\mu_x(d\mathbf{x}_s) = \hat{\phi}(\mathbf{x}_t) \tag{3.3d}$$

Thus because of the stationarity of k we have that the operator $T_{\mathcal{K}}$ is invariant under translations. Thus the transfer function of the operator can be found to be the spectral density $T_{\mathcal{K}}\phi(\mathbf{x}_t) \xrightarrow{\mathcal{L}|_{s=iw}=\mathcal{F}} T_{S(\omega)}\phi(\mathbf{x}_t)$:

$$\mathcal{F}T_{\mathcal{K}}\phi(\mathbf{x}_s) = \int k(\mathbf{r})\phi(\mathbf{x}_s)\mu_x(d\mathbf{x}_s)e^{-i\omega'\mathbf{r}}d\mathbf{r} \tag{3.4a}$$

$$\overset{\text{fubini-tonelli}}{=} \int_{\mathcal{X}}\int_{\mathcal{X}} k(\mathbf{r})e^{-i\omega'\mathbf{r}}d\mathbf{r}\phi(\mathbf{x}_s)\mu_x(d\mathbf{x}_s) \tag{3.4b}$$

$$= \int_{\mathcal{X}} S(\omega)\phi(\mathbf{x}_s)\mu_x(d\mathbf{x}_s) \tag{3.4c}$$

Following Solin and Särkkä (2014) by the isotropy of the kernel we have $S(\omega) \equiv S(||\omega||)$ which is also the transfer function for the kernel operator. Furthermore $S(||\omega||) = S(\sqrt{||\omega||^2}) = S(f(||\omega||^2)) = S(||\omega||^2)$ and if we assume the necessary amount of regularity we can write the polynomial expansion:

$$S(||\omega||) = \sum_{j=0}^{\infty} a_j (||\omega||^2)^j \tag{3.5}$$

We refer to the space $L^2(\mathbb{R}^d) := \{f : \mathbb{R}^d \to \mathbb{R} : \int_{\mathbb{R}^d} |f(\mathbf{x})|^2 d\mathbf{x} < \infty\}$ in the usual manner with the usual $L^2(\mathbb{R}^d)$ inner product $< f, g >_2 = \int_{\mathbb{R}^d} f(\mathbf{x}) g(\mathbf{x}) d\mathbf{x}$ and the norm $||f||_2 = \sqrt{< f, g >}$. Let $(\mathcal{D}(\delta), \delta)$ be the Laplacian defined as $(\delta f)(\mathbf{x}) = \sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_j^2}$ with f in the domain $\mathcal{D}(\delta) := W_s^2(\mathbb{R}^d)$ (Sobolev space consisting of all functions $f \in L^2(\mathbb{R}^d)$ with weak derivatives $\frac{\partial^s f}{\partial x_j^s} \in L^2(\mathbb{R}^d)$ up to order s that have finite $L^2(\mathbb{R}^d)$ norms). We know $C_c^\infty(\mathbb{R}^d)$ is dense in $W_s^2(\mathbb{R}^d) \subset L^2(\mathbb{R}^d)$. We only need to consider $s = 2$ for the laplace operator then δf lies in $L^2(\mathbb{R}^d)$ and we have $\delta : W_2^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$. Let $\mathcal{F}$ be the Fourier operator and note that for $g(\mathbf{x}) = e^{i\omega'\mathbf{x}}$ we have $(\delta g)(\mathbf{x}) = -||\omega||^2 e^{i\omega'\mathbf{x}}$. Then:

$$\delta f(\mathbf{x}) = \delta (2\pi)^{-d} \int (\mathcal{F}f)(\omega) e^{i\omega'\mathbf{x}} \mu(d\omega) \text{ Fourier duality} \tag{3.6a}$$

$$= \delta (2\pi)^{-d} \int (\mathcal{F}f)(\omega) g(\mathbf{x}) d\omega \tag{3.6b}$$

$$= (2\pi)^{-d} \int (\mathcal{F}f)(\omega) \delta g(\mathbf{x}) d\omega \tag{3.6c}$$

$$= (2\pi)^{-d} \int -||\omega||^2 (\mathcal{F}f)(\omega) e^{i\omega'\mathbf{r}} d\omega \tag{3.6d}$$

Where we let the positive finite measure μ be absolute continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and assume that the interchanging of the Laplacian and the integral is allowed for the function f. Given the unique representation of $\delta f(\mathbf{x}) = (2\pi)^{-d} \int (\mathcal{F}\delta f)(\omega) e^{i\omega'\mathbf{x}} d\omega$ we have that $(\mathcal{F}\delta f)(\omega) = -||\omega||^2 (\mathcal{F}f)(\omega)$. Thus the Laplacian is diagonalized by the Fourier transform and can be seen as a Fourier multiplier. Then going back to Solin and Särkkä (2014) we can say that the transfer function of the Laplacian is $\delta f(\mathbf{x}) \to h(\omega)\hat{f}(\omega), h(\omega) = -||\omega||^2$. Then we can take the Fourier transform of equation 3.5 to get a pseudo-differential operator representation of $T_{\mathcal{K}}$:

$$T_{\mathcal{K}} = \sum_{j=0}^{\infty} a_j (-\delta)^j \tag{3.7}$$

Now if we consider the eigenvalue problem of Laplacian with boundary conditions on the compact subset $\Omega \subset \mathbb{R}^d$:

$$-\delta \phi_j(\mathbf{x}) = \lambda_j \phi_j(\mathbf{x}) \text{ if } \mathbf{x} \in \Omega \tag{3.8a}$$

$$\phi_j(\mathbf{x}) = 0 \text{ if } \mathbf{x} \in \underbrace{\partial\Omega}_{\text{Boundary}} \tag{3.8b}$$

Where $\phi_j(\mathbf{x}) \in L^2(\mathbb{R}^d)$ are the eigenfunctions of δ with corresponding eigenvalues $\lambda_j \in \mathbb{R}$. This is a symmetric boundary condition such that $< \delta f, g >=< f, \delta g >$ with known solutions and there are other conditions with this property for example the Neuman conditions. For $\Omega = \bigtimes_{k=1}^{d+e}[-L_k, L_k]$, including inputs $\mathbf{z} = (\mathbf{x}, \mathbf{u}), \mathbf{x} \in \mathbb{R}^d, \mathbf{u} \in \mathbb{R}^e$, the solution is:

$$\lambda_j := \lambda_{j_1,\ldots,j_{d+e}} = \sum_{k=1}^{d+e} \left(\frac{\pi j_k}{2L_k}\right)^2 \tag{3.9a}$$

$$\phi_j(\mathbf{z}) := \phi_{j_1,\ldots,j_{d+e}}(\mathbf{z}) = \prod_{k=1}^{d+e} \sqrt{\frac{1}{L_k}} \sin\left(\frac{\pi j_k (z_k + L_k)}{2L_k}\right) \tag{3.9b}$$

We have that the negative Laplacian operator is also positive definite under the Dirichlet boundary condition, and the combination of symmetry and the boundedness of the negative operator implies that the eigenvalues are all real and positive (assuming it has eigenvalues) and that the eigenfunctions are orthonormal with respect to $< \cdot, \cdot >_2$. Then an orthonormal basis for the formal kernel assigned to the negative Laplacian (Mercer's theorem 4) can be formed.

We let $(\mathcal{X} = \Omega, \mathcal{B} = \mathcal{B}(\mathbb{R}^d)$ and μ_x the usual Lebesgue measure. Writing the operator as an integral transform $(-\delta f)(\mathbf{x}_s) = \int_\Omega l(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_s) d\mathbf{x}_s$ for $-\delta : W_2^2(\Omega) \to L^2(\Omega), f \in W_2^2(\Omega)$ (note that $W_2^2(\Omega) \subset L^2(\Omega)$), and by the previously mentioned properties of the operator Mercer's Thm gives us $l(\mathbf{x}_t, \mathbf{x}_s) = \sum_k \lambda_k \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s)$. Note that by the orthonormality of the basis we can consider formal powers of the operator $(-\delta^p f)(\mathbf{x}_s) = \int_\Omega l(\mathbf{x}_t, \mathbf{x}_s)^p f(\mathbf{x}_s) d\mathbf{x}_s$ with $-\delta^p : W_{2p}^2(\Omega) \to L^2(\Omega), f \in W_{2p}^2(\Omega)$ as:

$$l^p(\mathbf{x}_t, \mathbf{x}_s) = \sum_k \lambda_k^p \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s), s \in \mathbb{N} \tag{3.10}$$

Then we can write:

$$a_0 f(\mathbf{x}_s) + a_1(-\delta f)(\mathbf{x}_s) + a_2((-\delta)^2 f)(\mathbf{x}_s) + \ldots \tag{3.11a}$$

$$= \left[\sum_{j=0}^{\infty} a_j (-\delta)^j\right] f(\mathbf{x}_s) \tag{3.11b}$$

$$= \int_\Omega a_0 f(\mathbf{x}_s) d\mathbf{x}_s + \int_\Omega a_1 l(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_s) d\mathbf{x}_s + \int_\Omega a_2 l^2(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_s) d\mathbf{x}_s + \ldots \tag{3.11c}$$

$$= \sum_{j=0}^{\infty} \int_\Omega a_j l^j(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_s) d\mathbf{x}_s \tag{3.11d}$$

$$= \int_\Omega \sum_{j=0}^{\infty} a_j l^j(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_s) d\mathbf{x}_s \tag{3.11e}$$

Where we assume $\sum_{j=0}^{\infty} \int_\Omega |a_j l^j(\mathbf{x}_t, \mathbf{x}_s) f(\mathbf{x}_s)| d\mathbf{x}_s < \infty$ and apply Fubini-Tonelli because summation can be viewed as integration with respect to the counting measure. We can

write each $a_j l^j$ in equation 3.11e as in equation 3.10 to get:

$$\sum_{j=0}^{\infty} a_j l^j(\mathbf{x}_t, \mathbf{x}_s) = \sum_{j=0}^{\infty} a_j \left[\sum_k \lambda_k^j \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s) \right] \tag{3.12a}$$

$$= \lim_{N\to\infty} \sum_{j=0}^{N} \left[\sum_k a_j \lambda_k^j \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s) \right] \tag{3.12b}$$

$$= \sum_k \left[\lim_{N\to\infty} \sum_{j=0}^{N} a_j \lambda_k^j \right] \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s) \tag{3.12c}$$

$$= \sum_k \left[\sum_{j=0}^{\infty} a_j \lambda_k^j \right] \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s) \tag{3.12d}$$

Again we assume $\lim_{N\to\infty} \sum_{j=0}^{N} |\sum_k a_j \lambda_k^j \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s)| < \infty$ with respect to the counting measure. Note that the sums in equations 3.7 and 3.11c coincide, thus we can fill in $T_{\mathcal{K}}$ for the sum in equation 3.11c. Then from the integral operator form as in equation 2.7 of $T_{\mathcal{K}}$ we can see that we can approximate the kernel on the domain Ω:

$$k(\mathbf{x}_t, \mathbf{x}_s) \approx \sum_{j=0}^{\infty} a_j l^j(\mathbf{x}_t, \mathbf{x}_s) \tag{3.13a}$$

$$= \sum_k \left[\sum_{j=0}^{\infty} a_j \lambda_k^j \right] \phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s) \tag{3.13b}$$

By comparing the sum in brackets in equation 3.13b to that of equation 3.5 we can write use the same polynomial expansion for $S(\sqrt{\lambda_k})$ where $\lambda_k = ||\omega||^2$. For notational ease we write the function $S(\sqrt{\lambda_k})$ as $S(\lambda_k)$ and we get the following approximation (including the hyper-parameters) for the kernel:

$$k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f) \approx \sum_{k=1}^{m} S_{\boldsymbol{\theta}_f}(\lambda_k)\phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s) \tag{3.14}$$

Where $S_{\boldsymbol{\theta}}(\cdot)$ is the only term dependent on $\boldsymbol{\theta}$. The domain is restricted to be some compact $\Omega \subset \mathbb{R}^D$ and symmetric boundary conditions (Dirichlet boundary conditions) are assumed. Thus even in the case of an infinite terms on the right in equation 3.14 the boundary conditions and the compactness of the domain does not allow for equality in the equation. It is also important to note that in the implementations we make sure that the inputs are always far away from the boundary inside the $d+e$-dimensional cube. We can use the approximation in equation 3.14 to form the Karhunen–Loève (see 1) expansion of $f(\mathbf{x}_t) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f)$:

$$f(\mathbf{x}_t) \approx \sum_{k=1}^{m} \mathcal{N}(\mathbf{0}, S_{\boldsymbol{\theta}_f}(\lambda_k))\phi_k(\mathbf{x}_t) \tag{3.15}$$

Ch3. Theorem 1 (Karhunen–Loève Theorem)**.** *Let $T, \mathcal{X} \subset \mathbb{R}^d$ with T a compact index set, let $\mathbf{X} = \{X_t, t \in T\}$ be a centered stochastic process mapping values from the probability space $(\Omega, \mathcal{A}, p)$ to $(\mathcal{X}, \mathcal{B}(\mathbb{R}))$. Let the kernel $k : T \times T \to \mathbb{R}$ be such that an orthonormal*

basis $\{\phi_j\}_{j=1}^{\infty}$ *of* $L^2(\mathcal{X})$ *can be formed by the eigenfunctions of its corresponding operator with respective eigenvalues* $\{\lambda_j\}_{j=1}^{\infty}$ *(Theorem 4). Then* $\mathbf{X}$ *admits the representation:*

$$X_t = \sum_{k=1}^{\infty} Z_k \phi_k(\mathbf{x}_t) \tag{3.16}$$

Where Z_k *are zero mean, uncorrelated random variable with covariance equal to* λ_k*.*

Although the expansion requires a centred stochastic process, theoretically we could also implement the mean function for $f(\mathbf{x}_t) \sim \mathcal{GP}(m(\mathbf{x}_t;\boldsymbol{\theta}_f), k(\mathbf{x}_t,\mathbf{x}_s;\boldsymbol{\theta}_f))$:

$$f(\mathbf{x}_t) - m(\mathbf{x}_t;\boldsymbol{\theta}_f) \approx \sum_{k=1}^{m} \mathcal{N}(\mathbf{0}, S_{\boldsymbol{\theta}_f}(\lambda_k))\phi_k(\mathbf{x}_t) \tag{3.17a}$$

$$f(\mathbf{x}_t) \approx m(\mathbf{x}_t;\boldsymbol{\theta}_f) + \sum_{k=1}^{m} \mathcal{N}(\mathbf{0}, S_{\boldsymbol{\theta}_f}(\lambda_k))\phi_k(\mathbf{x}_t) \tag{3.17b}$$

We can thus write the state equation 2.43 as :

$$\mathbf{x}_{t+1} = m(\mathbf{x}_t;\boldsymbol{\theta}_f) + \underbrace{\begin{bmatrix} w_{1,1} & \cdots & w_{1,m} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,m} \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} \phi_1(\mathbf{x}_t) \\ \vdots \\ \phi_m(\mathbf{x}_t) \end{bmatrix}}_{\boldsymbol{\phi}(\mathbf{x}_t)} + \boldsymbol{\eta}_t \tag{3.18}$$

However in practical situations we often don't have such strong prior knowledge as to be able to specify the form of the mean function and in the case of structure discovery we do not need to to impose such prior assumptions. Alternatively it would be possible to also infer the mean function using a basis function expansion (Blight and Ott 1975), but we choose to set the mean function to zero so that in the implementation we can test how well, without imposing the stylized structure, it is able to capture certain facts we know about financial time series (see application chapter). Note again that this certainly does not imply a zero mean posterior, as the posterior mean lying in the RKHS contains the corresponding kernel shifted by the mean (see regression example). If we now include the inputs in the notation, we have:

$$k((\mathbf{x}_t,\mathbf{u}_t),(\mathbf{x}_s,\mathbf{u}_s);\boldsymbol{\theta}_f) \approx \sum_{k=1}^{m} S_{\boldsymbol{\theta}_f}(\lambda_k)\phi_k(\mathbf{x}_t,\mathbf{u}_t)\phi_k(\mathbf{x}_s,\mathbf{u}_t) \tag{3.19a}$$

$$f(\mathbf{x}_t,\mathbf{u}_t) \approx \sum_{k=1}^{m} \mathcal{N}(\mathbf{0}, S_{\boldsymbol{\theta}_f}(\lambda_k))\phi_k(\mathbf{x}_t,\mathbf{u}_t) \tag{3.19b}$$

Note that in the formulation above covariance between the states is allowed for depending on the spectral density structure (more on this in the next subsection). We can rephrase the

prior model of equation 2.34 including the hyper-priors with a similar joint prior:

$$\mathbf{x}_0 \sim \mathbb{P}_0(\mathbf{x}_0) \tag{3.20a}$$

$$\boldsymbol{\theta}_f \sim \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_f) \tag{3.20b}$$

$$f(\mathbf{x}_t, \mathbf{u}_t) \sim \mathcal{GP}(0, k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f)) \Leftrightarrow f(\mathbf{x}_t, \mathbf{u}_t) \approx \sum_{k=1}^{m} \mathcal{N}(\mathbf{0}, S_{\boldsymbol{\theta}_f}(\lambda_k))\phi_k(\mathbf{x}_t, \mathbf{u}_t) \tag{3.20c}$$

$$f(\mathbf{x}_t, \mathbf{u}_t) := \mathbf{f}_{t+1} \tag{3.20d}$$

$$\mathbf{Q} \sim \mathbb{P}_Q(\mathbf{Q}) \tag{3.20e}$$

$$\mathbf{x}_{t+1} = \underbrace{\begin{bmatrix} w_{1,1} & \cdots & w_{1,m} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,m} \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} \phi_1(\mathbf{x}_t, \mathbf{u}_t) \\ \vdots \\ \phi_m(\mathbf{x}_t, \mathbf{u}_t) \end{bmatrix}}_{\boldsymbol{\phi}(\mathbf{x}_t, \mathbf{u}_t)} + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t | \mathbf{Q} \sim \mathcal{N}(\boldsymbol{\eta}_t | 0, \mathbf{Q}) \tag{3.20f}$$

$$\mathbf{x}_{t+1} | \mathbf{f}_{t+1}, \mathbf{Q} \sim \mathcal{N}(\mathbf{x}_{t+1} | \mathbf{f}_{t+1}, \mathbf{Q}) \tag{3.20g}$$

$$\mathbf{y}_t | \mathbf{x}_t \sim \mathbb{P}_y(\mathbf{y}_t | \mathbf{x}_t) \tag{3.20h}$$

Note that the we could also place a GP prior on the measurement function in exactly the same manner, however this formulation suffices for the purposes of chapter chapter 5. Furthermore at the cost of complexity in this formulation $f(\mathbf{x}_t, \mathbf{u}_t) = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}_t, \mathbf{u}_t)$ is non separable, which results in the expansion order being $m^{\dim(\mathbf{x}_t)+\dim(\mathbf{u}_t)}$ rather than $m_x^{\dim(\mathbf{x}_t)} + m_u^{\dim(\mathbf{u}_t)}$. This would allow for non-additive state functions to be identifiable, i.e. functions that are not of the form $f(\mathbf{x}_t, \mathbf{u}_t) \neq g(\mathbf{x}_t) + h(\mathbf{u}_t)$, but can be relaxed in higher dimensional cases. For the volatility application we simply set $\mathbf{x}_0 = 0$ or $\mathbf{x}_0 = \log(\widehat{\mathbb{V}ar(y_t)})$.

So we have semi-parametrized the SSM and returned in a sense to the basis function expansion view discussed in chapter two (section 2.4), as in equation 3.19b $w_{kl} \sim \mathcal{N}(0_k, S_{k,\boldsymbol{\theta}_f}(\lambda_l))$. We are in fact back to placing priors on the expansion weights, but for these specific distribution on the weights as well as these basis functions we have that the priors correspond to Gaussian Process regression as $L \to \infty$ for $\Omega = [-L, L]$ and $m \to \infty$ *for any stationary kernel*. We can control the trade-off between speed and accuracy similarly to the sparse framework (Titsias 2009). We retain all the probabilistic properties that allow for uncertainty propagation that full Gaussian process priors over functions have.

With regards to the discussion about the truncation of the expansion, suppose we are in the basis function expansion case where for simplicity we consider a compact set $\Omega \in \mathbb{R}^d$ with a set of orthogonal basis functions $\{\phi_l\}_{l=1}^{\infty}$ spanning the function space on Ω. A classical example is the Fourier basis spanning $L^2(\Omega)$. Then a finite set of basis functions of size m would have to be chosen for approximation. Rather than truncating the expansion by hand the Gaussian Process approximation puts priors on the weights of the basis functions, and as such regularizes the expansion to prevent or at least reduce over-fitting. As mentioned before GP regression can be viewed as a special case of a possibly infinite expansion where the imaginative order is determined by the data. Then here we can set m as large as the amount

of computational burden that can be carried, and as can be seen in the analysis section the data determines which posterior weights will be distributed very near zero (see for example Figure 6.5). Thus we continue the theme of allowing the model to be data driven. Another way of looking at the truncation problem is noting that finding the mode of the posterior $\mathbb{P}_y(\mathbf{W}|\mathbf{y}_{1:T})$ corresponds to regularized maximum likelihood estimation (section 2.5), and thus unneeded weights are penalized to reduce over-fitting.

From this analysis we are unable to find a way of employing more general covariance structure such as the one of Kom Samo and S. Roberts (2015). However it does give us insight into how the kernel enters the approximation and how we can customize it and to get the most out of it. Thus we would like to enrich the language of possible kernels that we can easily implement in the form of their spectra. Another important question is how can we can best set up a kernel that maps from a multidimensional space, is computationally efficient and ideally allows for interactions between the states, and we find the answer in next subsection.

3.1.2 Covariance structure

The expressiveness of the model depends on the covariance function of the GP and for example the most popular one, the Squared Exponential, is in general not able to capture complex covariance structures and performs as a smoothing interpolator (A. G. Wilson and Adams 2013). Kernels with limited expressive power, of which their spectral densities are used within the reduced rank approximation method, means that we are estimating an approximation to a model that is already limited in its ability to capture complex structures in the data. A way of increasing the ability of the GP model of capturing more complex covariance structures and embedding these in the function approximations is to employ combinations of kernels (D. Duvenaud 2014). If we would like to be able to do this in the reduced rank approximation we would like to obtain the spectral density of combinations of kernels. The question then arises if the combination of spectral densities of a set of kernels corresponds to the spectral density of the combination of that set of kernels.

We consider kernel addition and multiplication here and in the literature we often find proofs of validity from the feature map perspective. As an example let $(\mathcal{X}, \mu_x)$ be a measure space and assume the existence of a map $\phi : \mathcal{X} \to L^2(\mu_x)$ such that for the mapping $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ we have $k(\mathbf{x}_t, \mathbf{x}_s) =< \phi(\mathbf{x}_t), \phi(\mathbf{x}_s) >_2$, with $\mathbf{x}_t, \mathbf{x}_s \in \mathcal{X}$, then using Mercer's theorem (4) we can easily show the sum and product of kernels to be a valid kernels by showing that they are PSD. However, having noted this, we take the random field approach in this section and setup straight forward proofs that we use later. From this perspective moving on to decomposing the random fields into combinations of stochastic processes for multidimensional inputs seems more intuitive.

Ch3. Theorem 2 (Sum of covariance functions over the same space)**.** *Let $T, \mathcal{X} \subset \mathbb{R}^d$ and let $\mathbf{X} = \{X_t, t \in T\}$ be a random field mapping values from the probability space $(\Omega, \mathcal{A}, p)$ to $(\mathcal{X}, \mathcal{B}(\mathbb{R}^d))$ with its trajectories $X_t(\omega) : T \to \mathcal{X}$ being vector valued functions (in some desirable space), for some $\omega \in \Omega$. Then its covariance function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$ can be defined as $K(\mathbf{x}_t, \mathbf{x}_s) = \sum_{l=1}^n k_l(\mathbf{x}_t, \mathbf{x}_s)$ for symmetric PSD covariance functions $k_l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d\times d}$*

Proof. Let $\mathbf{X} := \sum_{l=1}^n \mathbf{Z}_l$, with $\mathbf{Z}_l$ independent random fields over the same space $(\Omega, \mathcal{A}, p)$ with PSD covariance functions $k_l(\mathbf{x}_t, \mathbf{x}_s)$ so that:

$$\forall \omega \in \Omega \text{ and } \forall t \in T,\ X_t(\omega) = \sum_{l=1}^{n} Z_t^{(l)}(\omega) \tag{3.21}$$

Then, using the expectation definition of covariance, we have that for the marginals for all $t, s \in T$ and $\omega \in \Omega$ the covariances are given by:

$$\mathbb{E}[(X_t(\omega) - \mu_t)(X_s(\omega) - \mu_s)'] = \mathbb{E}[(\sum_{l=1}^{n} Z_t^{(l)}(\omega) - \sum_{l=1}^{n} \mu_t^{(l)})(\sum_{k=1}^{n} Z_s^{(k)}(\omega) - \sum_{k=1}^{n} \mu_s^{(k)})'] \tag{3.22}$$

$$= \sum_{l=1}^{n}\sum_{k=1}^{n} \mathbb{E}[Z_t^{(l)}(\omega) Z_s^{(k)}(\omega)'] - \sum_{l=1}^{n} \mu_t^{(l)} \mu_s^{(l)'} \tag{3.23}$$

$$\overset{\text{By independence}}{=} \sum_{l=1}^{n} \mathbb{E}[Z_t^{(l)}(\omega) Z_s^{(l)}(\omega)'] - \sum_{l=1}^{n} \mu_t^{(l)} \mu_s^{(l)'} \tag{3.24}$$

$$= \sum_{l=1}^{n} \mathbb{C}ov(Z_t^{(l)}(\omega), Z_s(\omega)^{(l)}) \tag{3.25}$$

Where $\mu_t^{(l)} := \mathbb{E}[Z_t^{(l)}(\omega)]$. □

Note that for the approximation $k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f) \approx \sum_{k=1}^m S_{\boldsymbol{\theta}_f}(\lambda_k)\phi_k(\mathbf{x}_t)\phi_k(\mathbf{x}_s)$ (Equation 3.14) we consider the spectral density of $k_{\boldsymbol{\theta}_f} : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ (or to $\mathbb{R}$ and not $\mathbb{R}^{d\times d}$) which we use inside the approximation.

The above theorem can be used to construct sum kernels, but given that we did not have time to play with various kernel combinations over the same space, what it did bring is additive kernels over multiple spaces (D. K. Duvenaud et al. 2011). The direct sum of kernels $k_l : \mathcal{X}_l \times \mathcal{X}_l \to \mathbb{R}$ denoted as $\bigoplus_{l=1}^d k_l$ is also valid by the above theorem. We can also allow for interaction terms but this makes for a complex structure (D. K. Duvenaud et al. 2011) and we found it in practice to not work as well as the spectral density of the convolution kernel, which we find later on. The interaction terms are incorporated by allowing the cross correlations between $Z_t^{(l)}(\omega)$ and $Z_s^{(k)}(\omega)$ to be non zero, where $Z_t^{(l)}(\omega)$ lies in $\mathcal{X}_l$ and $Z_s^{(k)}(\omega)$ lies in $\mathcal{X}_k$.

From the linearity of the Fourier transform and the spectral isometry we get the corresponding sum of spectral densities.

Ch3. Proposition 1 (Sums of spectral densities)**.** *Let $\mathcal{X} \subset \mathbb{R}^d$, $T \subset \mathbb{R}^d$, $\mathbf{x}_t, \mathbf{x}_s \in \mathcal{X}$ and $t, s \in T$. Suppose we have n valid PSD kernels $k_l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, indexed by T, which we denote $k_l(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_{k_l}) \equiv k_l(\mathbf{r}; \boldsymbol{\theta}_{k_l})$ for all $l = 1, \ldots, n$ with $\mathbf{r} = \mathbf{x}_t - \mathbf{x}_s$. Let $S_l(\omega)$be the corresponding spectral density of kernel k_l then we have that the spectral density of the sum of the kernels is the sum of the spectral density of the kernels.*

Proof. For the sum of the kernels we have that:

$$\sum_{l=1}^{n} k_l(\mathbf{r}; \boldsymbol{\theta}_{k_l}) = k(\mathbf{r}; \boldsymbol{\theta}_k)$$

Suppose that the kernels are stationary, thus $k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_k)$ for some stochastic process $\mathbf{X}$ is stationary. If we were to view the process $\mathbf{X}$ as a sum of stationary processes $(\mathbf{X}_i)$ we would have that these processes are jointly weakly stationary. We assume this spectral measure μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ with density S which allows us to write:

$$k(\mathbf{r}; \boldsymbol{\theta}_k) = (2\pi)^{-d} \int_{\mathbb{R}^D} S(\omega) e^{i\omega' \mathbf{r}} d\omega$$

And thus its Fourier Dual (Wiener-Khintchine theorem):

$$\begin{aligned} S(\omega) &= \int_{\mathcal{X}} \left[\sum_{l=1}^{n} k_l(\mathbf{r}; \boldsymbol{\theta}_{k_l}) \right] e^{-i\omega' \mathbf{r}} d\mathbf{r} \\ &= \sum_{l=1}^{n} \int_{\mathcal{X}} k_l(\mathbf{r}; \boldsymbol{\theta}_{k_l}) e^{-i\omega' \mathbf{r}} d\mathbf{r} \\ &= \sum_{l=1}^{n} S_l(\omega) \end{aligned}$$

Where by the validity and stationarity of the kernels k_l they all have a Fourier dual S_l. Note that ω here refers to the frequency spectrum variable and not an event. □

This is indirectly incorporated in the approximate framework in the following way:

$$k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f) \approx \sum_{l=1}^{n} \sum_{k_l=1}^{m} S^{(l)}_{\boldsymbol{\theta}_f}(\lambda_{k_l}) \phi_{k_l}(\mathbf{x}_t) \phi_{k_l}(\mathbf{x}_s) \tag{3.26a}$$

$$= \sum_{k=1}^{m} \sum_{l=1}^{n} S^{(l)}_{\boldsymbol{\theta}_f}(\lambda_k) \phi_k(\mathbf{x}_t) \phi_k(\mathbf{x}_s) \tag{3.26b}$$

Now if we look at the product of kernels we can again quickly see its validity. Again we consider proving it from the stochastic process perspective.

Ch3. Theorem 3 ((Hadamard) Product of covariance functions over the same space)**.** *Let $T, \mathcal{X} \subset \mathbb{R}^d$ and let $\mathbf{X} = \{X_t, t \in T\}$ be a random field mapping values from the probability space $(\Omega, \mathcal{A}, p)$ to $(\mathcal{X}, \mathcal{B}(\mathbb{R}^d))$ with its trajectories $X_t(\omega) : T \to \mathcal{X}$ being vector valued functions (in some desirable space), for some $\omega \in \Omega$. Then its covariance function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$ can be defined as $K(\mathbf{x}_t, \mathbf{x}_s) = \bigcirc_{l=1}^{n} k_l(\mathbf{x}_t, \mathbf{x}_s)$ for symmetric PSD $k_l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{d \times d}$*

Proof. Let $\mathbf{X} := \bigcirc_{l=1}^{n} \mathbf{Z}_l$, with $\mathbf{Z}_l$ independent random fields over the same space $(\Omega, \mathcal{A}, p)$ with PSD covariance functions $k_l(\mathbf{x}_t, \mathbf{x}_s)$ so that:

$$\forall \omega \in \Omega \text{ and } \forall t \in T,\ X_t(\omega) = \bigcirc_{l=1}^{n} Z_t^{(l)}(\omega) \tag{3.27}$$

Then, using the expectation definition of covariance, we have that, for the finite dimensional distributions, for all $t, s \in T$ and $\omega \in \Omega$ the covariances are given by:

$$\begin{aligned}
\mathbb{E}[(X_t(\omega) - \mu_t)(X_s(\omega) - \mu_s)'] &= \mathbb{E}[X_t(\omega)X_s(\omega)'] - \mu_t\mu_s' && (3.28)\\
&= \mathbb{E}[\bigcirc_{l=1}^{n} Z_t^{(l)}(\omega)\ \bigcirc_{l=1}^{n} Z_s^{(l)}(\omega)'] && (3.29)\\
&- \mathbb{E}[\bigcirc_{l=1}^{n} Z_t^{(l)}(\omega)]\mathbb{E}[\bigcirc_{l=1}^{n} Z_s^{(l)}(\omega)'] && (3.30)\\
&= \{\mathbb{E}[\prod_{l=1}^{n} Z_{t,i}^{(l)}(\omega) \prod_{l=1}^{n} Z_{s,j}^{(l)}(\omega)]\}_{i=1,j=1}^{d,d} && (3.31)\\
&- \{\mathbb{E}[\prod_{l=1}^{n} Z_{t,i}^{(l)}(\omega)]\mathbb{E}[\prod_{l=1}^{n} Z_{s,j}^{(l)}(\omega)']\}_{i=1,j=1}^{d,d} && (3.32)\\
\overset{\text{By independence}}{=} &\{\prod_{l=1}^{n} \mathbb{E}[Z_{t,i}^{(l)}(\omega)Z_{s,j}^{(l)}(\omega)]\}_{i=1,j=1}^{d,d} && (3.33)\\
&- \{\prod_{l=1}^{n} \mathbb{E}[Z_{t,i}^{(l)}(\omega)]\mathbb{E}[Z_{s,j}^{(l)}(\omega)']\}_{i=1,j=1}^{d,d} && (3.34)\\
&= \bigcirc_{l=1}^{n} \mathbb{C}ov(Z_t(\omega)^{(l)}, Z_s(\omega)^{(l)}) && (3.35)
\end{aligned}$$

Note that if the independent fields are Gaussian then the product can certainly not expected to be Gaussian, but by virtue of the product kernel being symmetric and PSD we have that there exists a Gaussian process with the product covariance function. This existence is implied by Kolmogorov's extension theorem. □

Given time-constraints on the experimentation with product kernels whats of value here, besides being able to create a more rich language of covariance functions for the reduced rank GP-SSM model, we can set up tensor product kernels over multiple spaces. Because the covariance functions are closed under multiplication it is easy to see that they are also closed under the tensor product (for a proof from the symmetric and positive definiteness perspective see (corollary 1.13 (Van Den Berg et al. 2012)). However we assumed independence here, furthermore the implementation of such product kernels does not come as easy as with the direct sum kernels. When we consider a product of spectral densities we don't have the same properties as with the sums of spectral densities. I quickly encountered this by heuristically employing the same reasoning as in Equation 3.26 with the sum kernel. We see that for n=2:

$$k(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_f) \approx \sum_{k_1=1}^{m} S_{\boldsymbol{\theta}_f}^{(1)}(\lambda_{k_1})\phi_{k_1}(\mathbf{x}_t)\phi_{k_1}(\mathbf{x}_s) \cdot \sum_{k_2=1}^{m} S_{\boldsymbol{\theta}_f}^{(2)}(\lambda_{k_2})\phi_{k_2}(\mathbf{x}_t)\phi_{k_2}(\mathbf{x}_s) \tag{3.36a}$$

The kernel in the above equation as a product of two approximation does not involve simply taking the product over the spectral densities, rather we have a product over polynomials. If

we just look at the spectral densities we have $\prod_{l=1}^{n}\sum_{k_1=1}^{m} S_{\boldsymbol{\theta}_f}^{(l)}(\lambda_{k_l})$, where it is obvious we have a convolution type of problem. This is the indirectly result of the spectra being Fourier transforms. It turns out this is a standard result in the signal processing literature, which brings us to the convolution theorem that is in line with what we encounter in Equation 3.36. In short the theorem says that the Fourier transform of a point-wise product (as in 3) is the convolution of the Fourier transforms. But it also turns out that from the same theorem we get the bonus that the Fourier transform of the convolution of inverse Fourier transforms of functions is the product of those functions. If we recall from Bochner's Theorem (Equation 2.29) that the kernel of a weakly stationary mean-square continuous random process is the inverse Fourier transform $k = \mathcal{F}^{-1}S$ then we can see that the product of spectral densities can be interpreted as $\bigcirc_{l=1}^{n} S^{(l)} = \mathcal{F}(\bigstar_{l=1}^{n} k_l)$ by the convolution theorem, where $\bigstar$ is the convolution symbol. This is easily seen if we suppose we have symmetric PSD, weakly stationary and mean square continuous processes with kernels $k_l : \mathcal{X}\times\mathcal{X} \to \mathbb{C}$ indexed by T such that $k_l(\mathbf{r}) = (2\pi)^{-d}\int_{\mathbb{R}} e^{i\omega'\mathbf{r}}\mu(d\omega)$ where μ is assumed to be absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$ and $l = 1, 2$ (without loss of generality). Then the Fourier dual of the convolution kernel $K = k_1 \star k_2$ by the Wiener-Khintchine theorem is:

$$
\begin{aligned}
S(\omega) &= \int_{\mathcal{X}} K(\mathbf{r}) e^{-i\omega'\mathbf{r}} d\mathbf{r} && \text{(3.37a)}\\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k_1(\mathbf{r}-\mathbf{s}) k_2(\mathbf{s}) e^{-i\omega'\mathbf{r}} d\mathbf{s} d\mathbf{r} && \text{(3.37b)}\\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} k_1(\mathbf{r}-\mathbf{s}) e^{-i\omega'(\mathbf{r}-\mathbf{s})} k_2(\mathbf{s}) e^{-i\omega'\mathbf{s}} d\mathbf{s} d\mathbf{r} && \text{(3.37c)}\\
&= S_1(\omega)\cdot S_2(\omega) \text{ where } \mathbf{r}-\mathbf{s} \text{ is a new variate} && \text{(3.37d)}
\end{aligned}
$$

However this is under the assumption that that the kernel $K = k_1 \star k_2$ is valid. Luckily we find a proof of this in the spatial statistics literature(Thm. 1 (Majumdar and Gelfand 2007)). Whats more is that we can view the convolution kernel $K_{lk} = k_l \star k_k$ with $l, k \in \{1, \ldots, n\}$ as the cross-covariance function and k_{ll} as the covariance function (Majumdar and Gelfand 2007). This is not shown for convolution of the type k_l over $\mathcal{X}_l$ and k_k over $\mathcal{X}_k$, however intuitively this looks like a direct result. The problem with these kernel convolutions is that the integrals are often intractable, however with this approximation framework we only need the spectral densities, which we can take to be the point-wise product over the product spaces. In fact now the class of "separable" (product) kernels become hard to employ because these are convolutions of spectral densities.

