Dynamic Covariance Modelling in High Dimensions

Haosheng Tan

SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE BY RESEARCH

SCHOOL OF MATHEMATICS AND STATISTICS

COLLEGE OF SCIENCE AND ENGINEERING

July 27, 2026

Abstract

Covariance is a fundamental measure to describe the dependency structure between two random variables and covariance modelling is significant in a wide range of applications. Existing systems mainly assume the covariance is static even though covariance structure is typically dynamic in real-world settings, which indicates the need of dynamic covariance modelling. Two main approaches are mainly adopted in covariance parameterisation: covariance-function-based methods and matrix-decomposition-based methods. Covariance-function-based approaches describe the covariance structure with parsimonious parameters, thereby effectively solve the the rapid growth in the number of parameters with the increment of covariance dimension but they may fail to guarantee positive definiteness in non-stationary settings. Covariance decomposition methods, on the other hand, offer unconstrained factors that guarantee positive-definite output but suffer from the quadratic increase of estimated parameters and poor statistical interpretability. As a result, methods that simultaneously ensure positive definiteness and maintain parsimonious parameterisations remain largely unexplored. Motivated by these limitations, we propose a flexible estimation framework for covariance functions by adopting positive-definite regularisations of the objective functions. This framework imposes penalties on eigenvalues or the determinant of the estimated covariance matrix, thereby enforcing positive-definiteness with no additional estimated parameters required. This framework effectively addresses the positive-definite issue for covariance functions while maintaining its strengths, which makes it suitable for high-dimensional dynamic covariance modelling. Simulation studies and real-data applications demonstrate that the proposed framework consistently improves estimation accuracy while ensuring positive-definite covariance estimates across a range of stationary and non-stationary covariance structures.

Contents

Chapter 1

Introduction

Covariance is a fundamental statistical concept that characterises the dependence structure between variables and plays a crucial role in multivariate analysis. In many real-world applications, such as multivariate probabilistic forecasting, covariance structures are dynamic, presenting patterns that evolve over time and accurate estimation require capturing this specific dynamics. Despite this prevalence, a general and flexible framework for modelling high-dimensional dynamic covariance structures remains underdeveloped. Existing approaches to describe covariance structure primarily fall into two categories: covariance-function-based methods (Browell et al., 2022; Gneiting et al., 2006; Higdon et al., 2023; Bonat and Jørgensen, 2016) and unconstrained matrix decomposition techniques (Gioia et al., 2022; Pinheiro and Bates, 1996). Covariance-function-based methods offer explicit representations of dependence relationships while requiring the estimation of a relatively small number of parameters. By contrast, matrix decomposition methods permit greater modelling flexibility through unconstrained parameterisation, but at the cost of a quadratic growth in the number of parameters to be estimated.

One significant limitation of covariance-function-based methods is that, under unconstrained optimisation, they do not guarantee the positive definiteness of the estimated covariance matrix in non-stationary settings, which occurs frequently in practice. Since positive definiteness is crucial for valid covariance matrices, this limitation can cause instability and make estimation unfeasible. To address this issue, we propose a positive-definite estimation framework suitable for non-stationary covariance functions such that the estimated covariance matrices throughout stochastic optimisation are enforced to be positive-definite, thereby enabling the application of robust optimisation algorithms commonly used in machine learning.

This thesis provides a comprehensive overview of both static and dynamic covariance estimation methods, covering isotropic and non-stationary spatial covariance settings. Primary methodologies for covariance matrix estimation are reviewed alongside their associated challenges. The main contributions of this work include the development of a constrained optimisation framework that guarantees positive-definite solutions for non-stationary covariance functions within stochastic optimisation procedures. Empirical results demonstrate that the proposed framework consistently improves goodness of fit, achieving an average accuracy improvement of 37% while reducing estimation variability for both the powered exponential covariance function and its extended variants. These results indicate strong generalisation performance and improved estimation accuracy. In addition, we are interested in the time-varying covariance modelling for multivariate probabilistic

forecasting tasks including various regional variables (see Chapter 4), on which we evaluate the proposed method with detailed results and analysis.

1.1 Covariance Matrices

Before introducing dynamic covariance modelling, we first review the definition of covariance matrices and covariance functions. The general concept of covariance functions as well as their link to stochastic process are briefly discussed. Covariance is a statistic describing the relationship between two random variables, and the covariance matrix is a multivariate statistic that organizes all the variances and covariances for a set of random variables into a single, square matrix. Consider two random variables X_1 and X_2. Covariance is defined as

$$\mathrm{Cov}(X_1, Y_2) = \mathbb{E}\left[(X_1 - \mathbb{E}(X_1))\,(X_2 - \mathbb{E}(X_2))\right] \tag{1.1}$$

where $\mathbb{E}(X_1)$ and $\mathbb{E}(X_2)$ are the expectation of X_1 and X_2, respectively. This statistics describes how two variables change together and indicates their relationship. If $\mathrm{Cov}(X_1, X_2) > 0$, then values above the expectation of X_1 indicate values above the expectation of X_2 while $\mathrm{Cov}(X_1, X_2) < 0$ implies the increment of X_1 leads to a decrease of X_2. $\mathrm{Cov}(X_1, X_2) = 0$ indicates there is no correlated relationship between X_1 and X_2. In multivariate cases, define a random vector $X = [X_1, X_2, \ldots, X_N]$ consisting of N random variables. The covariance matrix Σ is defined as

$$\Sigma = \mathbb{E}\left[(X - \mathbb{E}(X))\,(X - \mathbb{E}(X))^\top\right] \tag{1.2}$$

where $\mathbb{E}(X)$ is the expectation of random vector. The diagonal of Σ denotes the variances of the corresponding random variable, and the off-diagonal entries are the covariances between pairs of the variables. Covariance matrix is symmetric according to the definition, meaning $\Sigma = \Sigma^\top$, and positive semi-definite, meaning that for any vector $\alpha \in R^d$, the quadratic form

$$\alpha^\top \Sigma \alpha = \mathrm{Var}(\alpha^\top X) \tag{1.3}$$

is non-negative. Eigenvalues and eigenvectors are two significant concepts to indicate the positive definiteness of a square matrix and their definitions are given in Definition (1.4)

Eigenvalues and Eigenvectors Let $\Sigma \in R^{d\times d}$ be a square covariance matrix. A non-zero vector $\mathbf{v} \in R^d$ is called an eigenvector of Σ if there exists a scalar $\lambda \in R$ such that:

$$\Sigma \mathbf{v} = \lambda \mathbf{v}, \tag{1.4}$$

where λ is an eigenvalue. $\Sigma\mathbf{v}$ represents the transformation of the vector $\mathbf{v}$ under matrix Σ and the condition implies that the action of Σ on $\mathbf{v}$ scales it by factor of λ without changing its direction (Jolliffe, 2002). Eigenvectors represent directions in the vector space that remain invariant under the linear transformation induced by Σ. Specifically, the transformation may stretch or compress a vector, but it does not change its direction. The associated eigenvalues quantify the extent of this stretching or compression. The properties stated in (1.3) further imply that all eigenvalues of Σ are greater than or equal to zero, indicating that Σ is positive semi-definite. In this context, eigenvalues characterise the magnitude of variance explained, where λ_i denotes the variance captured by the i-th eigenvector, or equivalently, the i-th principal component. Consequently, if all such scalar values are non-negative, as shown in (1.3), it can be claimed that Σ is positive semi-definite.

1.2 Covariance Function

A covariance function $C(\cdot)$ is a generalisation of covariance from a fixed set of random variables to random processes defined over a continuous domain. Define a stochastic process $Z(x)$ for every x in domain D, its covariance function is defined as

$$C(x_1, x_2) = \text{Cov}(Z(x_1), Z(x_2)) = \mathbb{E}\left[(Z(x_1) - \mu(x_1))\,(Z(x_2) - \mu(x_2))\right] \tag{1.5}$$

where $\mu(x) = \mathbb{E}[Z(x)]$ is the mean function of the process. The demonstration of how covariance function description a covariance matrix is shown as:

$$\Sigma = \begin{bmatrix} C(x_1, x_1) & C(x_1, x_2) & \cdots & C(x_1, x_d) \\ C(x_2, x_1) & C(x_2, x_2) & \cdots & C(x_2, x_d) \\ \vdots & \vdots & \ddots & \vdots \\ C(x_d, x_1) & C(x_d, x_2) & \cdots & C(x_d, x_d) \end{bmatrix} \tag{1.6}$$

Covariance functions measure how the values of the process at different domain values co-vary, and it is symmetric according to the definition. Also, covariance functions are positive semi-definite, which indicates for any subspaces of the domain, the subsets of the output values from the function form a covariance matrix. Any samples from the continuous process form covariance matrix of the corresponding random vector. For covariance matrix estimation, stationarity is a significant component which is defined as (1.7).

Stationarity Stationarity in stochastic processes refers to the property where the statistical characteristics of the process remain constant over time (Miller and Papoulis, 1966). For a general random field or stochastic process indexed by x, stationarity is defined with respect to shifts in the index variable. For instances, a process Z is stationary when the mean $\mu = \mathbb{E}[Z(x)]$ and variance σ^2 are constant and the covariance between x and y depends only on their separation $x - y$:

$$\text{Cov}(Z(x), Z(y)) = C(x - y). \tag{1.7}$$

In real-world scenarios, the stationarity can be expressed either in spatial or temporal domain. For spatial stationary stochastic process, the separation $x - y$ is defined as the spatial distance while temporal stationarity indicates that covariance between two timestamp is only dependent on the temporal gap. In this thesis, we only consider spatial stationarity and assume there is no stationarity assumption on temporal covariance, which means x and y are both spatial locations. And we are interested in dynamic covariance modelling that is time-varying spatial covariance. Covariance in spatial stationarity can be categorised into the following two types:

Isotropic Covariance Matrix An isotropic process is stationary and exhibits properties that are identical in all directions. In spatial statistics, this means that the spatial dependence between observations depends solely on the distance between them, not on the direction. Regarding the isotropic covariance matrix, the elements are based only on the separation r of two corresponding random variables $Z(x)$ and $Z(y)$ where $Z(x), x \in D$ and $Z(y), y \in D$ are random processes.

$$\text{Cov}(Z(x), Z(y)) = C(Z(x), Z(x + r)) = C(r) \tag{1.8}$$

$$r = \|y - x\| \tag{1.9}$$

For stationary spatial covariance matrix, we assume there is a fixed number of spatial locations and the process Z is assumed to be spatially stationary and isotropic but temporally non-stationary. The isotropic covariance matrix example is visualized in Figure 1.1

Non-stationary Covariance Matrix A non-stationary process is one whose statistical properties, such as mean, variance, and autocorrelation, change over time or space. This contrasts with stationary processes, where these properties remain constant over time or space shifts. In a non-stationary matrix, the covariance of two random variables is not only dependent on the separation r but also the absolute position in D, which is shown as

$$\text{Cov}(Z(x), Z(y)) = C(x, y) \tag{1.10}$$

Non-stationary covariance is a more general case in spatial-temporal applications where certain locations may preserve dominating covariance with other locations notwithstanding high 'separation'. A non-stationary covariance matrix is visualized in Figure 1.1. Covariance functions for these two circumstances are discussed in Section 2.1.1

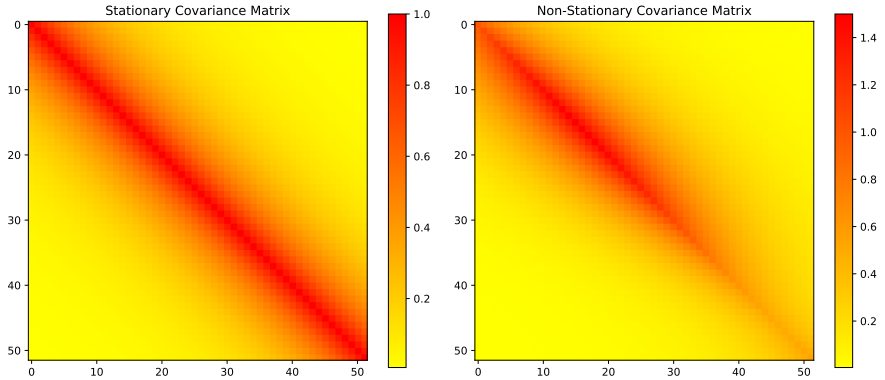

Figure 1.1: The visualisation of stationary covariance matrix (left) and non-stationary covariance matrix (right)

1.3 Applications of Covariance Modelling

Covariance modelling is widely used in various real-world application spanning Neurocience, Finance, Wind Power Forecasting and others. This section summarises several applications and corresponding methods from the literatures. For the literature review perspective, this section will include applications both on time-varying spatial covariance and temporal covariance. For the later chapters of simulations and real-world applications, we focus on the time-varying spatial covariance modelling for probabilistic forecasting tasks in Energy Sector.

1.3.1 Electricity Demand Forecasting

Electricity Demand Forecasting is a multivariate probabilistic forecasting task that require dynamic covariance modelling as prediction involving multiple regions. Net load forecasting is an indispensable process for power sectors and has been increasingly unpredictable due to the elusive

pattern of the energy industry. The infrastructure is being upgraded rapidly and renewable energy generation and usage are given increasing priority, which makes the supply, demand and prices become more volatile and less predictable than ever before (Hong et al., 2016). Thus, probabilistic forecasting that quantifies the uncertainty is becoming a major task in electricity net-demand forecasting and short-term probabilistic forecasting is increasingly needed in modern electricity distribution system (Hong and Fan, 2016). Electricity networks are transforming from a centralized system, where electricity is generated from large power plants and transmitted to certain regions, to a decentralized mode, where the power generating system is directly connected to the distribution system (Hong et al., 2016). Such a net-connected power generative system requires power flow forecasting across regions and spatially coherent joint forecasts instead of performing forecasting region by region. Therefore, covariance estimation is a part of electricity demand forecasting and short-term probabilistic forecasting requires dynamic covariance modelling.

Among predictors for covariance modelling in electricity demand forecasting, additive models are popular as they get high performance while maintaining high explainability (Hong et al., 2016). For probabilistic forecasting, covariates are included to estimate the distribution parameters; usually they are mean and covariance for multivariant Gaussian assumption. While the number of regions becomes large, the parameters to be estimated grows quadratically, making it unreasonable to select a subset of covariates for each of the parameter manually. Gradient Boosting, which allows to fit model to the training data while performing features selection, is adopted by Gioia et al. and this method provides an practical solution to non-parametric model estimating a large amount of distribution parameters.

This application is the one we apply regularised non-stationary covariance functions on and will be discussed in Chapter 4. In general, probabilistic short-term electricity demand forecasting shows a great need of dynamic covariance modelling and some previous work adopt Modified Cholesky Decomposition and additive models for covariance parameterisation and estimation. Various and novel dynamic covariance methods are expected to be applied on this application. This thesis adopts the proposed covariance estimation framework for non-stationary covariance functions to a electricity net load data covariance modelling task. Multivariate probabilistic metrics, such as Energy Score and p-Variogram Score, are used to evaluate whether accurate covariance modelling improve forecasting accuracy.

1.3.2 Neurocience

Covariance estimation is applied in the bio-medical field. One significant application is to estimate the functional connectivity between brain regions. Inverse covariance models are applied in estimating the functional connectivity in the human brain. Colclough et al. (2018) propose a hierarchical structure in estimating the precision matrix of multi-subject functional connectivity. The covariance modelling method they used allows each element of the precision matrix to have its own prior, which imposes sparsity of the matrix. And they use a linked regression model for modelling each column of the precision matrix. Their Bayesian framework introduces both strongly and weakly sparse priors to model functional brain connectivity, enabling flexible regularisation of precision matrices. By jointly estimating individual- and group-level networks, the method improves inference accuracy, particularly with limited data. Utilizing a conditional linear regression formulation, the model supports efficient block-Gibbs sampling, ensuring scalability to large neuroimaging datasets such as fMRI and MEG.

Chen and Leng (2016) study the sparse dynamic covariance matrix and they propose dynamic covariance models by replacing the weight of sample covariance function $\frac{1}{n}$ to a kernel function of dynamical variable, e.g. $K(t)$ for time-variate kernel function. This model is applied for attention deficit hyperactivity disorder (ADHD) study. Their method estimates blood flow signal of different regions of interest(ROI) in brain and manages to capture temporally varying information in fMRI data, which provide insight into the fundamental workings of brain networks.

1.3.3 Portfolio optimisation

Estimating the covariance matrix of equity returns is a pivotal exercise in empirical finance, informing portfolio construction and the assessment of asset-pricing theories. One publicly available example is Monte Prediction—a weekly challenge where participants forecast the joint distribution of next week's returns for eleven financial sector ETFs. Conventional covariance matrix of stock returns is estimated by optimally weighted average of two existing estimators: the sample covariance matrix and single-index covariance matrix capturing the dependencies between two stocks in single-index model. One challenge comes with the lack of samples where the number of stocks is the same order of magnitude as the number of historical returns per stock, which leads to singular sample covariance matrix. Ledoit and Wolf (2003) address this problem through the definition the optimal shrinkage intensity by minimizing a loss function that does not involve the inverse of the covariance matrix. The general form of their method is

$$\hat{\Sigma} = \alpha F + (1 - \alpha)S \tag{1.11}$$

where S is the sample covariance matrix and F is the single-index model covariance matrix F. This method combines unbiased covariance S and low variance F using weighted average, striking a balance between bias and variance. It provides a rigorous and automatic way to control overfitting without relying on arbitrary factor choices and it solve the high-dimensional covariance matrix estimation in financial econometrics. But this method assumes the covariance structure is constant over time and dynamic changes in volatility is not captured. Ledoit and Wolf (2017) extend this method by introducing non-linear shrinkage. It apply shrinkage function to each eigenvector individually and propose a portfolio selection-specific loss function based on out-of-sample portfolio variance. However, this method remains static and estimates a single covariance matrix $\hat{\Sigma}$. GARCH-based methods (Yu et al., 2020) are also adopted to capture the dependencies between stocks and markets as well as events dynamically. Multivariate GARCH could be simplified as

$$r_t = \mu + \epsilon_t, \ \epsilon_t | \mathcal{F}_{t-1} \sim N(0, H_t) \tag{1.12}$$

where μ is the unconditional mean vector and ϵ_t is the vector of shocks or residuals. H_t is a dynamic $N \times N$ matrix at time t which is the core to be estimated. This type of method is widely used in capturing dynamics in finance because it efficiently captures the high and low volatility and transmission of shocks between markets. But this method also come across the parameterisation explosion problem since the number of parameters increases at rate $O(N^2)$ or even worse when the assets increases. Therefore, a more computationally efficient way is yet to be adopted in portfolio optimisation.

1.3.4 Wind Power Forecasting

Wind power forecasting, as one of the typical forecasting tasks relies on spatial and temporal features. Owing to the complexity of the various decision-making problems at hand, it is preferable that the forecasts provide decision-makers not only with an expected value for future power generation, but also with associated estimates of prediction uncertainty (Tastu et al., 2015). Multivariate probabilistic forecasting has many use cases in power system operation and energy trading. Spatial dependency is important for managing network constraints, and temporal dependency for scheduling storage and conventional generation (Browell et al., 2022). It is essential for the decision-making scenario to have wind power forecasting in different temporal and spatial scales and wind power forecasting is indispensable in wind storage system as well as making price decision.

Previous probabilistic wind power forecasting generates expected values along with uncertainties at each site individually, which generates distribution for each site instead of estimating a joint probability distribution. These forecasting methods are partly due to the covariates used for modelling since most of the covariates are site-specific (Tastu et al., 2015). However, spatial and temporal dependencies of different wind farms at each time are ignored, which results in a less accurate prediction and it also make it challenging to track error propagation since wind power forecast errors propagate in time and in space under the influence of meteorological conditions. Based on the probabilistic forecasting made by each of the site, Tastu et al. (2015) obtain multivariate probability densities by Gaussian Copula that links different variables with its own marginal with a Gaussian inverse accumulative probability distributive function. They manage to capture temporal and spatial dependencies by treating each lead time and location as a single random variable. They also illustrates wind power temporal correlation is limited within a short temporal window and regional dependencies are extremely sparse. Based on this observation, they estimate the precision matrix by setting the structure of precision matrix parameterisation manually. Nevertheless, it still remains static as the estimated covariance is within a limited locations and lead time. Although a sliding window strategy makes it adaptive to different temporal regions, different seasons for example, it is also a static method for short-term forecasting and requires re-estimating while temporal features change. Browell et al. (2022) propose an additive framework and apply it on dynamic wind power covariance modelling. It leverages covariance function for covariance matrix parameterisation and exhibits a more explicit model for dynamic estimation. But ensuring positive definite is a thorny problem in non-stationary situation and an effective constrained optimisation is critical.

1.3.5 Ecology

Covariance estimation is fundamental to understanding complex ecological dynamics, particularly in modelling species interactions (Latimer et al., 2006), environmental influences, and temporal dependencies. Static covariance models are used when systems are assumed to be time-invariant or observations are cross-sectional, whereas dynamic models capture the evolution of covariances over time. Recent advances include Bayesian hierarchical modelling (Hooten et al., 2015), spatio-temporal models (Subba Rao, 2012), and machine learning-enhanced covariance estimators (Hengl et al., 2017). Applications range from biodiversity modelling and habitat selection to epidemic tracking and climate-driven processes. However, most of these applications utilize static covariance estimation and fail to capture the dynamic structure of covariance matrix.

Regarding dynamic covariance estimation, Wood (2010) study the case of chaotic or near-chaotic ecological system where the joint probability of observed data is extremely irregular, which makes

it impossible to compute and optimise. A framework called 'Synthetic Likelihood Framework' is used to estimate covariance dynamically. The mechanism of this framework is leveraging a model with certain parameters to generate approximate data with the observation. Then, the empirical covariance and other statistics are obtained from the generated data to form a joint probability distribution, which is used to optimise the ecological model. This method requires estimating covariance matrix to capture long-term structure, which is not similar to short-term forecasting. Hefley et al. (2017) presents a flexible framework for modelling spatial the temporal covariance on ecological data. Correlation function is used to directly specify a covariance structure and spectral decomposition 2.1.2 to obtain basis vectors which are used for basis functions added to the mean of the model as shown in (1.13), similar to Generalised Additive Model:

$$y(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta} + \sum_{k=1}^{K} \alpha_k \phi_k(\mathbf{s}) + \varepsilon(\mathbf{s}) \tag{1.13}$$

$$\mathrm{Cov}\left(y(\mathbf{s}_i), y(\mathbf{s}_j)\right) = \phi(\mathbf{s}_i)^\top \boldsymbol{\Sigma}_\alpha \, \phi(\mathbf{s}_j) \tag{1.14}$$

where $y(\mathbf{s})$ is the response at spatial location $\mathbf{s}$ and ϕ_k is the k-th basis function. The basis function vector is denoted as $\phi(\mathbf{s}) = [\phi_1(\mathbf{s}), \phi_2(\mathbf{s}), \cdots, \phi_K(\mathbf{s})]$ and Σ_α denotes the covariance matrix of all the basis functions under the assumption $\alpha \sim \mathcal{N}(0, \Sigma_\alpha)$ where $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_K]$. This means that the correlation between two random variables changes according to how similar their spatial or temporal basis-function representations are. However, this structure fail to capture dynamic changes in time-varying correlation and relies on basis functions to embed autocorrelation.

