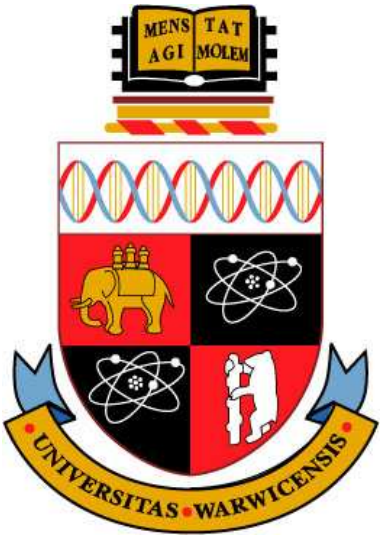

Parameter Estimation and Model Fitting

of

Stochastic Processes

by

Fan Zhang

Thesis

Submitted to the University of Warwick

in partial fulfillment of the requirements

for admission to the degree of

Doctor of Philosophy

Centre for Scientific Computing & Mathematics Institute

June 2011

THE UNIVERSITY OF WARWICK

Contents

Acknowledgments

Foremost, I would like to express my deepest gratitude to my supervisors, Professor Andrew Stuart and Doctor Anastasia Papavasiliou, for their continuous support through my PhD study, research and professional development, for the motivation, guidance, patience and knowledge they have provided through my PhD study. This thesis can not have been made possible without their guidance and tutorship. So, here again, I express my most sincere gratitude to their assistance.

I would also like to thank the members of the Applied Maths and Statistics seminar group and research fellows for their fruitful thoughts, critical but constructive comments and advices, assistance and help, especially Konstantinos Zygalakis, Simon Cotter, Yvo Pokern, Jochen Voss and David White.

I am also thankful to the financial support provided by the University Warwick Vice-Chancellor's studentship, and the internship opportunity provided by International Monetary Fund to practice my learnings.

Last but not least, I would like to thank my parents for their continuing support through my life, and for the opportunity to study at Warwick.

Declarations

The work contained in this thesis is original, except as acknowledged, and has not been submitted previously for a degree at any university. To the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made.

Abstract

Multiscale methods such as averaging and homogenization have become an increasingly interesting topic in stochastic time series modelling. When applying the averaged/homogenized processes to applications such as parameter estimation and filtering problems, the resulting asymptotic properties are often weak. In this thesis, we focus on the above mentioned multiscale methods applied on Ornstein-Uhlenbeck processes. We find that the maximum likelihood based estimators for the drift and diffusion parameters derived from the averaged/homogenized systems can use the corresponding marginal multiscale data as observations, and still provide a strong convergence to the true value as if the observations are from the averaged/homogenized systems themselves. The asymptotic distribution for the estimators are studied in this thesis for the averaging problem, while that of the homogenization problem exhibit more difficulties and will be an interest of future work. In the case when applying the multiscale methods to the Kalman filter of Ornstein-Uhlenbeck systems, we study the convergence between the marginal covariance and marginal mean of the full scale system and those of the averaged/homogenized systems, by measuring their discrepancies.

In Part III, we study real world projects of time series modelling in the field of econometrics. Chapter 7 presents a modelling project on interest rate time series from the well known Nelson-Siegel yield curve model. The methodology shows a development from standard Vector Autoregressive model to Bayesian based heteroscedastic regression model. Gibbs sampling is used as the Monte Carlo method. Chapter 8 presents a model comparison in modelling a portfolio of economic indices between constant correlation GARCH and Dynamic Conditional Correlation GARCH models. It compares the two models suitability in capturing the effect of “volatility clustering”.

List of Figures

Chapter 1

Introduction

The problem of parameter estimation for autoregressive (AR) type time series has long been a very popular topic. A large amount of literature focuses on parameter estimation problems for AR type models under different setups. In this thesis, we present parameter estimation strategies of AR type models within the framework of Ornstein Uhlenbeck stochastic differential equations in Part I and II, where the data are provided continuously, except when discretization is necessary. We present the problem within the Bayesian framework in Part III, where the data are provided discretely since it is based on real world applications.

Parameter estimation forms an essential part of the statistical inference methodologies, especially for standard models such as Ornstein Uhlenbeck processes. In recent literatures, such as [1, 2, 69, 72], model fitting of multiscale data has become a popular topic in this area, since the finite dimensional data with different scales often become inconsistent with model at small scales when applying standard statistical inference methods. The discrepency between the estimated and true values of the parameters of the model could deviate significantly. Furthermore, the methods presented in [72] gives a general set of models, but a weak convergence for the estimators. For applications in real practice, this motivates us to study the asymptotic behaviour of the estimators in a strong sense for the model at small scales. For ease of approach, we focus on the Ornstein Uhlenbeck processes as a point of attack, since the problem of parameter estimation for Ornstein-Uhlenbeck (OU) processes has been extensively studied in the literature. Discussions of maximum likelihood estimators for the drift and diffusion parameters of an OU process and their asymptotic properties can be found in [11, 54, 66].

The problem of parameter estimation for stochastic differential equation within the multiscale framework had already been studied for some certain types of multiscale setup. In [6, 7], the authors discuss maximum likelihood estimation of drift and diffusion parameters of a scalar OU process x_t, when data is observed from a scaled smoothed OU process

$y_t = \frac{1}{\epsilon}\int_{t-\epsilon}^{t} x_s ds$. They conclude that the observation time step Δ and total number of observations N should both be functions of ϵ, in order to preserve asymptotic consistency and efficiency. In addition, [5] also constructs an adaptive subsampling scheme to be applied in a Triad model. Another paper discussing parameter estimation for OU processes in the multiscale framework is [1]. It studies the problem of estimating integrated diffusion under the existence of microstructure noise. It assumes the hidden process follows an Itô diffusion process, and tries to estimate the integrated diffusion parameter, while observing data with microstructure additive noise. It proposes a subsampling and aggregating scheme to ensure the consistency and statistical efficiency of the estimator.

In Part I, we focus on a different set up of the multiscale framework, which is discussed in detail in [78]. Within this framework, weak convergence of drift estimator for a general type of Itô SDE is discussed in [73]. Weak drift and diffusion estimatiors for a Langevin equation is discussed in [79]. In this thesis, we observe data from the slow variable of a two (averaging) or three (homogenization) time-scale system of OU stochastic differential equations, and estimate the drift and diffusion parameters for the coarse-grained equation for the slow variable. We will show that the maximum likelihood based drift and diffusion estimators are asymptotically consistent, in a strong sense.

After we have investigated the behaviour of maximum likelihood estimators for the data with small scales in a finite dimensional multiscale framework, it comes natural to us that we want to further utilize the feature of averaging and homogenization in a wider area of applications. One of the most popular area in stochastic modelling is filtering. Studying the behaviour of multiscale filtering can be useful in many areas, such as analysis of signal processing, dynamical systems and meterology and oceanic modelling. Multiscale filtering can aid accurate estimate of the small scale component, which reveals the microscopic stochastic nature of the data, while also significantly reduces the demand for computational resources, which make simultaneous estimates more cheaply achievable or even from impossible to possible. These reasons directly motivate us to investigate this methodology in the context of averaging and homogenization.

Though multiscale filtering is a recently developed topic, it has already been studied in many literatures. In [62], the authors studied mathematical strategies for filtering turbulent dynamical systems. The approach involves the synergy of rigorous mathematical guidelines, exactly solvable nonlinear models with physical insight, and novel cheap algorithms with judicious model errors to filter turbulent signals with many degrees of freedom. [76] studied nonlinear filtering problems for the optimal filter of the slow component of a two-scale system, where the fast scale component is ergodic.

Applications of multiscale linear filter has been extensively studied in the context of signal processing, such as [75, 91]. [41] presented a complete study of the limiting

behaviour of the homogenization problem following SPDEs of the form

$$dp^\epsilon(t,x) = L^\epsilon(t)p^\epsilon(t,x)dt + M^\epsilon(t)p^\epsilon(t,x)dW_t$$

where

$$L^\epsilon = \bigtriangledown_{x_i}(a^{ij}(x/\epsilon, Z^\epsilon/\epsilon) \bigtriangledown_{x_j} \cdot)$$

and

$$M_k^\epsilon(t) = h_k(x_t/\epsilon, Z_t^\epsilon/\epsilon)$$

for which Z_t^ϵ follows $dZ_t^\epsilon = f(Z^\epsilon/\epsilon)dt + QdW_t$. [65] discussed filtering problems where the underlying SDE follows

$$dx = g(x/\epsilon)xdt + \sigma(x/\epsilon)dw$$

with observations taken from $dz = h(x/\epsilon)xdt$, for which g, σ and h lie on a unit torus on $\mathbb{R}^n$.

In Part II, we focus on the problem of linear filtering for the multiscale system studied in Part I. The observations we consider contain two sources of discrepancies, the discrepancy between the slow part of the multiscale system and the averaged/homogenized process, and the discrepancy between the actual model and the observed data. We apply Kalman filter to the contaminated observation from the slow part of the multiscale system, to show that the marginal Kalman filter for the slow part of the system converges to the filtered distribution of the averaged/homogenized process.

Part III studies the problem of AR type derived time series models fitted in real world applications. Autoregressive model is one of the standard tool in time series data analysis. Autoregressive models are the most established models in time series forecasting. In this part, we integrate the autoregressive models with Bayesian methods and implemented through Markov Chain Monte Carlo to build specific time series hierarchical models for time series forecasting.

In Chapter 7, we will show a step-by-step construction of the AR model from a standard Vector Autoregressive model to a Bayesian Heteroscedastic Regression model, implemented using Monte Carlo methods. The underlying data is the nominal interest rates from the Nelson Siegel yield curve model. The reason behind the choice of Bayesian method is that we believe the nature of the time series has been distorted significantly by the recent crisis, which made standard regression not plausible. We believe that Bayesian updating scheme is a good feature that can be added to the model to make plausible predictions.

In Chapter 8, we fit the GARCH type of the AR model to diagnose the volatility structure of a group of multivariate time series data, under the presence of volatility clus-

tering. The underlying data is a portfolio of 8 indices representing a wide area of economic aspects. Since the aim of this project is to reconstruct the cross-correlation between components of the time series under the presence of volatility clustering, we opt for the GARCH model, and compare the performance of the Constant Correlation (CC) and Dynamic Conditional Correlation (DCC) modifications of the model. The reason we choose GARCH model is that it is the best tool in volatility modelling. We also expect that the DCC version of the GARCH model would represent the evolution of the correlations across the indices.

In this thesis, we use C or c to denote an arbitrary constant which can vary from occurrence to occurrence. For the simplicity of notation, we will write x_n (or y_n, X_n) instead of $x(n\delta)$ (resp. $y(n\delta)$, $X(n\delta)$) for the discretized process. To simplify on notation, we may sometimes omit the tensor product sign $\otimes$, where we actually mean matrix/vector product as tensor products. Similarly, will often use $(\cdot)^2$ to denote a square under tensor product, and for any matrix $\sqrt{m}$, we define m as $m = \sqrt{m}\sqrt{m}^*$. When data is observed discrete, we use δ to denote the time increment observations are provided, and Δ to denote the time increment observations are taken.

Part I

PARAMETER ESTIMATION FOR MULTISCALE ORNSTEIN-UHLENBECK PROCESSES

Chapter 2

The Ornstein-Uhlenbeck (OU) Process

A vector valued Ornstein Uhlenbeck (OU) process is defined as the solution of a stochastic differential equation of the form

$$\frac{dX}{dt} = a(X - \mu) + \sqrt{\sigma}\frac{dW}{dt} \, . \tag{2.1}$$

When the drift matrix a is negative definite, and the diffusion matrix σ is positive definite diagonal, the process is ergodic. The solution X can be written in closed form,

$$X(t) = (I - e^{at})\mu + e^{at}X(0) + \int_0^t e^{a(t-s)}\sqrt{\sigma}dW_s \, . \tag{2.2}$$

When this process is ergodic, we take the limit in time,

$$\lim_{t\to\infty} X(t) = \mu + \lim_{t\to\infty}\int_0^t e^{a(t-s)}\sqrt{\sigma}dW_s \, . \tag{2.3}$$

This is clearly a Gaussian random variable, with mean μ, and variance σ_∞, for which,

$$\mathrm{vec}(\sigma_\infty) = (-a \oplus -a)^{-1}\mathrm{vec}(\sigma), \tag{2.4}$$

where $\oplus$ denotes the Kronecker sum, and $\mathrm{vec}(\cdot)$ denotes the vectorization of the matrix by stacking its columns into a single column vector. Using this invariant property, the drift and diffusion parameters can be easily estimated. The following known results can be found in [11, 54, 66].

Theorem 2.1. *Assume we are given continuous observations from an Ornstein Uhlenbeck process X defined in* (2.1)*. Then the maximum likelihood estimator for the drift parameter*

$\hat{a}_T$, defined as

$$\hat{a}_T = \left(\int_0^T dX \otimes X \right) \left(\int_0^T X \otimes X dt \right)^{-1} \tag{2.5}$$

is asymptotically unbiased and converges almost surely to a as $T \to \infty$. It is also asymptotically normal, as

$$\sqrt{T} \left(\hat{a}_T - a \right) \overset{D}{\to} \mathcal{N} \left(0, \sigma_\infty \right) \ as \ T \to \infty \, .$$

Theorem 2.2. *Assume we are given a discretized realization $X_n = X(n\delta)$ from an Ornstein-Uhlenbeck process defined in* (2.1) *with time step δ. The maximum likelihood estimator $\hat{\sigma}_\delta$, defined as*

$$\hat{\sigma}_\delta = \frac{1}{T} \sum_{n=0}^{N-1} \left(X_{n+1} - X_n \right) \otimes \left(X_{n+1} - X_n \right) \, . \tag{2.6}$$

is asymptotically unbiased and converges almost surely to σ as $\delta \to 0$, while T is fixed, and $X_n = X(n\delta)$. In addition, it is asymptotically normal as

$$\frac{1}{\sqrt{\delta}} (\hat{\sigma}_\delta - \sigma) \overset{D}{\to} \mathcal{N} \left(0, \frac{2\sigma^2}{T} \right) \ as \ \delta \to 0 \, .$$

Some key steps in the proof of these theorems are the following:

$$\int_0^T X(t) \otimes dW_t \overset{D}{\to} \mathcal{N} \left(0, T\sigma_\infty \right) \ \text{ as } \ T \to \infty \, ,$$

and

$$\int_0^T X(t) \otimes X(t) dt \to T\sigma_\infty \ \text{ a.s., } \ \text{ as } \ T \to \infty \, .$$

In Part I, we use the estimators defined in (2.5) and (2.6) to fit the data coming from equations (3.1a) in Chapter 3 and (4.1a) in Chapter 4. Our main goal is to study their asymptotic properties. In chapter 3, we discuss parameter estimation problem in the averaging setup, where the data comes from equation (3.1a); while in chapter 4, we study the parameter estimation problem in the homogenization setup corresponding to equation (4.1a).

Another result which will be useful to us is an extended version of the maximal inequality result from Theorem 2.5 in [34]. The theorem states that the expected supremum of a stopped scalar OU process is bounded in terms of the drift and its stopping time. We convert this result to suit a vector valued OU process.

Theorem 2.3. *Let $(X(t))_{t \geq 0}$ be the Ornstein-Uhlenbeck process solving* (2.1) *with $X(0) = x_0$, where W is a standard Brownian Motion. Then there exists universal constant $C > 0$,*

such that

$$\mathbb{E}\left(\sup_{0\le t\le T}\|X(t)\|^2\right) \le C\frac{\log\left(1+\max_i(|D_{ii}|)T\right)}{\min_i(|D_{ii}/\Sigma_{ii}|)}$$

where $a = PDP^{-1}$ is the eigenvalue diagonalization of the drift matrix a, which only has real eigenvalues; D is the diagonal matrix of eigenvalues of a with the scales of eigenvalues sorted in increasing magnitude from top-left to lower-right entries. Note that P is the normalized matrix of corresponding eigenvectors. Finally, $\Sigma = P^{-1}\sigma(P^{-1})^*$. $\|\cdot\|$ is the Euclidean norm.

A maximal bound for complex valued OU processes is discussed in Theorem A.1 in [74].

Proof. We prove this theorem by doing a linear transformation for $X(t)$. By assumption, we can write $a = PDP^{-1}$. Let $X'(t)$ be

$$X'(t) = P^{-1}X(t),$$

Then, we can rewrite equation (2.1) as

$$dX'(t) = DX'(t)dt + P^{-1}\sqrt{\sigma}dW_t.$$

Since $P^{-1}\sqrt{\sigma}W_t$ is a linear combination of the vector valued Brownian motion W_t, we can define a new Brownian motion, by defining a positive definite symmetric matrix $\Sigma = P^{-1}\sigma(P^{-1})^*$,

$$\sqrt{\Sigma}dW'_t = P^{-1}\sqrt{\sigma}dW_t\ .$$

Furthermore, by the time change property of Brownian motions, we can rescale W'_t as

$$\sqrt{\Sigma_{ii}}dW'_{t/\Sigma_{ii}} = d\tilde{W}_{i,t}$$

where Σ_{ii} is the ith entry on the diagonal of Σ. We can rewrite the OU process in scalar form, for equation i,

$$dX'(t/\Sigma_{ii})_i = (D_{ii}/\Sigma_{ii})X'(t/\Sigma_{ii})_i dt + d\tilde{W}_{i,t}.$$

Notice that the $\tilde{W}_{i,t}$ and $\tilde{W}_{j,t}$ for $i \neq j$ are correlated, however, this does not undermine the assumption of Theorem 2.5 in [34], since $\tilde{W}_{i,t}$ is a standard Brownian motion. We apply

Theorem 2.5 in [34] to each transformed equation above, we have

$$\begin{aligned}
& \mathbb{E}\left(\sup_{0\le t\le T}|(X'(t))_i|\right) \\
= \;& \mathbb{E}\left(\sup_{0\le t\le \Sigma_{ii}T}|(X'(t/\Sigma_{ii}))_i|\right) \\
\le \;& C\sqrt{\frac{\log\left(1+|\frac{D_{ii}}{\Sigma_{ii}}\Sigma_{ii}|T\right)}{|D_{ii}/\Sigma_{ii}|}} \\
\le \;& C\sqrt{\frac{\log\left(1+|D_{ii}|T\right)}{|D_{ii}/\Sigma_{ii}|}}\,.
\end{aligned}$$

From the transformation, we know

$$\begin{aligned}
& \mathbb{E}\left(\sup_{0\le t\le T}|X(t)_i|\right) \\
\le \;& \|P_i\|\mathbb{E}\left(\sup_{0\le t\le T}|X'(t)_i|\right) \\
\le \;& \|P_i\|\mathbb{E}\left(\sup_{0\le t\le T\Sigma_{ii}}|X'(t/\Sigma_{ii})|\right)\,.
\end{aligned}$$

where P_i is the ith row of P, which are normalized eigenvectors. Consequently we have the result,

$$\mathbb{E}\left(\sup_{0\le t\le T}\|X(t)\|\right) \le C\max_i(\|P_i\|)\sqrt{\frac{\log\left(1+\max_i(|D_{ii}|)T\right)}{\min_i\left(|D_{ii}/\Sigma_{ii}|\right)}}\,,$$

hence,

$$\mathbb{E}\left(\sup_{0\le t\le T}\|X(t)\|^2\right) \le C\max_i(\|P_i\|^2)\frac{\log\left(1+\max_i(|D_{ii}|)T\right)}{\min_i(|D_{ii}/\Sigma_{ii}|)}\,.$$

Since P_i are normalized eigenvectors of a, $\|P_i\|^2=1$, thus

$$\mathbb{E}\left(\sup_{0\le t\le T}\|X(t)\|^2\right) \le C\frac{\log\left(1+\max_i(|D_{ii}|)T\right)}{\min_i(|D_{ii}/\Sigma_{ii}|)}\,.$$

□

Chapter 3

Parameter Estimation for the Averaged Equation of a Multiscale OU Process

3.1 Introduction

In this chapter, we consider the following fast/slow system of stochastic differential equations

$$\frac{dx}{dt} = a_{11}x + a_{12}y + \sqrt{q_1}\frac{dU}{dt} \tag{3.1a}$$

$$\frac{dy}{dt} = \frac{1}{\epsilon}\left(a_{21}x + a_{22}y\right) + \sqrt{\frac{q_2}{\epsilon}}\frac{dV}{dt} \tag{3.1b}$$

for which $x \in \mathcal{X}$, $y \in \mathcal{Y}$. We may take $\mathcal{X}$ as $\mathbb{R}^{d_1}$ and Y as $\mathbb{R}^{d_2}$. The case where $\mathcal{X}$ is $\mathbb{T}^{d_1}$, and $\mathcal{Y}$ is $\mathbb{T}^{d_2}$ is discussed in [78]. We assume we observe data generated by the projection onto the x coordinate of the system. We also make the following assumptions.

Assumptions 3.1.
We assume that

(i) *U, V are independent standard Brownian motions;*

(ii) *q_1, q_2 are positive definite diagonal matrices;*

(iii) *$0 < \epsilon \ll 1$;*

(iv) *the system's drift matrix*

$$\begin{pmatrix} a_{11} & a_{12} \\ \frac{1}{\epsilon}a_{21} & \frac{1}{\epsilon}a_{22} \end{pmatrix}$$

only have negative real eigenvalues when ϵ is sufficiently small;

(v) *$x(0)$ and $y(0)$ are independent of U and V, $(x(0), y(0))$ is under the invariant measure of system* (3.1)*, and $\mathbb{E}\left(\|x(0)\|^2 + \|y(0)\|^2\right) < \infty$.*

Remark 3.2. *Assumption 3.1(iv) guarantees the ergodicity of the system* (3.1) *when ϵ is small.*

Remark 3.3. *Though we assumed the whole system* (3.1) *to be ergodic through assumption 3.1(iv). In other words, we assumed a_{22} and $a_{11} - a_{12}a_{22}^{-1}a_{21}$ to be negative definite. We believe it may be relaxed to complex eigenvalues with negative real parts. Drift matrix with complex eigenvalues may be of interest to further work.*

Remark 3.4. *Assumption 3.1(ii) assumes diagonal matrices for the diffusion parameters q_1 and q_2, which ensures independence of Brownian motions. However, we believe that q_1 and q_2 being positive definite symmetric should be sufficient to guarantee the same results in this chapter, since Brownian motions can be rescaled in time and linearly combined to obtain an equivalent Brownian motion in distribution with diagonal diffusion matrix. We make this assumption for simplicity of notation.*

In what follows, we will refer to the following equation as the **averaged equation** for equation (3.1a),

$$\frac{dX}{dt} = \tilde{a}X + \sqrt{q_1}\frac{dU}{dt}\,, \tag{3.2}$$

where

$$\tilde{a} = a_{11} - a_{12}a_{22}^{-1}a_{21}\,. \tag{3.3}$$

In the rest of this chapter,

- we take observations from the multiscale system (3.1a);
- we first show that the discrepancy between the trajectories from the slow part x of the multiscale system (3.1a) and the averaged equation (3.2) is of order $\mathcal{O}(\sqrt{\epsilon})$ in the L^2 sense, in Section 3.2;
- we then show that using observations from the multiscale system (3.1a) and applying them to the drift estimator $\hat{a}_T$, defined in (2.5), we can correctly estimate the drift $\tilde{a}$ of the averaged equation (3.2) in Section 3.3, and study the asymptotic normality of the estimator in Section 3.4;
- we also show that using observations from the multiscale system (3.1a) and applying them to the diffusion estimator $\hat{\sigma}_\delta$, defined in (2.6), we can correctly estimate the diffusion parameter q_1 of the averaged equation (3.2) in Section 3.5, and study the asymptotic normality of the estimator in Section 3.6;

- finally a numerical example is studied to illustrate our findings in Section 3.7.

3.2 The Paths

In this section, we show that the projection of system (3.1) onto the x coordinate converges in a strong sense to the solution X of the averaged equation (3.2). Our result extends that of Theorem 17.1 in [78], where the state space $\mathcal{X}$ is restricted to $\mathbb{T}$ and the averaged equation is deterministic. Assuming that the system is an OU process, the domain can be extended to $\mathbb{R}$ and the averaged equation can be stochastic. We prove the following lemma first.

Lemma 3.5. *Suppose that (x, y) solves (3.1a) and Assumptions 3.1 are satisfied. Then, for finite $T > 0$, and ϵ small*

$$\mathbb{E} \sup_{0\leq t\leq T} \left(\|x(t)\|^2 + \|y(t)\|^2\right) = \mathcal{O}\left(\log(1 + \frac{T}{\epsilon})\right) \tag{3.4}$$

where $\|\cdot\|$ is the vector norm, and the order is in terms of ϵ.

Proof. We look at the system of SDEs as,

$$d\mathbf{x}_t = \mathbf{a}\mathbf{x}_t dt + \sqrt{\mathbf{q}} dW_t \tag{3.5}$$

where

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \mathbf{a} = \begin{pmatrix} a_{11} & a_{12} \\ \frac{1}{\epsilon}a_{21} & \frac{1}{\epsilon}a_{22} \end{pmatrix} \text{ and } \mathbf{q} = \begin{pmatrix} q_1 & 0 \\ 0 & \frac{q_2}{\epsilon} \end{pmatrix}.$$

We try to characterize the magnitude of the eigenvalues of $\mathbf{a}$. To find the eigenvalues, we require

$$\det(\mathbf{a} - \lambda I) = 0 \, .$$

For block matrices, the equation above can be rearranged to,

$$\det\left(\frac{1}{\epsilon}a_{22} - \lambda I\right) \det\left((a_{11} - \lambda I) - a_{12}(\frac{1}{\epsilon}a_{22} - \lambda I)^{-1}\frac{1}{\epsilon}a_{21}\right) = 0 \, .$$

First, we set the first determinant equal to zero:

$$\det\left(\frac{1}{\epsilon}a_{22} - \lambda I\right) = \frac{1}{\epsilon^{d_2}} \det\left(a_{22} - \epsilon\lambda I\right) = 0 \, .$$

By definition, $\epsilon\lambda$ are the eigenvalues of a_{22}, thus they are of order $\mathcal{O}(1)$. Consequently, we have d_2 (not necessarily distinct) real eigenvalues of order $\mathcal{O}(\frac{1}{\epsilon})$.

If the determinant of the second matrix is zero, we have

$$\det\left((a_{11}-\lambda I)-a_{12}(\frac{1}{\epsilon}a_{22}-\lambda I)^{-1}\frac{1}{\epsilon}a_{21}\right)=0\ .$$

We apply Taylor expansion on $(a_{22}-\epsilon\lambda I)^{-1}$ at $\epsilon=0$. We have,

$$(a_{22}-\epsilon\lambda I)^{-1}=a_{22}^{-1}+\epsilon\lambda a_{22}^{-2}+\mathcal{O}(\epsilon^2)\ .$$

We substitute the above expansion into the determinant,

$$\det\left(a_{11}-a_{12}(a_{22}^{-1}+\epsilon\lambda a_{22}^{-2}+\mathcal{O}(\epsilon^2))a_{21}-\lambda I\right)=0\ ,$$

and we find it is equivalent to finding the eigenvalues of a perturbed matrix of $\tilde{a}$,

$$\det\left(\tilde{a}-\epsilon\lambda(a_{12}a_{22}^{-2}a_{21})-\mathcal{O}(\epsilon^2)-\lambda I\right)=\det\left(\tilde{a}+\mathcal{O}(\epsilon)-\lambda I\right)=0\ .$$

By Theorem 2 in [42], on the eigenvalues of a perturbed matrix, we know that the corresponding d_1 (not necessarily distinct) real eigenvalues are of order $\mathcal{O}(1)$. Therefore, we can decompose $\mathbf{a}$ as

$$\mathbf{a}=PDP^{-1}\text{ with }D=\begin{pmatrix}D_1 & 0\\ 0 & \frac{1}{\epsilon}D_2\end{pmatrix}$$

where D is the diagonal matrix, for which $D_1\in\mathbb{R}^{d_1\times d_1}$ and $D_2\in\mathbb{R}^{d_2\times d_2}$ are diagonal blocks of the eigenvalues of order $\mathcal{O}(1)$. We also define $\Sigma=P^{-1}\mathbf{q}(P^{-1})^*$. Using Lemma 9.14 in Appendix 9.11, we have the ratio between diagonal elements of D and Σ is always

$$D_{ii}/\Sigma_{ii}=\mathcal{O}(1)\ .$$

We apply Theorem 2.3 to the system of equations (3.5). We have

$$\mathbb{E}\left(\sup_{0\le t\le T}\|\mathbf{x}(t)\|^2\right)\le C\frac{\log\left(1+\max_i(|D_{ii}|)T\right)}{\min_i(|D_{ii}/\Sigma_{ii}|)}\ .$$

Since $D_{ii}/\Sigma_{ii}=\mathcal{O}(1)$, $\max_i|D_{ii}|=\mathcal{O}(\frac{1}{\epsilon})$, we have

$$\mathbb{E}\left(\sup_{0\le t\le T}\|\mathbf{x}(t)\|^2\right)=\mathcal{O}\left(\log(1+T/\epsilon)\right).$$

Since $\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$, we get

$$\mathbb{E}\left(\sup_{0\le t\le T}\left(\|x(t)\|^2 + \|y(t)\|^2\right)\right) = \mathcal{O}\left(\log(1+\frac{T}{\epsilon})\right) .$$

This completes the proof. □

Theorem 3.6. *Let Assumptions 3.1 hold for system* (3.1)*. Suppose that x and X are solutions of (3.1a) and* (3.2) *respectively, corresponding to the same realization of the U process and $x(0) = X(0)$. Then, x converges to X in L^2. More specifically,*

$$\mathbb{E}\sup_{0\le t\le T}\|x(t) - X(t)\|^2 \le c(\epsilon^2\log(\frac{T}{\epsilon}) + \epsilon T)e^T ,$$

when T is fixed finite, the above bound can be simplified to

$$\mathbb{E}\sup_{0\le t\le T}\|x(t) - X(t)\|^2 = \mathcal{O}(\epsilon).$$

Proof. For auxiliary equations used in the proof, please refer to the construction in [78]. The generator of system (3.1) is

$$\mathcal{L}_{avg} = \frac{1}{\epsilon}\mathcal{L}_0 + \mathcal{L}_1,$$

where

$$\begin{aligned} \mathcal{L}_0 &= (a_{21}x + a_{22}y)\cdot\nabla_y + \frac{1}{2}q_2 : \nabla_y\nabla_y \\ \mathcal{L}_1 &= (a_{11}x + a_{12}y)\cdot\nabla_x + \frac{1}{2}q_1 : \nabla_x\nabla_x \end{aligned}$$

To prove that the L^2 error between the solutions $x(t)$ and $X(t)$ is of order $\mathcal{O}(\sqrt{\epsilon})$, we first need to find the function $\Phi(x,y)$ which solves the Poisson equation

$$-\mathcal{L}_0\Phi = a_{11}x + a_{12}y - \tilde{a}x \ , \ \int_{\mathcal{Y}}\Phi\rho(y;x)dy = 0; \tag{3.6}$$

where $\rho(y;x)$ is the invariant density of y in (3.1b) with x fixed. In this case, the partial differential equation (3.6) is linear and can be solved explicitly

$$\Phi(x,y) = \Phi(y) = -(a_{12}a_{22}^{-1})y. \tag{3.7}$$

Applying Itô formula to $\Phi(x,y)$, we get

$$\frac{d\Phi}{dt} = \frac{1}{\epsilon}\mathcal{L}_0\Phi + \mathcal{L}_1\Phi + \frac{1}{\sqrt{\epsilon}}\sqrt{q_2}\nabla_y\Phi\frac{dV_t}{dt},$$

and substituting into (3.1a) gives

$$\begin{aligned}\frac{dx}{dt} &= (\tilde{a}x - \mathcal{L}_0\Phi) + \sqrt{q_1}\frac{dU_t}{dt} \\ &= \tilde{a}x - \epsilon\frac{d\Phi}{dt} + \epsilon\mathcal{L}_1\Phi + \sqrt{\epsilon}\sqrt{q_2}\nabla_y\Phi\frac{dV_t}{dt} + \sqrt{q_1}\frac{dU_t}{dt}. \end{aligned} \tag{3.8}$$

Define

$$\theta(t) := (\Phi(x(t),y(t)) - \Phi(x(0),y(0))) - \int_0^t (a_{11}x(s) + a_{12}y(s))\cdot\nabla_x\Phi ds.$$

From (3.7), we see that Φ does not depend on x and thus

$$\begin{aligned}\theta(t) &= \Phi(x(t),y(t)) - \Phi(x(0),y(0)) \\ &= -(a_{12}a_{22}^{-1})(y(t)-y(0)). \end{aligned} \tag{3.9}$$

Now define

$$\begin{aligned}M(t) &:= -\int_0^t \sqrt{q_2}\nabla_y\Phi(x(s),y(s))dV_s \\ &= -\int_0^t \sqrt{q_2}(a_{12}a_{22}^{-1})^* dV_s. \end{aligned}$$

Itô isometry gives

$$\mathbb{E}\|M(t)\|^2 = ct \tag{3.10}$$

The solution of (3.1a) in the form of (3.8) is

$$x(t) = x(0) + \int_0^t \tilde{a}x(s)ds + \epsilon\theta(t) + \sqrt{\epsilon}M(t) + \sqrt{q_1}\int_0^t dU_s\ .$$

Also, from the averaged equation (3.2), we get

$$X(t) = X(0) + \int_0^t \tilde{a}X(s)ds + \sqrt{q_1}\int_0^t dU_s\ .$$

Let $e(t) = x(t) - X(t)$. By assumption, $e(0) = 0$ and

$$e(t) = \int_0^t \tilde{a}\,(x(s) - X(s))\,ds + \epsilon\theta(t) + \sqrt{\epsilon}M(t)\,. \tag{3.11}$$

Then,

$$\|e(t)\|^2 \leq 3\|\tilde{a}\int_0^t e(s)ds\|^2 + 3\epsilon^2\|\theta(t)\|^2 + 3\epsilon\|M(t)\|^2\,.$$

By applying Lemma 3.5 on (3.11), the Burkholder-Davis-Gundy inequality (see Appendix 9.2) and Hölder's inequality (see Appendix 9.3), we get

$$\begin{aligned} & \mathbb{E}\left(\sup_{0\leq t\leq T}\|e(t)\|^2\right) \\ \leq\ & c\left(\int_0^T \mathbb{E}\|e(s)\|^2 ds + \epsilon^2\log(\frac{T}{\epsilon}) + \epsilon T\right) \\ \leq\ & c\left(\epsilon^2\log(\frac{T}{\epsilon}) + \epsilon T + \int_0^T \mathbb{E}\sup_{0\leq u\leq s}\|e(u)\|^2 ds\right). \end{aligned}$$

By Gronwall's inequality, we deduce that

$$\mathbb{E}\left(\sup_{0\leq t\leq T}\|e(t)\|^2\right) \leq c(\epsilon^2\log(\frac{T}{\epsilon}) + \epsilon T)e^T.$$

When T is fixed, we have

$$\mathbb{E}\left(\sup_{0\leq t\leq T}\|e(t)\|^2\right) = \mathcal{O}\left(\epsilon\right).$$

This completes the proof. □

3.3 The Drift Estimator

Suppose that we want to estimate the drift of the process X described by (3.2), but we only observe a solution $\{x(t)\}_{t\in(0,T)}$ of (3.1a). According to the previous theorem, x is a good approximation of X, so we replace X in the formula of the MLE (2.5) by x. In the following theorem, we show that the error we will be making is insignificant, in a sense to be made precise.

Theorem 3.7. *Suppose that x is the projection to the x-coordinate of a solution of system* (3.1) *satisfying Assumptions 3.1. Let $\hat{a}_T^\epsilon$ be the estimate we get by replacing X in* (2.5) *by*

x, *i.e.*

$$\hat{a}_T^\epsilon = \left(\int_0^T dx \otimes x\right)\left(\int_0^T x \otimes x dt\right)^{-1}. \tag{3.12}$$

Then,

$$\lim_{\epsilon\to 0}\lim_{T\to\infty} \mathbb{E}\|\hat{a}_T^\epsilon - \tilde{a}\|^2 = 0\,.$$

Proof. We define

$$I_1 = \frac{1}{T}\int_0^T dx \otimes x \text{ and } I_2 = \frac{1}{T}\int_0^T x \otimes x dt.$$

By ergodicity, which is guaranteed by Assumptions 3.1 *(iii) and (iv)*

$$\lim_{T\to\infty} I_2 = \mathbb{E}(x\otimes x) = C \neq 0 \; a.s.,$$

which is a constant invertible matrix. We expand dx using Itô formula applied on Φ as in (3.8):

$$I_1 = J_1 + J_2 + J_3 + J_4 + J_5$$

where

$$J_1 = \frac{1}{T}\int_0^T \tilde{a}x \otimes x dt$$

$$J_2 = \frac{\epsilon}{T}\int_0^T d\Phi \otimes x$$

$$J_3 = \frac{\epsilon}{T}\int_0^T \mathcal{L}_1\Phi \otimes x dt$$

$$J_4 = \frac{\sqrt{\epsilon}}{T}\int_0^T \nabla_y\Phi\sqrt{q_2}dV_t \otimes x$$

$$J_5 = \frac{1}{T}\sqrt{q_1}\int_0^T dU_t \otimes x$$

It is obvious that

$$J_1 = \tilde{a}I_2.$$

Since Φ is linear in y, and by Itô isometry, we get

$$\begin{aligned}\mathbb{E}\left(\|J_4\|^2\right) &= \frac{c\epsilon}{T}\mathbb{E}\|\frac{1}{T}\int_0^T dV_t \otimes x(t)\|^2 \\ &= \frac{c\epsilon}{T}\mathbb{E}\left(\frac{1}{T}\int_0^T \|x(t)\|^2 dt\right)\end{aligned}$$

by ergodicity, we have

$$\mathbb{E}\left(\|J_4\|^2\right) = \frac{c\epsilon}{T}\,.$$

Similarly for J_5,

$$\begin{aligned}\mathbb{E}\left(\|J_5\|^2\right) &= \frac{c}{T}\mathbb{E}\|\frac{1}{T}\int_0^T dU_t \otimes x(t)\|^2 \\ &= \frac{c}{T}\mathbb{E}\left(\frac{1}{T}\int_0^T \|x(t)\|^2 dt\right) \\ &= \frac{c}{T}\end{aligned}$$

We know Φ is independent of x, so

$$J_3 \equiv 0.$$

Finally, using (3.7) and (3.1b) we break J_2 further into

$$J_2 = -\frac{1}{T}\int_0^T (a_{12}a_{22}^{-1})(a_{21}x + a_{22}y) \otimes x dt - \frac{a_{12}a_{22}^{-1}\sqrt{\epsilon q_2}}{T}\int_0^T dV_t \otimes x$$

Again, using Itô isometry and ergodicity, we bound the L^2 norm of the second term by

$$\mathbb{E}\|\frac{a_{12}a_{22}^{-1}\sqrt{\epsilon q_2}}{T}\int_0^T dV_t \otimes x\|^2 \leq \frac{c\epsilon}{T}.$$

By ergodicity, the first term converges in L^2 as $T \to \infty$,

$$-\frac{a_{12}a_{22}^{-1}}{T}\int_0^T (a_{21}x + a_{22}y) \otimes x dt \to -a_{12}\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x\right).$$

We write the expectation as

$$\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x\right) = \mathbb{E}_{\rho^\epsilon}\left(\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x|x\right)\right)$$

Clearly, the limit of ρ^ϵ conditioned on x is a normal distribution with mean $-a_{22}^{-1}a_{21}x$ by (2.3). Thus, we see that

$$\lim_{\epsilon\to 0}\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x\right) = 0.$$

Putting everything together, we see that

$$\lim_{\epsilon\to 0}\lim_{T\to\infty}(I_1 - \tilde{a}I_2) = 0 \quad \text{in} \quad L^2$$

Since the denominator I_2 of $\hat{a}_T^\epsilon$ converges almost surely, the result follows. □

3.4 Asymptotic Normality for the Drift Estimator

We extend the proof of theorem 3.7 to prove asymptotic normality for the estimator $\hat{a}_T^\epsilon$. We have seen that

$$\hat{a}_T^\epsilon - \tilde{a} = (J_2 + J_4 + J_5)I_2^{-1}.$$

We will show that

$$\sqrt{T}\left(\hat{a}_T^\epsilon - \tilde{a} + a_{12}\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x\right)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_\epsilon^2\right)$$

and compute the limit of σ_ϵ^2 as $\epsilon \to 0$. First we apply the Central Limit Theorem for martingales to J_4 and J_5 (see [36]). We find that

$$\sqrt{T}J_4 \xrightarrow{D} \mathcal{N}\left(0, \sigma(4)_\epsilon^2\right) \text{ as } T \to \infty,$$

where

$$\sigma(4)_\epsilon^2 = \epsilon a_{12}a_{22}^{-1}q_2 a_{22}^{-1^*}a_{12}^*\mathbb{E}_{\rho^\epsilon}(x \otimes x);$$

and

$$\sqrt{T}J_5 \xrightarrow{D} \mathcal{N}\left(0, \sigma(5)_\epsilon^2\right) \text{ as } T \to \infty,$$

where

$$\sigma(5)_\epsilon^2 = q_1\mathbb{E}_{\rho^\epsilon}(x \otimes x).$$

We write $J_2 = J_{2,1} + J_{2,2}$ where

$$J_{2,1} = -\frac{a_{12}a_{22}^{-1}}{T}\int_0^T (a_{21}x + a_{22}y) \otimes x dt \text{ and } J_{2,2} = -\frac{a_{12}a_{22}^{-1}\sqrt{\epsilon q_2}}{T}\int_0^T dV \otimes x.$$

Once again, we apply the Central Limit Theorem for martingales to $J_{2,2}$ and we find

$$\sqrt{T}J_{2,2} \xrightarrow{D} \mathcal{N}\left(0, \sigma(2,2)_\epsilon^2\right) \text{ as } T \to \infty$$

where

$$\sigma(2,2)_\epsilon^2 = \epsilon a_{12}a_{22}^{-1}q_2 a_{22}^{-1^*}a_{12}^*\mathbb{E}_{\rho^\epsilon}(x \otimes x).$$

Finally, we apply the Central Limit Theorem for functionals of ergodic Markov Chains to $J_{2,1}$ (see [16]). We get

$$\sqrt{T}\left(J_{2,1} + a_{12}\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x\right)\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma(2,1)_\epsilon^2\right)$$

as $T \to \infty$, where

$$\begin{aligned}\sigma(2,1)_\epsilon^2 &= \int_{\mathcal{X}\times\mathcal{Y}} \xi(x,y)\xi(x,y)^* \rho_\epsilon(x,y)dxdy \\ &+ 2\int_{\mathcal{X}\times\mathcal{Y}} \xi(x,y) \int_0^\infty (P_t^\epsilon \xi)(x,y)dt\rho_\epsilon(x,y)dxdy\end{aligned}$$

with

$$\xi(x,y) = -\left(a_{12}a_{22}^{-1}a_{21}x + a_{12}y\right) \otimes x + \mathbb{E}\left(a_{12}a_{22}^{-1}a_{21}x + a_{12}y\right) \otimes x$$

and

$$(P_t^\epsilon \xi)(x,y) = \mathbb{E}\left(\xi(x(t),y(t))|x(0)=x, y(0)=y\right).$$

Putting everything together, we get that as $T \to \infty$,

$$\sqrt{T}(J_2 + J_4 + J_5) \to X_{2,1} + X_{2,2} + X_4 + X_5$$

in law, where $X_i \sim \mathcal{N}(0, \sigma(i)_\epsilon^2)$ for $i \in \{\{2,1\},\{2,2\},4,5\}$. Finally, we note that the denominator I_2 converges almost surely as $T \to \infty$ to $\mathbb{E}_{\rho^\epsilon}(x(t) \otimes x(t))$. It follows from Slutsky's theorem that as $T \to \infty$,

$$\sqrt{T}\left(\hat{a}_T^\epsilon - \tilde{a} + a_{12}\mathbb{E}_{\rho^\epsilon}\left((a_{22}^{-1}a_{21}x + y) \otimes x\right)\right) \to X_\epsilon$$

in law, where

$$X_\epsilon = (X_{2,1} + X_{2,2} + X_4 + X_5)(\mathbb{E}_{\rho^\epsilon}(x(t) \otimes x(t)))^{-1} \sim \mathcal{N}(0, \sigma_\epsilon^2).$$

It remains to compute $\lim_{\epsilon\to 0} \sigma_\epsilon^2$. We have already seen that $\sigma(2,2)_\epsilon^2 \sim \mathcal{O}(\epsilon)$ and $\sigma(4)_\epsilon^2 \sim \mathcal{O}(\epsilon)$. Thus, we need to compute

$$\begin{aligned}&\lim_{\epsilon\to 0} \mathbb{E}\left((X_{2,1} + X_5) \otimes (X_{2,1} + X_5)\right) \\ =\ &\lim_{\epsilon\to 0} \mathbb{E}\left(X_{2,1} \otimes X_{2,1} + X_{2,1} \otimes X_5 + X_5 \otimes X_{2,1} + X_5 \otimes X_5\right)\end{aligned}$$

First, we see that

$$\lim_{\epsilon\to 0} \mathbb{E}(X_5 \otimes X_5) = q_1 \lim_{\epsilon\to 0} \mathbb{E}_{\rho_\epsilon}(x \otimes x) = q_1 \mathbb{E}(X \otimes X) = q_1 q_1^\infty$$

for which the variance of the invariant distribution of X is defined as $\mathrm{vec}(q_1^\infty) = (-\tilde{a} \oplus -\tilde{a})^{-1}\mathrm{vec}(q_1)$.