Given that the direct sum and convolution kernels over multiple spaces have easy spectral density forms, these are the ones we got to experiment with. We performed these preliminary experiments in the simulation setting of section 6.2, and the convolution kernel spectral density far outperformed the additive one, which might very well be because of the cross covariances being accounted for in the product spectral density (dual of the convolution kernel).

Now that we have settled on the structure of the spectral densities we employ, we can look at the convergence of the approximate model, for which we use the convergence theorem of Solin and Särkkä (2014) with a trivial modification to include the inputs.

Ch3. Theorem 4 (Convergence GP-SSM approximation)**.** *Let* $\mathbf{x}_t \in \mathbb{R}^d, \mathbf{u}_t \in \mathbb{R}^e$ *and* $T, \Omega \subset \mathbb{R}^{d+e}$*. Let* k *be stationary kernel on* $\Omega \times \Omega$*, indexed by* T*, such that we have:*

$$k(\mathbf{r}) = (2\pi)^{-d+e} \int_{\mathbb{R}^{d+e}} e^{i\omega' \mathbf{r}} \mu(d\omega) \tag{3.38}$$

And $\Omega = \bigtimes_{k=1}^{d+e}[-L_k, L_k]$*,* $\mathbf{r} = \mathbf{z}_t - \mathbf{z}_s$ *where* $\mathbf{z}_t = (\mathbf{x}_t, \mathbf{u}_t)$*. We assume* μ *to be absolutely continuous with respect to the Lebesgue measure on* $\mathbb{R}^{d+e}$ *so that the spectral density* $S(\omega)$ *exists. Let* $S(\omega)$ *be two times differentiable with bounded derivatives. Let the integral* $\int_{\mathbb{R}^{d+e}} S(\omega)d\omega < \infty$*, which in this case is equivalent to having the single variable integrals be finite. Furthermore we assume that the inputs at which the kernel is evaluated are contained in* $\Omega = \bigtimes_{k=1}^{d+e}[-\hat{L}_k, \hat{L}_k]$ *with* $\hat{L}_k < L_k$*. Let*

$$k_m(\mathbf{z}_t, \mathbf{z}_s) = \sum_{k=1}^{m} S(\lambda_k)\phi_k(\mathbf{z}_t)\phi_k(\mathbf{z}_s) \tag{3.39}$$

Where ϕ_k *and* λ_k *are defined as in Equation 3.9. Then , if* $\frac{m}{L_k} \to \infty$ *when* $m, L_1, \ldots, L_{d+e} \to \infty$*, there exists a constant* C_{d+e} *such that:*

$$|k(\mathbf{z}_t, \mathbf{z}_s) - k_n(\mathbf{z}_t, \mathbf{z}_s)| \leq \frac{C_{d+e}}{\min_k L_k} + \frac{1}{\pi^{d+e}} \int_{||\omega|| \geq \frac{\pi m}{2 \min_k L_k}} S(\omega)d\omega \tag{3.40}$$

Which implies that uniformly:

$$\lim_{L_1,\ldots,L_{d+e}\to\infty} \left[\lim_{m\to\infty} k_m(\mathbf{z}_t, \mathbf{z}_s)\right] = k_(\mathbf{z}_t, \mathbf{z}_s) \tag{3.41}$$

Furthermore the uniform convergence of the prior kernel implies uniform convergence of the posterior mean and covariance as $m, L_1, \ldots, L_{d+e} \to \infty$*.*

Proof. See Solin and Särkkä (2014) A.2 □

The two kernels we consider in this thesis are that of the Matern class (Stein 1999) and the Gaussian mixture (A. G. Wilson and Adams 2013). For the application we consider $\mathcal{X}_l \subset \mathbb{R}$ and consider the spectral density of a convolution kernel over the product space $\bigtimes_{l=1}^{d+e} \mathcal{X}_l$ and the spectral density $S(\omega) = \prod_{l=1}^{d+e} S_l(\omega_l)$. The general requirements for the Matern class and Spectral mixture kernel are clearly met. Furthermore given finite real valued states and inputs for our application we can always set $\hat{L}_k > \max\{\max_{t=1,\ldots,T}\{|\mathbf{x}_t|, |\mathbf{y}_t|\}\} + C$ with $C \in \mathbb{R}_+$. We can let the requirement $\frac{n}{L_k} \to \infty$ when $n, L_1, \ldots, L_{d+e} \to \infty$ hold by assumption and then the only requirements that are left are: $|D^i S(\omega)| < \infty, i = 1.2$, where the D stands for the differentiation operator, and $\int_{\mathbb{R}^{d+e}} S(\omega)d\omega < \infty$. We begin with the Matern class spectral density which we take to be $S^{mat}(\omega) = \prod_{l=1}^{d+e} S_l^{mat}(\omega_l)$ with:

$$S_l^{mat}(\omega_l) = \frac{2\pi^{\frac{1}{2}}\Gamma(\nu + \frac{1}{2})(2\nu)^\nu}{\Gamma(\nu)\ell^{2\nu}} \left(\frac{2\nu}{\ell^2} + 4\pi^2\omega_l^2\right)^{-(\nu+\frac{1}{2})} \tag{3.42}$$

For $\nu, \ell > 0$, this function (rational, polynomial, denominator never zero) is clearly smooth, and note again that the scaling is chosen such that, as $\nu \to \infty$, the Matern class kernel converges to the squared exponential (SE) kernel with its corresponding spectral density (Stein 1999) with continuous derivatives of order n. Thus given that also $\ell < \infty$ the derivatives exist and are bounded. Next the question arises whether or not for the spectral density of the convolution kernel $\int_{\mathbb{R}^{d+e}} S(\omega)d\omega < \infty$ also holds:

$$\int_{\mathbb{R}^{d+e}} S^{mat}(\omega)d\omega = \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \prod_{l=1}^{d+e} S_l^{mat}(\omega_l)d\omega_1 \cdots d\omega_{d+e} \tag{3.43a}$$

$$= \int_{\mathbb{R}} S_1^{mat}(\omega_1) \cdots \int_{\mathbb{R}} S_{d+e}^{mat}(\omega_{d+e})d\omega_1 \cdots d\omega_{d+e} \tag{3.43b}$$

Thus we only look at $\int_{-\infty}^{\infty} S_l^{mat}(\omega_l)d\omega_l$ and note that spectral density in Equation 3.42 corresponds to the Matern class random process, and by virtue of it being a valid Gaussian process its variance exists and if we look at its spectral representation we have:

$$k(0) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{i\omega r}\Big|_{r=0} S_l^{mat}(\omega_l)d\omega_l = (2\pi)^{-1} \int_{-\infty}^{\infty} S_l^{mat}(\omega_l)d\omega_l < \infty \tag{3.44}$$

For the Gaussian Mixture spectral density its even easier because $S_l^{mix}(\omega_l)d\omega_l$ is a sum of symmetrized scale location Gaussians. Let $\Phi_l^i(\omega_l|\mu_i, \sigma_i^2) = \mathcal{N}(\omega_l|\mu_i, \sigma_i^2) + \mathcal{N}(-\omega_l|\mu_i, \sigma_i^2)$ then clearly $\frac{d^2}{d\omega_l^2}\Phi_l^i(\omega_l|\mu_i, \sigma_i^2) = \frac{d^2}{d\omega_l^2}\mathcal{N}(\omega_l|\mu_i, \sigma_i^2) + \frac{d^2}{d\omega_l^2}\mathcal{N}(-\omega_l|\mu_i, \sigma_i^2)$ which exists and is bounded. Then the same holds also holds for the mixture $S_l^{mix}(\omega_l) = \sum_{i=1}^{q} a_i\Phi_l^i(\omega_l|\mu_i, \sigma_i^2)$. Its also easy to see that the density is integrable by being the sum of Gaussians.

Finally note that for the spectral density we place a scaling factor σ_l at each dimension as can be read in section subsection 3.2.3.

3.1.3 Adaptive regularity

Note that the decay rate of the power spectrum influences the smoothness of the sample paths, or in other words the tails of spectral density in Fourier space determine the regularity of the associated stochastic process. This brings us to the last question we have with regards to the spectral density we employ, namely allowing for adaptive regularity given data (section 2.7) so that the prior sample paths might match some true function better. For stationary processes again this is related to the covariance function $k(0)$ at zero. The properties shown here are for the full GP-prior based model rather than the approximate one.

Suppose μ has a finite second moment:

$$\int \omega^2 \mu(d\omega) < \infty \tag{3.45}$$

Then we have that the existence of a finite second moment is equivalent to having a mean square differentiable process (Stein 1999) (p27). Thus if the spectral measure of the stationary $\mathbf{X}$ on $(\Omega, \mathcal{B}, p)$ has a finite second moment then for all $t \in T \subset \mathbb{R}, s \in \Omega$ we have:

$$\frac{X_{t+h}(s) - X_t(s)}{h} \to X_t'(s) \text{ in } L^2(p) \text{ as } h \to 0 \tag{3.46}$$

The derivative of this stationary process is again stationary and has spectral measure $|\omega|^2\mu(ds)$. It can be seen that this result follows because the spectral representation of the derivative process at the origin is proportional to the second spectral moment. In the same spirit we can look at how the ν_l's determine the regularity of the process associated with the Matern convolution kernel. By of the previous result then:

$$\int_{\mathcal{S}} \omega^{2M} \mu(d\omega) < \infty \tag{3.47}$$

the above is equivalent to the corresponding random process being M times mean square differentiable. For the Matern process $S_l^{mat}(\omega_l)$ has $2M$ moments if and only if $\nu_l > M$. I was unable to find a proof of this in the literature, but we can easily show this:

Ch3. Proposition 2 (Mean Square differentiability if the univariate Matern process)**.** *Let $T, \mathcal{X} \subset \mathbb{R}$ and let $\mathbf{X} = \{X_t, t \in T\}$ be a Matern process mapping values from the probability space $(\Omega, \mathcal{A}, p)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. The spectral representation of its covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is as in Equation 3.42. Then $\mathbf{X}$ is mean square differentiable of order M if and only if $\nu > M$*

Proof. Let $0 < \ell, \nu < \infty$ and $C = \frac{2\pi^{\frac{1}{2}}\Gamma(\nu+\frac{1}{2})(2\nu)^\nu}{\Gamma(\nu)\ell^{2\nu}}$. For the if implication suppose $\nu = M + \delta$ with $\delta > 0$. Note that $\int_{-\infty}^{\infty} \omega^{2M}\mu(d\omega) = \int_{-\infty}^{0} \omega^{2M}\mu(d\omega) + \int_0^{\infty} \omega^{2M}\mu(d\omega)$ and without loss of generality we consider $\int_0^{\infty} \omega^{2M}\mu(d\omega)$. Furthermore note that $C\int_0^1 \omega^{2M}\left(\frac{2\nu}{\ell^2} + 4\pi^2\omega_l^2\right)^{-(\nu+\frac{1}{2})} d\omega$ is clearly finite so we look at $C\int_1^{\infty} \omega^{2M}\left(\frac{2\nu}{\ell^2} + 4\pi^2\omega_l^2\right)^{-(\nu+\frac{1}{2})} d\omega$. Note that:

$$\begin{aligned}
\omega^{2M}\left(\frac{2\nu}{\ell^2} + 4\pi^2\omega^2\right)^{-(\nu+\frac{1}{2})} &\leq \omega^{2M}\left(1 + \omega^2\right)^{-(\nu+\frac{1}{2})} && \text{(3.48a)}\\
&\leq \omega^{2M}\omega^{-2(\nu+\frac{1}{2})} && \text{(3.48b)}\\
&= \omega^{2M-2\nu-1} = \omega^{2M-2M-\delta-1} = \omega^{-\delta-1} && \text{(3.48c)}
\end{aligned}$$

And we have that $\int_1^{\infty} \omega^{-\delta-1} d\omega < \infty$ for all $\delta > 0$, hence by comparison:

$$C\int_1^{\infty} \omega^{2M}\left(\frac{2\nu}{\ell^2} + 4\pi^2\omega^2\right)^{-(\nu+\frac{1}{2})} d\omega < \infty \text{ for } \nu > M \tag{3.49}$$

The same goes for the integral on $(\infty, 0]$. Then by the finiteness of the $2M$-th moment we have that the corresponding process is M times mean square differentiable (Stein 1999) (p27). For the only if implication suppose the process is M times mean square differentiable

which is equivalent to the existence and finiteness of the $2M$-th order derivative of the kernel at the origin (Adler 1981) (p27). Note that the spectral measure of the M-th derivative is $|\omega|^2\mu(d\omega)$ for any finite, symmetric Borel measure on $\mathbb{R}$. Then we have by the spectral representation:

$$\frac{d^M}{dr^M}k(r)\Big|_{r=0} = (2\pi)^{-1}\int_{-\infty}^{\infty} e^{i\omega r}\Big|_{r=0}\omega^{2M}S^{mat}(\omega)d\omega \quad = (2\pi)^{-1}\int_{-\infty}^{\infty}\omega^{2M}S^{mat}(\omega)d\omega < \infty \tag{3.50a}$$

This implies that $\nu > M$ and we can see this by contra-position, suppose than $\nu = M - \delta$ with $\delta \geq 0$ then:

$$\lim_{s\to\infty}\int_0^s \frac{\omega^{2M}}{\left(\frac{2\nu}{\ell^2} + 4\pi^2\omega^2\right)^{(M-\delta+\frac{1}{2})}}d\omega = \infty \tag{3.51}$$

The above follows because, for $M - \delta \geq 1/2$, the highest power in the expansion of the denominator is $2M - 2\delta + 1$ and given that the numerator is of the power $2M$, if we would divide the numerator and the expansion in the denominator by ω^{2M} we would have that the numerator is 1 and the highest power in the denominator is $1 - 2\delta$ which means that integral diverges given that all other powers are lower. If $M - \delta < 1/2$ then we can see that the expression in the integral is larger than $\frac{\omega^{2M}}{(\frac{2\nu}{\ell^2})^{(M-\delta+\frac{1}{2})}+(4\pi^2\omega^2)^{(M-\delta+\frac{1}{2})}}$ and by dividing the numerator and denominator of this expression by ω^{2M} we can see that its integral diverges. The same hold for the integral in Equation 3.51 on the interval $(-\infty, 0]$. □

In higher dimensions we have that the relevant $2M$-th order partial derivative of the covariance function exists if and only if the corresponding random field is mean square differentiable of order M (Adler 1981) (p27). For stationary process processes we only need to look at the partial derivatives of kernel at the origin, for $\mathbf{r} = \mathbf{x}_t - \mathbf{x}_s$:

$$\frac{\partial^{2M}k(\mathbf{r})}{\partial^2_{r_{i_1}}\cdots\partial^2_{r_{i_M}}}\Big|_{\mathbf{r}=0} \tag{3.52}$$

Then the M-th order partial derivative of the process exists as a limit in $L^2(p)$. By the spectral isometry we can also look at the spectral moments:

$$\omega_{i_1,\ldots,i_{d+e}} := \int_{R^{d+e}}\omega_1^{i_1}\cdots\omega_{d+e}^{i_{d+e}}\mu(d\omega) \tag{3.53}$$

Where μ is the measure as in the spectral representation of Bochner's theorem. Note that by stationarity odd-ordered moments are zero in case they exists, that is:

$$\omega_{i_1,\ldots,i_{d+e}} = 0 \text{ when } \sum_{j=1}^{d+e} i_j = 2k+1 \tag{3.54}$$

We are interested in finding some condition on the spectral measure of the convolution kernel that influences the regularity of the corresponding process and this can be found

by noting that having $2M$ partial derivatives of the covariance at the origin is equivalent to the existence of the $2M$-th order spectral moment (Adler 1981) (p31). Then letting $\mathbf{X} = \{X_t, t \in T\}$ be a random field mapping values from the probability space $(\Omega, \mathcal{A}, p)$ to $(\mathcal{X}, \mathcal{B})$, with $T \subset \mathbb{R}^{d+e}$, we have for the variance

$$\mathbb{E}\left[\frac{\partial^{\sum_{i=1}^{M}\alpha_i} X_t}{\partial^{\alpha_1} t_1 \cdots \partial^{\alpha_M} t_{d+e}} \cdot \frac{\partial^{\sum_{j=1}^{K}\beta_i} X_t}{\partial^{\beta_1} t_1 \cdots \partial^{\beta_K} t_{d+e}}\right] \tag{3.55}$$

with $\alpha_i, \beta_i \in \{0, 1, 2, \ldots\}$. We can look at:

$$(-1)^{\sum_{i=1}^{M}\alpha_i} \frac{\partial^{\sum_{i=1}^{M}\alpha_i + \sum_{j=1}^{K}\beta_i}}{\partial^{\alpha_1}\mathbf{r}_1 \cdots \partial^{\alpha_M}\mathbf{r}_{d+e} \cdots \partial^{\beta_1}\mathbf{r}_1 \cdots \partial^{\beta_K}\mathbf{r}_{d+e}} k(\mathbf{r})\Big|_{\mathbf{r}=0} \tag{3.56}$$

which exists if and only if (Adler 1981) (p31):

$$\int_{\mathbb{R}^{d+e}} \omega_1^{\alpha_1+\beta_1} \cdots \omega_{d+e}^{\alpha_M+\beta_K} < \infty \mu(d\omega) \tag{3.57}$$

Note again that odd numbered spectral moments and odd-ordered derivatives of the kernel are identically zero. The formulation of course depends on the relation between the orders d, e, M, K. For our purposes we let $M = K$ and $\alpha_i, \beta_i = 1$, and $d + e$ can be either larger of smaller than $2M$.

Ch3. Theorem 5 (Mean square differentiability of the Matern process with the convolution kernel)**.** *Let $T, \mathcal{X} \subset \mathbb{R}^{d+e}$ and let $\mathbf{X} = \{X_t, t \in T\}$ be a Matern process mapping values from the probability space $(\Omega, \mathcal{A}, p)$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let The spectral representation of its covariance function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be:*

$$k(\mathbf{r}) = (2\pi)^{-d+e} \int_{\mathbb{R}^{d+e}} e^{i\omega'\mathbf{r}} S^{mat}(\omega) d\omega = (2\pi)^{-d+e} \int_{\mathbb{R}^{d+e}} e^{i\omega'\mathbf{r}} \prod_{l=1}^{d+e} S_l^{mat}(\omega_l) d\omega \tag{3.58}$$

Where S_l is as in Equation 3.42, then if $d + e \leq 2M$ the process is mean square differentiable of order M if $\nu_l > M$ for $l = 1, \cdots, d + e$. If $d + e > M$ then the process is mean square differentiable of order M if, for ν_{i_j} corresponding to the degrees of freedom of S_{i_j} as in Equation 3.42 and i_j corresponding to the index of the partial derivatives as in Equation 3.52, we have $\nu_{i_j} > 1$ for $j = 1, \ldots, M$.

Proof. We look at the spectral moments, then if $d + e \leq 2M$, by Equation 3.57:

$$\int_{\mathbb{R}^{d+e}} \omega_1^{2M} \cdots \omega_{d+e}^{2M} \mu_{d\omega} = \int_{\mathbb{R}^{d+e}} \omega_1^{2M} \cdots \omega_{d+e}^{2M} S^{mat}(\omega) d\omega \tag{3.59a}$$

$$\overset{\text{by proposition 2}}{=} \int_{\mathbb{R}} \omega_1^{2M} S_1^{mat}(\omega_1) \cdots \int_{\mathbb{R}} \omega_{d+e}^{2M} S_{d+e}^{mat}(\omega_{d+e}) d\omega_1 \cdots d\omega_{d+e} < \infty \tag{3.59b}$$

$$\text{if } \nu_l > M \text{ for } l = 1 \ldots, d + e \tag{3.59c}$$

If $d+e > 2M$, then by Equation 3.52 and Equation 3.57:

$$\int_{\mathbb{R}^{d+e}} \omega_{i_1}^2 \cdots \omega_{i_M}^2 \mu_{d\omega} = \int_{\mathbb{R}^{d+e}} \omega_{i_1}^2 \cdots \omega_{i_M}^2 S^{mat}(\omega) d\omega \tag{3.60a}$$

$$= \int_{\mathbb{R}} \cdots \int_{\mathbb{R}} \omega_{i_1}^2 \cdots \omega_{i_M}^2 \prod_{l=1}^{d+e} S_l^{mat}(\omega_l) d\omega_{i_1} \cdots d\omega_{d+e} \tag{3.60b}$$

$$\overset{\text{by proposition 2}}{=} \int_{\mathbb{R}} \omega_{i_1}^2 S_{i_1}^{mat}(\omega_{i_1}) \cdots \int_{\mathbb{R}} \omega_{i_M}^2 S_{i_M}^{mat}(\omega_{i_M}) \tag{3.60c}$$

$$\cdots \int_{\mathbb{R}} S_{d+e}^{mat}(\omega_{d+e}) d\omega_{i_1} \cdots d\omega_{i_M} \cdots d\omega_{d+e} < \infty \tag{3.60d}$$

$$\text{if } \nu_{i_j} > 1 \text{ for } j = 1 \ldots, M \tag{3.60e}$$

Then we have that the finiteness of the spectral moments is equivalent to having $2M$ partial derivatives of the kernel at the origin which in turn is equivalent to the process being mean square differentiable of order M as in Equation 3.52, Equation 3.57, Equation 3.56 □

Thus in short we can let the regularity be adaptive in the convolution kernel, for the Matern process, by letting the ν_l's be learned from the data (see section 2.7 as to why), thereby hopefully allowing for the credible regions of the posterior we define to have better frequentist coverage.

For the spectral mixture we did not have time to perform a similar analysis. But we do note that the tails of a Gaussian mixture is clearly adaptable depending on the mixture, hence so is the regularity of the corresponding prior. For example A. G. Wilson and Adams (2013) show how the heavy tails of a rational quadratic in Fourier space are modelled by the Gaussian mixture, with two components with large periods (i.e. large $\frac{1}{\sigma_i^2}$) and two with short to large length-scales $\frac{1}{\ell_i}$ respectively.

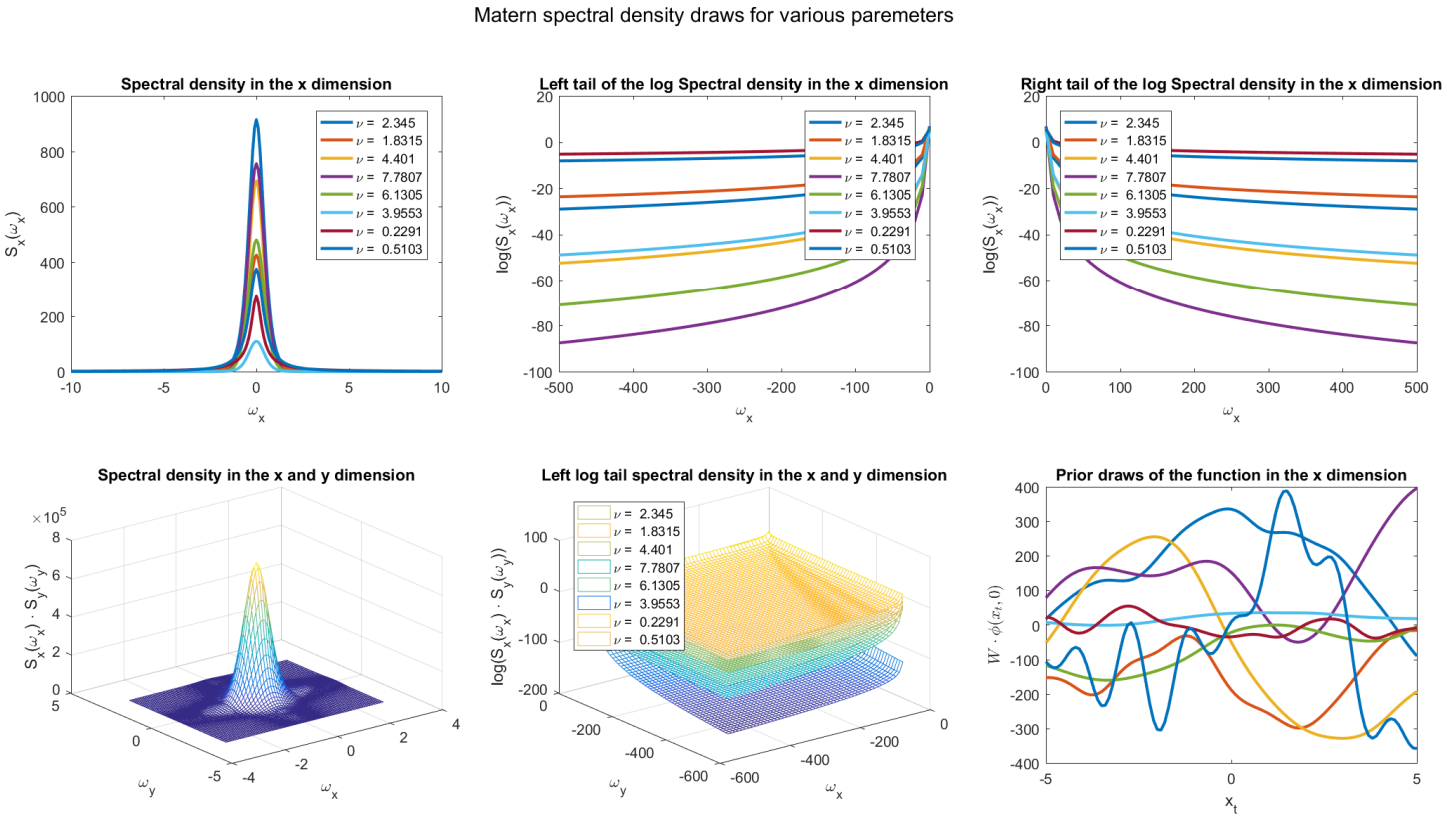
Hyperpriors: I experimented with various hyper-priors for the Matern spectral density parameters. For the Matern scaling parameters σ_l I considered the following weakly informative and truncated priors: Half-student-t (3 <deg. freedom < 7), Half-Cauchy, and the Half-Normal prior in the setting of section section 6.2. Following Gelman and Hill (2006) I avoided the use of a inverse Gamma prior on this parameter, because it might might place an unnecessary amount of probability mass over values that are outside of a plausible posterior, thereby skew it. In the end the truncated prior $\sigma_l \sim \mathcal{HN}(\sigma_l|0, 50)$ showed the best results, although it is not robust. It might be that the Cauchy tails are too thick, so that when the data is not informative enough posterior sampling becomes troublesome. For the mixture weights, a_i that correspond to the signal variance, of the Gaussian mixture spectral density we did the same thing. On the length-scales we placed an informative prior $\ell \sim \mathcal{N}(\ell|10, 1)$, because this is a hard to identify parameter (see evaluation of the blocked Gibbs sampler section). For the inverse length-scales of the spectral mixture components, $\frac{1}{\sigma_i^2}$ we did the same thing. For the regularity parameter ν after experimenting

with various weakly informative priors, such as the Half-Cauchy and the Half-Normal, we found an exponential prior with a rate of 10 to work best in the setting of section section 6.2.

For the frequencies of the spectral mixture kernel I followed, in an ad-hoc fashion, the hierarchical example of of A. G. Wilson (2014), and placed a normal prior on the locations of the mixture components. The mean vector of the prior is a uniformly distributed vector on the interval $[4, 200]$ and the covariance matrix is diagonal with prior-$\sigma_i^2 = 20$.

In Figure 3.1 we see various prior draws of the Matern product spectral density and see the effect the regularity parameter has on the tails of the spectral density.

Figure 3.1: Prior draws of the spectral density and prior draws of the cross section of the state-function



3.2 Estimation and inference in the approximate model

Given the prior model of 3.20 learning constitutes finding a joint posterior over the hyper-parameters $\boldsymbol{\theta}_f, \mathbf{Q}$, the weights $\mathbf{W}$ and the hidden states $\mathbf{x}_{1:T}$ where we know $\mathbf{y}_{1:T}$ and $\mathbf{x}_0$ can be fixed for simplicity. As Frigola, Lindsten, Schön, and C. Rasmussen (2013) a blocked Gibbs scheme is employed for this purpose, however Svensson, Solin, et al. (2015) rely, instead of any marginalisation, only on data augmentation (Tanner and Wong 1987). For Y. Wu et al. (2014) and Frigola, Lindsten, Schön, and C. Rasmussen (2013) marginalisation alleviates the burden of needing to obtain a posterior over the function values as well as the

unknown states in the unsupervised learning setting. However data augmentation is in part made possible by the linear structure of the approximate model, where the state equation 3.20 has the form of a linear time-invariant (LTI) SSM much like the one considered by Wills et al. (2012). Note also that the blocked Gibbs sampler alleviates the problem in the face of strong dependence between the variables of the marginal draws of the vanilla Gibbs algorithm, which is inevitable given the chain structure of the states for example.

Algorithm 3.1 Blocked Gibbs Scheme

Input: $\mathbf{y}_{1:T}, \mathbf{u}_{1:T}, \phi(\cdot), \mathbb{P}_y(\cdot), \mathbb{P}_Q(\cdot), \mathbb{P}_{\theta_f}(\cdot), \mathbb{P}_W(\cdot)$
Initialize: $\mathbf{x}_{0:T}[0], \mathbf{W}[0], \mathbf{Q}[0], \boldsymbol{\theta}_f[0]$
  for $k \in \{0, 1, \ldots, K\}$ **do**
    Sample $\mathbf{x}_{0:T}[k+1]|\mathbf{W}[k], \mathbf{Q}[k], \boldsymbol{\theta}_f[k]$ *(PGAS Markov kernel)*
    Sample $\mathbf{Q}[k+1]|\mathbf{x}_{0:T}[k+1], \mathbf{W}[k], \boldsymbol{\theta}_f[k]$ *(Equation 3.74c)*
    Sample $\mathbf{W}[k+1]|\mathbf{Q}[k+1], \mathbf{x}_{0:T}[k+1], \boldsymbol{\theta}_f[k]$ *(Equation 3.74b)*
    Sample $\boldsymbol{\theta}_f[k+1]|\mathbf{W}[k+1], \mathbf{Q}[k+1], \mathbf{x}_{0:T}[k+1]$ *(Metropolis-within-Gibbs)*
  end for
Output: K Samples form the joint posterior $\mathbb{P}(\mathbf{x}_{0:T}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f|\mathbf{y}_{1:T})$
Note that when comparing models, the seed is the same for all models and the Mersenne-Twister algorithm is used for stream generation.