1.3.6 Epidemiology

In analyzing dynamic infectious disease data, Prashad (2025) leverage Bayesian estimation approach for time-varying covariance matrices in multivariate state-space models. A multivariate state-space model could be written as

$$y_t = X_t \beta_t + \epsilon_t, \;\; \epsilon_t \sim \mathcal{N}(0, \Sigma_t) \tag{1.15}$$

where X_t is the observation design matrix and β_t is the hidden (latent) state vector. ϵ_t denotes the noise that follows Gaussian distribution with dynamic covariance matrix Σ_t. Prashad (2025) assume observation noise covariance matrix Σ_t is modelled as time-varying and given a prior inverse Wishart distribution $\Sigma_t \sim IW(n_0, n_0 S_0)$ where n_0 is the degrees of freedom and S_0 denotes the scale matrix. The posterior distribution of $\Sigma_t | y_{1:t} \sim \mathcal{IW}(n_t, n_t S_t)$ where the degrees of freedom and scale matrix is updated over time by (1.16) and (1.17)

$$n_t = n_{t-1} + 1 \tag{1.16}$$

$$S_t = S_{t-1} + (y_t - X_t m_t)(y_t - X_t m_t)^\top \tag{1.17}$$

where m_t is the mean of the posterior distribution of the hidden state β_t and $X_t m_t$is the predicted mean of observation y_t. The update rule is run recursively and dynamic covariance modelling demonstrates better estimation quality on relationship between COVID cases and environmental covariates compared to static methods.

1.3.7 Smoothing

As we mention in section 1.2, covariance function can be used to describe the correlation within a stochastic process including stationary and isotropic cases. It is commonly used in a Gaussian process problem where each stochastic processes $Z(x)$ has zero mean function $\mu(x) = 0$ and the covariance between x and y is captured by the covariance function $\text{Cov}(Z(x), Z(y))$. Apart from Gaussian process, covariance function is also used for solving Stochastic partial differential equation (SPDE). SPDE involves stochastic processes and differential operators, such as the first derivative operator. SPDE states that for some differential operator D, a stochastic process f is a SPDE solution to the equation:

$$Df = \epsilon \tag{1.18}$$

where x is a normal random variable with mean zero and variance ϵ. This equation also means the first derivative of f has zero mean and finite variance at every point and the value of derivative at any two points x and y is uncorrelated. Consider the variance ϵ is controlled by certain parameter τ by:

$$Df = \frac{\epsilon}{\tau} \tag{1.19}$$

Then parameter τ controls the 'smoothness' of function f. Consider the mathematical solution to the equation (1.18) is given by:

$$f(x) = \int w(x-u)\epsilon(u)d(u) \tag{1.20}$$

where w is a Green's function (Cabada, 2014) derived from D, which is a weighting function such that the value of the stochastic process at x is a weighted sum of white noise process. This is similar to Generalised Additive Model where each value of the function is a weighted sum of basis functions. For uncorrelated cases, w is known as:

$$w(d) = \begin{cases} 0 & d \neq 0 \\ \infty & d = 0 \end{cases} \tag{1.21}$$

and if $w(d) = 1$ for all d, then the whole process is a constant. Between these two extremes are weighted functions that reproduce correlations over different ranges and the covariance function for point x and point y is shown as

$$c(x, y) = \int w(x-u)w(y-u)du \tag{1.22}$$

This equation shows that the solution to SPDE has a covariance structure that is induced by the choice of D, which means covariance function estimation methods may be potentially adopted in the field of solving SPDE.

1.3.8 Summary of Applications

In conclusion, Covariance estimation plays a foundational role across a wide range of scientific and engineering disciplines, serving as a critical tool for characterizing dependencies in multivariate

systems. In neuroscience, it enables the modelling of functional connectivity across brain regions, capturing both static and dynamic structures. In finance, covariance modelling underpins portfolio optimisation, with methods evolving from static shrinkage estimators to dynamic GARCH-based approaches to account for time-varying volatility. In renewable energy applications, such as wind power and electricity demand forecasting, dynamic covariance estimation facilitates joint probabilistic forecasting by capturing spatio-temporal dependencies essential for reliable decision-making. Similarly, ecological systems and epidemiological models benefit from dynamic covariance structures to model interactions and temporal evolution under complex uncertainty. Even in mathematical frameworks such as SPDEs, covariance functions define the smoothness and dependency structure of stochastic processes. Despite diverse domains, a common challenge persists: balancing model expressiveness and computational tractability, particularly in high-dimensional or dynamic settings. Continued advancements in structured, scalable, and dynamic covariance estimation are essential to improving forecasting accuracy, model interpretability, and decision support across these application areas.

1.4 Challenges in Dynamic Covariance Estimation

Based on the discussion of covariance modelling and covariance parameterisation, the task of covariance modelling can be generally divided into two sub-tasks: parameterisation of the covariance matrix, and parameter estimation. The main challenges of parameterisation include parameter explosion, whereby the number of parameters to be estimated increases quadratically with the number of variables ($d(d+1)/2$ in particular), and the lack of statistical interpretability. Covariance modelling methods typically incorporate the effects covariates on covariance structure while simultaneously satisfying the strict positive definiteness constraint. Enforcing this constraint during optimisation often introduces inflexibility and may result in poor statistical properties (Pinheiro and Bates, 1996). In addition, because covariance is a second-order statistic and therefore not directly observable, limited sample availability can present a substantial challenge for dynamic covariance estimation. For example, a time-varying covariance matrix may have only a single observed realisation at each time point, making it difficult to fit the model and to evaluate the estimated covariance matrix.

Chapter 2

Background Methodology

This chapter reviews existing methodology in covariance factorization, covariance modelling and evaluation. It begins by establishing a systematic framework for the parameterisation of covariance matrices, a central problem in multivariate statistical modelling. The covariance matrix encodes marginal variability and dependence structure simultaneously, yet its estimation is intrinsically constrained by symmetry and positive definiteness, rendering naive parameterisations infeasible in moderate to high dimensions. To address these structural constraints, this section reviews two principal classes of approaches: parametric covariance functions and matrix decomposition–based representations. The former characterise dependence through structured functional forms—encompassing both stationary and non-stationary, static and dynamic formulations—while the latter enforce positive definiteness through algebraic factorisations such as spectral and Cholesky-type decompositions. By examining these complementary perspectives, this section clarifies the theoretical foundations, modelling trade-offs, and practical implications of covariance parameterisation, thereby providing the methodological basis for subsequent developments in dynamic and high-dimensional covariance modelling.

Existing covariance modelling approaches in both static and dynamic settings are also summarised in this chapter, connecting structural parameterisations to practical estimation frameworks. For static covariance estimation, methods such as regression-based partial correlation modelling, graphical lasso regularisation, Bayesian formulations, and shrinkage estimators are examined. These approaches often operate on the precision matrix to characterise conditional dependence and typically rely on sparsity assumptions or penalisation to ensure tractability in high-dimensional contexts, though they generally presume time-invariant dependence structures. The discussion then extends to dynamic covariance modelling, including Generalized Additive Covariance frameworks, ARCH/GARCH-type volatility models, and Dynamic Conditional Correlation (DCC) models, which allow second-order dependence to evolve over time or with covariates. While additive and regression-based approaches emphasise flexibility and covariate integration, volatility-driven models focus on temporal persistence and computational scalability. Together, these methodologies highlight the trade-offs among flexibility, interpretability, and stability, thereby motivating the development of more general and principled dynamic covariance modelling strategies.

In addition to covariance parameterisation and covariance estimation methods, this chapter also review establishes principles for evaluating covariance models in both static and dynamic contexts, with particular emphasis on their role in multivariate probabilistic forecasting. Unlike pointwise

error metrics, covariance-oriented evaluation criteria assess the quality of the inferred dependence structure, uncertainty quantification, and joint distributional accuracy. Likelihood-based measures, such as the log-likelihood and logarithmic score, provide strictly proper scoring rules and align naturally with maximum likelihood and Bayesian inference. Complementary metrics with the "true" covariance matrix observable, including weighted squared error and Kullback–Leibler divergence, quantify discrepancies in variance–covariance structure from matrix and distributional perspectives, respectively. Since we are interested in applying dynamic covariance modelling on a probabilistic forecasting tasks, multivariate proper scoring rules such as the Energy Score and p-Variogram Score are introduced as well and they further evaluate calibration and dependence representation in predictive distributions. Collectively, these metrics provide a comprehensive toolkit for assessing covariance estimation quality, structural fidelity, and predictive performance, thereby ensuring rigorous comparison of modelling approaches across both static and time-varying settings.

2.1 Parameterisation of Covariance Matrix

This section introduce two principle approaches for parameterising the covariance matrix: covariance functions and matrix decomposition-based representation. Both their static and dynamic formulations are summarised.

2.1.1 Parametric Covariance Functions

Parametric covariance functions are commonly used parameterisation to describe covariance structure with a small number of parameters. The covariance functions defined in equation (1.5) employ a single functional form to represent the overall structure of the covariance matrix in time-invariant settings. Such parameterisations involve only a limited set of parameters to be estimated, thereby effectively mitigating the issue of parameter proliferation as the dimensionality d increases, while retaining clear statistical interpretability. Consequently, parametric covariance functions are computationally attractive for large-scale covariance modelling. Parametric covariance functions can be divided into stationary and non-stationary cases:

Stationary Covariance Functions Define a stochastic process as $\{Z(x); x \in D\}$ where D is the parameter space of the stochastic process. For each fixed $x \in D$, the value $Z(x)$ is a random variable. According to the Definition 1.7, stationary covariance format can be described as

$$C(Z(x_1), Z(x_2)) = f(r; \xi) \tag{2.1}$$

$$r = \|x_1 - x_2\| \tag{2.2}$$

where ξ denotes all the parameters of the covariance function and its value is determined solely by separation r of random variables $Z(x_1)$ and $Z(x_2)$. For dynamic version where covariance is time-varying, stationary covariance function can be described as

$$C\left(Z_t(x_1), Z_t(x_2)\right) = f(r; \xi_t) \tag{2.3}$$

$$r = \|x - y\| \tag{2.4}$$

where $Z_t(x_1)$ and $Z_t(x_2)$ denote the variables and the function relies on time-variate parameters ξ_t to capture the dynamics. In practical applications, such as probabilistic forecasting, covariance structure ofthen varies with respect to certain covariates. Dynamic covariance functions are

therefore employed to incorporate covariate information efficiently during the model fitting process. Commonly used covariance functions along with their corresponding parameters to be estimated, are summarised in Table 2.1.

Class	**Function** $C(r;\xi)$	**Parameters** ξ
Powered Exponential	$\sigma^2 e^{-(\theta r)^p}$	$0 < p \leq 2;\ \theta > 0;\ \sigma^2 \geq 0$
Whittle–Matérn	$\sigma^2 \dfrac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2}\nu\dfrac{r}{\rho})^\nu K_\nu(\sqrt{2}\nu\dfrac{r}{\rho})$	$\nu > 0;\ \rho > 0;\ \sigma^2 \geq 0$
Cauchy	$\sigma^2\big(1 + (\theta r)^2\big)^{-\gamma}$	$\gamma > 0;\ \theta > 0;\ \sigma^2 \geq 0$
Spherical	$C(r) = \begin{cases} \sigma^2\left(1 - \frac{3r}{2\theta} + \frac{r^3}{2\theta^3}\right), & 0 \leq r \leq \theta, \\ 0, & r > \theta \end{cases}$	$\theta > 0;\ \sigma^2 \geq 0$
Canonic Periodic	$\sigma^2 \exp\Big(\frac{2\sin^2(\frac{\theta r}{2})}{\phi}\Big)$	$\sigma^2 \geq 0;\ \theta > 0;\ \phi > 0$

Table 2.1: Common covariance functions and their parameters.

The powered exponential function is taken as an illustrative example. The static stationary powered exponential function is defined as

$$f(r;\xi) = \sigma^2 e^{-(\theta r)^\gamma}, \xi = \{\sigma, \theta, \gamma\}, 0 < \gamma \leq 2 \tag{2.5}$$

where σ denotes the estimated standard deviation, θ is a scale parameter associated with the separation distance, and γ is a shape parameter. The resulting covariance depends solely on the separation r, reflecting the stationarity assumption.

The powered exponential covariance function is commonly employed in stationary covariance modelling due to its parsimonious parameterisation, involving only a small number of parameters (σ, θ, and γ in this case). Each of these parameters may be modelled as a function of covariates, thereby introducing a degree of flexibility. While this formulation provides a fundamental and widely used covariance model, alternative functions, such as the Whittle–Matérn covariance function, can offer greater flexibility.

Stationary covariance functions guarantee that the resulting covariance matrix is positive definite. Their dynamic extensions can partially accommodate non-stationary behaviour by allowing parameters such as θ_t to vary, for example, to capture spatial rather than purely temporal characteristics. However, such approaches remain limited in their ability to represent fully dynamic, non-stationary covariance structures. In spatio-temporal applications, the covariance between two time-varying random variables often depends on more than their separation alone, a dependence that stationary and weakly dynamic covariance functions may fail to capture adequately.

Non-stationary Covariance Functions Non-stationary covariance functions enable the modelling of non-stationarity in dynamic covariance structures. In a dynamic setting, a non-stationary covariance function can be expressed as

$$C(Z_t(x_1), Z_t(x_2)) = f(\theta(x_1, x_2, \mathbf{x}_{x_1 x_2, t}); \xi_t) \tag{2.6}$$

where $\theta(\cdot)$ denotes a predictor defined on a pair of random variables, capturing the non-stationary characteristics of the underlying stochastic process and $\mathbf{x}_{x_1 x_2, t}$ denote the covariates selected for

modelling covariance of $Z(x_1)$ and $Z(x_2)$. Other time-constant parameters are defined as ξ. Under this formulation, the estimated covariance no longer depends solely on the separation between $Z_t(x_1)$ and $Z_t(x_2)$, but also on the specific pairing of the random variables. An example of non-stationary powered exponential function case is given by

$$f(x, y, \mathbf{x}_{x_1x_2,t}; \xi_t) = \sigma_t^2 e^{-(\theta(x_1,x_2,\mathbf{x}_{x_1x_2,t}))^{\gamma_t}},\ \xi_t = \{\sigma_t, \theta(\cdot), \gamma_t\},\ 0 < \gamma_t \leq 2 \tag{2.7}$$

where the original scale parameter θ is replaced by a predictor $\theta(\cdot)$ in order to capture non-stationary behaviour. More generally, predictors may be introduced for any subset of the time-varying parameters in ξ_t, and different subsets of covariates can be incorporated depending on the specific application.

Despite their flexibility, non-stationary covariance functions do not, in general, guarantee a positive-definite covariance matrix. The use of predictors for individual off-diagonal entries may introduce numerical instability, potentially resulting in ill-conditioned covariance matrices. This limitation raises the necessity of a principled positive-definite estimation framework tailored to non-stationary covariance functions. Chapter 3 introduces a novel estimation framework for fitting non-stationary covariance functions that ensures positive definiteness while improving goodness of fit.

2.1.2 Matrix Decomposition

Matrix decomposition provides a widely used approach for parameterising covariance matrices. A key challenge in covariance modelling is the well-known positive-definiteness constraint (Pourahmadi, 2011) and one common strategy for addressing this constraint is to obtain unconstrained parameters through appropriate matrix decompositions. The underlying rationale of many such parameterisations can be expressed as

$$\Sigma = L^\top L \tag{2.8}$$

where $L \in \mathbb{R}^{d\times d}$ is a full-rank matrix. Different structural choices or constraints imposed on L give rise to different parameterisations of the covariance matrix Σ (Pinheiro and Bates, 1996). It is common to decompose a covariance matrix into components representing marginal variances and dependence (or correlation) structure. Regression-based models are often applied to the logarithm of the variance components, thereby removing the positivity constraint on variances. To dynamically modelling the covariance matrix Σ_t using this decomposition given by a subset of covariate $\mathbf{x}_{ij,t}$, the modelling formulation is given by:

$$l_{i,j} = f(i, j, \mathbf{x}_{ij,t}), \tag{2.9}$$

where $l_{i,j}$ denotes the i-th row and j-th column entry of factor L_t and $f(\cdot)$ is the predictor. The covariate $\mathbf{x}_{ij,t}$ denotes the specific subset of time-varying covariates selected for the i-th row and j-th column element. Such decompositions facilitate flexible modelling while ensuring the positive definiteness of the resulting covariance matrix by construction. The following sections review and discuss several matrix decomposition techniques for covariance matrix parameterisation.

Variance-correlation Decomposition

The variance-correlation decomposition of the covariance matrix expresses the matrix Σ as $\Sigma = D_\sigma R D_\sigma$ where D_σ is the diagonal matrix of standard deviations and $R = (\rho_{i,j})$ is the correlation

matrix. By modelling the logarithm of the diagonal elements of R, the positivity constraint is removed. In contrast, the correlation matrix R remains subject to non-trivial constraints: all diagonal elements must equal one, and all off-diagonal elements must lie in the interval $[-1, 1]$, with the additional requirement that P be positive definite. These constraints complicate statistical modelling, particularly for methods such as Generalised Linear Models (GLMs), and the variance–correlation decomposition does not, in itself, reduce the number of parameters to be estimated.

Spectral Decomposition

Spectral decomposition, also known as eigenvalue decomposition is widely used for principle component analysis (PCA). For a covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, the decomposition is given by

$$\Sigma = P\Lambda P^{-1} = \sum_{i=1}^{p} \lambda_i e_i e_i^\top \tag{2.10}$$

where Λ is a diagonal matrix containing the eigenvalues of Σ, and P is an orthogonal matrix whose columns are the corresponding normalised eigenvectors. Here, λ_i denotes the i-th eigenvalue and e_i the associated eigenvector. Due to the positive definiteness constraint on Σ, all the eigenvalues must be strictly positive. Modified forms of the spectral decomposition have been proposed to yield unconstrained parameterisations of symmetric matrices; one such formulation is presented in equation (2.11)

$$\log \Sigma = P \log(\Lambda) P^{-1} = \sum_{i=1}^{p} e_i e_i^\top \log \lambda_i. \tag{2.11}$$

Note that Σ and $\log \Sigma$ share the same eigenvectors. The dynamic modelling for the spectral decomposition formulation can be expressed as:

$$\log \Sigma_t = P_t \log(\Lambda_t) P_t^{-1} = \sum_{i=1}^{p} e_{t,i} e_{t,i}^\top \log \lambda_{t,i}. \tag{2.12}$$

$$e_{t,i}^d = f_d(\mathbf{x}_t^i), \lambda_{t,i} = f(\mathbf{x}_t^i) \tag{2.13}$$

where $e_{t,i}$ is the i-th time-variate eigenvalue and $e_{t,i}^d$ is the d-th element of the eigenvector $e_{t,i}$. The $\mathbf{x}_t^i$ denotes the time-varying covariate of the estimator. While this transformation yields an unconstrained symmetric matrix, the individual entries of $\log \Sigma$ no longer admit a straightforward statistical interpretation. This limitation highlights a fundamental trade-off between achieving unconstrained parameterisations and preserving statistical interpretability in covariance modelling.

Givens parameterisation

Following parameterisation like Spectral Decomposition, Givens parameterisation can be applied on the orthogonal matrix to improve numerical stability and obtain well-conditioned estimates. The Givens parameterisation represents an orthogonal matrix as the product of Givens rotations $G(i, j, \theta)$ each of which is a sparse orthogonal matrix that performs a planar rotation by an angle

θ in the (i, j)-plane, affecting only the ith and jth coordinates, as shown in (2.14).

$$G(i,j,\theta) = \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & \cos\theta & -\sin\theta & & \\ & & \sin\theta & \cos\theta & & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} \tag{2.14}$$

Since spectral decomposition requires modelling an orthogonal eigenvector matrix of dimension $d \times d$, applying Givens parameterisation provides a structured way to represent this matrix as

$$\Sigma = G_1 G_2 \cdots G_k \tag{2.15}$$

where the rotation angles θ_k constitute the model parameters. An orthogonal matrix in $\mathbb{R}^{d\times d}$ can be fully parameterised using $d(d-1)/2$ Givens rotation angles, which is substantially fewer than the d^2 unconstrained parameters of a general matrix, while respecting orthogonality by construction. In addition, Givens parameterisation can enhance numerical stability when sampling or estimating covariance matrices in probabilistic models, including Gaussian processes and Bayesian hierarchical models. However, compared with direct Cholesky or spectral decompositions, Givens parameterisation entails more complex implementation and additional bookkeeping to manage the sequence of rotations. Moreover, it does not yield a direct statistical interpretation in terms of marginal variances and correlations, in contrast to more conventional covariance parameterisations.

Cholesky Decomposition

Cholesky Decomposition is easily derived by setting L from(2.8) to be an upper triangular matrix. This parameterisation is computationally simple and stable, but the Cholesky factor is not unique for a single Σ. Thus, the more general Cholesky Decomposition of a positive-definite matrix is used in the form of Log-Cholesky by imposing a logarithm on the diagonal elements of the Cholesky factor, such that

$$\Sigma = CC' \tag{2.16}$$

where C is a unique lower-triangular matrix with positive diagonal entries. This parameterisation is computationally efficient and numerically stable, and the uniqueness is guaranteed by enforcing positivity on the diagonal elements. Yet, the resulting Cholesky factor has a poor statistical interpretation Pinheiro and Bates (1996) due to the logarithm implementation, but its transformation $L_m = CD^{-1}$ could lead to an unconstrained and statistically interpretable factor as

$$\Sigma = CD^{-1}DDD^{-1}C' = L_m D^2 L_m^\top \tag{2.17}$$

$$T\Sigma T' = D^2 \tag{2.18}$$

where D is the diagonal of C and $L_m = CD^{-1}$. T is the inverse of L_m. This representation is known as the modified Cholesky decomposition (MCD). In this formulation, both T and D are unconstrained and admit meaningful statistical interpretations: the entries of T correspond to regression coefficients obtained from sequential linear regressions, while D^2 contains the innovation variances.