To compute $\lim_{\epsilon\to 0} \mathbb{E}(X_{2,1}^2)$ first we set $\tilde{y} = a_{22}^{-1}a_{21}x + y$. Then, $(x, \tilde{y})$ is also an ergodic process with invariant distribution $\tilde{\rho}_\epsilon$ that converges as $\epsilon \to 0$ to $\mathcal{N}(0, q_1^\infty) \otimes$

$\mathcal{N}(0, q_2^\infty)$, for which the variance of the invariant distributions q_1^∞ and q_2^∞ are computed following (2.4),

$$\mathrm{vec}(q_1^\infty) = (-\tilde{a} \oplus -\tilde{a})^{-1}\mathrm{vec}(q_1)\ ,\ \mathrm{vec}(q_2^\infty) = (-a_{22} \oplus -a_{22})^{-1}\mathrm{vec}(q_2).$$

Since $\xi(x, \tilde{y}) = -a_{12}\tilde{y} \otimes x$, it follows that

$$\lim_{\epsilon\to 0} \mathbb{E}_{\rho^\epsilon}(\xi(x, \tilde{y}) \otimes \xi(x, \tilde{y})) = a_{12}\sqrt{q_2^\infty}q_1^\infty\sqrt{q_2^\infty}^* a_{12}^*.$$

In addition, as $\epsilon \to 0$, the process $\tilde{y}$ decorrelates exponentially fast. Thus

$$\lim_{\epsilon\to 0}(P_t^\epsilon \xi)(x, y) = a_{12}\mathbb{E}(X(t)|X(0) = x)\mathbb{E}(\tilde{y}) \equiv 0$$

for all $t \geq 0$. As $t \to \infty$, the process $(x, \tilde{y})$ also converges exponentially fast to a mean-zero Gaussian distribution and thus the integral with respect to t is finite. We conclude that the second term of $\sigma(2, 1)_\epsilon^2$ disappears as $\epsilon \to 0$ and thus

$$\lim_{\epsilon\to 0} \mathbb{E}(X_{2,1} \otimes X_{2,1}) = a_{12}\sqrt{q_2^\infty}q_1^\infty\sqrt{q_2^\infty}^* a_{12}^*.$$

Finally, we show that

$$\lim_{\epsilon\to 0} \mathbb{E}(X_{2,1} \otimes X_5) = 0.$$

Clearly, X_5 is independent of $\tilde{y}$ in the limit, since it only depends on x and U. So,

$$\lim_{\epsilon\to 0} \mathbb{E}(X_{2,1} \otimes X_5) = \lim_{\epsilon\to 0} \mathbb{E}\left(\mathbb{E}(X_{2,1} \otimes X_5|x)\right)$$

and

$$\lim_{\epsilon\to 0} \mathbb{E}\left(\mathbb{E}(X_{2,1}|x)\right) = 0$$

for the same reasons as above. Similar calculations give

$$\lim_{\epsilon\to 0} \mathbb{E}(X_5 \otimes X_{2,1}) = 0.$$

Thus

$$\lim_{\epsilon\to 0} \sigma_\epsilon^2 = \left(q_1 q_1^\infty + a_{12}\sqrt{q_2^\infty}q_1^\infty\sqrt{q_2^\infty}^* a_{12}^*\right)(q_1^\infty)^{-2}\ . \tag{3.13}$$

We have proved the following

Theorem 3.8. *Suppose that x is the projection to the x-coordinate of a solution of system* (3.1) *satisfying Assumptions 3.1. Let $\hat{a}_T^\epsilon$ be as in* (3.12)*. Then, as $T \to \infty$,*

$$\sqrt{T}\left(\hat{a}_T^\epsilon - \tilde{a}\right) \xrightarrow{D} \mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2),$$

where μ_ϵ and σ_ϵ are dependent on ϵ, whilst $\mu_\epsilon \to 0$ and σ_ϵ^2 converges to the limit in (3.13) *as $\epsilon \to 0$.*

3.5 The Diffusion Estimator

Suppose that we want to estimate the diffusion parameter of the process X described by (3.2), but we only observe a solution $\{x(t)\}_{t\in(0,T)}$ of (3.1a). As before, we replace X in the formula of the MLE (2.6) by x. In the following theorem, we show that the estimator is still consistent in the limit.

Theorem 3.9. *Suppose that x is the projection to the x-coordinate of a solution of system* (3.1) *satisfying Assumptions 3.1. We set*

$$\hat{q}_\delta^\epsilon = \frac{1}{T}\sum_{n=0}^{N-1} (x_{n+1} - x_n) \otimes (x_{n+1} - x_n) \tag{3.14}$$

where $x_n = x(n\delta)$ is the discretized x process, $\delta \leq \epsilon$ is the discretization step and $T = N\delta$ is fixed. Then, for every $\epsilon > 0$

$$\lim_{\delta\to 0} \mathbb{E}\|\hat{q}_\delta^\epsilon - q_1\|^2 = 0 \,,$$

more specifically,

$$\mathbb{E}\|\hat{q}_\delta^\epsilon - q_1\|^2 = \mathcal{O}(\delta) \,.$$

Proof. We rewrite $x_{n+1} - x_n$ using discretized (3.1a),

$$x_{n+1} - x_n = \int_{n\delta}^{(n+1)\delta} \sqrt{q_1} dU_s + \hat{R}_1^{(n)} + \hat{R}_2^{(n)} \tag{3.15}$$

where

$$\begin{aligned} \hat{R}_1^{(n)} &= a_{11}\int_{n\delta}^{(n+1)\delta} x(s)ds \\ \hat{R}_2^{(n)} &= a_{12}\int_{n\delta}^{(n+1)\delta} y(s)ds \end{aligned}$$

We let $\xi_n = \frac{1}{\sqrt{\delta}}\left(U_{(n+1)\delta} - U_{n\delta}\right)$. Since U is a Brownian motion, $\{\xi_n\}_{n\geq 0}$ is a sequence of independent standard Gaussian random variables. We write

$$\int_{n\delta}^{(n+1)\delta} \sqrt{q_1} dU_s = \sqrt{q_1\delta}\xi_n.$$

We can write the estimator as

$$\begin{aligned}
\hat{q}_\delta^\epsilon &= q_1 \frac{1}{N}\sum_{n=0}^{N-1} \xi_n^2 \\
&+ \frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1} \xi_n \otimes (\hat{R}_1^{(n)} + \hat{R}_2^{(n)}) \\
&+ \frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1} (\hat{R}_1^{(n)} + \hat{R}_2^{(n)}) \otimes \xi_n \\
&+ \frac{1}{N\delta}\sum_{n=0}^{N-1} (\hat{R}_1^{(n)} + \hat{R}_2^{(n)})^2
\end{aligned}$$

Hence, we can expand the error as

$$\begin{aligned}
\mathbb{E}\left(\hat{q}_\delta^\epsilon - q_1\right)^2 &\leq C\mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1}\xi_n^2 - 1\right)^2 && \text{(3.16a)} \\
&+ C\frac{q_1}{N^2\delta}\mathbb{E}\left(\sum_{n=0}^{N-1}\xi_n \otimes (\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right)^2 && \text{(3.16b)} \\
&+ C\frac{q_1}{N^2\delta}\mathbb{E}\left(\sum_{n=0}^{N-1}(\hat{R}_1^{(n)} + \hat{R}_2^{(n)}) \otimes \xi_n\right)^2 && \text{(3.16c)} \\
&+ C\frac{1}{N^2\delta^2}\mathbb{E}\left(\sum_{n=0}^{N-1}(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})^2\right)^2 && \text{(3.16d)}
\end{aligned}$$

It is straightforward for line (3.16a),

$$\mathbb{E}\left(\frac{1}{N}\sum_{n=0}^{N-1}\xi_n^2 - 1\right)^2 = c\delta\ .$$

By Assumptions 3.1(v), and Hölder inequality, we have,

$$\begin{aligned}
\mathbb{E}(\hat{R}_1^{(n)})^2 &= a_{11}^2\mathbb{E}\left(\int_{n\delta}^{(n+1)\delta} x(s)ds\right)^2 && \text{(3.17)} \\
&\leq ca_{11}^2\delta\int_{n\delta}^{(n+1)\delta}\mathbb{E}x(s)^2 ds \\
&\leq c\delta^2\ .
\end{aligned}$$

It is similar for $\mathbb{E}(\hat{R}_2^{(n)})^2$,

$$\begin{aligned}\mathbb{E}(\hat{R}_2^{(n)})^2 &= a_{12}^2\mathbb{E}\left(\int_{n\delta}^{(n+1)\delta} y(s)ds\right)^2 \\ &\leq ca_{12}^2\delta\int_{n\delta}^{(n+1)\delta}\mathbb{E}y(s)^2ds \\ &\leq c\delta^2\ .\end{aligned} \tag{3.18}$$

Since $\hat{R}_1^{(n)}$ and $\hat{R}_2^{(n)}$ are Gaussian random variables, we have $\mathbb{E}(\hat{R}_1^{(n)}+\hat{R}_2^{(n)})^4 = C\delta^4$, so line (3.16d) is of order $\mathcal{O}(\delta^2)$. For line (3.16b), we need to get the correlation between $\hat{R}_i^{(n)}$ for $i\in\{1,2\}$ and ξ_n. We write system (4.1) in integrated form,

$$x(s) = x_n + a_{11}\int_{n\delta}^{s} x(u)du + a_{12}\int_{n\delta}^{s} y(u)du + \sqrt{q_1}\int_{n\delta}^{s} dU_u \tag{3.19}$$

$$y(s) = y_n + \frac{a_{21}}{\epsilon}\int_{n\delta}^{s} x(u)du + \frac{a_{22}}{\epsilon}\int_{n\delta}^{s} y(u)du + \frac{\sqrt{q_2}}{\epsilon}\int_{n\delta}^{s} dV_u \tag{3.20}$$

We substitute (3.19) and (3.20) into $\hat{R}_1^{(n)}$ and $\hat{R}_2^{(n)}$ respectively,

$$\begin{aligned}\hat{R}_1^{(n)}+\hat{R}_2^{(n)} &= \int_{n\delta}^{(n+1)\delta} a_{11}x(s)+a_{12}y(s)ds \\ &= a_{11}x_n\delta + a_{12}y_n\delta \\ &+ \left(a_{11}^2+\frac{1}{\epsilon}a_{12}a_{21}\right)\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} x(u)duds \\ &+ \left(a_{11}a_{12}+\frac{1}{\epsilon}a_{12}a_{22}\right)\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} y(u)duds \\ &+ a_{11}\sqrt{q_1}\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} dU_u ds \\ &+ a_{12}\frac{\sqrt{q_2}}{\epsilon}\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} dV_u ds\end{aligned}$$

Using this expansion, we find,

$$
\begin{aligned}
& \mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right) \\
= \;& \mathbb{E}\left(\xi_n(a_{11}x_n\delta + a_{12}y_n\delta)\right) && \text{(3.21a)} \\
+ \;& \mathbb{E}\left(\xi_n\left(\left(a_{11}^2 + \frac{1}{\epsilon}a_{12}a_{21}\right)\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} x(u)duds\right)\right) && \text{(3.21b)} \\
+ \;& \mathbb{E}\left(\xi_n\left(a_{11}a_{12} + \frac{1}{\epsilon}a_{12}a_{22}\right)\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} y(u)duds\right) && \text{(3.21c)} \\
+ \;& \mathbb{E}\left(\xi_n\left(a_{11}\sqrt{q_1}\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} dU_u ds\right)\right) && \text{(3.21d)} \\
+ \;& \mathbb{E}\left(\xi_n\left(a_{12}\frac{\sqrt{q_2}}{\epsilon}\int_{n\delta}^{(n+1)\delta}\int_{n\delta}^{s} dV_u ds\right)\right) && \text{(3.21e)}
\end{aligned}
$$

By the definition of ξ_n, line (3.21a) is zero. By substituting (3.19) and (3.20) into lines (3.21b) and (3.21c) respectively and iteratively, we know they are of orders $\mathcal{O}(\delta^2)$. By definition of ξ_n, we know that line (3.21d) is of order $\mathcal{O}(\delta^{\frac{3}{2}})$. By independence between U and V, line (3.21e) is zero. Therefore,

$$\mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right) = \mathcal{O}(\delta^{\frac{3}{2}}) .$$

Thus,

$$\mathbb{E}\left(\xi_n^2(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})^2\right) = \mathcal{O}(\delta^3) .$$

When $m < n$, we have,

$$
\begin{aligned}
& \mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\xi_m(\hat{R}_1^{(m)} + \hat{R}_2^{(m)})\right) \\
= \;& \mathbb{E}\left(\mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\xi_m(\hat{R}_1^{(m)} + \hat{R}_2^{(m)})|\mathcal{F}_{n\delta}\right)\right) \\
= \;& \mathbb{E}\left(\xi_m(\hat{R}_1^{(m)} + \hat{R}_2^{(m)})\mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})|\mathcal{F}_{n\delta}\right)\right) \\
= \;& \mathbb{E}\left(\xi_m(\hat{R}_1^{(m)} + \hat{R}_2^{(m)})\right)\mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right) \\
= \;& \mathcal{O}(\delta^3) .
\end{aligned}
$$

When $m > n$, the same result holds. Thus we have that line (3.16b) is of order $\mathcal{O}(\delta^2)$. By symmetry, line (3.16c) has the same order of $\mathcal{O}(\delta^2)$. Therefore, we have for equation (3.16),

$$\mathbb{E}\left(\hat{q}_\delta^\epsilon - q_1\right)^2 = \mathcal{O}(\delta) .$$

This completes the proof. □

3.6 Asymptotic Normality for the Diffusion Estimator

To examine the asymptotic normality of the diffusion estimator, we use the decomposition of $\hat{q}_\delta^\epsilon$ in the proof of Theorem 3.9,

$$\begin{aligned}
\delta^{-\frac{1}{2}}(\hat{q}_\delta^\epsilon - q_1) &= \delta^{-\frac{1}{2}} q_1 (\frac{1}{N}\sum_{n=0}^{N-1}\xi_n^2 - I) && \text{(3.22a)}\\
&+ \delta^{-\frac{1}{2}}\frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1}\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)}) && \text{(3.22b)}\\
&+ \delta^{-\frac{1}{2}}\frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1}(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\xi_n && \text{(3.22c)}\\
&+ \delta^{-\frac{1}{2}}\frac{1}{N\delta}\sum_{n=0}^{N-1}(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})^2 && \text{(3.22d)}
\end{aligned}$$

Since

$$\lim_{\delta\to 0}\delta^{-\frac{1}{2}} q_1 (\frac{1}{N}\sum_{n=0}^{N-1}\xi_n^2 - I) = \lim_{N\to\infty}\frac{q_1}{\sqrt{T}}\frac{1}{\sqrt{N}}\sum_{n=0}^{N-1}(\xi_n^2 - I)$$

It follows from Central Limit Theorem for sum of multivariate i.i.d random variables, as $\delta \to 0$,

$$\lim_{\delta\to 0}\delta^{-\frac{1}{2}} q_1 (\frac{1}{N}\sum_{n=0}^{N-1}\xi_n^2 - I) \xrightarrow{D} \mathcal{N}(0, 2\frac{q_1^2}{T})$$

We have shown that $\mathbb{E}\left(\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right) = \mathcal{O}(\delta^{\frac{3}{2}})$, so line (3.22b) has mean

$$\mathbb{E}\left(\delta^{-\frac{1}{2}}\frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1}\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right) = \mathcal{O}(\delta^{\frac{1}{2}})\ .$$

Using $\mathbb{E}\left(\sum_{n=0}^{N-1}\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right)^2 = \mathcal{O}(\delta)$, we find the second moment of (3.22b),

$$\mathbb{E}\left(\delta^{-\frac{1}{2}}\frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1}\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)})\right)^2 = \mathcal{O}(\delta)\ .$$

Thus when δ is small,

$$\delta^{-\frac{1}{2}}\frac{\sqrt{q_1}}{N\sqrt{\delta}}\sum_{n=0}^{N-1}\xi_n(\hat{R}_1^{(n)} + \hat{R}_2^{(n)}) \sim \mathcal{N}(\mathcal{O}(\delta^{\frac{1}{2}}), \mathcal{O}(\delta))\ .$$

By symmetry, same result holds for line (3.22c). Finally, for line (3.22d), using (3.17) and (3.18), we have

$$\mathbb{E}\left(\delta^{-\frac{1}{2}}\frac{1}{N\delta}\sum_{n=0}^{N-1}(\hat{R}_1^{(n)}+\hat{R}_2^{(n)})^2\right)=\mathcal{O}(\delta^{\frac{1}{2}})\,,$$

and,

$$\mathbb{E}\left(\delta^{-\frac{1}{2}}\frac{1}{N\delta}\sum_{n=0}^{N-1}(\hat{R}_1^{(n)}+\hat{R}_2^{(n)})^2\right)^2=\mathcal{O}(\delta)\,.$$

Thus,

$$\delta^{-\frac{1}{2}}\frac{1}{N\delta}\sum_{n=0}^{N-1}(\hat{R}_1^{(n)}+\hat{R}_2^{(n)})^2\sim\mathcal{N}(\mathcal{O}(\delta^{\frac{1}{2}}),\mathcal{O}(\delta))\,.$$

Putting all terms together, we have

$$\delta^{-\frac{1}{2}}\left(\hat{q}_\delta^\epsilon-q_1\right)\xrightarrow{D}\mathcal{N}(0,\frac{2q_1^2}{T})\,. \tag{3.23}$$

We have proved the following,

Theorem 3.10. *Under the conditions of Theorem 3.9 and with the same notation, it holds that*

$$\delta^{-\frac{1}{2}}\left(\hat{q}_\delta^\epsilon-q_1\right)\xrightarrow{D}\mathcal{N}(0,\frac{2q_1^2}{T})\text{ as }\delta\to 0\,.$$

3.7 Numerical Example

We show our findings in this chapter through the a numerical example. The multiscale system of interest is

$$\frac{dx}{dt} = -x+y+\sqrt{2}\frac{dU_t}{dt} \tag{3.24a}$$

$$\frac{dy}{dt} = \frac{1}{\epsilon}(-x-y)+\sqrt{\frac{2}{\epsilon}}\frac{dV_t}{dt}, \tag{3.24b}$$

The averaged equation is

$$\frac{dX}{dt}=-2X+\sqrt{2}\frac{dU_t}{dt}. \tag{3.25}$$

We first examine the convergence of the drift estimator in Theorem 3.7. We fix the scale parameter at $\epsilon=2^{-9},2^{-6}$ and 2^{-3}, observation time increment $\delta=2^{-10}$, and let the number of observations N increase from 2^{11} to 2^{18}. For each set of the parameters, we sample 100 paths using the exact solution.

We first show the consistency of the estimator by plotting the L^2 norm of the errors

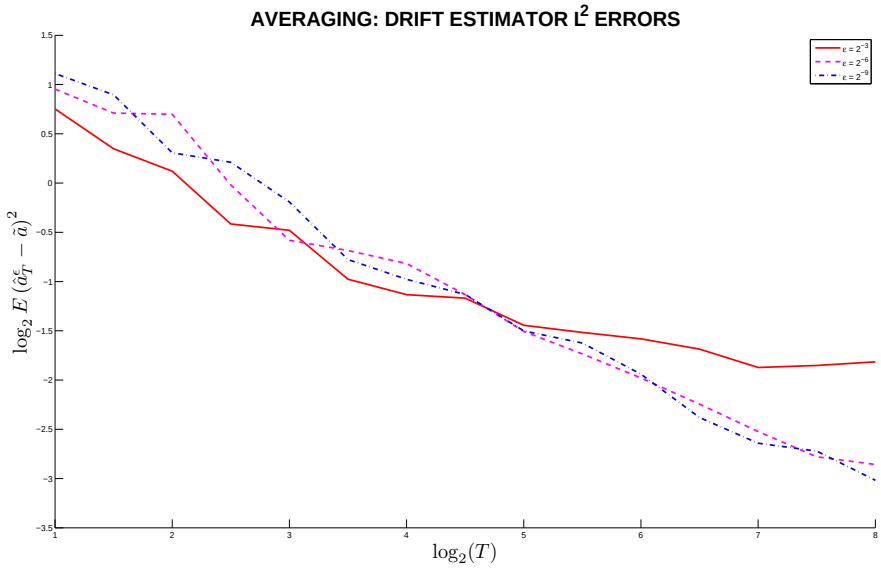

Figure 3.1: Averaging: Consistency of Estimator $\hat{a}_T^\epsilon$

$(\hat{a}_T^\epsilon - \tilde{a})$, in Figure 3.1. We see that when $T = N\delta$ is short, the estimation error from observations of different scale parameter ϵ's are similar. When time is large, the error with small scale parameter ϵ continues to decrease at a constant rate.

We then show the asymptotic variance of the estimator by plotting the distribution of the time adjusted errors $\sqrt{T}(\hat{a}_T^\epsilon - \mathbb{E}(\hat{a}_T^\epsilon))$ with $\epsilon = 2^{-9}$, in Figure 3.2. The asymptotic variance is computed using (3.13), which is 6 in our case. The red lines are the 2.5 and 97.5 quantiles of the adjusted errors, the blue lines are the expected confidence intervals of the adjusted errors. When observation time T is large, the confidence intervals of the simulated errors are contained in the expected confidence intervals.

We then examine the convergence of the diffusion estimator in Theorem 3.9. We fix the total time horizon $T = N\delta = 1$, and the scale parameter $\epsilon = 2^{-9}, 2^{-6}$ and 2^{-3}. We decrease the observation time increment δ from 2^{-9} to 2^{-17}. For each set of parameters, we sample 100 paths using the exact solution.

We first show the consistency of the estimator by plotting the L^2 norm of the errors $(\hat{q}_\delta^\epsilon - q_1)$, in Figure 3.3. We see that as δ gets small, the estimation error from observations of different scale parameter ϵs are similar. This shows that the error is irrelevant to what value the scale parameter takes, and are always converging to zero.

We then show the asymptotic variance of the estimator by plotting the distribution of the δ adjusted errors $\delta^{-\frac{1}{2}}(\hat{q}_\delta^\epsilon - \mathbb{E}(\hat{q}_\delta^\epsilon))$ with $\epsilon = 2^{-9}$, in Figure 3.4. The variance is computed using (3.23), which is 8 in our example. The red lines are the 2.5 and 97.5

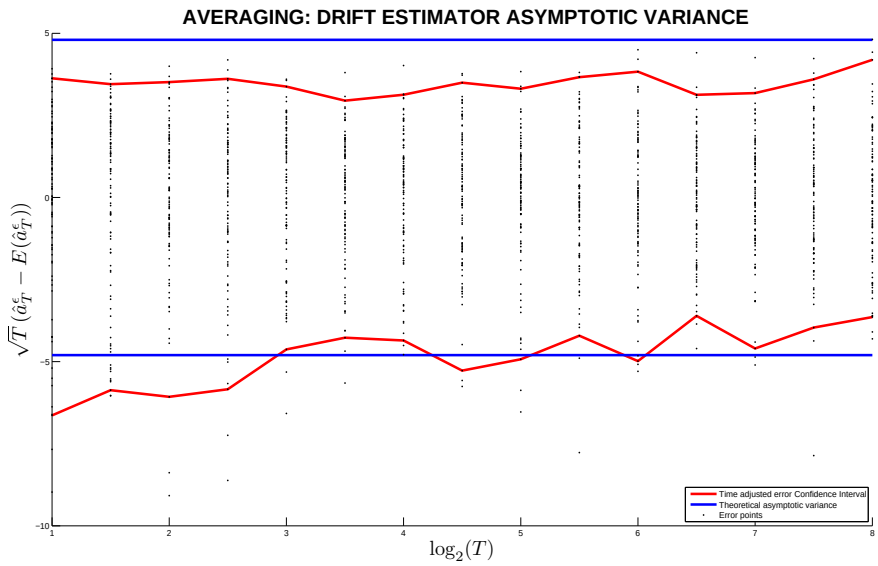

Figure 3.2: Averaging: Asymptotic Normality of $\hat{a}_T^\epsilon$

quantiles of the adjusted errors, the blue lines are the expected confidence intervals of the adjusted errors. We see that the confidence intervals of the simulated errors and and the expected confidence intervals agree.

3.8 Conclusion

In this chapter, we have verified asymptotic properties of the maximum likelihood estimators for the drift (2.5) and diffusion (2.6) parameters of an OU process, while observing data from the slow part of a multiscale system (3.1). We have verified that the discrepancy between the solution of the averaged equation (3.2) and the slow part of the system (3.1a), in the L^2 sense, is small when ϵ is small. In summary,

- we take continuous observations from the multiscale system (3.1a) x ;
- we have shown that the mismatch between trajectories of x and X is asymptotically small if ϵ is small;
- we have shown that the maximum likelihood estimator $\hat{a}_T^\epsilon$ converges to $\tilde{a}$ as $T \to \infty$ and $\epsilon \to 0$, and the asymptotic distribution of the estimator;
- we have shown that the maximum likelihood estimator $\hat{q}_\delta^\epsilon$ converges to $\tilde{q}$ as $\delta \to 0$, and the asymptotic distribution of the estimator. We notice the behaviour of $\hat{q}_\delta^\epsilon$ is not

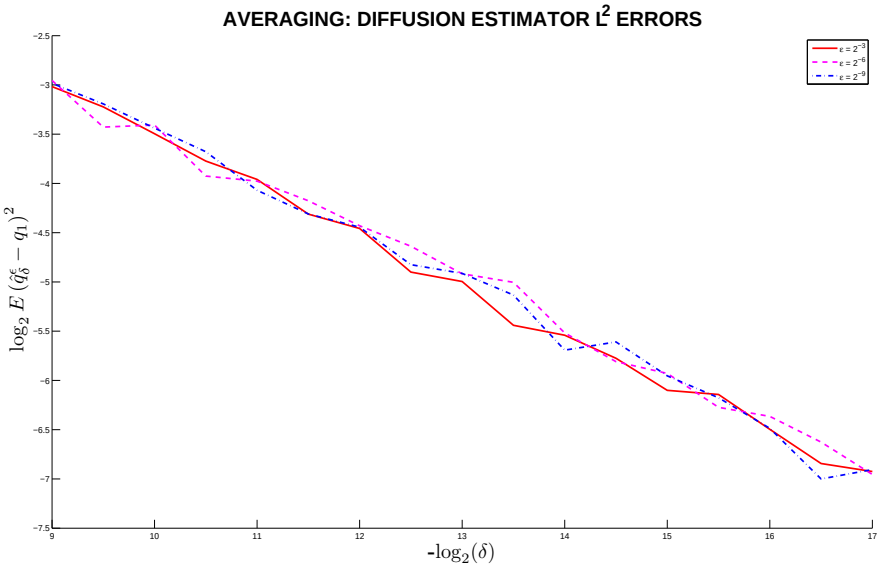

Figure 3.3: Averaging: Consistency of $\hat{q}_\delta^\epsilon$

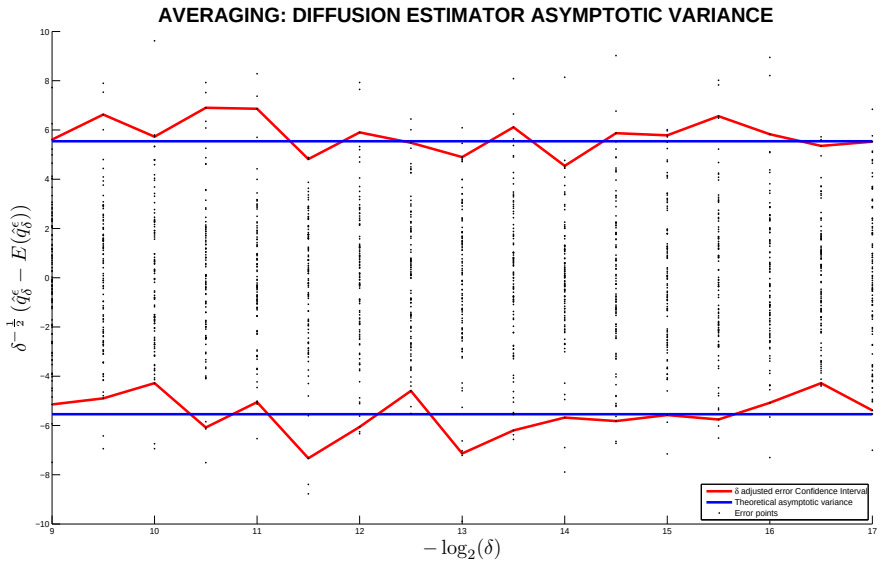

Figure 3.4: Averaging: Asymptotic Normality of $\hat{q}_\delta^\epsilon$

related to ϵ;

In future works, when possible, we can further relax the assumptions imposed on the drift and diffusion matrices, possibly in the ways addressed in the remarks to the assumptions.

Chapter 4

Parameter Estimation for the Homogenized Equation of a Multiscale OU Process

4.1 Introduction

In this chapter we consider the following fast/slow system of stochastic differential equations

$$\frac{dx}{dt} = \frac{1}{\epsilon}(a_{11}x + a_{12}y) + (a_{13}x + a_{14}y) + \sqrt{q_1}\frac{dU}{dt} \tag{4.1a}$$

$$\frac{dy}{dt} = \frac{1}{\epsilon^2}(a_{21}x + a_{22}y) + \sqrt{\frac{q_2}{\epsilon^2}}\frac{dV}{dt} \tag{4.1b}$$

for which $x \in \mathcal{X}$ $y \in \mathcal{Y}$. We may take $\mathcal{X}$ as $\mathbb{R}^{d_1}$ and $\mathcal{Y}$ as $\mathbb{R}^{d_2}$. The case where $\mathcal{X}$ is $\mathbb{T}^{d_1}$, and $\mathcal{Y}$ is $\mathbb{T}^{d_2}$ is discussed in [78]. We assume we observe data generated by the projection onto the x coordinate of the system. We also make the following assumptions.

Assumptions 4.1.
We assume that

(i) U, V *are independent Brownian motions;*

(ii) q_1, q_2 *are positive definite diagonal matrices;*

(iii) $0 < \epsilon \ll 1$*;*

(*iv*) *the system's drift matrix*

$$\begin{pmatrix} \frac{1}{\epsilon}a_{11} + a_{13} & \frac{1}{\epsilon}a_{12} + a_{14} \\ \frac{1}{\epsilon^2}a_{21} & \frac{1}{\epsilon^2}a_{22} \end{pmatrix}$$

only have negative real eigenvalues when ϵ is sufficiently small;

(*v*) a_{21} *invertible;*

(*vi*) $x(0)$ *and* $y(0)$ *are independent of* U *and* V, $(x(0), y(0))$ *is under the invariant measure of system* (3.1)*, and* $\mathbb{E}\left(\|x(0)\|^2 + \|y(0)\|^2\right) < \infty$.

Remark 4.2. *In assumption 4.1(iv), we have assumed the whole system* (4.1) *to be ergodic when ϵ is sufficiently small. This condition can be decomposed to a_{22} and $a_{13} - a_{14}a_{22}^{-1}a_{21}$ only have negative real eigenvalues; and $a_{11} - a_{12}a_{22}^{-1}a_{21} = 0$, which ensures the fast scale term in (4.1a) vanishes.*

Remark 4.3. *Assumption 4.1(v) is necessary in our setup, however, the result could still hold when a_{21} has determinant zero, a scalar example is discussed by Papavasiliou in [20] for diffusion estimates.*

Remark 4.4. *As in Remark 3.4 for the case of averaging, q_1 and q_2 in Assumption 4.1(ii) can also be relaxed to positive definite matrices to guarantee same result.*

In what follows, we will refer to the following equation as the **homogenized equation** for system (4.1),

$$\frac{dX}{dt} = \tilde{a}X + \sqrt{\tilde{q}}\frac{dW}{dt} , \tag{4.2}$$

where

$$\tilde{a} = a_{13} - a_{14}a_{22}^{-1}a_{21} , \tag{4.3}$$

and

$$\tilde{q} = q_1 + a_{12}a_{22}^{-1}q_2{a_{22}^{-1}}^*a_{12}^* . \tag{4.4}$$

In the rest of this chapter,

- we take observations from the multiscale system (4.1a);
- we first show that the discrepancy between the trajectories from the slow part x of the multiscale system (4.1a) and the homogenized equation (4.2) is of order $\mathcal{O}(\epsilon\sqrt{\log(\epsilon)})$ in the L^2 sense, in Section 4.2;
- we then show that using observations from the multiscale system (4.1a) and applying them to the drift estimator $\hat{a}_T$, defined in (2.5), we can correctly estimate the drift $\tilde{a}$

of the homogenized equation (4.2) in Section 4.3 by subsampling the observations at proper rates;

- we also show that using observations from the multiscale system (4.1a) and applying them to the diffusion estimator $\hat{\sigma}_\delta$, defined in (2.6), we can correctly estimate the diffusion parameter $\tilde{q}$ of the homogenized equation (4.2) in Section 4.4 by subsampling the observations at proper rates;

- finally a numerical example is studied to illustrate our findings in Section 4.5.

The convergence of the homogenized system is different from that of the averaging systems. For each given time series of observations, the paths of the slow process converge to the paths of the corresponding homogenized equation. However, we will see that in the limit $\epsilon \to 0$, the likelihood of the drift or diffusion parameter is different depending on whether we observe a path of the slow process generated by (4.1a) or the homogenized process (4.2) (see also [73, 78, 79]).

4.2 The Paths

The following theorem extends Theorem 18.1 in [78], which gives weak convergence of paths on $\mathbb{T}^d$. By limiting ourselves to the OU process, we extend the domain to $\mathbb{R}^d$ and prove a stronger mode of convergence. We prove the following lemma first.

Lemma 4.5. *Suppose that (x, y) solves (4.1a) and Assumptions 4.1 are satisfied. Then, for fixed finite $T > 0$ and small ϵ,*

$$\mathbb{E} \sup_{0 \le t \le T} \left(\|x(t)\|^2 + \|y(t)\|^2 \right) = \mathcal{O}\left(\log(1 + \frac{T}{\epsilon^2}) \right) \tag{4.5}$$

where $\|\cdot\|$ is the vector norm, and the order is in terms of ϵ.

Proof. We look at the system of SDEs as,

$$d\mathbf{x}_t = \mathbf{a}\mathbf{x}_t dt + \sqrt{\mathbf{q}} dW_t \tag{4.6}$$

where,

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}, \mathbf{a} = \begin{pmatrix} \frac{1}{\epsilon}a_{11} + a_{13} & \frac{1}{\epsilon}a_{12} + a_{14} \\ \frac{1}{\epsilon^2}a_{21} & \frac{1}{\epsilon^2}a_{22} \end{pmatrix} \text{ and } \mathbf{q} = \begin{pmatrix} q_1 & 0 \\ 0 & \frac{1}{\epsilon^2}q_2 \end{pmatrix}.$$

We try to characterize the magnitude of the eigenvalues of $\mathbf{a}$. To find the eigenvalues, we require

$$\det(\mathbf{a} - \lambda I) = 0 \, .$$

We either have the characteristic polynomial

$$\det\left(\frac{1}{\epsilon^2}a_{22} - \lambda I\right) = 0\,,$$

or

$$\det\left((\frac{1}{\epsilon}a_{11} + a_{13} - \lambda I) - (\frac{1}{\epsilon}a_{12} + a_{14})(\frac{1}{\epsilon^2}a_{22} - \lambda I)^{-1}\frac{1}{\epsilon^2}a_{21}\right) = 0\,.$$

First, we set the first determinant equal to zero:

$$\det\left(\frac{1}{\epsilon^2}a_{22} - \lambda I\right) = \frac{1}{\epsilon^{2d_2}}\det\left(a_{22} - \epsilon^2\lambda I\right) = 0\,.$$

By definition, $\epsilon^2\lambda$ are the eigenvalues of a_{22}, thus they are of order $\epsilon^2\lambda = \mathcal{O}(1)$. Consequently, we have d_2 (not necessarily distinct) real eigenvalues of order $\mathcal{O}(\frac{1}{\epsilon^2})$.

If the determinant of the second matrix is zero, we have

$$\det\left((\frac{1}{\epsilon}a_{11} + a_{13} - \lambda I) - (\frac{1}{\epsilon}a_{12} + a_{14})(\frac{1}{\epsilon^2}a_{22} - \lambda I)^{-1}\frac{1}{\epsilon^2}a_{21}\right) = 0\,. \tag{4.7}$$

Rearranging the matrix we have,

$$\det\left((\frac{1}{\epsilon}a_{11} + a_{13} - \lambda I) - (\frac{1}{\epsilon}a_{12} + a_{14})(a_{22} - \epsilon^2\lambda I)^{-1}a_{21}\right) = 0\,.$$

We apply Taylor expansion on $f(\epsilon^2) = (a_{22} - \epsilon^2\lambda I)^{-1}$ at $\epsilon = 0$. We have,

$$f(\epsilon^2) = a_{22}^{-1} + \epsilon^2\lambda a_{22}^{-2} + \epsilon^4\lambda^2 a_{22}^{-3} + \mathcal{O}(\epsilon^6) = a_{22}^{-1} + \mathcal{O}(\epsilon^2)\,.$$

We substitute the Taylor expansion into the determinant,

$$\begin{aligned}
& \det\left((\frac{1}{\epsilon}a_{11} + a_{13} - \lambda I) - (\frac{1}{\epsilon}a_{12} + a_{14})(a_{22} - \epsilon^2\lambda I)^{-1}a_{21}\right) \\
= \;& \det\left((\frac{1}{\epsilon}a_{11} + a_{13} - \lambda I) - (\frac{1}{\epsilon}a_{12} + a_{14})(a_{22}^{-1} + \mathcal{O}(\epsilon^2))a_{21}\right) \\
= \;& \det\left(\frac{1}{\epsilon}(a_{11} - a_{12}a_{22}^{-1}a_{21}) + (\tilde{a} - \lambda I) + \mathcal{O}(\epsilon)\right) \\
= \;& \det\left((\tilde{a} - \lambda I) + \mathcal{O}(\epsilon)\right) \\
= \;& 0
\end{aligned}$$

It is equivalent to finding the eigenvalues of a perturbed matrix of $\tilde{a}$. By Theorem 2 on page 137 in [42], on the eigenvalues of a perturbed matrix, we have that the corresponding d_1 (not necessarily distinct) real eigenvalues of order $\mathcal{O}(1)$. Therefore, we can decompose $\mathbf{a}$

as

$$\mathbf{a} = PDP^{-1} \text{ with } D = \begin{pmatrix} D_1 & 0 \\ 0 & \frac{1}{\epsilon^2} D_2 \end{pmatrix}$$

where D is the diagonal matrix, for which $D_1 \in \mathbb{R}^{d_1 \times d_1}$ and $D_2 \in \mathbb{R}^{d_2 \times d_2}$ are diagonal blocks of eigenvalues of order $\mathcal{O}(1)$. We also define $\Sigma = P^{-1}\mathbf{q}(P^{-1})^*$. Using Lemma 9.14 in Appendix 9.11, we have the ratio between diagonal elements of D and Σ is always

$$D_{ii}/\Sigma_{ii} = \mathcal{O}(1) \ .$$

We apply Theorem 2.3 to the system of equations (4.6). We have

$$\mathbb{E}\left(\sup_{0\le t\le T} \|\mathbf{x}(t)\|^2\right) \le C \frac{\log\left(1 + \max_i(|D_{ii}|)T\right)}{\min_i(|D_{ii}/\Sigma_{ii}|)} \ .$$

Since $D_{ii}/\Sigma_{ii} = \mathcal{O}(1)$, $\max_i |D_{ii}| = \mathcal{O}(\frac{1}{\epsilon^2})$, we have

$$\mathbb{E}\left(\sup_{0\le t\le T} \|\mathbf{x}(t)\|^2\right) = \mathcal{O}\left(\log(1 + \frac{T}{\epsilon^2})\right) \ .$$

Since $\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix}$, we get

$$\mathbb{E}\left(\sup_{0\le t\le T} (\|x(t)\|^2 + \|y(t)\|^2)\right) = \mathcal{O}\left(\log\left(1 + \frac{T}{\epsilon^2}\right)\right) \ .$$

This completes the proof. □

Theorem 4.6. *Let Assumptions 4.1 hold for system* (4.1)*. Suppose that x and X are solutions of (4.1a) and* (4.2) *respectively. (x, y) corresponds to the realization (U, V) of Brownian motion, while X corresponds to the realization*

$$W_\cdot = \tilde{q}^{-\frac{1}{2}} \left(\sqrt{q_1} U_\cdot - a_{12} a_{22}^{-1} \sqrt{q_2} V_\cdot\right) \tag{4.8}$$

and $x(0) = X(0)$. Then x converges to X in L^2. More specifically,

$$\mathbb{E} \sup_{0\le t\le T} \|x(t) - X(t)\|^2 \le c\left(\epsilon^2 \log(\frac{T}{\epsilon}) + \epsilon^2 T\right) e^T \ ,$$

when T is fixed finite, the above bound can be simplified to

$$\mathbb{E} \sup_{0\le t\le T} \|x(t) - X(t)\|^2 = \mathcal{O}(\epsilon^2 \log(\epsilon)).$$

Proof. We rewrite (4.1b) as

$$(a_{22}^{-1}a_{21}x(t) + y(t))dt = \epsilon^2 a_{22}^{-1} dy(t) - \epsilon a_{22}^{-1}\sqrt{q_2}dV_t \ . \tag{4.9}$$

We also rewrite (4.1a) as

$$\begin{aligned} dx(t) &= \frac{1}{\epsilon}a_{12}(a_{22}^{-1}a_{21}x(t) + y(t))dt + a_{14}(a_{22}^{-1}a_{21}x(t) + y(t))dt \\ &\quad +(a_{13} - a_{14}a_{22}^{-1}a_{21})x(t)dt + \sqrt{q_1}dU_t \\ &= \left(\frac{1}{\epsilon}a_{12} + a_{14}\right)(a_{22}^{-1}a_{21}x(t) + y(t))dt \\ &\quad +\tilde{a}x(t)dt + \sqrt{q_1}dU_t \ . \end{aligned} \tag{4.10}$$

Replacing $(a_{22}^{-1}a_{21}x(t) + y(t))dt$ in (4.10) by the right-hand-side of (4.9), we get

$$\begin{aligned} dx(t) &= \epsilon(a_{12} + \epsilon a_{14})a_{22}^{-1}dy(t) - a_{12}a_{22}^{-1}\sqrt{q_2}dV_t - \epsilon a_{14}a_{22}^{-1}\sqrt{q_2}dV_t \\ &\quad +\tilde{a}x(t)dt + \sqrt{q_1}dU_t \\ &= \tilde{a}x(t)dt + \epsilon(a_{12} + \epsilon a_{14})a_{22}^{-1}dy(t) \\ &\quad +\sqrt{\tilde{q}}dW_t - \epsilon a_{14}a_{22}^{-1}\sqrt{q_2}dV_t \ . \end{aligned} \tag{4.11}$$

Thus

$$\begin{aligned} x(t) &= x(0) + \int_0^t \tilde{a}x(s)ds + \sqrt{\tilde{q}}W_t \\ &\quad +\epsilon(a_{12} + \epsilon a_{14})a_{22}^{-1}(y(t) - y(0)) - \epsilon a_{14}a_{22}^{-1}\sqrt{q_2}V_t \ . \end{aligned} \tag{4.12}$$

Recall that the homogenized equation (4.2) is

$$X(t) = X(0) + \int_0^t \tilde{a}X(s)ds + \sqrt{\tilde{q}}W_t \ . \tag{4.13}$$

Let $e(t) = x(t) - X(t)$. Subtracting the previous equation from (4.12) and using the assumption $X(0) = x(0)$, we find that

$$\begin{aligned} e(t) &= \tilde{a}\int_0^t e(s)ds \\ &\quad +\epsilon\left((a_{12} + \epsilon a_{14})a_{22}^{-1}(y(t) - y(0)) - a_{14}a_{22}^{-1}\sqrt{q_2}V_t\right) \ . \end{aligned} \tag{4.14}$$

Applying Lemma 4.5, we find an ϵ-independent constant C, such that

$$\mathbb{E}\left(\sup_{0\le t\le T}\|y(t)\|^2\right) \le C\log(\frac{T}{\epsilon})\ .$$

By Cauchy Schwarz,

$$\mathbb{E}\left(\sup_{0\le t\le T}\|e(t)\|^2\right) \le c\left(\int_0^T \mathbb{E}\|e(s)\|^2 ds + \epsilon^2\log(\frac{T}{\epsilon}) + \epsilon^2 T\right)\ . \tag{4.15}$$

By the integrated version of the Gronwall inequality, we deduce that

$$\mathbb{E}\left(\sup_{0\le t\le T}\|e(t)\|^2\right) \le c\left(\epsilon^2\log(\frac{T}{\epsilon}) + \epsilon^2 T\right)e^T\ . \tag{4.16}$$

When T is finite, we have

$$\mathbb{E}\left(\sup_{0\le t\le T}\|e(t)\|^2\right) = \mathcal{O}\left(\epsilon^2\log(\epsilon)\right)\ .$$

This completes the proof. □

4.3 The Drift Estimator

As in the averaging case, a natural idea for estimating the drift of the homogenized equation is to use the maximum likelihood estimator (2.5), replacing X by the solution x of (4.1a). However, in the case of homogenization we do not get asymptotically consistent estimates if we do not subsample our observations [73, 78, 79]. To achieve a correct estimate, we must subsample the data: we choose Δ, our time interval for observation, according to the value of the scale parameter ϵ and solve the estimation problem for discretely observed diffusion processes, see [73, 78, 79].