3.2.1 Sequential Monte Carlo (Particle Filter)

Given $\mathbf{W}[k], \mathbf{Q}[k], \boldsymbol{\theta}_f[k]$ the distribution of interest is $\mathbb{P}(\mathbf{x}_{0:T}[k+1]|\mathbf{y}_{1:T})$ where we omit explicitly conditioning on the weights and hyper-parameters. First we give a quick review of the Sequential Monte Carlo sampler, also referred to as the Particle Filter (PF) and then move on to the PGAS Markov kernel. As explained in chapter 12 of the book of Durbin and Koopman (2012) in the most basic form of the SMC we apply importance sampling (M. N. Rosenbluth and A. W. Rosenbluth 1955) sequentially at each time step while retaining the previous selection of $\mathbf{x}^i_{1:t-1}$ as well as making sure that the new draw is consistent with the importance density over the entire trajectory. In a sense one can think of SMC as a simulation based Kalman Filter for more general SSM's. There are many references on importance sampling, both practical guides ((Durbin and Koopman 2012) (Rubinstein and Kroese 2011)) as well those containing more theoretical background (Glynn and Iglehart 1989) and these can be consulted for an in depth treatment of the topic. The distribution of interest:

$$\mathbb{P}(\mathbf{x}_{0:T}|\mathbf{y}_{1:T}) = \frac{\mathbb{P}(\mathbf{x}_{0:T}, \mathbf{y}_{1:T})}{\int_{\Omega^{T+1}} \mathbb{P}(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}) d\mathbf{x}_{0:T}} \propto \mathbb{P}(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}) \tag{3.61a}$$

$$\mathbb{P}(\mathbf{x}_{0:T}, \mathbf{y}_{1:T}) = \mathbb{P}_0(\mathbf{x}_0) \prod_{t=0}^{T} \mathbb{P}(\mathbf{x}_{t+1}|\mathbf{x}_t) \prod_{t=1}^{T} \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t) \tag{3.61b}$$

$$= \mathbb{P}_0(\mathbf{x}_0) \prod_{t=0}^{T} \mathcal{N}(\mathbf{x}_{t+1}|\mathbf{f}_{t+1}, \mathbf{Q}) \prod_{t=1}^{T} \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t) \tag{3.61c}$$

Where the joint distribution in Equation 3.61c is visibly not Gaussian. In filtering terms for each $t \in \{1, \ldots, T\}$ the target density is $\mathbb{P}(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$. The point-mass approximation (δ_x is the Dirac measure) has the form:

$$\hat{\mathbb{P}}(d\mathbf{x}_{0:t}|\mathbf{y}_{1:t}) = \sum_{i=1}^{N} \frac{w_t^i}{\sum_{s=1}^{N} w_t^s} \delta_{\mathbf{x}_{0:t}^i}(d\mathbf{x}_{0:t}) \tag{3.62}$$

Where $\mathbb{P}(d\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ is the distribution defined on the measurable space $(\Omega^{t+1}, \mathcal{B}^{t+1})$ admitting the density $\mathbb{P}(\mathbf{x}_{0:t}|\mathbf{y}_{1:t})$ with respect to the σ-finite dominating measure $d\mathbf{x}_{0:t}$. Note that Ω is a compact subset of $\mathbb{R}^d$ and it can have the usual Borel σ-algebra and Ω^{t+1} its. The weights $\{w_t^i\}_{i=1}^N$ and the *particles* $\{\mathbf{x}_t^i\}_{i=1}^N$ represent a so-called weighted particle system $\{w_t^i, \mathbf{x}_{0:t}\}_{i=1}^N$. The particles represent point-masses in the state-space and each particle is a hypothesized state with the weights representing a probability on that hypothesis. The particle system is carried to time step $t+1$ by sampling $\{a_t^i, \mathbf{x}_{t+1}^i\}_{i=1}^N$ from the proposal transition mapping:

$$M(a_t, \mathbf{x}_{t+1}|\{w_t^i, \mathbf{x}_{0:t}^i\}_{i=1}^N) = \frac{w_t^{a_t}}{\sum_{s=1}^{N} w_t^s} r(\mathbf{x}_{t+1}|\mathbf{x}_{0:t}^{a_t}) \tag{3.63a}$$

$$r(\mathbf{x}_{t+1}^i|\mathbf{x}_{0:t}^{a_t^i}) \sim \mathcal{N}(\mathbf{f}_{t+1}^{a_t^i}, \mathbf{Q}) \tag{3.63b}$$

$$\mathbf{f}_{t+1}^{a_t^i} = \mathbf{W}\boldsymbol{\phi}(\mathbf{x}_t^{a_t^i}, \mathbf{u}_t) \tag{3.63c}$$

Where $r(\mathbf{x}_{t+1}|\mathbf{x}_{0:t}^{a_t})$ is the proposal density, which we take to be Gaussian centered around the GP prediction at time $t+1$ with variance $\mathbf{Q}$. In the above re-sampling then is performed through the ancestral indices written $\{a_t^i\}_{i=1}^N$ of the particles $\{\mathbf{x}_{t+1}^i\}_{i=1}^N$, where an ancestor refers to the particle on which $\mathbf{x}_{t+1}^i$ was conditioned at its inception. Then including these ancestral paths the trajectory of a particle $\mathbf{x}_{t+1}^i$ can be defined as $\mathbf{x}_{0:t+1}^i = (\mathbf{x}_{0:t}^{a_t^i}, \mathbf{x}_{t+1}^i)$. The weights are computed according to:

$$w_{t+1}^i = \frac{\mathbb{P}(\mathbf{x}_{0:t+1}^i, \mathbf{y}_{1:t+1})}{\mathbb{P}(\mathbf{x}_{0:t}^i, \mathbf{y}_{1:t}) r(\mathbf{x}_{t+1}^i|(\mathbf{x}_{0:t}^{a_t^i})} \tag{3.64a}$$

$$\overset{\text{By Equation 3.61c}}{=} \frac{\mathcal{N}(\mathbf{x}_{t+1}^i|\mathbf{f}_{t+1}^{a_t^i}, \mathbf{Q})\mathbb{P}_y(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^i)}{r(\mathbf{x}_{t+1}^i|\mathbf{x}_{0:t}^{a_t^i})} \tag{3.64b}$$

$$\overset{\text{By Equation 3.63b}}{=} \mathbb{P}_y(\mathbf{y}_{t+1}|\mathbf{x}_{t+1}^i) \tag{3.64c}$$

Where as mentioned before we simply initialize with $\mathbf{x}_0^i = 0$ or $\mathbf{x}_0^i = \log(\widehat{\mathbb{V}ar(y_t)})$ for practical ease and we have the $\mathbb{P}_y(\mathbf{y}_{t+1}|\mathbf{x}_{t+1})$ from the choice of the model. With all the necessary ingredients of the SMC sampler in place the PGAS algorithm can be set up, which also addresses the poor mixing property of the Particle Gibbs (PG) (Andrieu et al. 2010) when we inevitably are faced with path degeneracy in the SMC (Chopin and Singh 2015). An in depth review of the SMC is given by Schön et al. (2015), and for econometrics by Durbin and Koopman (2012).

3.2.2 PGAS Markov Kernel

Three main ingredients define the PGAS Markov Kernel, of which the first two involve the use of a reference trajectory, which in our case is the trajectory drawn in the previous iteration of the Blocked Gibbs scheme $\mathbf{x}_{0:T}[k]$. Firstly after an iteration of the SMC within the PG(AS) J is sampled with $J \sim \mathbb{P}(J = i) \propto w_T^i$ and $\mathbf{x}_{0:T}[k+1] = \mathbf{x}_{0:T}^J$ mapping $\mathbf{x}_{0:T}[k]$ to $\mathbb{P}(\mathbf{x}_{0:T}[k+1] = \mathbf{x}_{0:T}^i)$ on Ω^T. In this sense then it defines a Markov kernel on $(\Omega^T, \mathcal{B}^T, \mu)$. Secondly whereas in the PF $\{a_t^i, \mathbf{x}_{t+1}^i\}_{i=1}^N$ is drawn from the transition mapping M independently, in the PGAS these samples are drawn conditioned on keeping the reference trajectory throughout the algorithm. To this end sampling is performed from the proposal transition mapping M for $i \in \{1, \ldots, N-1\}$ and $\mathbf{x}_t^N$ is set to $\mathbf{x}_t[k]$. Then by theorem 5 of Andrieu et al. (2010) for any $N \geq 1$ the PG(AS) has as its stationary distribution the target smoothing distribution of Equation 3.61a. These two features also make it so that the SMC in the PGAS is often referred to as a conditional particle filter (CPF).

But of course the number of iterations needed to get close to this stationary distribution depends on the mixing properties, which are in general poor in the face of path degeneracy. This is where the third ingredient, the AS step, in the PGAS comes in, where a new value for the ancestral index variable a_t^N is sampled. At $t \geq 1$ the aim is to assign an ancestral trajectory to to the path $\mathbf{x}_{t+1:T}[k]$ through the index connected to the particles $\{\mathbf{x}_{0:t}^i\}_{i=1}^N$. This is done by sampling a_t^N with a probability mass proportional to the importance weight ${\tilde{w}_t}^i$ that connects $\mathbf{x}_{t+1:T}[k]$ to a particle $\mathbf{x}_{0:t}^i$, where by theorem 1 of Lindsten et al. (2014) we have:

$$
\begin{aligned}
{\tilde{w}_t}^i &:= w_t^i \frac{\mathbb{P}([\mathbf{x}_{0:t}^i, \mathbf{x}_{t+1:T}[k]], \mathbf{y}_{1:T})}{\mathbb{P}(\mathbf{x}_{0:t}^i, \mathbf{y}_{1:t})} && \text{(3.65a)} \\
&= w_t^i \mathbb{P}(\mathbf{x}_{t+1:T}[k], \mathbf{y}_{t+1:T} | \mathbf{x}_t^i) \propto w_t^i \mathcal{N}(\mathbf{x}_{t+1}[k] | \mathbf{f}_{t+1}^i, \mathbf{Q}) && \text{(3.65b)} \\
\mathbf{f}_{t+1}^i &= \mathbf{W} \boldsymbol{\phi}(\mathbf{x}_t^i, \mathbf{u}_t) && \text{(3.65c)}
\end{aligned}
$$

To understand the importance weights from a Bayesian perspective we can view the importance weight w_t^i as the prior on the particle $\mathbf{x}_t^i$ and $\mathcal{N}(\mathbf{x}_{t+1}[k] | \mathbf{f}_{t+1}^i, \mathbf{Q})$ as the likelihood of obtaining $\mathbf{x}_{t+1}[k]$ given $\mathbf{x}_t^i$. Then the product is proportional to the posterior probability that the history of $\mathbf{x}_{t+1}[k]$ was $\mathbf{x}_t^i$. Then the PGAS Markov Kernel is given by algorithm 3.2. Note that the convergence results given by Lindsten et al. (2014) do not depend on the asymptotic behavior of the SMC sampler. Note also that the computational complexity of this step is $\mathcal{O}(NT)$ thus only considering this step in the blocked Gibbs scheme the computational complexity in the approximate model is $\mathcal{O}(NmT^2)$, which is good gain compared to the $\mathcal{O}(NT^4)$. Furthermore this is only for one learning step in a rolling window application for example the gain is $\mathcal{O}(NT^5)$ to $\mathcal{O}(NmT^3)$. The batch nature of this sampling algorithm is undesirable and methods for allowing on-line inference are developed, SONIG (Bijl et al. 2016) for example, but most depend on the Nyström eigendecomposition and variational methods which require more data for estimation.

Algorithm 3.2 Particle Gibbs with Ancestral Sampling Markov Kernel

Input: $\mathbf{y}_{1:T}, \mathbf{u}_{1:T}, \phi(\cdot), N, \mathbf{x}_{0:T}[k]$
  $\{\mathbf{x}_0^i\}_{i=1}^{N-1} = \log(\widehat{\mathbb{V}ar(y_t)})$
  $\mathbf{x}_0^N = \mathbf{x}_0[k]$
  for $t \in \{1, 2, \ldots, T\}$ **do**
    Set $\{w_t^i\}_{i=1}^N$ according Equation 3.64c
    Sample $\{a_t^i, \mathbf{x}_{t+1}^i\}_{i=1}^{N-1} \sim M(a_t, \mathbf{x}_{t+1} | \{w_t^i, \mathbf{x}_{0:t}^i\}_{i=1}^{N-1})$
    Set $\mathbf{x}_{t+1}^N = \mathbf{x}_{t+1}[k]$
    Sample a_t^N with $\mathbb{P}(a_t^N = j) \propto \tilde{w}_t{}^j \propto w_t^j \mathcal{N}(\mathbf{x}_{t+1}^N | \mathbf{f}_{t+1}^j, \mathbf{Q})$
    Set $\{\mathbf{x}_{0:t+1}^i\}_{i=1}^N = \{(\mathbf{x}_{0:t}^{a_t^i}, \mathbf{x}_{t+1}^i)\}_{i=1}^N$
  end for
  Sample $J \sim \mathbb{P}(J = i) \propto w_T^i$ and set $\mathbf{x}_{0:T}[k+1] = \mathbf{x}_{0:T}^J$
Output: $\mathbf{x}_{0:T}[k+1]$

A Note on the marginal distribution $\mathbb{P}(\mathbf{y}_{1:T})$. As in the PF within the PGAS the unnormalized weights of Equation 3.64c can be used to approximate the marginal likelihood of the observed data $\mathbb{P}(\mathbf{y}_{1:T}|\mathcal{M})$ given the model, where $\mathcal{M}$ includes design choices such as hyper-prior choices and the order of the expansion m. In particular the marginal distribution is of help when comparing various expansion orders m of the model in Equation 3.20. In this paragraph, for clarity, we do explicitly condition on the prior weights, the variance of the idiosyncratic state errors and the hyper-prior parameters where we let $(\mathbf{W}[k], \mathbf{Q}[k], \boldsymbol{\theta}_f[k]) := \Psi[k]$, however we omit conditioning on the model for notational ease. Thus in the Blocked Gibbs scheme at each iteration $k \in \{1, \ldots, K\}$ we first aim to obtain an estimate of $\mathbb{P}(\mathbf{y}_{1:T}|\Psi[k])$ with the states integrated out. We can write:

$$\mathbb{P}(\mathbf{y}_{1:T}|\Psi[k]) = \prod_{t=1}^{T} \mathbb{P}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \Psi[k]) \tag{3.66a}$$

$$\mathbb{P}(\mathbf{y}_t|\mathbf{y}_{1:t-1}, \Psi[k]) = \int_\Omega \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t, \Psi[k]) \mathbb{P}(\mathbf{x}_t|\mathbf{y}_{0:t-1}, \Psi[k]) d\mathbf{x}_t \tag{3.66b}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t^i, \Psi[K]) \tag{3.66c}$$

$$\overset{\text{Equation 3.64c}}{=} \frac{1}{N} \sum_{i=1}^{N} w_t^i \tag{3.66d}$$

Where we can obtain w_t^i from the PGAS kernel after re-sampling and before drawing $J \sim \mathbb{P}(J = i)$. Thus we finally get:

$$\hat{\mathbb{P}}(\mathbf{y}_{1:T}|\Psi[k]) = \prod_{t=1}^{T} \left(\frac{1}{N} \sum_{i=1}^{N} w_t^i \right) \tag{3.67a}$$

$$\log \hat{\mathbb{P}}(\mathbf{y}_{1:T}|\Psi[k]) = \sum_{t=1}^{T} \log \left(\frac{1}{N} \sum_{i=1}^{N} w_t^i \right) \tag{3.67b}$$

The next step the is to integrate out $\Psi[k]$, which is possible naively by just taking the average of the likelihoods for each draw of $\Psi[k]$. However, we follow Gelfand and Dey (1994) and estimate the reciprocal of the marginal distribution as follows:

$$\hat{\mathbb{P}}(\mathbf{y}_{1:T})^{-1} = \frac{1}{K}\sum_{k=1}^{K}\frac{g(\Psi[k])}{\hat{\mathbb{P}}(\mathbf{y}_{1:T}|\Psi[k])\mathbb{P}(\Psi[k])} \tag{3.68a}$$

Where $\mathbb{P}(\Psi[k])$ is the joint prior on the weights and hyper-parameters, which is available analytically. For the implementation of $g(\Psi[k])$ we follow Gelfand and Dey (1994) exactly.

3.2.3 Sampling Q and W

Given $\mathbf{x}_{0:T}[k+1]$ then the next steps in the Blocked Gibbs sampler is obtaining the samples $\mathbf{Q}[k+1]|\mathbf{x}_{0:T}[k+1], \mathbf{W}[k], \boldsymbol{\theta}_f[k]$ and $\mathbf{W}[k+1]|\mathbf{Q}[k+1], \mathbf{x}_{0:T}[k+1], \boldsymbol{\theta}_f[k]$. For this we employ the Matrix Normal Inverse Wishart distribution as a joint prior on the weights and the state error covariances, because of the conjugacy due to the model in Equation 3.20 being Gaussian and linear in the basis functions (Wills et al. 2012). The posterior $\mathbb{P}(\mathbf{Q}, \mathbf{W}|\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}, \boldsymbol{\theta}_f[k])$ is proportional to $\mathbb{P}(\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}|\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f[k])\mathbb{P}(\mathbf{Q}, \mathbf{W}|\boldsymbol{\theta}_f[k])$ (Bayes Thm.). Omitting conditioning on $\boldsymbol{\theta}_f$ the hierarchical prior assumptions on the weights and the state error covariances can be written as:

$$\begin{aligned}
\mathbb{P}(\mathbf{Q}, \mathbf{W}) &= \mathbb{P}(\mathbf{Q})\mathbb{P}(\mathbf{W}|\mathbf{Q}) && (3.69a)\\
&\equiv \mathcal{IW}(\mathbf{Q}|\ell, \boldsymbol{\Lambda})\mathcal{MN}(\mathbf{W}|\mathbf{M}, \mathbf{Q}, \mathbf{V}) && (3.69b)\\
&= \frac{|\boldsymbol{\Lambda}|^{\ell/2}|\mathbf{Q}|^{-(d+\ell+1)/2}}{2^{d\ell/2}\Gamma_d(\ell/2)}\exp\left(-\frac{1}{2}\mathrm{tr}(\mathbf{Q}^{-1}\boldsymbol{\Lambda})\right) && (3.69c)\\
&\times \frac{|\mathbf{V}|^{d/2}}{(2\pi)^{dm}|\mathbf{Q}|^{m/2}}\exp\left(-\frac{1}{2}\mathrm{tr}\left[(\mathbf{W}-\mathbf{M})'\mathbf{Q}^{-1}(\mathbf{W}-\mathbf{M})\mathbf{V}\right]\right)
\end{aligned}$$

And

$$\begin{aligned}
-2\log\mathbb{P}(\mathbf{Q}, \mathbf{W}) &= -2\log\mathbb{P}(\mathbf{Q}) - 2\log\mathbb{P}(\mathbf{W}|\mathbf{Q}) && (3.70a)\\
&\propto (d+\ell+m+1)\log|\mathbf{Q}|\\
&+ \mathrm{tr}\left(\mathbf{Q}^{-1}(\boldsymbol{\Lambda}+\mathbf{W}\mathbf{V}\mathbf{W}')\right) && (3.70b)
\end{aligned}$$

Where we let $\mathbf{M} = 0$. For the Inverse Wishart distribution, preliminary experimentation with various parameter values for the degrees of freedom ℓ and the positive definite scale matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{d\times m}$ in a simulation setting (see section 6.2) has shown $\boldsymbol{\Lambda} = \mathrm{diag}([1, \ldots, 1]), \ell = 3$ to work best in most trials. The aim is to keep the prior somewhat informative by positioning the scale matrix towards not too high valued covariances in the parameter space. At the same time the prior is not too informative by setting the degrees of freedom low. A problem here

could be the dependence between the variances and the covariances, which can exaggerate correlation when variances are high, therefore for higher dimensional states (deep learning application) we set the degrees of freedom higher to say 50 for correlation shrinkage. Other priors might work better but the conjugacy property is what makes this a will-do prior. For the Matrix Normal prior the row covariance matrix $\mathbf{V}^{-1}$ links the spectral density of the kernel corresponding to that of the GP prior in Equation 3.20 where it is set to:

$$\mathbf{V}^{-1} = \text{diag}(\{S_{\boldsymbol{\theta}_f}(\lambda_i)\}_{i=1}^{m})^{-1} \tag{3.71}$$

Note that as can be seen in Equation 3.69c the marginal variances of the weights are scaled by both the scaling factors in the spectral densities within $\mathbf{V}^{-1}$ but also by $\mathbf{Q}$. Similar to the scaling factor in the spectral densities the scaling by $\mathbf{Q}$ is constant along the rows, therefore it is possible to use an overall scaling factor that can determine the influence of the prior (i.e. the amount of L^2 regularization). Preliminary experimentation did show that, when in the multi dimensional case, allowing for varying scaling factors for the product of spectral densities over the compact product space $\Omega = [-L_1, L_1] \times \cdots \times [-L_{d+e}, L_{d+e}]$ performed better in terms of RMSE and predictive log likelihood in a simulation setting.

Then we have an expression for the prior distribution in the unnormalized posterior $\mathbb{P}(\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}|\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f[k])\mathbb{P}(\mathbf{Q}, \mathbf{W}|\boldsymbol{\theta}_f[k])$ and as for the conditional joint likelihood it can be noticed that $\mathbb{P}_y(\mathbf{y}_t|\mathbf{x}_t)$ in equation Equation 3.61c does not depend on either $\mathbf{Q}$ or $\mathbf{W}$. Therefore the expression $\mathbb{P}(\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}|\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f[k])$ is itself proportional to $\mathbb{P}(\mathbf{x}_{0:T}[k+1]|\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f[k])$ then again without conditioning on the spectral density parameters we get (M West and Harrison 1997):

$$\begin{aligned}
-2\log\mathbb{P}(\mathbf{x}_{0:T}[k+1]|\mathbf{Q}, \mathbf{W}) &= T\log|\mathbf{Q}| && \text{(3.72a)}\\
&\quad + \text{tr}\left(\mathbf{Q}^{-1}(\Phi - \mathbf{W}\mathbf{P}' - \mathbf{P}\mathbf{W}' + \mathbf{W}\Sigma\mathbf{W}')\right)\\
\Phi &= \sum_{t=0}^{T} \mathbf{x}_{t+1}[k+1]\mathbf{x}_{t+1}[k+1]' && \text{(3.72b)}\\
\mathbf{P} &= \sum_{t=0}^{T} \mathbf{x}_{t+1}[k+1]\boldsymbol{\phi}(\mathbf{x}_t[k+1], \mathbf{u}_t)' && \text{(3.72c)}\\
\Sigma &= \sum_{t=0}^{T} \boldsymbol{\phi}(\mathbf{x}_t[k+1], \mathbf{u}_t)\boldsymbol{\phi}(\mathbf{x}_t, \mathbf{u}_t)' && \text{(3.72d)}
\end{aligned}$$

Then in combination with the prior in Equation 3.70a we get the unnormalized log-posterior:

$$\begin{aligned}
-2\log\mathbb{P}(\mathbf{Q}, \mathbf{W}|\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}) &\propto -2\log(\mathbb{P}(\mathbf{x}_{0:T}[k+1]|\mathbf{Q}, \mathbf{W})\mathbb{P}(\mathbf{Q}, \mathbf{W}) && \text{(3.73a)}\\
&\propto (d+\ell+m+1)\log|\mathbf{Q}|\\
&\quad + \text{tr}[\mathbf{Q}^{-1}(\boldsymbol{\Lambda} + \Phi - \mathbf{P}(\Sigma+\mathbf{V})^{-1}\mathbf{P}'\\
&\quad + (\mathbf{W} - \mathbf{P}(\Sigma+\mathbf{V})^{-1})\mathbf{Q}^{-1}(\mathbf{W} - \mathbf{P}(\Sigma+\mathbf{V})^{-1})')]
\end{aligned}$$

Equation 3.73a can be recognized to also be a $\mathcal{MNIW}$ distribution (Wills et al. 2012) and we get:

$$\mathbb{P}(\mathbf{W}, \mathbf{Q}|\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}) \equiv \mathcal{MNIW}(\mathbf{Q}, \mathbf{W}|\mathbf{P}(\Sigma+\mathbf{V})^{-1}, \mathbf{Q}, (\Sigma+\mathbf{V})^{-1}, \\ \mathbf{\Lambda}+\Phi-\mathbf{P}(\Sigma+\mathbf{V})^{-1}\mathbf{P}', \ell+T) \tag{3.74a}$$

$$\mathbb{P}(\mathbf{W}[k+1]|\mathbf{Q}[k+1], \mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}) \equiv \mathcal{MN}(\mathbf{W}|\mathbf{P}(\Sigma+\mathbf{V})^{-1}, \mathbf{Q}, (\Sigma+\mathbf{V})^{-1}) \tag{3.74b}$$

$$\mathbb{P}(\mathbf{Q}[k+1]|\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}) \equiv \mathcal{IW}(\mathbf{Q}|\ell+T, \mathbf{\Lambda}+\Phi-\mathbf{P}(\Sigma+\mathbf{V})^{-1}\mathbf{P}') \tag{3.74c}$$

Note that in most of the above explicitly conditioning on the inputs has also been omitted to the relief notational burden.

3.2.4 Sampling the spectral density parameters

Following Svensson, Solin, et al. (2015) we employ a random walk Metropolis Hastings algorithm for sampling from the posterior over the spectral density parameters resulting in a so called Metropolis-within-Gibbs step. The posterior over the spectral density parameters is:

$$\mathbb{P}(\boldsymbol{\theta}_f|\mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}, \mathbf{Q}[k+1], \mathbf{W}[k+1]) \tag{3.75a}$$

$$= \mathbb{P}(\boldsymbol{\theta}_f, \mathbf{x}_{0:T}[k+1], \mathbf{Q}[k+1], \mathbf{W}[k+1]|\mathbf{y}_{1:T}) \times C \tag{3.75b}$$

$$\propto \mathbb{P}_{\boldsymbol{\theta}_f}(\boldsymbol{\theta}_f)\mathbb{P}(\mathbf{Q}[k+1]|\mathbf{x}_{0:T}[k+1], \boldsymbol{\theta}_f, \mathbf{y}_{1:T}) \tag{3.75c}$$

$$\times \mathbb{P}(\mathbf{W}[k+1]|\mathbf{Q}[k+1], \mathbf{x}_{0:T}[k+1], \boldsymbol{\theta}_f, \mathbf{y}_{1:T}) \tag{3.75d}$$

Where all the needed densities are analytically available. Then we draw the proposal parameter $\boldsymbol{\theta}_f^*$ from $q(\boldsymbol{\theta}_f^*|\boldsymbol{\theta}_f[k])$ where the proposal density is just a random walk $AR(1)$. We can then either accept or reject the proposal with probability:

$$\alpha = \min\left(1, \frac{\mathbb{P}(\boldsymbol{\theta}_f^*, \mathbf{x}_{0:T}[k+1], \mathbf{Q}[k+1], \mathbf{W}[k+1]|\mathbf{y}_{1:T})q(\boldsymbol{\theta}_f[k]|\boldsymbol{\theta}_f[k])}{\mathbb{P}(\boldsymbol{\theta}_f[k], \mathbf{x}_{0:T}[k+1], \mathbf{Q}[k+1], \mathbf{W}[k+1]|\mathbf{y}_{1:T})q(\boldsymbol{\theta}_f^*|\boldsymbol{\theta}_f[k])}\right) \tag{3.76}$$

In case we reject the proposal we set $\boldsymbol{\theta}_f[k+1] = \boldsymbol{\theta}_f[k]$.

4
Deep State-Space Models

Contents

4.1 Deep Architectures

From the connections made between the GP prior based estimation and the one layer neural network we can ask the question whether or not we can formulate a multilayer version of a Gaussian processes State-space model. As mentioned in the section on kernel design in chapter 3, kernels such as the squared exponential provide limited structure discovery performance, and designing more complex kernels can alleviate this. However the inherent architecture is still shallow, and as such these models can be incapable of capturing certain possible types of complex structures from the data (Bengio and LeCun 2007). Inspired by neural networks with deep architectures that are popular these days, but require a large amount of data, adding latent auto-regression to the GP-SSM might be another way of achieving better structure discovery. This idea is inspired by the recurrent Gaussian processes of Mattos, Damianou, et al. (2016). At the same time the probabilistic nature of the GP-SSM would mean that we can perform inference in the deep architecture. In machine learning terms we can view the state-space model as an instance of reinforcement learning and in econometrics terms we can view the deep GP-SSM as a non-parametric counterpart to the hierarchical state-space models of Durbin and Koopman (2012). We restrict ourselves to real outputs, let $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^d$ and $\mathbf{u}_t \in \mathcal{U} \subset \mathbb{R}^e$ for $i = 1, \ldots, I$ transition functions we formulate the deep GP-SSM as:

$$f^i(\mathbf{x}_t^i) \sim \mathcal{GP}(0, k^i(\mathbf{x}_t, \mathbf{x}_s; \boldsymbol{\theta}_{f^i})) \quad \mathbf{x}_0^i \sim \mathbb{P}_0(\mathbf{x}_0) \text{ for all } i \tag{4.1a}$$

$$\mathbf{x}_{t+1}^1 = f^1(\mathbf{x}_t^1, \mathbf{u}_t) + \boldsymbol{\eta}_t^1 \quad \boldsymbol{\eta}_t^1 \sim \mathcal{N}(\boldsymbol{\eta}_t^1 | 0, \mathbf{Q}^1) \tag{4.1b}$$

$$\mathbf{x}_{t+1}^i = f^i(\mathbf{x}_t^i, \mathbf{x}_{t+1}^{i-1}, \mathbf{u}_t) + \boldsymbol{\eta}_t^i \quad \boldsymbol{\eta}_t^i \sim \mathcal{N}(\boldsymbol{\eta}_t^i | 0, \mathbf{Q}^i) \text{ for } 1 < i \leq I \tag{4.1c}$$

$$y_t | \mathbf{x}_t^I \sim \mathbb{P}_y(y_t | \mathbf{x}_t^I) \tag{4.1d}$$

For estimation we use the same Blocked Gibbs sampler, which we find to not be such a good idea in terms of computation and performance and hence in the application we are limited to only one additional layer.

5 Application and proposed models

Contents

5.1 Volatility and the leverage effect

Volatility modeling can affect various areas of finance such as the pricing of derivatives or risk management where the volatility process can be used a risk measure. Usually we concern ourselves with conditional volatility $\sigma_t^2|\mathcal{F}_{t-1}$ given the σ-algebra induced by information available at $t-1$, where for brevity we omit explicitly conditioning on $\mathcal{F}_{t-1}$ when writing σ_t^2. This conditional volatility is defined as the standard deviation of a distribution (a stochastic process in fact) over returns or percentage growths (or deltas) of an underlying asset (given $\mathcal{F}_{t-1}$) and as such it is inherently unobserved. The classic GARCH models (without ARMA components) usually have the form:

$$\mathbf{y}_t|\mathcal{F}_{t-1} \sim \mathbb{P}(\mathbf{y}_t|\mu_t, \sigma_t^2) \tag{5.1}$$

with some deterministic specification for the σ_t^2 process (see appendix). For an overview of volatility models see Poon and Granger (2003). In the case of the discretized log normal SV model the log of σ_t^2 follows an AR(1) process. Let $x_t = \log(\sigma_t^2)$ then the uni-variate discrete log-normal SV-model has the form (Kim et al. 1998):

$$x_{t+1}|\mathcal{F}_{t-1} \sim \mathcal{N}(x_{t+1}|\mu + \phi x_t, Q) \qquad y_t|x_t \sim \mathcal{N}(y_t|0, \exp(x_t)) \tag{5.2}$$

Where the two processes are assumed to be correlated. Note also that the predictive distribution:

$$\mathbb{P}(y_t|\mathcal{F}_{t-1}, \boldsymbol{\theta}) = \int_{-\infty}^{\infty} \mathbb{P}(y_t|x_t, \mathcal{F}_{t-1}, \boldsymbol{\theta})\mathbb{P}(x_t|\mathcal{F}_{t-1}, \boldsymbol{\theta})dx_t \tag{5.3a}$$

$$= \int_{-\infty}^{\infty} \mathcal{N}(y_t|0, \exp(x_t))\mathcal{N}(x_t|\mu + \phi x_{t-1}, Q)dx_t \tag{5.3b}$$

is not Gaussian. In the literature we did not encounter many stochastic volatility methods employing Gaussian Process priors, and when we did the method was significantly different then those in this thesis. Furthermore Robinson (2001) shows a clear link between SV models and the Gaussian Process covariance function corresponding to it. As such the modeling of the discretized underlying volatility process can in this case be done quite intuitively using a GP state-space model. In fact it would then also be possible to model the measurement function using a GP and allow a general mean function to be learnt as well. We did not have time to experiment with this, but it certainly is interesting to do. It would also be possible to add in the measurement function an additional GP-regression $g(t)$ as a function of t which as mentioned in chapter two nests an ARMA model.

Of importance to this thesis are stylized facts such as volatility clustering and the so-called leverage effect, because at the moment these are well accepted in stock and index time series. It is not the we are interested in the leverage-effect as a phenomenon rather we can test whether or not the models we propose can find these structural properties. The interest in the relationship between returns and volatility dates back to Black (1976) and Christie (1982) and, although underlying forces and theories behind the phenomenon are debatable (Figlewski and X. Wang 2000; Bouchaud et al. 2001; Modigliani and Miller 1958), the negative relationship between returns and volatility is well-established. For these reasons the data chosen for the empirical part in chapter 7 is chosen in a systematic way.

The first index that turned up without much delving was the NIKKEI 225 (Bekaert and G. Wu 2000), and Engle and V. K. Ng (1993) already found evidence of the leverage effect in another Japanese index (TOPIX) form 1986 to 1995. Thus we expect the NIKKEI 225 to also show this behavior in the same period . The S&P 500 is also known to exhibit the asymmetric behavior between returns and volatility (Whaley 2000), therefore we use data from this index during the period 2007 to 2014. We also use data from two stocks of companies from two different industries (ABB, PEPSICO) during different periods. What we want to see is if we obtain different results in the functional form estimation of the transition function in the GP-SSM. The data $\mathbf{y}_t$ is taken to be the percentage growth, as in definition 1 with P_t being the adjusted daily closing price. As a final note, we also wish to see a clear dependence relationship between the conditional variances overtime for the clustering stylized fact.

5.2 RR-GPSV and Deep RR-GPSV

Although the leverage effect in the SV-model is incorporated through the correlation between the errors of the states and those of the measurements, we incorporate the leverage effect as feedback into a control system akin to how it is implemented in GARCH models. Let

x_t, y_t be real valued then the model is: [RR-GPSV model]

$$x_0 \sim \mathbb{P}_0(x_0) \tag{5.4a}$$

$$\boldsymbol{\theta}_f \sim \mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_f) \tag{5.4b}$$

$$f(x_t, y_t) \sim \mathcal{GP}(0, k(x_t, x_s; \boldsymbol{\theta}_f)) \Leftrightarrow f(x_t, y_t) \approx \sum_{k=1}^{m} \mathcal{N}(0, S_{\boldsymbol{\theta}_f}(\lambda_k))\phi_k(x_t, y_t) \tag{5.4c}$$

$$f(x_t, y_t) := \mathbf{f}_{t+1} \tag{5.4d}$$

$$Q \sim \mathbb{P}_Q(Q) \tag{5.4e}$$

$$x_{t+1} = \underbrace{\begin{bmatrix} w_{1,1} & \cdots & w_{1,m} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,m} \end{bmatrix}}_{\mathbf{W}} \underbrace{\begin{bmatrix} \phi_1(x_t, y_t) \\ \vdots \\ \phi_m(x_t, y_t) \end{bmatrix}}_{\boldsymbol{\phi}(x_t, y_t)} + \eta_t \quad \eta_t | Q \sim \mathcal{N}(\eta_t | 0, Q) \tag{5.4f}$$

$$\tag{5.4g}$$

$$x_{t+1} | \mathbf{f}_{t+1}, Q \sim \mathcal{N}(x_{t+1} | \mathbf{f}_{t+1}, Q) \tag{5.4h}$$

$$y_t | x_t \sim \mathcal{N}(y_t | 0, \exp(x_t)) \tag{5.4i}$$

An important distinction between the model of Y. Wu et al. (2014) is that we impose less structure on it, which they do through the mean function and structurally affect the minimizer of the RKHS that can be generated. The decision to impose less structure is risky given that we in fact have a lot of prior knowledge about certain properties of financial time series and we are interested in estimation in the presence of not too much data. But it allows us to answer an important question with regards to the models we propose. Given the stylised facts we know about financial time series, the leverage effect and clustering in particular, we can test whether or not the models we propose are in fact able to capture these structures.