Owing to these properties, several covariance estimation methods impose sparsity-inducing penalties on the MCD factors, thereby encouraging sparsity in the implied precision matrix. While this makes MCD a powerful and flexible parameterisation, it still suffers from parameter proliferation in high dimensions and depends critically on an a priori ordering of the random variables.

While MCD provides a promising way for covariance parameterisation, modelling the precision matrix directly provides an alternative means of reducing the effective number of parameters. The entries of the precision matrix encode conditional dependencies between variables and are often substantially sparser than those of the covariance matrix. The Cholesky decomposition and its modified variants can also be applied to the precision matrix, yielding

$$\Sigma^{-1} = \hat{T}^\top D^{-2} \hat{T} \tag{2.19}$$

where the entries of $\hat{T}$ and D are unconstrained, which facilitates the modelling. The dynamic formulation of MCD is described as:

$$\eta_{i>j,t} = f_{i,j,t}(i, j, \mathbf{x}_{i,j}) \tag{2.20}$$
$$\eta_{i,i,t} = f_{i,i,t}(i, j, \mathbf{x}_{i,j}) \tag{2.21}$$

where $\eta_{i>j,t}$ represent the off-diagonal element of $\hat{T}$ and $\eta_{i,i,t}$ is the diagonal element of D. $f_{i,j,t}$ denotes the predictor of each $\eta_{i>j,t}$ and $f_{i,i,t}$ represents the predictor of each $\eta_{i,i,t}$. A key limitation of this class of parameterisations is that it is not permutation invariant: the resulting decomposition, and hence the inferred dependency structure, depends on the ordering of the variables. Consequently, Cholesky-based approaches are most appropriate for data with a natural ordering, such as time-indexed or spatially ordered observations.

2.2 Existing Covariance Modelling Approaches

In this section, we introduce covariance estimation, encompassing both static and dynamic covariance modelling. Dynamic covariance modelling constitutes an important class of covariance estimation problems, particularly in real-world applications where dependence structures evolve over time. We present the formal formulations of both static and dynamic covariance estimation. Existing covariance estimation methods applicable to these settings are reviewed, and their principal limitations are summarised. For dynamic covariance predictors, we introduce both time-varying spatial covariance modelling and temporal covariance methods but we will only study spatial covariance modelling in the following chapters.

2.2.1 Static Covariance Estimation

Estimation of covariance matrices is a fundamental component of multivariate probabilistic modeling, underpinning tasks such as probabilistic forecasting and other statistical inference procedures by characterizing joint dependencies among random variables. In the context of Gaussian process models, estimating covariance parameters (e.g., the hyperparameters of the covariance function) is central to defining the dependence structure of the process and to making accurate probabilistic predictions, since the covariance function governs the prior and posterior uncertainty in the model. Static covariance estimation requires a covariance matrix $\Sigma = \mathrm{Cov}(X_1, X_2)$ to characterise the dependency of two random variables X_1 and X_2. The corresponding covariance function can

be expressed as in (1.5). Since the true underlying distribution is unobservable in practice, the covariance matrix is typically estimated using the sample covariance matrix $\tilde{\Sigma}$, defined as

$$\tilde{\Sigma} = \frac{1}{N-1} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}}) (\mathbf{x}_i - \bar{\mathbf{x}})^\top \tag{2.22}$$

where $\mathbf{x} \in \mathbb{R}^d$ denotes the d-dimensional observed sample vectors and $\bar{\mathbf{x}}$ represents the sample mean. However, when the number of observations N is substantially smaller than the dimensionality d, the sample covariance matrix becomes numerically unstable and provides a poor estimate of the true covariance structure. Commonly used covariance matrix estimators include the sample covariance estimator, maximum likelihood estimators, and shrinkage-based methods and they are summarised as follow:

Sample (Empirical) Covariance: Sample covariance matrix provides an unbiased estimate of the population covariance matrix and it is calculated through the observed data and it is expressed as (2.22). The resulting estimator is positive semi-definite but may become singular in high-dimensional settings (Franklin, 2005), particularly when the dimensionality d exceeds the number of observations N.

Maximum Likelihood Estimators: The maximum likelihood estimator (MLE) of the covariance matrix is derived under the assumption that the data follow a multivariate Gaussian distribution. Similar to the sample covariance matrix, the MLE of the covariance matrix is given by

$$\Sigma_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top. \tag{2.23}$$

This estimator is biased, as it systematically underestimates the true covariance matrix due to the use of the factor $1/N$ rather than $1/(N-1)$. The bias becomes more pronounced in small-sample settings, where the number of observations is limited relative to the dimensionality.

Shrinkage Estimators: Shrinkage estimators combine the sample covariance matrix with a structured target matrix in order to improve estimation accuracy and mitigate the singularity and instability issues commonly encountered in high-dimensional settings (Schäfer and Strimmer, 2005). A widely used shrinkage approach is the Ledoit–Wolf estimator (Ledoit and Wolf, 2004), defined as

$$\Sigma_{\text{LW}} = (1-\delta)\tilde{\Sigma} + \delta T, \tag{2.24}$$

where $\delta \in [0, 1]$ denotes the intensity of shrinkage and T is a predefined target matrix, typically chosen as a scaled identity matrix $T = \mu I$ with the scaling factor equal to the average variance of the sample. This estimator yields better-conditioned covariance estimates than the sample covariance matrix and, provided that the target matrix is positive definite and $\delta > 0$, guarantees positive definiteness of the resulting estimator. The shrinkage intensity δ can be estimated directly from the data in a data-driven and asymptotically optimal manner. While this adaptivity improves robustness, the performance of the estimator remains sensitive to the choice of the target matrix.

Graphical Lasso: Graphical Lasso (Glasso) is a regularized maximum-likelihood method for estimating a sparse precision matrix (i.e., the inverse covariance matrix) of a multivariate Gaussian distribution. It extends classical covariance estimation to high-dimensional settings where the number of variables d may be large relative to the number of observations n, and the sample covariance matrix may be singular or unstable. Rather than directly inverting the sample covariance matrix S, Glasso solves a convex optimization problem that maximizes a penalized log-likelihood with an ℓ_1 penalty on the precision matrix entries:

$$\hat{\Theta} = \underset{\Theta \succ 0}{\operatorname{argmax}} \left[\log\det(\Theta) - \operatorname{tr}(\tilde{\Sigma}\Theta) - \gamma \left\|\Theta\right\|_1 \right] \tag{2.25}$$

where $\Theta = \Sigma^{-1}$ is the true precision matrix and $\hat{\Theta} = \hat{\Sigma}^{-1}$ is the estimated precision matrix. The $\log\det(\Theta)$ and $\operatorname{tr}(S\Theta)$ come from the Gaussian log-likelihood. The $\left\|\Theta\right\|_1 = \Sigma_{i,j} \left\|\Theta_{i,j}\right\|$ is the ℓ_1 penalty encouraging entries of Θ to approach zero and γ is a regularisation parameter controlling sparsity. Setting off-diagonal entries of $\hat{\Theta}$ to zero corresponds to assuming conditional independence between variable pairs, aligning with the structure of a Gaussian graphical model. But its performance strongly depends on the choice of γ and it is highly sensitive to outliers, making estimation less robust. The coordinate-descent algorithm used in many Glasso implementations may produce asymmetric precision estimates due to inexact inversion procedures, especially when regularization is weak. This can lead to estimates that are not proper inverses of the covariance estimate and can even have negative or complex eigenvalues, undermining positive definiteness and validity in multivariate analyses.

The Glasso is a well-established penalized maximum likelihood estimator that induces sparsity in the precision matrix via an ℓ_1 penalty, and alternative penalized likelihood approaches — such as estimators based on the minimax concave penalty (MCP) or ridge-type ℓ_2 penalties — have also been proposed for static covariance or precision estimation. These methods principally differ in the form of the penalty function and in their consequent effects on the estimated covariance or precision matrix during optimization.

Bayesian Estimators: Bayesian estimators incorporate prior beliefs about the covariance structure within a Bayesian framework. Specifically, it treats the covariance matrix as a random variable and and update it using Bayes' theorem

$$p(\Sigma|X) = \frac{p(X|\Sigma)p(\Sigma)}{p(X)} \tag{2.26}$$

where X is a random variable of the observed data and $p(\Sigma)$ is the assumed prior. For covariance matrix estimation, a commonly adopted prior is the inverse-Wishart distribution (Kruschke, 2010)

$$\Sigma \sim \mathcal{W}^{-1}(\Psi, \nu) \tag{2.27}$$

where Ψ denotes the scale matrix and ν represents the degrees of freedom. In Bayesian analysis, the inverse-Wishart distribution serves as a conjugate prior for the covariance matrix Σ of a multivariate normal model, implying that the posterior distribution of Σ given the observed data also follows an inverse-Wishart distribution. This conjugacy offers substantial computational convenience. However, the inverse-Wishart prior imposes restrictive assumptions on the covariance structure—such as coupling marginal variances and correlations—which may introduce bias or limit modelling flexibility, particularly in high-dimensional or weakly informative settings.

2.2.2 Dynamic Covariance Predictors

Apart from the methods discussed above, a distinct class of approaches focuses on dynamically modelling the covariance matrix. In dynamic settings, the covariance structure evolves concerning a subset of covariates and, most commonly in practice, varies over time. Therefore, the general way is to apply any estimators with time-varying covariates to the existing covariance parameterisation, either covariance functions or matrix decomposition factors. This section summarises and discusses existing predictors used in dynamic covariance modelling in detail.

Generalized Additive Covariance

Generalized Additive Covariance(GAC) framework proposed by Browell et al. (2022) employs Generalised Additive Model (GAM) (Hastie and Tibshirani, 1986) to model the parameters of a covariance function. In this framework, a parametric covariance function is used to represent the covariance matrix, while GAMs are introduced to model the dependence of the covariance function parameters on covariates. The GAC framework is applicable to both isotropic (1.8) and non-stationary (1.10) settings, in which the covariance depends separation as well as other covariates. Such formulations are commonly used in spatial or temporal covariance modelling. The introduction of GAM and its smoothing penalty are as follow.

The GAM formulation used to model the covariance function parameters is given by

$$g(\xi_j) = A_j\beta_j + \sum_{i=1}^{k} f_{j,i}(x_t^{S_{j,i}}), j = 1, ..., m \tag{2.28}$$

where $g(\cdot)$ is a link function, ξ_j denotes the jth parameter of the covariance function, $A_j\beta_j$ represents the parametric component, and $f_{j,i}(\cdot)$ are smooth functions of covariates. Here, K denotes the number of smooth terms and m the number of covariance function parameters being modelled. Each smooth function $f_{j,i}$ is expressed as a linear combination of basis functions,

$$f_{j,i}(x_t^{S_{j,i}}) = \sum_{k=1}^{K} b_k^{j,i}(x_t^{S_{j,i}})\beta_k^{j,i} \tag{2.29}$$

This modelling strategy provides substantial flexibility for incorporating spatial or temporal covariates in spatio-temporal forecasting applications. In particular, the covariate vector x_t in (2.28) may depend on time, spatial distance, or other relevant predictors. The notation $x_t^{S_{j,i}}$ denotes a subset of covariates associated with the ith smooth term for the jth covariance parameter, allowing different subsets of covariates to influence different components of the covariance structure. Smooth covariate effects are widely used in probabilistic forecasting, and by leveraging these properties, GAMs offer a principled and flexible approach to dynamic covariance modelling. For GAM model fitting, as the dynamically estimated covariance matrix fits in the Multivariate Gaussian framework, the optimisation object is set to maximize the regularized Bayesian posterior log-density of the form

$$\ell(\beta) = \log p(\beta|y, \gamma) = \sum_{i=1}^{N} \log p(y^{(i)}|\beta) - \frac{1}{2}\beta^\top S^\gamma \beta \tag{2.30}$$

where the quadratic term $\beta^\mathsf{T} S^\gamma \beta$ acts as a smoothing penalty that controls the complexity of the smooth functions through the regularisation parameter γ. More details of smoothing penalty

are discussed in the next paragraph. Despite its flexibility, the GAC framework remains limited to stationary or isotropic covariance structures and does not naturally accommodate fully non-stationary dependence. Moreover, jointly modelling spatial and temporal covariance structures places strong constraints on the choice and design of the covariance function, presenting a significant challenge in more complex dynamic settings.

Apart from GLASSO regularisation (Friedman et al., 2008), smoothing regularisation is widely used for GAMs, to control the complexity of smooth functions and prevent overfitting. They achieve this by penalizing the "wiggliness" of the estimated functions, ensuring a balance between model flexibility and generalisation. The estimation of a smooth function $f(\cdot)$ involves a regularisation term shown as:

$$\mathcal{L}_\Phi = \mathcal{L}(f(\theta), \mathbf{y}) + \gamma \int \left[f^{''}(\theta)\right]^2 dx \tag{2.31}$$

where $\mathcal{L}(\cdot)$ denotes the loss function of $f(\theta)$ and the regularisation term ensure smooth by penalizing the second derivatives $f^{''}(\theta)$. A larger γ imposes a heavier penalty on curvature, leading to smoother (potentially linear) functions, while a smaller γ allows for more flexibility, potentially capturing more complex patterns in the data. In GAM, this smoothing regularisation can be written as:

$$\mathcal{L}_\Phi = \|\mathbf{y} - X\beta\|^2 + \gamma\beta^\top S\beta \tag{2.32}$$

where X is the design matrix with coefficient β and S is the penalty matrix derived from the basis functions and their derivatives. This smoothing technique is also applied in the simulation in Section 3.2.1.

Generalized Autoregressive Conditional Heteroskedasticity

Autoregressive conditional heteroskedasticity(ARCH) is a widely used statistical model for time series data temporal covariance modelling that describes the variance of the error as a function of previous error terms. They are particularly common in financial econometrics, where time series often exhibit volatility clustering and heteroskedastic behaviour. In an ARCH(q) model, the innovation process is expressed as

$$\epsilon_t = \sigma_t z_t, \qquad z_t \sim \text{i.i.d. } (0, 1), \tag{2.33}$$

where z_t is typically assumed to follow a standard normal or other zero-mean, unit-variance distribution. The mean and variance equations are given by

$$y_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i y_{t-i} + \epsilon_t \tag{2.34}$$

$$\sigma_t^2 = \beta_0 + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}^2 \tag{2.35}$$

where y_t denotes the time-varying response, q controls the lag order, and α_i and β_j govern the dynamics of the conditional mean and conditional variance, respectively.

In multivariate or high-dimensional settings, covariance modelling can be applied to the innovation vector $\boldsymbol{\epsilon}_t$, which is typically assumed to follow a conditional distribution with time-varying

covariance matrix Σ_t. In this context, ARCH-type models capture temporal non-stationarity in second-order moments through the evolution of Σ_t. When the variance parameters β_j are constant, the model implies a stationary conditional variance process, whereas allowing these parameters to vary over time introduces non-stationary volatility dynamics. For covariance stationarity to hold, the ARCH model requires

$$\sum_{j=1}^{q} \beta_j < 1 \tag{2.36}$$

which ensures that the unconditional variance is finite. ARCH models are effective at capturing short-term volatility persistence; however, they often require large lag orders to represent long-memory behaviour. This limitation motivates generalised extensions that incorporate lagged conditional variances directly into the model.

The Generalised Autoregressive Conditional Heteroskedasticity (GARCH) model extends the ARCH framework and is widely used to capture both short- and long-term volatility persistence in univariate and multivariate time series. By incorporating lagged conditional variances into the volatility equation (2.37)

$$\sigma_t^2 = \beta_0 + \sum_{j=1}^{q} \beta_j \epsilon_{t-j}^2 + \sum_{k=1}^{p} \gamma_k \sigma_{t-k}^2, \tag{2.37}$$

GARCH is allowed to capture the persistence of past volatility, enabling it to capture the slow decay of volatility shocks over time. The term $\sum_{k=1}^{p} \gamma_k \sigma_{t-k}^2$ captures persistence in conditional variance beyond what is achievable with ARCH models alone. Compared with ARCH, GARCH provides a more flexible and realistic representation of volatility dynamics and is therefore extensively used in financial applications, such as forecasting asset return volatility and estimating time-varying risk for portfolio optimisation. While multivariate GARCH models exist, extending GARCH-type frameworks to efficiently and reliably model high-dimensional time-varying covariance matrices remains computationally challenging and is an active area of research.

Dynamic Conditional Correlation (DCC) Models

Dynamic Conditional Correlation (DCC) models address the scalability issue by decomposing the conditional covariance matrix as

$$\Sigma_t = D_t R_t D_t \tag{2.38}$$

where D_t is a diagonal matrix of conditional standard deviations, typically modelled using univariate GARCH processes, and R_t is a time-varying correlation matrix. The dynamics of R_t are governed by a low-dimensional autoregressive structure, allowing correlations to evolve over time without incurring excessive parameter complexity. DCC can be interpreted as a two-stage extension of GARCH: 1. Conditional variances are modelled using standard univariate GARCH equations. 2. Conditional correlations are updated through a separate dynamic process. This separation enables DCC models to capture time-varying dependence structures while maintaining computational feasibility, making them popular in financial econometrics. However, DCC models rely on strong assumptions, such as homogeneous correlation dynamics across all asset pairs, and remain limited in their ability to incorporate exogenous covariates or complex non-stationary behaviour.

2.3 Evaluation

This section introduces evaluation metrics used for covariance modelling in both static and dynamic settings. As probabilistic forecasting is a key practical application of dynamic covariance modelling and is closely related to real-world problems (see Chapter 4 for details), commonly used evaluation metrics from the probabilistic forecasting literature are also reviewed.

Before introducing the evaluation metrics used in covariance modelling, we first restate the notation adopted throughout this section. Let $X \in \mathbb{R}^d$ denote the random variable of interest with the observation x^i, $i = 1, 2, \cdots, N$. Σ denotes a true covariance matrix that does not depend on time, and Σ_t indicates a true covariance matrix that is time-varying. We use $\hat{\Sigma}$ to denote the estimated covariance matrix and similarly, $\hat{\Sigma}_t$ for the time-variate estimated covariance matrix. The empirical covariance matrix is written as $\tilde{\Sigma}$ and its time-variate version as $\tilde{\Sigma}_t$. The covariance function is denoted as $C(\cdot)$ and it can capture time-varying structure where the incorporated covariate $\mathbf{x}_t$ is time-varying.

In this section, we introduce objective functions commonly used in covariance modelling the probabilistic forecasting. We assume the true covariance matrix Σ is accessible in this section to simplify the introduction, but in practice, the true covariance matrix is typically not observable and objective functions use empirical covariance $\tilde{\Sigma}$ to approximate the role of true covariance matrix.

2.3.1 Evaluation Metrics for Covariance Estimation

The assessment of covariance models demands evaluation criteria that go beyond pointwise prediction accuracy to examine the quality of the inferred dependence structure and uncertainty quantification. In contrast to simple error-based measures, covariance-oriented metrics evaluate how effectively a model represents joint variability, cross-correlations, and conditional relationships among variables. Well-designed loss functions discourage mischaracterisation of uncertainty by penalising both variance and covariance estimation errors. This section introduces several commonly adopted evaluation metrics for covariance modelling, including the Kullback–Leibler divergence, negative log-likelihood, and weighted mean squared error, highlighting their respective roles in assessing model performance in static and dynamic contexts.

Norm-based Matrix Discrepancy Measures

Distance-based metrics are commonly used in various estimations and for covariance estimation, three commonly used loss are Stein's entropy loss (L_1), quadratic loss (L_2), and the normalized Frobenius loss (L_3). The L_1 loss function advocated by James and Stein (1992) is shown as (2.39)

$$\mathcal{L}_1(\hat{\Sigma}, \Sigma) = \operatorname{tr}(\hat{\Sigma}\Sigma^{-1}) - \log\left|\hat{\Sigma}\Sigma^{-1}\right| - d \tag{2.39}$$

where d is the dimension of covariance matrix. The time-variate version for dynamic covariance modelling could be:

$$\mathcal{L}_1(\hat{\Sigma}_t, \Sigma_t) = \frac{1}{T}\sum_{i=1}^{T}\left(\operatorname{tr}(\hat{\Sigma}_i\Sigma_i^{-1}) - \log\left|\hat{\Sigma}_i\Sigma_i^{-1}\right| - d\right) \tag{2.40}$$

where T is the number of timestamps. This loss is also known as Stein's loss or entropy loss. It is invariant under nonsingular linear transformations, meaning that its value is unchanged under a

change of coordinate system. This property makes it particularly suitable for covariance matrices, whose interpretation depends on the scaling and rotation of variables. Furthermore, for multivariate Gaussian distributions, this loss has a direct connection with the Kullback–Leibler (KL) divergence between two Gaussian distributions. Specifically, when the mean vectors are identical, minimizing the KL divergence between $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu, \hat{\Sigma})$ is equivalent to minimizing the Stein loss.

Unlike L_1 loss, quadratic loss directly measures the element-wise distance between two covariance matrices without assuming a probability distribution. The Frobenius norm based quadratic loss is defined as

$$\mathcal{L}_2(\hat{\Sigma}, \Sigma) = \sum_{i=1}^{d} \sum_{j=1}^{d} (\hat{\Sigma}_{i,j} - \Sigma_{i,j})^2. \tag{2.41}$$

This loss treats all covariance elements equally and penalizes overestimation and underestimation symmetrically. Compared with Stein's loss, the quadratic loss does not require matrix inversion, making it computationally more efficient and applicable even when the covariance matrix is close to singular. However, because all entries are weighted equally, L_2 loss is sensitive to the scale of variables. Large variance components may dominate the loss, causing estimation errors in small-variance directions to receive insufficient attention. Therefore, it is not invariant under linear transformations and is less suitable when variables have substantially different scales.

To address the dimensional dependence of the quadratic loss, Ledoit and Wolf (2004) propose a normalised variant of the squared Frobenius norm:

$$\mathcal{L}_3(\hat{\Sigma}, \Sigma) = \frac{1}{d} \left\| \hat{\Sigma} - \Sigma \right\|^2 = \frac{1}{d} \mathrm{tr}(\hat{\Sigma} - \Sigma)^2 \tag{2.42}$$

The normalization by d ensures that the loss remains comparable across different covariance dimensions. In particular, the Frobenius norm of the identity matrix becomes independent of the dimension, which facilitates comparison between covariance estimators with different matrix sizes Ledoit and Wolf (2004). Nevertheless, L_3 loss inherits the limitations of the Frobenius norm: it is not invariant to rescaling and gives equal importance to all covariance elements. Consequently, it may not accurately reflect the statistical importance of errors in different directions.