The maximum likelihood estimator for the drift of a homogenized equation converges after proper subsampling. We let the observation time interval Δ and the number of observations N both depend on the scaling parameter ϵ, by setting $\Delta = \epsilon^\alpha$ and $N = \epsilon^{-\gamma}$. We find the error is optimized in the L^2 sense when $\alpha = 1/2$. We will show that $\hat{a}_{N,\epsilon}$ converges to $\tilde{a}$ only if $\frac{\Delta}{\epsilon^2} \to \infty$, in a sense to be made precise later.

Theorem 4.7. *Suppose that x is the projection to the x-coordinate of a solution of system* (4.1) *satisfying Assumptions 4.1. Let $\hat{a}_{N,\epsilon}$ be the estimate we get by replacing X in* (2.5) *by*

x, *i.e.*

$$\hat{a}_{N,\epsilon} = \left(\frac{1}{N\Delta} \sum_{n=0}^{N-1} (x_{n+1} - x_n) \otimes x_n \right) \left(\frac{1}{N\Delta} \sum_{n=0}^{N-1} x_n \otimes x_n \Delta \right)^{-1} \tag{4.17}$$

Then,

$$\mathbb{E}\|\hat{a}_{N,\epsilon} - \tilde{a}\|^2 = \mathcal{O}(\Delta + \frac{1}{N\Delta} + \frac{\epsilon^2}{\Delta^2})$$

where $\tilde{a}$ *as defined in* (4.3)*. Consequently, if* $\Delta = \epsilon^\alpha$, $N = \epsilon^{-\gamma}$, $\alpha \in (0,1)$, $\gamma > \alpha$,

$$\lim_{\epsilon \to 0} \mathbb{E}\|\hat{a}_{N,\epsilon} - \tilde{a}\|^2 = 0 \, .$$

Furthermore, $\alpha = 1/2$ *and* $\gamma \geq 3/2$ *optimize the error.*

Before proving Theorem 4.7, we first find the magnitude of the increment of y over a small time interval Δ. Solving equation (4.1b), we have

$$\begin{aligned} y_{n+1} - y_n &= (e^{a_{22}\frac{\Delta}{\epsilon^2}} - I) y_n \\ &+ \frac{1}{\epsilon^2} \int_{n\Delta}^{(n+1)\Delta} e^{a_{22}\frac{(n+1)\Delta - s}{\epsilon^2}} x(s) ds \\ &+ \frac{1}{\epsilon} \int_{n\Delta}^{(n+1)\Delta} e^{a_{22}\frac{(n+1)\Delta - s}{\epsilon^2}} \sqrt{q_2} dV_s \, . \end{aligned} \tag{4.18}$$

By triangle inequality, we have

$$\begin{aligned} \mathbb{E}\|y_{n+1} - y_n\|^2 &\leq \|e^{a_{22}\frac{\Delta}{\epsilon^2}} - I\|^2 \mathbb{E}\|y_n\|^2 \\ &+ c\|e^{a_{22}\frac{\Delta}{\epsilon^2}} - I\|^2 \\ &+ \frac{1}{2}\|e^{2a_{22}\frac{\Delta}{\epsilon^2}}\|^2 \|q_2\|^2 \, . \end{aligned}$$

Since a_{22} is negative definite, thus,

$$\mathbb{E}\|y_{n+1} - y_n\|^2 = \mathcal{O}(e^{-\frac{\Delta}{\epsilon^2}} - 1) \, .$$

By definition $\Delta = \epsilon^\alpha$, and the property that $(e^{-\frac{\Delta}{\epsilon^2}} - 1) = \mathcal{O}(\frac{\Delta}{\epsilon^2})$ if $\frac{\Delta}{\epsilon^2}$ is small, the above equation can be rewritten as

$$\mathbb{E}\|y_{n+1} - y_n\|^2 = \mathcal{O}(\epsilon^{\max(\alpha-2,0)}) \, . \tag{4.19}$$

Proof. Define I_1 and I_2 as

$$I_1 = \frac{1}{N\Delta}\sum_{n=0}^{N-1}(x_{n+1} - x_n) \otimes x_n \ , \ I_2 = \frac{1}{N\Delta}\sum_{n=0}^{N-1} x_n \otimes x_n \Delta$$

By ergodic theorem, and since $N = \epsilon^{-\gamma}$, we have

$$\lim_{\epsilon \to 0} I_2 = \mathbb{E}\left(x_n \otimes x_n\right) = C \neq 0$$

which is a constant invertible matrix. Hence instead of proving

$$\mathbb{E}\|\hat{a}_{N,\epsilon} - \tilde{a}\|^2 = \mathcal{O}(\Delta^2 + \frac{1}{N\Delta} + \frac{\epsilon^2}{\Delta^2}) \ ,$$

we prove,

$$\mathbb{E}\|I_1 - \tilde{a} I_2\|^2 = \mathcal{O}(\Delta^2 + \frac{1}{N\Delta} + \frac{\epsilon^2}{\Delta^2}) \ .$$

We use the rearranged equation (4.11) of (4.1a) to decompose the error,

$$I_1 - \tilde{a} I_2 = J_1 + J_2 + J_3 + J_4 \ . \tag{4.20}$$

where

$$\begin{aligned}
J_1 &= \frac{1}{N\Delta}\sum_{n=0}^{N-1}\left(\tilde{a}\int_{n\Delta}^{(n+1)\Delta} x(s)ds - x_n\right) \otimes x_n \\
J_2 &= \frac{1}{N\Delta}\sum_{n=0}^{N-1}\left(\sqrt{\tilde{q}}\int_{n\Delta}^{(n+1)\Delta} dW_s\right) \otimes x_n \\
J_3 &= \frac{\epsilon}{N\Delta}\sum_{n=0}^{N-1}(a_{12} + \epsilon a_{14})a_{22}^{-1}\int_{n\Delta}^{(n+1)\Delta} dy(s) \otimes x_n \\
J_4 &= \frac{\epsilon}{N\Delta}\sum_{n=0}^{N-1} a_{14}a_{22}^{-1}\sqrt{q_2}\int_{n\Delta}^{(n+1)\Delta} dV_s \otimes x_n
\end{aligned}$$

By independence, Itô isometry and ergodicity, we immediately have

$$\begin{aligned}
\mathbb{E}\|J_2\|^2 &= \mathbb{E}\|\frac{\sqrt{\tilde{q}}}{N\Delta}\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta} dW_s \otimes x_n\|^2 \\
&= \frac{\tilde{q}}{N^2\Delta^2}\mathbb{E}\|\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta} dW_s \otimes x_n\|^2 \\
&\leq \frac{\tilde{q}}{N^2\Delta^2}N\mathbb{E}\|\int_{n\Delta}^{(n+1)\Delta} dW_s\|^2\mathbb{E}\|x_n\|^2 \\
&\leq \frac{\tilde{q}}{N^2\Delta^2}N\Delta\mathbb{E}\|x_n\|^2 \\
&= \mathcal{O}(\frac{1}{N\Delta}) ,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\|J_4\|^2 &= \mathbb{E}\|\frac{\epsilon C}{N\Delta}\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta} dV_s \otimes x_n\|^2 \\
&= \frac{\epsilon^2 C}{N^2\Delta^2}\mathbb{E}\|\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta} dV_s \otimes x_n\|^2 \\
&\leq \frac{\epsilon^2 C}{N^2\Delta^2}N\mathbb{E}\left(\|\int_{n\Delta}^{(n+1)\Delta} dV_s\|^2\right)\mathbb{E}(\|x_n\|^2) \\
&\leq \frac{\epsilon^2 C}{N^2\Delta^2}N\Delta\mathbb{E}(\|x_n\|^2) \\
&= \mathcal{O}(\frac{\epsilon^2}{N\Delta}) .
\end{aligned}$$

By Hölder inequality, and (4.19), we have,

$$\begin{aligned}
\mathbb{E}\|J_3\|^2 &= \mathbb{E}\|\frac{\epsilon C}{N\Delta}\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta} dy \otimes x_n\|^2 \\
&= \mathbb{E}\|\frac{\epsilon C}{N\Delta}\sum_{n=0}^{N-1}(y_{n+1}-y_n) \otimes x_n\|^2 \\
&\leq \frac{\epsilon^2}{N^2\Delta^2}\mathbb{E}\|\sum_{n=0}^{N-1}(y_{n+1}-y_n)\|^2\mathbb{E}\|\sum_{n=0}^{N-1}x_n\|^2 \\
&\leq \frac{\epsilon^2 C}{N^2\Delta^2}N(\epsilon^{\max(\alpha-2,0)})N\mathbb{E}\|x_n\|^2 \\
&= \mathcal{O}(\frac{\epsilon^2}{\Delta^2}) .
\end{aligned}$$

Finally, we find the squared error for J_1. We use the integrated form of equation (4.11) on time interval $[n\Delta, s]$ to replace $x(s)$

$$\begin{aligned}
\mathbb{E}\|J_1\|^2 &= \frac{\tilde{a}^2}{N^2\Delta^2}\mathbb{E}\|\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta}(x(s)-x_n)ds\otimes x_n\|^2 && (4.21)\\
&= \frac{\tilde{a}^2}{N^2\Delta^2}\mathbb{E}\|\sum_{n=0}^{N-1}(K_1^{(n)}+K_2^{(n)}+K_3^{(n)}+K_4^{(n)})\|^2 && (4.22)\\
& && (4.23)
\end{aligned}$$

where,

$$\begin{aligned}
K_1^{(n)} &= \tilde{a}\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s}x(u)duds\otimes x_n\ ,\\
K_2^{(n)} &= \epsilon(a_{12}+\epsilon a_{14})a_{22}^{-1}\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s}dy(u)ds\otimes x_n\ ,\\
K_3^{(n)} &= \sqrt{\tilde{q}}\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s}dW_u ds\otimes x_n\ ,\\
K_4^{(n)} &= \epsilon a_{14}a_{22}^{-1}\sqrt{q_2}\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s}dV_u ds\otimes x_n\ .
\end{aligned}$$

We immediately see that

$$\begin{aligned}
\mathbb{E}\|J_1\|^2 &= \frac{\tilde{a}^2}{N^2\Delta^2}\mathbb{E}\sum_{n=0}^{N-1}\|\sum_{i=1}^{4}K_i^{(n)}\|^2 && (4.24)\\
&+ \frac{\tilde{a}^2}{N^2\Delta^2}\mathbb{E}\sum_{m\neq n}\|\left(\sum_{i=1}^{4}K_i^{(n)}\right)\left(\sum_{j=1}^{4}K_j^{(m)}\right)\| && (4.25)
\end{aligned}$$

Remark 4.8. *We use the exact decomposition of $\mathbb{E}\|J_1\|^2$ by using* (4.24) *and* (4.25)*. This is essential in order to obtain more optimized subsampling rate for the drift estimator. For general L^p bound for the error, we can apply Hölder's inequality to decompose J_1 as,*

$$\begin{aligned}
\mathbb{E}\|J_1\|^p &= \frac{C}{N^p\Delta^p}\mathbb{E}\|\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta}(x(s)-x_n)ds\otimes x_n\|^p\\
&\leq \frac{C}{N^{p-1}\Delta^p}\mathbb{E}\|\sum_{n=0}^{N-1}\int_{n\Delta}^{(n+1)\Delta}(x(s)-x_n)ds\|^p\mathbb{E}\|x_n\|^p
\end{aligned}$$

which is used in [79]. Using this inequality will give an optimal subsampling rate of $\alpha = 2/3$, and achieves an over all L^1 error of order $\mathcal{O}(\epsilon^{1/3})$. However, this magnitude of overall

error is not optimal in L^2. We will show later that the optimal L^2 error can be achieved at the order of $\mathcal{O}(\epsilon^{1/2})$, using the exact decomposition shown above.

By Cauchy Schwarz inequality, we know for line (4.24),

$$\mathbb{E}\sum_{n=0}^{N-1}\|\sum_{i=1}^{4}K_i^{(n)}\|^2 \leq \sum_{n=0}^{N-1}\sum_{i=1}^{4}\mathbb{E}\|K_i^{(n)}\|^2 .$$

Using first order iterated integrals, we have

$$\begin{aligned}
\mathbb{E}\|K_1^{(n)}\|^2 &= \mathbb{E}\|\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s} x(u)duds \otimes x_n\|^2 \\
&\leq C\Delta\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s}\|x(u)\|^2 duds\|x_n\|^2 \\
&\leq C\Delta\int_{n\Delta}^{(n+1)\Delta}(s-n\Delta)^2 ds \\
&= \mathcal{O}(\Delta^4) .
\end{aligned}$$

Using (4.19), we have

$$\begin{aligned}
\mathbb{E}\|K_2^{(n)}ds\|^2 &= \mathbb{E}\|\epsilon C\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s} dy(u)ds \otimes x_n\|^2 \\
&\leq C\epsilon^2\mathbb{E}\|\int_{n\Delta}^{(n+1)\Delta}(y(s)-y(u))ds \otimes x_n\|^2 \\
&\leq C\epsilon^2\Delta\mathbb{E}\int_{n\Delta}^{(n+1)\Delta}\|y(s)-y(u)\|^2 ds\|x_n\|^2 \\
&\leq C\epsilon^2\Delta\mathbb{E}\int_{n\Delta}^{(n+1)\Delta}(e^{-\frac{s-n\Delta}{\epsilon^2}}-1)ds \\
&= \mathcal{O}\left(\epsilon^4(e^{-\frac{\Delta}{\epsilon^2}}-1)\right) .
\end{aligned}$$

For $K_3^{(n)}$, we have,

$$\begin{aligned}
\mathbb{E}\|K_3^{(n)}\|^2 &= \mathbb{E}\|\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s}\sqrt{\tilde{q}}dW_u ds \otimes x_n\|^2 \\
&\leq C\Delta\int_{n\Delta}^{(n+1)\Delta}\|\int_{n\Delta}^{s} dW_u\|^2 ds \\
&\leq C\Delta\int_{n\Delta}^{(n+1)\Delta}(s-n\Delta)ds \\
&= \mathcal{O}(\Delta^3) .
\end{aligned}$$

Since $K_4^{(n)}$ is similar to $K_3^{(n)}$, we have

$$\mathbb{E}\|K_4^{(n)}\|^2 = \mathcal{O}(\epsilon^2\Delta^3) \; .$$

Thus, for line (4.24), the order of the dominating terms are,

$$\mathbb{E}\sum_{n=0}^{N-1}\|\sum_{i=1}^{4} K_i^{(n)}\|^2 = \mathcal{O}(N\Delta^4 + N\epsilon^4(e^{-\frac{\Delta}{\epsilon^2}} - 1) + N\Delta^3) \; .$$

For line (4.25),

$$\mathbb{E}\sum_{m\neq n}\|(\sum_{i=1}^{4} K_i^{(n)})(\sum_{j=1}^{4} K_j^{(m)})\| \leq \sum_{m\neq n}\mathbb{E}\|\sum_{i=1}^{4} K_i^{(n)}\|\mathbb{E}\|\sum_{j=1}^{4} K_j^{(m)}\| \; .$$

We know,

$$\begin{aligned}
\mathbb{E}\|K_1^{(n)}\| &= \mathbb{E}\|C\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s} x(u)duds\| \\
&\leq C\mathbb{E}\left(\int_{n\Delta}^{(n+1)\Delta}(s-n\Delta)ds\right) \\
&= \mathcal{O}(\Delta^2) \; .
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\mathbb{E}\|K_2^{(n)}\| &= \epsilon C\mathbb{E}\left(\int_{n\Delta}^{(n+1)\Delta}(y(s)-y_n)ds\right) \\
&= \mathcal{O}(\epsilon\Delta) \; .
\end{aligned}$$

Since the integral of Brownian motions is Gaussian

$$\begin{aligned}
\mathbb{E}\|K_3^{(n)}\| &= C\mathbb{E}(\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s} dW_u ds) \\
&= C\mathbb{E}(\int_{n\Delta}^{(n+1)\Delta}(W(s)-W(n\Delta))ds) \\
&= C\mathbb{E}(\int_{n\Delta}^{(n+1)\Delta}W(s)ds - W(n\Delta)\Delta) \\
&= 0 \; .
\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}\|K_4^{(n)}\| &= C\epsilon\mathbb{E}(\int_{n\Delta}^{(n+1)\Delta}\int_{n\Delta}^{s} dV_u ds)\\ &= C\epsilon\mathbb{E}(\int_{n\Delta}^{(n+1)\Delta} V(s)ds - V(n\Delta)\Delta)\\ &= 0\,.\end{aligned}$$

Thus,

$$\mathbb{E}\|\sum_{i=1}^{4} K_i^{(n)}\| = \mathcal{O}(\Delta^2 + \epsilon\Delta)\,,$$

immediately we have for line (4.25),

$$\mathbb{E}\sum_{m\neq n}\|(\sum_{i=1}^{4} K_i^{(n)})(\sum_{j=1}^{4} K_j^{(m)})\| = \mathcal{O}(N^2\Delta^4 + N^2\epsilon^2\Delta^2)\,.$$

Putting all terms for J_1 together, we keep the dominating terms, and by assumption $N\Delta \to \infty$, and $\alpha < 2$ since $e^{-\frac{\Delta}{\epsilon^2}} \to 0$,

$$\begin{aligned}\mathbb{E}\|J_1\|^2 &\leq \frac{C}{N^2\Delta^2}(N\Delta^4 + N\epsilon^4(e^{-\frac{\Delta}{\epsilon^2}} - 1) + N\Delta^3)\\ &+ \frac{C}{N^2\Delta^2}(N^2\Delta^4 + N^2\epsilon^2\Delta^2)\\ &= \mathcal{O}(\frac{\Delta^2}{N} + \frac{\epsilon^4}{N\Delta^2}(e^{-\frac{\Delta}{\epsilon^2}} - 1) + \frac{\Delta}{N} + \Delta^2 + \epsilon^2)\\ &= \mathcal{O}(\frac{\epsilon^4}{N\Delta^2} + \Delta^2 + \epsilon^2)\,.\end{aligned}$$

Therefore, putting J_i's, $i \in \{1,2,3,4\}$, together, we have,

$$\begin{aligned}\mathbb{E}\|I_1 - \tilde{a}I_2\|^2 &\leq \sum_{i=1}^{4}\mathbb{E}\|J_i\|^2\\ &= \mathcal{O}(\frac{\epsilon^4}{N\Delta^2} + \Delta^2 + \epsilon^2)\\ &+ \mathcal{O}(\frac{1}{N\Delta})\\ &+ \mathcal{O}(\frac{\epsilon^2}{\Delta^2})\\ &+ \mathcal{O}(\frac{\epsilon^2}{N\Delta})\\ &= \mathcal{O}(\Delta^2 + \frac{1}{N\Delta} + \frac{\epsilon^2}{\Delta^2})\end{aligned}$$

We rewrite the above equation using $\Delta = \epsilon^\alpha$ and $N = \epsilon^{-\gamma}$,

$$\mathbb{E}\|I_1 - \tilde{a}I_2\|^2 = \mathcal{O}(\epsilon^{2\alpha} + \epsilon^{\gamma-\alpha} + \epsilon^{2-2\alpha}) \, .$$

It is immediately seen that $\alpha = \frac{1}{2}$ and $\gamma \geq 3/2$ optimize the error, and $\alpha \in (0, 1)$, the order of the error is

$$\mathbb{E}\|I_1 - \tilde{a}I_2\|^2 = \mathcal{O}(\epsilon) \, .$$

This completes the proof. □

4.4 The Diffusion Estimator

Just as in the case of the drift estimator, we define the diffusion estimator by the maximum likelihood estimator (2.6), where X is replaced by the discretized solution of (4.1a). More specifically, we define

$$\hat{q}_\epsilon = \frac{1}{N\Delta} \sum_{n=0}^{N-1} (x_{n+1} - x_n) \otimes (x_{n+1} - x_n) \tag{4.26}$$

where $x_n = x(n\Delta)$ is the discrete observation of the process generated by (4.1a) and Δ is the observation time interval.

Theorem 4.9. *Suppose that x is the projection to the x-coordinate of a solution of system* (4.1) *satisfying Assumptions 4.1. Let $\hat{q}_\epsilon$ be the estimate we get by replacing X in* (2.6) *by x, i.e.*

$$\hat{q}_\epsilon = \frac{1}{T} \sum_{n=0}^{N-1} (x_{n+1} - x_n) \otimes (x_{n+1} - x_n) \, .$$

Then

$$\mathbb{E}\|\hat{q}_\epsilon - \tilde{q}\|^2 = \mathcal{O}\left(\Delta + \epsilon^2 + \frac{\epsilon^4}{\Delta^2}\right)$$

where $\tilde{q}$ as defined in (4.4)*. Consequently, if $\Delta = \epsilon^\alpha$, fix $T = N\Delta$, and $\alpha \in (0, 2)$, then*

$$\lim_{\epsilon \to 0} \mathbb{E}\|\hat{q}_\epsilon - \tilde{q}\|^2 = 0 \, .$$

Furthermore, $\alpha = 4/3$ optimizes the error.

We first define

$$\sqrt{\Delta}\eta_n = \int_{n\Delta}^{(n+1)\Delta} dW_t \, .$$

Proof. We now prove Theorem 4.9. Using the integral form of equation (4.11),

$$\begin{aligned} x_{n+1} - x_n &= \int_{n\Delta}^{(n+1)\Delta} \sqrt{\tilde{q}} dW_s \\ &+ \hat{R}_1 + \hat{R}_2 + \hat{R}_3 \end{aligned} \tag{4.27}$$

where

$$\begin{aligned} \hat{R}_1 &= \tilde{a} \int_{n\Delta}^{(n+1)\Delta} x(s) ds \\ \hat{R}_2 &= \epsilon a_{14} a_{22}^{-1} \sqrt{q_2} \int_{n\Delta}^{(n+1)\Delta} dV_s \\ \hat{R}_3 &= \epsilon (a_{12} + \epsilon a_{14}) a_{22}^{-1} \int_{n\Delta}^{(n+1)\Delta} dy(s) \end{aligned}$$

We rewrite line (4.27) as

$$\int_{n\Delta}^{(n+1)\Delta} \sqrt{\tilde{q}} dW_s = \sqrt{\tilde{q}\Delta}\eta_n$$

where η_n are $\mathcal{N}(0, I)$ random variables.

For Δ and ϵ sufficiently small, by Cauchy-Schwarz inequality

$$\begin{aligned} \mathbb{E}\| c \int_{n\Delta}^{(n+1)\Delta} x(s) ds \|^2 &\leq c\mathbb{E} \int_{n\Delta}^{(n+1)\Delta} \|x(s)\|^2 ds \int_{n\Delta}^{(n+1)\Delta} ds \\ &\leq c\Delta\mathbb{E} \int_{n\Delta}^{(n+1)\Delta} \|x(s)\|^2 ds \\ &\leq c\Delta^2 \mathbb{E} \left(\sup_{n\Delta \leq s \leq (n+1)\Delta} \|x(s)\|^2 \right) \\ &= \mathcal{O}(\Delta^2) \end{aligned}$$

Therefore,

$$\mathbb{E}\|\hat{R}_1\|^2 = \mathcal{O}(\Delta^2)$$

By Itô isometry

$$\mathbb{E}\|\hat{R}_2\|^2 = \mathcal{O}(\epsilon^2\Delta)$$

Then we look at $\hat{R}_3$,

$$\mathbb{E}\|\hat{R}_3\|^2 = \epsilon^2 C \mathbb{E}\|y_{n+1} - y_n\|^2$$

By (4.19), we have

$$\mathbb{E}\|\hat{R}_3\|^2 = \mathcal{O}(\epsilon^{\max(\alpha,2)}) \tag{4.28}$$

We substitute $(x_{n+1} - x_n)$ into the estimator $\hat{q}_\epsilon$ in Theorem 4.9. We decompose the estimator's error as follows,

$$\begin{aligned}
\hat{q}_\epsilon - \tilde{q} &= \tilde{q}(\frac{1}{N}\sum_{n=0}^{N-1}\eta_n \otimes \eta_n - I) \\
&+ \frac{1}{T}\sum_{n=0}^{N-1}\sum_{i=1}^{3}\left(\hat{R}_i \otimes \hat{R}_i\right) \\
&+ \frac{1}{T}\sum_{n=0}^{N-1}\sum_{i=1}^{3}\hat{R}_i \otimes \sqrt{\tilde{q}\Delta}\eta_n \\
&+ \frac{1}{T}\sum_{n=0}^{N-1}\sum_{i=1}^{3}\sqrt{\tilde{q}\Delta}\eta_n \otimes \hat{R}_i \\
&+ \frac{1}{T}\sum_{n=0}^{N-1}\left(\sum_{i\neq j}\hat{R}_i \otimes \hat{R}_j\right) \\
&= R
\end{aligned}$$

Then we bound the mean squared error using Cauchy Schwarz inequality.

$$\begin{aligned}
\mathbb{E}\left(\hat{q}_\epsilon - \tilde{q}\right)^2 &\leq C\tilde{q}^2\mathbb{E}(\frac{1}{N}\sum_{n=0}^{N-1}\eta_n^2 - I)^2 && (4.29)\\
&+ C\sum_{i=1}^{3}\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_i^2\right)^2 && (4.30)\\
&+ C\sum_{i=1}^{3}\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_i \otimes \sqrt{\tilde{q}\Delta}\eta_n\right)^2 && (4.31)\\
&+ C\sum_{i=1}^{3}\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\sqrt{\tilde{q}\Delta}\eta_n \otimes \hat{R}_i\right)^2 && (4.32)\\
&+ C\sum_{i\neq j}\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\left(\hat{R}_i \otimes \hat{R}_j\right)\right)^2 && (4.33)
\end{aligned}$$

Notice that we use the simplified $(\cdot)^2$ notation in this section, what we mean is actually square by tensor product. By law of large numbers, it is easy to see that line (4.29) is of order $\mathcal{O}(\Delta)$.

In line (4.30), for $i \in \{1,2\}$, we have

$$\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_i^2\right)^2 = \frac{1}{T^2}N\sum_{n=0}^{N-1}\mathbb{E}(\hat{R}_i^2)^2.$$

Since $\mathbb{E}\|\hat{R}_1\|^2 = \mathcal{O}(\Delta^2)$, we have

$$\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_1^2\right)^2 = \mathcal{O}\left(N^2(\Delta^2)^2\right) = \mathcal{O}\left(\Delta^2\right);$$

since $\mathbb{E}\|\hat{R}_2\|^2 = \mathcal{O}(\epsilon^2\Delta)$, we have

$$\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_2^2\right)^2 = \mathcal{O}\left(N^2(\Delta\epsilon^2)^2\right) = \mathcal{O}(\epsilon^4).$$

It is different for $\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_3^2\right)^2$, by (4.19), we have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_3^2\right)^2 &= \frac{C\epsilon^4}{T^2}\mathbb{E}\left(\sum_{n=0}^{N-1}\left(y_{n+1}-y_n\right)^2\right)^2 \\
&\leq C\epsilon^4 N\sum_{n=0}^{N-1}\mathbb{E}\left(y_{n+1}-y_n\right)^4 \\
&= \mathcal{O}\left(\frac{\epsilon^{4+2\max(0,\alpha-2)}}{\Delta^2}\right) \\
&= \mathcal{O}\left(\frac{\epsilon^{\max(4,2\alpha)}}{\Delta^2}\right)
\end{aligned}$$

Adding up all terms for line (4.30), we have,

$$\sum_{i=1}^{3}\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_i^2\right)^2 = \mathcal{O}\left(\Delta^2+\epsilon^4+\frac{\epsilon^{\max(4,2\alpha)}}{\Delta^2}\right). \tag{4.34}$$

In line (4.31), for $i \in \{1,2\}$, we have

$$\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_i\otimes\sqrt{\tilde{q}\Delta}\eta_n\right)^2 \leq CN^2\Delta\mathbb{E}\left(\hat{R}_i\otimes\eta_n\right)^2 = CN\mathbb{E}\|\hat{R}_i\|^2$$

Since $\mathbb{E}\|\hat{R}_1\|^2 = \mathcal{O}(\Delta^2)$, we have

$$\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_1 \otimes \sqrt{\tilde{q}\Delta}\eta_n\right)^2 = \mathcal{O}(N\Delta^2) = \mathcal{O}(\Delta);$$

since $\mathbb{E}\|\hat{R}_2\|^2 = \mathcal{O}(\epsilon^2\Delta)$, we have

$$\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_2 \otimes \sqrt{\tilde{q}\Delta}\eta_n\right)^2 = \mathcal{O}(N\epsilon^2\Delta) = \mathcal{O}(\epsilon^2)\,.$$

Again, it is different for $\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_3 \otimes \sqrt{\tilde{q}\Delta}\eta_n\right)^2$ due to correlation between $\hat{R}_3^{(n)}$ and η_n. Using the expression from (4.18) by only considering the dominating terms, we have

$$\begin{aligned}
& \mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_3 \otimes \sqrt{\tilde{q}\Delta}\eta_n\right)^2 \\
= \;& \mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_3^2\left(\sqrt{\tilde{q}\Delta}\eta_n\right)^2\right) \\
+ \;& \mathbb{E}\left(\frac{1}{T^2}\sum_{m\neq n}\hat{R}_3^{(m)}\hat{R}_3^{(n)}\int_{m\Delta}^{(m+1)\Delta}\sqrt{\tilde{q}}dWs\int_{n\Delta}^{(n+1)\Delta}\sqrt{\tilde{q}}dWs\right)
\end{aligned}$$

By computing the order of the dominating terms and the martingale terms, when $m = n$,

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_3^2\left(\sqrt{\tilde{q}\Delta}\eta_n\right)^2\right) &= \frac{1}{T}\sum_{n=0}^{N-1}\Delta\mathbb{E}\left(\hat{R}_3^2\tilde{q}\eta_n^2\right) \\
&= \frac{1}{T}\mathbb{E}(\hat{R}_3^2\eta_n^2) \\
&= \mathcal{O}\left(\epsilon^{\max(\alpha,2)}\right)
\end{aligned}$$

and when $m < n$,

$$\begin{aligned}
& \mathbb{E}\left(\frac{1}{T^2}\sum_{m\neq n}\hat{R}_3^{(m)}\hat{R}_3^{(n)}\int_{m\Delta}^{(m+1)\Delta}\sqrt{\tilde{q}}dWs\int_{n\Delta}^{(n+1)\Delta}\sqrt{\tilde{q}}dWs\right) \\
\leq\ & CN^2\epsilon^2\mathbb{E}\left((y_{n+1}-,y_n)(y_{m+1}-y_m)\right. \\
\times\ & \left.\int_{n\Delta}^{(n+1)\Delta}dW_s'\int_{m\Delta}^{(m+1)\Delta}dW_s\right) \\
\leq\ & CN^2\epsilon^2\mathbb{E}\left((y_{n+1}-y_n)\int_{n\Delta}^{(n+1)\Delta}dW_s'\right. \\
\times\ & \left.\mathbb{E}\left((y_{m+1}-y_m)\int_{m\Delta}^{(m+1)\Delta}dW_s|\mathcal{F}_{m\Delta}\right)\right)
\end{aligned}$$

Using the expansion in (4.18), and using the dominating terms only,

$$\begin{aligned}
& \mathbb{E}\left((y_{m+1}-y_m)\int_{n\Delta}^{(n+1)\Delta}dW_s|\mathcal{F}_{m\Delta}\right) \\
=\ & \mathbb{E}\left(\left((e^{-\frac{\Delta}{\epsilon^2}}-1)y_m\right.\right. \\
+\ & \frac{1}{\epsilon^2}\int_{m\Delta}^{(m+1)\Delta}e^{-\frac{(m+1)\Delta-s}{\epsilon^2}}x(s)ds \\
+\ & \left.\left.\frac{1}{\epsilon}\int_{m\Delta}^{(m+1)\Delta}e^{-\frac{(m+1)\Delta-s}{\epsilon^2}}dV_s\right)\int_{m\Delta}^{(m+1)\Delta}dW_s|\mathcal{F}_{m\Delta}\right) \\
=\ & \mathcal{O}(\epsilon(e^{-\frac{\Delta}{\epsilon^2}}-1))
\end{aligned}$$

Therefore, when $m < n$, we have,

$$\begin{aligned}
& \mathbb{E}\left(\frac{1}{T^2}\sum_{m\neq n}\hat{R}_3^{(m)}\hat{R}_3^{(n)}\int_{m\Delta}^{(m+1)\Delta}\sqrt{\tilde{q}}dWs\int_{n\Delta}^{(n+1)\Delta}\sqrt{\tilde{q}}dWs\right) \\
=\ & \mathcal{O}(\frac{\epsilon^4}{\Delta^2}(e^{-\frac{\Delta}{\epsilon^2}}-1)^2) \\
=\ & \mathcal{O}(\epsilon^{4-2\alpha+2\max(\alpha-2,0)}) \\
=\ & \mathcal{O}(\epsilon^{\max(0,4-2\alpha)})
\end{aligned}$$

In the case $m > n$, the result is identical due to symmetry. Adding up all terms for line

(4.31),

$$\sum_{i=1}^{5}\mathbb{E}\left(\frac{1}{T}\sum_{n=0}^{N-1}\hat{R}_i\otimes\sqrt{\tilde{q}\Delta}\eta_n\right)^2 = \mathcal{O}\left(\Delta+\epsilon^2+\epsilon^{\max(\alpha,2)}+\epsilon^{2\max(0,2-\alpha)}\right) \tag{4.35}$$

Line (4.32) is symmetric with line (4.31), which we can conclude it has the same order in (4.35).

In line (4.33), we have

$$\sum_{i\neq j}\mathbb{E}\left(\sum_{n=0}^{N-1}\hat{R}_i\otimes\hat{R}_j\right)^2 \leq N\mathbb{E}\|R_i\|^2\mathbb{E}\|R_j\|^2$$

Substituting in the L^2 norms of each $\hat{R}_i$, $i\in\{1,2,3\}$, we have for line (4.33),

$$\begin{aligned}&\sum_{i\neq j}\mathbb{E}\left(\sum_{n=0}^{N-1}\hat{R}_i\otimes\hat{R}_j\right)^2\\ =\;&\mathcal{O}\left(\Delta^2\epsilon^2+\Delta\epsilon^{\max(\alpha,2)}+\epsilon^{2+\max(\alpha,2)}\right)\end{aligned} \tag{4.36}$$

Aggregating bounds (4.34), (4.35) and (4.36) for equation lines from (4.29) to (4.33) respectively, we have

$$\begin{aligned}&\mathbb{E}\|\hat{q}_\epsilon-\tilde{q}\|^2\\ =\;&\mathcal{O}(\Delta)\\ +\;&\mathcal{O}\left(\Delta^2+\epsilon^4+\frac{\epsilon^{\max(4,2\alpha)}}{\Delta^2}\right)\\ +\;&\mathcal{O}\left(\Delta+\epsilon^2+\epsilon^{\max(\alpha,2)}+\epsilon^{2\max(0,2-\alpha)}\right)\\ +\;&\left(\Delta^2\epsilon^2+\Delta\epsilon^{\max(\alpha,2)}+\epsilon^{2+\max(\alpha,2)}\right)\end{aligned}$$

It is clear that when $\alpha<2$,

$$\mathbb{E}\|\hat{q}_\epsilon-\tilde{q}\|^2=\mathcal{O}(\Delta+\epsilon^{4-2\alpha}+\epsilon^2).$$

The error is minimized when $\alpha=4/3$, which is of order

$$\mathbb{E}\|\hat{q}_\epsilon-\tilde{q}\|^2=\mathcal{O}\left(\epsilon^{\frac{4}{3}}\right)\ .$$

It is easy to see when $\alpha > 2$, the error explodes. This completes the proof. □

4.5 Numerical Example

We show our findings in this chapter through the a numerical example. The multiscale system of interest is

$$\frac{dx}{dt} = \frac{1}{\epsilon}(-x-y) + (-x+y) + \sqrt{2}\frac{dU}{dt} \tag{4.37a}$$

$$\frac{dy}{dt} = \frac{1}{\epsilon^2}(-x-y) + \frac{\sqrt{2}}{\epsilon}\frac{dV}{dt}, \tag{4.37b}$$

The homogenized equation is

$$\frac{dX}{dt} = -2X + \sqrt{4}\frac{dW}{dt}. \tag{4.38}$$

To justify the optimal subsampling rate for drift estimator, we simulate the multiscale system using the exact solution of the OU process. Each path is subsampled with $N = \epsilon^{1.5}$ number of observations at a time increment of $\Delta = \epsilon^{\alpha}$, with $\alpha \in [0.1, 1]$. We take $\epsilon = 2^{-4}, \ldots, 2^{-12}$. Each estimate is based on 100 paths. The initial condition is set at $(x_0, y_0) = (0, 0)$.

The L^2 norm of the errors from the drift estimator $\hat{a}_{N,\epsilon}$ at different subsampling rate and ϵ are plotted in Figure 4.1, we can find the optimal subsampling rate α is roughly between 0.5 and 0.7, which agrees with our choice of $\alpha = 1/2$. Figure 4.2 provides an alternative view of the 3D contour surface.

To justify the optimal subsampling rate for the diffusion estimator $\hat{q}_{\epsilon}$, we simulate the multiscale system using exact solution. Each path is generated over a fixed total time horizon of $T = 1$, at a very fine resolution with $\delta = 2^{-20}$, with available number of observations $N = 2^{20}$. Each estimate is based on 100 paths. We take the scale parameter $\epsilon = 2^{-2}, \ldots, 2^{-9.5}$, and test the diffusion estimator a sequence of subsampling rates α over each path at rates $[0.1, 2]$. When subsampling the observations, we make full use of each simulated path as introduced in [1] by setting the start of each subsampled sequence consecutively.

We find from Figure 4.3, that the L^2 norm of the error is minimized roughly within the interval of $\alpha = [1.2, 1.6]$, this agrees with our expectation of finding $\alpha = 4/3$ optimizes error. Figure 4.4 provides an alternative view of the 3D contour surface.

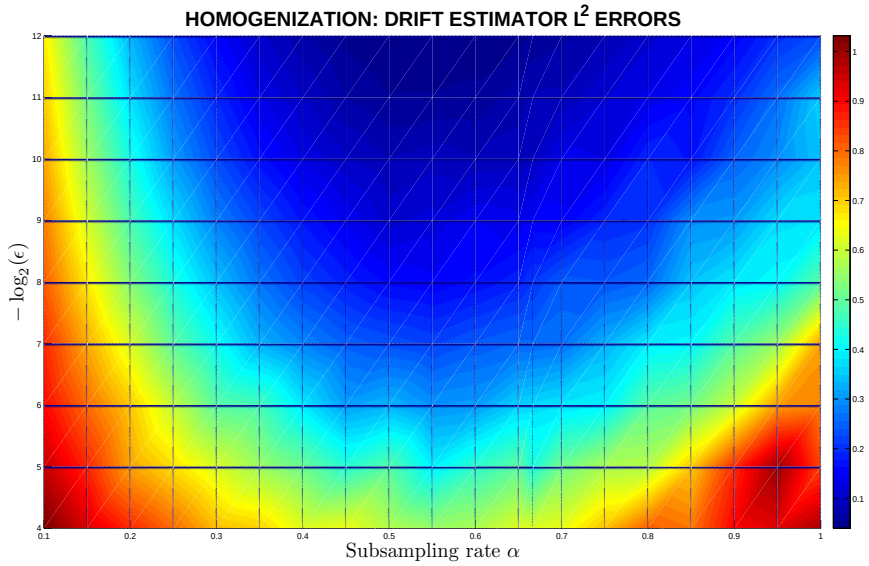

Figure 4.1: Homogenization: L^2 norm of $(\hat{a}_{N,\epsilon} - \tilde{a})$ for different ϵ and α

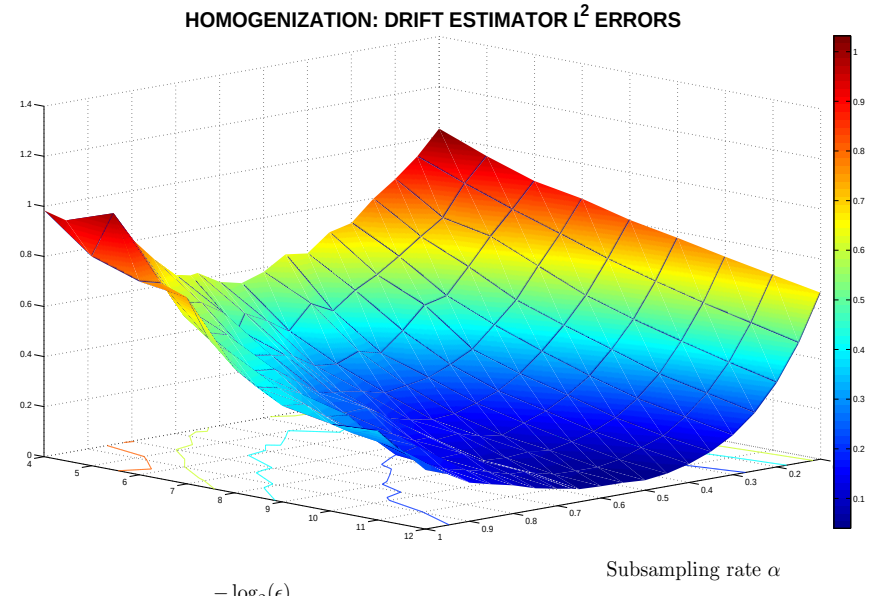

Figure 4.2: Homogenization: L^2 norm of $(\hat{a}_{N,\epsilon} - \tilde{a})$ for different ϵ and α (alternative view)

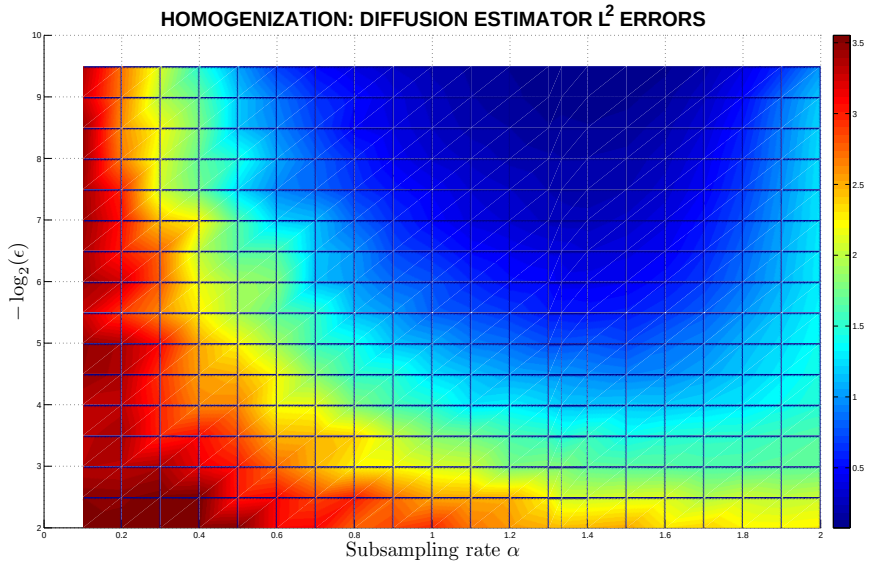

Figure 4.3: Homogenization: L^2 norm of $(\hat{q}_\epsilon - \tilde{q})$ for different ϵ and α

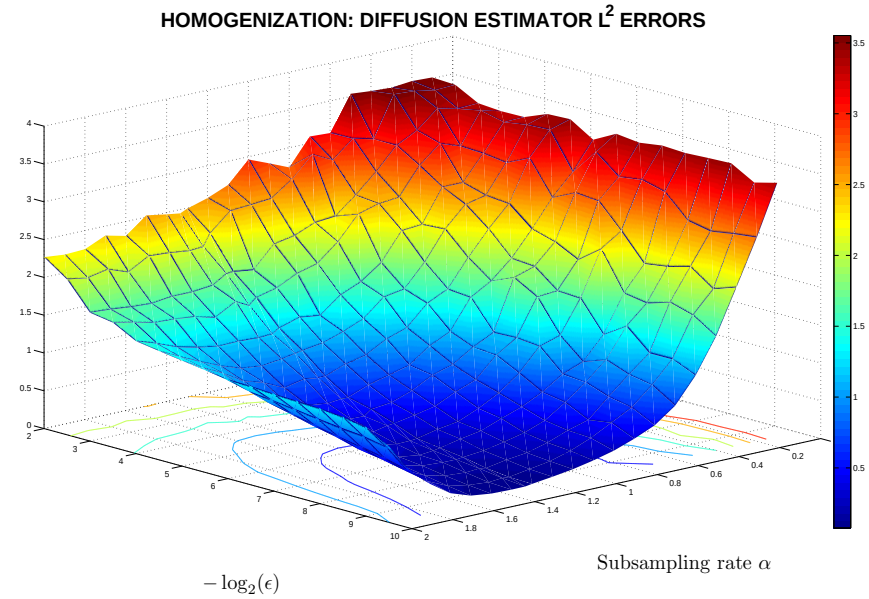

Figure 4.4: Homogenization: L^2 norm of $(\hat{q}_\epsilon - \tilde{q})$ for different ϵ and α (alternative view)

4.6 Conclusion

In this chapter, we have verified asymptotic limits of the maximum likelihood estimators for the drift (2.5) and diffusion (2.6) parameters of an OU process, while observing data from the slow part of a multiscale system (4.1). We have verified that the discrepancy between the solution of the homogenized equation (4.2) and the slow part of the system (4.1a), in the L^2 sense, is small when ϵ is small. In summary,

- we take discrete observations from the multiscale system (4.1a) x ;
- we have shown that the mismatch between trajectories of x and X is asymptotically small if ϵ is small;
- we have shown that the maximum likelihood estimator $\hat{a}_{N,\epsilon}$ converges to $\tilde{a}$ as $\epsilon \to 0$, with proper subsampling at time step $\Delta = \epsilon^{\alpha}$, and $N = \epsilon^{-\gamma}$. The values $\alpha = 1/2$ and $\gamma \geq 3/2$ optimize the error in L^2 sense;
- we have shown that the maximum likelihood estimator $\hat{q}_{\epsilon}$ converges to $\tilde{q}$ as $\epsilon \to 0$, with proper subsampling at time step $\Delta = \epsilon^{\alpha}$, and the total time horizon fixed at $T = N\Delta$. The values $\alpha = 4/3$ optimizes the error in L^2 sense;

We did not examine the asymptotic variances for the estimators as we did for the case of averaging. It is because we believe that the method we used in the case of averaging is not readily applicable for the case of homogenization, since it results in too many decomposed error terms, the correlations become too difficult to be accurately quantified. We believe we need better ways to decompose the error terms, and also better tools to quantify the limiting variances.