The Deep RR-GPSV is the same as the model in Equation 5.4, but includes the latent auto-regressions as in Equation 4.1, with $I = 2$. I implemented the Deep RR-GPSV in a very ad-hoc fashion, and did not experiment much with it, simply because there was not enough time. I only applied it in the empirical context of chapter 7, therefore it is merely a toy exhibition of the model's performance.

6 Finite Sample Analysis

Contents

Although the frequentist coverage of the credible regions is not that easy to establish asymptotically, even in our case where we have a prior model with adaptive regularity, we can ask ourselves how reliable these are in a finite sample setting. This is the aim of the subsequent sections. In the first section we evaluate the efficacy and efficiency of the sampler by simulating data from a known expansion with a known distribution over the weights and seek to answer how the various settings of the sampler influence its ability to recover the true distribution. In the second section we simulate from a state-space model with a known transition function. We study how well the credible sets of the posterior over the estimated function covers the true functional form we simulate from, given various expansion orders.

6.1 Evaluation of the Blocked Gibbs Algorithm

There are a couple of important questions with regards to the efficacy and efficiency of the Blocked Gibbs sampler when it comes to inference, with the most general one being how well can it recover a known distribution. The algorithm has quite a number of tunable settings, and we wish to analyse their effect on the estimation. In choosing these parameters a number of metrics are employed which are also evaluated. We also would like to know how well the estimation procedure performs when the size of the available data is no more than $T = 500$.

6.1.1 Simulation set-up

Data Generating Function. The DGP is set up in two steps. First we simulate data from:

$$\tilde{\mathbf{x}}_{t+1} \sim \tilde{\mathbb{P}}(\tilde{\mathbf{x}}_{t+1}|\tilde{f}(\tilde{\mathbf{x}}_t, \tilde{\mathbf{y}}_t), \mathbf{Q}) \qquad \tilde{\mathbf{y}}_t \sim \tilde{\mathbb{P}}(\tilde{\mathbf{y}}_t|0, \exp(\tilde{\mathbf{x}}_t)) \tag{6.1}$$

Then in order to have an interpretable DGP in the form of Equation 3.20 we obtain a postrior $\mathbb{P}_{post}(\tilde{\mathbf{x}}_{0:\tilde{T}}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f|\tilde{\mathbf{y}}_{0:\tilde{T}})$ by running the Blocked Gibbs sampler for output $\{\tilde{\mathbf{y}}_t\}_{t=0}^{\tilde{T}}$ and states $\{\tilde{\mathbf{x}}_t\}_{t=0}^{\tilde{T}}$ generated by the model in Equation 6.1, where in this case we set $\tilde{f}$ to be a simple wave function in both of its arguments. Then the DGP becomes:

$$\mathbf{x}_0 = \mathbf{u}_0 = \mathbf{y}_0 = 0 \qquad \mathbf{u}_t \equiv \mathbf{y}_t \tag{6.2a}$$

$$\{\mathbf{Q}^i, \mathbf{W}^i\}_{i=1}^{\tilde{K}} \sim \mathbb{P}_{post}(\tilde{\mathbf{x}}_{0:\tilde{T}}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f|\tilde{\mathbf{y}}_{0:\tilde{T}}) \tag{6.2b}$$

$$\bar{\mathbf{Q}} = \frac{1}{\tilde{K}} \sum_{i=1}^{\tilde{K}} \mathbf{Q}^i \tag{6.2c}$$

$$\bar{\mathbf{W}} = \frac{1}{\tilde{K}} \sum_{i=1}^{\tilde{K}} \mathbf{W}^i \tag{6.2d}$$

$$\mathbf{x}_{t+1} = \underbrace{\bar{\mathbf{W}} \begin{bmatrix} \phi_1(\mathbf{x}_t, \mathbf{u}_t) \\ \vdots \\ \phi_m(\mathbf{x}_t, \mathbf{u}_t) \end{bmatrix}}_{f(\mathbf{x}_t, \mathbf{u}_t)} + \boldsymbol{\eta}_t \qquad \boldsymbol{\eta}_t|\bar{\mathbf{Q}} \sim \mathcal{N}(\boldsymbol{\eta}_t|0, \bar{\mathbf{Q}}) \qquad t = 1, \ldots, T + 100 \tag{6.2e}$$

$$\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{y}_t|0, \exp(\mathbf{x}_t)) \qquad t = 1, \ldots, T + 100 \tag{6.2f}$$

The tilde's reflect the fact that an element either belongs to or is obtained in step one of the DGP setup, and the first 100 samples are purged as burn-in. This way we have control over the structure of the function that is generated by the weighted basis function expansion. In the first step of the DGP setup we have $\tilde{K} = 5000 - 1000$ (burn-in), $\tilde{T} = 500$. $\tilde{K}$ is the important parameter because it has to be sufficiently large so that a statistic such as the mean $\bar{\mathbf{W}}$ can be informative about the joint distribution in Equation 6.2b. In this case $\tilde{K}$ is not that large, but it suffices for the purposes of this section. $\tilde{T}$ is less important, because we do not care about how well $\tilde{f}$ is approximated in the first step, as long as the approximation resembles $\tilde{f}$. Note also that the weights contain information about the spectral density parameter distribution of the DGP. If we recall that $\mathbb{P}(\mathbf{W}[k+1]|\mathbf{Q}[k+1], \mathbf{x}_{0:T}[k+1], \mathbf{y}_{1:T}) \equiv \mathcal{MN}(\mathbf{W}|\mathbf{P}(\Sigma + \mathbf{V})^{-1}, \mathbf{Q}, (\Sigma + \mathbf{V})^{-1})$ and the row covariance matrix $\mathbf{V}^{-1}$ has the spectral density on its diagonal.

In this case we have set the spectral density to be that of the Matern convolution kernel over the product space of the state and the input. The expansion order in generating the posterior $\mathbb{P}_{post}(\tilde{\mathbf{x}}_{0:\tilde{T}}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f|\tilde{\mathbf{y}}_{0:\tilde{T}})$ was 4 basis functions for each dimension $m^2 = 4^2$ and the domain set to $[-6, 6] \times [-6, 6]$. When performing the evaluation of the sampler we set these to be the same, except for the domain which in that case becomes $[-8, 8] \times [-8, 8]$. When

evaluating the blocked Gibbs Sampler we set the hyper-priors over the spectral densities to be $\sigma_{\mathbf{x}_t}, \sigma_{\mathbf{y}_t} \sim \mathcal{N}(\sigma_\cdot|100, 1)$, $\ell_{\mathbf{x}_t}, \ell_{\mathbf{y}_t} \sim \mathcal{N}(\ell_\cdot|10, 1)$ and $\nu_{\mathbf{x}_t}, \nu_{\mathbf{y}_t} \sim \text{Exp}(\nu_\cdot|10)$. The length-scale parameter is known to be hard to identify therefore we impose structure on the prior function space by choosing an informative prior on the length-scale of the spectral density. An informative prior centered around 10 with a small variance restricts the function space to only contain functions that are neither very wiggly nor those that allow for extreme correlation between two consecutive points. We do the same for the scale/signal-SD parameters, where the high signal variance make it sensitive to variation in the small amount of data. The latter choice later turns out to not be such a wise one. As for taking means of the weights another approach could have been to compute the sequence of $\tilde{\mathbf{y}}_t^i$ for each k in $\{1, \ldots, \tilde{K}\}$ and take the mean of that. In both cases by taking the mean we make the implicit assumption here that the statistics contain enough information about the joint posterior over the weights, spectral density parameters, state error variances and the hidden states such that it can be identified. In an empirical setting we implicitly make this assumption as well so it is a fair one.

To summarize the simulated distribution over the function we plot the predictive mean of the function as in Equation 6.2e, which includes a summary of the distributions over the state error variances, the basis function expansion weights, as well the spectral density parameters. In addition we plot a credible region of the predictive distribution via standard deviations from the predictive mean using the draws from $\mathbb{P}_{post}(\tilde{\mathbf{x}}_{0:\tilde{T}}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f|\tilde{\mathbf{y}}_{0:\tilde{T}})$. In both cases we let $\mathbf{x}_t^*, \mathbf{y}_t^* \in \{-8, -7.9, -7.8, \ldots, 7.8, 7, 9, 8\}$ be input values for one-step ahead prediction of the state at $t + 1$. For the standard deviation of the draws we use:

$$\mathbb{V}ar[\hat{\mathbf{x}}_{t+1}] = \boldsymbol{\phi}(\mathbf{x}_t^*, \mathbf{y}_t^*)'\mathbb{V}ar[\mathbf{W}]\boldsymbol{\phi}(\mathbf{x}_t^*, \mathbf{y}_t^*) + \mathbb{V}ar[\boldsymbol{\eta}_t] \tag{6.3}$$

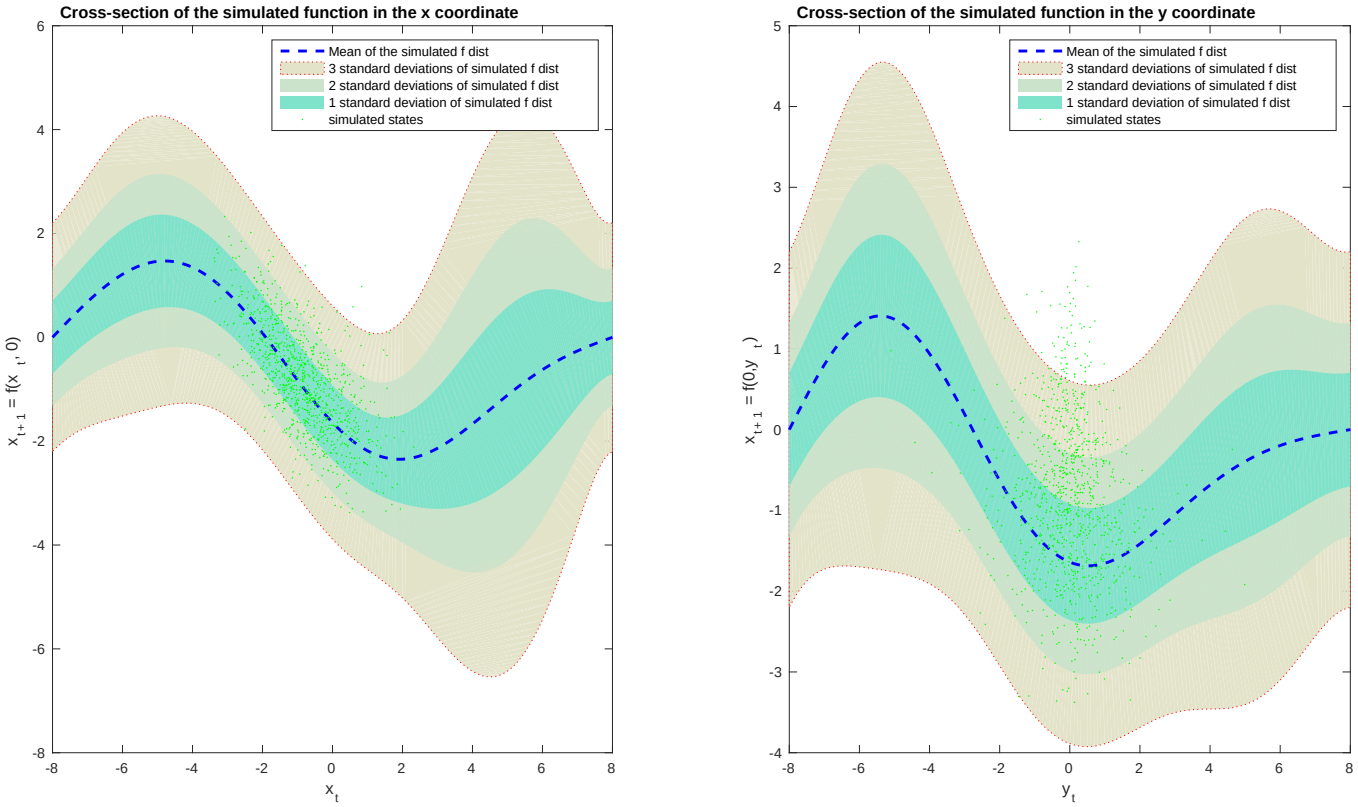
We can estimate $\mathbb{V}ar[W]$ by taking the empirical covariance and estimate $\mathbb{V}ar[\boldsymbol{\eta}_t]$ by $\bar{\mathbf{Q}}$. If $\mathbf{W}$ has $n > 1$ rows we can take:

$$\widehat{\mathbb{V}ar}[\mathbf{W}] = \text{diag}(\{\widehat{\mathbb{V}ar}[\mathbf{W}_i]\}_{i=1}^n) \tag{6.4}$$

This is not used here, but is used for the deep learning part where the state dimension is necessarily larger than 1. By taking the square root of the estimated variance we get the standard deviation, which can be used to find l times the standard deviation of the distribution over the function. With these statistics we summarize the joint distribution $\mathbb{P}_{post}(\tilde{\mathbf{x}}_{0:\tilde{T}}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f|\tilde{\mathbf{y}}_{0:\tilde{T}})$ (Figure 6.1) graphically.

Evaluation metrics Given that in this setting we have the simulated states, we can use these to compute the Root Mean Square Error (RMSE) to measure the performance of the learning algorithm. After estimating the model on data $\{\mathbf{y}_t\}_{t=0}^T$, we generate another

Figure 6.1: Summary of the distribution over the state function for input grid. Left we have $\mathbf{x}_{t+1} = f(\mathbf{x}_t^*, 0)$, and right $\mathbf{x}_{t+1} = f(0, \mathbf{y}_t^*)$ with the distributions computed according to Equation 6.3



data-set $\{\mathbf{y}_t^*\}_{t=1}^{T_{test}}, \{\mathbf{x}_t^*\}_{t=1}^{T_{test}}$ from the DGP, where $T_{test} = 100000$. Then at step t given $\mathbf{x}_t^*$ and $\mathbf{y}_t^*$ we predict $\hat{\mathbf{x}}_{t+1}$ using the estimated model and compare it to the true simulated state.

$$\text{RMSE} = \left(\frac{1}{T_{test} - 1} \sum_{t=1}^{T_{test}-1} (\hat{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1}^*)^2 \right)^{-1/2} \tag{6.5}$$

In order to take into account the whole distribution of the state equation in the evaluation we also consider evaluating the log-likelihood at each true state $\mathbf{x}_{t+1}^*$ and averaging those:

$$LL = \frac{1}{T_{test} - 1} \sum_{t=1}^{T_{test}-1} \log \mathcal{N}(\mathbf{x}_{t+1}^* | \hat{\mathbf{x}}_{t+1}, \widehat{\mathbb{V}ar}[\hat{\mathbf{x}}_{t+1}]) \tag{6.6}$$

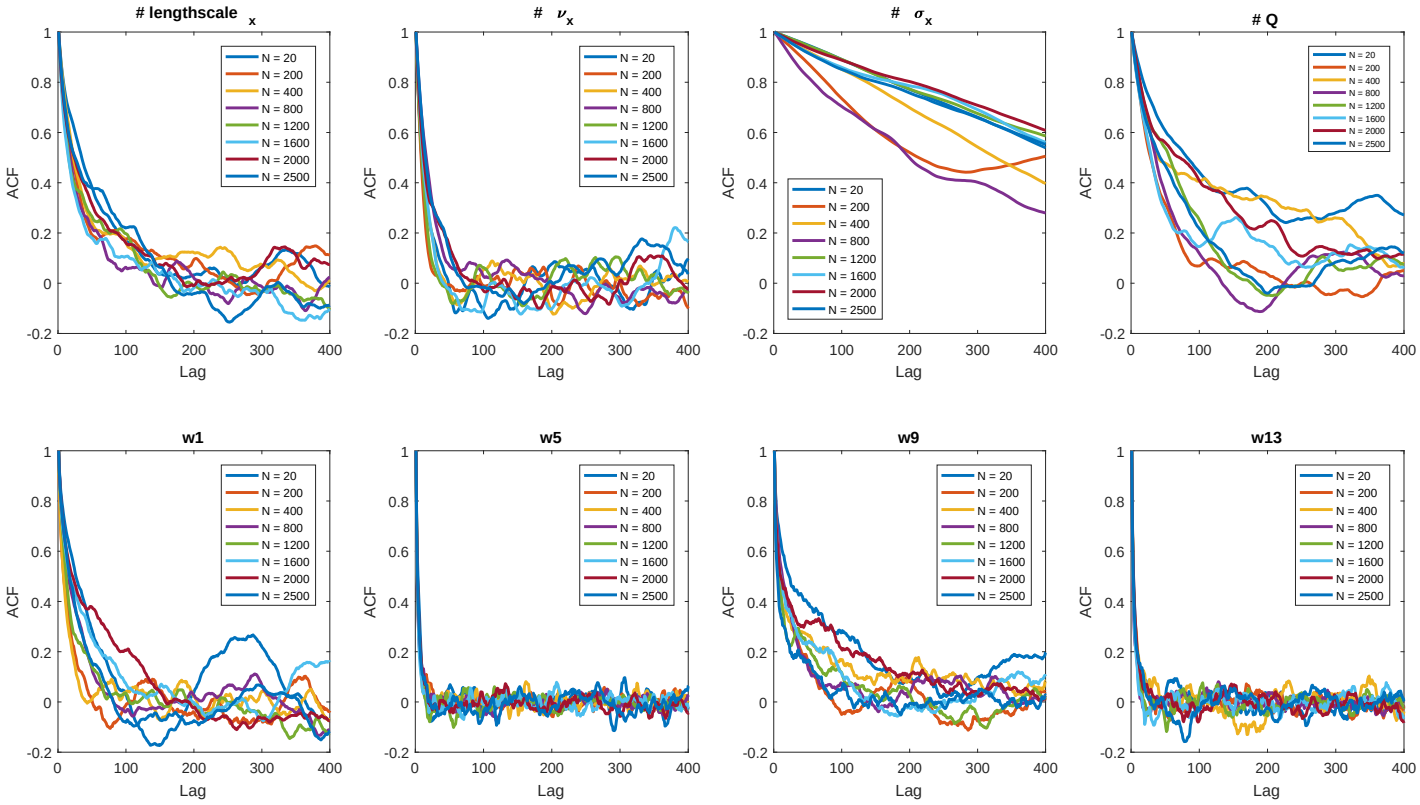
Where $\widehat{\mathbb{V}ar}[\hat{\mathbf{x}}_{t+1}]$ is the sample estimate of the variance in Equation 6.3. An interesting question we aim to answer is if these metrics reflect properties of the sampler such as its mixing. In the subsequent section in turns out that they do.

6.1.2 Mixing of the sampler and the number of particles

We run the Blocked Gibbs Sampler for data generated by the DGP Equation 6.2 in the second step with $K = 10000$, a burn-in of 1000 and initialise all the parameters with a value of 1. The sample size of the generated data T is 500. We begin with a plot of the ACF

for various draws from $\hat{\mathbb{P}}_{post}(\mathbf{x}_{0:T}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f | \mathbf{y}_{0:T})$, where we wish to see a quick decrease in the ACF for not too large lag values. This would indicate that there is not much relation between values in the Markov chain for draws not too far apart. We repeat the experiment for $N \in \{20, 200, 400, 800, 1200, 1600, 2000, 2500\}$ and $N_{mh} = 10$ Metropolis-within Gibbs runs, and for each N we repeat the experiment 5 times and average the values.

Figure 6.2: ACF plots of posteriors draws of the spectral density parameters in the state dimension, the state error variance, and a couple of the basis function expansion weights. Omitted are ℓ_y, σ_y, ν_y which look very similar as well as w_i for $i = 2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16$ which also closely resemble those plotted below. $T = 500$, $K = 10000 - 1000$,$N_{mh} = 10$



As can be seen in Figure 6.2 even for $N = 20$ the mixing is acceptable and for most of the parameters the ACF is indistinguishable between $N = 20$ and $N > 20$. As Lindsten et al. (2014) note this indicates that the performance of the sampler is close to one that would draw the states from the true joint distribution (an "ideal" Gibbs sampler), which is (almost) also the case for our Blocked Gibbs sampler. Hence its limitations are derivative of the mixing of the "ideal" Gibbs sampler rather than the PGAS kernel within the sampler. However for $\mathbf{Q}$ a setting such as $N = 20$ for example does not seem to suffice and for the spectral density scale parameter σ_x the gradient of the ACF line is not sharp for small lag values, with any N, indicating bad mixing. Although not plotted in the figure the same holds for the scale parameter σ_y, and this issue is addressed later on. From the plots in Figure 6.2 $N = 200$ seems to be the smallest choice for the number of particles that works well. Note that an N

between 20 and 200 could also be possible, which would reduce computation, and that the plots for the other parameters are omitted for brevity and are all available by request.

Additionally as a metric for the mixing we compute the Inefficiency Factors (IF) (Giordani et al. 2011) for all the parameters. The IF can be interpreted as a factor that indicates the need for $K\times$IF draws from the Markov Chain to resemble drawing K i.i.d. samples from the posterior.

$$\text{IF} = 1 + 2\sum_{i=1}^{\infty} \widehat{\text{ACF}}(i) \tag{6.7}$$

The equation above is directly derived by taking the variance of dependent draws from the Markov Chain and assuming stationarity. $\widehat{\text{ACF}}(i)$ is the empirical ACF at lag i, and the terms in the sum are cut-off using a Parzen window (Andrews 1991):

$$k(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 \text{ for } 0 \leqslant |x| \leqslant \frac{1}{2} \\ 2(1 - |x|^3) \text{ for } \frac{1}{2} \leqslant |x| \leqslant 1 \\ 0 \text{ otherwise} \end{cases} \tag{6.8}$$

We compute the IF for each hyper-parameter and weight after which we take the average.

Another method to compare efficiency is the consistent batch means method. This method relies on requirements for a Central Limit Theorem for the Markov chain, from which we draw our samples, to be fulfilled (For details (Jones et al. 2006)). In that case we have $\sqrt{K}\left(\frac{1}{K}\sum_{i=1}^{K} X_i - \mathbb{E}[X]\right) \to^d \mathcal{N}(0, \Sigma)$ as $K \to \infty$. For $\{X_i\}_{i=1}^{K}$ draws from the Markov chain:

$$Y_j := \frac{1}{b}\sum_{i=(j-1)b}^{jb-1} X_i \qquad \text{for } j = 1, \ldots, a \tag{6.9a}$$

$$\hat{\Sigma}_{bm} = \frac{b}{a-1}\sum_{j=1}^{a}\left(Y_j - \frac{1}{K}\sum_{i=1}^{K} X_i\right)\left(Y_j - \frac{1}{K}\sum_{i=1}^{K} X_i\right)' \tag{6.9b}$$

If the number of batches a and their sizes b are fixed consistency cannot be obtained, however if we let both increase along with the sample size then there are results available for consistency. Based on these results (Jones et al. 2006; Vats et al. 2015) we let $b = \sqrt{N}$. Furthermore we also compute the log-likelihood and RMSE metrics mentioned earlier for each N. For each N the experiment has been repeated 5 times and the averages are reported.

In table Table 6.1 the results are visible and seem to be in line with the plots in Figure 6.2, where the inefficiency estimate is in fact the lowest for $N = 200$ and the batch means variance estimate is the third lowest for $N = 200$. The RMSE and the LL give somewhat less conclusive results. What can be taken away from these two is that the RMSE seems to be the highest for $N = 20$ and acceptable for $N = 200$, and the same holds for the LL.

Table 6.1: IF are averaged over the parameters. The batch means variance is the average of the diagonals of $\hat{\Sigma}_{bm}$. Note that this is a point estimate of the asymptotic variance of $\sqrt{K}\left(\frac{1}{K}\sum_{i=1}^{K} X_i - \mathbb{E}[X]\right)$ and only serves as a measure for efficiency comparisons. Values are averages over 5 experiments for each $N.T = 500$, $K = 10000 - 1000$, $N_{mh} = 10$. The best value for the metric is given in bold/italics, the second best in bold only and the third in italics only.

	IF	**BM-$\mathbb{V}ar$**	**RMSE**	**LL**
$\boldsymbol{N = 20}$	6,9907	43,5114	0,1696	-0,5444
$\boldsymbol{N = 200}$	***3,3970***	*26,2951*	0,1595	-0,5369
$\boldsymbol{N = 400}$	6,2370	32,7810	***0,1484***	-0,5395
$\boldsymbol{N = 800}$	**3,5075**	***20,9504***	**0,1499**	-0,5349
$\boldsymbol{N = 1200}$	4,7020	41,3447	0,1654	-0,5432
$\boldsymbol{N = 1600}$	*4,1793*	**25,4282**	0,1600	*-0,5313*
$\boldsymbol{N = 2000}$	6,2724	31,1036	*0,1539*	**-0,5304**
$\boldsymbol{N = 2500}$	4,4483	49,7104	0,1551	***-0,5303***

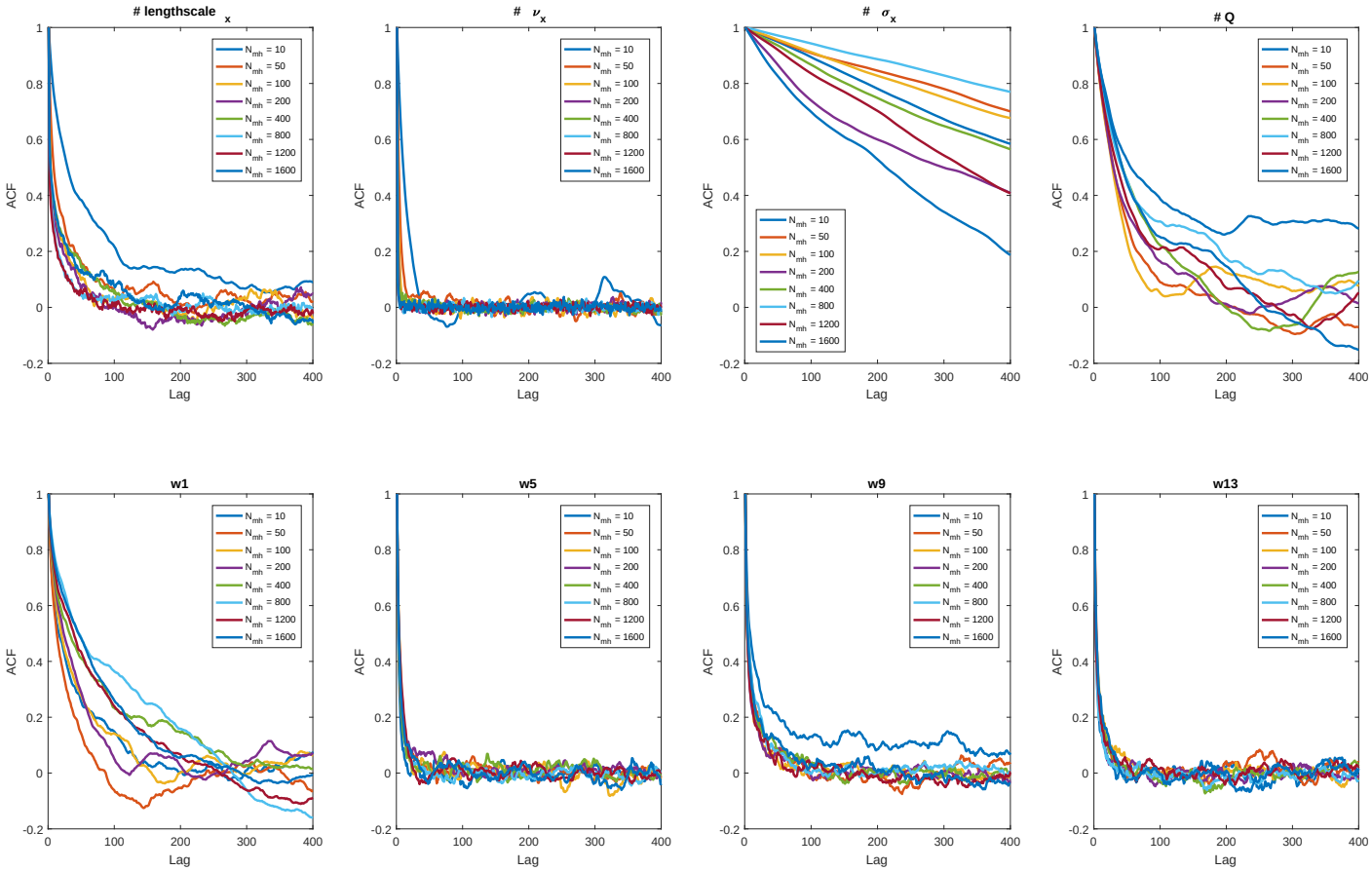
6.1.3 Number of Metroplois-within-Gibbs runs

The most important tuning parameter of the random walk Metropolis-within-Gibbs algorithm, besides the number of runs, is the variance of the random walk innovations. Given that we do not wish to allow for large degrees of freedom for the length-scale of the GP prior, we set the random walk innovation variance low to 0.1 for both dimensions of the spectral density, and for all others a value of 1 is used. For tuning the parameters we follow P. Neal and G. Roberts (2006) and aim for an average acceptance rate between 0.2 and 0.3 (see Table 6.2), similar to a full Metropolis-Hastings algorithm with more than 5 components. Although the block components of the Metropolis-within-Gibbs update are not identically distributed and we don't use sub-block updating, as in their paper, after experimenting with various random walk innovation variance values, we indeed found those to that correspond to an acceptance rate of roughly 0.22 to give the best results.

As for the number of runs, looking at the plots in Figure 6.3 it looks as though from $N_{mh} \geq 50$ increasing the number of runs does not necessarily result in better mixing of the sampler (not considering the spectral density scale parameter for now). In Table 6.2 the batch means variance estimate reflect a notion close to the implication of Theorem 1 of Sherlock et al. (2016) where, because the Blocked Gibbs samplers target density is the joint posterior and not an accurate conditional, a large number of Metropolis-within-Gibbs runs is not necessarily advantageous. But in Table 6.2 the efficiency seems to worsen with increasing N_{mh}. The IF metric does not indicate that much harm done by a large number Metropolis iterations in terms of efficiency (except of course computational burden). In fact after close inspection we find the evil doers in the large BM variances to be the spectral density scale parameter draws (addressed later on).

For most variables $N_{mh} \in \{50, 100, 200\}$ seems to result in the best mixing, but there are some subtleties to be noticed here and these become clear in the next figures. We

Figure 6.3: ACF plots of posteriors draws of the spectral density parameters in the state dimension, the state error variance, and a couple of the basis function expansion weights. Omitted are ℓ_y, σ_y, ν_y which look very similar as well as w_i for $i = 2, 3, 4, 6, 7, 8, 10, 11, 12, 14, 15, 16$ which also closely resemble those plotted below. $T = 500$, $K = 10000 - 1000$, $N = 200$

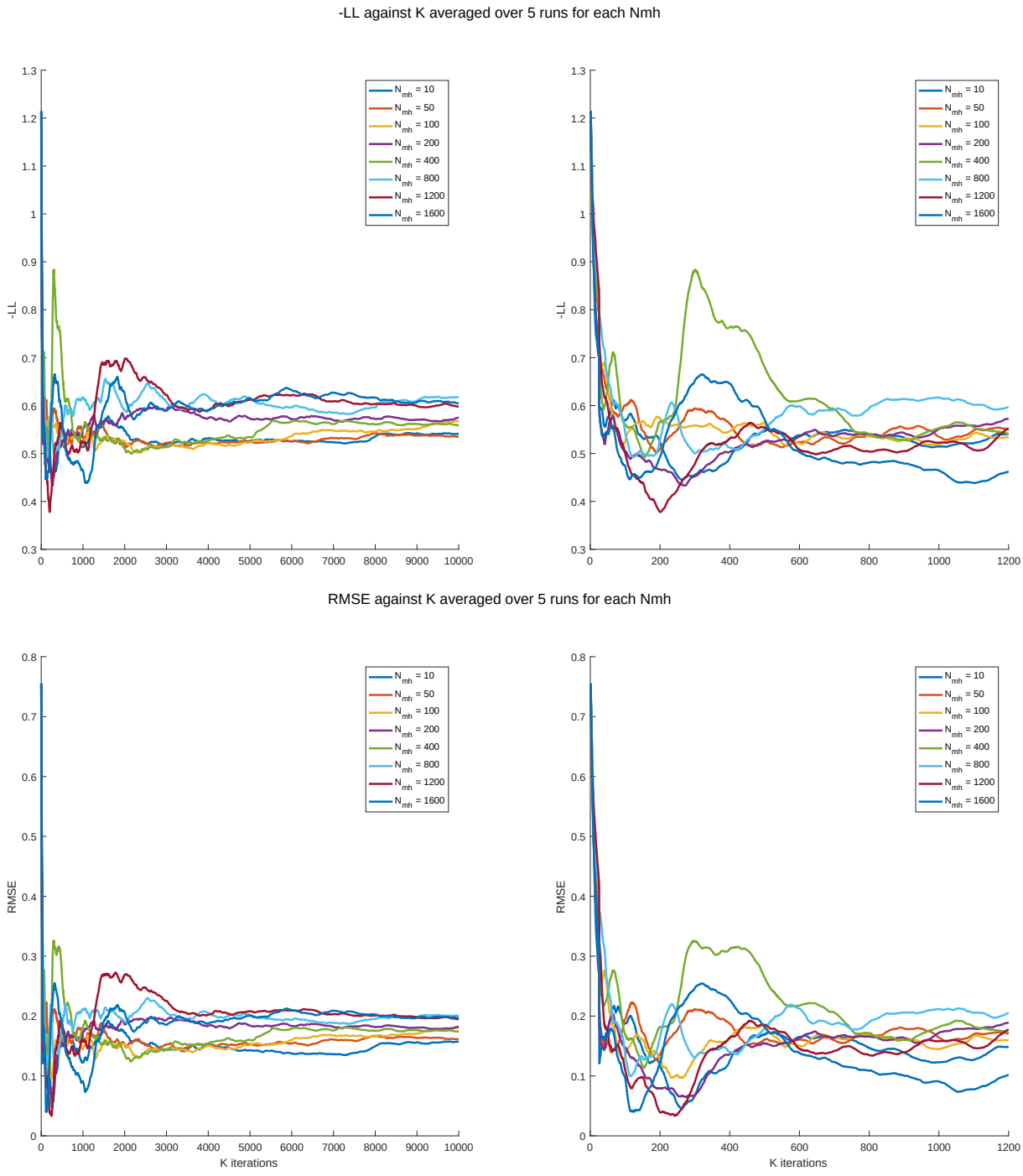


delve into the subtleties beginning with plotting the LL and RMSE against the number of iterations for various Metropolis-within-Gibbs runs.