Log-likelihood

Under the assumption of a Gaussian distribution, the (log-)likelihood is a widely used metric for evaluating the accuracy of covariance estimation. It measures how probable the observed data are under a specified probabilistic model and therefore provides a principled criterion for assessing both mean and covariance estimates. Given observed random vector $\mathbf{x}_t \in \mathbb{R}^d$ at time t, the log-likelihood for a single observation is given by

$$\mathcal{L}(\mu_t.\Sigma_t|\mathbf{x}_t) = \log \prod_{i=1}^{d} f_{\mathcal{N}}(x_t^i|\mu_t, \Sigma_t) = -\frac{Nd}{2} \log(2\pi) - \frac{N}{2} \log|\Sigma_t| - \frac{1}{2} \sum_{i=1}^{N} (x_t^i - \mu_t)^\top \Sigma_t^{-1} (x_t^i - \mu_t) \tag{2.43}$$

where $f_{\mathcal{N}}$ is the probability density function of the assumed distribution according to the assumption and μ_t denotes the time-varying mean. The log-likelihood has a strong theoretical foundation and constitutes a strictly proper scoring rule under the assumed distribution, ensuring that the true

data-generating covariance is optimal in expectation. Its close alignment with maximum likelihood and Bayesian inference makes it a natural and coherent choice for both model estimation and evaluation, particularly in dynamic settings involving time-varying covariance structures. However, its effectiveness depends critically on the validity of the Gaussian assumption; departures such as heavy tails or skewness may lead to misleading assessments, and the metric is sensitive to outliers. In high-dimensional contexts, numerical instability may also arise from matrix inversion and determinant computation when covariance estimates are ill-conditioned. Nevertheless, when Gaussian assumptions are reasonable, the log-likelihood remains a central and interpretable metric for covariance modelling within likelihood-based frameworks.

Weighted Square Error

Least Squares Error (LSE) is widely used when an explicit objective function is available. In covariance modelling, a weighted least squares (WLS) criterion is often adopted to quantify discrepancies between the estimated covariance matrix and a known or reference covariance matrix by comparing their individual entries. This approach allows differential emphasis to be placed on specific components of the covariance structure, which is particularly useful when correlations of varying magnitudes exhibit different levels of importance or reliability. Let $\Sigma = (C(R_{i,j})) \in \mathbb{R}^{d\times d}$ denote the true covariance matrix and $\hat{\Sigma}$ is the estimated covariance matrix, the weighted square error is defined as

$$\mathcal{L}_{WLS} = \frac{1}{d(d-1)} \sum_{i\neq j} \left(\frac{\Sigma - \hat{C}(R_{i,j};\xi)}{1 - \hat{C}_{corr}(R_{i,j};\xi)} \right)^2 \tag{2.44}$$

where $C(R_{i,j})$ is the d-dimensional empirical covariance with $R_{i,j}$ representing the separation matrix where entry at i-th row and j-th column denotes the distance between variable X_i and variable X_j. The estimated covariance is denoted as $\hat{C}(R_{i,j};\xi)$ and $\hat{C}_{corr}(R_{i,j};\xi)$ denotes the estimated correlation. This loss function leverages a weight dependent on correlation such that higher loss is gained for covariance distance between variables with high correlation. In the case of dynamic modelling, this loss function could be written as

$$\mathcal{L}_{WLS} = \frac{1}{d(d-1)} \frac{1}{T} \sum_{t=1}^{T} \sum_{i\neq j} \left(\frac{\Sigma_t - \hat{C}(R_{i,j};\xi_t)}{1 - \hat{C}_{corr}(R_{i,j};\xi_t)} \right)^2 \tag{2.45}$$

where ξ_t denotes the temporal covariates and T is the number of timestamp. In practices, the true covariance matrix Σ_t is not observable and instead, empirical covariance matrix is used and the weighted least square loss function become

$$\mathcal{L}_{WLS} = \frac{1}{d(d-1)} \frac{1}{T} \sum_{t=1}^{T} \sum_{i\neq j} \left(\frac{[\tilde{\Sigma}_t - \hat{C}(R_{i,j};\xi_t)]_{i,j}}{1 - [\hat{C}_{corr}(R_{i,j};\xi_t)]_{i,j}} \right)^2 . \tag{2.46}$$

This loss function can be explained as the weighted distance between empirical covariance, which is the $\tilde{\Sigma}_t$, and the estimated covariance, the $\hat{C}(R_{i,j};\xi_t)$. Empirical covariance can be replaced by sample covariance but may lead to instability, especially in sample insufficient scenarios. Notice that this loss function does not consider the variance, and incorporating losses on variance introduces

the full weighted loss:

$$\mathcal{L}_{WLS} = \frac{1}{d^2 T} \sum_{t=1}^{T} \sum_{i,j} \left(\frac{\tilde{\Sigma}_t - \hat{C}(R_{i,j}; \xi_t)}{1 - \hat{C}_{corr}(R_{i,j}; \xi_t)} \right)^2 . \tag{2.47}$$

Kullback-Leibler Divergence

The Kullback-Leibler (KL) Divergence is a general metric to measure how well a probability density function approximates the other one. Although it is asymmetric, it still is a proper scoring rule widely used to quantify the distributional approximation error. Browell et al. Browell et al. (2022) employ KL divergence to evaluate how closely an estimated covariance matrix approximates the true time-varying covariance in simulation study. Assuming zero-mean Gaussian distributions with different covariance matrices, the KL divergence is defined as

$$\mathcal{L}_{KL}(\Sigma_t, \hat{\Sigma}_t) = \sum_x P(X_t) \log \frac{P(X_t)}{Q(X_t)}, P \sim N(0, \Sigma_t), Q \sim N(0, \hat{\Sigma}_t) \tag{2.48}$$

where Σ_t denotes the true covariance matrix at time t and $\hat{\Sigma}_t$ is its estimate. For Gaussian distributions, this divergence admits a closed-form expression and directly penalizes errors in both variance and correlation structure. While KL divergence is not symmetric and depends on the assumed parametric form, it provides an informative metric for assessing covariance estimation accuracy and is widely used in machine learning. For probabilistic forecasting where the 'true' covariance matrix is not observable, KL-divergence may only be adapted to sample covariance generated by a single realisation, which makes it noisy and may not be a suitable criterion in practice.

2.3.2 Evaluation Metrics for Probabilistic Forecasting

Evaluating multivariate probabilistic forecasts requires scoring rules that jointly assess marginal accuracy, dependence structure, and distributional calibration. Unlike pointwise error metrics, proper scoring rules reward forecasts that assign high probability mass to the realized outcomes while penalizing misrepresentation of uncertainty and dependency. This section reviews several widely used evaluation metrics for probabilistic forecasting and dynamic covariance modelling, including the Energy Score, p-Variogram Score, and logarithmic score, each emphasizing different aspects of forecast quality.

Energy Score

The Energy Score (ES) is a widely used proper scoring rule for evaluating multivariate probabilistic forecasts, particularly in power and renewable energy forecasting. It assesses both the accuracy of forecasted trajectories and the dispersion of the predictive distribution, thereby capturing how well the ensemble of predicted scenarios aligns with the observed outcome. For a single forecast–observation pair, the Energy Score is defined as

$$\mathcal{L}_{\text{ES}} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{M} \sum_{m=1}^{M} \left\| X_i - \hat{X}_i^{(m)} \right\|_2 - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{n=1}^{M} \left\| \hat{X}_i^{(m)} - \hat{X}_i^{(n)} \right\|_2 \right]. \tag{2.49}$$

where $\|\cdot\|_2$ denotes the norm ℓ_2 and X_i is the observed trajectory. The $\hat{X}_i^{(m)}$ is the m-th forecasted trajectory sampled from the predictive multivariate distribution and the total amount of sampled trajectories is M. The first term measures the closeness of the forecast ensemble to the observation, while the second term penalizes excessive dispersion by evaluating the pairwise distances among forecasted trajectories. Together, these components balance sharpness and calibration, encouraging accurate yet appropriately uncertain predictions.

p-Variogram Score

The p-Variogram Score (VS-p) is a proper scoring rule that designed to evaluate the dependency structure of multivariate probabilistic forecasts. Rather than focusing on marginal distributions, VS-p compares the empirical variogram of the observation with the variogram implied by the forecast distribution. It is defined as

$$\mathrm{VS}_p(F_t, X_t) = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}\left(|X_{t,i} - X_{t,j}|^p - \frac{1}{M}\sum_{m=1}^{M}\left|\hat{X}_{t,i}^{(m)} - \hat{X}_{t,j}^{(m)}\right|^p\right)^2. \tag{2.50}$$

where $X_t \in \mathbb{R}^d$ denotes the observed multivariate outcome and $\hat{X}_{t,i}$ denotes the i-th element of $\hat{X}_t$, which is a random vector drawn from the predictive distribution F_t. w_{ij} are non-negative weights specifying the importance of each component pair, and $p > 0$ is a user-defined exponent. The score is derived from the transformation

$$T_{ij}(x) = |X_i - X_j|^p \tag{2.51}$$

which computes the p-th power of the absolute difference between component pairs. This transformation mitigates the influence of marginal shifts and emphasizes the relative dependence structure, making VS-p particularly suitable for assessing spatial, temporal, or spatio-temporal covariance forecasts.

Logarithmic Score

The Logarithmic Score (Log Score) evaluates probabilistic forecasts by directly assessing the predictive density at the observed outcome. Unlike the Energy Score, it does not require sampling from the predictive distribution and is therefore computationally efficient. It is defined as

$$S(p_t(X_t)) = -\frac{1}{T}\sum_{t=1}^{T}\ln(p_t(X_t)) \tag{2.52}$$

where $p_t(\cdot)$ is the predictive density at time t and X_t is the corresponding observation. The Log Score is a strictly proper and local scoring rule, meaning it depends only on the predictive density evaluated at the realized observation Bröcker and Smith (2007). Moreover, it aligns naturally with maximum likelihood–based parameter estimation. However, due to the logarithmic transformation, it is highly sensitive to outliers and may perform poorly in the presence of noisy or heavy-tailed data.

2.4 Summary

In this chapter, we introduce the parameterisation of the covariance matrix and review existing covariance modelling approaches in both static and dynamic settings. We also summarise current evaluation metrics for covariance estimation and a closely related application, probabilistic forecasting. Covariance matrix parameterisation falls broadly into two categories: parametric covariance functions, encompassing both stationary and non-stationary formulations, and unconstrained matrix decompositions, such as the modified Cholesky decomposition. Non-stationary covariance functions offer a parsimonious representation of evolving dependence structures, yet ensuring that the resulting estimates remain positive definite poses a substantial challenge. This highlights the need for an estimation framework that guarantees positive definiteness in the non-stationary setting—a gap we address in detail in Chapter 3.

Chapter 3

Flexible estimation of non-stationary covariance functions

In this chapter, we develop a flexible estimation framework for non-stationary covariance functions that addresses both of these requirements. The framework combines additive models for the parameters of a structured covariance specification with a suite of regularisation strategies designed to enforce positive definiteness throughout the optimisation procedure. We begin by reviewing the objective functions commonly employed in covariance model fitting, including likelihood-based and loss-based formulations. We then turn to the central methodological contribution of this chapter: a systematic treatment of techniques for ensuring positive definiteness. These range from smoothing penalties and eigenvalue-based regularisation to barrier methods drawn from interior-point optimisation, each offering a different trade-off between computational cost, numerical stability, and the strength of the positive definiteness guarantee. Finally, we present a simulation study in which data are generated from known non-stationary covariance models. The experiments demonstrate that the proposed regularisation strategies, the log-barrier penalty and the one-sided Huber loss on eigenvalues in particular, not only ensure that the estimated covariance function remains positive definite across the entire index domain, but also lead to improved estimation accuracy relative to their unregularised counterparts.

3.1 Ensuring Positive Definiteness

To address the positive-definiteness issue arising in non-stationary covariance modelling, we introduce a regularisation framework that enforces the positive-definite constraint during optimisation. Regularisation is a fundamental technique in statistics and machine learning for improving generalisation performance and stabilising solutions to ill-posed or high-dimensional estimation problems. It is typically implemented by augmenting the objective function with a penalty term that constrains the parameter space and discourages undesirable solutions.

In its general form, regularisation modifies the optimisation problem by adding a penalty functional to the loss. For example, ℓ_2 regularisation incorporates the squared Euclidean norm of the

parameter vector into the objective, shrinking parameter magnitudes and promoting numerical stability. This regularisation technique provides a potential solution to maintain the structural validity in non-stationary covariance modelling. Since non-stationary covariance functions do not automatically preserve positive definiteness during optimisation, constrained or structure-inducing regularisation is required. We therefore propose a positive-definite regularisation scheme that explicitly penalises violations of positive definiteness, thereby guiding the optimisation procedure toward admissible covariance estimates while retaining modelling flexibility.

When optimisation algorithms are applied directly to covariance estimation, particularly under non-stationary parameterisations, the positive-definiteness constraint poses a substantial challenge. In such settings, unconstrained optimisation may yield symmetric matrices that are not admissible covariance matrices. Consequently, the estimation problem must be formulated as a constrained optimisation task to ensure structural validity

$$\underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(\hat{\Sigma}(\theta), \Sigma), \forall \hat{\lambda}_i \geq 0 \tag{3.1}$$

where $\hat{\lambda}_i$ denotes the i-th eigenvalue of the estimated covariance matrix $\hat{\Sigma}(\theta)$. The eigenvalue constraints guarantee that $\hat{\Sigma}(\theta)$ remains positive definite throughout optimisation. A common strategy for satisfying this constraint is Cholesky reparameterisation whereby the covariance matrix is expressed in terms of its Cholesky factor and the optimisation is performed over the unconstrained elements of this factor rather than over the covariance matrix itself. While this approach ensures positive definiteness by construction, it increases the number of estimated parameters that scales with the square of matrix dimension whereas the covariance function can be relatively parsimonious and fixed complexity. To address this limitation, we introduce a positive-definite regularisation framework designed specifically for covariance functions, which encourages the estimated covariance matrices to remain positive definite while preserving the flexibility of the functional parameterisation.

Analogous to norm-based regularisation that imposes restrictions on the norm of model's parameters, positive definiteness regularisation structural penalties that encourage the estimated covariance matrix to remain positive definite. In general, however, an explicit closed-form relationship between positive definiteness and the underlying model parameters is unavailable, particularly when the covariance matrix is generated through a complex non-stationary covariance function. A direct and practical strategy is therefore to penalise violations of the defining properties of positive definiteness. Two common approaches are: (i) introducing a penalty that enforces positivity of the eigenvalues, and (ii) incorporating a determinant-based penalty motivated by the fact that a positive-definite matrix must have strictly positive determinant. These ideas lead to regularised objectives of the form

$$\mathcal{L}_\Phi(\Sigma, \hat{\Sigma}(\theta)) = \mathcal{L}(\Sigma, \hat{\Sigma}(\theta)) + \gamma\Phi(\hat{\Lambda}(\hat{\Sigma}(\theta))) \tag{3.2}$$

$$\mathcal{L}_\Phi(\Sigma, \hat{\Sigma}(\theta)) = \mathcal{L}(\Sigma, \hat{\Sigma}(\theta)) + \gamma\Phi(\det(\hat{\Sigma}(\theta))) \tag{3.3}$$

where $\mathcal{L}(\Sigma, \hat{\Sigma}(\theta))$ is the unregularized loss function and $\Phi(\hat{\Lambda}(\hat{\Sigma}(\theta)))$ is a penalty functional defined on the vector of estimated eigenvalues $\hat{\Lambda}$ of $\hat{\Sigma}(\theta)$, and $\gamma > 0$ controls the strength of regularisation. By appropriately designing $\Phi(\hat{\Lambda}(\hat{\Sigma}(\theta)))$, penalising negative or near-zero eigenvalues for example, the optimisation procedure can be guided toward admissible covariance estimates while preserving flexibility in the underlying parameterisation. The following content of this section discusses 4 regularisations used to encourage positive-definite output.

3.1.1 One-side Huber Loss

Penalising the smallest eigenvalue of the estimated covariance matrix provides a direct mechanism for encouraging positive definiteness. If the minimum eigenvalue is driven toward positive values, the spectrum of the matrix is correspondingly shifted away from the non–positive region, thereby improving the likelihood that the matrix becomes positive definite. Consequently, the regularisation function $\Phi(\min(\hat{\lambda}))$ is supposed to should impose strong penalties when the estimated smallest eigenvalue is negative or close to zero, while applying little or no penalty when it is sufficiently positive. To achieve this behaviour, we adopt a regularisation term based on a one-sided Huber loss. The Huber loss combines quadratic and linear penalties, providing robustness while maintaining stable gradients during optimisation. The regularisation function is defined as

$$\Phi(\hat{\lambda}) = r(u) = \begin{cases} \frac{1}{2}u^2 & 0 < u \leq \alpha \\ \alpha(u - \frac{1}{2}\alpha) & u > \alpha \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

where $u = \epsilon - \hat{\lambda}_{\min}$ with $\hat{\lambda}_{\min}$ denoting the smallest estimated eigenvalue of $\hat{\Sigma}(\theta)$. The hyperparameter $\alpha > 0$ as well as $\epsilon > 0$ control the behaviour of the penalty. Specifically, ϵ defines the threshold below which eigenvalues begin to incur a penalty, while α determines the transition point between quadratic and linear growth of the loss. Under this formulation, eigenvalues larger than ϵ incur no regularisation penalty, ensuring that sufficiently positive eigenvalues remain unaffected. When $\hat{\lambda}_{\min} < \epsilon$, the penalty increases according to the distance between $\hat{\lambda}_{\min}$ and ϵ. For moderate violations, the quadratic term provides smooth gradients that facilitate optimisation. For larger violations (i.e., when $u > \alpha$), the loss grows linearly, preventing excessively large gradients that could destabilise training. In practice, ϵ is typically chosen as a small positive constant (near zero) to encourage strictly positive eigenvalues while avoiding distortion of well-behaved spectra due to significant penalty. The illustration of this one-sided Huber loss is shown in Figure (3.1) $\epsilon = 1$ and $\alpha = 5$ are set for visualisation. Negative eigenvalues are penalised increasingly as they move further below the threshold, while extremely negative values transition to a linear penalty to avoid numerical instability. The empirical effect of this positive-definiteness regularisation is further discussed in Section 3.2.2.

3.1.2 Interior-Point Methods (Barrier Methods)

While the one-sided Huber loss directly penalises violations of the positive definiteness condition through the smallest eigenvalue, it relies on spectral decomposition and targets only the minimum eigenvalue of the matrix. An alternative strategy is to incorporate the structural constraint of positive definiteness directly into the objective function through interior-point methods, also known as barrier methods. These methods incorporate the constraints of an optimisation problem directly into the objective function through a barrier term that discourages the optimisation trajectory from approaching the boundary of the feasible set. Instead of explicitly enforcing constraints, the barrier function implicitly keeps the iterates within the interior of the feasible region throughout optimisation. Such methods are widely used in convex optimisation and operate by progressively approximating the constrained problem using a sequence of unconstrained problems Boyd and Vandenberghe (2004).

In the context of covariance estimation, the feasible set consists of the cone of positive definite matrices Nesterov and Nemirovskii (1994). A natural barrier for this set is the logarithmic determinant function, which diverges when the determinant approaches zero. Since the determinant of

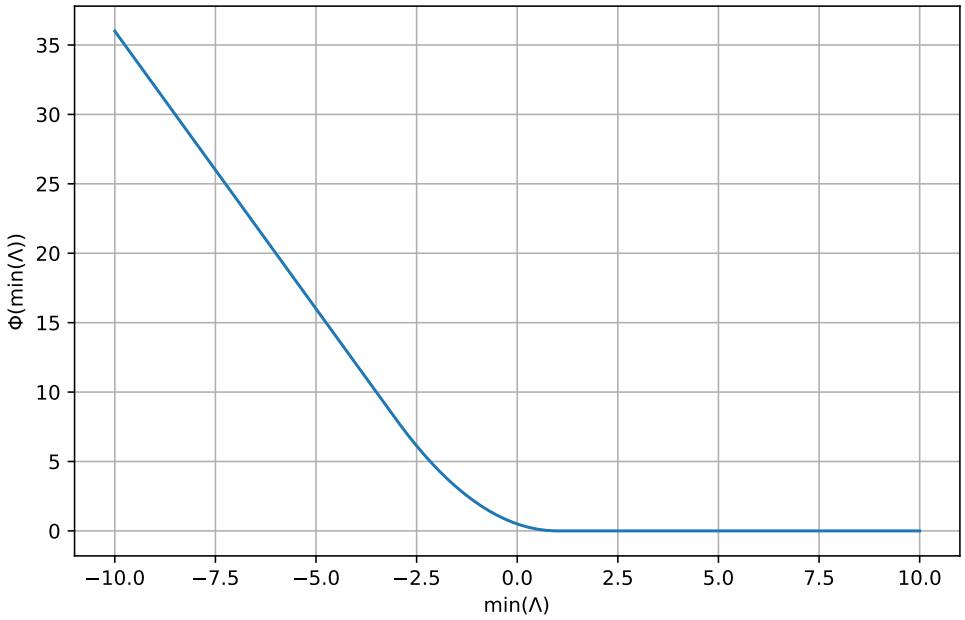

Figure 3.1: Visualisation of One-sided Huber Loss(3.4). A regularisation function to penalize the eigenvalues of the predicted covariance matrix. In this visualisation example, $\epsilon = 1.0$, $\alpha = 4.0$ is used and a clear turning point can be observed at $\min(\Lambda) = 1.0$.

a positive definite matrix equals the product of its eigenvalues, it approaches zero when the matrix approaches the boundary of the positive semidefinite cone. Consequently, the logarithmic barrier imposes an increasingly large penalty as the covariance matrix becomes nearly singular, thereby preventing the optimisation procedure from leaving the positive definite region. Following this principle, the regularised objective function can be written as

$$\mathcal{L}_{\Phi}(\hat{\Sigma}(\theta), \Sigma) = \mathcal{L}(\hat{\Sigma}(\theta), \Sigma) - \gamma \log\Big(\det(\hat{\Sigma}(\theta))\Big)\,, \tag{3.5}$$

where $\mathcal{L}(\hat{\Sigma}(\theta), \Sigma)$ denotes the original loss function and $\gamma > 0$ is the barrier parameter controlling the strength of the constraint. The logarithmic barrier grows rapidly when $\det(\hat{\Sigma}(\theta))$ approaches zero, which corresponds to one or more eigenvalues approaching zero. As a result, matrices that are close to singular receive large penalties, encouraging the optimisation algorithm to remain within the interior of the positive definite cone.