In future works, when possible, we can further relax the assumptions imposed on the drift and diffusion matrices, possibly in the ways addressed in the remarks to the assumptions.

Part II

FILTERING FOR MULTISCALE PROCESSES

Chapter 5

Averaging and Kalman Filter

In Part I, we have shown the behaviour of multiscale methods of averaging and homogenization applied to the drift and diffusion estimation problem of Ornstein Uhlenbeck (OU) processes. Since the path of the slow part of the OU process can be approximated closely by the averaged/homogenized process as we have shown in Sections 3.2 and 4.2 respectively, it is natural that we would like to take advantage of this property in other applications. In this chapter, we will integrate the method of averaging with the linear filtering problem. We compare the behaviour of the Kalman filter [9, 22, 47, 68] for a multiscale OU system with that for the averaged system. We will look at the behaviour of the Kalman filter for a system of multiscale OU process, as well as the Kalman filter for the averaged process. Our goal is to show that the marginal Kalman filtered distribution for the slow part of the multiscale OU system approximates the filtered distribution from the averaged process using the data from the multiscale system.

We derive the Kalman filter for the multiscale OU system in section 5.1, and then introduce the Kalman filter for the averaged process in section 5.2. We discuss the convergence between Kalman filters for the multiscale system and the averaged process in section 5.3. A numerical example is discussed in section 5.4.

5.1 Kalman Filter for the Multiscale System

Recall the multiscale system (3.1) satisfying Assumptions 3.1,

$$\begin{aligned} \frac{dx}{dt} &= a_{11}x + a_{12}y + \sqrt{q_1}\frac{dU}{dt} \\ \frac{dy}{dt} &= \frac{1}{\epsilon}\left(a_{21}x + a_{22}y\right) + \sqrt{\frac{q_2}{\epsilon}}\frac{dV}{dt} \end{aligned}$$

with initial condition $(x(0)^*, y(0)^*)^* \sim \mathcal{N}(m_0, v_0)$.

We rewrite the above system as

$$\frac{d\mathbf{x}}{dt} = \mathbf{a}\mathbf{x} + \sqrt{\mathbf{q}}\frac{dW}{dt}, \quad \mathbf{x}(0) \sim \mathcal{N}(m_0, v_0), \tag{5.2}$$

where,

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix},$$

$$\mathbf{a} = \mathbf{a}_0 + \frac{1}{\epsilon}\mathbf{a}_1 = \begin{pmatrix} a_{11} & a_{12} \\ 0 & 0 \end{pmatrix} + \frac{1}{\epsilon}\begin{pmatrix} 0 & 0 \\ a_{21} & a_{22} \end{pmatrix},$$

and

$$\mathbf{q} = \mathbf{q}_0 + \frac{1}{\epsilon}\mathbf{q}_1 = \begin{pmatrix} q_1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\epsilon}\begin{pmatrix} 0 & 0 \\ 0 & q_2 \end{pmatrix}$$

for which $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\mathbf{x} \in \mathcal{X} \oplus \mathcal{Y}$. We take $\mathcal{X} = \mathbb{R}^{d_1}$, and $\mathcal{Y} = \mathbb{R}^{d_2}$. Suppose we observe z, a noise contaminated integral of $\mathbf{x}$, which follows the SDE

$$\frac{dz}{dt} = h\mathbf{x} + \sqrt{\tau}\frac{dw}{dt}, \quad z(0) = 0, \tag{5.3}$$

where $h = (h_1, h_2)$ for which $h_1 \in \mathbb{R}^{l \times d_1}$, $h_2 \in \mathbb{R}^{l \times d_2}$, τ is invertible and W, w are independent standard Brownian motions. Equation (5.3) shows that the observation is a linear transformation of the hidden process $\mathbf{x}$, contaminated by Gaussian noise. Notice that $\mathbf{x}$ is Gaussian from equation (5.2). Under this setup, the conditional distribution of $\mathbf{x}|z$ is also Gaussian and is characterized by a mean $m(t)$ and covariance matrix $v(t)$. These two quantities satisfy a pair of closed nonlinear ODEs, known as the Kalman filter [68]:

$$\frac{dv}{dt} = \mathbf{a}v + v\mathbf{a}^* - vh^*\tau^{-1}hv + \mathbf{q}, \tag{5.4a}$$

$$dm = \mathbf{a}mdt + (vh^*\tau^{-1})(dz - hmdt). \tag{5.4b}$$

with initial conditions $v(0)$ and $m(0)$

Our interest is the conditional distribution of the slow part of the multiscale system x given observations z, which is the marginal of the Gaussian distribution $\mathcal{N}(m(t), v(t))$. We set

$$v_x = I_{d_1} v I_{d_1}^*, \tag{5.5a}$$

$$m_x = I_{d_1} m, \tag{5.5b}$$

where

$$I_{d_1} = \left(\begin{array}{ccc|ccc} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \end{array} \right) \begin{matrix} \\ \\ \\ \end{matrix} \Big\} d_1 \quad . \tag{5.6}$$

$$\underbrace{\qquad\qquad}_{d_1} \quad \underbrace{\qquad\qquad}_{d_2}$$

and $*$ denotes transposition of a matrix or the adjoint of an operator.

5.2 Kalman Filter for the Averaged Process

Now consider the averaged equation for the slow process in (5.2). From Chapter 3, we know that the slow process x of (5.2) can be averaged to the following SDE

$$\frac{dX}{dt} = \tilde{a}X + \sqrt{q_1}\frac{dU}{dt}\ ,\ X(0) = x(0) \sim \mathcal{N}(M_0, V_0) \tag{5.7}$$

where $\tilde{a} = a_{11} - a_{12}a_{22}^{-1}a_{21}$.

The observations taken from z should be close to Z from the SDE below, if ϵ is small,

$$\frac{dZ}{dt} = \tilde{h}X + \sqrt{\tau}\frac{dw}{dt} \quad , \quad Z(0) = 0, \tag{5.8}$$

where $\tilde{h} = h_1 - h_2 a_{22}^{-1} a_{21}$.

The conditional distribution $X(t)|Z(t)$ is Gaussian and is characterized by a mean $M(t)$ and covariance matrix $V(t)$. The corresponding Kalman filter can be derived as the following coupled SDEs,

$$\frac{dV}{dt} = \tilde{a}V + V\tilde{a}^* - V\tilde{h}^*\tau^{-1}\tilde{h}V + q_1, \tag{5.9a}$$

$$dM = \tilde{a}Mdt + (V\tilde{h}^*\tau^{-1})(dZ - \tilde{h}Mdt). \tag{5.9b}$$

We will show in Section 5.3, that if we feed the observations from (5.3) to the Kalman filter (5.9), with identical initial conditions, the Gaussian distribution characterized by (5.5) converges to (5.9), as $\epsilon \to 0$.

5.3 The Convergence of the Kalman Filters

We will prove in this section, as ϵ gets small, the distribution of Kalman filter described in (5.5) converges to the distribution of the Kalman filter described (5.9), given that we take $Z(t) = z(t)$ in equation (5.9b).

Theorem 5.1. *Consider the scale separated Ornstein-Uhlenbeck system* (5.2)*, and let Assumptions 3.1 hold and only consider the scalar processes when* $d_1 = d_2 = 1$*. By feeding noisy observations* $\{z(s)\}_{0\leq s\leq t}$ *from equation* (5.3) *to both the marginal filtered distribution of the multiscale system* $\mathcal{N}(m_x(s, z(s)), v_x(s))$ *and the filter distribution of the averaged equation* $\mathcal{N}(M(s, z(s)), V(s))$*, then, for every* $1 \gg \epsilon > 0$*,*

(i) $\|v_x(s) - V(s)\| = \mathcal{O}(\epsilon)$*, for any* $s \in [0, t]$*. given initial condition* $v_x(0) = V(0)$*;*

(ii) $\mathbb{E}\left(m_x(s, z(s)) - M(s, z(s))\right)^2 = \mathcal{O}(\epsilon^2)$ *for any* $s \in [0, t]$*, given initial condition* $m_x(0, z(0)) = M(0, z(0))$*.*

Before going to the main result of this chapter, we first define some linear operators for an arbitrary symmetric matrix m:

$$\begin{aligned}\mathcal{L}_0 m &= \mathbf{a}_0 m + m\mathbf{a}_0^*,\\ \mathcal{L}_1 m &= \mathbf{a}_1 m + m\mathbf{a}_1^*,\\ \mathcal{L}^\epsilon m &= \mathbf{a} m + m\mathbf{a}^* = \mathcal{L}_0 m + \frac{1}{\epsilon}\mathcal{L}_1 m.\end{aligned}$$

and for simplicity, denote

$$S = h^*\tau^{-1}h\ . \tag{5.10}$$

We will write the covariance matrix v in terms of its scalar entries,

$$v = \begin{pmatrix} v_{11} & v_{12}\\ v_{21} & v_{22}\end{pmatrix}$$

since v characterizes covariance, $v_{21} = v_{12}^*$.

Proof. Proof of *(i)* in Theorem 5.1.

Recall the definition for the variance (5.4a) written in simplified notation defined above

$$\begin{aligned}\frac{dv}{dt} &= \mathbf{a}v + v\mathbf{a}^* - vh^*\tau^{-1}hv + \mathbf{q}\\ &= \mathcal{L}_0 v + \frac{1}{\epsilon}\mathcal{L}_1 v - vSv + \left(\mathbf{q}_0 + \frac{1}{\epsilon}\mathbf{q}_1\right)\end{aligned} \tag{5.11}$$

By rearranging the equation above, we have

$$\mathcal{L}_1 v = -\mathbf{q}_1 + \epsilon\left(\frac{dv}{dt} - \mathcal{L}_0 v + vSv - \mathbf{q}_0\right) \tag{5.12}$$

Using the block representation of v, we have a system of 3 equations, for which the

equation corresponds to the lower-left block can be written as follows,

$$\begin{aligned} a_{21}v_{11} + a_{22}v_{21} &= \epsilon\Big(\frac{dv_{21}}{dt} - (v_{21}a_{11} + v_{22}a_{12}) \\ &+ ((v_{21}S_{11} + v_{22}S_{12})v_{11} + (v_{21}S_{12} + v_{22}S_{22})v_{21}) \\ &- q_1) \end{aligned}$$

The above equation can be solved approximately for v_{21} as a function of v_{11},

$$v_{21} = -a_{22}^{-1}a_{21}v_{11} + \epsilon C \tag{5.13}$$

We now try to find an ϵ-independent bound for the C in equation (5.13). By writing the symmetric square matrices in vector forms, we define $w_v = a_{21}v_{11} + a_{22}v_{12} = \epsilon a_{22}C$, and $\vec{b} = (a_{21}, a_{22}, 0)$,

$$\vec{v} = \begin{pmatrix} v_{11} \\ v_{12} \\ v_{22} \end{pmatrix}, \vec{q}_0 = \begin{pmatrix} q_1 \\ 0 \\ 0 \end{pmatrix}, \vec{q}_1 = \begin{pmatrix} 0 \\ 0 \\ q_2 \end{pmatrix}.$$

We can write

$$\frac{dw_v}{dt} = \vec{b}\frac{d\vec{v}}{dt} = \vec{b}\left(\vec{a}_1\vec{v} + \frac{1}{\epsilon}\vec{a}_2\vec{v} + \vec{q}_0 + \frac{1}{\epsilon}\vec{q}_1 - \vec{F}(\vec{v})\right)$$

for which $\vec{F}(\vec{v})$ is the corresponding vectorized form of matrix vSv, and

$$\vec{a}_1 = \begin{pmatrix} 2a_{11} & 2a_{12} & 0 \\ 0 & a_{11} & a_{12} \\ 0 & 0 & 0 \end{pmatrix}, \vec{a}_2 = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 2a_{21} & 2a_{22} \end{pmatrix}.$$

We find that for the fast scale terms with order $\mathcal{O}(\frac{1}{\epsilon})$,

$$\vec{b}\vec{a}_2\vec{v} = (a_{21}, a_{22}, 0)\begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 2a_{21} & 2a_{22} \end{pmatrix}\begin{pmatrix} v_{11} \\ v_{12} \\ v_{22} \end{pmatrix} = a_{22}w_v$$

and

$$\vec{b}\vec{q}_1 = (a_{21}, a_{22}, 0)\begin{pmatrix} 0 \\ 0 \\ q_2 \end{pmatrix} = 0$$

Therefore, we can write the differential equation for w_v as

$$\begin{aligned}\frac{dw_v}{dt} &= \frac{1}{\epsilon}a_{22}w_v + \vec{b}\vec{q}_0 + \vec{b}\vec{a}_1\vec{v} - \vec{b}\vec{F}(\vec{v}) \\ &= \frac{1}{\epsilon}a_{22}\left(w_v + \epsilon\vec{G}(v)\right)\end{aligned}$$

where $\vec{G}(\vec{v}) = a_{22}^{-1}\left(\vec{b}\vec{q}_0 + \vec{b}\vec{a}_1\vec{v} - \vec{b}\vec{F}(\vec{v})\right)$.

Solving this nonhomogeneous first order linear equation, we find the solution

$$w_v(t) = \epsilon\int_0^t e^{\frac{a_{22}}{\epsilon}(t-s)}\vec{G}(\vec{v}(s))ds\ .$$

Since the unconditional variance of the invariant measure of a multiscale OU system of the form (3.1) is finite and independent of time and ϵ, we have $\lim_{t\to\infty}\mathbb{E}(x^2(t)) \leq c$ and $\lim_{t\to\infty}\mathbb{E}(y^2(t)) \leq c$, for c some finite constant independent of t and ϵ. Immediately, we also know that the covariance $\mathbb{E}(x(t)y(t)) \leq \sqrt{\mathbb{E}(x^2(t))\mathbb{E}(y^2(t))} \leq c$. Hence we know that the unconditional covariance matrix of the multiscale system is finite, and by the solution to an OU SDE as described in (2.2) and the invariant variance as in (2.4), we can immediately deduce that the invariant variance is the upper bound for the variance of $x(t)$ and $y(t)$. We know that $v(t)$ is the conditional covariance matrix of the coupled system $(x(t), y(t))$ given $\{z(s)\}_{0\leq s\leq t}$. We know the conditional variance is always less than or equal to the unconditional variance, and since the unconditional variance is bounded above by the invariant variance, we conclude that $v(t) \leq C$, for some C independent of t and ϵ.

Since $\vec{G}(\vec{v}(t))$ is a continuous quadratic function of $v(t)$, and $a_{22} < 0$, we can obtain an upper bound for the following integral,

$$\|\int_0^t e^{\frac{a_{22}}{\epsilon}(t-s)}\vec{G}(\vec{v}(s))ds\| \leq \sup_{0\leq s\leq t}\|\vec{G}(\vec{v}(s))\|\|\int_0^t e^{\frac{a_{22}}{\epsilon}(t-s)}ds\| \leq c$$

for which the $\|\cdot\|$ is a vector norm, and c is an ϵ-independent constant. Therefore, we proved that $w_v(t)$ is of order $\mathcal{O}(\epsilon)$, and consequently,

$$v_{21} = -a_{22}^{-1}a_{21}v_{11} + \epsilon C$$

where C is of order $\mathcal{O}(1)$.

According to (5.11), we have the equation for the top-left block as follows,

$$\begin{aligned}\frac{dv_{11}}{dt} &= ((a_{11}v_{11}+a_{12}v_{21})+(v_{11}a_{11}+v_{12}a_{12}))\\ &+ \frac{1}{\epsilon}\cdot 0\\ &- ((v_{11}S_{11}+v_{12}S_{21})v_{11}+(v_{11}S_{12}+v_{12}S_{22})v_{21})\\ &+ q_1\end{aligned}$$

By substituting the solution of v_{21} from equation (5.13) into the above equation, and considering $v_{21}=v_{12}$, and the definitions of $\tilde{h}$ and $\tilde{a}$, we have

$$\begin{aligned}\frac{dv_{11}}{dt} &= (a_{11}v_{11}+a_{12}v_{12})+(v_{11}a_{11}+v_{12}a_{12})\\ &- \left(v_{11}h_1\tau^{-1}h_1+v_{12}h_2\tau^{-1}h_1\right)v_{11}-\left(v_{11}h_1\tau^{-1}h_2+v_{12}h_2\tau^{-1}h_2\right)v_{12}\\ &+ q_1\\ &= \left(a_{11}-a_{12}a_{22}^{-1}a_{21}+\epsilon C\right)v_{11}+v_{11}\left(a_{11}-a_{21}a_{22}^{-1}a_{12}+\epsilon C\right)\\ &- \left(v_{11}h_1+v_{12}h_2\right)\tau^{-1}h_1v_{11}-\left(v_{11}h_1+v_{12}h_2\right)\tau^{-1}h_2v_{12}\\ &+ q_1\\ &= \left(a_{11}-a_{12}a_{22}^{-1}a_{21}+\epsilon C\right)v_{11}+v_{11}\left(a_{11}-a_{12}a_{22}^{-1}a_{21}+\epsilon C\right)\\ &- v_{11}\left(h_1-a_{21}a_{22}^{-1}h_2+\epsilon C\right)\tau^{-1}h_1v_{11}-v_{11}\left(h_1-a_{21}a_{22}^{-1}h_2+\epsilon C\right)\tau^{-1}h_2v_{12}\\ &+ q_1\\ &= \left(a_{11}-a_{12}a_{22}^{-1}a_{21}+\epsilon C\right)v_{11}+v_{11}\left(a_{11}-a_{12}a_{22}^{-1}a_{21}+\epsilon C\right)\\ &- v_{11}\left(h_1-a_{21}a_{22}^{-1}h_2+\epsilon C\right)\tau^{-1}\left(h_1-h_2a_{22}^{-1}a_{21}+\epsilon C\right)v_{11}\\ &+ q_1\\ &= (\tilde{a}+\epsilon C)v_{11}+v_{11}(\tilde{a}+\epsilon C)-v_{11}(\tilde{h}+\epsilon C)\tau^{-1}(\tilde{h}+\epsilon C)v_{11}+q_1\\ &= \tilde{a}v_{11}+v_{11}\tilde{a}-v_{11}\tilde{h}\tau^{-1}\tilde{h}v_{11}+q_1+\epsilon C\ .\end{aligned}$$

As $\epsilon\to 0$, this equation converges to the equation for V in (5.9a). Since $v_{11}=v_x$ and noting that we are considering v_{11} and V to be scalars, we can show that

$$\begin{aligned}\frac{d(v_x-V)}{dt} &= \frac{d(v_{11}-V)}{dt}\\ &= 2\tilde{a}(v_{11}-V)-\tilde{h}\tau^{-1}\tilde{h}(v_{11}^2-V^2)+\epsilon C\\ &= \left(2\tilde{a}-\tilde{h}\tau^{-1}\tilde{h}(v_{11}+V)\right)(v_{11}-V)+\epsilon C\end{aligned}$$

The solution to the above equation bounded as

$$\begin{aligned}\|(v_x - V)(t)\| &\leq \|\int_0^t \sup_{0\leq s\leq t}\left((e^{2\tilde{a}-\tilde{h}\tau^{-1}\tilde{h}(v_{11}(s)+V(s))}\right)\epsilon C(t)ds\| \\ &\leq \epsilon C(t)\|\int_0^t e^{2\tilde{a}}ds\| \\ &\leq \epsilon C(t)\end{aligned}$$

where $C(t)$ is a constant for every t, which is independent of ϵ .

The proof for statement *(i)* of Theorem 5.1 is complete. □

We then prove the convergence between the mean of the marginal distribution of the filtered multiscale OU system and the filtered mean of the averaged distribution.

Proof. Proof of *(ii)* in Theorem 5.1.

We write m in the form of its scalar entries,

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix}$$

Recall the definition for the filtered mean (5.4b)

$$dm = \mathbf{a}mdt + (vh^*\tau^{-1})(dz - h^*mdt) \tag{5.14}$$

Rearranging the above equation, we have

$$\mathbf{a}_1 m = \epsilon\left(\frac{dm}{dt} - \mathbf{a}_0 m - (vh^*\tau^{-1})(\frac{dz}{dt} - hm)\right)$$

By rewriting the above equation in block representation, we have a system of 2 equations, the equation for the lower block,

$$a_{21}m_1 + a_{22}m_2 = \epsilon\left(\frac{dm_2}{dt} - (vh^*\tau^{-1})(\frac{dz}{dt} - hm)\right)$$

We have the solution of m_2 as a function of m_1

$$m_2 = -a_{22}^{-1}a_{21}m_1 + \epsilon C \tag{5.15}$$

We now try to find a bound for C in equation (5.15). We define $w_m = a_{21}m_1 +$

$a_{22}m_2 = \epsilon a_{22}C$ and $\vec{b} = (a_{21}, a_{22})$. We can write the differential equation for w_m as

$$\frac{dw_m}{dt} = \vec{b}\frac{dm}{dt} = \vec{b}\left(\mathbf{a}_0 m(t) + \frac{1}{\epsilon}\mathbf{a}_1 m(t) + (v(t)h^*\tau^{-1})\left(\frac{dz}{dt} - hm(t)\right)\right)$$

We find that for the fast scale term with order $\mathcal{O}(\frac{1}{\epsilon})$,

$$\vec{b}\mathbf{a}_1 m = (a_{21}, a_{22})\begin{pmatrix} 0 & 0 \\ a_{21} & a_{22}\end{pmatrix}\begin{pmatrix} m_1 \\ m_2\end{pmatrix} = a_{22}w_m$$

Therefore, we can write the differential equation for w_m as

$$\begin{aligned}\frac{dw_m}{dt} &= \frac{1}{\epsilon}\vec{b}\mathbf{a}_1 m(t) + \vec{b}\left(\mathbf{a}_0 m(t) + (v(t)h^*\tau^{-1})\left(\frac{dz}{dt} - hm(t)\right)\right) \\ &= \frac{1}{\epsilon}a_{22}\left(w_m + \epsilon a_{22}^{-1}\vec{b}\left(\mathbf{a}_0 m(t) + (v(t)h^*\tau^{-1})\left(\frac{dz}{dt} - hm(t)\right)\right)\right)\end{aligned}$$

We substitute in the definition for $\frac{dz}{dt}$ from equation (5.3), we have

$$\frac{dw_m}{dt} = \frac{1}{\epsilon}a_{22}\left(w_m + \epsilon a_{22}^{-1}\vec{b}\left(\mathbf{a}_0 m(t) + (v(t)h^*\tau^{-1})\left(h\mathbf{x}(t) + \sqrt{\tau}\frac{dw}{dt} - hm(t)\right)\right)\right)$$

where w is a Brownian motion.

Solving this nonhomogeneous first order linear equation, we have the solution,

$$\begin{aligned}& w_m(t) \qquad (5.16)\\ =\ & \epsilon\int_0^t e^{\frac{a_{22}}{\epsilon}(t-s)}a_{22}^{-1}\vec{b}\left(\mathbf{a}_0 m(s) + (v(s)h^*\tau^{-1})(h\mathbf{x}(s) - hm(s))\right)ds \\ +\ & \epsilon\int_0^t e^{\frac{a_{22}}{\epsilon}(t-s)}a_{22}^{-1}\vec{b}(v(s)h^*\tau^{-1})\sqrt{\tau}dw_s\end{aligned}$$

We have proved that $\mathbb{E}\sup_{0\le s\le t}(\|x\|^2+\|y\|^2)$ is bounded in Lemma 3.5, so $\mathbb{E}(\mathbf{x}^2(s))$ is bounded. Since the coupled system $(x(s), y(s))$ has invariant mean of zero, and finite invariant variance, by Itô isometry, we see that $\mathbb{E}(w_m^2)$ is of order $\mathcal{O}(\epsilon^2)$, consequently, C in equation (5.15) has a bound in L^2, ie. $\mathbb{E}(C^2) = \mathcal{O}(1)$.

We then substitute the solution for m_2 in equation (5.15) and v_{11} in equation (5.13)

into the upper block of the rearranged equation (5.14)

$$\begin{aligned} dm_1 &= (a_{11}m_1 + a_{12}m_2)dt + (vh^*\tau^{-1})(dz - (h_1m_1 + h_2m_2)dt) \\ &= (\tilde{a}m_1 + a_{22}^{-1}w_m(t))dt + (v_{11}\tilde{h}\tau^{-1} + \epsilon C(t))(dz - (\tilde{h}m_1 + h_2a_{22}^{-1}w_m(t))dt) \\ &= \tilde{a}m_1dt + (v_{11}\tilde{h}\tau^{-1})(dz - \tilde{h}m_1dt) \\ &+ a_{22}^{-1}w_m(t)dt \\ &+ \epsilon C(t)(dz - (\tilde{h}m_1 + h_2a_{22}^{-1}w_m(t))dt) \\ &- (v_{11}\tilde{h}\tau^{-1})h_2a_{22}^{-1}w_m(t)dt \end{aligned}$$

Notice $m_1 = m_x$, we can express the difference between $m_x(t, z(t))$ and $M(t, z(t))$,

$$\begin{aligned} d(m_x(t,z(t)) - M(t,z(t))) &= \left(\tilde{a} - v_{11}\tilde{h}\tau^{-1}\tilde{h}\right)(m_1 - M)dt \\ &+ a_{22}^{-1}w_m dt \\ &+ \epsilon C(t)(dz - (\tilde{h}m_1 + h_2a_{22}^{-1}w_m(s))ds) \\ &- (v_{11}(s)\tilde{h}\tau^{-1})h_2a_{22}^{-1}w_m(s)ds \end{aligned}$$

Hence we have the solution

$$\begin{aligned} (m_x - M)(t,z(t)) &= \int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}^2\tau^{-1})(t-s)}a_{22}^{-1}w_m(s)ds \\ &+ \epsilon C(t)\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}^2\tau^{-1})(t-s)}h\mathbf{x}dt \\ &+ \epsilon C(t)\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}^2\tau^{-1})(t-s)}\sqrt{\tau}dw \\ &- \epsilon C(t)\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}^2\tau^{-1})(t-s)}\tilde{h}w_m(s)ds \\ &- \epsilon C(t)\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}^2\tau^{-1})(t-s)}h_2a_{22}^{-1}w_m(s)ds \\ &- \int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}^2\tau^{-1})(t-s)}v_{11}(s)\tau^{-1}\tilde{h}^2a_{22}^{-1}w_m(s)ds \end{aligned}$$

Since we know that $\mathbb{E}(w_m^2) = \mathcal{O}(\epsilon^2)$, using that $v_{11}(t) \geq 0$ for all t, we have $e^{\tilde{a}-v_{11}\tilde{h}^2\tau^{-1}} \leq e^{\tilde{a}}$, and v_{11} has a finite upper bound since it is the conditional variance compared to the invariant invariance of x. Using these facts, and by Cauchy-Schwarz inequality, we have

$$\begin{aligned} \mathbb{E}(m_x(t,z(t)) - M(t,z(t)))^2 &\leq \epsilon^2C(t) + \epsilon^2C(t) + \epsilon^2C(t) + \epsilon^4C(t) + \epsilon^4C(t) + \epsilon^2C(t) \\ &\leq \epsilon^2C(t) \end{aligned}$$

This completes the proof for statement *(ii)* of Theorem 5.1. □

5.4 Numerical Example

We illustrate our idea in Theorem 5.1 through a simple numerical example. The system of interest is

$$d\mathbf{x} = \mathbf{a}\mathbf{x}dt + \sqrt{\mathbf{q}}dU\ ,\ x_0 \sim N(0, I), \tag{5.17}$$

$$dz = h\mathbf{x}dt + \sqrt{\tau}dW\ ,\ z(0) = 0 \tag{5.18}$$

We observe $z(t)$, to model the conditional distribution of $\mathbf{x}(t)|z(t)$. We know this distribution is Gaussian, so we apply the Kalman filter to estimate the mean and variance of the distribution. We plot the direct Kalman filter (5.4) applied to the system, and plot the marginal mean and variance as described in (5.5). On the other hand, we model it through the averaged system

$$dX = \tilde{a}Xdt + \sqrt{q_1}dU\ ,\ X_0 \sim \mathcal{N}(0, I) \tag{5.19}$$

We know the conditional distribution of $X(t)|z(t)$ is Gaussian, so we plot the mean and variance of this conditional distribution from the averaged Kalman filter (5.9). In the numerical example below, the simulation of the SDE is via euler method without subsampling. We take following values for the parameters,

$$\mathbf{a} = \begin{pmatrix} -.5 & 1 \\ \frac{1}{\epsilon} & -\frac{1}{\epsilon} \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{\epsilon} \end{pmatrix}, h = (1, 1)$$

$\tilde{a} = 0.5$, $q_1 = 1$, $\tau = 0.1$, $\epsilon = 2^{-10}$, $\delta = 2^{-12}$, $n = 2^{13}$.

In Figure 5.1, we see the actual path of slow process from the multiscale system (blue line) and the path of the averaged process (red line) closely follow each other. This illustrates the convergence of the paths for averaged process stated in Theorem 3.6. The two standard deviation confidence intervals from the Kalman filters, both direct (cyan lines) and averaged (green lines), almost lie directly on top of each other. They both provide a good support for the actual path of the slow process x from the OU system.

In Figure 5.2, we plot the convergence of the Kalman filter variances. We see that the variance of marginal distribution $x|z$ (dashed cyan line) and $X|z$ (dashed green line) are almost the same. The speed of convergence of the variances are very fast and stable. The variance of $y|z$ (dash-dotted cyan line) is also plotted for illustration.

In Figure 5.3, we plot the squared error $(m_x(t) - M(t))^2$, from which we see that the error quickly converges to zero. The size of the error when settled is of scale 10^{-6},

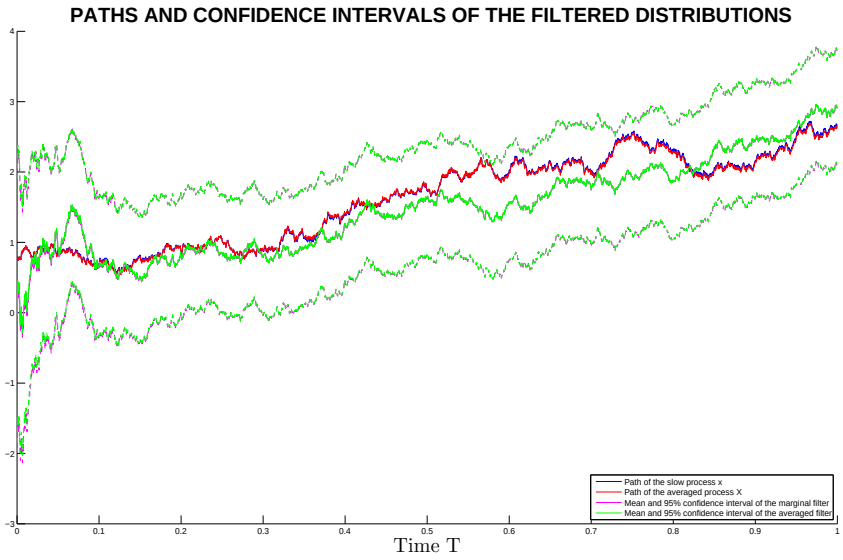

Figure 5.1: Paths and 95% confidence intervals of the filtered distributions

which is in line with our expectation of order $\mathcal{O}(\epsilon^2)$, for $\epsilon = 2^{-10}$.

In Figure 5.4, we plot the distance $\|v_x(t) - V(t)\|$, from which we see that the error quickly converges to a constant. The size of the constant is of observed at an order of $\mathcal{O}(\epsilon)$, which is in line with our expectations.

We conclude this is a good evidence to support Theorem 5.1. The advantage of this application of Kalman filter on averaged process mainly lies in reduction of observations and computational complexity when solving similar problems. There is no need to observe the entire multiscale system, but just the corresponding marginal data, and hence reduces computational time and requirement on computational resources.

5.5 Conclusion

In this chapter, we tried to integrate the multiscale method of averaging and Kalman filtering. We have proved that the slow part of the marginal filtered distribution of the multiscale system converges to the filter distribution of averaged process, given observations from the multiscale system contaminated with Gaussian noise, and ϵ is small. In summary,

- We take observations from the noisy observation (5.3);
- we use these observation to substitute into the mean and variance of the Kalman filter derived from the averaged process (5.9);

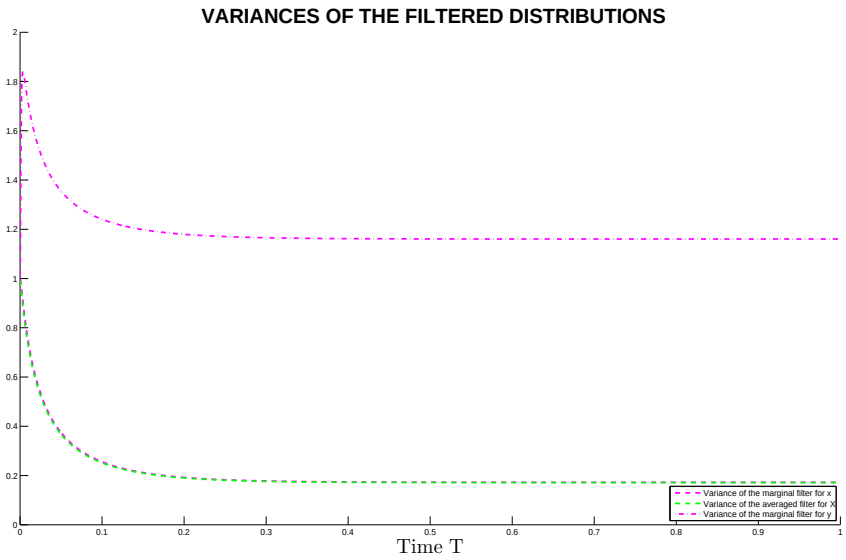

Figure 5.2: Variances of the filtered distributions

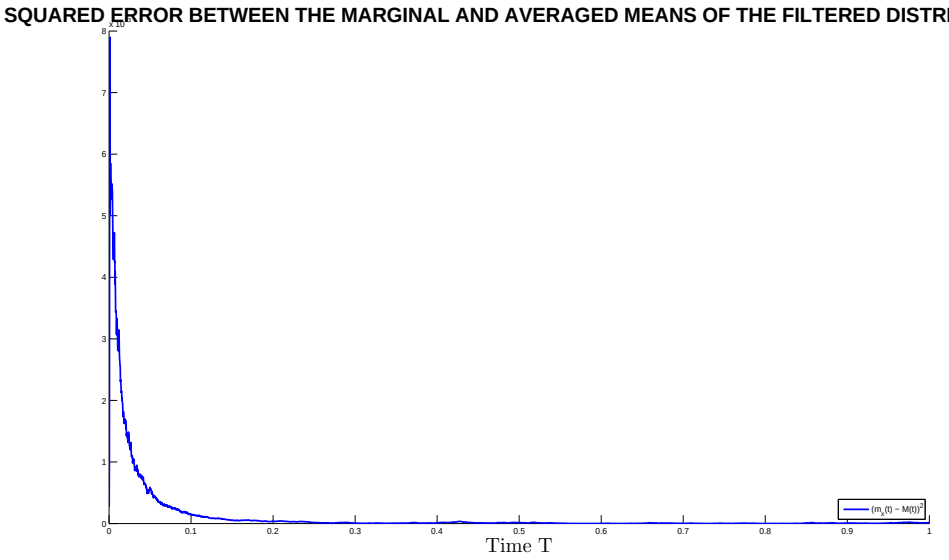

Figure 5.3: Squared error between the means of the marginal and averaged filtered distributions

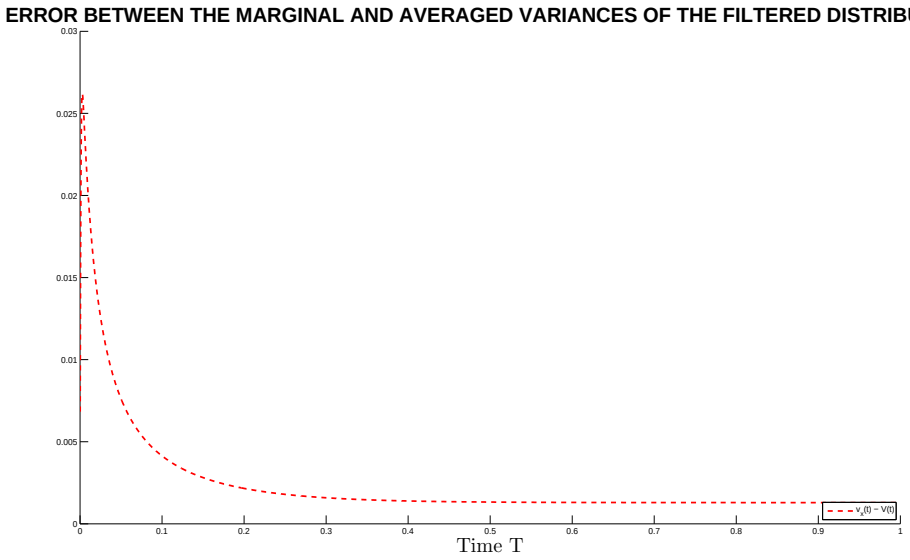

Figure 5.4: Error between the variances of the marginal and averaged filtered distributions

- we show that the Gaussian distributions characterized by (5.9) and (5.5) closely follow each other.

In Theorem 5.1, we proved the convergence between the marginal distribution of the Kalman filter applied on the whole coupled system and the distribution of that Kalman filter applied on the averaged process, when the coupled system is made of two scalar OU processes. However, we expect similar result to hold for any finite d_1 and d_2.

Chapter 6

Homogenization and Kalman Filter

In this chapter, we will compare the behaviour of the Kalman filter applied to a multiscale OU system with that applied to the homogenized process. We will look at the behaviour of the Kalman filter for a system of multiscale OU processes, as well as the Kalman filter for the homogenized process. Our goal is to show that the filtered marginal distribution for the slow part of the multiscale OU system approximates the filtered distribution of the homogenized process.

We derive the Kalman filter for the multiscale OU system in section 6.1, and then introduce the Kalman filter for the homogenized process in section 6.2. Then we discuss the convergence between the filtered marginal distribution for the slow part of the multiscale system and the filtered distribution for the homogenized process in section 6.3. A numerical example is discussed in section 6.4.

6.1 Kalman Filter for the Multiscale System

Recall the multiscale system (4.1) satisfying Assumptions 4.1,

$$\begin{aligned} \frac{dx}{dt} &= \frac{1}{\epsilon}(a_{11}x + a_{12}y) + (a_{13}x + a_{14}y) + \sqrt{q_1}\frac{dU}{dt} \\ \frac{dy}{dt} &= \frac{1}{\epsilon^2}(a_{21}x + a_{22}y) + \sqrt{\frac{q_2}{\epsilon^2}}\frac{dV}{dt} \end{aligned}$$

with initial condition $(x(0)^*, y(0)^*)^* \sim \mathcal{N}(m_0, v_0)$.

We rewrite the above system as

$$\frac{d\mathbf{x}}{dt} = \mathbf{a}\mathbf{x} + \sqrt{\mathbf{q}}\frac{dW}{dt}, \quad \mathbf{x}(0) \sim N(m_0, v_0) \tag{6.2}$$

where

$$\mathbf{x} = \begin{pmatrix} x \\ y \end{pmatrix},$$

$$\mathbf{a} = \mathbf{a}_0 + \frac{1}{\epsilon}\mathbf{a}_1 + \frac{1}{\epsilon^2}\mathbf{a}_2 = \begin{pmatrix} a_{13} & a_{14} \\ 0 & 0 \end{pmatrix} + \frac{1}{\epsilon}\begin{pmatrix} a_{11} & a_{12} \\ 0 & 0 \end{pmatrix} + \frac{1}{\epsilon^2}\begin{pmatrix} 0 & 0 \\ a_{21} & a_{22} \end{pmatrix},$$

and

$$\mathbf{q} = \mathbf{q}_0 + \frac{1}{\epsilon^2}\mathbf{q}_1 = \begin{pmatrix} q_1 & 0 \\ 0 & 0 \end{pmatrix} + \frac{1}{\epsilon^2}\begin{pmatrix} 0 & 0 \\ 0 & q_2 \end{pmatrix}.$$

for which $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\mathbf{x} \in \mathcal{X} \oplus \mathcal{Y}$. We take $\mathcal{X} = \mathbb{R}^{d_1}$, and $\mathcal{Y} = \mathbb{R}^{d_2}$. Suppose we observe z, a noise contaminated integral of $\mathbf{x}$, which follows the SDE

$$\frac{dz}{dt} = h\mathbf{x} + \sqrt{\tau}\frac{dw}{dt}, \quad z(0) = 0, \tag{6.3}$$

where $h = (h_1, h_2)$, for which $h_1 \in \mathbb{R}^{l \times d_1}$, $h_2 \in \mathbb{R}^{l \times d_2}$; τ invertible. W, w are independent standard Brownian motions of appropriate dimensions. Equation (6.3) shows that the observation is a linear transformation of the hidden process $\mathbf{x}$, contaminated by Gaussian noise. Notice that $\mathbf{x}$ is Gaussian from equation (6.2). Under this setup, the conditional distribution of $\mathbf{x}|z$ is also Gaussian and is characterized by a mean $m(t)$ and covariance matrix $v(t)$. These two quantities satisfy a pair of closed nonlinear ODEs,

$$\frac{dv}{dt} = \mathbf{a}v + v\mathbf{a}^* - vh^*\tau^{-1}hv + \mathbf{q} \tag{6.4a}$$

$$dm = \mathbf{a}mdt + (vh^*\tau^{-1})(dz - hmdt)\,. \tag{6.4b}$$

Our interest is the conditional distribution of the slow part of the multiscale system x given observations z, which is the marginal of the Gaussian distribution $\mathcal{N}(m(t), v(t))$. We set

$$m_x = I_{d_1} m \tag{6.5a}$$

$$v_x = I_{d_1} v I_{d_1}^* \tag{6.5b}$$

where I_{d_1} is as in (5.6).

6.2 Kalman Filter for the Homogenized Process

Now consider the homogenized equation for the slow process in (6.2). From Chapter 4, we know that the "slow" process x of (6.2) can be homogenized to the following SDE

$$\frac{dX}{dt} = \tilde{a}X + \sqrt{\tilde{q}}\frac{dU'}{dt} \quad , \quad X(0) = x(0) \sim \mathcal{N}(M_0, V_0), \tag{6.6}$$

where $\tilde{a} = a_{13} - a_{14}a_{22}^{-1}a_{21}$, and $\tilde{q} = q_1 + a_{12}a_{22}^{-1}q_2\left(a_{22}^{-1}\right)^* a_{12}^*$. The observations taken from z should be close to Z from the SDE below, if ϵ is small,

$$\frac{dZ}{dt} = \tilde{h}X + \sqrt{\tau}\frac{dw}{dt} \quad , \quad Z(0) = 0, \tag{6.7}$$

where $\tilde{h} = h_1 - h_2 a_{22}^{-1}a_{21}$.