In Figure 6.4, desirably, a steep decline in both RMSE and -LL is visible as function of K for any number of N_{mh}. One important take away is that increasing the number of Metropolis-within-Gibbs (N_{mh}) runs has the effect of reaching the minimum -LL and RMSE for smaller number of Blocked Gibbs runs (K). But for higher N_{mh} it seems that the RMSE and -LL are worse than for smaller ones as the number of K increases. This is also visible in Table 6.2 where the long run values are given and the LL, RMSE, IF, and the BM variance only get worse as the number of Metropolis-within-Gibbs are increased.

If we look at the distribution identification of the various hyper-parameters and weights in Figure 6.5 and compare a long run of the blocked Gibbs sampler with a short one, we observe the spectral density scale parameters are the most troublesome ones. It seems as though as K increases the variance of the posterior marginal of the spectral density scale parameter only gets larger, indicating draws from a non-stationary series. Indeed this is in line with the bad mixing of σ_x and σ_x in both Figure 6.2 and Figure 6.3.

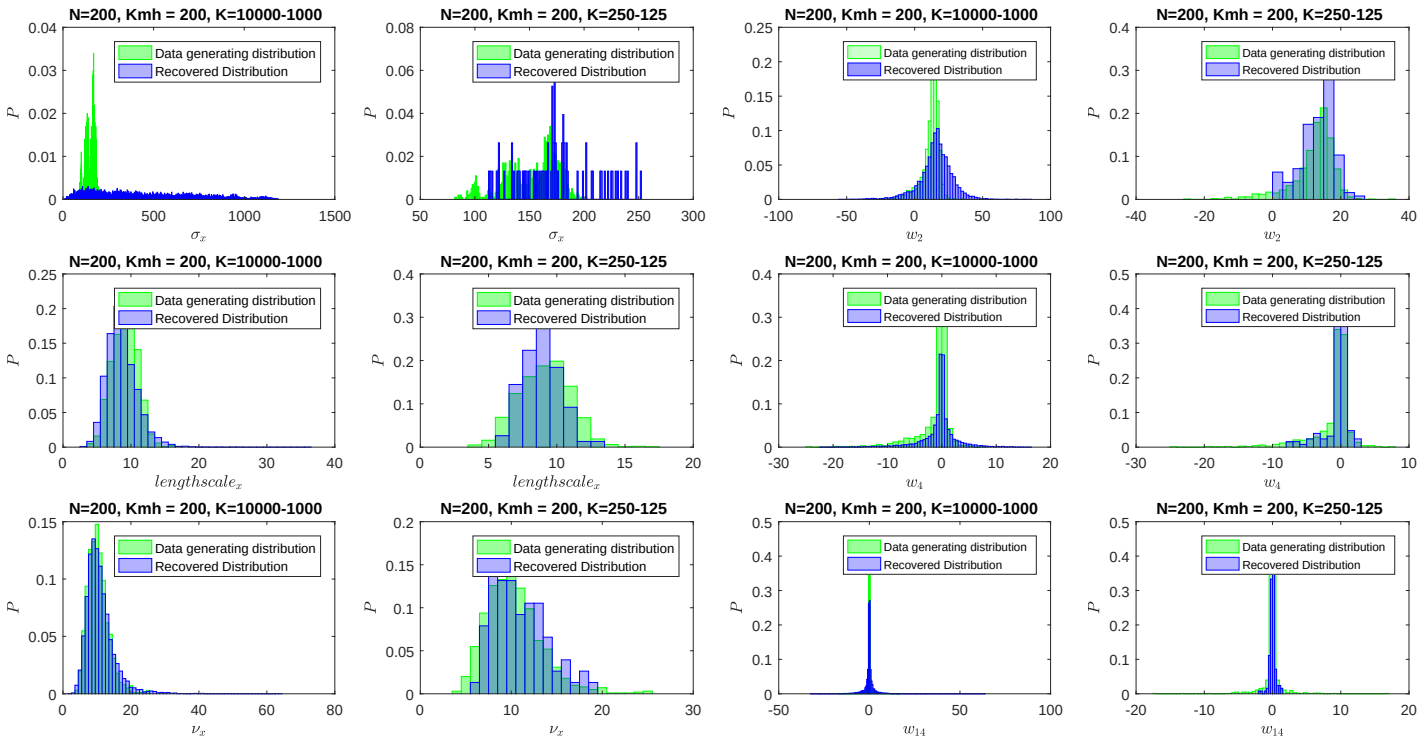
Figure 6.4: The negative LL and RMSE plotted against K for various N_{mh}. For each N_{mh} the experiment is performed 5 times, after which the LL and RMSE is averaged. $T = 500$, $N = 200$



In Figure 6.6 we can see that the distribution summary as explained in the simulation set-up section indeed summarized the distribution identification of the DGP where in the short run it is clearly much better than in the long run.

Looking at Figure 6.4 we see that in there is a clear distinction between short run and

Figure 6.5: Histograms of the DGP distributions versus those from the joint posterior to compare the long and the short run. Because of the large amount of possible parameters 6 representative ones are chosen. Omitted are ℓ_y, σ_y, ν_y which look very similar as well as w_i for $i = 1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16$ which also closely resemble those plotted below. In this figure $Kmh := N_{mh}$. $T = 500$



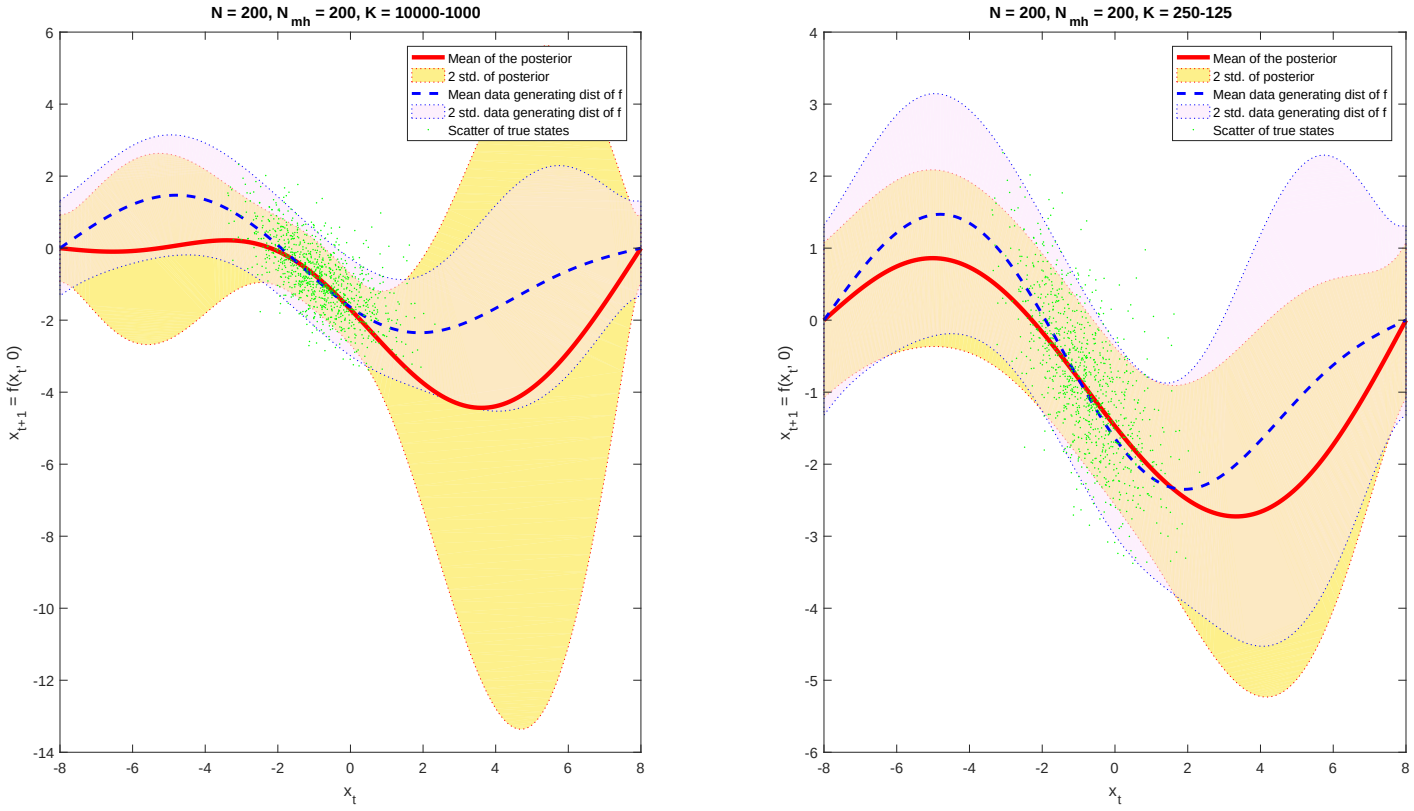
long run performance of the sampler and we see that in the short run $N_{mh} = 200$ had better LL and RMSE values than $N_{mh} = 50$ and this is also visible in the histograms in Figure 6.7.

For a subset of the weights and hyper-parameters the trace-plots are given in Figure 6.8, making sure to include the spectral density scales. As with all the other plots the subset is chosen to be representative of all the parameters. For example the histograms of all other non-zero centered weight distribution are similar to w_2 and the same thing holds for the zero-centred weight distributions. The trace plots in Figure 6.8 reflect the non-stationary behavior of the spectral density scale chain. After some experimentation with the Metroplis-within-Gibbs tuning parameter and various hyper-priors it seemed that a different hyper-prior choice has a lot of impact. We found the relatively uninformative prior $\sigma_x\sigma_y \sim \mathcal{HN}(\sigma.|0, 50)$ to work best. The half-Cauchy prior resulted in allowing too large scales due to its thick tails and a flat prior resulted in bad identification. The trace plots with the new prior are given in Figure 6.9. With $N = 200, N_{mh} = 200, T = 500$ we found in the long run, $K = 10000 - 1000$, the RMSE and LL to be 0.0821 and -0.5281 respectively and in the short run, $K = 10000 - 1000$, 0.0908 and -0.5230 respectively. The samplers efficiency is much better with $IF = 2.8650$ and BM-$\mathbb{V}ar = 10.4380$. We can also conclude that the LL

Table 6.2: IF are averaged over the parameters. The batch means variance is the average of the diagonals of $\hat{\Sigma}_{bm}$. Note that this is a point estimate of the asymptotic variance of $\sqrt{K}\left(\frac{1}{K}\sum_{i=1}^{K} X_i - \mathbb{E}[X]\right)$ and only serves as a measure for efficiency comparisons. The acceptance rate are avregaed over the number of Blocked-Gibbs updates. $T = 500$, $K = 10000 - 1000$, $N = 200$. The values are averages over 5 experiments for each N_{mh}. The best value for the metric is given in bold/italics, the second best in bold only and the third in italics only.

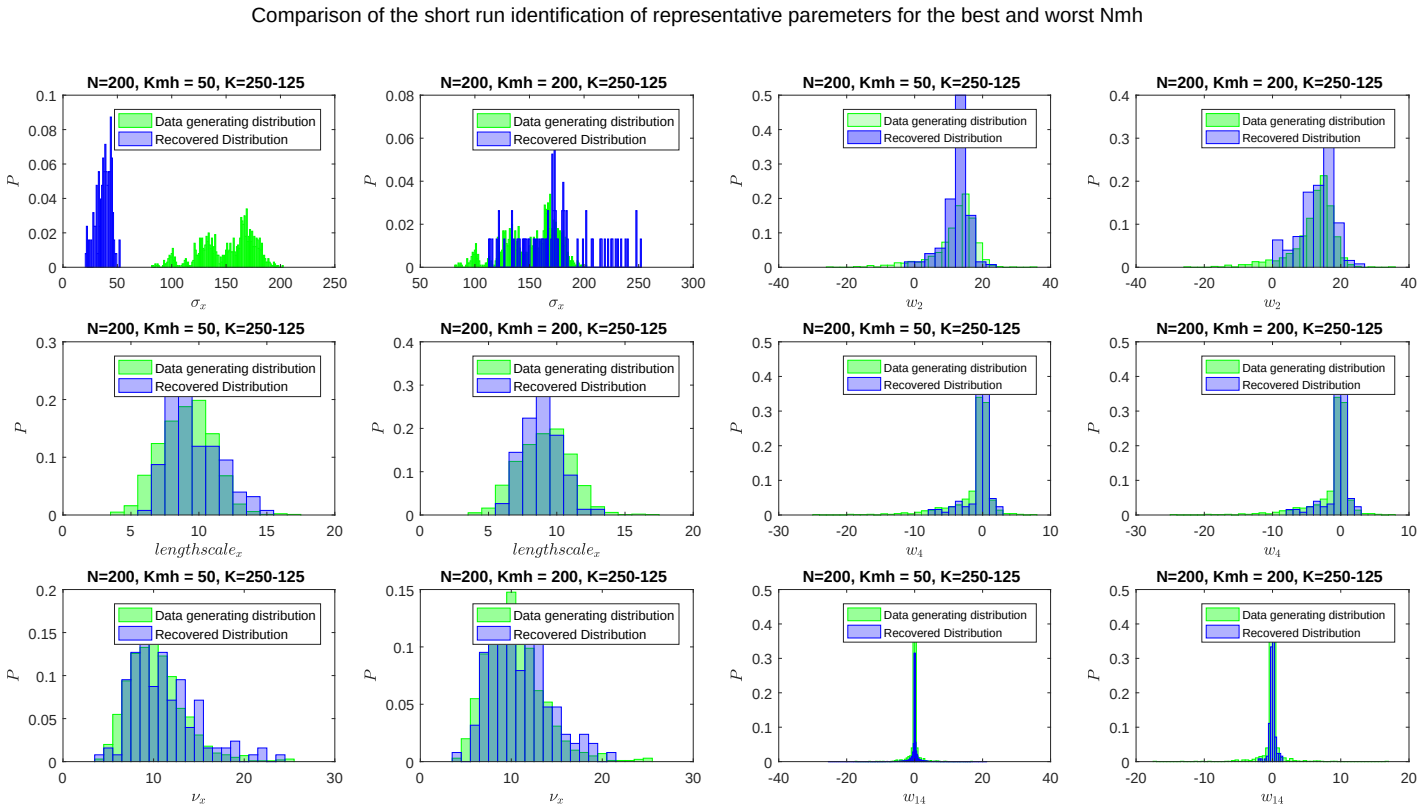
	IF	**BM-$\mathbb{V}ar$**	**RMSE**	**LL**	**AvgAR**
$N_{mh} = 10$	3,3970	***26,2951***	***0,1595***	**-0,5369**	0,2049
$N_{mh} = 50$	**3,3768**	**66,9391**	**0,1608**	***-0,5346***	0,2121
$N_{mh} = 100$	***3,2565***	134,4351	0,1800	-0,5666	0,2259
$N_{mh} = 200$	*3,8579*	*99,1656*	0,1819	-0,5743	0,2269
$N_{mh} = 400$	4,5420	141,7283	*0,1738*	*-0,5583*	0,2223
$N_{mh} = 800$	5,4195	281,6407	0,2004	-0,6164	0,2295
$N_{mh} = 1200$	4,4482	206,0086	0,1945	-0,5968	0,2259
$N_{mh} = 1600$	5,0793	204,6472	0,1971	-0,6048	0,2270

Figure 6.6: Summary of the predictive distribution of the estimated model, with $\hat{f}$ in only its $\mathbf{x}_t$ argument, versus that of the DGP to compare the identification in the short and long run. Note that this summary implicitly includes information about all the distributions of the parameters and we only plot one cross-section for brevity given that the two cross sections look similar. $T = 500$.



and RMSE are good indicators of drawing from the target stationary distribution and from Figure 6.7 we know that they are also good indicators of the DGP distribution identification.

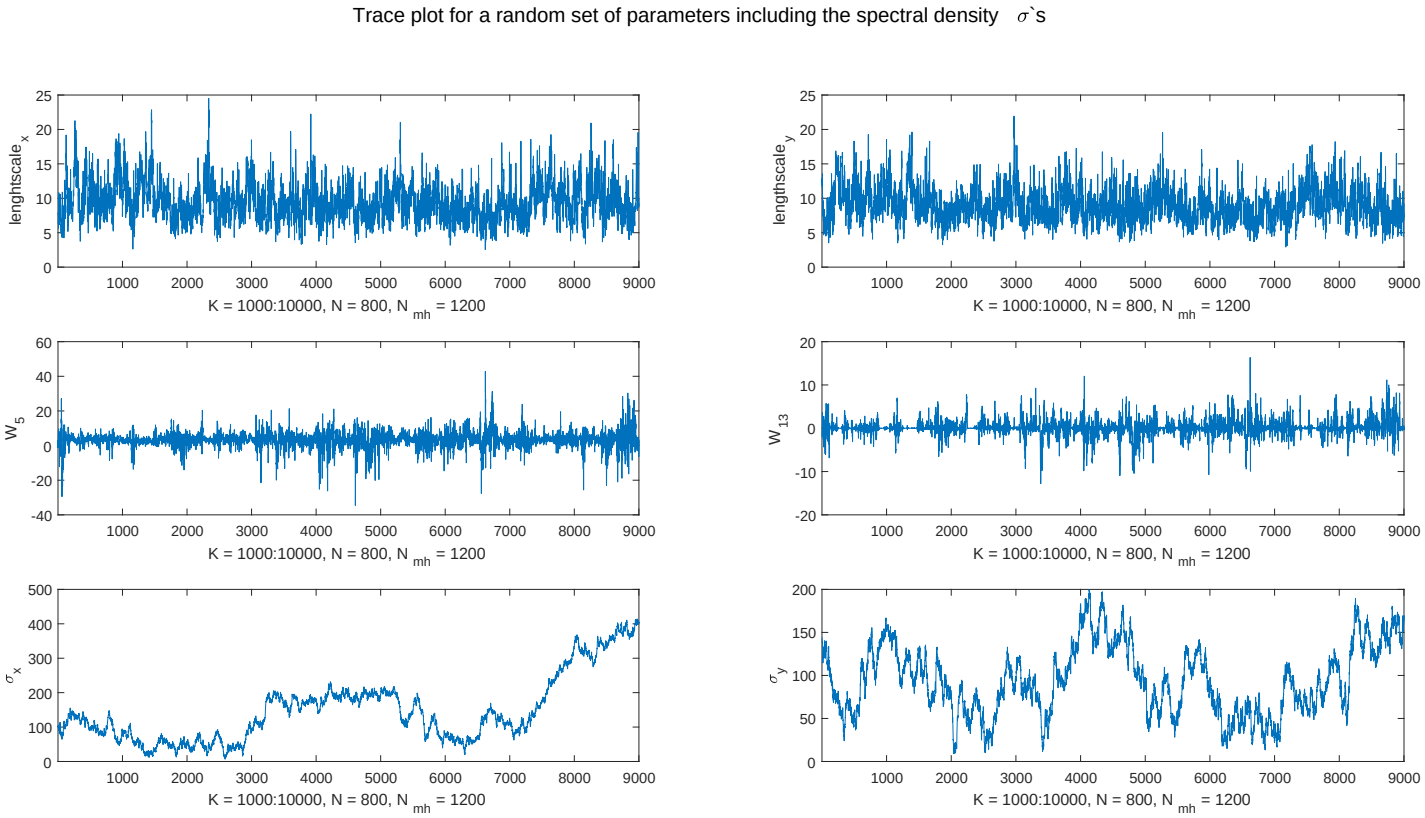
Figure 6.7: Histograms of the DGP distributions versus those from the joint posterior to compare the best and worst N_{mh} setting in the short run. Because of the large amount of possible parameters 6 representative ones are chosen. Omitted are ℓ_y, σ_y, ν_y which look very similar as well as w_i for $i = 1, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 16$ which also closely resemble those plotted below. $T = 500$



As a final test we check how much the estimation is enhanced by keeping everything the same and increasing the sample size to $T = 10000$. We find the $IF = 3.0081$ and BM-$\mathbb{V}ar = 13.9027$. The RMSE and LL become 0.0519 and -0.5192 and a summary plot of the estimated distribution is given in Figure 6.10. It is visible that both the short and long run estimations are sufficient now, with the focus being the estimation around the available data.

Conclusions on the sampling scheme From the analysis a number conclusions can be drawn. First it is clear that it is possible to recover the distributions over the DGP parameters sufficiently well with the algorithm setting set to minimal and the identification becomes increasingly good as the data and settings such as the number of particles is increased. With regards to the sample size, T=500 seems to be sufficient for acceptable estimation results. Note that the aim is not perfect identification, but rather the ability to have the estimation algorithm work with a small sample size, with reliable uncertainty quantification. Also a regularizing mechanism is visible where the posterior marginals over unneeded weights are distributed tightly around zero.

Figure 6.8: Trace plots of some the parameters in the joint posterior, where clearly the bad mixing of the spectral density scale parameters does not result in draws from a stationary distribution. For $T = 500$.

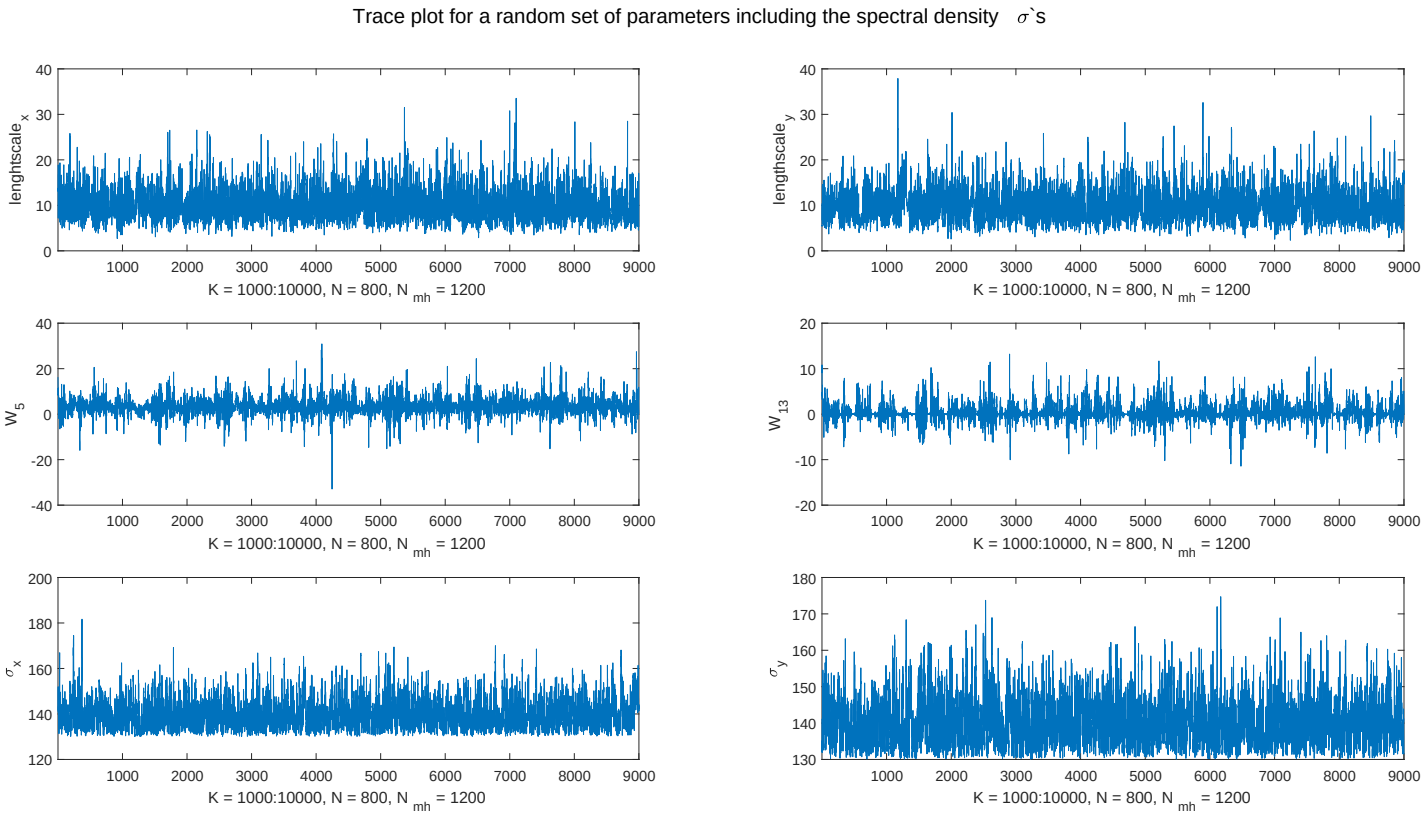


With regards to the distribution summaries in Figure 6.10 and in Figure 6.7, they are good indicators of the identification performance summarizing the recovery of the distributions as in Figure 6.5.

As for the settings as a default, in a similar model setting, satisfactory results can be obtained with a relatively small number of particles such as $N = 200$ as a default. For the number or blocked-Gibbs runs K a trade-off between the K and the number of Metropolis-within-Gibbs runs N_{mh} can be made, where I found for $N_{mh} = 200$ and a number $K \in \{200, 201, \ldots, 299, 300\}$ to be satisfactory. Increasing K is more expensive then increasing N_{mh}, although the latter is less risky. Then from here on for $K \geq K_{min}$ we compute the RMSE and LL and choose the K and burn-in to be the one that gives the most suitable results. Where K_{min} in this problem setting is 200 but can be different for other models such as when we have more basis weights or a larger number of spectral density parameters.

This brings us to an important note about the blocked Gibbs settings. Although from this analysis we can take away some defaults for the next sections, these would not in general be the optimal ones. We have, however, verified the metrics used in this section. Using the IF or BM-$\mathbb{V}ar$ and some fit metric the algorithm can be optimized for mixing and identification and this can be done for each model. We can also build a dictionary of various

Figure 6.9: Trace plots of some the parameters in the joint posterior, but now with the Half-Normal prior for the spectral density scale parameter. For $T = 500$.

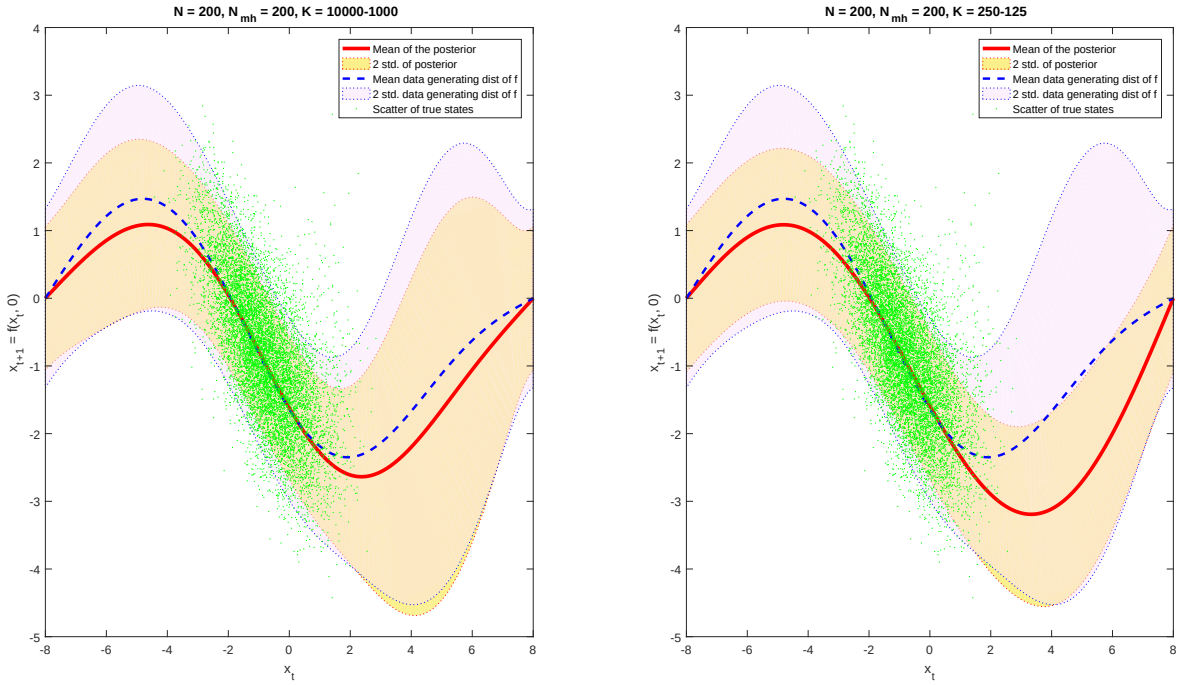


spectral densities and hyper-priors for optimization. The optimization can be achieved using methods similar to those in this thesis (see Shahriari et al. (2016)).

Spectral mixture kernel Within the spectral density of convolution kernel we also experimented with placing the Gaussian mixture in the input dimension. Where we set the number of mixture components $q = 10$, and use the same priors for the signal variances, frequencies and inverse length-scales as discussed in chapter 3.

After performing a similar analysis, which we omit for brevity here, we found that we need more K, N, M_{mh} to obtain similar performance as for the Matern spectral density, which is a price to be paid for a higher degree of generality. Furthermore the distribution over the μ's of the spectral mixture kernel is highly multi-modal making estimation hard. The fact that the amount of spectral density parameters increases by $3 \times q$, which in turn increases the needs for K, N, M_{mh} make it computationally more costly. We found $N = N_{mh} = 400$ with $400 \leq K \leq 1000$ to produce similar results as with the Matern kernel in both dimensions of the spectral density.

Figure 6.10: Summary of the predictive distribution of the estimated model, with $\hat{f}$ in only its $\mathbf{x}_t$ argument, versus that of the DGP to compare the identification in the short and long run. Note that this summary implicitly includes information about all the distributions of the parameters. For $T = 10000$.



6.2 Simulations with state function of the form $f : \mathbb{R}^2 \to \mathbb{R}$

In this section we expose the model and learning algorithm to a more general setting, where the DGP is now directly the function as in step 1 of the DGP-setup in the previous section (Equation 6.1). The aim is to see whether the MCMC settings found in the previous section provide sufficient performance results, to find how the basis function expansion order m influences estimation, what the minimal m is, and how the different spectral densities compare in terms of estimation performance as well as computation time.

DGP We simulate $\{\mathbf{y}_t, \mathbf{x}_t\}_{t=0}^{T+100}$ following:

$$\mathbf{x}_0 = \mathbf{y}_0 = 0 \tag{6.10a}$$
$$\mathbf{x}_{t+1} = f_i(\mathbf{x}_t, \mathbf{y}_t) + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim \mathcal{N}(0, 0.4) \tag{6.10b}$$
$$f_i(\mathbf{x}_t, \mathbf{y}_t) = g_i(\mathbf{x}_t) + h_i(\mathbf{y}_t) \quad t = 1, \dots, T + 100, i = 1, 2 \tag{6.10c}$$
$$g_i(\mathbf{x}_t) = \alpha_i \mathbf{x}_t \quad \alpha_1 = 0.5, \alpha_2 = 0.3 \tag{6.10d}$$
$$h_1(\mathbf{y}_t) = -0.05\mathbf{y}_t - 2.5\frac{\mathbf{y}_t}{1 + \mathbf{y}_t^2} \tag{6.10e}$$
$$h_2(\mathbf{y}_t) = -0.3 \cdot \mathbb{1}_{\mathbf{y}_t<0}(\mathbf{y}_t) + 0.15 \cdot \mathbb{1}_{\mathbf{y}_t\geq 0}(\mathbf{y}_t) \tag{6.10f}$$
$$\mathbf{y}_t|\mathbf{x}_t \sim \mathcal{N}(\mathbf{y}_t|0, \exp(\mathbf{x}_t)) \tag{6.10g}$$

Where of course only $\{\mathbf{y}_t\}_{t=100}^{T+100}$ is used as data and the first 100 samples are purged as burn-in. Note that in this case the DGP state functions $f_i, i = 1, 2$ are in fact separable and we experimented with both separable and non-separable model formulations in the estimation procedure and both worked fine. We do make the choice however to keep the model formulation as in Equation 3.20. The second function resembles in form the asymmetry as in the stylized leverage effect in financial time series literature and is inspired by the one-dimensional benchmark function of Frigola, Y. Chen, et al. (2014b). The first function is the negative of a popular benchmark function for non-liner dynamical system identification (Gordon et al. 1993), which in this simulation case constitutes a function in only one input dimension of the state-function. In both cases the state function in asymmetrical in its input term.

6.2.1 Matern/Matern Spectral density

With the spectral density being that of the convolution kernel wit the Matern class specification in both its dimensions, $200 \leq K \leq 300$, $N = 200$, $N_{mh} = 200$ and $T = 500$ we run the Blocked Gibbs sampler for data generated by the DGP with f_1 as its state function. The spectral density hyper-priors are $\sigma_x, \sigma_y \sim \mathcal{HN}(\sigma_\cdot|0, 50)$, $\ell_x, \ell_y \sim \mathcal{N}(\ell_\cdot|10, 1)$ and $\nu_x, \nu_y \sim \text{Exp}(\nu_\cdot|10)$. We set up a simple heuristic algorithm that starts with $m = l^{\dim(\mathbf{x}_T)+\dim(\mathbf{y}_T)}$ and then increases l by 1 until the 6th step at which point it relates the performance of the next iteration with the previous ones. It stops if the next iteration is not better than the previous 6 in terms of RMSE. The results in Table 6.3 suggest that for 8^2 basis functions we obtain sufficient results, which is indicated by this number being the minimum of the best 5 expansion truncations.

Besides having some quantitative idea of the predictive performance, noting that for each input $(\mathbf{x}_t, \mathbf{y}_t)$ the next state is time-invariably determined by f_1, we can plot estimated states at $t + 1$ for inputs on a grid fed into $\hat{f}_1 = \bar{\mathbf{W}}\boldsymbol{\phi}$, where $\bar{\mathbf{W}}$ is the mean of the posterior draws. We look at the properties of the estimation for $m = 8^2$. This is done in Figure 6.11 and the estimations are relatively accurate, more importantly the uncertainty is increased at

regions of the state-space where there is little (filtered) data available. In the same figure the two distributions over the cross section of the predicted function surface $(\hat{\mathbf{x}}_{t+1}, \mathbf{x}_t, \mathbf{y}_t)$ with $\mathbf{x}_t = 0$ we see a dramatic difference between the lowest expansion order $m = 2^2$ and $m = 8^2$. A similar difference is not found for the cross section of the predicted function surface with $\mathbf{y}_t = 0$. We do observe better uncertainty quantification in the right lower panel.