During optimisation, the barrier parameter γ is normally decreased gradually, allowing the solution to approach the boundary of the feasible region while maintaining numerical stability. When γ is large, the barrier strongly restricts the search space and ensures strictly positive eigenvalues. As γ decreases, the optimisation progressively approximates the original constrained problem. This path-following strategy is a defining characteristic of interior-point methods and allows the algo-

rithm to approach the optimal solution while preserving the positive definiteness of the estimated covariance matrix throughout the optimisation process.

3.1.3 Lagrange Multiplier

The Lagrange multiplier framework shares similarities with regularisation techniques in that the constraint is incorporated directly into the objective function. However, unlike standard regularisation approaches where the penalty weight is typically selected through cross-validation, the Lagrange multiplier is treated as an additional variable to be estimated as part of the optimisation procedure. This allows the constraint to be enforced more explicitly rather than indirectly through a fixed penalty coefficient. Suppose that the objective function $\mathcal{L}(\theta)$ is to be optimised subject to an equality constraint $g(\theta) = 0$. The constrained problem can be reformulated through the Lagrangian function

$$\mathcal{L}(\theta, \alpha) = \mathcal{L}(\theta) - \alpha g(\theta), \tag{3.6}$$

where α is the Lagrange multiplier associated with the constraint. At an optimal solution, the gradient of the objective function is proportional to the gradient of the constraint function, yielding the first-order condition

$$\nabla\mathcal{L}(\theta) = \alpha\nabla g(\theta). \tag{3.7}$$

This condition implies that no feasible descent direction exists that simultaneously improves the objective function while satisfying the constraint Boyd and Vandenberghe (2004). For positive definiteness constraints in covariance estimation, the constraint function can be defined based on the eigenvalues of the estimated covariance matrix. Let $\hat{\Lambda} = (\hat{\Lambda}_1, \ldots, \hat{\Lambda}_d)$ denote the vector of estimated eigenvalues of $\hat{\Sigma}(\theta)$. Positive definiteness requires that all eigenvalues are strictly positive, which can be expressed as

$$g(\hat{\Lambda}) = \min(\hat{\Lambda}) \geq 0. \tag{3.8}$$

To incorporate this inequality constraint within the Lagrangian framework, a common approach is to introduce a slack variable that converts the inequality constraint into an equality constraint. Specifically, let

$$g(\hat{\Lambda}) - r = 0, \qquad r \geq 0, \tag{3.9}$$

where r is a non-negative slack variable representing the distance to the boundary of the feasible region. To eliminate the explicit inequality constraint on r, it can be parameterised as an exponential function $r = \exp(z)$, which guarantees positivity for any real-valued z. Consequently, the set of optimisation variables becomes (θ, α, z). In practice, pure Lagrangian formulations may suffer from numerical instability. To alleviate this issue, the augmented Lagrangian method introduces an additional quadratic penalty term that stabilises the optimisation process. The augmented Lagrangian can be written as

$$\mathcal{L}_{\text{aug}} = \mathcal{L}(\hat{\Sigma}(\theta), \Sigma) - \alpha g(\hat{\Lambda}(\hat{\Sigma}(\theta))) + \frac{\rho}{2} g(\hat{\Lambda}(\hat{\Sigma}(\theta)))^2, \tag{3.10}$$

where $\rho > 0$ is a penalty parameter controlling the strength of the quadratic regularisation. This formulation combines the advantages of penalty methods and Lagrange multipliers, improving convergence behaviour while maintaining constraint enforcement.

Although Lagrangian approaches provide a principled framework for constrained optimisation, their application to dynamic covariance modelling may be computationally challenging in high-dimensional settings. The introduction of additional multipliers and auxiliary variables increases the number of optimisation parameters, which may significantly increase computational cost. Moreover, the constraint function based on eigenvalues involves spectral decomposition, which is nonlinear and may lead to complicated gradient calculations. These challenges limit the practicality of the Lagrange multiplier approach for large-scale covariance estimation problems.

3.1.4 Shrinking Regularisation

Shrinking regularisation provides a simple and computationally efficient strategy for enforcing positive definiteness of covariance matrices. The core idea is to modify a potentially indefinite symmetric matrix by shrinking it toward a positive-definite target matrix, most commonly the identity matrix. In its simplest form, this is achieved by adding a small multiple of the identity matrix to the estimated covariance matrix. Such diagonal inflation improves numerical conditioning and guarantees positive definiteness when the shrinkage parameter is sufficiently large. This technique is widely used in covariance estimation and ridge-type regularisation methods.

In the context of covariance modelling, the regularised covariance estimator can be expressed as

$$\hat{\Sigma}(\theta) = \hat{\Sigma}_{raw}(\theta) + kI\sigma \tag{3.11}$$

where the $\hat{\Sigma}_{raw}(\theta)$ denotes the predicted raw covariance matrix and I denotes the identity matrix. σ is a minor positive value and k is dependent on $\hat{\Sigma}_{raw}(\theta)$, indicating iterative addition of σ on the diagonal. Intuitively, the addition of $k\sigma I$ shifts all eigenvalues of the matrix by the same amount, thereby increasing the smallest eigenvalue and moving the matrix away from the boundary of the positive semidefinite cone. However, this method also suffers from complicated gradient calculation when using gradient-based optimisation as k is implicitly determined by $\hat{\Sigma}_{\text{raw}}(\theta)$ and therefore depends on the model parameters θ. To address this issue, we adopt the straight-through estimator (STE), a commonly used technique for approximating gradients through non-differentiable operations in deep learning. The straight-through estimator approximates the derivative of the shrinkage operation by treating it as an identity mapping during gradient calculation. Consequently, the gradient of the regularised covariance matrix is approximated as

$$\frac{\partial \hat{\Sigma}(\theta)}{\partial \theta} \simeq \frac{\partial \hat{\Sigma}_{raw}(\theta)}{\partial \theta}. \tag{3.12}$$

Under this approximation, the optimisation process ignores the gradient contribution from the shrinkage operation and propagates gradients only through the raw covariance prediction. Although this approximation introduces bias in the gradient direction, it greatly simplifies the optimisation procedure and enables efficient training using standard gradient-based algorithms. It is worth noting that this approximation becomes exact when the shrinkage parameter is independent of θ, i.e., when a fixed diagonal inflation $k\sigma I$ is applied. In practice, however, the adaptive formulation is preferred because it automatically adjusts the amount of shrinkage required to restore positive definiteness.

Overall, the combination of shrinking regularisation and the straight-through estimator provides a computationally inexpensive and easily interpretable approach for maintaining positive definiteness in non-stationary covariance modelling. The added term $k\sigma I$ can be interpreted as the minimal diagonal adjustment required to move the estimated matrix back into the positive definite cone while preserving the structure of the raw covariance estimate.

3.1.5 Choice of Positive-definite Regularisations

To conclude, the positive-definite regularisations introduced above either imposing penalty on the eigenvalues of the estimated covariance matrix or punishing negative determinant. One-side Huber loss penalty relies on calculating the smallest eigenvalue of the covariance matrix while Interior-Point methods construct a barrier around the boundary of non-positive-definite space in the parameters space, encouraging large positive determinant of the covariance matrix. These two methods are easily to be adopted and can be controlled via regularisation strength γ, which do not directly change the regularised estimation framework in stochastic optimisation process. Shrinking regularisation acts as a fall-back solution to map ill-posed covariance matrix to nearest positive-definite space when out-of-distribution covariates are given, ensuring the estimated matrix is positive-definite. Lagrange Multiplier introduces an additional estimated parameter for each constraint inequality, such that estimation framework become more flexible but unstable while estimating a series of additional parameters simultaneously. Therefore, the empirical study of the Lagrange Multiplier regularisation is not discussed in this thesis and the following simulation experiment as well as the application in Chapter 4 employ one-side Huber loss, barrier method and shrinking regularisation as a positive-definite insurance. Since the value of regularisation strength γ may significantly affect the estimated performance, we empirically select the best γ on the test data in the following simulation and real-world application instead of using cross-validation, which is a recommended penalty strength selection process to balance singualrity of the estimated covariance and overfitting.

3.2 Simulation Experiment

To evaluate the proposed positive-definite penalty on covariance functions, simulation studies are conducted using synthetic data. The simulations consider several scenarios, including dynamic covariance modelling under a linear evolution assumption and isotropic covariance estimation. In addition, static non-stationary covariance estimation is also examined. All simulations adopt parametric covariance functions that represent the covariance structure through a single covariance function, which factorizes the covariance matrix based on pairwise distances. In geostatistical modelling, a random process is commonly represented as $Z(s,t)$, where s denotes the spatial location and t represents time. The spatio-temporal covariance structure can therefore be expressed through a covariance function describing the dependence between observations separated in space and time. In the following simulations, we only discuss the spatial stationarity and ignore the temporal stationarity, and according to the definition of stationarity, the isotropic covariance function in spatial covariance modelling depends only on the spatial lag Δs, written as

$$C(\|\Delta s\|) = \mathrm{Cov}(Z(s,t), Z(s+\Delta s,t)) \tag{3.13}$$

where Δs denote the spatial separation. In this formulation, the covariance depends solely on the magnitude of the separation rather than the absolute locations, which characterizes an isotropic and stationary covariance structure. In contrast, under a non-stationary setting the covariance function depends on the absolute spatial and temporal locations of the observations. The covariance can therefore be written more generally as

$$C(s_1, s_2) = \mathrm{Cov}(Z(s_1,t), Z(s_2,t)). \tag{3.14}$$

To evaluate the estimation framework and its generalisation, two widely used parametric covariance families are considered: the powered exponential covariance function (3.15) and the Matern (3.17)

covariance function. For the powered exponential covariance function, a scaled sigmoid transformation is used to constrain the exponent parameter within a stable range during optimisation. The modified covariance function is defined as

$$C(r;\xi) = \sigma^2 e^{-(\theta r)^{\delta(\kappa)}} \tag{3.15}$$

$$\delta(\kappa) = \frac{1}{1+e^{-\kappa}} \tag{3.16}$$

where r denotes the spatial distance between two locations and $\delta(\cdot)$ is the sigmoid function used to bound the exponent parameter. The Matern covariance function is defined as

$$C(r;\xi) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2}\nu\frac{r}{\rho}\right)^{\nu} K_\nu \left(\sqrt{2}\nu\frac{r}{\rho}\right) \tag{3.17}$$

$$\rho(x) = \frac{1}{1+e^{-x}} \tag{3.18}$$

where ρ denotes the range parameter controlling spatial correlation decay, ν is the smoothness parameter, K_ν is the modified Bessel function of the second kind, and $\Gamma(\cdot)$ is the Gamma function. We use cubic GAM to estimate the parameter θ for powered exponential function (3.15) and parameter ρ for Matern covariance function (3.17). Specifically, the number of basis functions is set to be 100 and the cubic GAM is defined as

$$\theta(\mathbf{x}_t) = \mathbf{B}(\mathbf{x}_t)\boldsymbol{\beta}^\top + \mathbf{b} = \sum_{m=1}^{p}\sum_{n=1}^{B} \beta_{m,n}\,\phi_{m,n}(x_{m,t}) + b \tag{3.19}$$

$$\rho(\mathbf{x}_t) = \mathbf{B}(\mathbf{x}_t)\boldsymbol{\beta}^\top + \mathbf{b} = \sum_{m=1}^{p}\sum_{n=1}^{B} \beta_{m,n}\,\phi_{m,n}(x_{m,t}) + b \tag{3.20}$$

where $\mathbf{x}_t$ denotes the covariate vector and $\mathbf{B}(\mathbf{x})$ denotes the basis matrix constructed for covariates. Each scalar $x_{m,t}$ corresponds to the element of covariate vector $\mathbf{x}_t \in \mathbb{R}^p$, which is the value of m-th covariate observed at t timestamp. The coefficient vector $\boldsymbol{\beta}$ and intercept $\mathbf{b}$ are estimated and p is the number of covariates while B denotes the number of basis functions. Basis function n of covariate m is defined as $\phi_{m,n,t}(x_m)$. The notation $\theta_t = \theta(\mathbf{x}_t)$ and $\rho_t = \rho(\mathbf{x}_t)$ denote time-varying parameters for powered exponential and Matern covariance function, respectively.

We conduct two simulation studies in this section including dynamic isotropic covariance estimation and static non-stationary covariance estimation. Positive-definiteness is guarantee in isotropic setting and therefore, positive-definite regularisations are not examined in this simulation, which mainly focus on evaluating the capability of covariance function to capture the dynamic pattern in time-varying covariance modelling. While in the static non-stationary situation, positive-definite regularisations discussed in the previous section are employed and their empirical results are analysed.

3.2.1 Dynamic Isotropic Covariance Estimation

The following simulations focus on spatial dynamic covariance estimation, where the spatial distance is interpreted as the element-wise distance in the covariance matrix. Specifically, the spatial separation for the covariance entry $\mathrm{Cov}_t(Z(s_i,t), Z(s_j,t))$ is defined as $r = |i-j|$, where r denotes

the spatial distance, and i and j represent the row and column indices of the covariance matrix, respectively. Since this is in dynamic isotropic setting, the covariance matrix is spatially stationary at each time and non-stationary in temporal domain. In general, the covariance structure evolves with the temporal covariate x_t. The covariance matrix estimated at time t is therefore generated by a parametric covariance function of the form $C(r, x_t; \xi)$, where ξ denotes the set of parameters to be estimated. The following sections describe the simulation framework in detail.

In this simulation, the powered exponential covariance function is adopted, as defined in (3.15). To simplify the estimation procedure, the parameters σ and κ are fixed as constants, and the only parameter to be estimated is θ, which is assumed to vary as a function of the temporal covariate x_t. The synthetic data are generated under the assumption that

$$\theta_t = \sin(2\pi x_t) + 2. \tag{3.21}$$

To better reflect practical situations in which the true covariance matrices are unobservable and the covariance structure at each time point must be inferred from limited observations, we generate 5000 samples from multivariate normal distributions governed by the true covariance matrices. First, a realization $x_t \in \mathbb{R}^{10}$ of $X_t \sim U(0, 1)$ is sampled to represent the temporal covariate at time t. The corresponding covariance matrix at each time point is then generated using the powered exponential covariance function shown in (3.15), with θ specified by (3.21). These time-varying covariance matrices are denoted by Σ_t. The full weighted least squares loss defined in (2.46) is used as the base objective function for parameter estimation. We apply no positive definite regularisations to the weighted least square error as parametric covariance functions remain positive-definite in stationary settings. To simulate the real application that the real covariance matrix is not accessed during optimisation, empirical covariance matrix $\tilde{\Sigma}_t$ is used in the objective function. Since the number of basis functions is set to 100, which provides provides substantial flexibility at each knot, smoothing penalties are necessary to avoid overfitting. Therefore, a second-derivative smoothing penalty is introduced to control the roughness of the estimated function. The full objective function can be written as

$$\mathcal{L} = \frac{1}{d^2 T} \sum_{t=1}^{T} \left(\frac{\tilde{\Sigma}_t - \hat{C}(R_{i,j}; \xi_t)}{1 - \hat{C}_{corr}(R_{i,j}; \xi_t)} \right)^2 + \gamma \beta^\top S \beta \tag{3.22}$$

where γ is the regularisation strength and β represents the vector of cubic GAM coefficient parameters. The matrix S is a positive-definite smoothing penalty matrix from the second derivatives of the basis functions. This matrix quantifies the roughness of the fitted function, penalising excessive curvature and encouraging smoother estimates. In spline-based GAMs, such penalties are commonly formulated using the integrated squared second derivative of the smooth function, which directly quantifies its 'wiggliness'. Positive-definite regularisation is not employed in this task due to isotropic setting. Optimisation is performed using the Adam algorithm to improve numerical stability during training. Since this is a simulation to the multivariate probabilistic forecasting task, instead of the accuracy of the estimated covariance matrix, we are also interested to discover if more accurate dynamic covariance modelling improve the forecasting accuracy. Therefore, we use KL-divergence (2.48) to evaluate the accuracy of the estimated covariance matrix and leverage Energy Score (2.49) as well as p-Variogram Score (2.50) to evaluate the prediction accuracy. The fitted parameter θ is visualised in Figure 3.2.

From the results, the model fitted demonstrates high fitting accuracy with an appropriate smoothing regularisation strength, which indicates that the covariance function estimated using

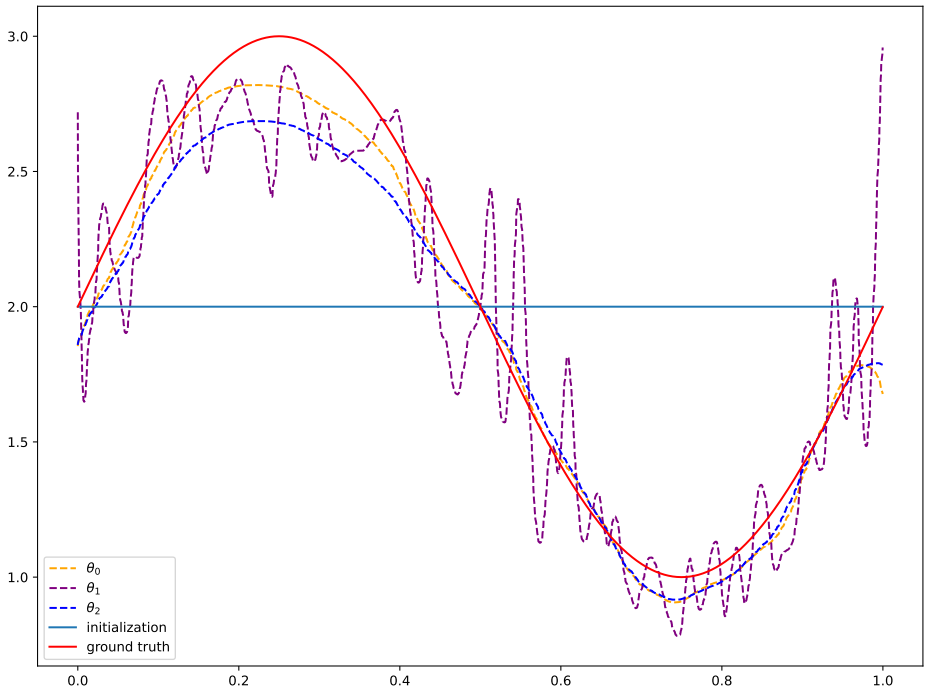

Figure 3.2: Fitted θ result of dynamic isotropic covariance estimation. All the estimated θ is derived from powered exponential function with different regularisation strength: θ_0: fitted θ $\gamma = 0$; θ_1: fitted θ with $\gamma = 0.5$; θ_2:fitted θ with $\gamma = 1$; Log-likelihood is used as optimisation objective

a Generalized Additive Model (GAM) effectively captures the dynamic behaviour of the covariance structure. It is evident that without smoothing regularisation, the cubic GAM overfits the samples due to high flexibility provided by large number of basis functions (see $\gamma = 0$ in Figure 3.2). However, with the penalty is strengthened (large γ in Figure 3.2), the fitted θ is gradually away from the ground truth assumption. To determine an appropriate level of smoothing for the given dataset, a simulation study is conducted with the regularisation parameter γ ranging from 0 to 3. The corresponding results are presented in Figure 3.3.

From the regularisation evaluation results, the loss exhibits a similar trend to the KL divergence as the regularisation strength varies, whereas the Energy Score and the 0.5-Variogram Score fluctuate more noticeably across different regularisation levels. A clear U-shaped pattern is observed for both the loss and KL divergence, indicating the presence of an optimal regularisation strength around 0.375. In contrast, the Energy Score reaches its minimum at approximately 1.375, while the Variogram Score attains its lowest value near 0.2. Since the performance differences within the interval $\gamma \in [0.25, 0.5]$ are relatively small, and the Energy Score does not deteriorate substantially

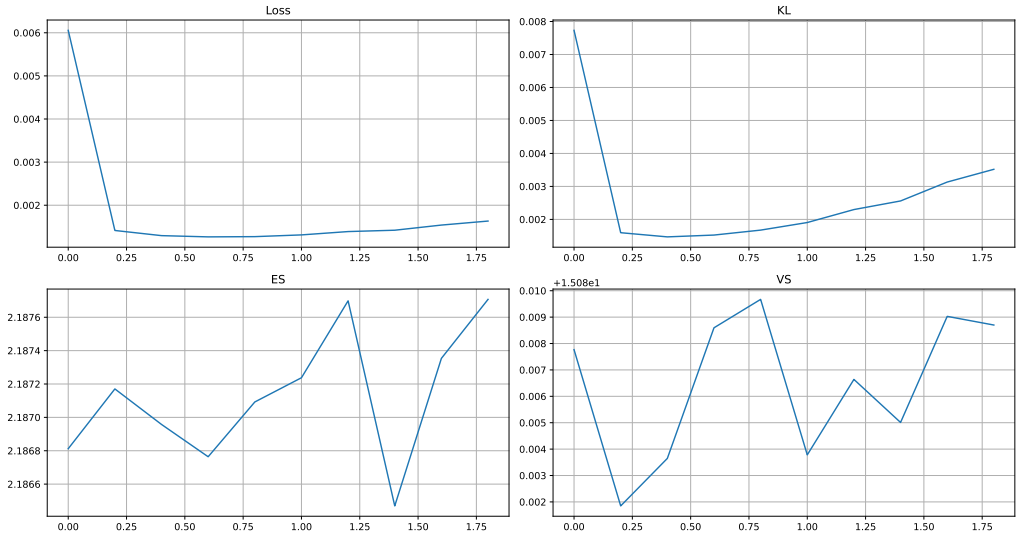

Figure 3.3: Regularisation Strength Evaluation. The x-axis denotes the regularisation strength and the y-axis denotes the evaluation metric value. Loss denotes the log-likelihood objective function and KL denotes KL-divergence under multivariate normal distribution. ES denotes Energy Score and VS represents 0.5-Variogram Score. The number of forecasted trajectories is 500 to reduce sampling bias for Energy Score and 0.5-Variogram Score.

compared with its minimum at 1.375, a regularisation strength of $\gamma = 0.4$ is selected as a balanced choice. This value achieves favourable performance across multiple evaluation metrics while maintaining model stability. The quantitative results are summarised in Table 3.1. Since there is no significant discrepancy among different γ setting at scale for Energy Score and 0.5-Variogram Score, Negative Log-likelihood is the main objective in this simulation study. To further reduce the bias introduced by sampling variability, the distributions of the evaluation metrics under different numbers of sampling trajectories are presented in Figure 3.4. It is evident that both evaluation metrics converge when the number of forecast trajectories reaches approximately 300, indicating reduced variability caused by sampling. This suggests that using more than 300 trajectories is sufficient for reliable evaluation.