The conditional distribution of $X(t)|Z(t)$ is Gaussian and is characterized by a mean $M(t)$ and covariance matrix $V(t)$. The corresponding Kalman filter can be derived as the following ODEs,

$$\frac{dV}{dt} = \tilde{a}V + V\tilde{a}^* - V\tilde{h}^*\tau^{-1}\tilde{h}V + \tilde{q} \tag{6.8a}$$

$$dM = \tilde{a}Mdt + (V\tilde{h}^*\tau^{-1})(dZ - \tilde{h}Mdt)\,. \tag{6.8b}$$

We will show in Section 6.3, that if we feed the observations from (6.3) to the Kalman filter (6.8), with identical initial conditions, the Gaussian distribution characterized by (6.5) converges to (6.8), as $\epsilon \to 0$.

6.3 The Convergence of the Kalman Filters

We will prove in this section, as ϵ gets small, the distribution of the Kalman filter described in (6.5) converges to the filtered distribution of the homogenized process described in (6.8), given that the we take $Z(t) = z(t)$ in equation (6.8b).

As in the averaging case, before going to the main result, we first define some linear operators for an arbitrary symmetric matrix $\mathbf{m}$:

$$\begin{aligned}
\mathcal{L}_0\mathbf{m} &= \mathbf{a}_0\mathbf{m} + \mathbf{m}\mathbf{a}_0^* \\
\mathcal{L}_1\mathbf{m} &= \mathbf{a}_1\mathbf{m} + \mathbf{m}\mathbf{a}_1^* \\
\mathcal{L}_2\mathbf{m} &= \mathbf{a}_2\mathbf{m} + \mathbf{m}\mathbf{a}_2^* \\
\mathcal{L}^\epsilon\mathbf{m} &= \mathbf{a}\mathbf{m} + \mathbf{m}\mathbf{a}^* = \mathcal{L}_0\mathbf{m} + \frac{1}{\epsilon}\mathcal{L}_1\mathbf{m} + \frac{1}{\epsilon^2}\mathcal{L}_2\mathbf{m}
\end{aligned}$$

and for simplicity, denote

$$S = h^* \tau^{-1} h$$

which is positive definite symmetric.

Theorem 6.1. *Consider the scale separated Ornstein-Uhlenbeck system* (6.2)*, and let Assumptions 4.1 hold and only consider the scalar processes when* $d_1 = d_2 = 1$*. By feeding noisy observations* $\{z(s)\}_{0 \le s \le t}$ *from equation* (6.3) *to both the marginal filtered distribution of the multiscale system* $\mathcal{N}(m_x(s, z(s)), v_x(s))$ *and the filtered distribution of the homogenized equation* $\mathcal{N}(M(s, z(s)), V(s))$*, then for every* $1 \gg \epsilon > 0$*,*

(*i*) $\|v_x(s) - V(s)\| = \mathcal{O}(\epsilon^2)$*, for any* $s \in [0, t]$*, given initial condition* $v_x(0) = V(0)$*;*

(*ii*) $\mathbb{E}\left(m_x(s, z(s)) - M(s, z(s))\right)^2 = \mathcal{O}(\epsilon^2)$*, for any* $s \in [0, t]$*, given initial condition* $m_x(0, z(0)) = M(0, z(0))$.

Before going to the main result of this chapter, we first define some linear operators for an arbitrary symmetric matrix m:

$$\begin{aligned}
\mathcal{L}_0 m &= \mathbf{a}_0 m + m \mathbf{a}_0^* \\
\mathcal{L}_1 m &= \mathbf{a}_1 m + m \mathbf{a}_1^* \\
\mathcal{L}_2 m &= \mathbf{a}_2 m + m \mathbf{a}_2^* \\
\mathcal{L}^\epsilon m &= \mathbf{a} m + m \mathbf{a}^* = \mathcal{L}_0 m + \frac{1}{\epsilon} \mathcal{L}_1 m + \frac{1}{\epsilon^2} \mathcal{L}_2 m
\end{aligned}$$

and for simplicity, denote

$$S = h^* \tau^{-1} h \, .$$

Proof. Proof of statement *(i)* in Theorem 6.1.

Recall the definition for the variance (6.4a) written in simplified notation defined above

$$\begin{aligned}
\frac{dv}{dt} &= \mathbf{a} v + v \mathbf{a}^* - v h^* \tau^{-1} h v + \mathbf{q} \\
&= \mathcal{L}_0 v + \frac{1}{\epsilon} \mathcal{L}_1 v + \frac{1}{\epsilon^2} \mathcal{L}_2 - v S v + \left(\mathbf{q}_0 + \frac{1}{\epsilon} \mathbf{q}_1 \right)
\end{aligned} \tag{6.9}$$

By rearranging the equation above, we have

$$\mathcal{L}_2 v = -\mathbf{q}_1 + \epsilon^2 \left(\frac{dv}{dt} - \mathcal{L}_0 v + v S v - \mathbf{q}_0 \right) - \epsilon \mathcal{L}_1 v \tag{6.10}$$

By looking at the lower-right block of equation (6.10), we have

$$\begin{aligned}(a_{21}v_{12}+a_{22}v_{22})+(v_{21}a_{21}+v_{22}a_{22}) &= -q_2 \\ &+ \epsilon^2\left(\frac{dv_{22}}{dt}-0+(vSv)_{22}-0\right) \\ &- \epsilon\cdot 0\end{aligned}$$

Thus we have the solution of v_{22} as a function of v_{12},

$$v_{22} = -a_{22}^{-1}a_{21}v_{12} - \frac{1}{2}a_{22}^{-1}q_2 + \frac{\epsilon^2}{2}a_{22}^{-1}\left(\frac{dv_{22}}{dt}+(vSv)_{22}\right) \tag{6.11}$$

By looking at the top-right block of equation (6.10), we have

$$\begin{aligned}v_{11}a_{21}+v_{12}a_{22} &= \epsilon^2\left(\frac{dv_{12}}{dt}-(a_{13}v_{21}+a_{14}v_{22})+(vSv)_{12}\right) \\ &- \epsilon(a_{11}v_{21}+a_{12}v_{22})\end{aligned} \tag{6.12}$$

By substituting the solution of v_{22} into equation (6.12), and by $a_{11}-a_{12}a_{22}^{-1}a_{21}=0$ in Remark 4.2 and note that $\tilde{a}=a_{13}-a_{14}a_{22}^{-1}a_{21}$, we have,

$$\begin{aligned}v_{12} &= -\frac{a_{21}}{a_{22}}v_{11} \\ &+ \epsilon\frac{a_{12}}{2a_{22}^2}q_2 \\ &+ \epsilon^2\left(\frac{dv_{12}}{dt}-(\tilde{a}v_{12}-\frac{a_{14}q_2}{2a_{22}})\right) \\ &+ \epsilon^3\frac{1}{2a_{22}^2}\left(\frac{dv_{22}}{dt}+(vSv)_{22}\right) \\ &+ \epsilon^4\frac{a_{14}}{2a_{22}^2}\left(\frac{dv_{22}}{dt}+(vSv)_{22}\right)\ .\end{aligned}$$

Hence we have the solution for v_{12} as a function of v_{11},

$$v_{12} = -\frac{a_{21}}{a_{22}}v_{11} + \epsilon\frac{a_{12}q_2}{2a_{22}^2} + \epsilon^2 C\ . \tag{6.13}$$

We now try to find an ϵ-independent bound for C in the above equation. By writing the symmetric square matrices in vector forms, we define $w_v = a_{21}v_{11}+a_{22}v_{12}-\epsilon\frac{a_{12}}{2a_{22}}q_2 = \epsilon^2 a_{22}C$, and $\vec{b}=(a_{21},a_{22},0)$,

$$\vec{v}=\begin{pmatrix}v_{11}\\v_{12}\\v_{22}\end{pmatrix},\ \vec{q}_0=\begin{pmatrix}q_1\\0\\0\end{pmatrix},\ \vec{q}_1=\begin{pmatrix}0\\0\\q_2\end{pmatrix}.$$

We can write

$$\frac{dw_v}{dt} = \vec{b}\frac{d\vec{v}}{dt} = \vec{b}\left(\vec{a}_0\vec{v} + \frac{1}{\epsilon}\vec{a}_1\vec{v} + \frac{1}{\epsilon^2}\vec{a}_2\vec{v} + \vec{q}_0 + \frac{1}{\epsilon^2}\vec{q}_1 - \vec{F}(\vec{v})\right)$$

for which $\vec{F}(\vec{v})$ is the corresponding vectorized form of matrix vSv, and

$$\vec{a}_0 = \begin{pmatrix} 2a_{13} & 2a_{14} & 0 \\ 0 & a_{13} & a_{14} \\ 0 & 0 & 0 \end{pmatrix},\ \vec{a}_1 = \begin{pmatrix} 2a_{11} & 2a_{12} & 0 \\ 0 & a_{11} & a_{12} \\ 0 & 0 & 0 \end{pmatrix},\ \vec{a}_2 = \begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 2a_{21} & 2a_{22} \end{pmatrix}.$$

We find that for the fast scale terms with order $\mathcal{O}(\frac{1}{\epsilon^2})$,

$$\vec{b}\vec{a}_2\vec{v} = (a_{21}, a_{22}, 0)\begin{pmatrix} 0 & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 2a_{21} & 2a_{22} \end{pmatrix}\begin{pmatrix} v_{11} \\ v_{12} \\ v_{22} \end{pmatrix} = a_{22}w_v + \epsilon\frac{a_{12}}{2a_{22}}q_2\ ,$$

and

$$\vec{b}\vec{q}_1 = (a_{21}, a_{22}, 0)\begin{pmatrix} 0 \\ 0 \\ q_2 \end{pmatrix} = 0\ .$$

Therefore, we can write the differential equation for w_v as

$$\frac{dw_v}{dt} = \frac{a_{22}}{\epsilon^2}\left(w_v + \epsilon\frac{a_{12}q_2}{2a_{22}^2} + \epsilon^2\frac{\vec{b}}{a_{22}}\left(\vec{a}_0\vec{v} + \frac{1}{\epsilon}\vec{a}_1\vec{v} - \vec{F}(\vec{v}) + \vec{q}_0\right)\right).$$

Solving this nonhomogeneous first order linear equation, we find the solution

$$\begin{aligned} w_v(t) &= \epsilon\int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)}\left(\frac{a_{12}q_2}{2a_{22}^2} + \frac{\vec{b}}{a_{22}}\vec{a}_1\vec{v}(s)\right)ds \\ &+ \epsilon^2\int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)}\frac{\vec{b}}{a_{22}}\left(\vec{a}_0 v(s) - \vec{F}(\vec{v}) + \vec{q}_0\right)ds\ . \end{aligned}$$

Since the unconditional variance of the invariant measure of a multiscale OU system in the form of (4.1) is finite and independent of time and ϵ, we have $\lim_{t\to\infty}\mathbb{E}(x^2(t)) \le c$ and $\lim_{t\to\infty}\mathbb{E}(y^2(t)) \le c$, for c some finite constant independent of t and ϵ. Immediately, we also know that the covariance $\mathbb{E}(x(t)y(t)) \le \sqrt{\mathbb{E}(x^2(t))\mathbb{E}(y^2(t))} \le c$. Hence we know that the unconditional covariance matrix of the multiscale system is finite, and by the solution to an OU SDE as described in (2.2) and the invariant variance as in (2.4), we can immediately deduce that the invariant variance is the upper bound for the variance of $x(t)$ and $y(t)$. We know that $v(t)$ is the conditional covariance matrix of the coupled system

$(x(t), y(t))$ given $\{z(s)\}_{0\le s\le t}$. We know the conditional variance is always less than or equal to the unconditional variance, and since the unconditional variance is bounded above by the invariant variance, we conclude that $v(t) \le C$, for some C independent of t and ϵ.

Since $\vec{F}(\vec{v}(t))$ is a continuous quadratic function of $v(t)$, and $a_{22} < 0$, we can obtain an upper bound for the following integrals,

$$\begin{aligned} & \| \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} \frac{\vec{b}}{a_{22}} \left(\vec{a}_0 v(s) - \vec{F}(\vec{v(s)}) + \vec{q}_0 \right) ds \| \\ \le \quad & \frac{\vec{b}}{a_{22}} \sup_{0\le s\le t} \| \vec{a}_0 v(s) - \vec{F}(\vec{v(s)}) + \vec{q}_0 \| \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} ds \\ \le \quad & \epsilon^2 C(t) \, . \end{aligned}$$

$$\begin{aligned} & \| \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} \left(\frac{a_{12}q_2}{2a_{22}^2} + \frac{\vec{b}}{a_{22}} \vec{a}_1 \vec{v}(s) \right) ds \| \\ \le \quad & \sup_{0\le s\le t} \| \frac{a_{12}q_2}{2a_{22}^2} + \frac{\vec{b}}{a_{22}} \vec{a}_1 \vec{v}(s) \| \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} ds \\ \le \quad & \epsilon^2 C(t) \, . \end{aligned}$$

for which the $\|\cdot\|$ is a vector norm, and $C(t)$ is a constant for every t but ϵ-independent. Therefore, we proved that $w_v(t)$ is of order $\mathcal{O}(\epsilon)$, and consequently,

$$v_{12} = -\frac{a_{21}}{a_{22}} v_{11} + \epsilon \frac{a_{12}q_2}{2a_{22}^2} + \epsilon^3 C(t) \, .$$

where $C(t)$ is of order $\mathcal{O}(\frac{1}{\epsilon})$.

We finally look at the top-left block of equation (6.10),

$$\begin{aligned} \frac{dv_{11}}{dt} \quad = \quad & (a_{13}v_{11} + a_{14}v_{21}) + (v_{11}a_{13} + v_{12}a_{14}) \\ + \quad & \frac{1}{\epsilon} \left((a_{11}v_{11} + a_{12}v_{21}) + (v_{11}a_{11} + v_{12}a_{12}) \right) \\ + \quad & \frac{1}{\epsilon^2} \cdot 0 \\ - \quad & (v_{11}h_1\tau^{-1}h_1 + v_{12}h_2\tau^{-1}h_1)v_{11} - (v_{11}h_1\tau^{-1}h_2 + v_{12}h_2\tau^{-1}h_2)v_{21} \\ + \quad & q_1 \, . \end{aligned}$$

Substituting in the solution of v_{12} from (6.13), and considering the definitions of $\tilde{a}$,

$\tilde{h}$ and assumption $a_{11} - a_{12}a_{22}^{-1}a_{21} = 0$ in Remark 4.2, we have

$$\begin{aligned}
\frac{dv_{11}}{dt} &= (\tilde{a} + \epsilon^3 C(t))v_{11} + v_{11}(\tilde{a} + \epsilon^3 C(t)) \\
&+ \frac{a_{12}^2}{a_{22}^2} q_2 + \epsilon^3 C(t) \\
&+ 0 \\
&- v_{11}(\tilde{h} + \epsilon^3 C(t))\tau^{-1}(\tilde{h} + \epsilon^3 C(t))v_{11} \\
&+ q_1 \\
&= \tilde{a}v_{11} + v_{11}\tilde{a} - v_{11}\tilde{h}\tau^{-1}\tilde{h}v_{11} + \tilde{q} + \epsilon^3 C(t) \, .
\end{aligned}$$

Since $v_{11} = v_x$ where $C(t)$ in the above equation is of order $\mathcal{O}(\frac{1}{\epsilon})$. This equation converges to the equation for V in (6.8a). Noting that we are considering v_{11} and V to be scalars, we can show that

$$\begin{aligned}
\frac{d(v_x - V)}{dt} &= \frac{d(v_{11} - V)}{dt} \\
&= 2\tilde{a}(v_{11} - V) - \tilde{h}\tau^{-1}\tilde{h}(v_{11}^2 - V^2) + \epsilon^3 C(t) \\
&= \left(2\tilde{a} - \tilde{h}\tau^{-1}\tilde{h}(v_{11} + V)\right)(v_{11} - V) + \epsilon^3 C(t) \, .
\end{aligned}$$

The solution to the above equation is bounded as

$$\begin{aligned}
\|(v_x - V)(t)\| &\leq \|C(t)\epsilon^3 \int_0^t \sup_{0\leq s\leq t} \left(e^{2\tilde{a} - \tilde{h}\tau^{-1}\tilde{h}(v_{11}(s) + V(s))}\right) ds\| \\
&\leq \epsilon^3 C(t)\| \int_0^t e^{2\tilde{a}} ds\| \\
&\leq \epsilon^3 C(t) \, .
\end{aligned}$$

Notice $C(t)$ is of order $\mathcal{O}(\frac{1}{\epsilon})$ here, therefore,

$$\|(v_x - V)(t)\| = \mathcal{O}(\epsilon^2) \, .$$

The proof for statement *(i)* of Theorem 6.1 is complete. □

Proof. Proof of *(ii)* in Theorem 6.1.

We write m as a block matrix of its scalar entries,

$$m = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \, .$$

Recall the definition for the filtered mean (6.4b)

$$dm = \mathbf{a}mdt + (vh^*\tau^{-1})(dz - h^*mdt) \; . \tag{6.14}$$

Rearranging the above equation, we have

$$\mathbf{a}_2 m = \epsilon^2 \frac{dm}{dt} - \epsilon^2 \mathbf{a}_0 m - \epsilon \mathbf{a}_1 m - \epsilon^2 (vh^*\tau^{-1})(\frac{dz}{dt} - hm) \; .$$

By rewriting the above equation in block representation, we have a system of 2 equations, the equation for the lower block,

$$a_{21}m_1 + a_{22}m_2 = \epsilon^2 \frac{dm_2}{dt} - 0 - 0 - \epsilon^2 (vh^*\tau^{-1})(\frac{dz}{dt} - hm) \; .$$

We then have the solution of m_2 as a function of m_1

$$m_2 = -a_{22}^{-1} a_{21} m_1 + \epsilon^2 C \; . \tag{6.15}$$

We now try to find a bound for C in equation (6.15). We define $w_m = a_{21}m_1 + a_{22}m_2 = \epsilon^2 C$ and $\vec{b} = (a_{21}, a_{22})$. We can write the differential equation for w_m as

$$\frac{dw_m}{dt} = \vec{b}\frac{dm}{dt} = \vec{b}\left(\mathbf{a}_0 m + \frac{1}{\epsilon}\mathbf{a}_1 m + \frac{1}{\epsilon^2}\mathbf{a}_2 m + (v(t)h^*\tau^{-1})(\frac{dz}{dt} - hm)\right) \; .$$

We find that for the fast scale term with order $\mathcal{O}(\frac{1}{\epsilon^2})$,

$$\vec{b}\mathbf{a}_2 m = (a_{21}, a_{22}) \begin{pmatrix} 0 & 0 \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = a_{22} w_m \; .$$

Therefore, we can write the differential equation for w_m as

$$\begin{aligned} \frac{dw_m}{dt} &= \frac{1}{\epsilon^2}\vec{b}\mathbf{a}_2 m + \vec{b}\left(\mathbf{a}_0 m + \frac{1}{\epsilon}\mathbf{a}_1 m + (v(t)h^*\tau^{-1})(\frac{dz}{dt} - hm)\right) \\ &= \frac{1}{\epsilon^2} a_{22} \left(w_m + \epsilon^2 a_{22}^{-1}\vec{b}(\mathbf{a}_0 m + \mathbf{a}_1 m + (vh^*\tau^{-1})(\frac{dz}{dt} - hm))\right) \; . \end{aligned}$$

We substitute in the definition for $\frac{dz}{dt}$ from equation (6.3), we have

$$\frac{dw_m}{dt} = \frac{1}{\epsilon^2} a_{22} \left(w_m + \epsilon^2 a_{22}^{-1}\vec{b}(\mathbf{a}_0 m + \frac{1}{\epsilon}\mathbf{a}_1 m + (vh^*\tau^{-1})(h\mathbf{x} + \sqrt{\tau}\frac{dw}{dt} - hm))\right) \; .$$

where w is a Brownian motion.

Solving this nonhomogeneous first order linear equation, we have the solution,

$$\begin{aligned} w_m(t) &= \epsilon^2 \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} \left(\mathbf{a}_0 m(s) + (v(s)h^* \tau^{-1})(h\mathbf{x}(s) - hm(s)) \right) ds \quad (6.16) \\ &+ \epsilon \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} \mathbf{a}_1 m(s) ds \\ &+ \epsilon^2 \int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} (v(s)h^* \tau^{-1}) \sqrt{\tau} dw_s \ . \end{aligned}$$

Finding the L^2 bounds term by term above, we have

$$\begin{aligned} & \mathbb{E} \left(\int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} \left(\mathbf{a}_0 m(s) + (v(s)h^* \tau^{-1})(h\mathbf{x}(s) - hm(s)) \right) ds \right)^2 \\ \leq \ & t\mathbb{E} \left(\int_0^t e^{\frac{2a_{22}}{\epsilon^2}(t-s)} \left(C(\mathbf{a}_0 - v(s)h^* \tau^{-1} h)^2 m^2(s) + Cv(s)h^* \tau^{-1} h \mathbf{x}^2(s) \right) ds \right) \\ \leq \ & t\epsilon^2 \ . \end{aligned}$$

We have proved that $\mathbb{E} \sup_{0 \leq s \leq t} (\|x\|^2 + \|y\|^2)$ is bounded in Lemma 4.5, so $\mathbb{E}(\mathbf{x}^2(s))$ is bounded. Since the coupled system $(x(s), y(s))$ has invariant mean of zero, and finite invariant variance, we see that $\mathbb{E}m^2(s)$ is bounded. We have

$$\mathbb{E} \left(\int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} \left(\mathbf{a}_0 m(s) + (v(s)h^* \tau^{-1})(h\mathbf{x}(s) - hm(s)) \right) ds \right)^2 \leq \epsilon^2 C(t) \ ,$$

$$\mathbb{E} \left(\int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} \mathbf{a}_1 m(s) ds \right)^2 \leq \epsilon^2 C(t) \ ,$$

and

$$\mathbb{E} \left(\int_0^t e^{\frac{a_{22}}{\epsilon^2}(t-s)} a_{22}^{-1} \vec{b} (v(s)h^* \tau^{-1}) \sqrt{\tau} dw_s \right) \leq \epsilon^2 C(t) \ .$$

Hence, for equation (6.16)

$$\mathbb{E}(w_m(t))^2 \leq \epsilon^4 C(t) \ . \quad (6.17)$$

We then substitute the solution for m_2 from equation (6.15) and v_{11} into the upper

block of the rearranged equation (6.14)

$$\begin{aligned}
dm_1 &= (a_{13}m_1 + a_{14}m_2)dt \\
&+ \frac{1}{\epsilon}(a_{11}m_1 + a_{12}m_2)dt \\
&+ 0 \\
&+ (vh\tau^{-1})(dz - (h_1m_1 + h_2m_2)dt) \\
&= (\tilde{a}m_1 + a_{14}a_{22}^{-1}w_m(t))dt \\
&+ \frac{1}{\epsilon}(a_{12}a_{22}^{-1}w_m(t))dt \\
&+ ((v_{11} + \epsilon C(t))\tilde{h}\tau^{-1})(dz - \tilde{h}m_1(t) - h_2a_{22}^{-1}w_m(t))dt) \\
&= \tilde{a}m_1dt + (v_{11}\tilde{h}\tau^{-1})(dz - \tilde{h}m_1dt) \\
&+ a_{14}a_{22}^{-1}w_m(t)dt \\
&+ \frac{1}{\epsilon}(a_{12}a_{22}^{-1}w_m(t))dt \\
&+ \epsilon C(t)\tilde{h}\tau^{-1}(dz - (\tilde{h}m_1(t) - h_2a_{22}^{-1}w_m(t))dt) \\
&- v_{11}\tilde{h}\tau^{-1}h_2a_{22}^{-1}w_m(t)dt\ .
\end{aligned}$$

Notice $m_1 = m_x$, and by assumption, we have the same initial conditions $m_x(0, z(0)) = M(0, z(0))$, we have the difference,

$$\begin{aligned}
d(m_x(t, z(t)) - M(t, z(t))) &= \left(\tilde{a} - v_{11}\tilde{h}\tau^{-1}\tilde{h}\right)(m_x - M)dt \\
&+ a_{14}a_{22}^{-1}w_m(t)dt \\
&+ \frac{1}{\epsilon}(a_{12}a_{22}^{-1}w_m(t))dt \\
&+ \epsilon C(t)\tilde{h}\tau^{-1}(dz - (\tilde{h}m_1(t) - h_2a_{22}^{-1}w_m(t))dt) \\
&- v_{11}\tilde{h}\tau^{-1}h_2a_{22}^{-1}w_m(t)dt\ .
\end{aligned}$$

We have the solution,

$$\begin{aligned}(m_x - M)(t, z(t)) &= a_{14}a_{22}^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}w_m(s)ds \\ &+ \frac{1}{\epsilon}a_{12}a_{22}^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}w_m(s)ds \\ &+ \epsilon C(t)\tilde{h}\tau^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}dz \\ &- \epsilon C(t)\tilde{h}\tau^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}(\tilde{h}m_1(s) - h_2a_{22}^{-1}w_m(s))ds) \\ &- \tilde{h}\tau^{-1}h_2a_{22}^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}v_{11}(s)w_m(s)ds\ .\end{aligned}$$

In order to use Cauchy-Schwarz inequality, we use the solution (6.17) to find the magnitude of each term in the above equation.

$$\begin{aligned}&\mathbb{E}\left(a_{14}a_{22}^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}w_m(s)ds\right)^2 \\ \leq\ & C\mathbb{E}\left(\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}w_m(s)ds\right)^2 \\ \leq\ & C\mathbb{E}\left(\int_0^t e^{\tilde{a}(t-s)}w_m(s)ds\right)^2 \\ \leq\ & Ct\mathbb{E}\left(\int_0^t e^{2\tilde{a}(t-s)}w_m^2(s)ds\right) \\ \leq\ & Ct\epsilon^4\ .\end{aligned}$$

$$\begin{aligned}&\mathbb{E}\left(\frac{1}{\epsilon}a_{12}a_{22}^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}w_m(s)ds\right)^2 \\ \leq\ & \frac{C}{\epsilon^2}\mathbb{E}\left(\int_0^t e^{\tilde{a}(t-s)}w_m(s)ds\right) \\ \leq\ & \frac{C}{\epsilon^2}\epsilon^4 \\ \leq\ & C\epsilon^2\ .\end{aligned}$$

$$
\begin{aligned}
& \mathbb{E}\left(\epsilon C(t)\tilde{h}\tau^{-1}\int_0^t e^{(\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h})(t-s)}dz\right)^2 \\
\leq\ & \epsilon^2 C(t)\mathbb{E}\left(\int_0^t e^{\tilde{a}(t-s)}dz\right)^2 \\
\leq\ & \epsilon^2 C(t)\mathbb{E}\left(\int_0^t e^{\tilde{a}(t-s)}(h\mathbf{x}(s)ds+\sqrt{\tau}dw_s)\right)^2 \\
\leq\ & \epsilon^2 C(t)\left(C(t)+Ct\right) \\
\leq\ & C(t)\epsilon^2\ .
\end{aligned}
$$

Similarly

$$
\mathbb{E}\left(\epsilon C(t)\tilde{h}\tau^{-1}\int_0^t e^{\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h}}(\tilde{h}m_1(s)-h_2a_{22}^{-1}w_m(s))ds\right)^2 \leq \epsilon^2 C(t)\ ,
$$

and

$$
\mathbb{E}\left(\tilde{h}\tau^{-1}h_2a_{22}^{-1}\int_0^t e^{\tilde{a}-v_{11}(s)\tilde{h}\tau^{-1}\tilde{h}}v_{11}(s)w_m(s)ds\right)^2 \leq \epsilon^4 C(t)\ .
$$

Hence, by Cauchy-Schwarz inequality,

$$
\begin{aligned}
\mathbb{E}(m_x(t,z(t))-M(t,z(t)))^2 &\leq Ct\epsilon^4+C\epsilon^2+C(t)\epsilon^2+C(t)\epsilon^2+C(t)\epsilon^4 \\
&\leq C(t)\epsilon^2\ .
\end{aligned}
$$

where as $\epsilon\to 0$, the L^2 difference is small. This completes the proof for statement *(ii)* of Theorem 6.1. □

6.4 Numerical Example

We illustrate our idea in Theorem 6.1 through a simple numerical example. The system of interest is

$$
\begin{aligned}
d\mathbf{x} &= \mathbf{a}\mathbf{x}dt+\sqrt{\mathbf{q}}dW \quad , \quad x_0\sim N(0,I), && (6.18)\\
dz &= h\mathbf{x}dt+\sqrt{\tau}dw \quad , \quad z(0)=0 && (6.19)
\end{aligned}
$$

We observe $\{z(s)\}_{0\leq s\leq t}$, to model the conditional distribution of $x(t)$. We know this distribution is Gaussian, so we apply the Kalman filter to estimate the mean and variance of the distribution. We plot the marginal mean and variance of the filtered distribution

as described in (6.5). On the other hand, we model it through the homogenized system

$$dX \quad = \quad \tilde{a}Xdt + \sqrt{\tilde{q}}dU' \quad , \quad X(0) \sim (M_0.V_0) \tag{6.20}$$

We know the conditional distribution of $X(t)|z(t)$ is Gaussian, so we plot the mean and variance of this conditional distribution from the homogenized Kalman filter (6.8). In the numerical example below, the simulation of the SDE is via euler method without subsampling. We take the following values for the parameters,

$$\mathbf{a} = \begin{pmatrix} -\frac{1}{\epsilon} - 1 & -\frac{1}{\epsilon} + 1 \\ -\frac{1}{\epsilon^2} & -\frac{1}{\epsilon^2} \end{pmatrix}, \mathbf{q} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{2}{\epsilon^2} \end{pmatrix}, h = (1, \tfrac{1}{2}),$$

$\tilde{a} = -2$, $\tilde{q} = 2$, $\tau = 0.1$, $\epsilon = 2^{-7}$, $\delta = 2^{-15}$, $n = 2^{15}$.

In Figure 6.1, we see the actual path of slow process from the multiscale system (blue line) and the path of the homogenized process (red line) closely follow each other. This illustrates the convergence of the paths for averaged process stated in Theorem 4.6. The two standard deviation confidence intervals from the Kalman filters, both direct (cyan lines) and homogenized (green lines), closely follow each other. They both provide a good support for the actual path of the slow process x from the OU system.

In Figure 6.2, we plot the convergence of the Kalman filter variances. We see that the variance of marginal distribution $x|z$ (dashed cyan line) and $X|z$ (dashed green line) are almost the same. The speed of convergence of the variances are relatively slower compared to those from the averaged system, but it is very much convincing that the convergence and stability is well obtained. The variance of $y|z$ (dash-dotted cyan line) is also plotted for illustration.

In Figure 6.3, we plot the squared error $(m_x(t) - M(t))^2$, from which we see that the errors are controlled at below the order of 10^{-4}, and are still decreasing, which is in line with our expectation of an order $\mathcal{O}(\epsilon^2) = 2^{-14} \approx 10^{-4.2}$.

In Figure 6.4, we plot the distance $\|v_x(t) - V(t)\|$, from which we see that the error is rough of order 10^{-3}, and our expectation is that it is of order $\mathcal{O}(\epsilon^2) = 2^{-14} \approx 10^{-4.2}$.

We conclude this is a good evidence to support Theorem 6.1. The advantage of this application of Kalman filter on homogenized process is the same as those discussed in Section 5.4.

6.5 Conclusion

In this chapter, we tried to integrate the multiscale method of homogenization and Kalman filtering. We have proved that the slow part of the marginal filtered distribution of the multi-scale system converges to the filter distribution of homogenized process, given observations

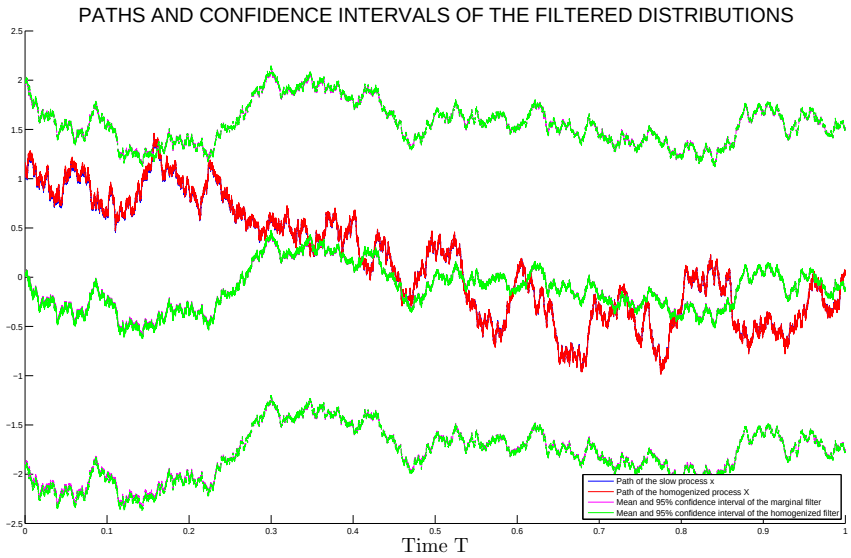

Figure 6.1: Paths and 95% confidence intervals of the filtered distributions

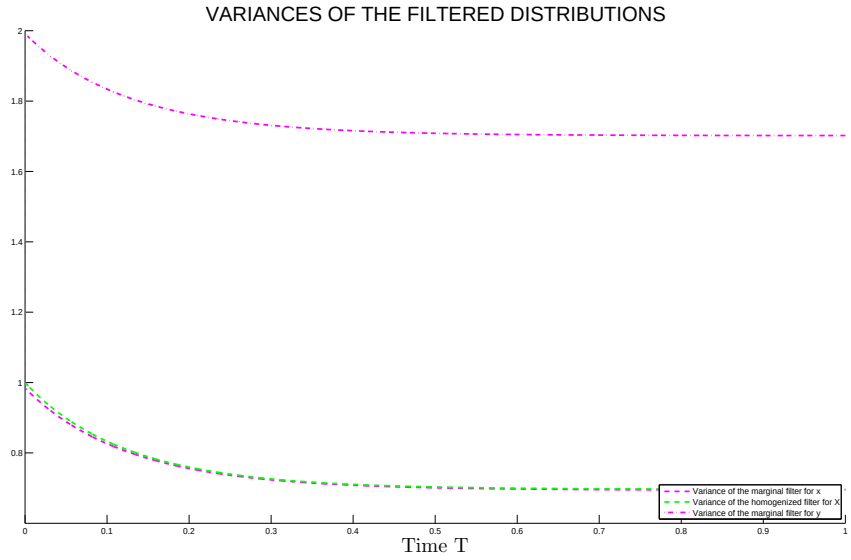

Figure 6.2: Variances of the filtered distributions

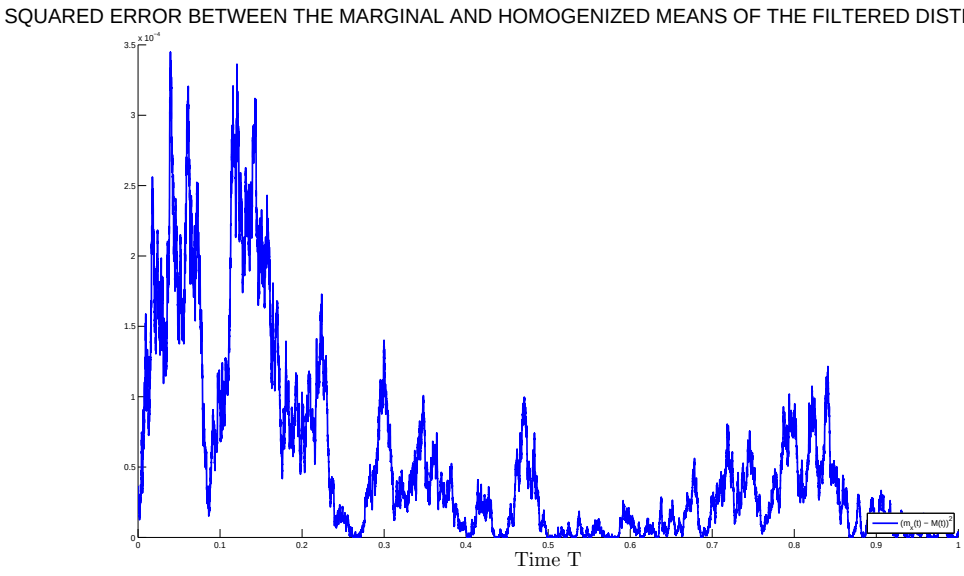

Figure 6.3: Squared error between the means of the marginal and homogenized filtered distributions

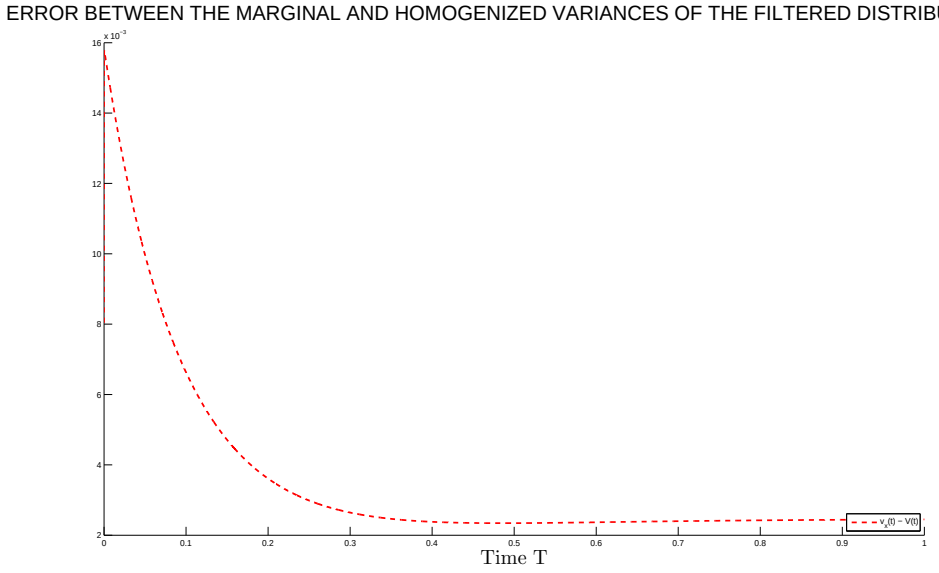

Figure 6.4: Error between the variances of the marginal and homogenized filtered distributions

from the multiscale system contaminated with Gaussian noise, and ϵ is small. Unlike the parameter estimation for the homogenized OU system in Chapter 4, here we observe no need for subsampling. In summary,

- We take observations from the noisy observation (6.3);
- we use these observation to substitute into the mean and variance of the Kalman filter derived from the homogenized process (6.8);
- we show that the Gaussian distributions characterized by (6.8) and (6.5) closely follow each other.

Similar to the result achieved in averaging, in Theorem 6.1, we proved the convergence between the marginal distribution of the Kalman filter applied on the whole coupled system and the distribution of that Kalman filter applied on the homogenized process, when the coupled system is made of two scalar OU processes. However, we expect similar result to hold for any finite d_1 and d_2.

Part III

EXTERNAL PROJECTS

In this part of the thesis, we will study applications of two types of stochastic models in real world. We will study how a standard vector autoregressive model is developed to a heteroscedastic Bayesian autoregressive model when fitting into a time series data in Chapter 7. Then we will study the Generalized Autoregressive Conditional Heteroscedastic (GARCH) model, equipped with a constant and a time dependent correlation matrix, for a portfolio of 8 asset classes in Chapter 8.

The recent financial crisis has significantly challenged the assumptions and applicabilities of some of the empirically tested stochastic models. In particular, mean reverting models such as vector autoregressive models were preferred by econometric researchers and monetary policy makers. However, they no longer show convincing arguments in simulating future economic scenarios, since the volatile and swiftly changing economic environment do not adhere to models with constant parameters anymore. Bayesian vector autoregressive models have come of more popularity due to its "self updating" ability and its autoregressive based property. We will discuss about the application of Bayesian autoregressive models to the time series of interest rate factors in the Nelson-Siegel yield curve model in Chapter 7. Along the development of the model from standard vector autoregressive model to the hierarchical heteroscedastic regression model, we will incorporate the numerical method of Markov Chain Monte Carlo, Gibbs sampler in specific. Comparison of the models will be based on mean forecasts based on each model specification. Since the model is essentially targeted at forecasting, we fix the estimated parameters on their mean values, while ignoring the uncertainty band for the estimated parameters.

Also due the unexpected and rapid change in the recent economic atmosphere, vulnerability to risks (volatility) has come to the core concern when making investment decisions. Inter-governmental institutions, such as the International Monetary Fund, are especially sensitive to economic and market volatilities. The Generalized Autoregressive Conditional Heteroscedastic (GARCH) Model has been a popular and essential tool in modelling volatility, we study two types of multivariate GARCH model, the Constant Correlation and Dynamic Conditional Correlation GARCH models, in Chapter 8. The Constant Correlation GARCH assumes a constant correlation matrix among assets class, and the Dynamic Conditional Correlation GARCH model assumes a time dependent correlation matrix. We will compare the two setups of the GARCH model by fitting them into our data. Comparing to our focus on mean estimates in Chapter 7, we concentrate on the distribution in 8. This is due to the target of this modelling project is to analyze the portfolio risk structure, the aim is to replicate the volatility and the cross asset correlation time series according to historical data.

All models in this part are coded in MATLAB. We thank the publicly licenced "Econometric Tool Box" developed by James LeSage, et al [55] for the great efficiency

and convenience it provides to the implementation of our modelling process. I would also like to thank the International Monetary Fund for providing the opportunity and support for these projects.

Chapter 7

Bayesian Vector Autoregressive Models

Vector Autoregressive (VAR) models are of particular interest to economic researchers and policy makers. This is due to the mathematical simplicity and proven credibility of regression models. Autoregressive models, specifically with short memories, are usually mean reverting, and trend deterministic under normal conditions. This is true, for an economic statistic, across (usually very) long time horizon. However, the recent economic recession presented a distorted picture for the economic scenario based on autoregressive models. For this reason, we apply Bayesian Vector Autoregressive (BVAR) models and their derivatives to our problem of simulating the Nelson-Siegel factors. The methods of Bayesian Vector Autoregressions are discussed in detail in [59, 88].

In section 7.1 we introduce the Nelson-Siegel yield curve model, from which the time series of factors are taken. From section 7.2 to section 7.5, we discuss how the time series model for the Nelson-Siegel interest rate factors (ie. the long, medium and short term interest rate parameters of Nelson-Siegel model) can be developed from the standard VAR(1) to the Bayesian heteroscedastic regression model. In section 7.6, we discuss the convergence tests for the MCMC samplers, which take important place in the Gibbs sampling for Bayesian VAR model studied in section 7.4, and the Bayesian heteroscedastic regression model studied in section 7.5. Finally, in section 7.7, we discuss the results from each model in the development process, the reason a specific model is preferred, and the convergence test results if the model is MCMC based. For the MCMC based models in sections 7.4 and 7.5, though we will sample the entire the posterior space for the parameters of interest, we will focus on the mean estimates of the parameters since our aim is to implement those estimated models to simulate data, therefore, we will not discuss the uncertainty bands of the estimated parameters.

7.1 The Nelson-Siegel Factors

The Nelson-Siegel model [23, 24, 67] is the most widely used model for yield curve modelling of government bonds (ie. US Treasury bills) by central banks and monetary policy researchers. It takes the form

$$\text{Yield}(t) = y_0(t) + y_1(t)\frac{1-e^{-\lambda\tau}}{\lambda\tau} + y_2(t)\left(\frac{1-e^{-\lambda\tau}}{\lambda\tau} - e^{-\lambda\tau}\right) \tag{7.1}$$

where λ is a decay factor, which is assumed to be a constant across the time horizon due to its low volatility; τ is the bond maturity parameter; $y(t) = (y_0(t), y_1(t), y_2(t))$ is a three dimensional time series which the following models are based on.

In economic terms, y_0 is interpreted as the long run level of interest rates, y_1 is the short term component and y_2 for the medium term. They are usually fitted via least-squares or similar algorithms (see [23]).

In our problem, historical monthly $y(t)$ is supplied. It is the vector valued time series we try to model and simulate. Since short memory models are preferred by the policy makers, we restrict our models to have lag length of 1 (month). We also assume there is no cross-dependence between components of $y(t)$ at the same time due to observational restrictions, which means we can not forecast $y_i(t)$ given $y_j(t)$, $i \neq j$. The modelling scheme takes four steps. We start from the standard VAR(1) in section 7.2, then introduce the standard BVAR(1) in section 7.3, and its hierarchical Monte Carlo derivative in section 7.4. Finally, we modify the Monte Carlo BVAR(1) as a heteroscedastic linear regression model to give our best forecasting result in section 7.5.

We have time series data $y(t)$ on monthly frequency, ranging from August 1988 to April 2010 (as in Figure 7.7), and given λ and a range of different maturities τ. Our aim is to forecast $y(t)$ for 60 periods (ie. 5 years) from May 2010 using regression based models. We assume a constant λ for the entire 5-year simulation period, which only takes part when we are converting the simulated $y(t)$ to yield curve forecasts using equation (7.1). The Nelson-Siegel model (7.1) is only used as a map to convert the factors $y(t)$ to yields, and is irrelevant to the model development process discussed below.