Table 6.3: Estimation on $T = 500$ data generated with f_1 in Equation 6.10. The table is ordered by the RMSE with the smallest first. The lowest m^2 with the best RMSE and LL is 8^2. Run column gives the run $200 \leq K \leq 300$ at which the highest RMSE was achieved and sets K to that, then the burn-in is computed for the K. The spectral density consists of the Matern in both dimensions.

M	RMSE	LL	RUN	IF	BM-$\mathbb{V}ar$	TrainTimes	AR
12^2	0,2051	-0,7027	200	0,1188	2,6103	12,0831	0,1137
14^2	0,2187	-0,7584	216	0,1414	2,1091	18,5135	0,1052
11^2	0,2307	-0,7483	252	0,1859	2,5339	8,2220	0,1675
13^2	0,2468	-0,7589	227	0,0634	2,1404	16,4612	0,1612
8^2	0,2513	-0,7055	299	0,2444	5,3308	2,9304	0,2484
10^2	0,2550	-0,7404	234	0,1386	4,4032	5,2122	0,1238
9^2	0,2882	-0,7451	244	0,3465	8,8616	3,6802	0,2332
6^2	0,2978	-0,7992	267	0,9423	13,1012	1,9933	0,2568
7^2	0,3241	-0,7901	300	0,1089	11,1768	2,2537	0,1845
15^2	0,3260	-0,6897	237	0,0829	6,2093	21,7522	0,1264
4^2	0,3816	-0,8882	257	0,2435	12,9391	0,8701	0,2306
5^2	0,3894	-0,8556	274	0,4114	14,0666	1,5037	0,2066
2^2	0,4592	-0,9327	260	0,8787	13,1098	0,2721	0,6083
3^2	0,4640	-0,9374	237	0,8560	13,5691	0,3603	0,4844

In the upper two panels in Figure 6.11 it looks as though the signal variance is estimated to be too high resulting in the predicted function values to be sensitive to scarce data points in the outskirts of the state-space. We can delve into the joint posterior and look at the means of the marginals of the spectral density parameters. We then find that for $\sigma_x \sigma_y$ the values are around 10 suggesting indeed somewhat of a large signal variance. The regularity parameters ν_x, ν_y (degrees of freedom) are around 1 and 8 respectively indicating a desired level of smoothness if we look at the DGP. Given the intertwined relationship between filtering the states and estimating the state function we can also look closer at the filtering performance.

In Figure 6.12 we see the state marginal mean and two standard deviations from the mean of the joint posterior $\hat{\mathbb{P}}_{post}(\mathbf{x}_{0:T}, \mathbf{W}, \mathbf{Q}, \boldsymbol{\theta}_f | \mathbf{y}_{0:T})$. We see that the means are not too far away from the true states and the true states are almost always within the 2 standard deviations of the mean of the posterior.

In Figure 6.13 we visualise the mixing of the sampler for a random selection of weights and a couple of expansion orders, which although not optimal can be deemed to be acceptable.

Figure 6.11: Posterior distribution over the predicted function values of $\hat{\mathbf{x}}_{t+1} = \hat{f}_1(\mathbf{x}_t, \mathbf{y}_t)$ for $m = 2^2, 8^2$. The upper plot is the cross section of the surface of the estimated function with $\mathbf{x}_t = 0$ and the lower graph is the same idea but with $\mathbf{y}_t = 0$. The shaded areas are the areas between l times the standard deviations from the mean of the posterior on both sides.

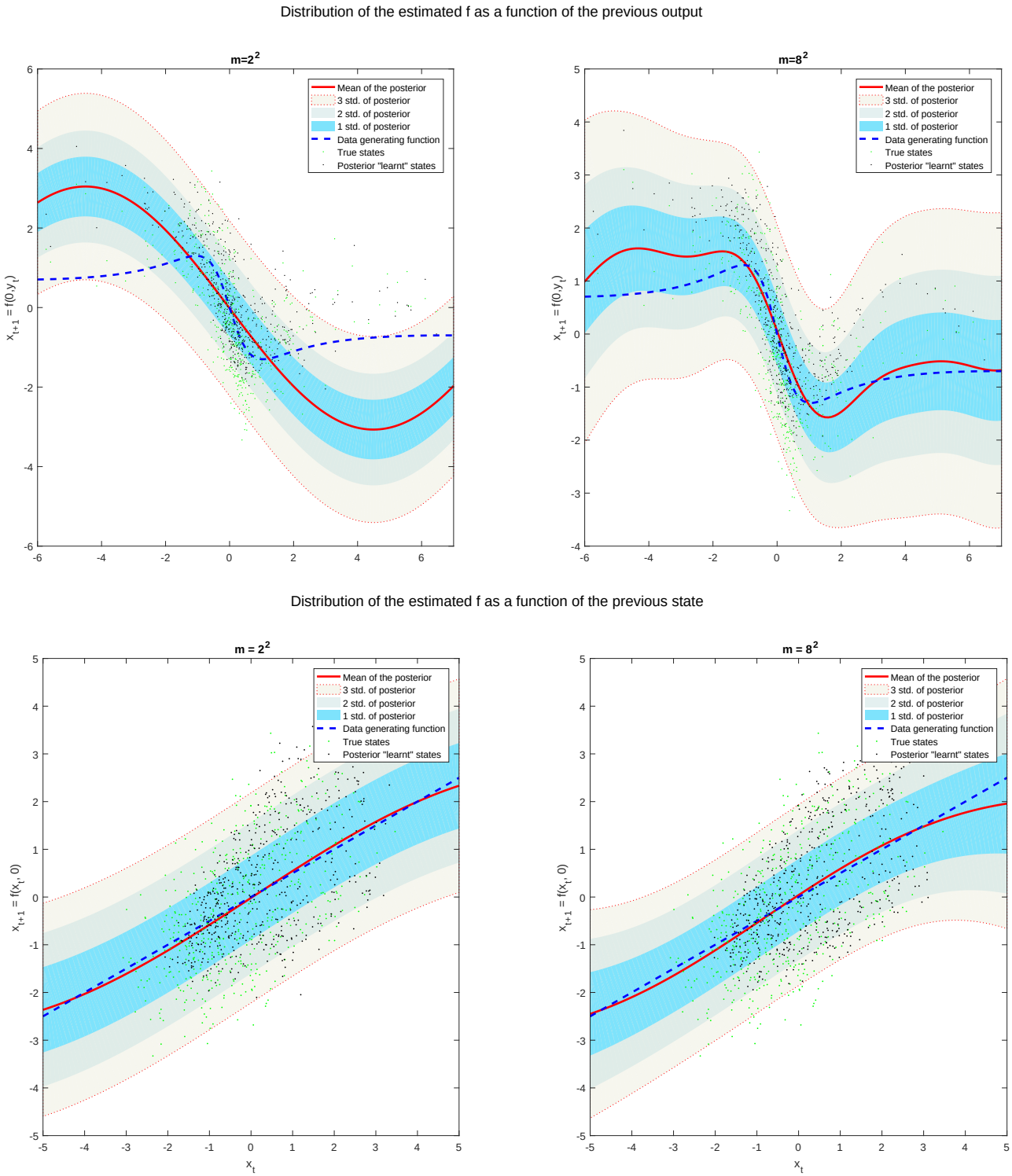
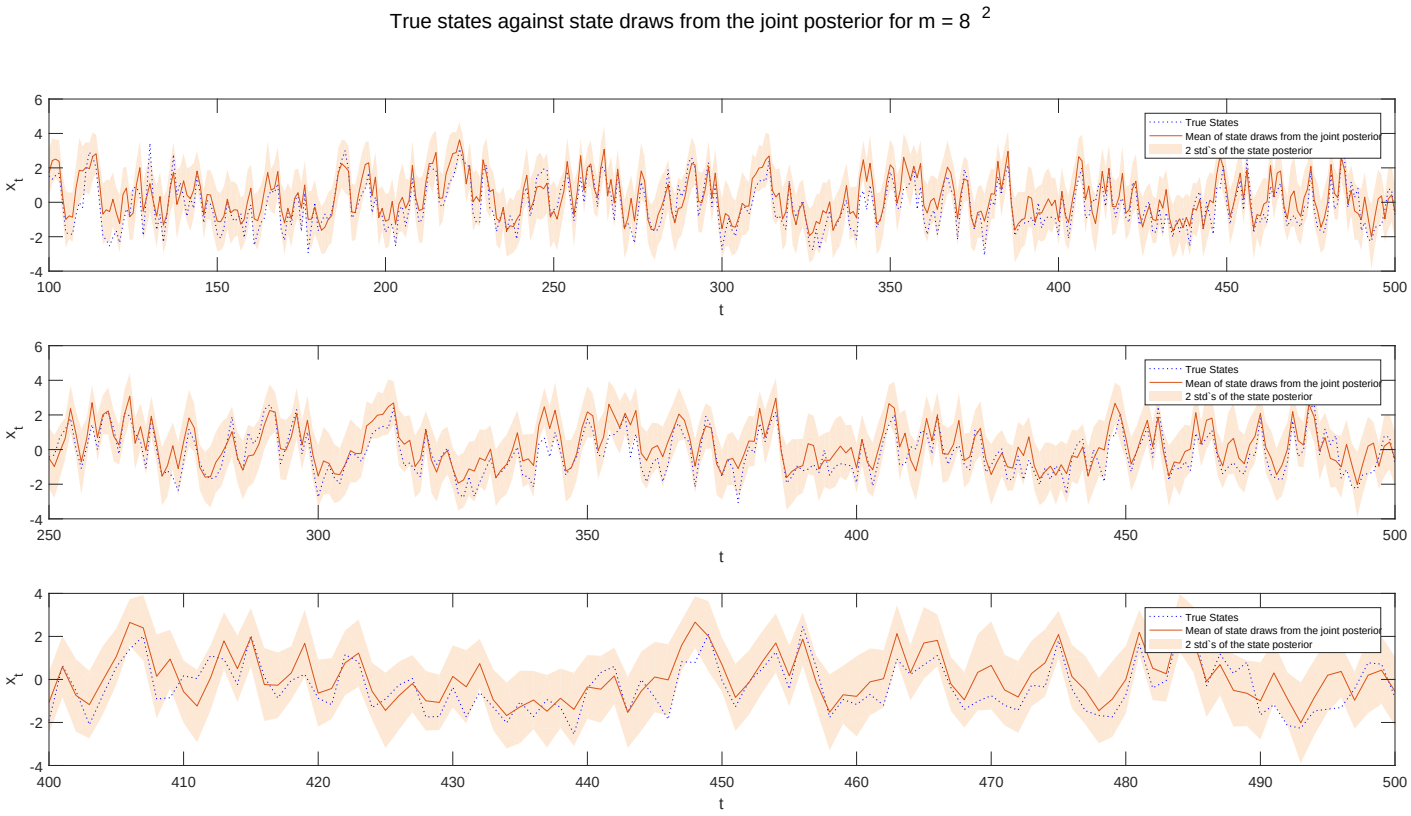


Figure 6.12: The means and 2 standard deviations from the mean of the state draws from the joint posterior are given. These state draws are obtained by the conditional particle filter, where at each step the state function estimate $\hat{f}_1(\mathbf{x}_t, \mathbf{y}_t)$ is used. $N = 200, N_{mh} = 200, K = 299, T = 500$.



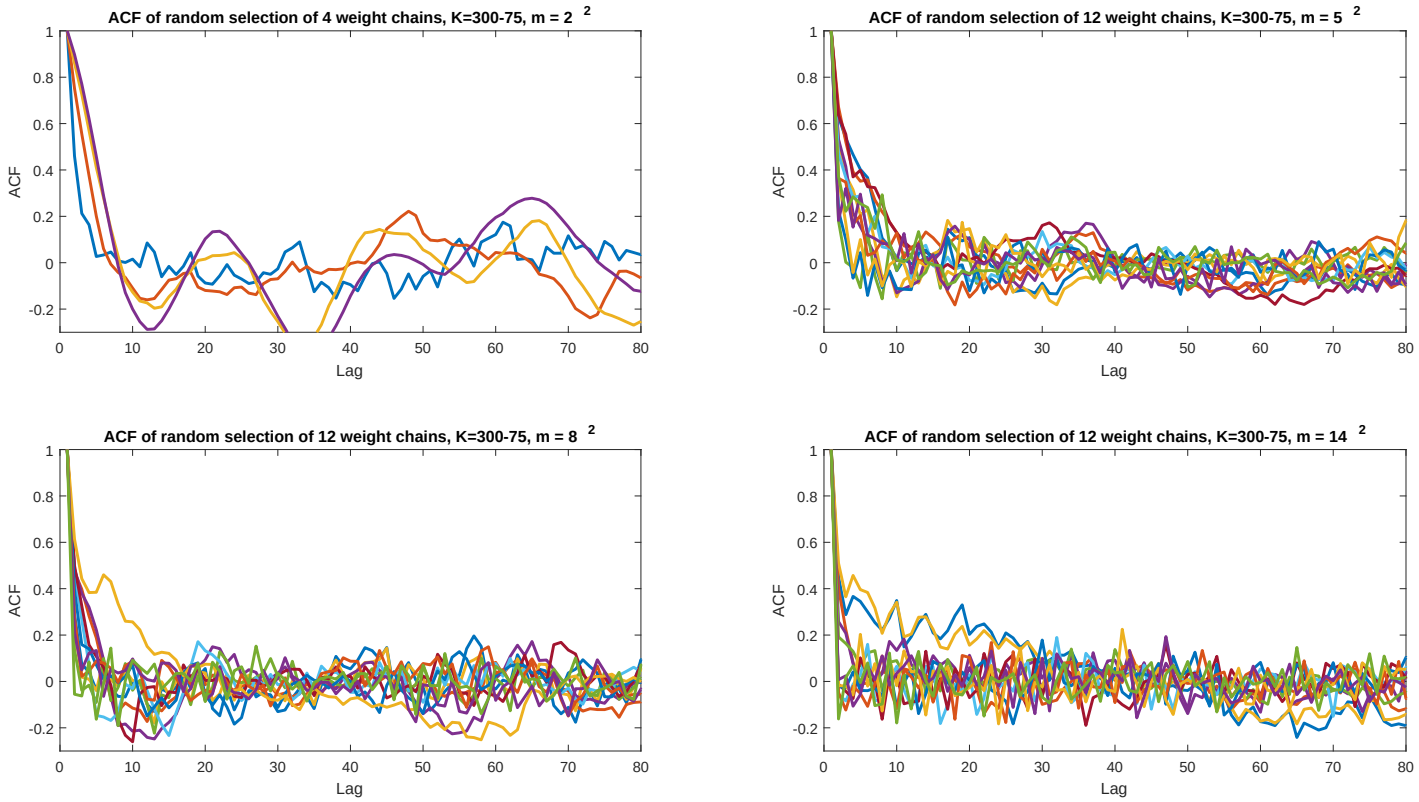
Conclusions of this subsection Given the results in table Table 6.4 and the figures in Figure 6.11 and Figure 6.12 we can conclude that: 1.) it does seem as though the number of K, N, M_{mh} found in the previous section are sufficient for descent estimation results. Furthermore 2.) for not too high of an expansion order, (8^2) for example, the estimation results are sufficient and the computation times are feasible. Higher orders of m^2 do seem to result in better estimation but are not needed, given that feasible computation times are the reason to approximate the full GP-SSM with a Bayesian basis function expansion.

For the second function f_2 in equation Equation 6.10 the results are in table Table 6.4. Again the acceptance rates and computation time are "acceptable". $m^2 = 6^2$ seems to be working well in terms of RMSE, LL, and computation time and the function estimation is given in Figure 6.13. The difference in estimation as a result of the expansion order seems more dramatic for this function. But again we can draw the same conclusions as for the estimation on data generated with f_1.

6.2.2 Matern/Gaussian mixture Spectral density

We begin with estimating the model with the Matern/Gaussian-mixture spectral density with data generated using the DGP of Equation 6.10 using f_1. The hyper-priors for the

Figure 6.13: For the estimation of data generated by f_1 in Equation 6.10 the ACF of a random number of weights are given to visualise the mixing behavior of the sampler, using the posterior draws. $N = 200, N_{mh} = 200, K = 299, T = 500$.



Matern kernel are the same as in the previous section, those for the Gaussian mixture kernel as in chapter 3 and the number of mixture components q is set to 10. The results are in table Table 6.5 for $N = 400$, $N_{mh} = 400$ and $T = 500$ and three things immediately stand out. Firstly we observe low acceptance rates of the Metropolis-within-Gibbs algorithm, secondly the long training times and thirdly the higher RMSE in comparison to Table 6.3. These results did not become much better when playing with the random walk Metropolis-within-Gibbs tuning parameter (the variance of the innovations of the random walk).

These results render the use of the Gaussian-mixture kernel in combination with the current learning algorithm as slow and bad converging. We would need a much higher K or/and N_{mh}, which would increase estimation time drastically. However the bad characteristics can be easily explained. Given that the number of spectral density parameters increases from 12 to 36 the resulting longer computation times are not surprising. Furthermore the μ's of the Gaussian mixture, or as mentioned in chapter 3 the frequencies, are hard to learn for the algorithm as the marginal likelihood (the constant in the Metropolis-within-Gibbs) is the most multimodal in these parameters. The Metropolis-within-Gibbs could for instance be modified to iterate over subsets of the hyper-prior components in blocks (P. Neal and G. Roberts 2006), or we could employ slice sampling (R. M. Neal 2003) rather than Metropolis-within-Gibbs. We did not have time to experiment with these.

Table 6.4: Estimation on $T = 500$ data generated with f_2 in Equation 6.10. The table is ordered by the RMSE with the smallest first. Run column gives the run $200 \leq K \leq 300$ at which the highest RMSE was achieved and sets K to that, then the burn-in is computed for the K. The spectral density consists of the Matern in both dimensions.

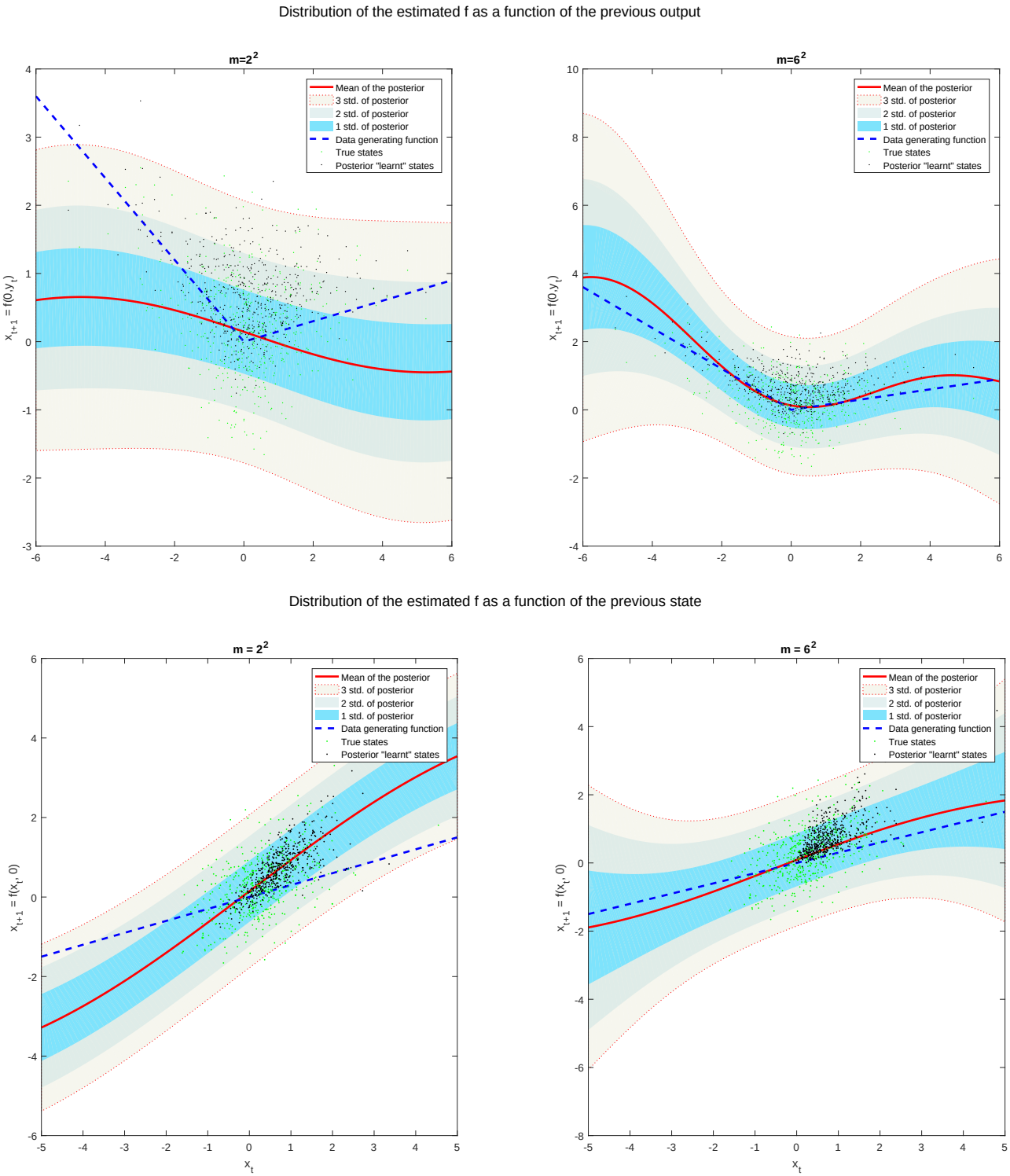
M	RMSE	LL	RUN	IFF	BM-$\mathbb{V}ar$	TrainTimes	AR
8^2	0,0962	-0,7089	245	0,4312	8,1160	4,6452	0,1968
6^2	0,1062	-0,6827	295	0,4890	14,1416	3,9352	0,2352
9^2	0,1076	-0,7135	280	0,3197	7,3779	6,5869	0,2310
3^2	0,1094	-0,7657	253	0,6740	21,8345	0,7897	0,5665
11^2	0,1107	-0,6951	236	0,3221	3,8877	10,6436	0,2377
12^2	0,1144	-0,5728	223	0,1847	4,4911	14,4589	0,1236
7^2	0,1147	-0,6247	231	0,4605	17,1908	2,5658	0,1788
5^2	0,1176	-0,7259	246	0,6995	16,5257	2,6470	0,2933
10^2	0,1258	-0,6867	200	0,2202	6,1588	8,2280	0,1459
13^2	0,1323	-0,6111	202	0,2198	4,5141	16,3440	0,1185
4^2	0,1554	-0,6028	300	0,7453	19,2209	1,2102	0,2306
2^2	0,2455	-0,7095	272	1,0432	12,2456	0,3871	0,2066

Conclusions of this subsection From this section we can conclude that the Matern/Gaussian-mixture is likely to not be useful in a rolling window setting. And also, although more general than the Matern/Matern spectral density, it is not so successful in combination with the Metropolis-within-Gibbs algorithm within our Blocked Gibbs scheme.

Table 6.5: Estimation on $T = 500$ data generated with f_1 in Equation 6.10. The table is ordered by the RMSE with the smallest first. Run column gives the run $400 \leq K \leq 1000$ at which the highest RMSE was achieved and sets K to that, then the burn-in is computed for the K. The spectral density consists of the Matern in the state dimension and the Gaussian mixture in the input dimension.

M	RMSE	LL	RUN	IF	BM-$\mathbb{V}ar$	TrainTimes	AR
12^2	0,2306	-0,6607	914	0,5681	6,1423	67,0997	0,0704
10^2	0,2805	-0,7521	538	0,2688	7,4784	47,5683	0,0728
6^2	0,3007	-0,8188	979	0,2966	9,3974	17,7987	0,0706
11^2	0,3095	-0,7715	477	0,5974	3,287	64,0851	0,0707
14^2	0,3182	-0,7116	435	0,4661	4,9591	87,3298	0,0703
13^2	0,3189	-0,8142	930	0,291	3,4597	104,5666	0,0707
9^2	0,339	-0,8741	465	0,1603	3,6745	46,1606	0,0730
8^2	0,3448	-0,8282	1000	1,3741	8,8067	40,8007	0,0712
7^2	0,3465	-0,8395	959	0,3153	7,1939	26,3989	0,0719
4^2	0,4227	-0,9376	494	3,1955	15,4925	10,1018	0,0712
5^2	0,4787	-1,0648	413	1,4922	6,7382	16,1833	0,0705
3^2	0,4824	-1,0685	590	2,3798	11,26	4,8456	0,0706
2^2	1,1308	-1,6086	476	0,2315	8,7136	2,2131	0,0708

Figure 6.14: Posterior distribution over the predicted function values of $\hat{\mathbf{x}}_{t+1} = \hat{f}_2(\mathbf{x}_t, \mathbf{y}_t)$ for $m = 2^2, 6^2$. The upper plot is the cross section of the surface of the estimated function with $\mathbf{x}_t = 0$ and the lower graph is the same idea but with $\mathbf{y}_t = 0$. The shaded areas are the areas between l times the standard deviations from the mean of the posterior on both sides.



7 Empirical study

Contents

7.1 In-sample Estimation

The questions we aim to answer here is how well does the model in combination with the sampling algorithm perform in terms of structure discovery and computation time. In particular, is it able to capture the leverage effect and volatility clustering in the estimated transition function, and do that in an acceptable amount of estimation time. Also we want to compare the models with various expansion orders $m = l^{\dim(\mathbf{x}_t)+\dim(\mathbf{y}_t)}$ and finally compare the spectral densities on data. The hyper-priors on the spectral density parameters are the same as in the simulation section.

In Table 7.1 the results are visible for both indices and both stocks. We compare the various basis function expansion orders using the marginal likelihood, but rather than using 2 times the log of the Bayes factor as Kass and Raftery (1995) suggest we heuristically give preference to the highest marginal likelihood unless pairwise comparisons are too similar. As in the simulation setting, and as expected, we immediately notice the low acceptance rates of the Metropolis-within-Gibbs for the model with the Matern/Mixture (Mixture is the spectral density of the Gaussian mixture kernel) spectral density for all indices and stocks. We also notice the long training times for this model. Furthermore If we compare the best model of the subset of models with the Matern/Matern spectral density with those with the Matern/Mixture spectral density, for all stocks and indices considered, we find strong evidence in favor of the best Matern/Matern spectral density. The reasons for this have been discussed in the simulation chapter. If we take the sample size of the data in Table 7.1 into consideration then the lowest expansion order in the top 3 best models all have acceptable computation times.

Table 7.1: Comparisons between expansion orders as well as spectral densities using the log of the marginal likelihood. Note that if $2 \times (\text{MargLL(Model-A)} - \text{MargLL(Model-B)}) \geq 6$ there is strong evidence in favor of model A.

NIKKEI 225

Spectral Density	Matern/Matern						Matern/Mixture				
M	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**	**M**	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**
8^2	-3545,5061	0,4450	16,9766	22,0711	0,5822	7^2	-3555,3297	0,7939	4,9161	33,3887	0,0842
10^2	-3553,1382	0,1932	15,6013	30,8991	0,4184	8^2	-3561,9545	0,2672	2,7234	45,9852	0,0818
7^2	-3556,3949	0,5870	21,3129	18,1350	0,4296	10^2	-3575,7377	0,9331	4,5288	61,4584	0,0537
6^2	-3557,8098	0,3833	23,4303	14,9485	0,3198	5^2	-3581,2763	1,2532	13,5545	27,3606	0,0345
9^2	-3559,7272	0,3440	16,6364	26,7598	0,2401	4^2	-3609,3366	2,0065	19,4731	19,7598	0,0208
5^2	-3561,0299	0,9640	35,2560	12,7226	0,2198	6^2	-3631,2742	0,3252	1,9010	30,5890	0,0216
4^2	-3565,2753	1,3861	41,4145	9,8894	0,1709	9^2	-3655,9509	0,3574	12,7967	59,4213	0,0266
T = 2464	1986-01-07 to 1995-12-29										
Tuning	$K = 300$	$N = 400$	$N_{mh} = 200$				$K = 700$	$N = 400$	$N_{mh} = 400$		

S&P 500

Spectral Density	Matern/Matern						Matern/Mixture				
M	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**	**M**	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**
7^2	-2982,5559	0,3301	11,4865	17,3525	0,4464	4^2	-2988,7245	1,8531	18,9549	18,8870	0,0809
9^2	-2983,5337	0,4286	13,0383	24,5493	0,4873	5^2	-2995,5511	0,4278	5,0600	22,4176	0,0811
4^2	-2984,1655	1,0679	42,8883	8,2796	0,3676	6^2	-3013,9229	0,3499	5,1636	34,8918	0,0807
10^2	-2984,5080	0,5181	20,3355	27,9928	0,2914	9^2	-3026,2246	0,3373	3,8973	57,2700	0,0423
8^2	-2987,6290	0,4110	18,6092	20,8193	0,2432	7^2	-3034,7724	0,4706	7,7050	33,7809	0,0548
6^2	-2990,1642	0,6097	29,3235	10,7441	0,1609	8^2	-3041,8243	0,2229	4,1970	44,8114	0,0213
5^2	-2991,1016	0,8399	29,1361	9,5557	0,1380	10^2	-3075,8503	0,5417	7,4407	62,8865	0,0220
T = 2264	2006-01-04 to 2014-12-31										
Tuning	$K = 300$	$N = 200$	$N_{mh} = 200$				$K = 700$	$N = 400$	$N_{mh} = 400$		

ABB

Spectral Density	Matern/Matern						Matern/Mixture				
M	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**	**M**	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**
10^2	-2770,9243	1,3706	49,0657	18,4869	0,4754	4^2	-2793,9973	0,9692	7,0307	15,6392	0,1025
4^2	-2779,5096	0,4393	19,6284	2,8172	0,4664	7^2	-2796,3069	0,2164	3,2329	20,6881	0,0712
9^2	-2780,7336	0,1122	24,2782	15,4953	0,4466	8^2	-2867,1293	0,4497	7,2259	24,4651	0,0826
7^2	-2781,7040	0,7149	39,4548	9,8573	0,4045	6^2	-2900,5108	0,5165	5,5070	17,8691	0,0719
6^2	-2783,7086	0,6262	47,6346	7,7976	0,3422	10^2	-2915,7044	0,2698	5,2546	33,5755	0,1134
5^2	-2785,6717	0,9657	63,2729	3,8424	0,2822	9^2	-2927,9461	0,5330	6,6149	28,4801	0,0816
8^2	-2787,4215	0,1759	32,7872	12,2054	0,1740	5^2	-2934,4255	0,6126	23,1011	16,9464	0,0907
T = 1259	2007-01-03 to 2011-12-30										
Tuning	$K = 300$	$N = 200$	$N_{mh} = 200$				$K = 700$	$N = 400$	$N_{mh} = 400$		

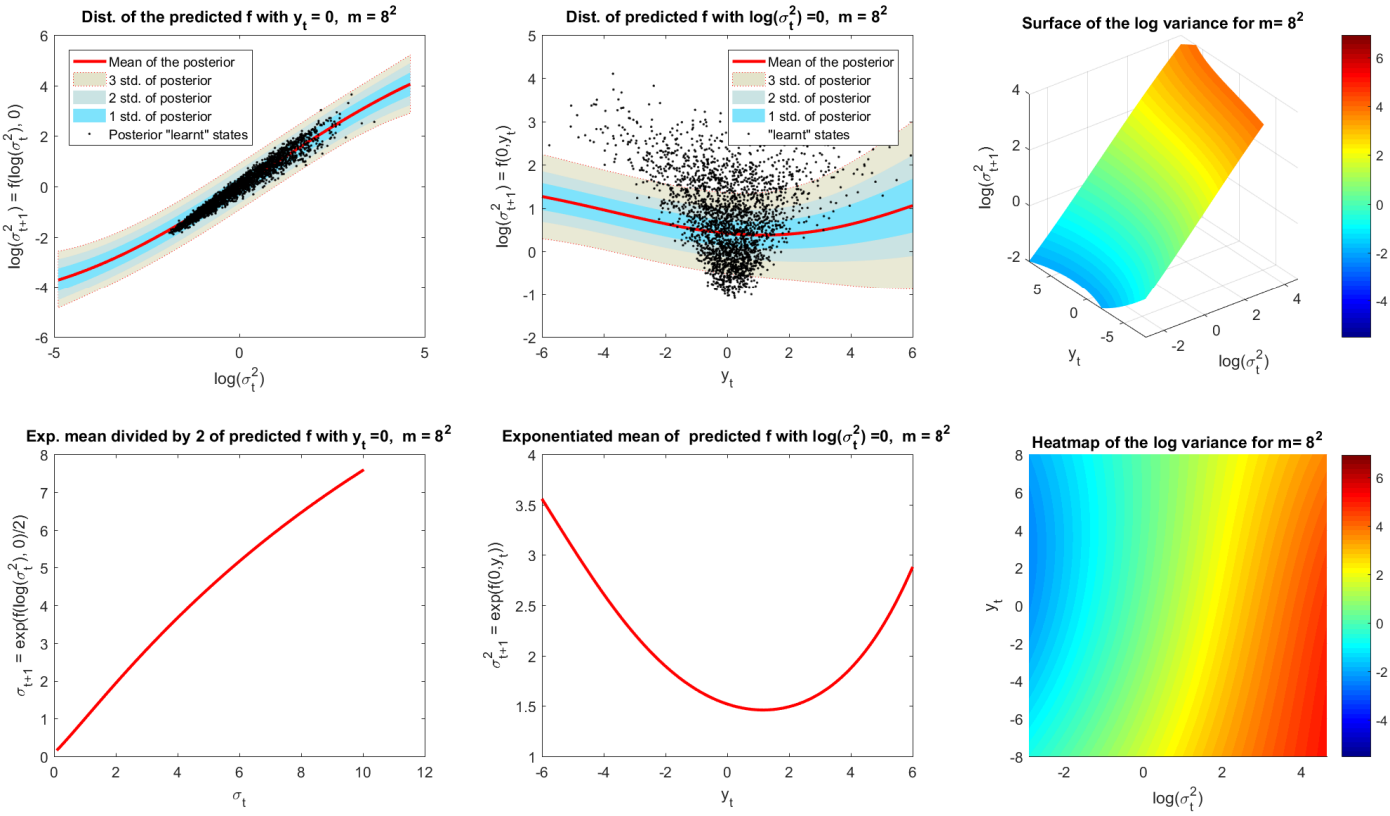
PEPSICO

Spectral Density	Matern/Matern						Matern/Mixture				
M	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**	**M**	**MargLL**	**IF**	**BM-$\mathbb{V}ar$**	**TrainTimes**	**AR**
7^2	-1715,2393	0,6735	41,8284	15,2337	0,5245	4^2	-1747,5967	0,9357	8,8554	16,4262	0,0808
5^2	-1717,6027	0,9709	37,5632	7,7500	0,4234	6^2	-1753,3001	0,7164	10,1557	30,1289	0,0835
10^2	-1721,8174	0,3588	12,2553	25,3842	0,3175	9^2	-1759,4601	0,4707	4,6087	39,1568	0,0811
9^2	-1725,5682	0,2872	18,9829	21,6299	0,2949	10^2	-1763,5770	0,4434	6,0569	47,1167	0,0514
4^2	-1726,5778	0,5203	25,0786	4,0586	0,1714	7^2	-1769,4350	0,5924	6,8588	31,2167	0,0415
6^2	-1727,3012	0,8731	22,0173	12,7664	0,1278	8^2	-1780,1563	0,4264	3,0263	32,6489	0,0341
8^2	-1727,4287	0,2410	16,3679	17,7999	0,1206	5^2	-1789,1551	0,3674	5,9569	25,9513	0,0275
T = 1505	2011-01-04 to 2016-12-23										
Tuning	$K = 300$	$N = 200$	$N_{mh} = 200$				$K = 700$	$N = 400$	$N_{mh} = 400$		

The first thing we conclude from these results is, as before, that the Matern/Mixture spectral density in not suitable for the models in the rolling window setting in combination with the current inference algorithm. This is because of the high computation times, and that the best truncations with this spectral density are consistently outperformed by those with a Matern/Matern one in terms of marginal likelihood.