These simulation results indicate that, with an appropriate smoothing penalty, covariance functions with a limited number of estimated parameters can effectively model isotropic covariance structures. This demonstrates the strong potential of the proposed framework for dynamic covariance modelling. In addition, smoothing regularisation helps prevent excessive oscillations in the estimated function and consequently encourages positive-definite covariance estimates in the stationary setting. The covariance functions used for estimating non-stationary covariance structures are discussed in the following section.

3.2.2 Static Non-stationary Covariance Estimation

In this section, a simulation study of static non-stationary covariance matrices is conducted to evaluate the effectiveness of the proposed positive-definite regularisation methods. In contrast

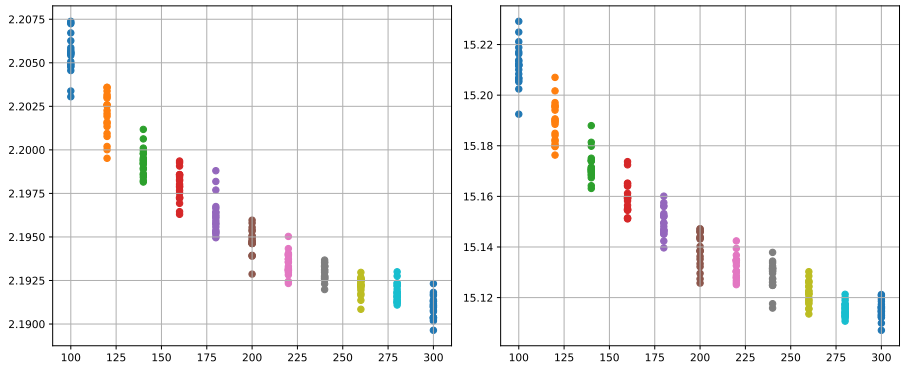

Figure 3.4: Evaluation metrics distributions over the number of forecasted trajectories. The left figure visualises the Energy Score distribution of different number of trajectories and the right figure visualises the distribution of 0.5-Variogram Score. $\gamma = 0.4$ is used for smoothing penalty strength.

Penalty Strength	**ES**	**KL**($1e^{-2}$)	**WLS**($1e^{-2}$)	**VS-0.5**
0	<u>2.1868</u>	0.773	0.606	15.088
0.2	2.1871	0.160	0.141	<u>15.082</u>
0.4	2.1870	<u>0.147</u>	0.129	15.084
0.6	<u>2.1868</u>	0.153	<u>0.127</u>	15.089
0.8	2.1871	0.168	<u>0.127</u>	15.090
1.0	2.1872	0.191	0.131	15.084

Table 3.1: Dynamic Isotropic Covariance Estimation Result. The underline highlights the best performance.

to stationary covariance structures, non-stationary covariance allows the dependence structure to vary across locations, meaning that covariance may change with spatial position rather than depending solely on separation distance. To examine the robustness and generalisation capability of the proposed estimation framework, both covariance functions, powered exponential (3.15) and Matern (3.17), are employed in the simulation study. In the case of static non-stationary setting, entries of covariance matrix with the same separation may differ according to their specific locations. To simulate this scenario and evaluate the generalisation of our estimation framework, we create non-stationary ground truth covariance matrices Σ in the following assumption:

$$\theta(l_1, l_2) = \frac{5}{1 + l_1 + l_2} \tag{3.23}$$

$$\rho(l_1, l_2) = \frac{5}{1 + l_1 + l_2} \tag{3.24}$$

Here, θ and ρ denote scalar parameters in the powered exponential covariance function and the Matern covariance function, respectively. The variables l_1 and l_2 represent the row and column indices corresponding to each element of the covariance matrix. All observations $X \sim \mathcal{N}(0, \Sigma)$ are sampled from a multivariate normal distribution with the true covariance matrix Σ. The scalar parameters of two covariance functions, θ and ρ, are estimated using a cubic GAM predictor

described as

$$\theta(x) = \mathbf{B}(x)\boldsymbol{\beta}^\top + \mathbf{b} = \sum_{n=1}^{B} \beta_n \, \phi_n(x) + b \tag{3.25}$$

$$\rho(x) = \mathbf{B}(x)\boldsymbol{\beta}^\top + \mathbf{b} = \sum_{n=1}^{B} \beta_n \, \phi_n(x) + b \tag{3.26}$$

where x denotes the value of $l_1 + l_2$ and we use set $\mathbf{x} = \{x = l_1 + l_2 | l_1 \in d, l_2 \in d\}$ to denote the vector containing all the potential values of $l_1 + l_2$. $\mathbf{B}$ denotes the basis matrix constructed from x and ϕ_n denotes the n-th basis function where $B = 100$ is the number of basis functions. The negative log-likelihood is adopted as the objective function and it is described as

$$\mathcal{L}_{ll}(\mu.\hat{\Sigma}|X) = \log \prod_{i=1}^{N} f_{\mathcal{N}}(X_i|\mu, \hat{\Sigma}) = -\frac{Nd}{2}\log(2\pi) - \frac{N}{2}\log\left|\hat{\Sigma}\right| - \frac{1}{2}\sum_{i=1}^{N}(X_i - \mu)^\top \hat{\Sigma}^{-1}(X_i - \mu) \tag{3.27}$$

where $f_{\mathcal{N}}$ is the probability density function of the multi-variate normal distribution and N is the number of observations. The sample mean μ is set to zero due to normal distribution assumption and $\hat{\Sigma}$ is the estimated covariance matrix. Adam optimiser is employed to improve optimisation stability. Due to stochastic optimisation and the absence of explicit constraints on the covariance function, the estimated covariance matrix $\hat{\Sigma}$ during optimisation may fail to remain positive definite. A valid covariance matrix must be positive semi-definite, meaning all eigenvalues are non-negative. To encourage positive definiteness in the estimated matrices, two regularisation approaches—namely the barrier method (3.5) and Huber Loss regularisation (3.4)—are imposed, and multiple regularisation strengths are evaluated. The overall objective function for Huber loss regularisation is described as

$$\mathcal{L}_{\text{huber}} = \mathcal{L}_{ll}(\mu.\hat{\Sigma}(\theta)|X) + \gamma\Phi(\hat{\lambda}) \tag{3.28}$$

$$\Phi(\hat{\lambda}) = r(u) = \begin{cases} \frac{1}{2}u^2 & 0 < u \leq \alpha \\ \alpha(u - \frac{1}{2}\alpha) & u > \alpha \\ 0 & \text{otherwise} \end{cases} \tag{3.29}$$

where $\hat{\lambda}$ denotes the smallest eigenvalue of $\hat{\Sigma}(\theta)$ and parameter u is calculated through $u = \epsilon - \hat{\lambda}$. Parameter $\epsilon = 1$ is set for the threshold of regularisation (see Figure 3.1 for detail). Parameter α is set to be 50 to increase the penalty and γ controls the regularisation strength on the overall loss. Similarly, the overall loss with the barrier method is written as

$$\mathcal{L}_{\text{barrier}} = \mathcal{L}_{ll}(\mu.\hat{\Sigma}|X) - \gamma \log\left(\det\left(\hat{\Sigma}\right)\right). \tag{3.30}$$

In addition to these two regularisation, the shrinking regularisation (3.11) is applied to $\hat{\Sigma}$ in case of non-positive-definite covariance matrix is estimated during optimisation and $\mathcal{L}_{\text{huber}}$ as well as $\mathcal{L}_{\text{barrier}}$ can not be calculated. This regularisation is applied only $\hat{\Sigma}$ become not positive-definite as an insurance due to the distortion of the estimated covariance matrix. Yet, during the simulation, this regularisation is never triggered with both regularised objective functions.

Since the covariance matrix Σ is static, trajectory-based evaluation metrics such as the Energy Score (2.49) and the p-Variogram Score (2.50) are not applicable, as no predictive trajectories

are generated. Instead, the fitted models are evaluated using KL divergence and the weighted mean square error. To better reveal the effect of both positive-definite regularisations, We use the non-regularised covariance function (i.e. $\gamma = 0$) as the baseline. The overall simulation setting is summarised in Table 3.2.

PD_regularisation	PD_ensurance	Covariance Function	Data Generation		
			θ func	num_samples	dim
Huber loss	Shrinking+STE	Powered Exponential	$\frac{5}{1+l_1+l_2}$	100000	14
Barrier Method	Shrinking+STE	Powered Exponential	$\frac{5}{1+l_1+l_2}$	100000	14
Huber loss	Shrinking+STE	Mattern	$\frac{5}{1+l_1+l_2}$	100000	14
Barrier Method	Shrinking+STE	Mattern	$\frac{5}{1+l_1+l_2}$	100000	14

Table 3.2: Experiment Running set. Objective function is set to be negative loglikelihood. θ func denotes the non-stationary assumption that the certain parameters follow, which are θ for powered exponential and ρ for Matern. The evaluation metrics of non-stationary setting are KL-divergence, Weighted Least Square Distance, and Log likelihood. Shrinking denotes the shrinking regularisation and STE denotes straight through estimator, which means the gradient backward propagation skip the shrinking regularisation entirely during Adam optimisation.

Negative log-likelihood (2.43) without additional regularisation terms, weighted least squares (2.47), and KL divergence are adopted as evaluation metrics. The negative log-likelihood measures the goodness of fit by assessing the probability of the observed data under the assumed model, while KL divergence quantifies the discrepancy between two probability distributions. Weighted least squares directly measures the estimated covariance structure distance to the ground truth covariance matrix. To determine the optimal strength of the positive-definite regularisation, each experiment is conducted using 16 different regularisation strengths ranging from 0.0 to 1.6 evenly. To mitigate potential sampling bias, each configuration is repeated 20 times. The aggregated results are presented in Table 3.3 and Table 3.4.

Regularisation	PD_ensurance	Neg Log-likelihood	WLS(10^{-5})	KL(10^{-4})	Function
Baseline	nearest	$1.566 * 10^5 \pm 1990.827$	732 ± 164	363 ± 188	PE
Huber loss	Shrinking+STE	$1.564 * 10^5 \pm 1627.0468$	$\mathbf{486 \pm 50.3}$	97.7 ± 128	PE
Barrier Method	Shrinking+STE	$\mathbf{1.563 * 10^5 \pm 590.360}$	521 ± 61.8	$\mathbf{74.5 \pm 25.6}$	PE

Table 3.3: Non-stationary Experiment Results. PE denotes powered exponential covariance function. WLS represents Weighted Least Square and KL records the KL divergence. The mean and standard deviation are reported for each metrics and the bold value indicates the best performance.

Compared with the baseline model, where no positive-definite regularisation is applied, both Huber loss regularisation and the Barrier Method consistently improve the goodness of fit in non-stationary covariance modelling. Among the evaluated approaches, the Barrier Method achieves the best overall performance. Specifically, it improves the weighted least squares metric (2.47) by approximately 28% and reduces the KL divergence by nearly 78% relative to the baseline.

Regularisation	PD_ensurance	Neg Log-likelihood	WLS(10^{-5})	KL(10^{-4})	Function
Baseline	nearest	$1.885 * 10^5 \pm 592.820$	7.7 ± 5.01	4.84 ± 3.03	MA
Huber loss	Shrinking+STE	$1.884 * 10^5 \pm 590.648$	2.15 ± 1.11	1.08 ± 0.351	MA
Barrier Method	Shrinking+STE	$\mathbf{1.884 * 10^5 \pm 590.210}$	$\mathbf{1.99 \pm 0.97}$	$\mathbf{1.04 \pm 0.385}$	MA

Table 3.4: Non-stationary Experiment Results. MA denotes Matern covariance function. WLS represents Weighted Least Square and KL records the KL divergence. The mean and standard deviation are reported for each metrics and the bold value indicates the best performance.

Huber loss regularisation also demonstrates substantial improvement, reducing the weighted least squares error by approximately 33% and the KL divergence by around 73% compared with the baseline model. Notably, these positive-definite regularisation methods improve not only the mean values of the evaluation metrics but also reduce their standard deviations, indicating improved optimisation stability and reduced sensitivity to stochastic training effects. Similar trends are observed when the Matern covariance function is used, demonstrating that the effectiveness of the proposed regularisation methods is consistent across different covariance function families. This result further supports the robustness and generalisability of the proposed covariance estimation framework.

3.3 Summary

This chapter develops a comprehensive and flexible framework for the estimation of non-stationary covariance functions, with a particular emphasis on ensuring positive definiteness throughout the modelling and optimisation process. The motivation arises from the inherent difficulty in modelling dynamic covariance structures, where both flexibility and structural validity must be simultaneously maintained.

The chapter begins by reviewing a range of objective functions commonly used in covariance estimation, including likelihood-based approaches such as the Kullback–Leibler divergence and log-likelihood under multivariate normality, as well as loss-based formulations such as Frobenius norm loss and weighted least squares. A central contribution of the chapter is the systematic treatment of positive-definiteness constraints in non-stationary covariance modelling. Since unconstrained optimisation may yield invalid covariance matrices, the estimation problem is formulated within a constrained framework. Rather than relying solely on reparameterisation techniques such as the Cholesky decomposition, which may be restrictive in functional settings, the chapter proposes a suite of regularisation-based approaches that directly penalise violations of positive definiteness. Four principal regularisation strategies are discussed and two of them, which are one-sided Huber loss and interior-point (barrier) method, are examined through simulation. One-sided Huber loss applied to the minimum eigenvalue provides a robust and computationally stable mechanism for penalising near-singular or indefinite matrices while methods incorporate a logarithmic determinant penalty, ensuring that the optimisation trajectory remains within the positive-definite cone.

The proposed framework is evaluated through a series of simulation studies covering both dynamic isotropic and static non-stationary covariance settings. Using parametric covariance families such as the powered exponential and Matern functions, the experiments demonstrate that the in-

corporation of appropriate regularisation not only ensures positive definiteness but also improves estimation accuracy and stability. In dynamic settings, the use of smoothing penalties within a generalised additive modelling framework effectively captures temporal variation in covariance structure while preventing overfitting. In static non-stationary scenarios, the proposed regularisation techniques enable reliable recovery of spatially varying covariance patterns.

Overall, the findings of this chapter highlight that enforcing positive definiteness is not merely a technical constraint but a critical component of robust covariance modelling. The integration of flexible functional modelling with carefully designed regularisation provides a practical and effective solution for estimating complex, non-stationary covariance structures in high-dimensional settings, which provides a solid base for the application introduced in the following chapter.

Chapter 4

Application on Net-load Data

This chapter investigates the practical performance of the proposed positive-definite regularised non-stationary covariance modelling framework in a real-world, high-dimensional energy forecasting setting. The primary objective of this application is to model and estimate a time-varying covariance matrix Σ_t governing the joint distribution of regional net-demand residuals, thereby capturing dynamic cross-regional dependence structures that evolve with exogenous covariates. In contrast to approaches that treat spatial units independently, we explicitly focus on the multivariate structure of the response vector and its covariate-dependent dynamics. The empirical study is conducted on half-hourly net-load data from 14 Grid Supply Point (GSP) groups in Great Britain over a five-year period (2014–2018). Electricity net-demand, measured at the transmission–distribution interface, represents the operational load after accounting for embedded generation and thus constitutes a key quantity for system balancing and short-term forecasting.

The choice of this dataset is motivated by several considerations. First, the energy system context inherently exhibits dynamic dependence: regional residual loads co-move due to shared meteorological drivers, transmission constraints, and synchronized demand patterns. Second, the half-hourly frequency over multiple years provides sufficient temporal depth to identify systematic covariance variation across daily and seasonal cycles. Third, prior studies such as Browell and Fasiolo (2021) and Gioia et al. (2022) have demonstrated the richness of this dataset for marginal and structured covariance modelling, yet existing analyses either treat regions independently or impose alternative parameterisations of Σ_t. This creates a well-defined benchmark environment in which the benefits of non-stationary covariance functions with explicit positive-definite regularisation can be rigorously evaluated.

The remainder of this chapter is organised as follows. Section 4.1 introduces the dataset and presents exploratory evidence of dynamic cross-regional dependence. Section 4.2 outlines the proposed non-stationary covariance formulations and their covariate-driven parameterisation. Section 4.3 reports empirical results, including feature selection, regularisation tuning, and predictive performance evaluation. Together, these analyses demonstrate the effectiveness of the proposed framework for modelling high-dimensional, time-varying covariance structures in energy systems.

4.1 Dataset Introduction

To demonstrate the actual application of positive-definite regularisations, examination on real-world data is performed. Positive-definite regularized non-stationary covariance functions are applied on net-demand residual covariance modelling task of 14 UK regions with a 5 years net-load data from 2014 to 2018 in half an hour interval. Electricity Net-demand is the load measured at the interface between transmission and distribution networks. In GB these interfaces are called Grid Supply Points (GSPs) and are grouped into 14 GSP groups(Gioia et al., 2022). We use $X_t \in \mathbb{R}^{14}$ to denote the temporal net load of 14 regions and $X_t = \{X_t | t = 1, 2, \cdots, N\}$ to denote a series of net load vectors at N time point where $X_{t,i,j}$ represent the net load at time i in region j. Chronological features, such as time of day and day of the year, and weather features (temperature, surface solar radiation, total precipitation) derived from the hourly day-ahead weather forecasts produced by the operational ECMWF-HRES model are included. Full covariates used in this experiment are listed in Table 4.1. Gridded weather predictions are summarized via the regional features. Assuming most of the weather features are highly correlated, such as temperatures of adjacent regions, weather features are preprocessed by taking the mean over all regions and the selected features are based on the feature selection process carried out by Browell et al. (2022). To estimate the covariance matrix among residual loads of 14 regions, geo-distance, which is the separation of two regions, need to be defined and is calculated through GIS boundaries provided by the National Energy System Operator.

Browell and Fasiolo (2021) analyse the aforementioned dataset by modelling the conditional distribution of $X_{t,i,j}$ independently for each of the $d = 14$ regions. Their methodology adopts a composite modelling framework in which residuals obtained from a Gaussian generalised additive model are subsequently modelled using linear quantile regression, while tail behavior is captured through a GAMLSS specification based on the generalized Pareto distribution. In this paper, we are interested in modelling the joint distribution of the d-dimensional response variable X_t dynamically under an assumption $X_t \sim \mathcal{N}(\mu_t, \Sigma_t)$. Specifically, we are dedicated to modelling the covariance matrix Σ_t and capture the temporal dynamics. Gioia et al. (2022) propose an additive modelling framework by adopting linear predictors to the Modified Cholesky Decomposition of Σ_t and carry out a boosting method to do covariates auto-selection. But this approach fails to solve the estimated parameter explosion, which indicates the number of estimated parameters grows quadratically when the dimension of covariance matrix increases, and sacrifices part of the statistical interpretability incurred by matrix decomposition. In this paper, we explore the estimation framework of using non-stationary covariance functions to estimate the covariate-dependent covariance matrix in dynamic.

Primary exploration of the dataset is shown below. To better reveal the dynamic pattern of net-load data, visualisation of empirical covariance matrices of different covariate values is present in Figure 4.1. A clear net demand residual covariance discrepancy is visualised for different mean solar radiation levels. Overall, in low solar radiation conditions, which are likely at nights, variance and covariance remain low while in high solar radiant, typically at day time, net load residuals at some regions are significant, such as the upper left part and region 9 and 10. This visualisation indicates that at net-load demand anticipation, covariance structure changes along certain covariates, justifying the need of dynamic covariance modelling. Figure 4.2 visualises the temporal trajectories of 14 regions within a 48 hours. A similar pattern is demonstrated for each of the region, showing 3 peaks at around 7am, 4pm and mid-might. East North Wales presents unique trend before 5am, indicating a low correlation to other regions. Different fluctuating trend among regions demonstrates a dynamic region-related structure of the net-load demand residuals, thereby underscoring

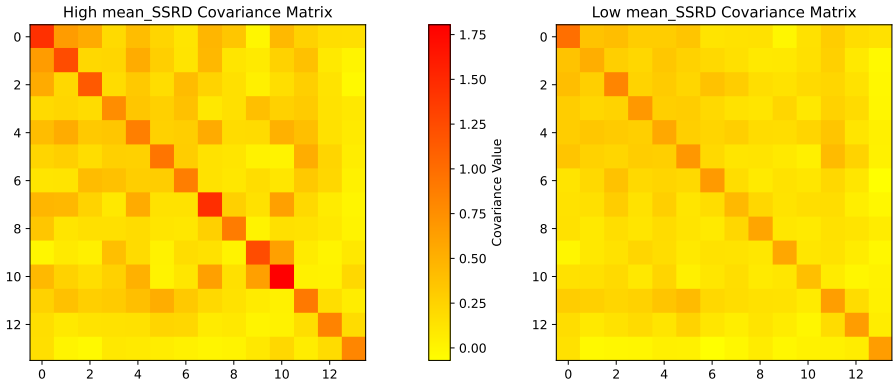

Figure 4.1: Empirical covariance matrices visualisation of different covariate(mean_SSRD) values. High mean_SSRD data is derived from the upper 75% quantile while the lower mean_SSRD data is chosen from lower 25% quantile

Table 4.1: Feature selection results. doy: day of year, moy: month of year, temp: temperature, SSRD: surface solar radiation diemeter, TP: total precipitation, wind 100m: wind speed measured in 100m/s

Selected Features	doy / moy /clock hour	mean temp	mean SSRD	mean TP	mean wind 100m
Value Type	categorical	continuous	continuous	continuous	continuous
Missing Value	0	0	0	0	0
Minimum	1/1/0	268.721	-4.687e-05	-6.810e-07	1.287
Maximum	366/12/23.5	301.447	2.828	0.00193	18.530

the necessity of modelling time-varying regional dependency for accurate real-time prediction.