7.2 Model 1: The Vector AR(1) process

As a start, we regress the time series data $\{y(t)\}_{0\leq t\leq N}$ against itself with order one time lag, using the standard Vector AR(1) model

$$Y = LY\Phi + \eta \tag{7.2}$$

where the data $Y = \{y(t)\}_{1 \leq t \leq N} \in \mathbb{R}^{N \times k}$, with L as the lag operator[1]; $\Phi \in \mathbb{R}^{k \times k}$; $\eta \in \mathbb{R}^{N \times k}$ is the random Gaussian error. The variable $N + 1$ takes the value 261, and is the number of equidistantly observed historical $y(t)$, ranging from August 1988 to April 2010; $k = 3$ is the number of variables. We want to estimate the parameter Φ.

By estimating the parameter Φ through ordinary least squares, we can make mean forecasts using the model. The mean forecasted $y(t)$ and selected yield curves are plotted in Figure 7.8. The results will be discussed in Section 7.7.

For simplicity of notation in later steps, we decompose the above equation as

$$y_i = LY\phi_i + \epsilon_i \tag{7.3}$$

where $i \in \{0, 1, 2\}$. The vector y_i represents a column vector of $Y = (y_0, y_1, y_2)$ in equation (7.2), and ϕ_i a corresponding column vector in $\Phi = (\phi_0, \phi_1, \phi_2)$, respectively. ϵ_i is a Gaussian random vector with mean zero and covariance matrix Σ_i. For the sake of simplicity, we omit the column indices i of y_i and ϕ_i. We write the regression model in this form to avoid the unnecessary complication when describing the variance of a matrix random variable, which is a three-dimensional tensor. This simplification only results a change of notation, it does not affect the correlation among ϕ_i's, nor ϵ_i's.

7.3 Model 2: The standard BVAR(1)

We now implement the standard Bayesian vector autoregressive model with lag 1 based on the VAR(1) set up (7.3) by adding more randomness to the parameters. We assume ϕ (shorthand notation for ϕ_i's in (7.3)) follow a distribution classified by the Bayesian framework described below. For Bayesian models, the choice of appropriate priors is essential. There are many choices of priors for BVAR estimation, they are studied extensively in [59, 85, 88]. However, we choose the non-informative Litterman prior, or Minnesota prior, for our standard BVAR(1) application.

The Litterman prior is the most classical and most widely used non-informative prior in BVAR applications. When implementing a BVAR model with the absence of prior knowledge of the time series, external information enters each equation marginally and deviates as time lags increase. The prior is not derived from any explicit economic theory but purely on common belief [59, 60], which assumes that current state has higher dependence with immediate past than the past further back in time. This property shows a very weak assumption is required to implement this prior, and a wide scope of problems this prior is able to adapt to. This advantage motivates our construction of the BVAR model. In the Lit-

[1]For our problem, the lag operator is defined as $LY = \{y(t)\}_{0 \leq t \leq N-1}$

terman prior, the parameters are assumed to be independent from each other. All parameters are assumed to have mean of zero except the coefficient on the first lag of the dependent variable, and with standard deviation decreasing as time lags increase.

The Litterman **prior** assumes a Gaussian distribution for ϕ

$$\phi \sim \mathcal{N}(\phi_0, V_0) \tag{7.4}$$

where the prior mean $\phi_0 = (0, \ldots, 0, 1, 0, \ldots, 0)^*$, with the i^{th} entry valued 1, and all other entries valued zero[2], which reflects the belief that the ith variable has an expected value equal to its immediate past, but not other variables.

The prior covariance takes the form:

$$V_0 = V_0^{\frac{1}{2}} \left(V_0^{\frac{1}{2}}\right)^* \tag{7.5}$$

$$\left(V_0^{\frac{1}{2}}\right)_{ij} = \theta l^{-\psi} w(i,j) \frac{\hat{\sigma}_i}{\hat{\sigma}_j}$$

where $\theta \in (0,1)$ is a parameter describing overall tightness of the variance. $l^{-\psi}$ is a lag decay function at lag l with rate $\psi \in (0,1)$. The function describes the shrinkage of the standard deviation with increasing lag length. In our problem $l = 1$, which gives a constant lag decay function. $w(i,j)$ a weight function describes the tightness of the prior for variable j in equation i of the system, relative to the tightness of its own lags of variable i in equation i. w is a symmetric matrix, which is chosen as a common preference [55]

$$w = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{pmatrix}$$

The diagonal 1s mean variable i has a larger weight in equation i, which takes a bigger role in describing itself, and other variables in this equation are equally less weighted. $\hat{\sigma}_i$ is the estimated standard error from a univariate autoregression for variable i. It is pre-estimated using $y_i = Ly_i\tilde{\phi} + \tilde{\epsilon}_i$, where the variance of $\tilde{\epsilon}$ is $\hat{\sigma}^2$. $\hat{\sigma}_i/\hat{\sigma}_j$ acts as a scaling factor adjusts for varying magnitude of the variables across equations i and j. The construction of V_0 through the factorization setup eases sampling of ϕ since it is assumed to follow a multivariate Gaussian distribution.

[2]For our problem, ϕ_0 takes one of the following values: $(1,0,0)^*$, $(0,1,0)^*$ or $(0,0,1)^*$.

We have the **likelihood** function for ϕ from the model (7.3)

$$L(\phi|Y) \propto \det(\Sigma)^{\frac{1}{2}} \exp\left(-\frac{1}{2}(y-(LY)\phi)^*\Sigma^{-1}(y-(LY)\phi)\right) \tag{7.6}$$

Hence we find the **posterior** by completing the square for the product of the prior (7.4) and the likelihood (7.6),

$$p(\phi|Y) \propto \exp\left(-\frac{1}{2}(\phi-\hat{\phi}_{Bayesian})^*\hat{V}^{-1}_{Bayesian}(\phi-\hat{\phi}_{Bayesian})\right) \tag{7.7}$$

where $\hat{\phi}$ and $\hat{V}$ are the maximum likelihood estimators of the posterior mean and variance for parameter ϕ. Therefore ϕ is updated by the lag 1 data LY according to

$$\hat{\phi}_{Bayesian} = \left(V_0^{-1} + (LY)^*\Sigma^{-1}(LY)\right)^{-1}\left(V_0^{-1}\phi_0 + (LY)^*\Sigma^{-1}y\right)$$

$$\hat{V}_{Bayesian} = \left(V_0^{-1} + (LY)^*\Sigma^{-1}(LY)\right)^{-1}$$

In summary, for the BVAR(1) model, we have the following framework:

- We want to estimate the parameter ϕ in (7.3), by maximizing it's posterior density;
- we use data $Y = (y_0(t), y_1(t), y_2(t))_{0\le t\le N}$ as input, segment it to the dependent variable $\{y_i\}_{1\le t\le N}$ for each run of the model for ϕ_i, and the independent variables $LY = (y_0(t), y_1(t), y_2(t))_{0\le t\le N-1}$;
- the prior density for ϕ as given by (7.4);
- the likelihood $L(\phi|Y)$ as given by (7.6);
- the posterior $p(\phi|Y)$ as given by (7.7).

We use the estimated ϕ to make 60-period mean forecasts of $y(t)$ using equation (7.3). The selected yield curves are shown in Figure 7.9. The results will be discussed in Section 7.7.

7.4 Model 3: BVAR(1) with Gibbs sampling

When the Litterman prior is chosen as a non-informative prior, the posterior distributions can be obtained in a closed form, the necessary estimation can be analytically obtained. However, in the case an informative prior is preferred and can be reasonably supplied, we may opt for the Gibbs sampling for the BVAR model since analytical results in Model 2

cannot be obtained here. Gelfand and Smith [29] proved that, if the functional form of the joint density of the random observations is known, using Gibbs sampler to draw a sequence of complete conditional distributions for all of its parameters, will converge in the limit to the true joint posterior distribution of the parameter.

Recall the setup in (7.3),

$$y = LY\phi + \epsilon \tag{7.8}$$

where y and (LY) retain the same property, but assume $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ is a vector of independently identically distributed Gaussian random errors. In the case we want to implement the hierarchical Gibbs sampled BVAR(1) model to estimate parameters (ϕ, σ), we first need some more assumptions.

Assumptions 7.1.
We assume, for the system (7.3)*:*

(i) *the prior density $\pi_\phi(\phi)$ for parameter ϕ, and the prior density $\pi_\sigma(\sigma)$ for parameter σ are independent, ie. the joint prior density $\pi(\phi, \sigma) = \pi_\phi(\phi)\pi_\sigma(\sigma)$;*

(ii) *$\pi_\phi(\phi)$ is Gaussian, for which the mean and variance can be expressed as linear combinations of ϕ;*

(iii) *we assume a non-informative diffuse prior for $\pi_\sigma(\sigma) \propto \frac{1}{\sigma}$, which is a continuous uniform distribution, with $\sigma > 0$.*

For Assumption 7.1 (ii), we mean there exists R such that

$$\pi_\phi(\phi) \propto \exp\left(-\frac{1}{2}\left(R\phi - r\right)' \Gamma^{-1} \left(R\phi - r\right)\right) \tag{7.9}$$

from which it is possible that $R \in \mathbb{R}^{m\times k}$ with $m \leq k$, and r and Γ are the prior mean and covariance of R, respectively. In the event $m \leq k$, the prior $\pi_\phi(\phi)$ in (7.11) is improper, since there are fewer number of equations than the number of variables. However, we can seek an alternative by factorising $\Gamma^{-1} = \tau'\tau$, and $r_\tau = \tau r$. This gives

$$\pi_\phi(\phi) \propto \exp\left(-\frac{1}{2}\left(\tau\phi - r_\tau\right)' \left(\tau\phi - r_\tau\right)\right). \tag{7.10}$$

Therefore, we have the joint **prior** density

$$\pi(\phi, \sigma) = \pi_\sigma(\sigma)\pi_\phi(\phi), \tag{7.11}$$

and the **likelihood** function

$$L(\phi, \sigma|Y) \propto \sigma^{-(N-1)} \exp\left(\frac{1}{2\sigma^2}\left(y-(LY)\phi\right)^*\left(y-(LY)\phi\right)\right), \tag{7.12}$$

to get the **posterior**

$$p(\phi, \sigma) \propto \sigma^{-N} \exp\left(\left(\phi-\hat{\phi}_{gibbs}\right)^* \hat{V}_{gibbs}^{-1}\left(\phi-\hat{\phi}_{gibbs}\right)\right), \tag{7.13}$$

which gives the maximum likelihood estimators

$$\hat{\phi}_{gibbs} = \left((LY)^*(LY)+\sigma^2\tau^*\tau\right)^{-1}\left((LY)^*y+\sigma^2\tau^*r_\tau\right), \tag{7.14}$$

and

$$\hat{V}_{gibbs} = \sigma^2\left((LY)^*(LY)+\sigma^2\tau^*\tau\right)^{-1}. \tag{7.15}$$

It is observed that the MLEs $\hat{\phi}_{gibbs}$ and $\hat{V}_{gibbs}$ depend on σ, which prevents it to have an analytical solution as in the standard BVAR model with Litterman prior. Theil and Goldberger [90] proposed that σ^2 to be initially estimated by the least squares method $\hat{\sigma}^2 = (y-(LY)\phi)^*(y-(LY)\phi)/N-k$.

The Gibbs sampler samples ϕ from the multivariate conditional Gaussian posterior distribution (7.13), with conditional mean (7.14) and variance (7.15), and samples σ from the posterior density

$$p(\sigma|\phi) \propto -\sigma^{-N}\exp\left(-\left(y-(LY)\phi\right)^*\left(y-(LY)\phi\right)\right)$$

which is identical to

$$\left(y-(LY)\phi\right)^*\left(y-(LY)\phi\right)/\sigma^2|\phi \sim \chi^2(N-1)$$

In summary, for the BVAR(1) model with Gibbs sampling, we have the following framework:

- We want to estimate the coefficient ϕ, and σ, the standard deviation for the error ϵ in (7.3), by maximizing their joint posterior density;
- we use data $Y = (y_0(t), y_1(t), y_2(t))_{0\le t\le N}$ as input, segment it to the dependent variable $\{y_i\}_{1\le t\le N}$ for each run of the model for ϕ_i, and the independent variables $LY = (y_0(t), y_1(t), y_2(t))_{0\le t\le N-1}$;
- the joint prior density for (ϕ, σ) is given by (7.11);
- the likelihood function $L(\phi, \sigma|Y)$ is given by (7.12);

- and the joint posterior $p(\phi, \sigma|Y)$ is given by (7.13).

Convergence tests of the MCMC method is monitored using various methods, which are discussed in Section 7.6. The 60-month mean forecasts of $y(t)$ made by the BVAR(1) with Gibbs sampler and selected yield curves are shown in Figure 7.10. The results will be discussed in Section 7.7.

7.5 Model 4: Bayesian heteroscedastic regression and Gibbs sampling

We now refer to the spatial autoregressive models from a Bayesian perspective. The method is discussed extensively in [56, 57].

The model (7.3) is considered as a case of heteroscedastic linear regression model with an informative prior.

$$y = LY\phi + \epsilon$$

with

Assumptions 7.2.
We assume:

(*i*) ϵ *follows a Gaussian distribution* $\mathcal{N}(0, \sigma^2 V)$*;*

(*ii*) *the variance of the error takes a diagonal form* $V = \mathrm{diag}(v_1, v_2, \ldots, v_N)$*;*

(*iii*) ϕ *has a Gaussian prior* $\mathcal{N}(\bar{\phi}, T)$*;*

(*iv*) σ *has a flat diffuse prior* $(1/\sigma)$*, which is a uniform distribution with* $\sigma > 0$*;*

(*v*) *each* v_i *has a prior* r/v_i *follows independent* $\chi^2(r)$ *prior distributions;*

(*vi*) r *follows a prior* $\Gamma(m, k)$.

In this model, y and LY are the same as defined in (7.3). In addition to the previous model, we assume, that ϵ has a non-constant variance. The relative variance $(v_1, v_2, \ldots, v_N)$ are assumed to be fixed at each discrete observational time, they are unknown parameters need to be estimated. Bayesian methods avoid the constrains from a degrees-of-freedom perspective when estimating N parameters $(v_1, v_2, \ldots, v_N)$ of V, and the $k + 1$ parameters of ϕ and σ, using N data observations, since we can rely on an informative prior for the V parameters. The prior r/v_i takes the form of $\chi^2(r)$ distribution as described in Assumption 7.2 *(v)*. This type of prior was first introduced by Lindley in [58]

as cell variances in the analysis of variance with multiple observations per cell. It has been discussed in [30] as well.

The prior assigned to each v_i can be understood as a distribution with unity mean and variance $2/r$. As r becomes large, all v_i will approach unity, giving an identity matrix for V, believing that outliers and non-constant variances do not exist; when V does not equal identity, it makes the model more robust to outliers and observations with large variances by assigning less weight to these observations.

Following the usual Bayesian methodology, we have the likelihood function,

$$L(\phi, \sigma, V|Y) \propto \sigma^{-(N-1)} \left(\prod_{i=1}^{N} \sqrt{(}v_i) \right)^{-1} \exp\left(\frac{1}{2\sigma^2} (y - (LY)\phi)^* V^{-1} (y - (LY)\phi) \right). \tag{7.16}$$

we then compute the posterior using the priors and likelihood function analytically, however, the posterior density may not be tractable. We seek Gibbs sampling to derive the posterior distribution.

We will use the following mutually conditional distributions for ϕ, σ and V to do our sampling.

$$\begin{aligned} \phi|(\sigma, V) &\sim \mathcal{N}\left(H\left((LY)^* V^{-1} y + \sigma^2 T^{-1} c\right), \sigma^2 H\right) && (7.17) \\ H &= \left((LY)^* V^{-1} (LY) + T^{-1}\right)^{-1} \\ \left[\frac{\sum_{i=1}^{N} (e_i^2/v_i)}{\sigma^2}\right] |(\phi, V) &\sim \chi^2(N) && (7.18) \\ \left[\frac{\sigma^{-2} e_i^2 + r}{v_i}\right] |(\phi, \sigma) &\sim \chi^2(r+1) && (7.19) \end{aligned}$$

where $e_i = y_i - (LY)_i^* \phi$.

The Gibbs sampling takes the following steps repeatedly:

(i) Start with arbitrary choices of initial values ϕ^0, σ^0 and v_i^0;

(ii) Sample ϕ^1 using (7.17) conditional on σ^0 and v_i^0;

(iii) Sample σ^1 using (7.18) conditional on ϕ^1 and v_i^0;

(iv) Sample v_i using (7.19) conditional on ϕ^1 and σ^1.

In summary, for the Bayesian heteroscedastic regression model with Gibbs sampling, we have the following framework:

- We want to sample the parameters (ϕ, σ, V), by doing Gibbs sampling iteratively;

```
MCMC CONVERGENCE diagnostics
Based on sample size =       1000
Autocorrelations within each parameter chain
Variable        Lag 1        Lag 5       Lag 10       Lag 50
phi 0          -0.023        0.000        0.035        0.003
phi 1           0.026       -0.003        0.039       -0.067
phi 2          -0.003        0.026       -0.023       -0.014
```

Figure 7.1: Autocorrelation Diagnostics for Model 3

- we use data $Y = (y_0(t), y_1(t), y_2(t))_{0 \le t \le N}$ as input, segment it to the dependent variable $\{y_i\}_{1 \le t \le N}$ for each run of the model for ϕ_i, and the independent variables $LY = (y_0(t), y_1(t), y_2(t))_{0 \le t \le N-1}$;
- the priors for each parameter are specified through a set of distributions, which are set out in Assumptions 7.2;
- the likelihood function is given in (7.16);
- the posteriors are specified iteratively in (7.17), the updating scheme follows the algorithm in 7.5.

Convergence tests of the MCMC method is monitored using various methods, which are discussed in Section 7.6. The 60-month mean forecasts of $y(t)$ from the heteroscedastic regression model and Gibbs sampling and selected yield curves are shown in Figure 7.11. The results will be discussed in Section 7.7.

7.6 Convergence Tests of the MCMC Samplers

For the Gibbs sampler models 3 and 4 discussed in Section 7.4 and 7.5, we assess the convergence of the sampler to the posterior distributions. For both of the samplers in Section 7.4 and 7.5, we used 1100 draws for each step, and discarded the first 100 as burn-in. We use the rest of the 1000 samples to be tested for convergence. Among various MCMC convergence diagnostics, we focus on the following diagnostic tests: autocorrelation, Raftery & Lewis, and Geweke's diagnostics. The diagnostic measures are explained in [43].

Autocorrelation is the most common approach to measure dependency among Markov Chain samples. Autocorrelation with lags 1, 5, 10 and 50 have been computed (see Figures 7.1 and 7.2), we find all of them being very small, which suggest the samples are well mixed.

Raftery & Lewis [32, 82, 83] diagnostics provides a practical tool for finding the minimum sample size required to reach a desired level of posterior distributional accuracy in terms of percentiles. We take the quantile of $q = 0.025$ to be our interest, with a precision

```
MCMC CONVERGENCE diagnostics
Based on sample size =       1000
Autocorrelations within each parameter chain
Variable        Lag 1        Lag 5       Lag 10       Lag 50
phi 0           0.049       -0.051       -0.038        0.002
phi 1           0.004       -0.034        0.011       -0.012
phi 2          -0.007        0.003        0.039        0.010
```

Figure 7.2: Autocorrelation Diagnostics for Model 4

```
Raftery-Lewis Diagnostics for each parameter chain
(q=0.0250, r=0.010000, s=0.950000)
Variable       Thin       Burn   Total(N)     (Nmin)      I-stat
phi 0             1          2        969        937       1.034
phi 1             1          2        969        937       1.034
phi 2             1          2        969        937       1.034
```

Figure 7.3: Raftery & Lewis test for Model 3

level of $r = 0.01$ associated with a probability $s = 95\%$. In Figures 7.3 and 7.4, the resulting statistics suggest that a total number of 969 draws are required to achieve the desired accuracy of $r = 0.01$ on the proposed 0.025 percentile estimation; and 937 draws are required if the draws are from an iid chain. Our chain consists of 1000 effective draws, which exceeds both of these requirements.

Geweke's [31] diagnostics tests if the mean estimates converges. It compares the means from the early and latter part of the Markov Chain. There are two groups of summary statistics been produced for the Geweke's diagnostics. The first group of statistics titled "Geweke Diagnostics for each parameter chain" shows the estimates of the numerical standard error (NSE) and relative numerical efficiency (RNE). RNE provides an indication of the number of draws that would be required to produce the same numerical accuracy if the draws had been sampled independently from the posterior distribution. The test produces estimates of iid chain and truncation of the periodgram window at 4%, 8% and 15%. The NSE and RNE based on an iid process provides a sample about the statistics. The 4%, 8% and 15% NSEs and RNEs do not base on iid assumption of the process. If they are significantly different, then we tend to believe in the non-iid nature, however, in our case

```
Raftery-Lewis Diagnostics for each parameter chain
(q=0.0250, r=0.010000, s=0.950000)
Variable       Thin       Burn   Total(N)     (Nmin)      I-stat
phi 0             1          2        969        937       1.034
phi 1             1          2        969        937       1.034
phi 2             1          2        969        937       1.034
```

Figure 7.4: Raftery & Lewis test for Model 4

in Figures 7.5 and 7.6, corresponding statistics are close to each other, which suggests a well-mix of our samples.

The second group of the statistics titled "Geweke Chi-squared test for each parameter chain" shows if the sample draws have reached an equilibrium based on the means of the first 20% and the last 50% of the sample. If the sample Markov Chain has reached an equilibrium, the means of the two portions of the sample should be roughly equal. From the results in Figures 7.5 and 7.6, we see that the means of the parameters are close enough to indicate a good convergence.

Therefore, the conclusion arise as both of the MCMC samplers: the Bayesian VAR(1) model with Gibbs sampler (Model 3) and the Heteroscedastic Regression Model (Model 4) have reach equilibrium with our required accuracy level given our sample size in the simulation.

7.7 Discussion of the Results

Before we make our choices on the model, we outline the fundamental assumptions we make on yield curve. We make two assumptions (or beliefs) on the behaviour of our targeting yield curves: 1) the yield curves being simulated are normal yield curves[3]. Normal yield curves always have a positive derivative in time. It reflects a rational economic expectation that the economy will grow in the future, and the associated inflation will rise in the future rather than fall. The positive derivative is also associated with the risks posed to the uncertainty of future inflation rate and the value of scheduled cash flows, which is compensated by higher yields for longer maturity. 2) We believe the yield curves for different forwards should not intersect significantly, and yield curves with shorter forward periods should mostly be dominated by those with longer forwards. This assumption holds because we believe that risks posed to the uncertainty in longer time is greater, hence should be compensated with higher yields. With these two expectations in mind, we discuss the results from the four models.

In Figure 7.7, we see that the long run interest level y_0 evolves at a rather stable level, while the short y_1 and medium y_2 term levels show relatively high correlation and high volatility. In all of the four forecast plots corresponding to our four models respectively, we show the 60-period mean forecast values of $y(t)$ in the upper plot, and the forecasted mean yield curves with four typical forward periods (ie. today, 1-year, 3-year and 5-year) in the lower plot.

The forecasted Nelson-Siegel factors, in Figure 7.8, using the VAR(1) model shows

[3]There are 3 types of yield curves: normal, inverted and flat. Most of the post-Great Depression yield curves have been normal, and there is enough reason for us to believe it should be normal

```
Geweke Diagnostics for each parameter chain
Variable          Mean      std dev      NSE iid      RNE iid
phi 0         0.013170     0.044212     0.001398     1.000000
phi 1        -0.012824     0.043320     0.001370     1.000000
phi 2         0.095037     0.337383     0.010669     1.000000

Variable       NSE 4%       RNE 4%       NSE 8%       RNE 8%      NSE 15%
phi 0         0.001385     1.019037     0.001484     0.887651     0.001400
phi 1         0.001192     1.321504     0.001161     1.391757     0.001044
phi 2         0.010099     1.115972     0.009766     1.193568     0.009268

  RNE 15%
 0.997201
 1.722246
 1.325150

Geweke Chi-squared test for each parameter chain
First 20% versus Last 50% of the sample
Variable            phi 0
NSE estimate          Mean       N.S.E.  Chi-sq Prob
i.i.d.            0.012691     0.001710     0.620908
4% taper          0.012616     0.001583     0.581897
8% taper          0.012449     0.001668     0.587570
15% taper         0.012216     0.001725     0.602652

Variable            phi 1
NSE estimate          Mean       N.S.E.  Chi-sq Prob
i.i.d.           -0.011670     0.001686     0.778119
4% taper         -0.011679     0.001436     0.738817
8% taper         -0.011693     0.001393     0.727647
15% taper        -0.011748     0.001301     0.698195

Variable            phi 2
NSE estimate          Mean       N.S.E.  Chi-sq Prob
i.i.d.            0.105432     0.012862     0.799946
4% taper          0.105512     0.013101     0.801667
8% taper          0.105007     0.011099     0.783266
15% taper         0.104849     0.010969     0.787452
```

Figure 7.5: Geweke diagnostics for Model 3

```
Geweke Diagnostics for each parameter chain
Variable          Mean      std dev      NSE iid      RNE iid
phi 0         0.998704     0.003682     0.000116     1.000000
phi 1         0.013190     0.010479     0.000331     1.000000
phi 2        -0.016624     0.009377     0.000297     1.000000

Variable       NSE 4%       RNE 4%       NSE 8%       RNE 8%
phi 0         0.000137     0.721950     0.000153     0.575885
phi 1         0.000343     0.932118     0.000358     0.855544
phi 2         0.000275     1.164783     0.000284     1.086397

 NSE 15%      RNE 15%
0.000163     0.512816
0.000347     0.912707
0.000242     1.506677

Geweke Chi-squared test for each parameter chain
First 20% versus Last 50% of the sample
Variable            phi 0
NSE estimate          Mean       N.S.E.  Chi-sq Prob
i.i.d.            0.998483     0.000142     0.029574
4% taper          0.998506     0.000137     0.029981
8% taper          0.998532     0.000137     0.039884
15% taper         0.998581     0.000122     0.049599

Variable            phi 1
NSE estimate          Mean       N.S.E.  Chi-sq Prob
i.i.d.            0.012646     0.000409     0.150961
4% taper          0.012773     0.000361     0.164952
8% taper          0.012805     0.000330     0.152876
15% taper         0.012885     0.000277     0.176683

Variable            phi 2
NSE estimate          Mean       N.S.E.  Chi-sq Prob
i.i.d.           -0.016230     0.000363     0.306988
4% taper         -0.016265     0.000292     0.229130
8% taper         -0.016356     0.000224     0.220883
15% taper        -0.016395     0.000203     0.272107
```

Figure 7.6: Geweke diagnostics for Model 4

clear mean reverting feature of the model, while the three components stays relatively isolated. The resulting yield curves converge quite quickly, which means for a bond with a maturity of 10 years, there is little difference in yield if we buy it today, in 1-year, 3-year's or 5-year's time. The results from VAR(1) does not directly violate the two beliefs we assume for yield curves, however, the tiny differences in yields among the four forward periods squeeze out the motivation to take such risks. This result is not encouraging, so we look to the Bayesian modification of this regression model.

The forecast made by the standard BVAR(1) model is plotted in Figure 7.9. The Bayesian structure with non-informative prior forecasted a converging medium and short term interest levels. However, the yield curves violate the second belief we assumed at the beginning of this chapter. The yield curves with longer forward periods should dominate those with shorter forward periods. In our case, yield curves, from top to bottom, should be in the order of 5-year, 3-year, 1-year and zero-year (today) forward periods. The Bayesian VAR(1) with non-informative prior presented us with a wrong (ie. non normal) structure of the yield curves. The reason for this problem lies behind the choice of the non-informative prior, we assumed a non-informative prior, with fixed mean and variance. The result shows that this model lies heavily on correct specification of the prior, which is something we are unable to supply except making a consensus choice. Therefore, we attempt an informative prior with Monte Carlo methods to forecast the time series.

Then we look at the hierarchical Bayesian VAR(1) with an informative prior, incorporated with Gibbs sampler, the result is shown in Figure 7.10. For each simulation step, we make 1100 draws, and discard the first 100 as burn-in. Then compare to the yield curves from the standard Bayesian VAR(1), we find yield curves follow a more reasonable order, however, the yield curves still intersect at some maturity, which is not desirable according to the second belief we have on yield curve behaviors. Though this modification of the Bayesian VAR(1) model has shown some improvement, the results still exhibit the problem of intersecting (non normal) yield curves. This inspired us that the main structure of the model gives approximate correct order of the yield curves, but may have some critically information in the error terms being missed out. This motivation turns us to the heteroscedastic modifications of this Bayesian VAR(1) model.

For the last Bayesian heteroscedastic VAR(1), with a set of informative priors and Gibbs sampler, the results are shown in Figure 7.11. The simulated Nelson-Siegel factors have shown a much better replicate of the correlation as in Figure 7.7. The yield curves have demonstrated a reasonable order for different forward periods.

The idea above is further illustrated by the time series on cross-correlations between y_0, y_1 and y_2. We observe the correlation time series on a 12-monthly (ie. 1 year) intervals. Figure 7.12 shows the historical correlation structure among the 3 factors. Figure

7.13 shows the simulated correlations based on VAR(1), which looks more like switching between -1 and +1 correlation. This certainly does not look good in replicating the historical correlations. Figure 7.14 shows the simulated correlations based on standard BVAR(1). It is observed that the correlation between y_1 and y_2 still resembles a switching phenomenon, while the correlation between y_0 and y_1 shows an evident periodicity, which is non-existent in the historical correlation. Figure 7.15 shows the simulated correlations based on BVAR(1) with Gibbs sampler. It is observed that switching phenomenon between -1 and +1 is not so evident under this model, the correlation between y_0 and y_1 has shown most significant improvement, whilst the other two showing clear periodicity which is not evident in history. Figure 7.16 shows the simulated correlations based on heteroscedastic regression model with Gibbs sampler. This result best replicates the historical distribution of correlations, there are no obvious switching between extreme constant, nor periodic behaviour.

7.8 Figures

7.9 Conclusion

In this chapter, we have developed from the standard VAR model, to the Bayesian VAR model by assuming the distribution for the coefficient ϕ; further, we implemented a hierarchical model by adding distributional assumption onto the error variance parameter σ, and sampled through the Gibbs sampler to retrieve estimates for the parameters; finally, by introducing more parameters to accurately specify the covariance matrix of the error, namely, the variance matrix V, and the auxiliary parameter r, we developed a Bayesian heteroscedastic regression model. For the last two models based on MCMC methods, we examined the convergence of the Gibbs samplers.

By comparing the simulated results, we conclude that the Bayesian heteroscedastic regression model best suits our purpose, which is to simulate yield curves. In summary, we choose this model based on the following two reasons: 1) it best satisfies the normal yield curve assumption and the non intersecting nature; 2) the simulated Nelson-Siegel interest rate factors best replicates the historical time series, both in terms of the mean reverting feature, volatility level and correlations.

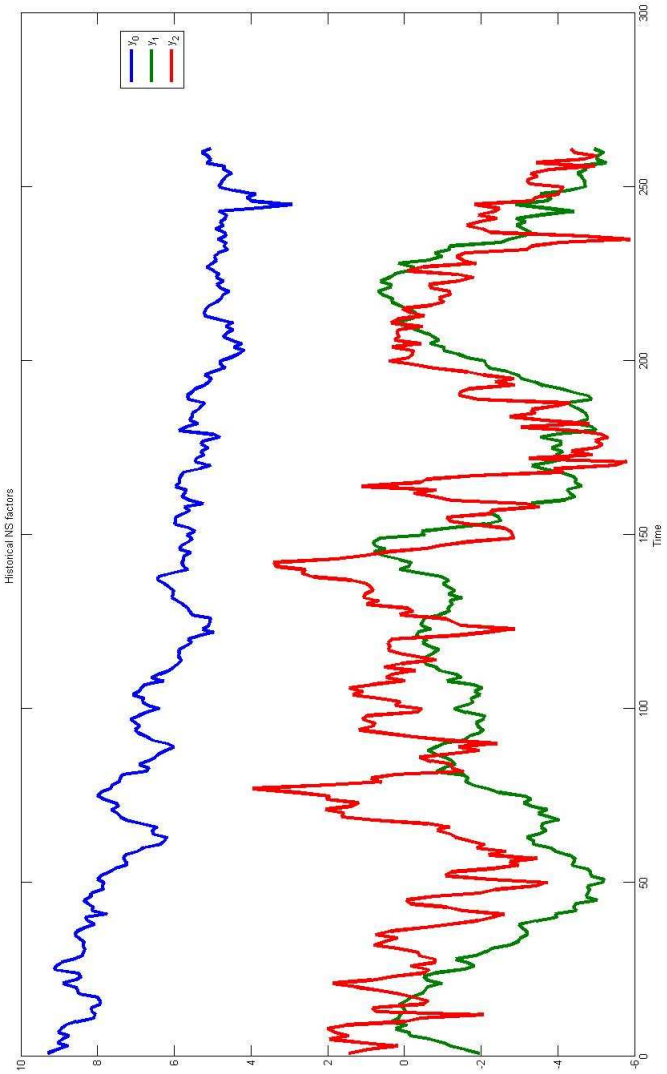

Figure 7.7: Historical values of the Nelson-Siegel Parameters

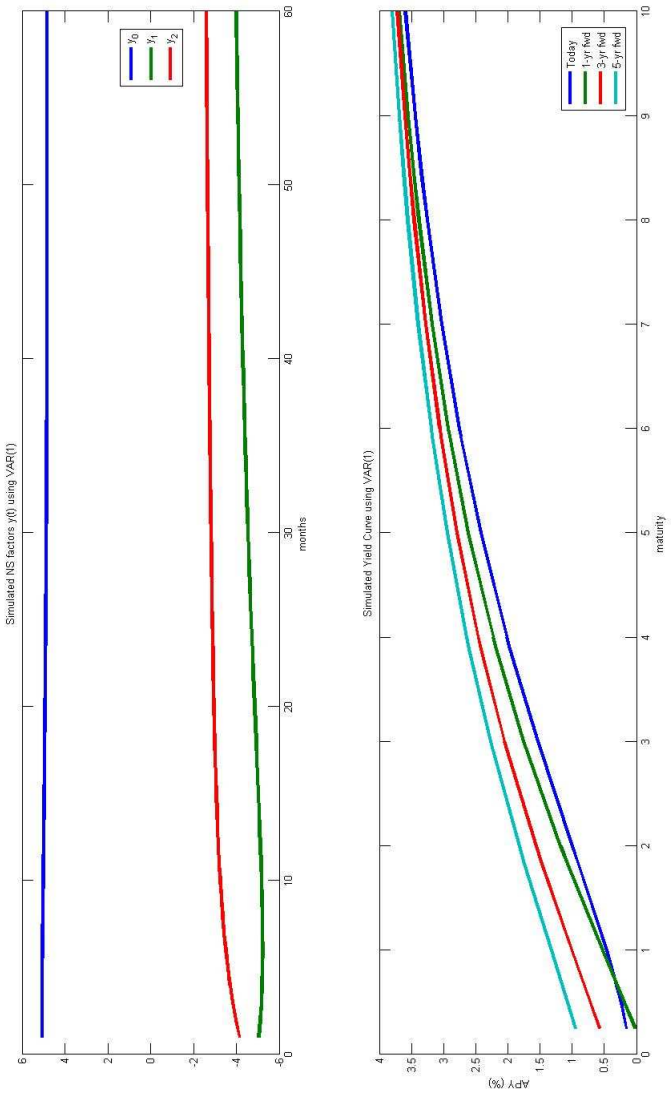

Figure 7.8: VAR(1) simulated Nelson-Siegel factors and Yield Curves

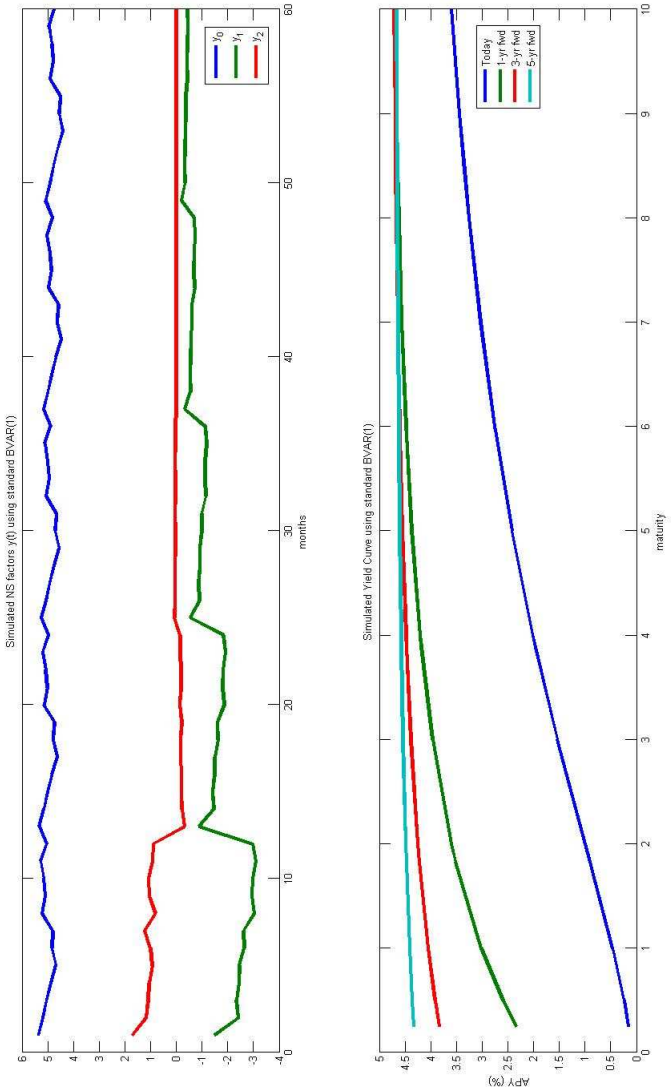

Figure 7.9: Standard BVAR(1) simulated Nelson-Siegel factors and Yield Curves

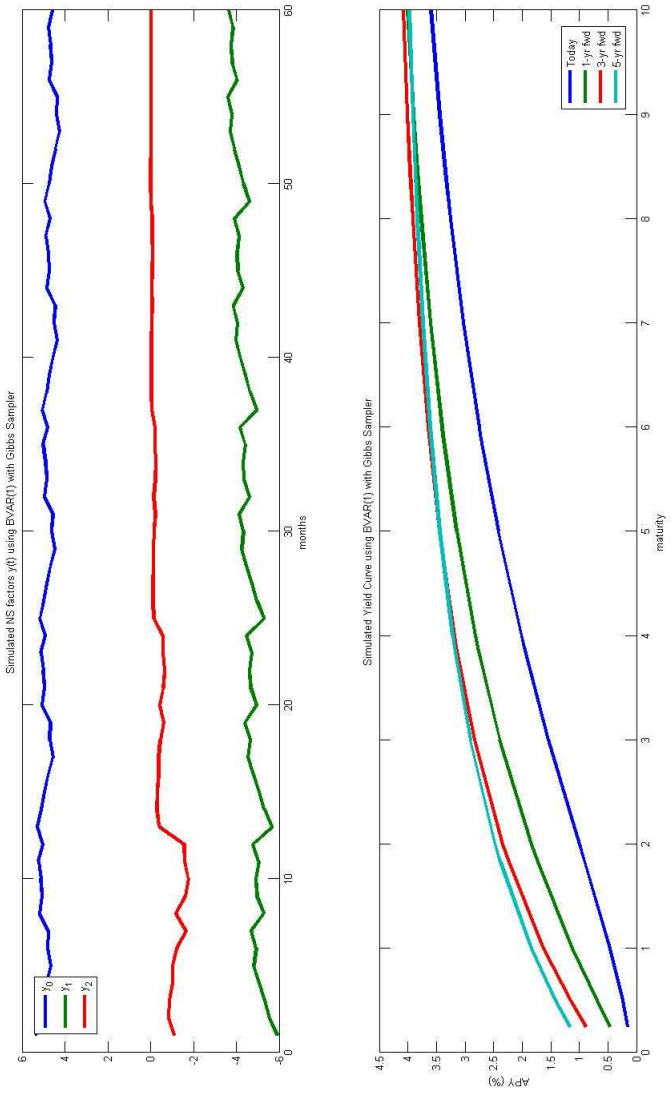

Figure 7.10: Simulated Nelson-Siegel factors and Yield Curves using Gibbs Sampled BVAR(1)

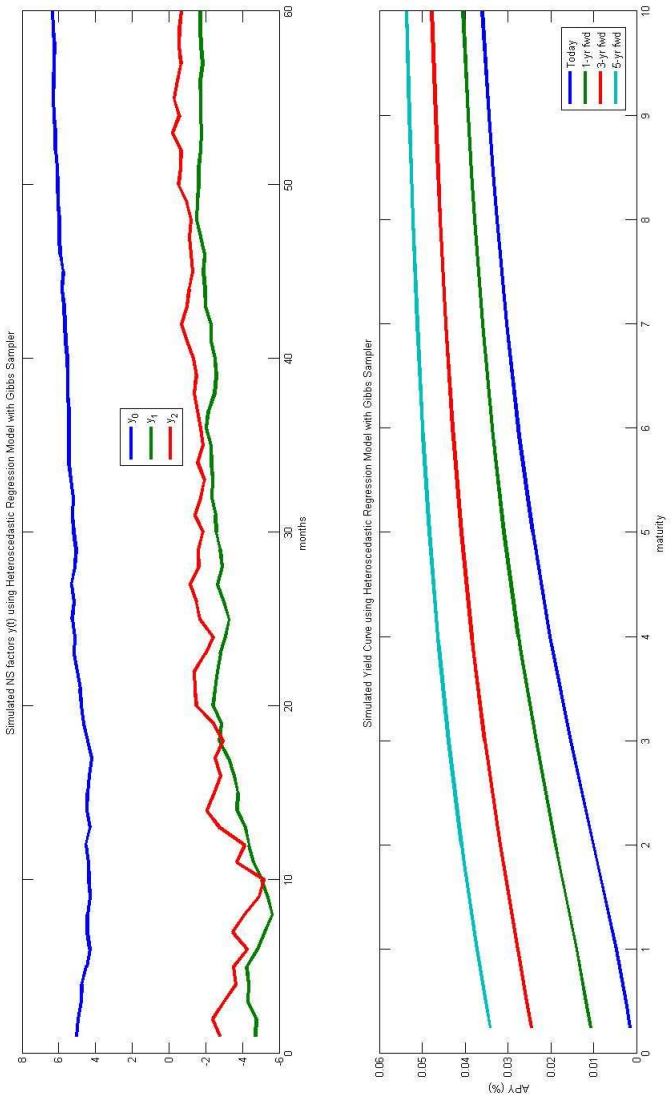

Figure 7.11: Simulated Nelson-Siegel factors and Yield Curves using Heteroscedastic Regression model with Gibbs Sampler

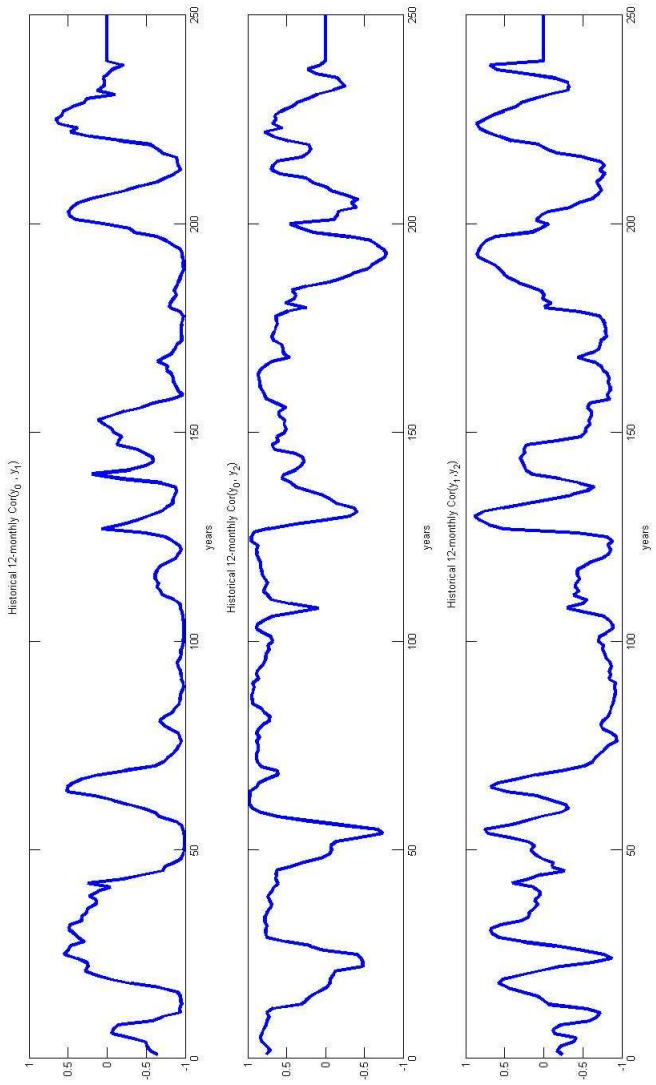

Figure 7.12: Historical Correlations

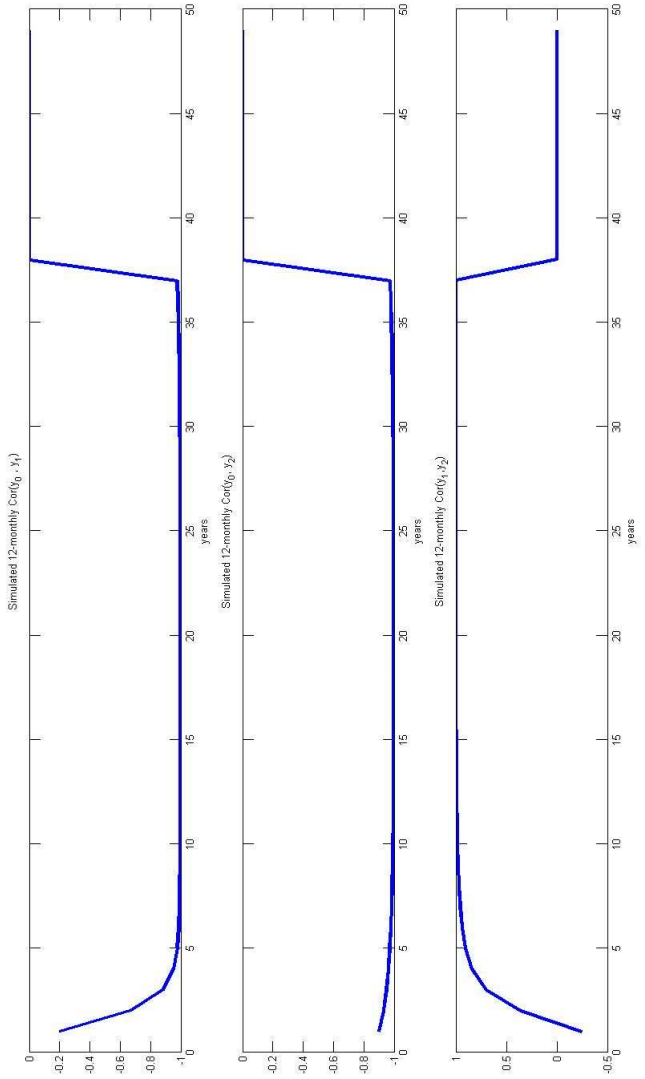

Figure 7.13: Simulated Correlations based on VAR(1)