Secondly now that we have four model choices for the two stocks and two indices considered, we can look inside the estimated model to see if the learnt structure adheres to the stylized facts discussed in chapter 5. For this as before we plot cross-sections of the predicted transition function surface as well relevant areas of the surface. Thus we plot for grid inputs $(\mathbf{x}_t^*, \mathbf{y}_t^*)$ the distribution over the predicted states $\hat{\mathbf{x}}_{t+1} = \hat{f}(\mathbf{x}_t^*, \mathbf{y}_t^*)$, where $\mathbf{x}_t = \log(\sigma_t^2)$ and $\mathbf{y}_t$ the percentage growth based on the adjusted daily closing price.

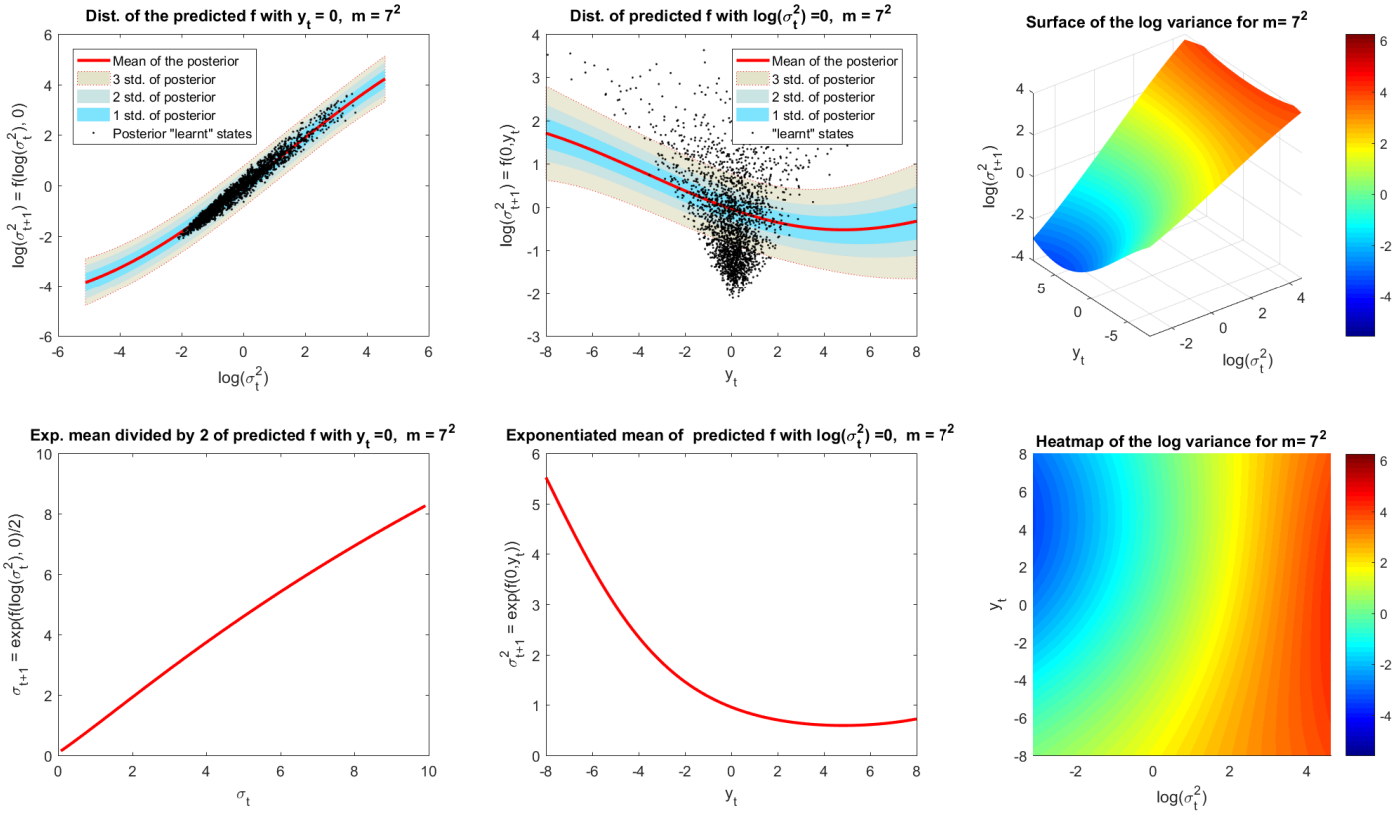
Figure 7.1: NIKKEI 225 results for the Matern/Matern spectral density. Given the large number of hyper-parameters and weights (71 for $m = 8^2$) the function surface cross-section summaries in the first two upper panels are the most condense yet informative way to provide the distribution over the predicted function as in Equation 6.3. The other panels are derivatives of these summaries. For $K = 300, N = 200, N_{mh} = 200, T = 2646$.



Beginning with the NIKKEI 225 index, in Figure 7.1 we have the relevant graphs. In the upper left panel the cross section of the predicted distribution of the log variance surface is given for $\mathbf{y}_t = 0$. In this panel we see that the functional form, although not linear, seems to be in line with the clustering stylized fact in financial time series. In the lower

left panel the exponent divided by two of the mean of same distribution is given. In the upper middle panel the cross section of the predicted distribution is given with $\log(\sigma_t^2) = 0$, which is where we expect to observe the leverage effect. The asymmetry of the estimated prediction as a function of previous percentage growth is clearly visible. It does look as though after $\mathbf{y}_t = 0$ and $0 < \mathbf{y}_t$ the estimated volatility keeps decreasing, and given my knowledge on financial time-series it is unclear to me whether or not this is reasonable. The upper and lower right panels show the surface of the predicted mean of the log variance surface in 3d and as a heat map. If we look at the heat map we also see the asymmetry for various levels of previous $\log(\sigma_t^2)$ where the level of the previous volatility seems to determine the intensity of the asymmetry. Higher previous volatility seems to results in higher asymmetry between the return and volatility.

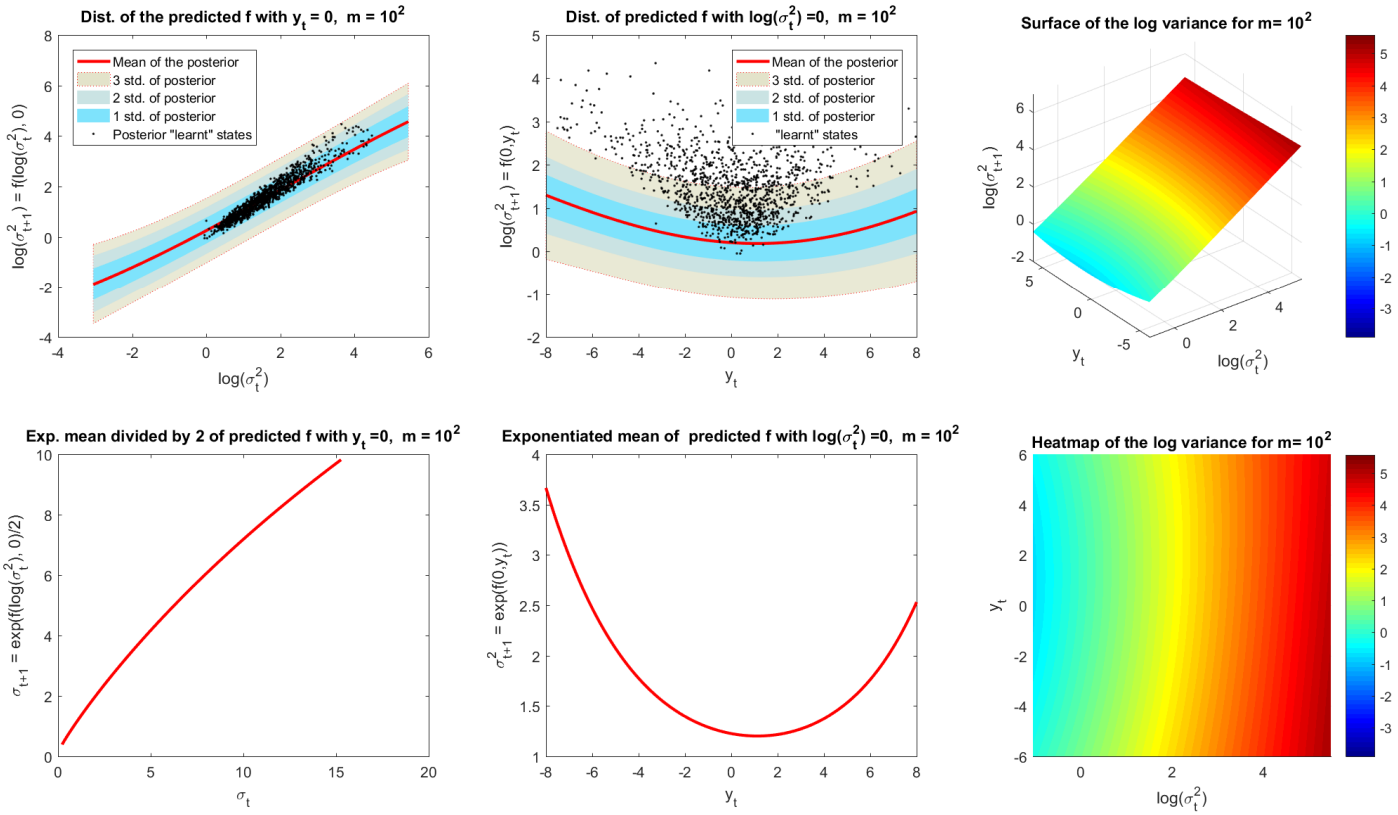
Figure 7.2: S&P 500 results for the Matern/Matern spectral density. For $K = 300, N = 200, N_{mh} = 200, T = 2264$.



In Figure 7.2 the same panels are given for the S&P 500 index and again the leverage effect is visible, as well as the clustering specification. Note that the slope of the volatility as a function of the previous volatility, in the left bottom panel, is almost 1 indicating a stronger persistence than in the previous index. Furthermore, the leverage effect seems to be more pronounced.

For the AAB stock in Figure 7.3 the same figures show a less pronounced leverage-effect. This is good to see because it shows that the model and estimation do seem to be capturing

Figure 7.3: ABB results for the Matern/Matern spectral density. For $K = 300, N = 200, N_{mh} = 200, T = 1259$.



the characteristics of the data. And finally the estimation results are given for the PEPSICO stock in Figure 7.4. The extracted states are graphed in Figure 7.5 and Figure 7.6.

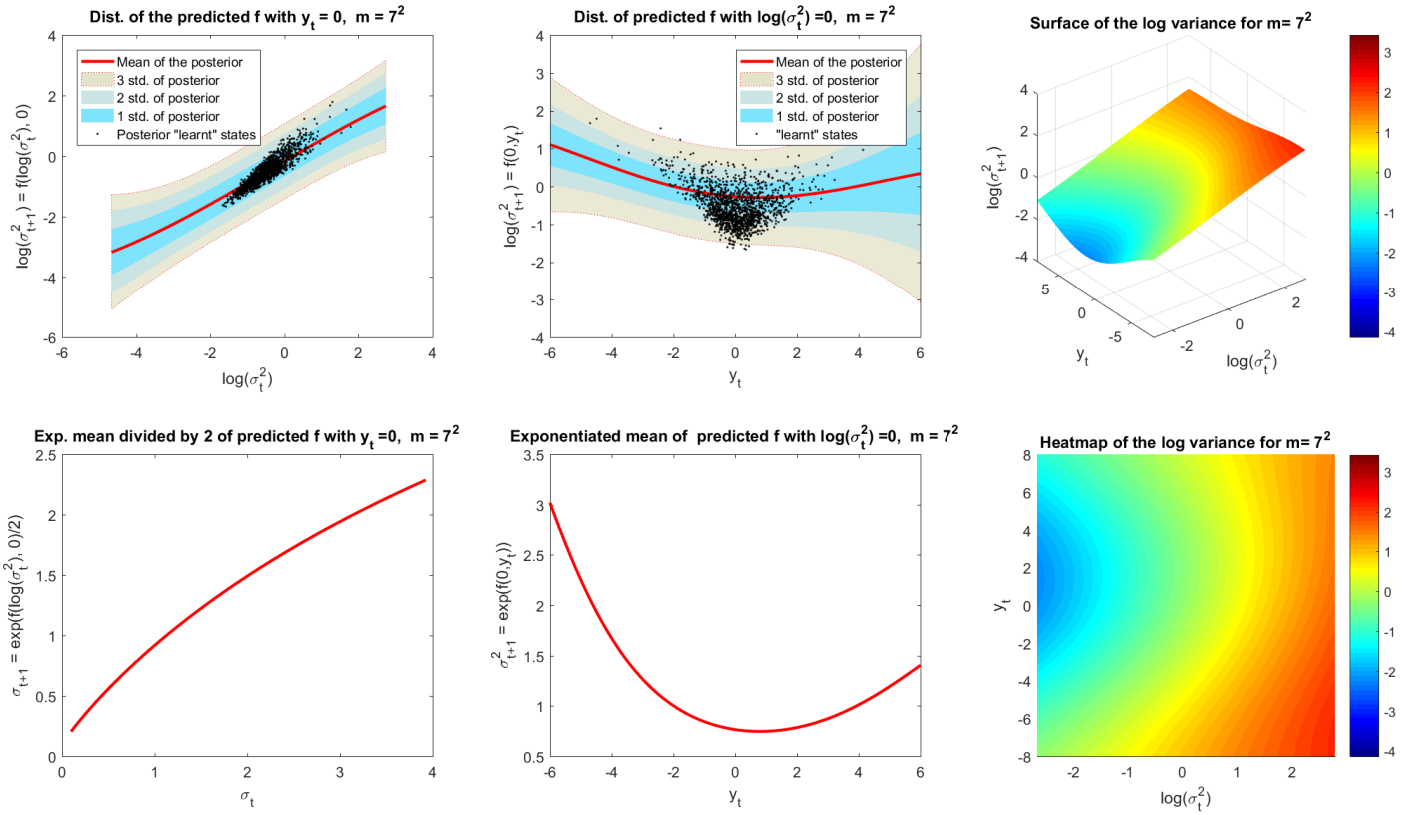
From these figures we can conclude that in terms of functional form estimation, for the graphed truncations between 7^2 and 10^2 the estimated forms reflect the leverage effect and clustering stylized facts known in financial time series.

7.2 Forecasting Performance

As a final test we can compare the forecasting performance of the RR-GPSV with that of a number of models from the GARCH family. In a comparison between various GARCH specifications Hansen and Lunde (2005) consider an exchange rates data-set and one consisting of stock returns. For the former the GARCH(1,1) was not outperformed by the other specifications whereas for the latter data-set the specifications that accommodated a leverage effect were superior. Furthermore for the stock returns data-set, on average, a Gaussian distribution for the returns showed better performance. From personal experience in a case-study[1] we find that although non-Gaussian specifications more often than not perform better, the difference in performance is not drastic. In light of all this we select the simple

[1] `https://www.dropbox.com/s/f3ti5m6r1ureuoy/Final_Report_Group02.pdf?dl=0`

Figure 7.4: PEPSICO results for the Matern/Matern spectral density. For $K = 300, N = 200, N_{mh} = 200, T = 1505$.



GARCH(1,1) model as well as two other models that accommodate a leverage effect which are the EGARCH(1,1) and the GJRGARCH(1,1) and in all cases we let the returns follow a Normal distribution (see appendix for implementation details). In this section we also employ the Deep RR-GPSV for forecasting, which was the only application time allowed for.

To evaluate the models we compared the one-step ahead predictive densities. Given $\{\mathbf{y}_t\}_{t=0}^{T}$ and the σ-algebra induced by information available at t, $\mathcal{F}_t$, we are interested in the predictive distribution:

$$p_{T+1}^{GP} = \mathbb{P}(\mathbf{y}_{T+1}|\mathcal{F}_t) = \int \mathbb{P}(\mathbf{y}_{T+1}|\mathbf{y}_{0:T}, \mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f, \mathbf{x}_{0:T})\mathbb{P}(\underbrace{\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f, \mathbf{x}_{0:T}}_{\mathcal{V}}|\mathbf{y}_{0:T})d_{\mathcal{V}} \qquad (7.1)$$

Recall that from the blocked Gibbs we obtain samples $\{\mathbf{Q}^i, \mathbf{W}^i, \boldsymbol{\theta}_f^i, \mathbf{x}_{0:T}^i\}_{i=1}^{K} \sim \hat{\mathbb{P}}(\underbrace{\mathbf{Q}, \mathbf{W}, \boldsymbol{\theta}_f, \mathbf{x}_{0:T}}_{\mathcal{V}}|\mathbf{y}_{0:T})$

thus we can approximate the integral with:

$$\mathbb{P}(\mathbf{y}_{T+1}|\mathcal{F}_t) \approx \frac{1}{K}\sum_{i=1}^{K}\mathbb{P}(\mathbf{y}_{T+1}|\mathbf{y}_{0:T}, \mathbf{Q}^i, \mathbf{W}^i, \boldsymbol{\theta}_f^i, \mathbf{x}_{0:T}^i) \tag{7.2a}$$

$$= \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\mathbf{y}_{T+1}|0, \exp(\mathbf{x}_{T+1}^i)) \tag{7.2b}$$

$$= \frac{1}{K}\sum_{i=1}^{K}\mathcal{N}(\mathbf{y}_{T+1}|0, \exp(\mathbf{W}_{\boldsymbol{\theta}_f^i}^i\phi(\mathbf{x}_T^i, \mathbf{y}_T) + \mathbf{Q}^i)) \tag{7.2c}$$

$$= \hat{p}_{T+1}^{GP} \tag{7.2d}$$

For the Gaussian GARCH methods we have that all the predictive distributions have the form:

$$\mathbf{y}_{T+1}|\hat{h}(\sigma_T^2), \mathcal{F}_t \sim \mathcal{N}(\mathbf{y}_{T+1}|0, \hat{h}(\sigma_T^2)) = \hat{p}_{T+1}^{GARCH} \tag{7.3}$$

Where $\hat{h}(\sigma_T^2)$ can be for example be $\hat{h}(\sigma_T^2) = \hat{\sigma}_T^2$ or $\hat{h}(\sigma_T^2) = \widehat{\log(\sigma_T^2)}$ for the EGARCH. The forecasting procedure is as follows: given a full sample $\{\mathbf{y}_t\}_{t=0}^{T_{sample}}$ we estimate the models on a subset of the data $\{\mathbf{y}_t\}_{t=i}^{j} \subset \{\mathbf{y}_t\}_{t=0}^{T_{sample}}$, the so called window, where $j - i + 1 = T_{window}$ is the window size. Then we evaluate the one-step-ahead predictive densities at the point $\mathbf{y}_{j+1}$, after which we set $i = i + 1, j = j + 1$ and repeat until $j + 1 = T_{sample}$. Thus we get predictive density values $\{\hat{p}_t^{GP}\}_{t=1}^{T_{sample}}$ and $\{\hat{p}_t^{GARCH}\}_{t=1}^{T_{sample}}$ and we wish to compare them. In alignment with the logarithmic scoring rule we take the log of these predictive densities, and Geweke and Amisano (2010) show that the predictive densities are directly comparable. Thus we compare the average value of these density values.

To test test whether or not the difference in predictive density values is significant we employ the Diebold-Mariano test (see Diebold and Mariano (2002) for details). For this we take the differences in Kullback-Leibler divergence between the predictive densities. In terms of Kullback-Leibler Divergence, under the assumption that there exists a true density p_t, we have $KL(\hat{p}_t^{GP}) = \mathbb{E}[\log(p_t) - \log(\hat{p}_t^{GP})]$. If we take the difference between $KL(\hat{p}_t^{GP})$ and $KL(\hat{p}_t^{GARCH})$ we have that:

$$\hat{d}_t = \log(\hat{p}_t^{GARCH}) - \log(\hat{p}_t^{GP}) \tag{7.4}$$

This sequence of differences is used to perform the Diebold-Mariano test.

In Table 7.2 the forecasting results are found for the two indices considered. For both indices, both the RR-GPSV and the Deep RR-GPSV significantly outperform the GARCH models in terms of one-step-ahead predictive density according to the DM-statistic. The Deep RR-GPSV in combination with the current Blocked Gibbs learning algorithm has bad performance/computation ratio when compared to the RR-GPSV, for this reason we discontinue the use of this model and only estimate the RR-GPSV for the stocks. Note the relatively small window size $T_{window} = 500$.

Table 7.2: Volatility forecast results for data of the two indices. AvgLL is the average of the predictive log likelihood over each one-step ahead forecast. DM_1 is the DM statistic with $KL^i = LL^i_{RR-GPSV} - LL^i_{*-GARCH}$ and DM_2 the same but with $KL^i = LL^i_{D-RR-GPSV} - LL^i_{*-GARCH}$. Avg-$TT_i$ is the average training/estimation time for each iteration of the rolling window, where $i = 1$ is for the RR-GPSV and $i = 2$ for the Deep RR-GPSV. f_1 is the transition function of the RR-GPSV as well as the first function for the Deep RR-GPSV in the iterative procedure over functions , and f_2 the second one. DM* indicates that $P < 0.01$.

NIKKEI 225

Model	RR-GPSV	Deep RR-GPSV	GARCH	EGARCH	GJRGARCH	Avg-TT_1	Avg-TT_2
AvgLL	-1,6571	-1,6511	-1,8535	-1,8455	-1,9170	3,8855	16,5402
DM_1			$6,1162^*$	$3,7231^*$	$7,8273^*$		
DM_2			$6,3001^*$	$4,2815^*$	$6,7901^*$		
Settings	$K = 300$, $N_1 = 400$, $N_{mh1} = 400$, $N_2 = 200$, $N_{mh2} = 50$						
	$T_{window} = 500$, $T_{sample} = 742$, 1993-01-05 to 1995-12-29						
	Spectral Density f_1: Matern/Matern, Spectral Density f_2: Matern/Matern, $m_1 = 7^2, m_2 = 6^3$						

S&P 500

Model	RR-GPSV	Deep RR-GPSV	GARCH	EGARCH	GJRGARCH	Avg-TT_1	Avg-TT_2
AvgLL	-1,7566	-1,8218	-2,3152	-2,0567	-2,0109	4,8696	19,1159
DM_1			$4,5764^*$	$5,2167^*$	$6,1953^*$		
DM_2			$3,7512^*$	$5,9101^*$	$4,8801^*$		
Settings	$K = 300$, $N_1 = 400$, $N_{mh1} = 400$, $N_2 = 200$, $N_{mh2} = 50$						
	$T_{window} = 500$, $T_{sample} = 756$, 2007-01-04 to 2009-12-31						
	Spectral Density f_1: Matern/Matern, Spectral Density f_2: Matern/Matern, $m_1 = 7^2, m_2 = 6^3$						

In Table 7.3 the forecasting results are given for two window sizes, $T_{windows} = 500, 1500$. Once more in all cases the RR-GPSV is significantly superior in terms of one-step-ahead predictive density according to the DM-statistic. For $T_{window} = 500$ the computation time is reasonable, whereas for $T_{window} = 1500$ the algorithm has to be adapted to allow for distributed computations and/or as the data size increases we can move away from sampling methods and employ the variational approach. For the ABB stock the average one-step-ahead predictive log likelihood value increases for all models as the window size is enlarged, whereas for the PEPSICO stock this is not the case with the GARCH models.

We can ask the question of whether or not the model estimates of the RR-GPSV are stable over time, as well as how this is affected by the sample size used for estimation. This question motivates estimating the model on a larger window size. To shed light on the question, at each 10th iteration i of the rolling window estimation we plot the means of the distribution over the state function predictions $\hat{\mathbf{x}}^i_{t+1} = \hat{f}^i(\mathbf{x}^*_t, 0) = \bar{W}^i\boldsymbol{\phi}(\mathbf{x}^*_t, 0)$ and $\hat{\mathbf{x}}^i_{t+1} = \hat{f}^i(0, \mathbf{y}^*_t) = \bar{W}^i\boldsymbol{\phi}(0, \mathbf{y}^*_t)$ for inputs $\mathbf{x}^*_t, \mathbf{y}^*_t$ on a grid. In the upper two panels of Figure 7.7, for the ABB stock, we see can see the transition function estimated for training with $T_{window} = 500$ and at the centres of the state space we do not see an unreasonable amount of variation. When comparing the upper panels with the lower ones we do see an increased degree of stability. The same observations can be made from the panels in

Table 7.3: Volatility forecast results for data from two stocks. AvgLL is the average of the predictive log likelihood over each one-step ahead forecast. DM is the DM statistic with $KL^i = LL^i_{RR-GPSV} - LL^i_{*-GARCH}$. Avg-TT is the average training/estimation time of the RR-GPSV for each iteration of the rolling window. Avg-TT of the ABB stock is likely to not be representative of the training time of the procedure having unique access to a single core. DM* indicates that $P < 0.01$.

ABB - small sample

Model	RR-GPSV	GARCH	EGARCH	GJRGARCH	Avg-TT
AvgLL	-2,3311	-2,5411	-2,5369	-2,5099	3,9552
DM		$10,2716^*$	$3,9095^*$	$7,1859^*$	
Settings	$K = 300$, $N = 400$, $N_{mh} = 400$				
$T_{window} = 500$, $T_{sample} = 756$, 2007-01-04 to 2009-12-31					
Spectral Density: Matern/Matern $m = 7^2$					

ABB - larger sample

Model	RR-GPSV	GARCH	EGARCH	GJRGARCH	Avg-TT
AvgLL	-1,6008	-1,9296	-1,9137	-1,8668	$\approx 40^{***}$
DM		$14,3979^*$	$14,2749^*$	$13,5245^*$	
Settings	$K = 300$, $N = 400$, $N_{mh} = 400$				
$T_{window} = 1500$, $T_{sample} = 1761$, 2007-01-04 to 2013-12-31					
Note that the avg-tt is unreliable for this table					
Spectral Density: Matern/Matern $m = 7^2$					

PEPSICO - small sample

Model	RR-GPSV	GARCH	EGARCH	GJRGARCH	Avg-TT
AvgLL	-1,1980	-1,2732	-1,2721	-1,2767	3,2010
DM		$7,0764^*$	$6,2378^*$	$7,6541^*$	
Settings	$K = 300$, $N = 400$, $N_{mh} = 400$				
$T_{window} = 500$, $T_{sample} = 752$, 2014-01-02 to 2016-12-23					
Spectral Density: Matern/Matern $m = 7^2$					

PEPSICO - larger sample

Model	RR-GPSV	GARCH	EGARCH	GJRGARCH	Avg-TT
AvgLL	-1,1639	-1,2750	-1,2750	-1,2752	16,6866
DM		$5,8566^*$	$5,9086^*$	$6,3999^*$	
Settings	$K = 300$, $N = 400$, $N_{mh} = 400$				
$T_{window} = 1500$, $T_{sample} = 1761$, 2010-01-05 to 2016-12-23					
Spectral Density: Matern/Matern $m = 7^2$					

Figure 7.8. We can conclude that, at least for the data sets we considered, a window of $T_{window} = 500$ is certainly not ideal for estimation stability, but is sufficient.

From the results of this section we find that it seems as though the GP-SSM is able to capture the leverage effect and clustering specification and it is different for various data sets. Although, given the number of data-sets considered for example, we cannot conclude that the RR-GPSV is superior to the other models considered, we certainly have verified that the model and learning algorithm are both feasible (for relatively small sample sized) and effective in a forecasting setting.

As a final point, note that all (except two) of the computations have been run on a single core of an i7-3740QM CPU and only the larger sample size estimations have been run on a server. For the server computations each iteration employed a single core of a XEON E5-2690. Given that any iteration has been run on a single core the average of the computation times is not affected by any parallelism.

Figure 7.5: The centre of the distribution over the states from the joint posterior. The blue dotted lines are the percentage growths based on the adjusted closing prices. The spectral density is the Matern/Matern. Upper 2 Panels: NIKKEI 225, lower 2 panels: S&P 500. $K = 300, N = 200, N_{mh} = 200$. $T_{NIKKEI} = 2646$, $T_{S\&P} = 2264$

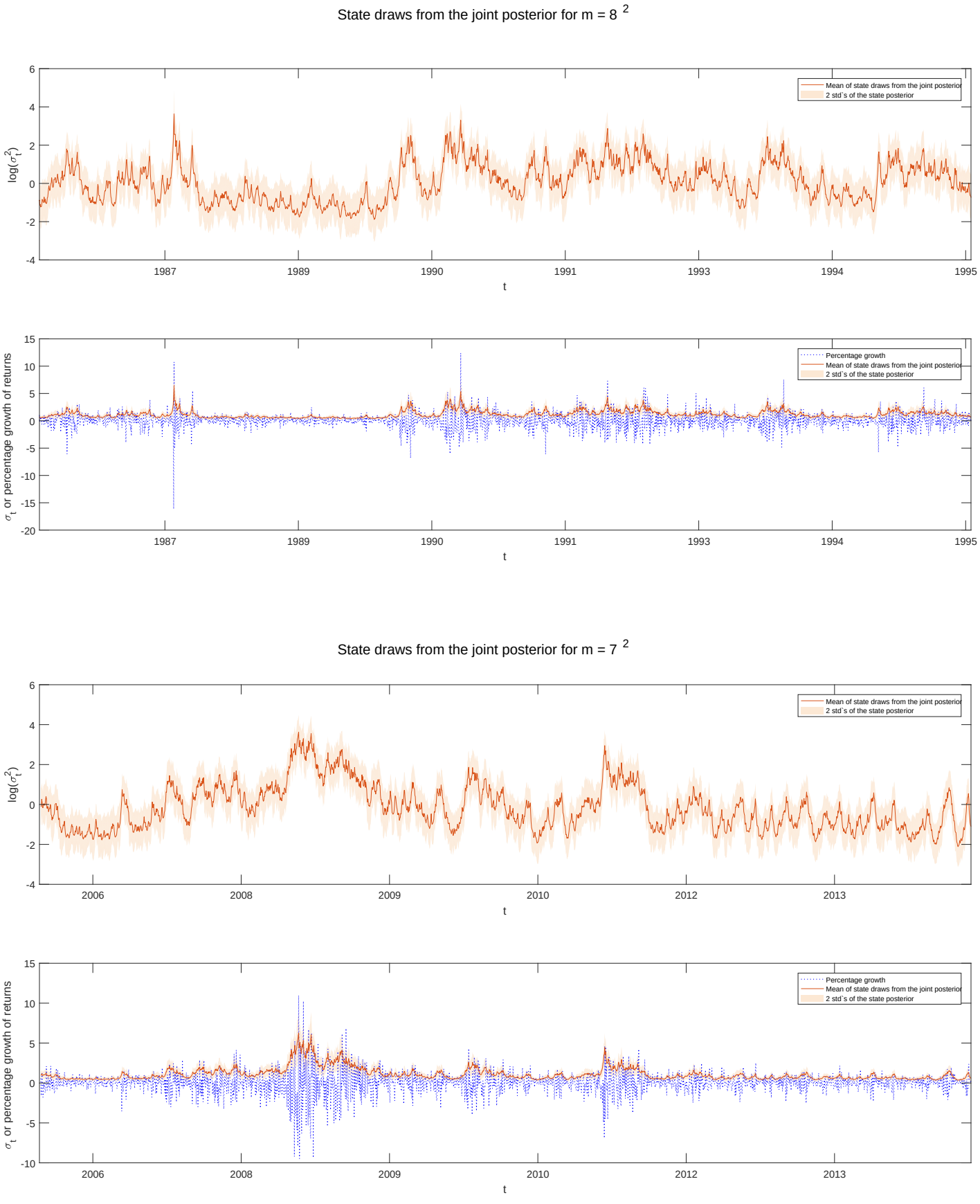


Figure 7.6: The centre of the distribution over the states from the joint posterior. The blue dotted lines are the percentage growths based on the adjusted closing prices. The spectral density is the Matern/Matern. Upper 2 Panels: ABB, lower 2 panels: PEPSICO. $K = 300, N = 200, N_{mh} = 200$. $T_{ABB} = 1259$, $T_{PEPSICO} = 1505$