4.2 Non-stationary covariance functions

Since this is a non-stationary dynamic covariance modelling task, the covariance functions employed must be capable of capturing both local non-stationarity and temporal variability. Consequently, rather than adopting a stationary covariance function, we employ its non-stationary extension. The general form of non-stationary covariance function can be written as

$$\mathrm{Cov}_t(Z_t(s_i), Z_t(s_j)) = C(r_{ij}, \theta(\mathbf{x}_{ij,t})) = C(|s_i - s_j|, \theta(\mathbf{x}_{ij,t})) \tag{4.1}$$

where $\mathbf{x}_{ij,t}$ denotes the the covariates selected for covariance between $Z_t(s_i)$ and $Z_t(s_j)$, and r_{ij} denotes their separation, which is $|s_i - s_j|$. We use Z_i and Z_j to represent $Z_t(s_i)$ and $Z_t(s_j)$ in the following discussion. The effect of separation r_{ij} is isotropic and the covariance function become non-stationary due to the effect of $\theta(\mathbf{x}_{ij,t})$ that not only depend on spatial separation between locations i and j. We examine non-stationary powered exponential and Matern function in this application. In particular, the non-stationary powered exponential covariance function is specified as

$$\mathrm{Cov}_t(X_{t,i}, X_{t,j}) = \sigma^2 \exp(-[\theta(\mathbf{x}_{ij,t})\, r_{ij}]^{\gamma}) \tag{4.2}$$

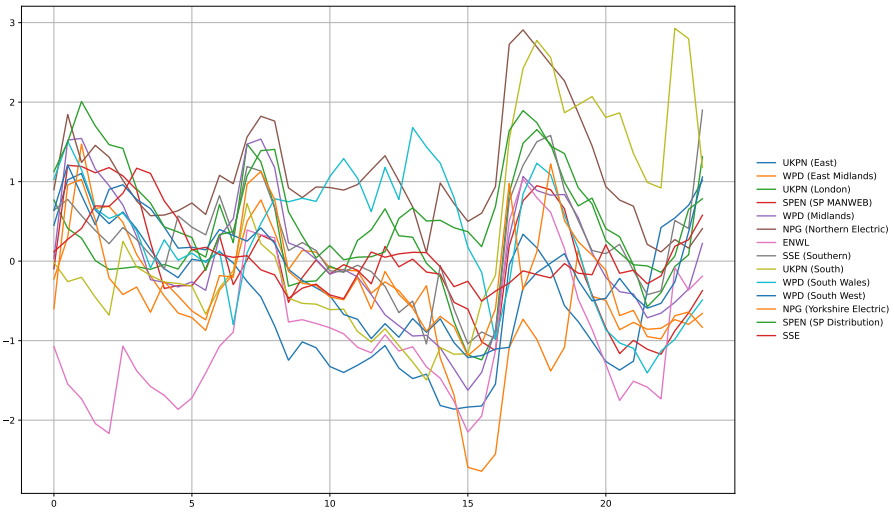

Figure 4.2: Temporal trajectory visualisation of each region. This figure visualise the 14 regions net-load residual trajectories within 24 hours. The x-axis denote the time of day in half-hour interval and the y-axis denote the residual value.

where r_{ij} denotes the geographical distance between the centroids of regions i and j. The variance parameter σ^2, the shape parameter γ, and the coefficients governing $\theta_t = \theta(\mathbf{x}_{ij,t})$ are estimated via optimisation. It is evident that this function become isotropic without $\theta(\mathbf{x}_{ij,t})$ which is able to capture the non-stationary behaviour of the covariance structure. It is modelled using a cubic GAM, given by

$$\theta_t = \theta(\mathbf{x}_{ij,t}) = \mathbf{B}(\mathbf{x}_{ij,t})\boldsymbol{\beta}^\top + \mathbf{b} = \sum_{m=1}^{p}\sum_{n=1}^{B} \beta_{m,n}\,\phi_{m,n}(x_{m,t}) + b \tag{4.3}$$

where $\mathbf{B}(\mathbf{x}_{ij,t})$ denotes the spline basis matrix evaluated at covariates $\mathbf{x}_{ij,t}$, b is the intercept, $\boldsymbol{\beta}$ is the vector of spline coefficients, p is the number of covariates, and B is the number of basis functions per covariate. Similar to the definition of (3.19), scalar $x_{m,t}$ is the m-th covariate value of the covariate vector $\mathbf{x}_{ij,t}$.

In the case of non-stationary Matern covariance function, we consider the Matern function proposed by Paciorek and Schervish (2006), which provides a principled way to construct valid non-stationary covariance functions by allowing spatially varying parameters. This class of models enables the covariance structure to adapt locally while maintaining global positive definiteness. The covariance function is defined as

$$\mathrm{Cov}_t(X_{t,i}, X_{t,j}) = \sigma^2 \frac{\sqrt{\ell_{i,t}\ell_{j,t}}}{\ell_{ij,t}} \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{r_{ij}}{\sqrt{\ell_{ij,t}}}\right)^{\nu} K_\nu\left(\sqrt{2\nu}\frac{r_{ij}}{\sqrt{\ell_{ij,t}}}\right), \quad i \neq j \tag{4.4}$$

where the spatially varying length-scale is defined as $\ell_t = \exp\left(-\theta(\mathbf{x}_{ij,t})\right)$ and $\theta(\cdot)$ is the cubic GAM defined in (4.3). The variance parameter σ^2 and the parameters of $\theta_t = \theta(\mathbf{x}_{ij,t})$ are estimated during optimisation.

The Matern covariance function is particularly flexible due to the smoothness parameter ν, which controls the differentiability of the underlying process. This flexibility, however, introduces sensitivity to the choice of ν, as different values can lead to substantially different model performance. Empirical results (see Figure 4.3) indicate that model performance varies significantly across ν. Therefore, ν is treated as a hyperparameter and selected via validation. In our experiments, the optimal value is found to be approximately $\nu = 0.3$, which achieves the best goodness-of-fit.

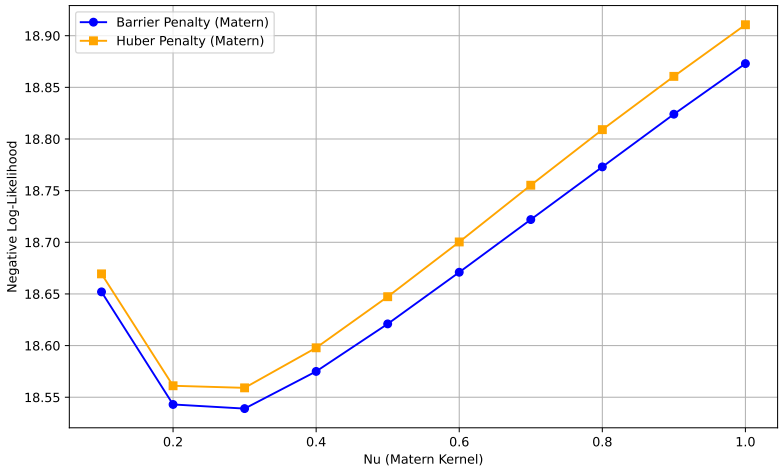

Figure 4.3: ν value evaluation results for non-stationary Matern covariance function. This figure visualise the loss value on different ν settings, demonstrating the sensitivity of this hyperparameter.

4.3 Application Results and Analysis

The dataset comprises five years of net-load residual observations. We partition the data chronologically, using the first four years (66,419 samples) as the training set and the remaining year (16,605 samples) as the test set. To mitigate the inclusion of weakly informative or redundant covariates that may degrade model performance, a forward feature selection procedure is employed. Specifically, the procedure iteratively evaluates each candidate covariate by adding it to the current feature set of the cubic GAM. At each step, the covariate that yields the greatest improvement in model performance is retained. This process continues until all covariates have been examined or no further performance gains are observed. During model fitting, the negative log-likelihood is adopted as the objective function, which is standard in probabilistic modelling as it corresponds to maximising the likelihood of the observed data under the model. Optimisation is performed using the Adam algorithm with a learning rate of 0.001. The number of sampling trajectories is set to 300. To ensure that this choice is sufficient and does not introduce sampling bias, we conduct a sensitivity analysis (see Figure 4.4), which empirically validates the adequacy of this setting. Model performance is evaluated using a combination of proper scoring rules, including negative log-likelihood that measures the accuracy of the estimated covariance matrix given observations

X_t, and metrics that are widely used for assessing the accuracy of multivariate probabilistic forecasts, which is the Energy Score and the 0.5-Variogram Score. The Energy Score evaluates both calibration and sharpness of the predictive distribution, while the p-Variogram Score evaluates the dependence structure and correlations in multivariate outputs. Lower values of these scores indicate better predictive performance. As a baseline, we adopt the empirical covariance matrix estimated from the test set. Furthermore, to isolate the effects of different covariance specifications and positive-definiteness regularisation, we additionally conduct experiments using non-regularised covariance functions.

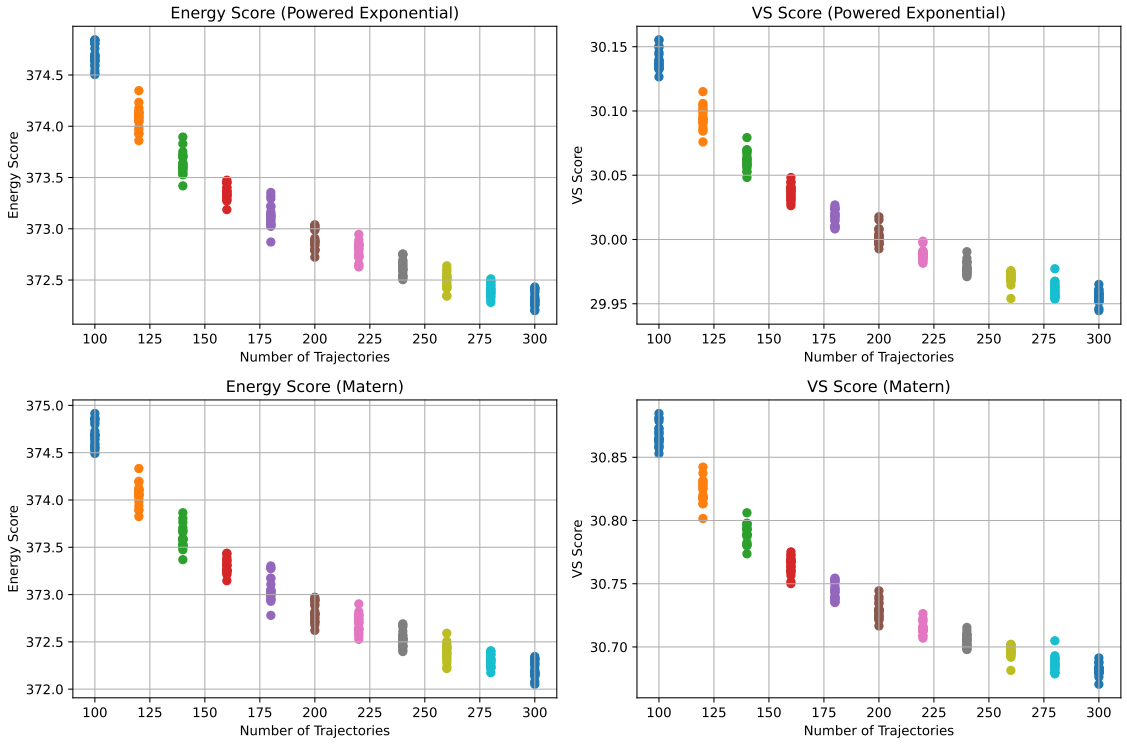

Figure 4.4: Number of trajectories evaluation distribution on both covariance functions. The x-axis denotes the number of trajectories used for sampling and the y-axis denotes the metric value. The first row demonstrate the distributions of Energy Score and 0.5-Variogram Score(VS Score) of powered exponential function while the second row depicts the results from Matern covariance function. The number of evaluation for each number of trajectories is set to be 300.

The experiment results are summarised in Table 4.2, which reports the best performance achieved under each configuration. The non-stationary powered exponential covariance function provides a more accurate representation of the underlying non-stationary covariance structure than the empirical covariance matrix estimated from the test set, thereby supporting the effectiveness of covariance function–based approaches for dynamic covariance modelling in practice. In contrast, the Paciorek–Schervish Matern covariance function exhibits degraded performance in terms of log-likelihood on the test set, despite achieving the best Energy Score under barrier-method reg-

ularisation. A possible interpretation is that its increased flexibility leads to overfitting of sample noise, while the positive-definite regularisation reduces the dispersion of the predictive distribution, which in turn improves the Energy Score (Pinson, 2013).

Compared with the non-regularised setting, both positive-definite regularisation strategies improve model fit across the two covariance functions. In particular, the barrier method consistently yields the strongest performance, achieving an average improvement of approximately 0.2% in the objective function and 1.4% in the 0.5-Variogram Score. The Huber loss–based regularisation also enhances performance, with an improvement of approximately 0.17% in negative log-likelihood. These results indicate that the proposed positive-definite regularisation framework is robust and adaptable across different covariance specifications. From a methodological perspective, incorporating constraints or penalties to enforce positive definiteness is known to stabilise covariance estimation and ensure valid solutions in high-dimensional settings. The observed discrepancy between the Energy Score and the 0.5-Variogram Score is consistent with the simulation results reported in Table 3.1, reflecting their differing sensitivities to marginal distributions and dependence structures, respectively. Our estimation framework also applies a positive-definite correction mechanism (referred to as the "positive-definite head") which corresponds to the shrinking regularisation described in Section 3.1.4. This mechanism is activated only when the estimated covariance matrix violates positive definiteness, as it introduces slight modifications to the covariance structure. In practice, however, this correction was never triggered when initialising with the empirical covariance matrix. Nevertheless, it serves as a safeguard to guarantee positive definiteness in cases where the primary regularisation methods fail. These results demonstrate that positive-definite regularisation not only enforces the structural validity of the estimated covariance matrices but also improves the accuracy and robustness of dynamic covariance estimation.

		Metrics		
Covariance Function	**Regularisation**	**Negative Log-likelihood**	**Energy Score**	**VS-0.5**
None(Empirical)	None	18.460	373.121	30.418
PE	None	18.427	373.133	29.845
PE	Barrier Method	**18.416**	**372.040**	29.932
PE	Huber loss	18.418	372.225	**29.791**
MA	None	18.570	373.143	**30.489**
MA	Barrier Method	**18.539**	371.923	30.659
MA	Huber loss	18.568	**371.878**	30.665

Table 4.2: The first 66,419 samples are used for training and the remaining 16,605 samples are used for testing. Initialisations are approximate empirical covariance matrix. The bold value indicates the best performance of each metric under each covariance function.

To more clearly assess the effect of the proposed positive-definite regularisation methods, we evaluate the objective function (penalised negative log-likelihood) across a range of regularisation strengths, using the optimally selected covariate subsets reported in Table 4.3. We use the negative loglikelihood evaluated on the test set to empirically select the optimal regularisation strength. Regularisation strengths from 0 to 1.0, in increments of 0.1, are considered, and the corresponding results are presented in Figure 4.5.

For the powered exponential covariance function, the optimal regularisation strength is observed at around 0.3. At this level, the barrier penalty achieves the global minimum of the objective

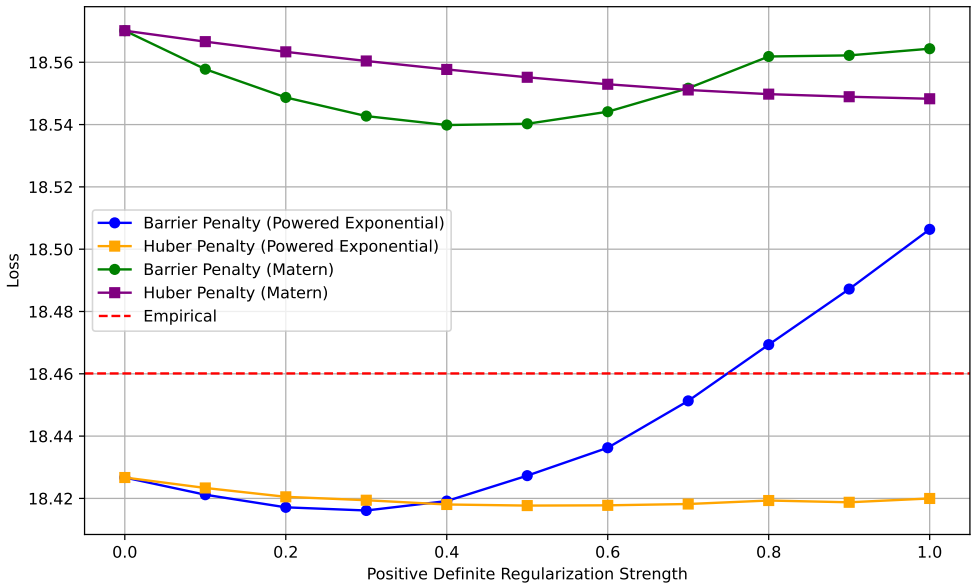

Figure 4.5: Positive-definite regularisation strength selection results. This figure visualise the effect of penalty strength on loss value of each covariance function

function, while the Huber penalty yields a modest improvement relative to the non-regularised case (i.e., regularisation strength equal to 0). Notably, the barrier penalty exhibits a high degree of sensitivity to the choice of regularisation strength: beyond 0.4, the objective value increases sharply, indicating potential over-penalisation and the overall trend of barrier penalty becomes a long-tailed U-shaped trajectory. In contrast, the Huber penalty demonstrates a more stable behaviour, with minor fluctuations around its minimum at the strength of 0.5. A similar pattern is observed for the Matern covariance function. The barrier penalty produces a pronounced U-shaped curve, attaining its minimum at a regularisation strength of approximately 0.4. Meanwhile, the Huber penalty leads to a gradual decrease in the objective value, approaching convergence as the strength increases towards 1.0. Although the full U-shaped behaviour is not observed within the evaluated range for the Huber penalty, the trend suggests that the minimum is approached asymptotically.

Overall, the barrier penalty consistently achieves superior performance when appropriately tuned, but its effectiveness is highly sensitive to the choice of regularisation strength due to the boundary design. By contrast, the Huber penalty provides more stable and robust performance across a wider range of strengths, albeit with slightly suboptimal results in comparison. These findings indicate that careful calibration of the regularisation strength remains an indispensable step in practical applications.

The feature selection results are summarised in Table 4.3. It is evident that for the powered exponential covariance function, nearly all candidate covariates are selected in the cubic GAM used to estimate $\theta(\cdot)$. In contrast, for the Matérn covariance function, only calendar-related covariates and averaged temperature are retained in the cubic GAM for estimating the range parameter ρ, as

the inclusion of additional covariates does not yield further improvements in model performance. It is interesting that all the best fitted Matern covariance functions converge to the same parameter settings with more covariates included, resulting in comparable objective values on the test set. This observation is consistent with the principle of feature selection in generalized additive models, where adding redundant or weakly informative covariates often provides negligible gains in out-of-sample performance. Among all the covariates, the day of year (`doy`) is consistently selected across all models, indicating that the covariance structure of the net-load residuals exhibits strong temporal variability. In addition, the averaged total precipitation is frequently selected at early stages of the forward selection procedure, suggesting that it contributes the most substantial marginal improvement among the candidate covariates. This aligns with the general mechanism of forward feature selection, which incrementally incorporates covariates that maximise model performance improvements at each step. To further investigate the how the covariance is affected by the individual covariate, we visualise the variation in estimated covariance values with respect to each covariate, as shown in Figure 4.6. These results are derived from the fitted powered exponential covariance function. Distinct patterns are observed for several covariates, including day of year (`doy`), clock hour (`clock hour`), averaged total precipitation (`mean TP`), averaged solar radiation (`mean SSRD`), averaged wind speed at 100m (`mean wind 100m`), and averaged temperature (`mean temp`). For each covariate, the regions corresponding to the top five highest variance levels are highlighted and illustrated, providing further insight into the heterogeneous effects of these variables on the covariance structure.

Table 4.3: Feature selection results. doy: day of year, moy: month of year, temp: temperature, SSRD: surface solar radiation diameter, TP: total precipitation, wind 100m: wind speed measured at 100m height

	Powered Exponential		Matern	
Covariate	Barrier	Huber	Barrier	Huber
doy	✓	✓	✓	✓
moy			✓	✓
clock hour	✓	✓		
mean temp	✓	✓		
mean SSRD	✓	✓		
mean TP	✓	✓		✓
mean wind 100m	✓	✓		

A prominent feature across all panels is that the covariance structure exhibits systematic and covariate-dependent variability, rather than remaining stationary. Specifically, both the magnitude and the shape of the estimated covariance trajectories vary substantially across covariates, indicating that inter-regional dependence is strongly modulated by exogenous drivers such as seasonal and daily environmental or operational cycles. Notably, residual covariance of certain regions recur across multiple panels, such as covariance between East Midlands and Southern, suggesting persistent inter-regional relationships that are nonetheless modulated in strength by different environmental drivers. The detailed analysis of each panel is described below.

The day of year panel reveals a clear seasonal pattern for all visualised covariance, exhibiting smooth, gradual variations over the annual cycle. The trajectories exhibit a broad bell-shaped

pattern in which the residual covariance tends to peak in the middle of the year and subsequently decline towards the end of the year. This pattern aligns with the typical seasonal dynamics observed in electricity demand and system behaviour, where aggregated load or residual load series often display statistically significant monthly and seasonal differences due to climatic, socio-economic, and behavioural factors that influence consumption patterns throughout the year. In our context, the elevated covariance during late spring and summer suggests higher synchronisation in net-load variability, which may reflect consistent concurrently strong influences across regions (e.g., similar responses to weather or demand patterns). In contrast, the clock hour panel highlights pronounced intra-day variability, with covariance values fluctuating sharply over the 24-hour cycle. Specifically, covariances attain local maxima between midnight and approximately 5am, corresponding to periods when residential and commercial activity is minimal, and a local minimum around mid-afternoon (approximately 3pm), which aligns to the pattern observed in previous work (Gioia et al., 2024) using matrix decomposition. This intra-daily pattern indicates that net-load residual behaviour is more homogeneous across regions late at night, while in the afternoon the load profiles diverge, leading to reduced covariance. The particularly dynamic diurnal behaviour observed in the covariance between the East Midlands and Southern regions suggests strong diurnal coupling driven by short-term operational or demand factors that introduce high-frequency variation in dependence structure. Together, these patterns demonstrate that the covariance structure of net-load residuals is shaped by both seasonal and daily cycles, with each covariate contributing distinct patterns of variability and reflecting their respective influences on the underlying processes.