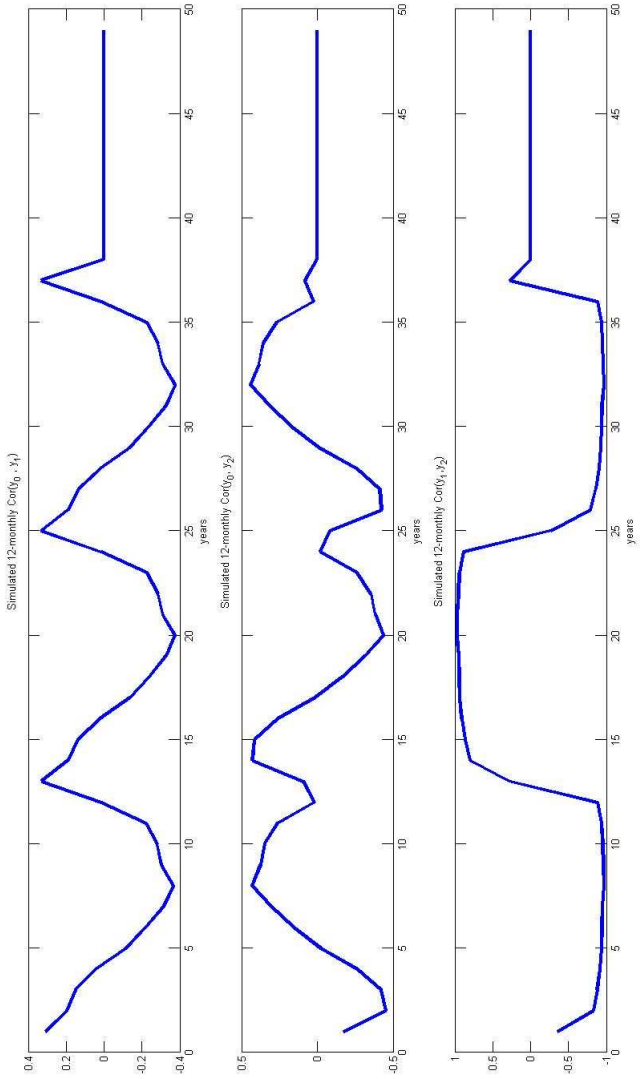

Figure 7.14: Simulated Correlations based on standard BVAR(1)

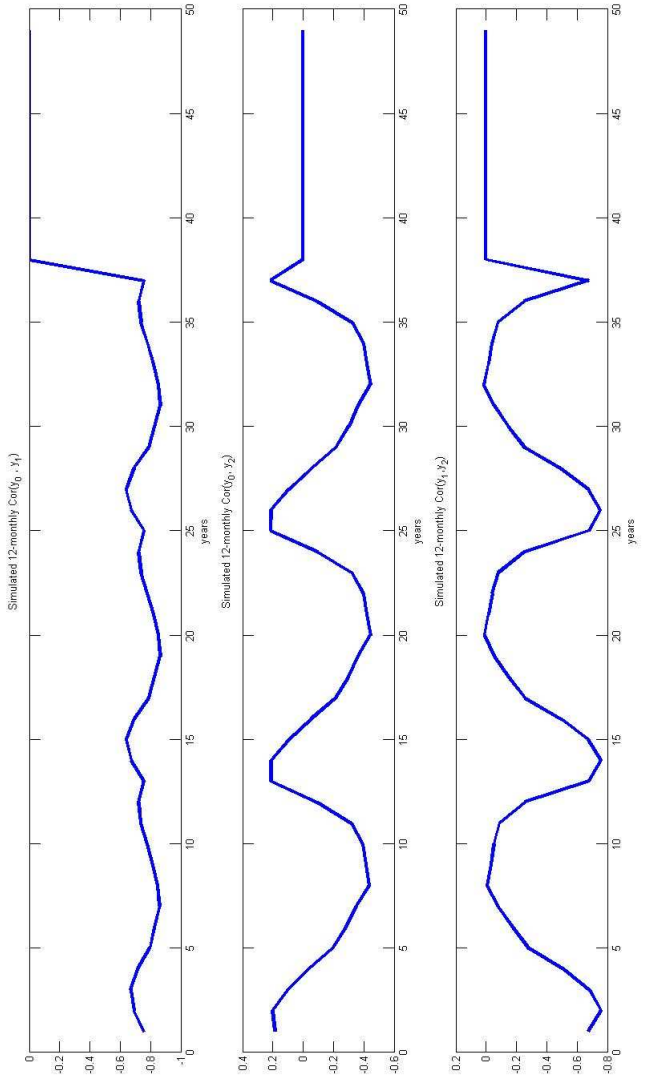

Figure 7.15: Simulated Correlations based on BVAR(1) with Gibbs Sampler

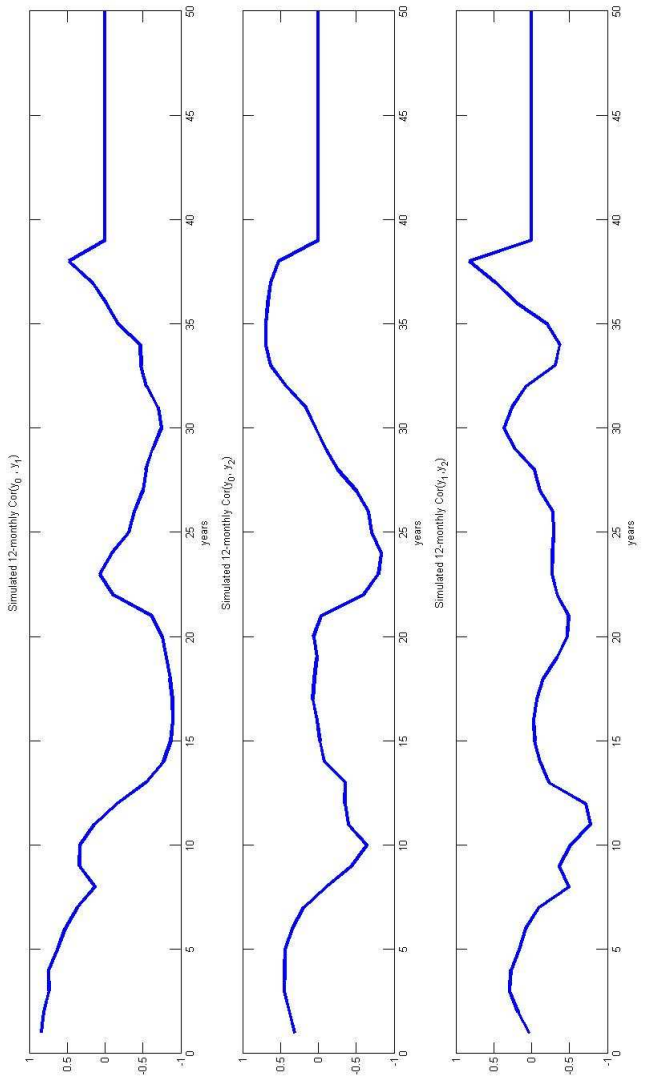

Figure 7.16: Simulated Correlations based on Heteroscedastic Regression Model with Gibbs Sampler

Chapter 8

Dynamic Conditional Correlation GARCH

The autoregressive models discussed in the previous chapter have shown inadequate capability in modelling time-varying volatility. We attempted to overcome some of these shortcomings with the Bayesian heteroscedastic regression models; in this chapter, we delve deeper into this modelling issue and investigate the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) model. This model has proved successful in predicting volatility changes.

The phenomena of clustering of volatilities exist in a wide range of economic and financial activities. The word "clustering of volatility"refers to observations that "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes"[63, 64]. For example, if we are given a set of uncorrelated, drift (or mean) corrected, economic time series data $\{x_t\}_{t\in[0,T]}$, the magnitude of the time series $\{|x_t|\}_{t\in[0,T]}$ displays a positive, significant and slowly decaying autocorrelation function: $\mathrm{corr}(|x_t|, |x_{t+\delta}|) > 0$ for various δ [64]. No real economic reason has been proved to explain the clustering behaviour of volatilities, despite the empirical success of the ARCH/GARCH models. This observation motivates us to apply these models to simulate economic risks. GARCH models require high frequency data, this is realisable for financial data, but not easily feasible in terms of economic and monetary data. This is one of the reasons we did not refer to this model in the previous chapter.

In section 8.1 we introduce the portfolio space, the data available and our objective. In section 8.2, we make a brief review of the multivariate GARCH model setup. We discuss the type of multivariate GARCH model with constant correlation matrix in section 8.3, and dynamic conditional correlation matrices in section 8.4. In section 8.5, we examine the goodness-of-fit of the proposed models using Kupiec Proportion-of-Failure test and

Christoffersen's Markov test. Finally, the simulated results are discussed in section 8.6.

8.1 The portfolio

The problem of interest here is a portfolio with a set of financial indices from very different asset classes. The portfolio space reflects the potential investment space for the reserves of the International Monetary Fund. The indices show the health of that specific asset class, and as a proxy to monitor the health of that economy. The data source is Bloomberg. The indices, discussed in this chapter are:

Index	Asset class	Weight[1]
Russell 3000	US Equity market index	30%
MSCI EAFA	Morgan Stanley Capital International	30%
	Europe, Australasia, Far East (International Equity index)	
US MBS	US Mortgage Backed Securities	11.5%
US Corp	US Corporate Debt	11.5%
HY	High Yield Bonds	12%
	(non-investment grade bond)	
Real Estate	Real Estate Prices	5%
Commodity	Global commodities (oil, gold, metal, etc)	0%
Cash	USD Cash equivalent assets	0%

The normalized[2] historical indices of the above assets are shown in Figure 8.3, from the plot, we see trend is not so obvious for most of the assets, so we opt away from autoregressive models, by referring to GARCH models on the (natural logged) returns of the assets to explore the nature of the volatilities. We plot the log returns in Figure 8.4 and the autocorrelation functions in Figure 8.5.

The data is provided on a weekly frequency from 3 April, 1992 to 12 June, 2009. Our objective is to find an appropriate volatility model to simulate returns for the indices. Assessment of the suitability of the model is based on the simulated cross asset correlation distribution and the empirical feasibility of the simulated returns. For this reason, we put more emphasis on the distribution closeness which are estimated by least squares in 8.3 and maximum likelihood in 8.4 respectively, instead of a mean estimate from Bayesian MCMC sampling as we did in the previous chapter. To measure the the closeness of the distributions, we implement goodness-of-fit tests, namely, Kupiec Proportion-of-Failure test and Christoffersen Markov test, which will be discussed in detail in section 8.5.

[2]Normalization here means all indices share the same starting value. eg. 1 or 100

8.2 The multivariate GARCH model

It is widely accepted in the industry that understanding and predicting the dependence in the co-movements of asset returns is important. It is observed that financial volatilities move together more or less closely over time across assets, geographical and industrial markets. Therefore, multivariate models provide a better way to explore the cross sectional (eg. asset classes, geographical distribution, markets, etc.) relevance than working with separate univariate models [70]. The GARCH model has been discussed in extensive detail in many literatures, such as [13, 14, 15], and multivariate GARCH model has also been explored extensively, such as [70, 86, 87]. In this chapter, we follow the notation used in [70].

The multivariate GARCH model is defined as

$$r_t = \mu_t + a_t, \tag{8.1a}$$
$$a_t = H_t^{\frac{1}{2}} z_t, \tag{8.1b}$$

where,

(i) $r_t \in \mathbb{R}^n$ is a vector of log returns at time t;

(ii) $a_t \in \mathbb{R}^n$ is a vector of mean-corrected returns at time t, ie. $\mathbb{E}(a_t) = 0$, $\mathrm{Cov}(a_t) = H_t$;

(iii) $\mu_t \in \mathbb{R}^n$ is a vector of expected value of the conditional r_t;

(iv) $H_t \in \mathbb{R}^{n \times n}$ is the conditional covariance matrix of a_t at time t;

(v) $H_t^{\frac{1}{2}}$ the Cholesky factorization of H_t[3];

(vi) $z_t \in \mathbb{R}^n$ is the vector of iid errors such that $\mathbb{E}(z_t) = 0$ and $\mathbb{E}(z_t z_t^T) = I$.

The idea of modelling conditional variance and correlation motivates us to decompose the covariance matrix as

$$H_t = D_t R_t D_t$$

where $D_t = \mathrm{diag}\left(h_{1t}^{\frac{1}{2}}, h_{2t}^{\frac{1}{2}}, \ldots, h_{1n}^{\frac{1}{2}}\right)$ is the conditional standard deviation, with h_{it} the (i,i)th element of H_t; R_t is the correlation matrix. We focus on two specifications of the multivariate GARCH model: GARCH with constant correlation matrix and GARCH with time-varying correlation matrix.

[3]The idea of using Cholesky decomposition is that it helps capture the correlation between independent random variables after we have obtained the covariance matrix. However, in Chapter 7, when specifying the prior covariance (7.5), we construct the covariance V_0 from it's factorized matrix, we did not restrict it to accord with Cholesky decomposition for modelling convenience.

We will apply multivariate GARCH models to the time series data, with short memories as in the previous chapter. Empirically, it is often sufficient to assume a GARCH(1,1) model than sophisticated volatility models with higher orders, since the marginal benefit we achieve is overcome by the additional complication when implementing higher order GARCH models, reason and examples are studied extensively in [37, 89]. In our modelling project, short memory is also an assumption we take upon request. Therefore, we will only look at the constant correlation and time-varying correlation multivariate GARCH(p,q) models with lags $p = 1$ and $q = 1$, where p is the order of the Autoregressive (AR) terms and q is the order of the autoregressive conditional heteroscedasticity (ARCH) terms.

8.3 Model 1: Constant Correlation Matrix

GARCH models with constant correlation matrix has a covariance matrix of the form

$$H_t = D_t R D_t \tag{8.2}$$

where the correlation matrix $R = [\rho_{ij}]$ is a constant positive definite matrix, and the diagonal standard deviation matrix D_t is time dependent. With $\rho_{ii} = 1$, $i = 1, \ldots, n$, the covariance matrix is given by

$$[H_t]_{ij} = h_{it}^{\frac{1}{2}} h_{jt}^{\frac{1}{2}} \rho_{ij} \ ,$$

and each conditional variance is modelled as

$$h_{it} = c_i + \sum_{q=1}^{Q_i} A_{i,q} a_{i,t-q}^{(2)} + \sum_{p=1}^{P_i} B_{i,p} h_{i,t-p}$$

where $c \in \mathbb{R}^n$ is a vector, $A_j, B_j \in \mathbb{R}^{n \times n}$ are diagonal matrices; $a_{t-j}^{(2)} = a_{t-j} \odot a_{t-j}$ is the element-wise product[4]. H_t is positive definite if elements of c, A_j and B_j are positive, since R is positive definite. When implementing this model, we use the time series data of index returns $(r_t)_{\{0 \le t \le T\}}$ of 8 assets, and aim to estimate the values of the constant correlation matrix R, and the time series of standard deviations D_t which are the components of the covariance time series H_t. In summary, for the Constant Correlation GARCH model, we follow the framework below,

- We estimate the time dependent covariance matrix H_t defined in (8.2), under the multivariate GARCH framework defined in (8.1), through the estimation of the constant correlation matrix R and the time series of standard deviations D_t;

[4]Element-wise product is defined as $(x \odot y)_{ij} = x_{ij} \cdot y_{ij}$.

- we use the time series data of the 8 asset classes as data input, by first apply mean correction, and then estimate covariance;
- the method used in the estimation is least squares.

The results are discussed later in Section 8.6.

8.4 Model 2: Dynamic Conditional Correlation GARCH

Now we look at the case when the correlation matrix R in (8.2) has time dependence, denoted as R_t in this section. This setup is defined as the Dynamic Conditional Correlation GARCH model, it was introduced in [27], and the multivariate case was discussed in [26]. According to the formulation in (8.1a), (8.1b) and (8.2), similar to the constant correlation case, we express the conditional variances as

$$h_{it} = \alpha_{i,0} + \sum_{q=1}^{Q_i} \alpha_{i,q} a_{i,t-q}^2 + \sum_{p=1}^{P_i} \beta_{i,p} h_{i,t-p}. \tag{8.3}$$

The results are discussed below in Section 8.6. We define the standard errors η_t as

$$\eta_t = D_t^{-1} a_t \sim \mathcal{N}(0, R_t).$$

To preserve the properties of positive definite symmetric properties of the covariance matrix H_t, we make the following assumptions,

Assumptions 8.1.
Assume:

(i) *R_t positive definite, so as to ensure H_t is positive definite;*

(ii) *all elements of R_t must be equal or less than 1 in magnitude, since it describes the correlation;*

(iii) *R_t is symmetric.*

In order to ensure Assumptions 8.1, R_t is decomposed into the following structure

$$R_t = Q_t^{*-1} Q_t Q_t^{*-1} \tag{8.4}$$

$$Q_t = (1 - a - b)\bar{Q} + a\eta_{t-1}\eta_{t-1}^T + bQ_{t-1} \tag{8.5}$$

where $\bar{Q}$ is the covariance matrix of the standard errors η_t, which can be estimated easily as mean-squared-error. Q_t^* is a diagonal matrix which takes the square root of the diagonal

elements of Q_t. It rescales the elements in Q_t to ensure assumption 8.1 (ii). Q_t, including the initial value Q_0, must be positive definite to ensure assumption 8.1 (i). a and b are scalars such that $a, b \geq 0$ and $a + b < 1$.

Maximum likelihood based estimation of the parameters of the DCC-GARCH model requires us to make assumptions on the distribution of the errors z_t. We assume standard Gaussian here. For ease of notation, we divide the parameters to two groups, $\phi_i = (\alpha_{i0}, \alpha_{i1}, \ldots, \alpha_{iq}, \beta_{i1}, \ldots, \beta_{ip})$ for parameters of the univariate GARCH model for the asset i, and $\psi = (a, b)$ for the correlation structure. The likelihood function for $a_t = H_t^{\frac{1}{2}} z_t$ is

$$L(\phi, \psi) = \prod_{t=1}^{T} (2\pi)^{-\frac{n}{2}} |H|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} a_t^T H_t^{-1} a_t\right), \tag{8.6}$$

replacing H_t using (8.2) the **log-likelihood** function becomes

$$l(\phi, \psi) = -\frac{1}{2} \sum_{t=1}^{T} \left(n \log(2\pi) + 2\log(|D_t|) + \log(|R_t|) + a_t^T D_t^{-1} R_t^{-1} D_t^{-1} a_t\right). \tag{8.7}$$

We then estimate the parameters in two steps for ϕ and ψ iteratively. We first replace R_t using the identity, which gives the likelihood function for ϕ,

$$\begin{aligned} l(\phi) &= -\frac{1}{2} \sum_{t=1}^{T} \left(n \log(2\pi) + \sum_{i=1}^{n} \left(\log(h_{it}) + \frac{a_{it}^2}{h_{it}}\right)\right) \\ &= \sum_{i=1}^{n} \left(-\frac{1}{2} \sum_{t=1}^{T} \left(\log(h_{it}) + \frac{a_{it}^2}{h_{it}} + C\right)\right) \end{aligned} \tag{8.8}$$

where C is a constant. Note that the second line of the above equation means that the log likelihood is the sum of the log-likelihoods of the univariate GARCH equations of n assets, meaning that each parameter can be determined separately for each asset.

Once we have the MLE $\hat{\phi}$ by maximizing (8.8), we also know h_{it} for each asset, so η_t and $\bar{Q}$ can be estimated. We then estimate ψ using the following conditional likelihood

$$l(\psi) = -\frac{1}{2} \sum_{t=1}^{T} \left(n \log(2\pi) + 2\log(|D_t|) + \log(|R_t|) + \eta_t^T R_t^{-1} \eta_t\right),$$

for which it is equivalent to maximizing,

$$l^*(\psi) = -\frac{1}{2} \sum_{t=1}^{T} \left(\log(|R_t|) + \eta_t^T R_t^{-1} \eta_t\right). \tag{8.9}$$

The maximum likelihood estimators under these pseudo likelihood yields consistent

and asymptotically normal estimates [21]. In conclusion, we use the index return data $(r_t)_{0\leq t\leq T}$, first apply mean correction (remove μ) and normalization (divide by standard deviation $D_t^{-1}a_t$) on the data, then estimate the correlation time series matrix R_t using the DCC-GARCH(1,1) model. In summary, we estimate the Dynamic Conditional Correlation GARCH model by following the framework below,

- We estimate the time dependent covariance matrix H_t defined in (8.3), under the multivariate GARCH framework defined in (8.1), through the estimation of the time series of correlation matrix R_t defined in (8.4) and the time series of standard deviations D_t defined as in section 8.3, the parameter set of interest is (ϕ, ψ) in specific;
- we use the time series data of the 8 asset classes as data input, by first apply mean correction, and then estimate covariance;
- the log-likelihood function of the parameters is given in (8.7);
- the estimation method is maximum likelihood, it is applied iteratively for ϕ by maximizing (8.8) for each parameter separately, and for ψ by maximizing (8.9).

The results are discussed below in Section 8.6.

8.5 Goodness-of-Fit Tests

Since the purpose of this modelling project is to analyze the risks the portfolio is exposed to, the goodness-of-fit of the models will not focus on the mean forecasts the model simulates, but on the risks underlying the simulated scenario. We apply two goodness-of-fit tests for the GARCH models discussed in this chapter, the Kupiec Proportion-of-Failure (PoF) test [12, 53] and the Christoffersen's Markov test [18].

We apply the Kupiec PoF test on the simulated Value-at-Risk (VaR)[5]. The Kupiec PoF test evaluates if the total number of exceptions (ie. the number of simulated returns falls below the proposed VaR) agrees with the expected number of exceptions. The Christoffersen's Markov test examines if the exceptions are independently distributed over time.

For each simulated data, it is either identified as an exception or not, so the number of exceptions follows a binomial distribution

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

where n is the total number of data, x is the number of exceptions observed, and p is the probability of getting an exception, under an assumed distribution.

[5]Here we mean 95% VaR, which is defined as the 5% quantile within the space the VaR lies.

The null hypothesis H_0 is defined as the expected proportion of violations is equal to α, where α is the significance level of the test.

Under the null hypothesis, the test statistic is

$$\begin{aligned}\text{Kupiec} &= 2\log\left(\frac{\binom{n}{x} p_{obs}^x (1-p_{obs})^{n-x}}{\binom{n}{x} \alpha^x (1-\alpha)^{n-x}}\right) \\ &= 2\log\left(\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}\right) - 2\log\left(\alpha^x(1-\alpha)^{n-x}\right)\end{aligned}$$

from which the test statistic follows $\chi^2(1)$ distribution when n is large.

If the estimated probability, p_{obs} is above the significance level, we accept the model, otherwise, we reject the model.

We then apply the Christoffersen's Markov test [18] to check if the exceptions are independently distributed over time. The null hypothesis H_0 is defined as the exceptions are independently distributed over time.

The test statistic is defined as

$$\begin{aligned}\text{Christoffersen} &= 2\log\left(\frac{(1-\pi_{01})^{n_{00}}\pi_{01}^{n_{01}}(1-\pi_{11})^{n_{10}}\pi_{11}^{n_{11}}}{\alpha^x(1-\alpha)^{n-x}}\right) \\ &= 2\log\left((1-\pi_{01})^{n_{00}}\pi_{01}^{n_{01}}(1-\pi_{11})^{n_{10}}\pi_{11}^{n_{11}}\right) - 2\log\left(\alpha^x(1-\alpha)^{n-x}\right)\end{aligned}$$

where n_{ij} is the number of transitions from state i to j, for $i, j \in \{0, 1\}$, which corresponds to non-exceptions and exceptions, the corresponding transition probabilities are $\pi_{ij} = n_{ij}/\sum_j n_{ij}$. The distribution of the test statistic converges to $\chi^2(1)$ distribution when n is large.

The Kupiec PoF test only quantifies the number of exceptions while ignores the timing and two-state switching process of the random variable. The Christoffersen's Markov test complements this weakness. Since both of these test statistics are likelihood ratios, and follow a $\chi^2(1)$ distribution, the sum of the two test statistics forms a more explanatory and powerful test statistic. However, the weakness of assuming the random variable follows a Markov process remains.

8.6 Discussion of Results

We first take a look at the normalised historical indices, Figure 8.3 gives us an overview of the normalized historical indices, and Figure 8.4 presents an overview of the historical

1.0000	0.7057	0.0499	0.0714	-0.0077	0.6223	0.1634	0.0780
0.7057	1.0000	0.0498	0.0677	-0.0285	0.4652	0.2690	0.0794
0.0499	0.0498	1.0000	0.7763	-0.0358	-0.0107	0.0007	0.0381
0.0714	0.0677	0.7763	1.0000	-0.0284	0.0449	0.0314	0.0172
-0.0077	-0.0285	-0.0358	-0.0284	1.0000	-0.0064	0.0359	-0.0421
0.6223	0.4652	-0.0107	0.0449	-0.0064	1.0000	0.1108	0.0640
0.1634	0.2690	0.0007	0.0314	0.0359	0.1108	1.0000	0.0406
0.0780	0.0794	0.0381	0.0172	-0.0421	0.0640	0.0406	1.0000

Figure 8.1: Historical correlation matrix of returns

log returns. We compare the two models from three aspects: the volatilities, correlation structure and the goodness-of-fit test statistics. For both models, we generated N=1000 paths, took the means to be our estimates. All the plots have been normalized to the same scale for the convenience of visual comparison.

From Figure 8.4, we observe that 4 assets presented significantly higher historical volatilities than others, they are “Russell 3000”, “MSCI EAFE”, “Real Estate” and “Commodities”. In addition, we observed significant recent volatility increase in “Russell 3000”, “MSCI EAFE”, “Real Estate”, “High Yield” and “Commodities”, we define them to be “highly volatile assets”. The degrees of volatility for “US MBS” and “US Corp” are relatively small and stable across time, we define them to be “stable assets”. The volatility for “Cash” is almost negligible.

The simulated indices using the Constant Correlation GARCH(1,1) is plotted in Figure 8.6 , and the simulated mean returns in Figure 8.7. Comparing Figures 8.4 and 8.7, it is easy to see that the volatilities of those highly volatile assets have been significantly underestimated, even the volatilities for the stable assets are underestimated. The only exception is “Real Estate”, of which the simulated returns retain a seemingly similar level of volatility.

Since we are assuming the constant correlation GARCH(1,1) model, we compare the correlation matrix across the entire time horizon. Figure 8.1 shows the historical correlation matrix, and Figure 8.2 shows the correlation matrix of the simulated returns using constant correlation GARCH(1,1) model. Both matrices are computed based on discrete observations provided. We see the discrepancies are significant, the two matrices are nowhere close to each other in terms of elementwise values. From the differences in correlation matrices, we believe the constant correlation matrix multivariate GARCH(1,1) model is not an appropriate choice for our data.

For the hypothesis testings, we take significance level of 95%, and compare the testing statistic with the $\chi^2(1)$ distribution’s critical value. We implement the hypothesis testing on the portfolio level of our model, which means, we aggregate the simulated asset returns to portfolio returns, and assess the null hypothesis on the mean returns of the

1.0000	0.4089	-0.0117	0.0030	-0.0336	0.2458	0.0316	-0.0133
0.4089	1.0000	-0.0452	-0.0212	-0.0340	0.2800	0.0603	-0.0110
-0.0117	-0.0452	1.0000	0.7936	-0.0605	0.0141	-0.0401	-0.0435
0.0030	-0.0212	0.7936	1.0000	-0.0300	0.0353	0.0054	-0.0507
-0.0336	-0.0340	-0.0605	-0.0300	1.0000	-0.0412	0.0711	-0.0494
0.2458	0.2800	0.0141	0.0353	-0.0412	1.0000	-0.0662	0.0019
0.0316	0.0603	-0.0401	0.0054	0.0711	-0.0662	1.0000	0.0654
-0.0133	-0.0110	-0.0435	-0.0507	-0.0494	0.0019	0.0654	1.0000

Figure 8.2: Simulated correlation matrix using CC-GARCH(1,1)

portfolio, since we are more interested on the risk of the given portfolio. For the constant correlation GARCH(1,1) model, the Kupiec PoF test statistic is 0.04, assuming Gaussian distribution for the simulated portfolio returns, which suggests the expected proportion of exceptions equals to the observed; and the Christoffersen's Markov test statistic is 0.3613, which suggests the exceptions are independently distributed over time. One specific weakness of these hypothesis tests on this problem is that we are evaluating the risks on the portfolio level, since the given portfolio has zero weights on some assets, it means that the test statistics have no representation for the simulated results of those zero weighted assets. However, we conclude that the constant correlation GARCH(1,1) model does not adequately present the data from the three above discussed aspects.

Now we refer to the dynamic conditional correlation GARCH(1,1) model. Figure 8.8 presents one sample path taken as means of N=1000 generations, and Figure 8.9 shows the simulated mean returns. From Figure 8.9, we observe that the volatilities of all assets are more normalized at a similar level, instead of the historically differently scaled volatilities as in Figure 8.4. This is expected, since most of the assets presented a significant increase in volatilities in recent period, which is observable in Figure 8.4.

The progressive correlations base on 12-week (3-month) periods for the entire time horizon is plotted in Figure 8.10. It shows cross correlations for all 8 asset indices. The red line indicates where the simulated results starts to be taken into account. We see that the correlations present a good autoregressive structure, both historically and in simulation. The simulated correlations shows the change in correlation structure across time, this is particularly obvious for the correlation between "Russell 3000" and "MSCI EAFE".

On the portfolio level, the Kupiec PoF test statistic is 0.0519, under the assumption that the simulated portfolio returns follow a Gaussian distribution. The Christoffersen's Markov test statistic is 2.7437. The critical value of $\chi^2(1)$ at 95% confidence, is 3.84, so the null hypothesis is accepted. Thus, we conclude that the dynamic conditional correlation GARCH(1,1) model outperforms the constant correlation GARCH(1,1) model in modelling the volatilities and correlations of our given portfolio.

8.7 Figures

8.8 Conclusion

In this chapter, we have compared the characteristics of the multivariate GARCH model with two types of correlation matrix. We studied the GARCH model with a constant correlation matrix for the mean corrected returns in 8.3, and the GARCH model with a time dependent correlation matrix in 8.4. We examined the goodness-of-fit of both of the models through the Kupiec PoF and Christoffersen Markov tests.

By comparing the simulated results, we conclude that the dynamic conditional correlation GARCH model better suits our purpose, which is to replicate the volatility and cross-asset correlation structure of the portfolio. Our conclusion is supported by the simulated results, which are plotted in the figures in the next section.