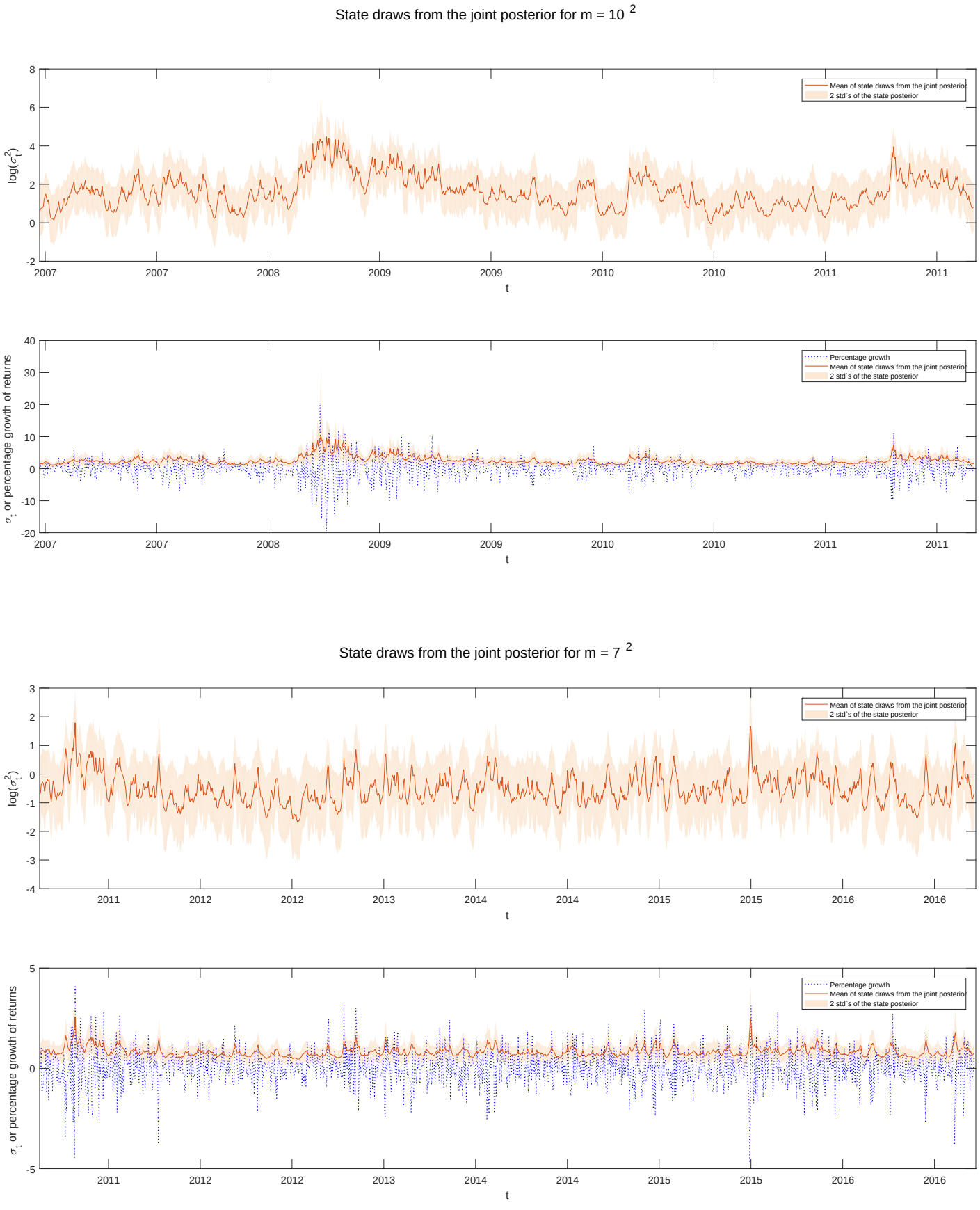


Figure 7.7: Comparison of the state function predictions over time for the ABB stock. In the first upper left panel the means of the distributions over $\hat{\mathbf{x}}^i_{t+1} = \hat{f}^i(\mathbf{x}^*_t, 0)$ is given for $\mathbf{x}^*_t$ on a grid and $\mathbf{x}_t = \log(\sigma^2_t)$. in the second upper panel the same is done for $\hat{\mathbf{x}}^i_{t+1} = \hat{f}^i(0, \mathbf{y}^*_t)$ for $\mathbf{y}^*_t$ on a grid representing percentage growth. For the two upper panels $\hat{f}^i$ is estimated in a rolling window setting for window i with size $T_i = 500$ over a horizon of 2007 to 2009 with $T = 756$. The lower two panels are the same except there $\hat{f}^i$ is estimated with a window of size $T_i = 1500$ over the horizon 2007 to 2013. All with $K = 300, N = 400, N_{mh} = 400, m = 7^2$

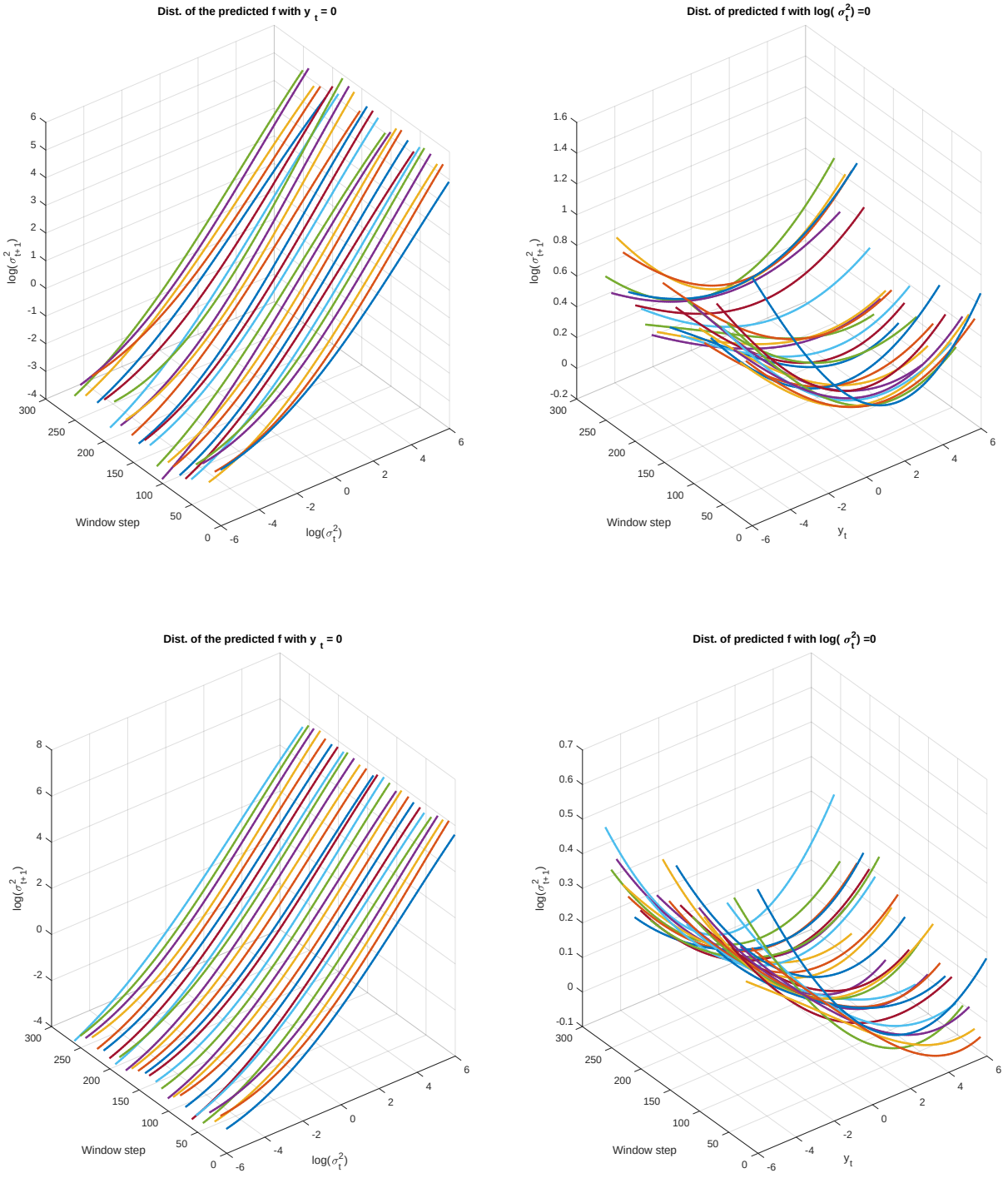
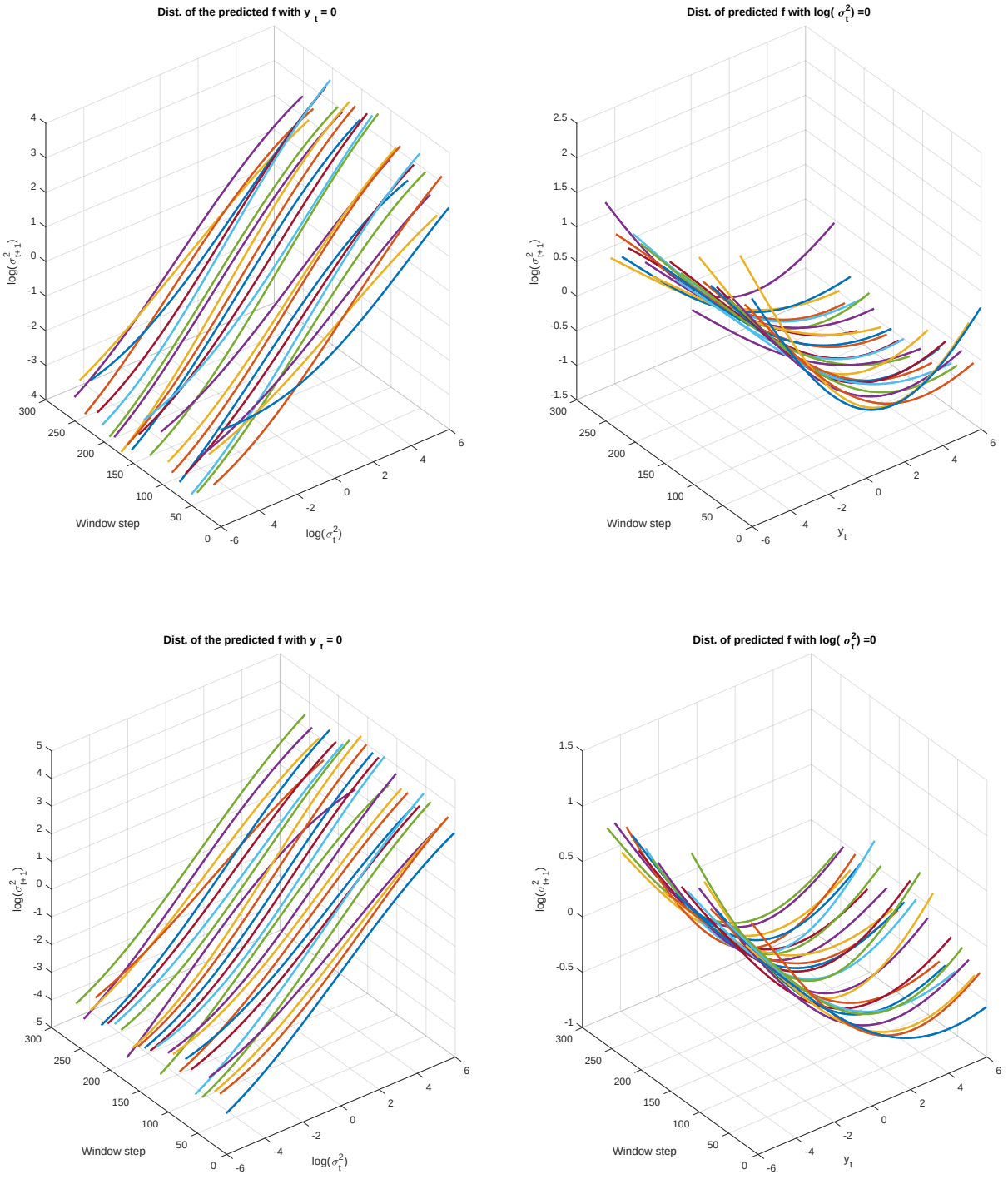


Figure 7.8: Comparison of the state function predictions over time for the PEPSICO stock. In the first upper left panel the means of the distributions over $\hat{\mathbf{x}}^i_{t+1} = \hat{f}^i(\mathbf{x}^*_t, 0)$ is given for $\mathbf{x}^*_t$ on a grid and $\mathbf{x}_t = \log(\sigma^2_t)$. in the second upper panel the same is done for $\hat{\mathbf{x}}^i_{t+1} = \hat{f}^i(0, \mathbf{y}^*_t)$ for $\mathbf{y}^*_t$ on a grid representing percentage growth. For the two upper panels $\hat{f}^i$ is estimated in a rolling window setting for window i with size $T_i = 500$ over a horizon of 2014 to 2016 with $T = 756$. The lower two panels are the same except there $\hat{f}^i$ is estimated with a window of size $T_i = 1500$ over the horizon 2010 to 2016. All with $K = 300, N = 400, N_{mh} = 400, m = 7^2$



8 Conclusion

Contents

8.1 Conclusion

This thesis began with the reconciliation of what is referred to as non-parametrics in econometrics and a subset of what nowadays is known as machine learning. It also delved into the frequentist and the Bayesian view on Gaussian Process based methods. We employed a Gaussian Process based non-parametric function as the state-equation of a Gaussian non-linear state-space model. After reviewing various methods for speeding up estimation in GP-based models we settled on the method introduced by Solin and Särkkä (2014). The choice rested on the objective of having an approximation method that allows for a hierarchical Bayesian treatment in small data circumstances. For this method we designed a convolution covariance structure over product spaces that performed better than an additive one, presumably because it accounts for cross covariances (Majumdar and Gelfand 2007). Because this kernel enters the model through its spectral representation it is easily employable. Within this convolution covariance structure many stationary covariance functions are employable, of which we employed that of the Matern class and the Gaussian Mixture. For the spectral representation of the Matern class as the uni-variate function at each dimension of the convolution covariance structure we found out that the regularity of the prior sample paths can be conveniently made adaptive. The idea of this adaptive regularity was to increase the reliability of the credible regions of the posterior in terms of frequentist coverage.

For estimating the reduced rank Gaussian Process state-space model a hierarchical Bayesian prior was set up and the posterior was approximated using a MCMC Scheme. To verify the finite sample performance of the learning algorithm we simulated from a known distribution over the basis function weights and the hyper-priors. We found that with a

relatively small sample size, small number of particles and a small number of Metropolis-within-Gibbs runs the identification of the true model distribution was sufficiently good. We also concluded that the intrinsic regularization mechanism, from a frequentist point of view, was effective. This in essence allows for the modeler to not worry too much about the truncation and set the order as high as the computational resources allow for. When comparing the performance of the sampler for the two covariance function specifications within the convolution kernel it was concluded that the MCMC algorithm is less effective for the Gaussian Mixture specification. This could partially be due to the large increase in spectral density parameters, and the multimodality of the marginal likelihood of the parameters in the directions of the means of the mixture, within the Metropolis-within-Gibbs step.

After performing a different kind of simulation study, where data was directly generated from a functional form specification, it was found that for relatively low expansion orders of 6 or 8 for each input of the transition function $f : \mathbb{R}^2 \to \mathbb{R}$ the true function was within 1 standard deviation from the mean of the posterior. Furthermore the mean of the posterior over the function looks sufficiently similar to the true function. Once more we found the implementation of the Mixture kernel in the convolution covariance structure, in combination with the sampling algorithm, to be inferior to the Matern one.

For the estimation of time varying conditional volatility, two models were introduced, namely the Reduced Rank GPSV and another version of it with an added layer of latent auto-regression. In an empirical setting we asked the question of whether or not the Reduced Rank GPSV was able to capture stylized facts in financial time series, without specifying the volatility equation in any way, and setting the mean function of the GP-based function to zero. In particular two characteristics were looked at, namely volatility clustering and the leverage effect, and both seemed to be expressed by the model after estimation. Finally we verified the forecasting performance of the models by comparing them to the Gaussian GARCH, EGARCH, and the GJRGARCH and both proposed models significantly outperformed these in terms of predictive log-likelihood. However the Deep Reduced Rank GPSV was very computationally inefficient in combination with the employed sampling algorithm, and its performance was similar to that of the basic Reduced Rank GPSV. We cannot make strong conclusions with regards to the effect of adding non-parametric latent auto-regression to the Reduced Rank GPSV other than that the current MCMC scheme employed does not seem to be sufficient for it.

8.2 Further Research possibilities

Many options for future research are available and we give only a small subset of these. To begin, in this thesis we have implemented the most simple version of the model, for example it would be possible to use additional explanatory variables within the covariance functions. Other obvious extensions include, adding more inputs such as realized volatility

in the transition equation and including long memory in the specification of the kernel. Furthermore multivariate models can also be made using the techniques in this thesis and experimentation with these is also a forthright next move.

It is of vital importance to note that the inference algorithm in combination with the approximation scheme is but one of many. In particular for higher dimensional problems, and when larger amounts of data are available, moving away from "exactly" sampling from the posterior of an approximate model using MCMC methods, towards approximating the posterior of an exact model is more beneficial. For this stochastic variational methods are widely popular as mentioned in the review on approximations. It is possible that within the latter paradigm adding additional non-parametric latent auto-regressive layers can be more effective as in Mattos, Dai, et al. (2015). A comparison between the various architectures can then be made, for example within a volatility estimation application. As mentioned before the batch inference nature of the employed sampling scheme in this thesis is undesirable, and research into on-line methods such as SONIG (Bijl et al. 2016) for example is important.

Whatever method is used in the end, there are certain algorithm settings that usually must be tuned, for example the number of particles and MCMC runs in this thesis. This means that although the interest lies in automatic modeling, one could still put a lot effort into tuning the algorithm settings. Therefore to complete the consolidated approach to automatic modelling the optimization of settings can be achieved using non-parametric Bayesian optimization ((Shahriari et al. 2016)). Research on the effectiveness of such a consolidated approach can then be conducted.

Appendices

A
Appendix

Contents

A.1 Volatility models

Appendix: Definition 1 (Percentage Growth)**.** *Suppose we have a series of observations* $\{P_t\}_{t=1}^T$*, and let* $\{r_t\}_{t=2}^T = \frac{P_t - P_{t-1}}{P_{t-1}}$ *be the proportional change of* P_t*. Then for small* r_t *we have that* $r_t \approx \log(P_t) - \log(P_{t-1})$*. Then we define the percentage growth as* $100 \cdot r_t$ *which we approximate by* $y_t = 100 \cdot (\log(P_t) - \log(P_{t-1}))$*.*

In the empirical applications it is this percentage growth that is used as the output y_t as well as the feedback of the SSM, and for P_t we take the adjusted daily closing prices for the stock.

A.1.1 Observation-driven Models

As a benchmark for the evaluation of the forecasting performance of the proposed models in this thesis a number of observation driven conditional variance models have been employed of which the specifications are given in this section. The Gaussian GARCH(1,1) following the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) framework, introduced by Bollerslev (1986) following the work of Engle (1982).

Appendix: Definition 2 (Gaussian GARCH(1,1))**.**

$$y_t = \sigma_t \epsilon_t \qquad \epsilon_t \sim \mathcal{NID}(0,1) \tag{A.1}$$

$$\sigma_{t+1}^2 = w_0 + w_1 y_t^2 + w_2 \sigma_t^2 \qquad w_0 > 0, w_1, w_2 \geq 0 \tag{A.2}$$

where $w_1 + w_2 < 1$*.*

Here the mean of the returns is not considered to be modelled with an offset in the y_t specification (which can follow an ARMA specification) for simplicity, however for all models the percentage growth series $\{y_t\}$ are de-meaned. Such models are referred to as observation driven because the update equation specifying the evolution is not stochastic (unlike the state equation of an SSM as considered in this paper). A downside to the GARCH(1,1) is the assumtion that negative and positive shocks have a symmetric effect on the condtional volatility, whereas a negative an asymmetric relationship between the returns and the volatility is an well-accepted stylised characteristic of conditional vo;atility models (McNeil et al. 2015). The EGARCH(1,1) (Nelson 1991) below aims to improve the accuracy of the GARCH by incorporating this stylized characteristic into the update equation.

Appendix: Definition 3 (Gaussian EGARCH(1,1))**.**

$$y_t = \sigma_t \epsilon_t \qquad \epsilon_t \sim \mathcal{NID}(0,1) \tag{A.3a}$$

$$\log \sigma_{t+1}^2 = w_0 + w_1 \frac{y_t}{\sigma_t} + w_2 \left(\frac{|y_t|}{\sigma_t} - \mathbb{E}\left[\frac{|y_t|}{\sigma_t} \right] \right) + w_3 \log \sigma_t^2 \qquad \mathbb{E}\left[\frac{|y_t|}{\sigma_t} \right] = \sqrt{\frac{2}{\pi}} \tag{A.3b}$$

Here the logarithm of the conditional variance is modeled and asymmetry in volatility clustering is captured in the w_2 term. We also employ the GJRGARCH (Glosten et al. 1993), which also addresses the asymmetry between returns and conditional volatility.

Appendix: Definition 4 (Gaussian GJRGARCH(1,1))**.**

$$y_t = \sigma_t \epsilon_t \qquad \epsilon_t \sim \mathcal{NID}(0,1) \tag{A.4a}$$

$$\sigma_{t+1}^2 = w_0 + w_1 y_t^2 + w_2 \mathbb{1}_{y_t<0} y_t^2 + w_3 \sigma_t^2 \tag{A.4b}$$

Where $w_0, w_2 > 0$*,* $w_1, w_3 \geq 0$ *and* $w_1 + w_3 + \frac{w_2}{2} < 1$*.*

Again here the leverage effect is captured by w_2.

For all these models the conditional likelihood is specified by the distribution on the percentage growth, for details see (Tsay 2010). The parameters are estimated by optimizing the average of this conditional likelihood using the `fmincon` function with the `sqp` solver. The initial values are taken to be the unconditional variance when available or else the variance estimate of the data.

Works Cited

Adler, RJ (1981). "The Geometry of Random Fields". In:

Agarwal, Deepak K and Alan E Gelfand (2005). "Slice sampling for simulation based fitting of spatial data models". In: *Statistics and Computing* 15.1, pp. 61–69.

Andrews, Donald WK (1991). "Heteroskedasticity and autocorrelation consistent covariance matrix estimation". In: *Econometrica: Journal of the Econometric Society*, pp. 817–858.

Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein (2010). "Particle markov chain monte carlo methods". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.3, pp. 269–342.

Aronszajn, Nachman (1950). "Theory of reproducing kernels". In: *Transactions of the American mathematical society* 68.3, pp. 337–404.

Barber, David, A Taylan Cemgil, and Silvia Chiappa (2011). *Bayesian time series models*. Cambridge University Press.

Bekaert, Geert and Guojun Wu (2000). "Asymmetric volatility and risk in equity markets". In: *Review of Financial Studies* 13.1, pp. 1–42.

Bengio, Yoshua, Yann LeCun, et al. (2007). "Scaling learning algorithms towards AI". In: *Large-scale kernel machines* 34.5, pp. 1–41.

Bijl, Hildo et al. (2016). "Online Sparse Gaussian Process Training with Input Noise". In: *arXiv preprint arXiv:1601.08068*.

Black, Fischer (1976). "Stuedies of stock price volatility changes". In:

Blight, BJN and L Ott (1975). "A Bayesian approach to model inadequacy for polynomial regression". In: *Biometrika* 62.1, pp. 79–88.

Bollerslev, Tim (1986). "Generalized autoregressive conditional heteroskedasticity". In: *Journal of econometrics* 31.3, pp. 307–327.

Bouchaud, Jean-Philippe, Andrew Matacz, and Marc Potters (2001). "Leverage effect in financial markets: The retarded volatility model". In: *Physical review letters* 87.22, p. 228701.

Box, George, Gwilym M Jenkins, and Gregory Reinsel (1994). *Time Series Analysis: Forecasting & Control*. Prentice Hall.

Chen, Xiaohong (2007). "Large sample sieve estimation of semi-nonparametric models". In: *Handbook of econometrics* 6, pp. 5549–5632.

Chopin, Nicolas, Sumeetpal S Singh, et al. (2015). "On particle Gibbs sampling". In: *Bernoulli* 21.3, pp. 1855–1883.

Christie, Andrew A (1982). "The stochastic behavior of common stock variances: Value, leverage and interest rate effects". In: *Journal of financial Economics* 10.4, pp. 407–432.

Deisenroth, Marc Peter and Jun Wei Ng (2015). "Distributed Gaussian Processes." In: *ICML*, pp. 1481–1490.

Diebold, Francis X and Robert S Mariano (2002). "Comparing predictive accuracy". In: *Journal of Business & economic statistics* 20.1, pp. 134–144.

Durbin, J. and S.J. Koopman (2012). *Time Series Analysis by State Space Methods: Second Edition*. Oxford Statistical Science Series. OUP Oxford. URL: `https://books.google.nl/books?id=fOq39Zh0olQC`.

Duvenaud, David (2014). "Automatic model construction with Gaussian processes". PhD thesis. University of Cambridge.

Duvenaud, David K, Hannes Nickisch, and Carl E Rasmussen (2011). "Additive gaussian processes". In: *Advances in neural information processing systems*, pp. 226–234.

Engle, Robert F (1982). "Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation". In: *Econometrica: Journal of the Econometric Society*, pp. 987–1007.

Engle, Robert F and Victor K Ng (1993). "Measuring and testing the impact of news on volatility". In: *The journal of finance* 48.5, pp. 1749–1778.

Figlewski, Stephen and Xiaozu Wang (2000). "Is the'Leverage Effect'a Leverage Effect?" In:

Freedman, David (1999). "On the Bernstein-von Mises theorem with infinite-dimensional parameters". In: *Annals of Statistics*, pp. 1119–1140.

Frigola, Roger, Yutian Chen, and Carl Rasmussen (2014b). "Variational Gaussian process state-space models". In: *Advances in Neural Information Processing Systems*, pp. 3680–3688.

Frigola, Roger, Fredrik Lindsten, Thomas B Schön, and Carl Rasmussen (2013). "Bayesian inference and learning in Gaussian process state-space models with particle MCMC". In: *Advances in Neural Information Processing Systems*, pp. 3156–3164.

Frigola, Roger, Fredrik Lindsten, Thomas B Schön, and Carl E Rasmussen (2014a). "Identification of Gaussian process state-space models with particle stochastic approximation EM". In: *IFAC Proceedings Volumes* 47.3, pp. 4097–4102.

Gelfand, Alan E and Dipak K Dey (1994). "Bayesian model choice: asymptotics and exact calculations". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 501–514.

Gelman, Andrew and Jennifer Hill (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Geweke, John and Gianni Amisano (2010). "Comparing and evaluating Bayesian predictive distributions of asset returns". In: *International Journal of Forecasting* 26.2, pp. 216–230.

Giordani, Paolo, Michael Pitt, and Robert Kohn (2011). "Bayesian inference for time series state space models". In:

Glosten, Lawrence R, Ravi Jagannathan, and David E Runkle (1993). "On the relation between the expected value and the volatility of the nominal excess return on stocks". In: *The journal of finance* 48.5, pp. 1779–1801.

Glynn, Peter W and Donald L Iglehart (1989). "Importance sampling for stochastic simulations". In: *Management Science* 35.11, pp. 1367–1392.

Gordon, Neil J, David J Salmond, and Adrian FM Smith (1993). "Novel approach to nonlinear/non-Gaussian Bayesian state estimation". In: *IEE Proceedings F (Radar and Signal Processing)*. Vol. 140. 2. IET, pp. 107–113.

Hamilton, James Douglas (1994). *Time series analysis*. Vol. 2. Princeton university press Princeton.

Hansen, Peter R and Asger Lunde (2005). "A forecast comparison of volatility models: does anything beat a GARCH (1, 1)?" In: *Journal of applied econometrics* 20.7, pp. 873–889.

Hartikainen, Jouni and Simo Särkkä (2010). "Kalman filtering and smoothing solutions to temporal Gaussian process regression models". In: *Machine Learning for Signal Processing (MLSP), 2010 IEEE International Workshop on*. IEEE, pp. 379–384.

Hjort, Nils Lid et al. (2010). *Bayesian nonparametrics*. Vol. 28. Cambridge University Press.

Johnstone, Iain M (2010). "High dimensional Bernstein-von Mises: simple examples". In: *Institute of Mathematical Statistics collections* 6, p. 87.

Jones, Galin L et al. (2006). "Fixed-width output analysis for Markov chain Monte Carlo". In: *Journal of the American Statistical Association* 101.476, pp. 1537–1547.

Kass, Robert E and Adrian E Raftery (1995). "Bayes factors". In: *Journal of the american statistical association* 90.430, pp. 773–795.

Kim, Sangjoon, Neil Shephard, and Siddhartha Chib (1998). "Stochastic volatility: likelihood inference and comparison with ARCH models". In: *The review of economic studies* 65.3, pp. 361–393.

Knapik, BT, AW van Der Vaart, and JH Van Zanten (2011). "Bayesian inverse problems with Gaussian priors". In: *The Annals of Statistics*, pp. 2626–2657.

Kom Samo, Yves-Laurent and Stephen Roberts (2015). "Generalized Spectral Kernels". In: *arXiv preprint arXiv:1506.02236*.

Kostantinos, N (2000). "Gaussian mixtures and their applications to signal processing". In: *Advanced signal processing handbook: theory and implementation for radar, sonar, and medical imaging real time systems*, pp. 3–1.

Lindsten, Fredrik, Michael I Jordan, and Thomas B Schön (2014). "Particle gibbs with ancestor sampling." In: *Journal of Machine Learning Research* 15.1, pp. 2145–2184.

Liu, Jane and Mike West (2001). "Combined parameter and state estimation in simulation-based filtering". In: *Sequential Monte Carlo methods in practice*. Springer, pp. 197–223.

MacKay, David JC (1998). "Introduction to Gaussian processes". In: *NATO ASI Series F Computer and Systems Sciences* 168, pp. 133–166.

Majumdar, Anandamayee and Alan E Gelfand (2007). "Multivariate spatial modeling for geostatistical data using convolved covariance functions". In: *Mathematical Geology* 39.2, pp. 225–245.

Mattos, César Lincoln C, Zhenwen Dai, et al. (2015). "Recurrent gaussian processes". In: *arXiv preprint arXiv:1511.06644*.

Mattos, César Lincoln C, Andreas Damianou, et al. (2016). "Latent Autoregressive Gaussian Processes Models for Robust System Identification". In: *IFAC-PapersOnLine* 49.7, pp. 1121–1126.

McNeil, Alexander J, Rüdiger Frey, and Paul Embrechts (2015). *Quantitative risk management: Concepts, techniques and tools*. Princeton university press.

Modigliani, Franco and Merton H Miller (1958). "The cost of capital, corporation finance and the theory of investment". In: *The American economic review*, pp. 261–297.

Neal, Peter, Gareth Roberts, et al. (2006). "Optimal scaling for partially updating MCMC algorithms". In: *The Annals of Applied Probability* 16.2, pp. 475–515.

Neal, Radford M (2003). "Slice sampling". In: *Annals of statistics*, pp. 705–741.

Nelson, Daniel B (1991). "Conditional heteroskedasticity in asset returns: A new approach". In: *Econometrica: Journal of the Econometric Society*, pp. 347–370.

Poon, Ser-Huang and Clive WJ Granger (2003). "Forecasting volatility in financial markets: A review". In: *Journal of economic literature* 41.2, pp. 478–539.

Quiñonero-Candela, Joaquin and Carl Edward Rasmussen (2005). "A unifying view of sparse approximate Gaussian process regression". In: *Journal of Machine Learning Research* 6.Dec, pp. 1939–1959.

Rakitsch, Barbara et al. (2013). "It is all in the noise: Efficient multi-task Gaussian process inference with structured residuals". In: *Advances in Neural Information Processing Systems*, pp. 1466–1474.

Rasmussen, Carl Edward (2006). "Gaussian processes for machine learning". In:

Robinson, Peter M (2001). "The memory of stochastic volatility models". In: *Journal of econometrics* 101.2, pp. 195–218.

Rosenbluth, Marshall N and Arianna W Rosenbluth (1955). "Monte Carlo calculation of the average extension of molecular chains". In: *The Journal of Chemical Physics* 23.2, pp. 356–359.

Rubinstein, Reuven Y and Dirk P Kroese (2011). *Simulation and the Monte Carlo method*. Vol. 707. John Wiley & Sons.

Schön, Thomas B et al. (2015). "Sequential Monte Carlo methods for system identification". In: *IFAC-PapersOnLine* 48.28, pp. 775–786.

Shahriari, Bobak et al. (2016). "Taking the human out of the loop: A review of bayesian optimization". In: *Proceedings of the IEEE* 104.1, pp. 148–175.

Sherlock, Chris, Alexandre Thiery, and Anthony Lee (2016). "Pseudo-marginal Metropolis–Hastings using averages of unbiased estimators". In: *arXiv preprint arXiv:1610.09788*.

Showalter, Ralph E (2010). *Hilbert space methods in partial differential equations*. Courier Corporation.

Snelson, Edward and Zoubin Ghahramani (2005). "Sparse Gaussian processes using pseudo-inputs". In: *Advances in neural information processing systems*, pp. 1257–1264.

Solin, Arno and Simo Särkkä (2014). "Hilbert space methods for reduced-rank Gaussian process regression". In: *arXiv preprint arXiv:1401.5508*.

Stein, Michael L (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

Svensson, Andreas and Thomas B Schön (2016). "A flexible state space model for learning nonlinear dynamical systems". In: *arXiv preprint arXiv:1603.05486*.

Svensson, Andreas, Arno Solin, et al. (2015). "Computationally efficient Bayesian learning of Gaussian process state space models". In: *arXiv preprint arXiv:1506.02267*.

Szabó, Botond, AW van der Vaart, JH van Zanten, et al. (2015). "Frequentist coverage of adaptive nonparametric Bayesian credible sets". In: *The Annals of Statistics* 43.4, pp. 1391–1428.

Tanner, Martin A and Wing Hung Wong (1987). "The calculation of posterior distributions by data augmentation". In: *Journal of the American statistical Association* 82.398, pp. 528–540.

Titsias, Michalis K (2009). "Variational Learning of Inducing Variables in Sparse Gaussian Processes." In: *AISTATS*. Vol. 5, pp. 567–574.

Tsay, Ruey S (2010). *Analysis of Financial Time Series*. John Wiley & Sons.

Turner, Ryan Darby (2012). "Gaussian Processes for state space models and change point detection". PhD thesis. University of Cambridge.

Van Den Berg, C, JPR Christensen, and P Ressel (2012). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Vol. 100. Springer Science & Business Media.

Van der Vaart, Aad W and J Harry van Zanten (2008). "Rates of contraction of posterior distributions based on Gaussian process priors". In: *The Annals of Statistics*, pp. 1435–1463.

Vats, Dootika, James M Flegal, and Galin L Jones (2015). "Multivariate output analysis for Markov chain Monte Carlo". In: *arXiv preprint arXiv:1512.07713*.

Wahba, Grace (1990). *Spline models for observational data*. SIAM.

Walter, Eric and Luc Pronzato (1997). *Identification of parametric models from experimental data*. Springer Verlag.

Wang, Jack, Aaron Hertzmann, and David M Blei (2005). "Gaussian process dynamical models". In: *Advances in neural information processing systems*, pp. 1441–1448.

West, M and J Harrison (1997). "Bayesian Forecasting and Dynamic Models. Springer-Verlag". In: *New York*.

Whaley, Robert E (2000). "The investor fear gauge". In: *The Journal of Portfolio Management* 26.3, pp. 12–17.

Wills, Adrian et al. (2012). "Estimation of linear systems using a Gibbs sampler". In: *IFAC Proceedings Volumes* 45.16, pp. 203–208.

Wilson, Andrew Gordon (2014). "Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes". In: *University of Cambridge*.

Wilson, Andrew Gordon and Ryan Prescott Adams (2013). "Gaussian Process Kernels for Pattern Discovery and Extrapolation." In: *ICML (3)*, pp. 1067–1075.

Wilson, Andrew et al. (2014). "Fast kernel learning for multidimensional pattern extrapolation". In: *Advances in Neural Information Processing Systems*, pp. 3626–3634.

Wu, Yue, José Miguel Hernández-Lobato, and Zoubin Ghahramani (2014). "Gaussian process volatility model". In: *Advances in Neural Information Processing Systems*, pp. 1044–1052.