Both the SSRD (solar radiation) panel and the total precipitation panel exhibit nonlinear, U-shaped relationships with the estimated covariance between regions. Specifically, for the visualised regional covariance pairs, covariance decreases at intermediate levels of solar radiation and precipitation and increases at both low and high extremes of these covariates. Such nonlinear behaviour is plausible given the complex and often nonmonotonic relationship between weather conditions and electricity system dynamics: clear or completely overcast conditions lead to low variability while partial cloud cover leads to variation and lower correlation. This is highly related to the solar power where extreme levels of solar radiation leads to adequate or absent power generation, which thereby affects the power consumption. In addition, solar radiation and electricity demand can correlate strongly during peak daylight hours, while at night solar radiation is absent, leading to distinct patterns in load and residual structures. Overall, weather-load relationships can be modulated by human activity patterns and power generation beyond simple linear dependence. The observed radiation pattern aligns reasonably with combined diurnal and seasonal cycles, where low SSRD corresponds to nighttime or winter periods and high SSRD corresponds to daytime or summer conditions with strong insolation. Under such regimes, electricity consumption tends to be lower in late night hours and higher during hot summer days, when cooling demand increases, thereby generating higher residual covariance across regions. In this context, periods of intermediate radiation levels may coincide with transitional times of day or variable meteorological conditions that dampen synchronous demand patterns.

The temperature panel reveals asymmetric effects on covariance of different regions. Covariance involved with Southern region (SSE) demonstrate a bell shape, reaching the peak at around 280 Kelvin. The covariance between SSE and East Midlands, and covariance of SSE and South West, both show a gradual decrease beyond this temperature, whereas the covariance between East Midlands and South West follows a similar trajectory with lower overall variance. In contrast, the covariance between SP distribution and the Southern region exhibits a distinct U-shaped pattern, reaching the minimal at approximately 285. These results indicate that temperature in-

fluences inter-regional dependencies in distinct ways, producing both scale and trend heterogeneity across different geographic pairs. Such differential sensitivity is consistent with empirical findings that temperature effects on electricity demand and residual structures can be region-specific and nonlinear.

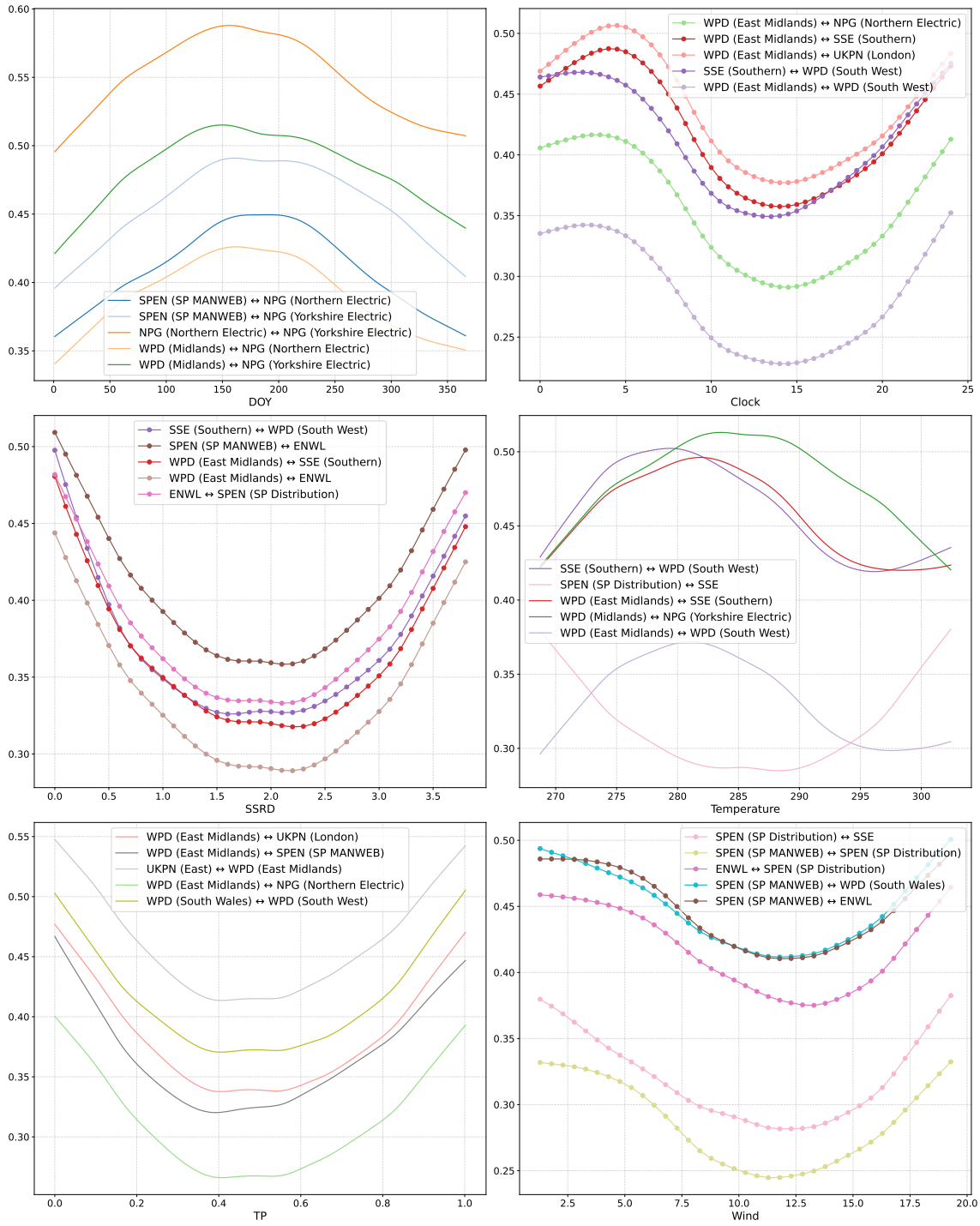

Figure 4.6: Covariance pattern variation with respect to individual covariates. This figure illustrates how the covariance estimated by a non-stationary powered exponential covariance function, varies as each covariate changes. Each subplot depicts the top-5 covariance corresponding to a specific covariate (indicated by the horizontal axis) and the y-axis denotes the covariance value. All remaining covariates are held fixed at zero during visualisation. The x-axis of the temperature panel denotes the Kelvin value of the temperature.

The wind panel demonstrates structured and region-specific variability, particularly evident in covariances involving SPEN and northern regions. Several covariance trajectories show monotonic or near-monotonic trends, indicating that increasing wind speed systematically alters inter-regional dependence. This reflects the role of wind as a spatially correlated driver, especially in regions with significant wind generation capacity, leading to coordinated fluctuations in residuals. Wind speed affects renewable generation and the residual load profile across regions through shared meteorological conditions, which can introduce coordinated fluctuations in net-load residuals. All trends shown in the wind panel display a minimum at around 12.5 m/s, suggesting that electricity consumption is more highly correlated when wind speed is at extreme levels—either very low or very high—relative to moderate winds. One plausible behavioural explanation is that embedded power generated by each wind farm behave in a similar pattern in both high or low wind speed situations, which are sufficient and meagre, and therefore increase the spatial correlation. While at intermediate level, the generated power is more variable and dominated by local effect, which shows low regional correlation. Another reason might be that extreme wind conditions can affect residential behaviour (e.g., leading people to stay indoors) and operational conditions (e.g., changing renewable generation patterns), thereby increasing synchronicity in load residuals. In contrast, at moderate wind speeds, behaviours and generation patterns may be more heterogeneous, leading to lower covariance in regional residuals.

In general, the fitted powered exponential covariance function manage to reveal the effect of different covariates on the covariance value which gives us more insight on electricity consumption pattern over factors. This also justify that non-stationary covariance functions is a feasible method for modelling dynamic non-stationary covariance structure with covariate included under our proposed positive-definite estimation framework and the fitted covariance functions are highly interpretable.

4.4 Summary

In this chapter, we present an empirical evaluation of the proposed positive-definite regularised modelling framework with non-stationary covariance functions in the context of high-dimensional net-load residual dynamic covariance modelling task. Using half-hourly data from 14 regions in Great Britain over a five-year period, the study has demonstrated the importance of modelling dynamic, covariate-dependent cross-regional dependence structures rather than treating regions independently. Exploratory analyses revealed clear evidence of non-stationarity in the covariance structure, driven by various covariates, thereby motivating the adoption of flexible non-stationary covariance functions. Powered exponential and Matern covariance functions are extended to incorporate covariate-driven non-stationarity via a cubic GAM predictor. Empirical results shows that under positive-definite regularisations, the non-stationary powered exponential function provides a robust and accurate representation of the evolving covariance structure, outperforming empirical benchmarks and alternative specifications in likelihood-based evaluation. While the Matern function offers greater flexibility, it is found to be more sensitive to hyperparameter tuning and still, achieves the best performance on Energy Score that evaluates the forecasted trajectories.

A central contribution of this chapter lies in the detailed analysis of how the covariance of specific region pairs varies with respect to individual covariates. By isolating the effect of each covariate, the results reveal that inter-regional dependence exhibits systematic, nonlinear, and highly heterogeneous responses to exogenous drivers. In particular, strong seasonal patterns are observed with respect to day-of-year, where covariance follows a smooth, bell-shaped trajectory peaking at

the middle of the year, indicating enhanced synchronisation during periods of similar climatic conditions. Intra-day variation, captured by clock hour, shows pronounced diurnal dynamics, with higher covariance during night-time periods and reduced dependence in the afternoon, reflecting changing demand heterogeneity throughout the day. Weather-related covariates further demonstrate complex nonlinear effects. Power generation related covariate exhibits a similar pattern. Both solar radiation and wind power show stronger covariance under high or low conditions, which indicates that the minimum present in the intermediate level is caused by noise of embedded power generation. Patterns of covariates like total precipitation and temperature reveal the consumption behaviour of resident in various weather conditions. These findings provide strong evidence that the covariance structure is not only time-varying but also highly interpretable when modelled as a function of individual covariates. The proposed framework therefore offers meaningful insights into the underlying drivers of inter-regional dependence, beyond purely predictive performance.

In addition, the evaluation of positive-definite regularisation methods demonstrates that both barrier and Huber-based penalties improve model stability and predictive accuracy consistently in different covariance functions. The barrier method achieves the best performance when carefully tuned, despite with higher sensitivity to regularisation strength, whereas the Huber approach provides more robust and stable behaviour across a wider parameter range. These results highlight the practical importance of enforcing positive definiteness in high-dimensional covariance estimation.

Overall, the results provide strong empirical support for the proposed framework as an effective and interpretable approach for modelling time-varying covariance structures in energy systems. By combining non-stationary covariance functions with positive-definite regularisation and detailed covariate-level interpretability, the framework offers a principled and practically valuable methodology for capturing complex dynamic dependencies in multivariate forecasting problems.

Chapter 5

Conclusions and Future Direction

This chapter summarises all the content introduced in this thesis including problem's background, proposed methods and experiments results.

5.1 Covariance Modelling Background

Covariance is established as a fundamental construct for capturing dependence in multivariate systems, and the discussion highlights the growing necessity of modelling time-varying, non-stationary covariance structures in practical applications. Existing modelling, approaches including matrix-decomposition-based and covariance-function-based methods, reveals an inherent trade-off between modelling flexibility, interpretability, and parameter scalability. Covariance function manage to describe covariance structure with parsimonious parameters but struggle to maintain positive-definite output in non-stationary dynamic covariance modelling. Cholesky matrix decomposition in contrast, parameterise the whole matrix with unconstrained factors but suffers from parameter explosion and poor statistical interpretability. While various tools, such as shrinkage and Bayesian estimators, support covariance estimation, they often struggle to simultaneously address key issues such as computational tractability, robustness, and the enforcement of positive definiteness, particularly in dynamic contexts. Furthermore, extensions to time-varying frameworks, such as Generalised Additive Covariance models and conditional heteroscedastic processes, introduce additional complexity in balancing adaptability with structural validity. These limitations underscore the central challenge of constructing flexible yet well-constrained covariance models, thereby motivating the need for a principled framework that can accommodate non-stationarity while ensuring stability and positive definiteness in high-dimensional applications.

5.2 Proposed Estimated Framework

With the theoretical foundations and methodological review presented in Chapters 1 and 2, we develop the core contribution of the thesis: a flexible estimation framework for non-stationary covariance functions that explicitly addresses the dual challenges of modelling adaptability and structural validity. It formulates covariance estimation as a constrained optimisation problem and introduces a suite of regularisation strategies—including eigenvalue-based penalties such as

the one-sided Huber loss, log-determinant barrier methods, Lagrangian approaches, and shrinkage techniques—to ensure positive definiteness throughout the estimation process while preserving functional flexibility. These methodological developments are integrated within an additive modelling framework, enabling covariance parameters to vary smoothly with covariates in dynamic settings. Through comprehensive simulation studies encompassing both dynamic isotropic and static non-stationary scenarios, we demonstrate that the proposed framework achieves more stable, accurate, and structurally valid covariance estimates than the compared baseline, with appropriate regularisation playing a critical role in balancing bias–variance trade-offs and preventing numerical instability. It therefore provides the methodological bridge between the conceptual motivations of earlier chapters and the empirical application that follows, where the proposed framework is applied to real-world data using non-stationary powered exponential and Matern covariance functions to model dynamic dependence structures in UK regional electricity residuals.

We finally provides a real-world validation of the proposed positive-definite regularised non-stationary covariance modelling framework by applying it to dataset from energy sector. We perform dynamic covariance modelling of the net-load demand across regions given the dataset, which is a fundamental part of probabilistic net-load forecasting. Building directly on the estimation methodology developed in Chapter 3, the framework is implemented using non-stationary powered exponential and Matern covariance functions, with covariate-dependent structures modelled through additive components to capture temporal and environmental variability. The empirical analysis demonstrates that dynamic cross-regional dependence is strongly influenced by exogenous drivers such as seasonal cycles, diurnal patterns, and weather conditions, thereby confirming the necessity of non-stationary modelling. In terms of predictive performance, the results show that covariance function–based approaches outperform empirical covariance estimates, while the incorporation of positive-definite regularisation—particularly through barrier and Huber-type penalties—consistently further improves both model stability and accuracy. Furthermore, the study highlights the interpretability of the proposed framework, as the estimated covariance structures reveal meaningful and heterogeneous relationships across regions and covariates. Overall, this real-world application demonstrates the practical effectiveness, robustness, and applicability of the proposed methodology in a complex, real-world setting, thereby substantiating its value as a general approach for dynamic covariance modelling.

5.3 Potential future directions

A natural direction for future research lies in the further development and systematic evaluation of positive-definite regularisation strategies within covariance modelling. While this thesis has demonstrated the effectiveness of approaches such as barrier methods, Huber-type penalties, and shrinkage-based corrections in ensuring structural validity and improving estimation stability, a more comprehensive investigation into alternative regularisation schemes is warranted. In particular, exploring adaptive or data-driven regularisation mechanisms, as well as hybrid approaches that combine multiple penalties, may yield improved robustness across varying data regimes. Moreover, a deeper analysis of the interaction between regularisation strength, optimisation dynamics, and model bias–variance trade-offs would provide valuable insights into the practical tuning of these methods, especially in high-dimensional and non-stationary settings.

Another promising avenue concerns the extension of modelling strategies for the parameters of covariance functions beyond the generalised additive modelling (GAM) framework adopted in this thesis. Although GAMs offer a flexible and interpretable approach for capturing smooth,

covariate-dependent variation, they may be limited in representing complex nonlinear interactions or high-dimensional feature spaces. Future work could therefore investigate alternative modelling paradigms, such as kernel-based methods, Gaussian process priors, or deep learning architectures, to parameterise covariance structures more expressively. Such extensions may enhance the ability to capture intricate dependence patterns, particularly in applications characterised by heterogeneous or highly nonlinear dynamics, while raising important questions regarding interpretability, computational scalability, and the preservation of positive definiteness under more complex parameterisations.

The theoretical properties of the proposed estimation framework remain an important area for further study. While the empirical results in both simulated and real-world settings demonstrate strong performance, a rigorous asymptotic analysis would provide a deeper understanding of estimator behaviour under increasing sample size and dimensionality. In particular, deriving consistency, convergence rates, and asymptotic distributions of the estimated covariance parameters would establish a formal statistical foundation for the methodology. Additionally, investigating how model performance scales with respect to the number of observations, the dimensionality of the covariance matrix, and the choice of covariance function would offer practical guidance for model selection and deployment. Such theoretical insights would complement the empirical findings of this thesis and contribute to the broader development of principled, scalable approaches for non-stationary covariance modelling.

Bibliography

J. Browell, C. Gilbert, and M. Fasiolo, "Covariance structures for high-dimensional energy forecasting," *Electric Power Systems Research*, vol. 211, 10 2022.

T. Gneiting, M. Genton, and P. Guttorp, "Geostatistical Space-Time Models, Stationarity, Separability, and Full Symmetry," 2006.

D. Higdon, J. Swall, and J. Kern, "Non-Stationary Spatial Modeling," in *Bayesian Statistics 6*, 2023.

W. H. Bonat and B. Jørgensen, "Multivariate covariance generalized linear models," *Journal of the Royal Statistical Society. Series C: Applied Statistics*, vol. 65, no. 5, 2016.

V. Gioia, M. Fasiolo, J. Browell, and R. Bellio, "Additive Covariance Matrix Models: Modelling Regional Electricity Net-Demand in Great Britain," 11 2022. [Online]. Available: http://arxiv.org/abs/2211.07451

J. C. Pinheiro and D. M. Bates, "Unconstrained parametrizations for variance-covariance matrices," Tech. Rep., 1996.

I. T. Jolliffe, "Principal Component Analysis, Second Edition," *Encyclopedia of Statistics in Behavioral Science*, vol. 30, no. 3, 2002.

I. Miller and A. Papoulis, "Probability, Random Variables, and Stochastic Processes," *Technometrics*, vol. 8, no. 2, 1966.

T. Hong, P. Pinson, S. Fan, H. Zareipour, A. Troccoli, and R. J. Hyndman, "Probabilistic energy forecasting: Global Energy Forecasting Competition 2014 and beyond," pp. 896–913, 2016.

T. Hong and S. Fan, "Probabilistic electric load forecasting: A tutorial review," *International Journal of Forecasting*, vol. 32, no. 3, pp. 914–938, 2016.

V. Gioia, M. Fasiolo, and R. Bellio, "A comparison of unconstrained parameterisations for additive mean and covariance matrix modelling."

G. L. Colclough, M. W. Woolrich, S. J. Harrison, P. A. Rojas López, P. A. Valdes-Sosa, and S. M. Smith, "Multi-subject hierarchical inverse covariance modelling improves estimation of functional brain networks," *NeuroImage*, vol. 178, 2018.

Z. Chen and C. Leng, "Dynamic Covariance Models," *Journal of the American Statistical Association*, vol. 111, no. 515, 2016.

O. Ledoit and M. Wolf, "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection," *Journal of Empirical Finance*, vol. 10, no. 5, 2003.

——, "Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets goldilocks," *Review of Financial Studies*, vol. 30, no. 12, 2017.

L. Yu, R. Zha, D. Stafylas, K. He, and J. Liu, "Dependences and volatility spillovers between the oil and stock markets: New evidence from the copula and VAR-BEKK-GARCH models," *International Review of Financial Analysis*, vol. 68, 2020.

J. Tastu, P. Pinson, and H. Madsen, "Space-time trajectories of wind power generation: Parametrized precision matrices under a Gaussian copula approach," in *Lecture Notes in Statistics*, vol. 217, 2015.

A. M. Latimer, S. Wu, A. E. Gelfand, and J. A. Silander, "Building statistical models to analyze species distributions," in *Ecological Applications*, vol. 16, no. 1, 2006.

M. B. Hooten, N. T. Hobbs, and A. M. Ellison, "A guide to Bayesian model selection for ecologists," *Ecological Monographs*, vol. 85, no. 1, 2015.

T. Subba Rao, "Statistics for Spatio-Temporal Data," *Journal of Time Series Analysis*, vol. 33, no. 4, 2012.

T. Hengl, J. M. De Jesus, G. B. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel, and B. Kempen, "SoilGrids250m: Global gridded soil information based on machine learning," *PLoS ONE*, vol. 12, no. 2, 2017.

S. N. Wood, "Statistical inference for noisy nonlinear ecological dynamic systems," *Nature*, vol. 466, no. 7310, 2010.

T. J. Hefley, K. M. Broms, B. M. Brost, F. E. Buderman, S. L. Kay, H. R. Scharf, J. R. Tipton, P. J. Williams, and M. B. Hooten, "The basis function approach for modeling autocorrelation in ecological data," *Ecology*, vol. 98, no. 3, 2017.

C. D. Prashad, "State-space modelling for infectious disease surveillance data: Dynamic regression and covariance analysis," *Infectious Disease Modelling*, vol. 10, no. 2, pp. 591–627, 6 2025.

A. Cabada, "Green's Functions in the Theory of Ordinary Differential Equations," 2014.

M. Pourahmadi, "Covariance estimation: The GLM and regularization perspectives," *Statistical Science*, vol. 26, no. 3, pp. 369–387, 2011.

J. Franklin, "The elements of statistical learning: data mining, inference and prediction," 2005.

J. Schäfer and K. Strimmer, "A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, 2005.

O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, vol. 88, no. 2, 2004.

J. K. Kruschke, "Bayesian data analysis," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 1, no. 5, 2010.

T. Hastie and R. Tibshirani, "Generalized additive models (with discussion)," *Statistical Science*, vol. 1, 1986.

J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 7 2008.

W. James and C. Stein, "Estimation with Quadratic Loss," 1992.

J. Bröcker and L. A. Smith, "Scoring probabilistic forecasts: The importance of being proper," *Weather and Forecasting*, vol. 22, no. 2, 2007.

S. S. Boyd and L. Vandenberghe, *Convex optimizationBoyd, S. S., & Vandenberghe, L. (2004). Convex optimization. Optimization Methods and Software (Vol. 25). Cambridge University Press. http://doi.org/10.1080/10556781003625177*, 2004, vol. 25, no. 3.

Y. Nesterov and A. Nemirovskii, *Interior-Point Polynomial Algorithms in Convex Programming*, 1994.

J. Browell and M. Fasiolo, "Probabilistic Forecasting of Regional Net-Load with Conditional Extremes and Gridded NWP," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, 2021.

C. J. Paciorek and M. J. Schervish, "Spatial modelling using a new class of nonstationary covariance functions," *Environmetrics*, vol. 17, no. 5, 2006.

J. Pinson, Pierre; Tastu, "Discrimination ability of the Energy score," Technical University of Denmark, Tech. Rep., 2013. [Online]. Available: https://backend.orbit.dtu.dk/ws/portalfiles/portal/56966842/tr13_15_Pinson_Tastu.pdf

V. Gioia, M. Fasiolo, J. Browell, and R. Bellio, "Additive Covariance Matrix Models: Modeling Regional Electricity Net-Demand in Great Britain," *Journal of the American Statistical Association*, pp. 1–13, 11 2024. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01621459.2024.2412361