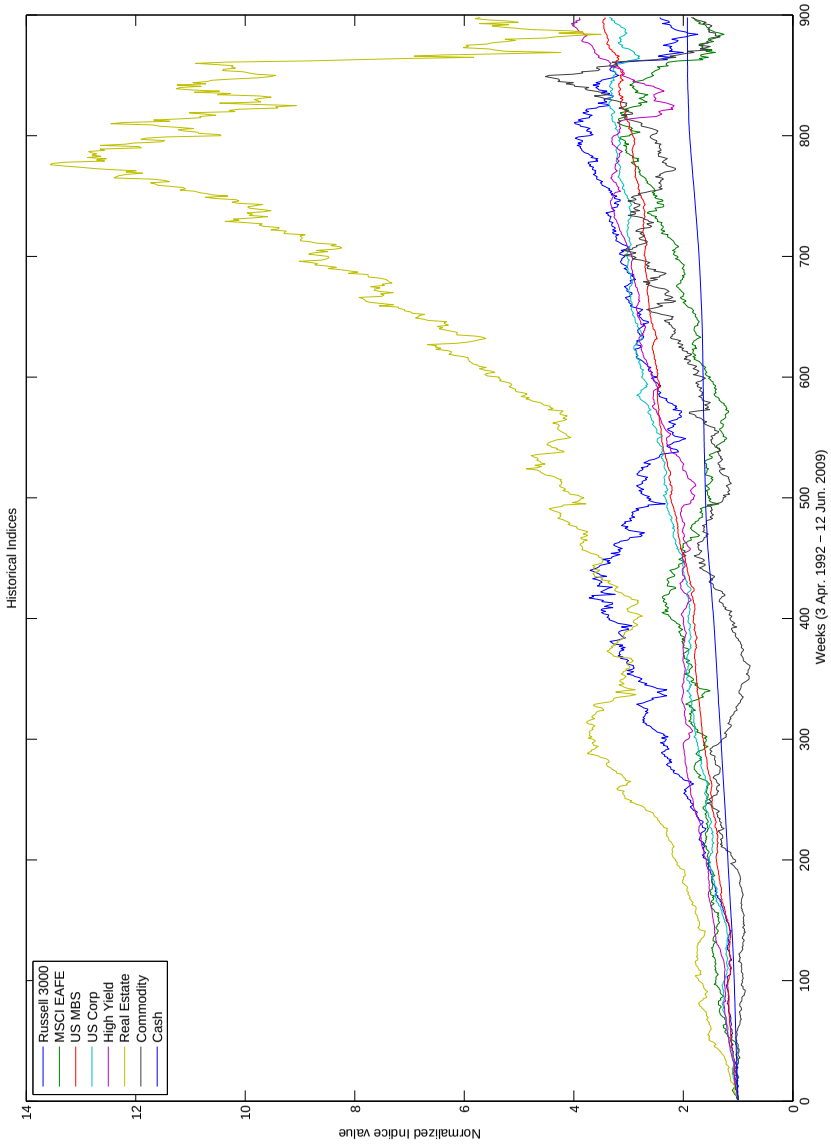

Figure 8.3: Historical price index of all assets

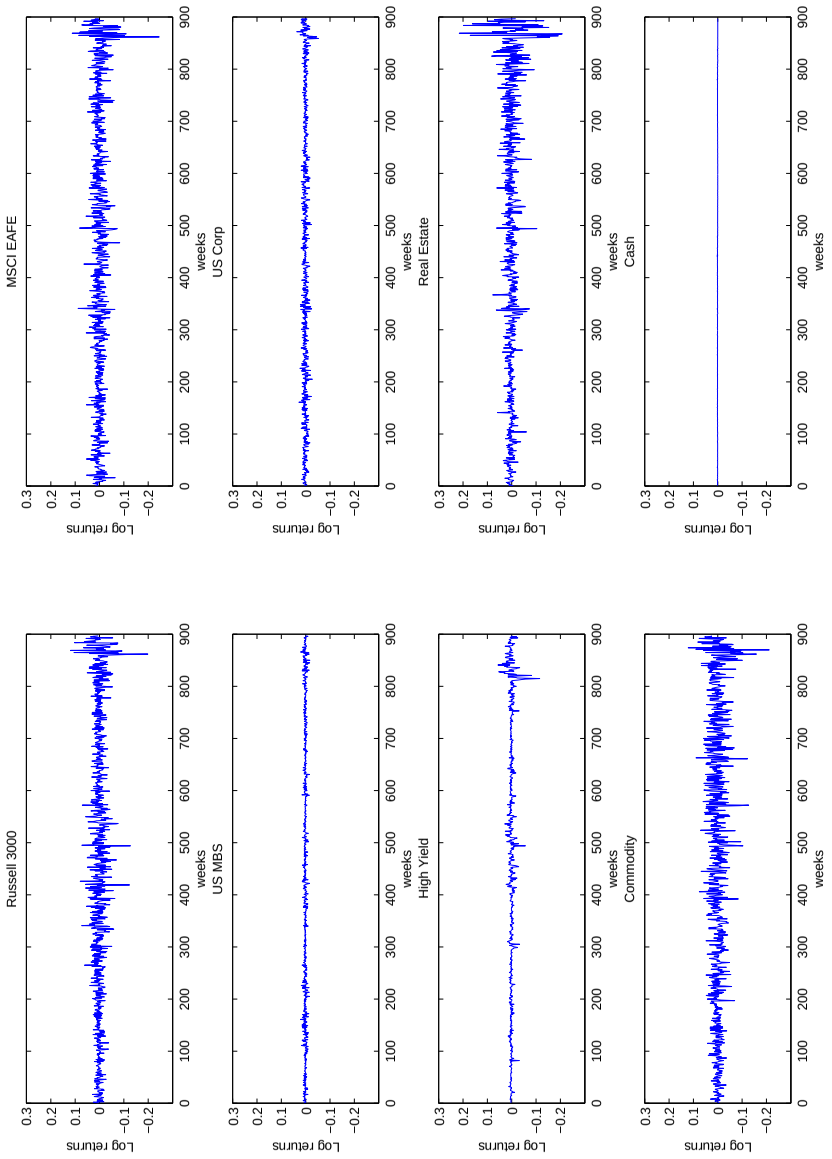

Figure 8.4: Historical log returns

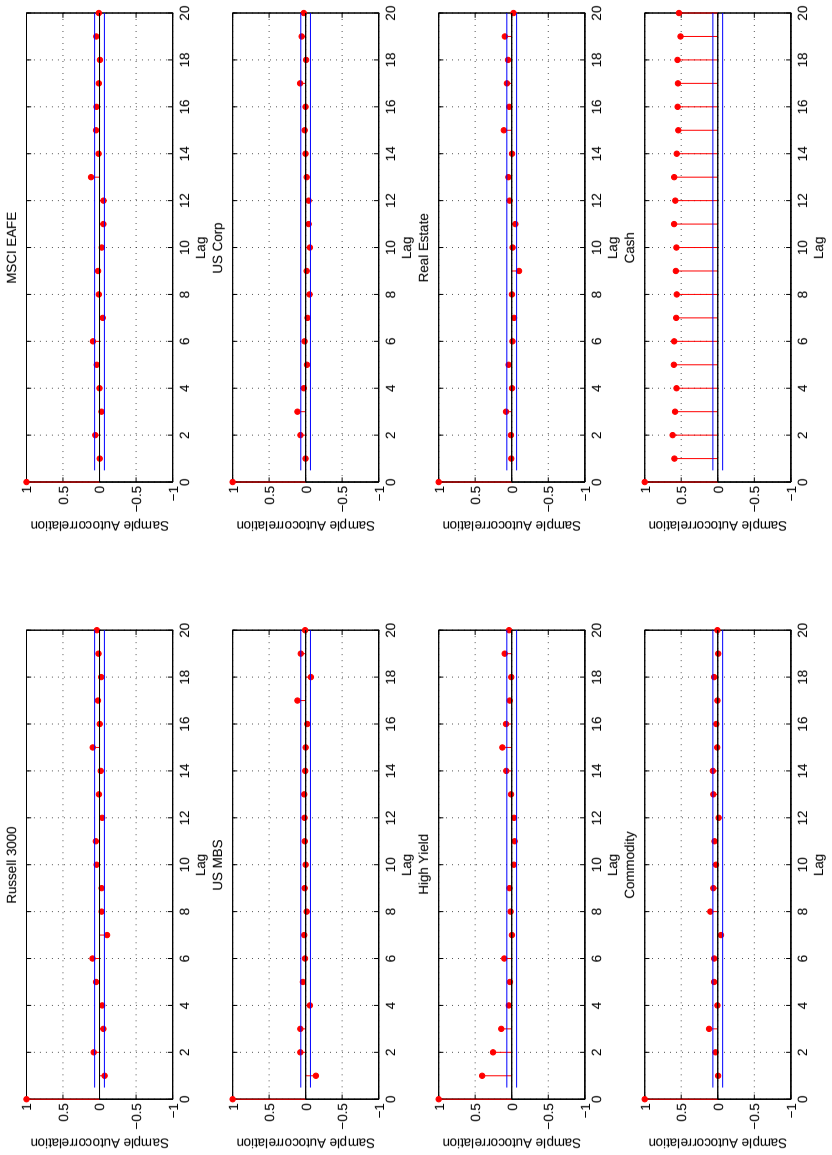

Figure 8.5: Autocorrelation function of historical returns

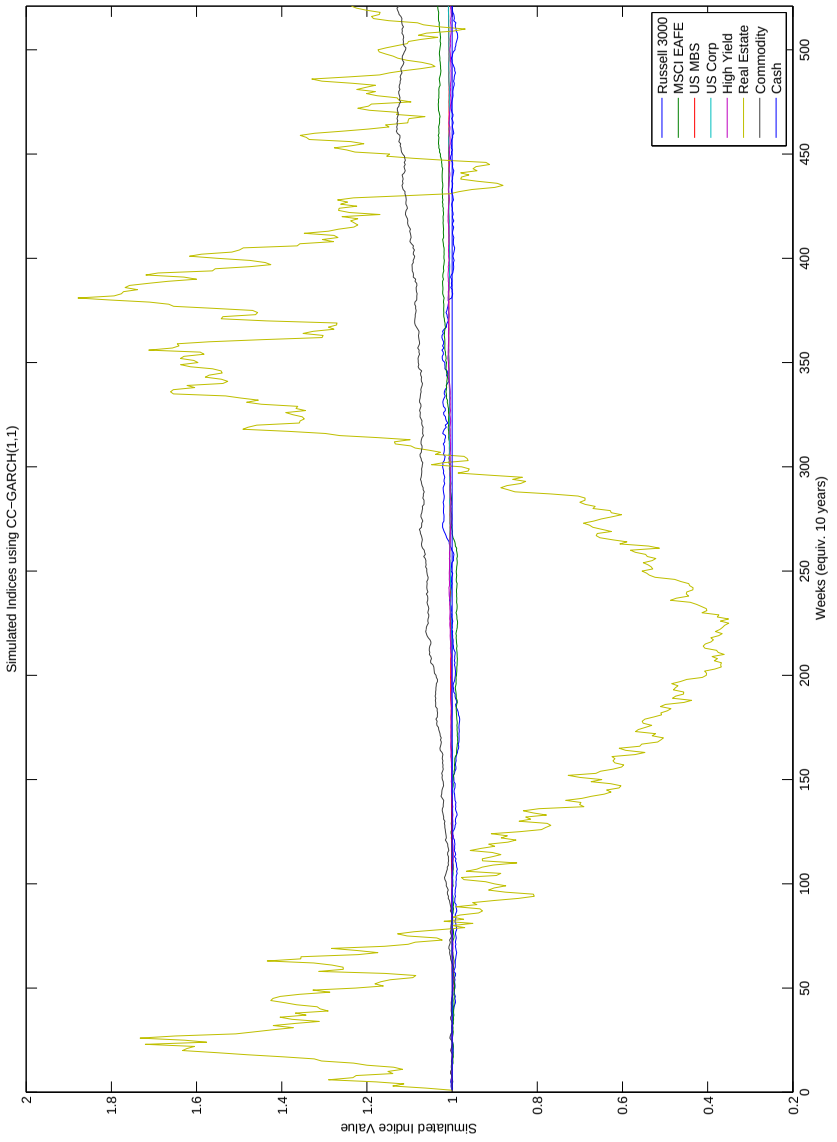

Figure 8.6: Price index simulated using CC-GARCH

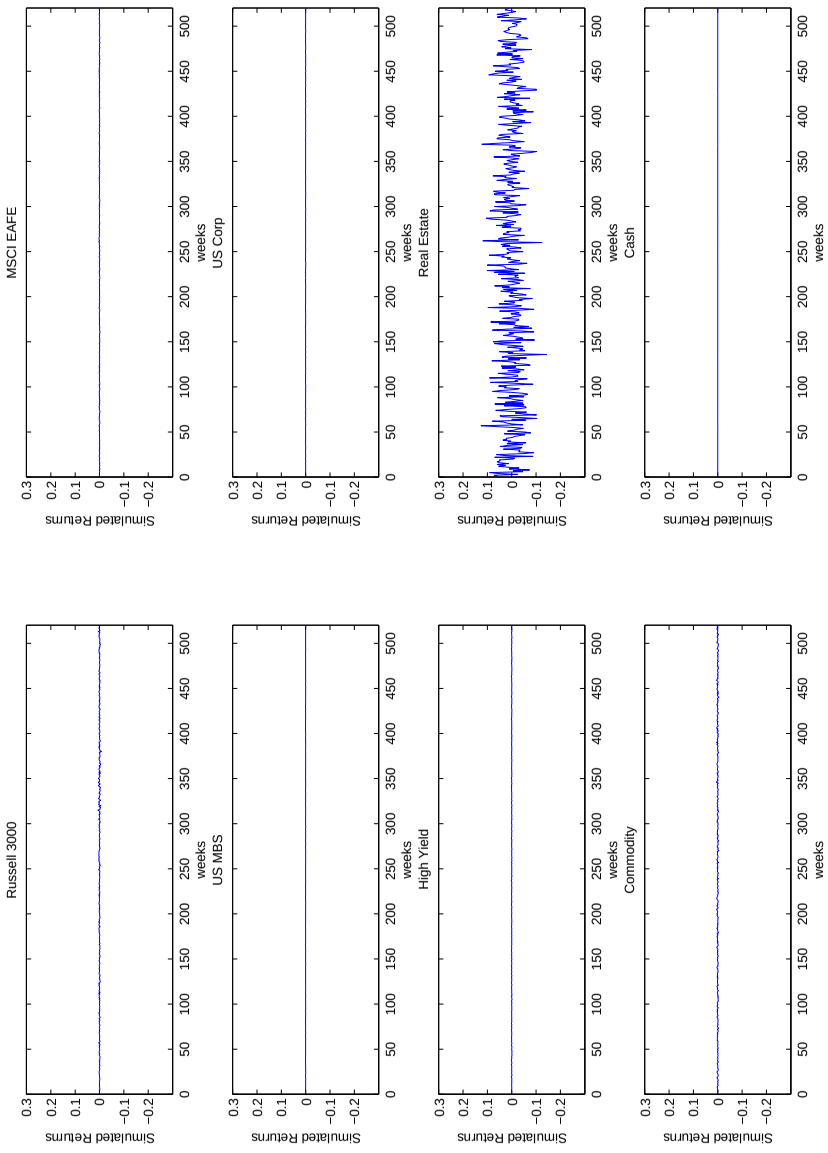

Figure 8.7: Returns simulated using CC-GARCH

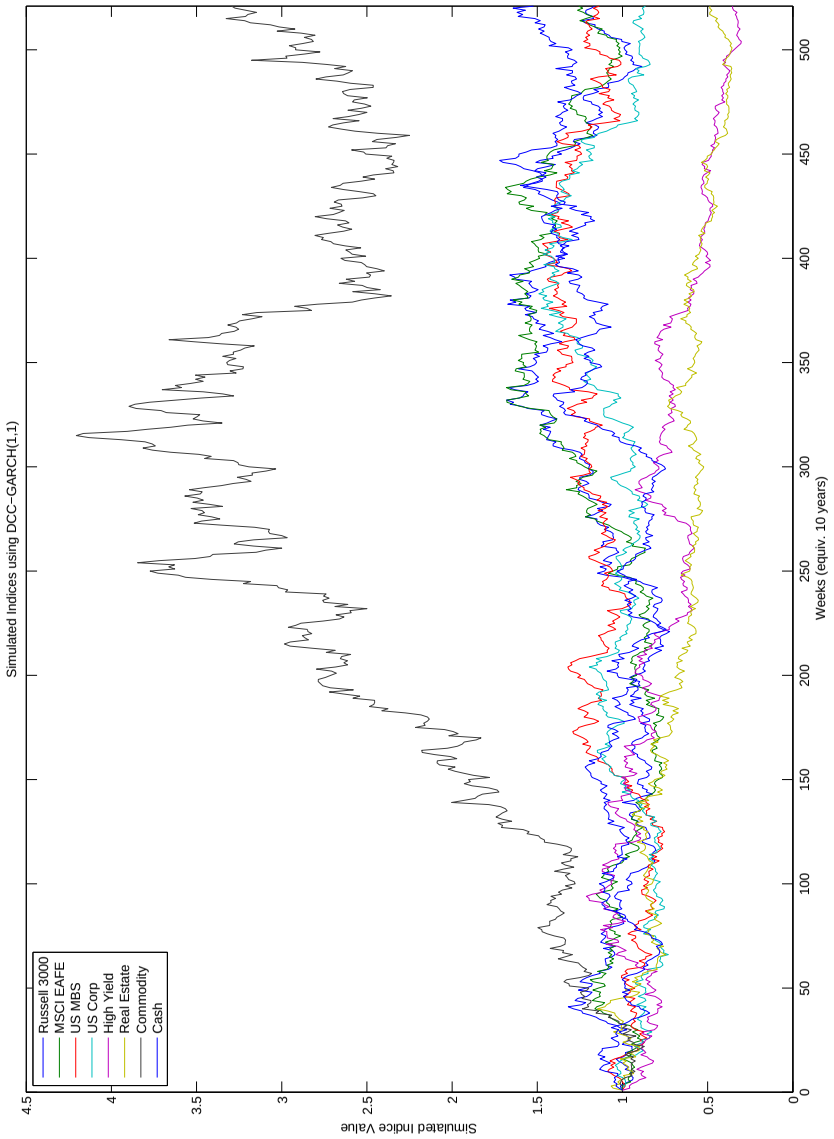

Figure 8.8: Price index simulated using DCC-GARCH

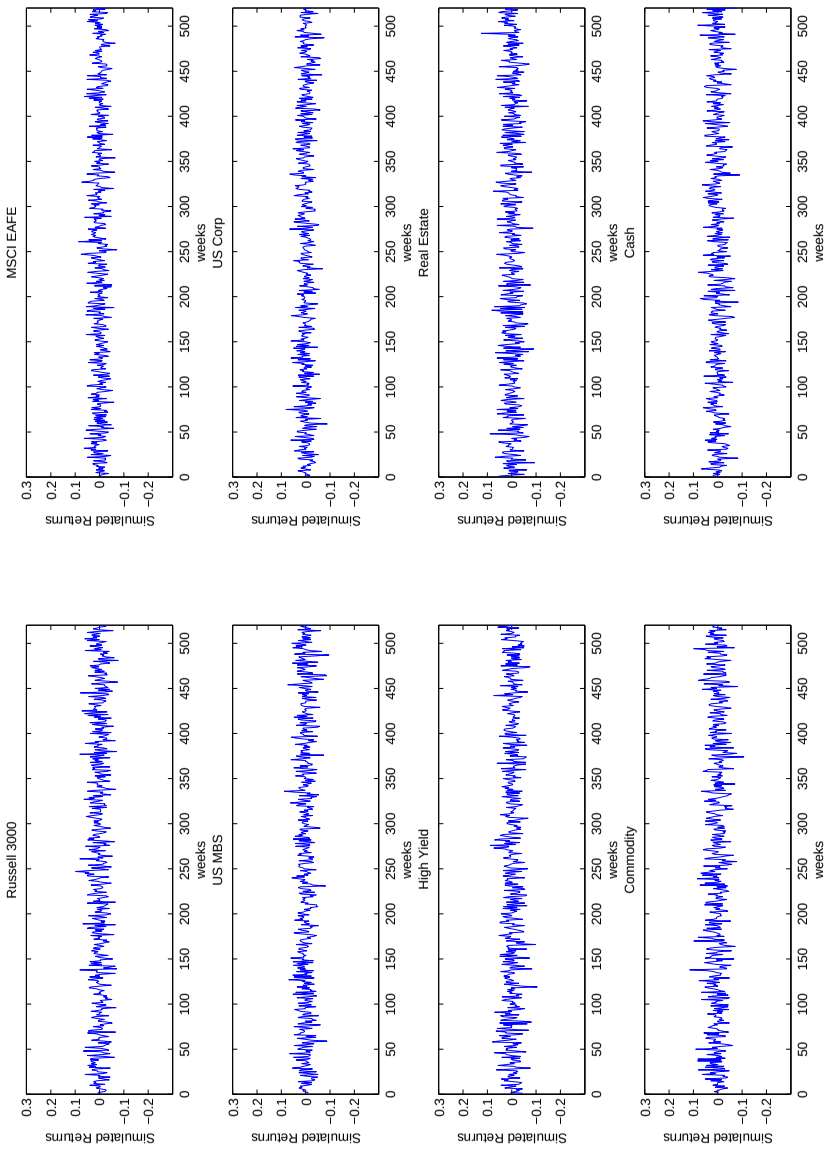

Figure 8.9: Returns simulated using DCC-GARCH

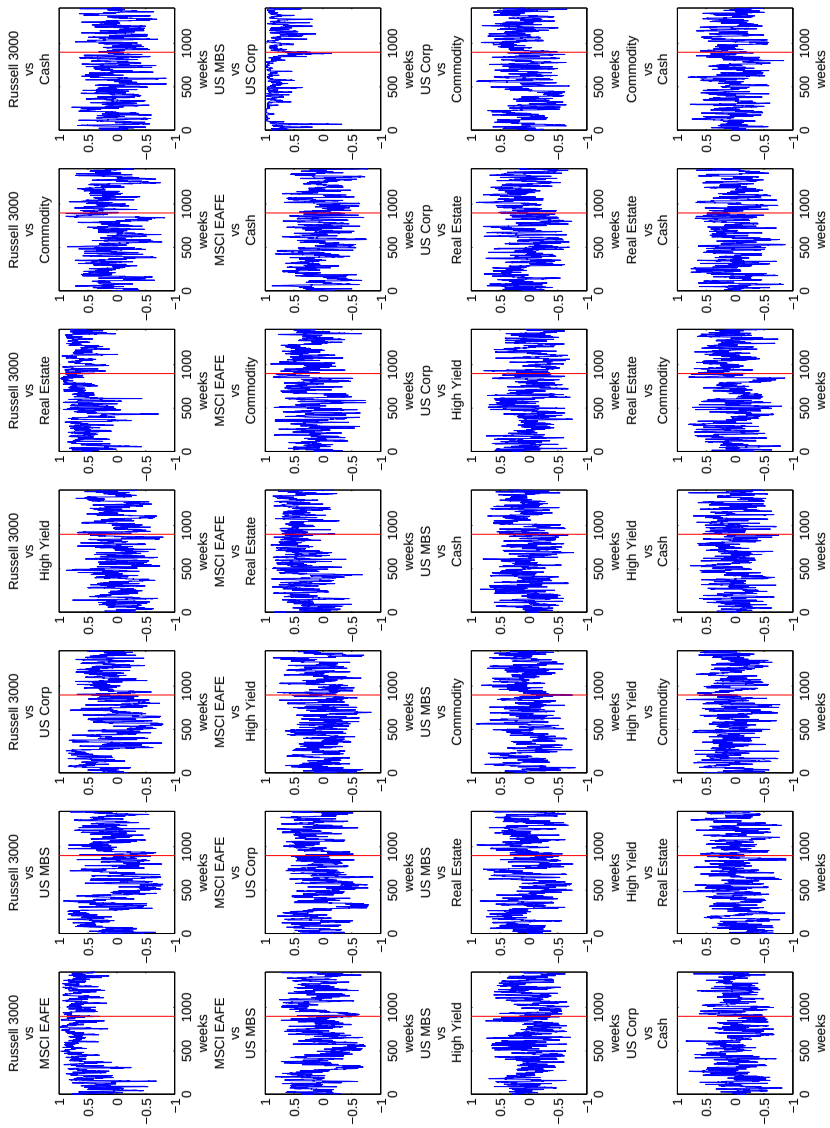

Figure 8.10: DCC-GARCH Simulated Correlation

Chapter 9

Appendix

9.1 Itô Formula

Lemma 6.5 in [78].

Consider the Itô SDE

$$\frac{dz}{dt} = h(z) + \gamma(z)\frac{dW}{dt}\,,\ Z(0) = z_0.$$

where $W(t)$ is a standard m-dimensional Brownian motion, $h : \mathcal{Z} \to \mathbb{R}^d$ is a smooth vector-valued function, and $\gamma : \mathcal{Z} \to \mathbb{R}^{m\times d}$ a smooth matrix-valued function. Let $\mathcal{Z}$ be $\mathbb{T}^d$, $\mathbb{R}^d$ or $\mathbb{R}^l \oplus \mathbb{T}^{d-l}$.

Define the generator of the SDE above as

$$\mathcal{L}v = h \cdot \bigtriangledown v + \frac{1}{2}\Gamma : \bigtriangledown \bigtriangledown \gamma$$

where $\Gamma(z) = \gamma(z)\gamma(z)^T$.

Lemma 9.1 (Itô Formula)**.** *Assume that both $h(\cdot)$ and $\gamma(\cdot)$ are globally Lipschitz on $\mathcal{Z}$ and that z_0 is a random variable independent of the Brownian motion $W(t)$, and satifying $\mathbb{E}|z_0|^2 < \infty$. Let $z(t)$ solve the above Itô SDE, and let $V \in C^2(\mathcal{Z}, \mathbb{R})$. Then the process $V(z(t))$ satisfies*

$$V(z(t)) = V(z(0)) + \int_0^t \mathcal{L}V(z(s))ds + \int_0^t \langle \bigtriangledown V(z(s)), \gamma(z(s))dW(s)\rangle.$$

9.2 Burkholder-Davis-Gundy Inequality

Theorem 3.22 in [78].

Consider the Itô stochastic integral

$$I(t) = \int_0^t f(s)dW(s),$$

where $W(t)$ is a d-dimensional Brownian motion and $f(s) \in \mathbb{R}^{m\times d}$. We assume that $f(t)$ is a random process, adapted to the filtration $\mathcal{F}_t$ generated by the process $W(t)$, and such that

$$\mathbb{E}\left(\int_0^T f(s)^2 ds\right) < \infty,$$

and define the quadratic variation process

$$\langle I\rangle_t = \int_0^t (f(s)\otimes f(s))ds.$$

Theorem 9.2 (Burkholder-Davis-Gundy Inequality)**.** *Consider the above Itô stochastic integral, a martingale with quadratic variation process $\langle I\rangle_t$. For every $p > 0$ there are positive constants $C^{\pm}$ such that*

$$C^-\mathbb{E}|\langle I\rangle_t|^{\frac{p}{2}} \le \mathbb{E}\left(\sup_{0\le s\le t}|I(s)|^p\right) \le C^+\mathbb{E}|\langle I\rangle_t|^{\frac{p}{2}}.$$

9.3 The Gronwall Inequality

Lemma 4.4 in [78] .

Lemma 9.3.

- *(Differential form) Let $\eta(t) \in C^1([0,T];\mathbb{R}^+)$ satisfy the differential inequality*

$$\frac{d\eta(t)}{dt} \le a\eta(t) + \psi(t)\,,\ \eta(0) = \eta,$$

where $a \in \mathbb{R}$ and $\psi \in L^1([0,T]);\mathbb{R}^+$. Then

$$\eta(t) \le \exp(at)\left(\eta + \int_0^t \exp(-as)\psi(s)ds\right)$$

for $t \in [0,T]$.

- *(Integral form) Assume that $\xi(t) \in C([0,T];\mathbb{R}^+)$ satisfies the integral inequality*

$$\xi \le a\int_0^t \xi(s)ds + b,$$

for some positive constants a and b. Then

$$\xi(t) \leq b \exp(at)\, t \in [0, T].$$

9.4 Central Limit Theorems

This section quotes central limit theorems for random variables, martingales, and functionals of erogodic Markov chains.

Theorem 9.4 (Central Limit Theorem)**.** *[78] Let $\{\xi_n\}_{n=1}^{\infty}$ be a sequence of i.i.d. random variables with mean zero and variance 1. Define*

$$S_n = \sum_{k=1}^{n} \xi_k$$

Then the sequence

$$X_n = \frac{1}{\sqrt{n}} S_n$$

converges in distribution to a standard normal random variable.

Theorem 9.5 (Martingale Central Limit Theorem)**.** *[78] Let $\{M(t) : R^+ \to R^d\}$ be a continuous square integrable martingale on a probability space $(\Omega, \mathcal{F}, \mu)$ with respect to a filtration $\{\mathcal{F}_t : t \geq 0\}$; let $\langle M \rangle_t$ denote its quadratic variation process. Assume that:*

- (i) *$M(0) = 0$;*
- (ii) *the process $M(t)$ has continuous sample paths and stationary increments;*
- (iii) *the scaled quadratic variation of $M(t)$ converges in $L^1(\mu)$ to some symmetric positive-definite matrix Σ:*

$$\lim_{t \to \infty} \mathbb{E}\left(|\frac{\langle M \rangle_t}{t} - \Sigma| \right) = 0$$

Then the process $1/\sqrt{t} M_t$ converges in distribution to an $\mathcal{N}(0, \Sigma)$ random variable. Furthermore, the rescaled martingale

$$M^{\epsilon}(t) = \epsilon M\left(\frac{t}{\epsilon^2}\right)$$

converges weakly in $C_{\mathbb{R}^d}$ to $\sqrt{\Sigma} W(t)$, where $W(t)$ is a standard d-dimensional Brownian motion and $\sqrt{\Sigma}$ denotes the square root of the matrix Σ.

Below we cite the Central Limit Theorem for functionals of ergodic Markov chains. The first theorem is part of Theorem 2.3 in [16], which characterizes the limiting distribution for an erogdic Markov chain; the second theorem is Theorem 3.1 in [16], which identifies the limiting variance of the limiting Gaussian distribution.

We first define $\{X_n\}_{n\geq 0}$ is an ergodic Markov chain with invariant distribution π and $\xi : E \to \mathbb{R}$ is a measurable function, where E is a general state space. Write

$$S_n = \sum_{k=0}^{n-1} \xi(X_k) \; n = 1, 2, \dots \; .$$

Theorem 9.6 (Central Limit Theorem for functionals of ergodic Markov chains)**.** *[16] Let $\{X_n\}_{n\geq 0}$ be an ergodic Markov chain with and assume that*

$$\int \xi^2 \pi(dx) < +\infty \; .$$

Then $S_n/\sqrt{n} \to \mathcal{N}(0, \sigma^2)$ in distribution for some $\sigma^2 \geq 0$.

Theorem 9.7. *Let $\{X_n\}_{n\geq 0}$ be an ergodic Markov chain with and assume that*

(*i*) $\int \xi^2(x)\pi(dx) < +\infty$ *;*

(*ii*) $\sum_{n=1}^{\infty} \int \xi(x) P^n \xi(x) \pi(dx)$ *converges.*

Then,

$$S_n/\sqrt{n} \to \mathcal{N}(0, \sigma^2) \text{ in distribution}$$

holds for some $\sigma^2 \geq 0$. Further,

$$\sigma^2 = \int \xi^2(x)\pi(dx) + 2 \int \sum_{n=1}^{\infty} \xi(x) P^n \xi(x) \pi(dx)$$

if (i) holds and (ii) is strengthened into

$$(ii') \; \sum_{n=1}^{\infty} \xi(\cdot) P^n \xi(\cdot) \text{ converges in } L^1(\pi) \; .$$

9.5 The Blockwise Matrix Inversion Formula

[19]

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{pmatrix}$$

9.6 Hölder's inequality

[38]

Let $x \in \mathbb{C}^n$, define, for $p \geq 1$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}$$

For $x, y \in \mathbb{C}^n$,

$$\sum_{i=1}^{n} |x_i||y_i| \leq \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |y_i|^{p'} \right)^{1/p'}.$$

9.7 The Continuous-time Ergodic Theorem

Lemma 9.8. *[46] Fix a measurable space S, let (T_t) be a measurable flow on S with invariant σ-field $\mathcal{I}$, and let ξ be a (T_t)- stationary random element in S. Consider a measurable function $f : S \to \mathbb{R}$ with $f(\xi) \in L^p$ for some $p \geq 1$. Then as $t \to \infty$,*

$$t^{-1} \int_0^t f(T_t \xi) ds \to \mathbb{E}[f(\xi)|\xi^{-1}\mathcal{I}] a.s. and in L^p$$

An immediate result from the above ergodic theorem would be: Let

$$I = \frac{1}{\sqrt{T}} \int_0^T \psi(z(t)) dW(t),$$

Then there exists a constant $C > 0 : \mathbb{E}|I|^2 \leq C$ for all $T > 0$.

Proof. Use the Itô isometry and invoke the Lipschitz continuity of ψ. □

9.8 Some Quoted Properties of Linear Operator

Some linear operator properties from [80]. The existence and uniqueness of the solution of initial value problem regarding a bounded linear operator.

Theorem 9.9. *[80][p104] If A is the infinitesimal generator of a differentiable semigroup then for every $x \in X$ the initial value problem of the X valued function $u(t)$*

$$\frac{du(t)}{dt} = Au(t) \quad , \quad u(0) = x \quad t > 0$$

has a unique solution.

Since an operator on a finite-dimensional normed space is a bounded linear operator if and only if it is a continuous linear operator [93]. Our $\mathcal{L}_0$, $\mathcal{L}_1$ and $\mathcal{L}_2$ in Chapters 3 and 4 are clearly continuous, hence bounded linear operators. The boundedness of the solution regarding a linear operator.

Theorem 9.10. *[80][p118] Let A be the infinitesimal generator of an analytic semigroup e^{At}. If*

$$\sigma = \sup\{\mathrm{Re}\lambda : \lambda \in \sigma(A)\} < 0$$

then there are constants $M \geq 1$ and $\mu > 0$ such that $\|e^{At}\| \leq Me^{-\mu t}$.

where σ here stands for the spectrum of the linear operator A.

9.9 Eigenvalues of A Simple Matrix

Lemma 9.11. *For a matrix M of the following structure:*

$$M = \begin{pmatrix} 0 & 0 \\ m_1 & m_2 \end{pmatrix}$$

with m_2 negative definite, top-left zero sub-matrix and m_2 squares matrices, and m_2 invertible. Then the zero eigenvalues from the top-left zero submatrix are simple (ie. with algebraic multiplicity 1), and all other eigenvalues negative.

Proof. We first prove that all the zero eigenvalues from the top-left submatrix are simple. Since the top-left zero matrix is a square matrix, we assume the dimension is k. We construct the following vector of size k,

$$e_l = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

with 1 as the l^{th} entry $l \in \{1, \ldots, k\}$.

Then we can construct an eigenvector v_l through each e_l,

$$v_l = \begin{pmatrix} e_l \\ -m_2^{-1} m_1 e_l \end{pmatrix}$$

It is easy to check by the definition of eigenvectors that v_l's are proper eigenvectors for eigenvalue zero,

$$Mv_l = 0 = 0v_l$$

It is also easy to see that all the eigenvectors v_l and $v_{l'}$ are orthogonal if $l \neq l'$,

$$\langle v_l; v_{l'} \rangle = 0 \quad (l \neq l').$$

Therefore, each zero eigenvalue has independent eigenvector, so the zero eigenvalues are simple. The second statement is straightforward, since m_2 is negative definite, all it's eigenvalues are negative. □

9.10 An Inequality of Matrix Norm

Definition 9.12. *[40] Let $\|\bullet\|$ be a vector norm on $\mathbb{C}^n$. Define $\|\bullet\|$ on $\mathbb{C}^{n\times n}$ by*

$$\|A\| = \sup_x \|Ax\|$$

where $x \in \mathbb{C}^n$.

Theorem 9.13. *[40] The matrix norm defined in Definition 9.12 satisfies*

$$\|Ax\| \leq \|A\|\|x\|$$

and

$$\|I\| = 1.$$

If $\mathcal{L}$ is a linear operator applied on a square matrix A, and the following holds

$$\|e^{\mathcal{L}}tA\| \leq C\|A\|$$

then we say

$$\|e^{\mathcal{L}}tA\| \leq C$$

where t denotes time, C is a constant changes from occurrence to occurrence.

9.11 Order Preservation on the Diagonal of a Matrix

In the first case, we let $\mathbf{a}$ and $\mathbf{q}$ be as defined in (3.5), with $\mathbf{a}$ decomposed as

$$\mathbf{a} = PDP^{-1},$$

where D is a diagonal matrix of eigenvalues of $\mathbf{a}$,

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & \frac{1}{\epsilon} D_2 \end{pmatrix} ,$$

and U is the corresponding matrix of eigenvectors; and the diagonal diffusion matrix can be written as

$$\mathbf{q} = \begin{pmatrix} q_1 & 0 \\ 0 & \frac{1}{\epsilon} q_2 \end{pmatrix} .$$

hence,

$$\mathbf{q}^{-1} = \begin{pmatrix} q_1^{-1} & 0 \\ 0 & \epsilon q_2^{-1} \end{pmatrix} .$$

and consequently,

$$D\mathbf{q}^{-1} = \begin{pmatrix} D_1 q_1^{-1} & 0 \\ 0 & D_2 q_2^{-1} \end{pmatrix} = \mathcal{O}(1) .$$

In the second case, we let $\mathbf{a}$ and $\mathbf{q}$ be as defined in (4.6), with $\mathbf{a}$ decomposed as

$$\mathbf{a} = PDP^{-1} ,$$

where D is a diagonal matrix of eigenvalues of $\mathbf{a}$,

$$D = \begin{pmatrix} D_1 & 0 \\ 0 & \frac{1}{\epsilon^2} D_2 \end{pmatrix} ,$$

and P is the corresponding matrix of eigenvectors; and the diagonal diffusion matrix can be written as

$$\mathbf{q} = \begin{pmatrix} q_1 & 0 \\ 0 & \frac{1}{\epsilon^2} q_2 \end{pmatrix} .$$

hence,

$$\mathbf{q}^{-1} = \begin{pmatrix} q_1^{-1} & 0 \\ 0 & \epsilon^2 q_2^{-1} \end{pmatrix} .$$

and consequently,

$$D\mathbf{q}^{-1} = \mathbf{q}D^{-1} = \begin{pmatrix} D_1 q_1^{-1} & 0 \\ 0 & D_2 q_2^{-1} \end{pmatrix} = \mathcal{O}(1) .$$

Therefore, for both averaging and homogenization problems, we have $D\mathbf{q}^{-1} = \mathcal{O}(1)$, where D is diagonal matrix of eigenvalues of the drift matrix $\mathbf{a}$, and $\mathbf{q}$ the diffusion matrix.

Lemma 9.14. *We let* $(\mathbf{a}, \mathbf{q})$ *be as defined in* (3.5) *or in* (4.6)*, let* $\Sigma = U^{-1}\mathbf{q}(U^{-1})^T$*, then the elements on the diagonal of matrix* $(D\Sigma^{-1})_{ii} = \mathcal{O}(1)$*, and thus* $D_{ii}/\Sigma_{ii} = \mathcal{O}(1)$*.*

Proof. By the definition of Σ, we know

$$\Sigma = P^{-1}\mathbf{q}(P^{-1})^T ;$$

We know immediately that Σ is a real symmetric matrix, hence it can be decomposed using Singular Value Decomposition,

$$\Sigma = VSV^{-1}$$

where V is unitary, and S is a diagonal matrix of singular values. Due to symmetry of Σ, the singular values $S_{ii} = |\Lambda ii|$, for which Λii are the eigenvalues of Σ, and are of same orders as $\mathbf{q}$, which is east to prove. Hence,

$$D\Sigma^{-1} = DVS^{-1}V^{-1} = VDS^{-1}V^{-1} .$$

It follows with our first result

$$(D\Sigma^{-1})_{ii} = \mathcal{O}(1) \tag{9.1}$$

Then we write Σ as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} .$$

for which $\Sigma_{11} \in \mathbb{R}^{d_1 \times d_1}$, $\Sigma_{22} \in \mathbb{R}^{d_2 \times d_2}$ and $\Sigma_{12} \in \mathbb{R}^{d_1 \times d_2}$.

Using block matrix inversion formula in Appendix 9.5, we have the diagonal blocks of Σ^{-1},

$$\begin{aligned} (\Sigma^{-1})_{11} &= (\Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{12}^T)^{-1} \\ (\Sigma^{-1})_{22} &= \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{12}^T)^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{aligned}$$

By equation (9.1), we have

$$(\Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{12}^T)^{-1} = D_1\Sigma_{11}^{-1} = \mathcal{O}(1)$$

from which we can conclude

$$\Sigma_{11} = \mathcal{O}(1) \quad , \quad \Sigma_{12}\Sigma_{22}\Sigma_{12}^T = \mathcal{O}(1) .$$

Again by equation (9.1),

$$\begin{aligned} & \mathcal{O}\left(\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}\Sigma_{12}^{T})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\right) \\ = \quad & \mathcal{O}\left(\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}C\Sigma_{12}\Sigma_{22}^{-1}\right) \end{aligned}$$

where C is an order $\mathcal{O}(1)$ matrix, and this block matrix is of order $\mathcal{O}(\epsilon)$ for averaging and $\mathcal{O}(\epsilon^2)$ for homogenization. By previous equation

$$\Sigma_{12}\Sigma_{22}\Sigma_{12}^{T} = \mathcal{O}(1)$$

we have

$$\Sigma_{12}C\Sigma_{12}\Sigma_{22}^{-1} = \mathcal{O}(1)\ .$$

therefore in averaging,

$$\Sigma_{22}^{-1} = \mathcal{O}(\epsilon) \quad so \quad \Sigma_{22} = \mathcal{O}(\frac{1}{\epsilon})$$

and in homogenization

$$\Sigma_{22}^{-1} = \mathcal{O}(\epsilon^2) \quad so \quad \Sigma_{22} = \mathcal{O}(\frac{1}{\epsilon^2})$$

Consequently, we have

$$D_{ii}/\Sigma_{ii} = \mathcal{O}(1)\ .$$

□

Bibliography

[1] Aït-Sahalia, Y.; Mykland, P. A.; Zhang, L. A Tale of Two Time Scales: Determining Integrated Volatility With Noisy High-Frequency Data *Journal of the American Statistical Association* **100** (472) (2005) 1394-1411

[2] Aït-Sahalia, Y.; Mykland, P. A.; Zhang, L. How Often To Sample a Continuous-Time Process in the Presence of Market Microstructure Noise *The Review of Financial Studies* **18** (2005) 351-416

[3] Aldrich, J.. A paper on the history of Maximum Likelihood. "R.A. Fisher and the making of maximum likelihood 1912-1922". *Statistical Science* **12** (3) (1997). 162-176. DOI:10.1214/ss/1030037906

[4] Anderson, D. F.; Cracium, G.; Kurtz, T.. Product-form Stationary Distributions for Deficiency Zero Chemical Reaction Networks. *Bulletin of Mathematical Biology* **72** (8) (2010) 1947-1970. DOI: 10.1007/s11538-010-9517-4

[5] Azencott, R; Beri, A; Timofeyev, I. Adaptive Sub-sampling for Parametric Estimation of Gaussian Diffusions. *J. Stat. Phys* **139** (6) (2010) 1066-1089

[6] Azencott, R; Beri, A; Timofeyev, I. Parametric Estimation of Stationary Stochastic Processes under Approximate Observability. *Journal of Statistical Physics* **144** (1) (2010) 150-170

[7] Azencott, R; Beri, A; Timofeyev, I. Sub-sampling and Parametric Estimation for Multiscale Dynamics. *SIAM J. Multiscale Modeling and Simulation* (2010)

[8] Bartels, R.H. ; Stewart, G.W. Solution of the matrix equation $AX + XB = C$. *Comm. ACM*. **15** (1972) 820-826.

[9] Bensoussan, A. Stochastic Control of Partially Observable Systems. Cambridge University Press 1992

[10] Bensoussan, A.; Lions, J.-L.; Papanicolaou. G. Asymptotic analysis for periodic structures. *Studies in Mathematics and its Applications* **5** (1978)

[11] Bishwal, J. P.N. Parameter Estimation in Stochastic Differential Equations Springer-Verlag Berlin 2008 ISBN 978-3-540-74447-4

[12] Blanco, C; Oks, M. Backtesting VaR models: Quantitative and Qualitative tests *The Risk Desk* **IV** (1) (2004)

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

[14] Bollerslev, T. Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model *The Review of Economics and Statistics* **72** (1990) 498-505

[15] Bollerslev, T; Engle, R.F.; Nelson, D.B. ARCH Models. *Handbook of Econometrics* **4** (1994) 2959-3038

[16] Chen, X. Limit theorems for functionals of ergodic Markov Chains with general state space. *Memoirs of the AMS* **129**, (664) (1999)

[17] Chorin, A. J.; Hald, O. H. Stochastic Tools in Mathematics and Science. Surveys and Tutorials in the Applied Mathematical Sciences (Volume 1). Springer, New-York, 2006. ISBN 0-387-28080-4.

[18] Christoffersen, P. Evaluating Interval Forecasts *International Economic Review* **39** (1998) 841 - 862

[19] Cormen, T. H. *et al.* Introduction to Algorithms. (Second Edition) MIT Press and McGraw-Hill, 2001. ISBN 0-262-03293-7

[20] Crisan, D. Stochastic Analysis 2010 Springer, 2010

[21] Cryer, J.; Chan, K. S. Time Series Analysis Springer 2008

[22] Davis, M.H.A. Linear Estimation and Stochastic Control Chapman and Hall 1977

[23] Diebold, F. X.; Li, C.. Forecasting the Term Structure of Government Bond Yields *Journal of Econometrics* **130** (2006) 337-364

[24] Diebold, F. X.; Li, C.; Yue, V. Z. Global Yield Curve Dynamics and Interactions: A Generalized Nelson-Siegel Approach *Journal of Econometrics* **146** (2008) 351-363

[25] Engle, R. Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of UK Inflation *Econometrica* **50** (1982)

[26] Engle, R. Dynamical conditional correlation - a simple class of multivariate GARCH models *National Bureau for Economic Research Working Paper* **2000-09** (2000)

[27] Engle, R.; Sheppard, K. Theoretical and empirical properties of dynamic conditional correlation multivariate GARCH *National Bureau for Economic Research Working Paper* **8554** (2001)

[28] Gardiner, C.W. Handbook of Stochastic Methods, for Physics, Chemistry and the Natural Sciences (Third Edition) Springer-Verlag, Berlin, 2003. ISBN 3-540-20882-8.

[29] Gelfand, A. E.; Smith, A.F.M. Sampling Based Approaches to Calculating Marginal Densities *Journal of the American Statisticl Association* **85** (1990) 398-409

[30] Geweke, J. Bayesian Treatment of the Independent Student *t* Linear Model *Journal of Applied Econometrics* **8** (1993) 19-40

[31] Geweke, J. Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments *Bayesian Statistics* **4** (1992)

[32] Gilks, W.R; Spiegelhalter, D.J; Richardson S. Markov Chain Monte Carlo in Practice Chapman and Hall, London, 1996

[33] Givon, D. ; Kevrekidis, I.G.; Kupferman, R. Strong convergence schemes of projective intregration schemes for singularly perturbed stochastic differential equations. *Comm. Math. Sci.* **4** (4) (2006) 707729

[34] Graversen, S. E.; Peskir, G. Maximal Inequalities for the Ornstein-Uhlenbeck Process *Proceedings of the American Mathematical Society* **128** (10) (2000) 3035-3041

[35] Grimmett, G.; Stirzaker, D. Probability and Random Processes (Third Edition) Oxford University Press, New York, 2001. ISBN 0-19-857222-0

[36] Hall, P.; Heyde, C.C Martingale limit theory and its application. Academic Press, New York, 1980. ISBN 0-12-319350-8

[37] Hansen, P. R.; Lunde, A. A forecast comparison of volatility models: Does anything beat a GARCH(1,1)? *Journal of Applied Econometrics* **20** (2005) 873-889

[38] Hardy, G.H.; Littlewood, J.E.;Plya, G. Inequalities. Cambridge University Press 1934

[39] Higham, D. J. ;Mattingly, J. C.;Stuart, A. (2001) Ergodicity for SDEs and Approximations: Locally Lipschitz Vector fields and Degenerate Noise. *Stochastic Processes and Applications* **101** (2002) 185–232

[40] Horn, R. A.; Johnson, C. R. Matrix Analysis Cambridge University Press 1985. ISBN 0-521-38632-2

[41] Ichihara, N. Homogenization for Stochastic Partial Differential Equations Derived from Nonlinear Filterings with Feedback. *J. Math. Soc. Japan* **57** (2005) 593603.

[42] Isaacson, E.; Keller, H. B. Analysis of Numerical Methods Wiley & Sons. 1966.

[43] Jackman, S. Bayesian Analysis for the Social Sciences Wiley Series in Probability and Statistics John Wiley & Sons 2009 ISBN 978-0-470-01154-6

[44] Joyce, J. "Bayes' Theorem" Stanford Encyclopedia of Philosophy. The Metaphysics Research Lab Center for the Study of Language and Information Stanford, 2003. ISSN 1095-5054 http://plato.stanford.edu/

[45] Jukić, D. The matrix of a linear operator in a pair of ordered bases. *Mathematical Communications* **2** (1997) 77-82

[46] Kallenberg, O. Foundations of Modern Probability (Second Edition) Springer, Berlin, 2002. ISBN 7-03-016672-8

[47] Kallianpur, G. Stochastic Filtering Theory Springer-Verlag, 1980

[48] Kato, T. Perturbation Theory for Linear Operators (Corrected Printing of the Second Edition) Springer-Verlag, 1980

[49] Kay, S. M. Fundamentals of Statistical Signal Processing: Estimation Theory. Prentice Hall, New Jersey, 1993. ISBN 0-13-345711-7.

[50] Kloeden, P. E. ; Platen, E. Numerical Solution of Stochastic Differential Equations. Springer-Verlag, New-York, 1992. ISBN 3-540-54062-8.

[51] Koval, V. A. On the Strong Law of Large Numbers for Multivariate Martingales with Continuous Time. *Ukranian Mathematical Journal* **53** (9) (2001) 1554-1560 ISSN: 0041-5995(Print),1573-9376 (Online)

[52] Küchler, U; Sørensen, M. A Note on Limit Theorems for Multivariate Martingales. *Bernoulli* **5** (3) (1999) 483-489

[53] Kupiec, P. Techniques for Verifying the Accuracy of Risk Measurement Model *Journal of Derivatives* **3** (1995) 73-84

[54] Kutoyants, Y. A. Statistical inference for ergodic diffusion processes. Springer Series in Statistics Springer-Verlag, London, 2004. ISBN 1-85233-759-1

[55] LeSage, J. P. Applied Econometrics using MATLAB and the "Econometric Toolbox". *supported by National Science Foundation BCS-0136229*

[56] LeSage, J. P. Theory and practice of spatial econometrics. *Unpublished manuscript available at www.spatial-econometrics.com* (1999)

[57] LeSage, J. P; Pace, R. K. Spatial and spatiotemporal econometrics. Chapman & Hall, London, 2004.

[58] Lindley, D. V. The Estimation of Many Parameters. *Foundations of Statistical Science* V.P.Godambe and D.A. Sprout (eds) Toronto, (1971)

[59] Litterman, R. Forecasting with Bayesian Vector Autoregressions: Five years of experience. *Journal of Business and Economic Statistics* **4** (1) (1986)

[60] Litterman, R. Specifying vector autoregressions for macroeconomic forecasting *Bayesian Inference and Decision Techniques with applications: Essays in Honor of Bruno De Finetti (North-Holland, Amsterdam)* (1986) 79-84.

[61] Liu, W. E, D.; Vanden-Eijnden, E. Analysis of multiscale methods for stochastic differential equations. *Comm. Pure Appl. Math.* **58** (11) (2005) 15441585

[62] Majda, A.J. ; Harlim, J.;Gershgorin, B. Mathematical Strategies for Filtering Turbulent Dynamical Systems *Discrete and Continuous Dynamical Systems* **27** (2) (2010) 441-486

[63] Mandelbrot, B. The Variation of Certain Speculative Prices. *Journal of Business* **36** (1963) 394

[64] Mandelbrot, B. New Methods in Statistical Economics *Journal of Political Economy* **71** (1963) 421-440.

[65] Marchetti, C. Asymptotic expansion in nonlinear filtering with homogenization *Appl. Math. Optim.* **41** (3) (2000) 403424.

[66] Miao, W. C. Estimation of diffusion parameters in diffusion processes and their asymptotic normality *International Journal of Comtemporary Mathematical Science* **1** (16) (2006) 763-776

[67] Nelson, C.R; Siegel, A.F. Parismonious Modelling of Yield Curves *Journal of Business* **60** (4) (1987) 473-489

[68] Øksendal, B. K. Stochastic Differential Equations (Fifth edition) Springer, Berlin, 1998. ISBN 3-540-63720-6.

[69] Olhede, S; Pavliotis, G. A. ; Sykulski, A. Multiscale Inference for High Frequency Data *preprint, 2008*

[70] Orskaug, E. Multivariate DCC-GARCH Model - With Various Error Distributions *Norsk Regnesentral* Publication ID: SAMBA/19/09 (2009)

[71] Papanicolaou, G C., Stroock, D W., Varadhan, S R S. Martingale approach to some limit theorems. In Papers from the Duke Turbulence Conference (Duke University, Durham, N.C., 1976), Paper No. 6, pages ii+120 pp. Duke Univ. Math. Ser., Vol. III. Duke Univ., Durham, N.C., (1977)

[72] Papavasiliou, A.; Pavlioitis, G. A.; Stuart, A. M. Parameter Estimation for Multiscale Diffusions *J. Stat. Phys.* **127** (4) (2007) 741-781

[73] Papavasiliou, A.; Pavlioitis, G. A.; Stuart, A. M. Maximum Likelihood Drift Estimation for Multiscale Diffusions *Stoch. Proc. Appl.* **119** (10) (2009) 3173-3210

[74] Pavliotis, G.A.; Stuart, A.M. White Noise Limits for Intertial Particles in a Random Field. *SIAM J. Multiscale Modeling and Simulation* **1** (4) 527-553 (2003)

[75] Parada, L. M.; Liang, X. Optimal multiscale Kalman filter for assimilation of near-surface soil moisture into land surface models. *Journal of Geophysical Research* **109** (2004)

[76] Park, J. H.; Sowers, R. B.; Namachchivaya, N. S. Dimensional Reduction in Nonlinear Filtering *Nonlinearity* **23** (2010) 305324

[77] Park, J. H.; Namachchivaya, N. S.; Sowers, R. B. A Problem in Stochastic Averaging of Nonlinear Filters. *Stoch. Dyn.* **8** (2008) 543560

[78] Pavliotis, G. A. ; Stuart, A. M. Multiscale Methods: Averaging and Homogenization. Series: Texts in Applied Mathematics , Vol. 53 Springer, New York, 2008. ISBN: 978-0-387-73828-4

[79] Pavliotis, G. A. ; Stuart, A. M. Parameter Estimation for Multiscale Diffusions *J. Stat. Phys.* **127** (2007) 741-781

[80] Pazy, A. Semigroups of Linear Operators and Applications to Partial Differential Equations, Applied Mathematical Sciences Volume 44, Springer-Verlag, New York, 1983. ISBN 0-387-90845-5

[81] Prakasa Rao, B. L. S. Semimartingales and their statistical inference. Monograph on Statistics and Applied Probability, 83. Chapman & Hall/CRC, Boca Raton, FL, 1999 ISBN 1-58488-008-2

[82] Raftery, A.E.; Lewis, S.M. One long run with Diagonstic: Implementation strategies for Markov Chain Monte Carlo *Statistical Science* **7** (1992) 493-497

[83] Raftery, A.E.; Lewis, S.M. How many iterations n the Gibbs Sampler? *Bayesian Statistics* **4** (1992) 963-773

[84] Reveuz, D.; Yor, M. Continuous martingales and Brownian motion (Third Edition) Springer-Verlag, Berlin, 1999. ISBN 3540643257 Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences] . **293**

[85] Sevinc, V.; Ergun, G. Usage of Different Prior Distributions in Bayesian Vector Autoregressive Models *Hacettepe Journal of Mathematics and Statistics* **38** (1) (2009) 85-93

[86] Sheppard, K. Multi-step estimation of multivariate GARCH models *Proceedings of the International ICSC Symposium: Advanced Computing in Financial Markets.* (2003)

[87] Silvennoinen, A.; Teräsvirta, T. Multivariate GARCH models *Handbook of Financial Time Series* Springer, 2008

[88] Sims, C. A.; Zha, T. Bayesian Methods for Dynamic Multivariate Models *International Economic Review* **39** (4) (1998) 949-68

[89] Teräsvirta, T. Two stylised facts and the GARCH (1,1) model. *Working Paper Series in Economics and Finance, Stockholm School of Economics* **96** (1996)

[90] Theil, H.; Goldberger, A. S. On Pure and Mixed Statitical Estimation in Economics. *International Economic Review* **2** 65-78

[91] Ungarala, S.; Bakshi, B. R. A Multiscale, Bayesian and error-in-variables approach for linear dynamic data rectification. *Computers and Chemical Engineering* **24** (2000) 445-451

[92] Vanden-Eijnden, E. Numerical Techniques for Multi-scale Dynamical Systems with Stochastic Effects. *Commun. Math. Sci.* **1** (2) (2003) 385391

[93] Walter, R. Functional Analysis(Second edition), Science / Engineering / Mathematics, McGraw Hill, 1991. ISBN 0-07-054236-8.