Covariance Structures for High-dimensional Energy Forecasting

Jethro Browell
School of Mathematics and Statistics
University of Glasgow, UK

Ciaran Gilbert
Dept. Electronic and Electrical Engineering
University of Strathclyde, UK

Matteo Fasiolo
School of Mathematics
University of Bristol, UK

***Abstract*—Forecasts of various quantities over multiple time periods and/or spatial expanses are required to operate modern power systems. Furthermore, probabilistic forecasts are necessary to facilitate economic decision-making and risk management. This gives rise to the challenge of producing forecasts which capture the dependency between variables, over time, and between locations. The Gaussian Copula has been widely used for multivariate energy forecasts and is scalable because the entire dependency structure is captured by a covariance matrix; estimating this covariance matrix in high dimensional problems remains a research challenge. Here we focus on parameterising this covariance matrix as a step towards more robust estimation and to enable conditioning on explanatory variables. We present a range of parametric structures and estimation strategies suitable for multivariate energy forecasting.**

***Index Terms*—Covariance Estimation; Dependency Modelling; Probabilistic Forecasting; Wind**

I. Introduction

When planning and operating weather-dependent energy systems one must consider patterns of spatial and temporal variation of supply and demand. Will solar generation coincide with high demand for space cooling? Will a wind drought exacerbate high winter demand? Is there sufficient network capacity to transmit renewable energy to demand centers? To answer these questions it is necessary to consider interdependency on the relevant spatial and temporal scales. On operational time scales, this means considering the spatial and temporal structure of forecast uncertainty; or loosely speaking, describing whether the impacts forecast errors are likely compound or alleviate one another. For instance, if a forecast error persists at one wind farm, will we see a similar error *with the same sign* at a neighbouring wind farm as well?

Copulas provide a mathematical framework for modelling dependency between random variables and have been widely applied to multivariate probabilistic energy forecasting [1]–[7]. The Gaussian copula in particular lends itself to this task in high-dimensional settings involving 10s or 100s of dimensions, which demands calibrated density forecasts as margins of the copula to be able to accurately estimate covariance parameters from data. Even with calibrated density forecast the parameters of the Gaussian copula can be challenging to estimate, which is the subject of this paper. The empirical covariance matrix can be calculated from training data but in the high-dimensional setting may be close to singular with finite training data. Several parametrisations have been proposed for wind power forecasting based on covariance functions, which guarantee positive-definiteness of the resulting covariance matrix, or parametric precision matrices [4]–[6]. However these are generally limited to isotropic structures (covariance depends on separation only, not specific time/location) and do not consider possible dynamics arising from dependence on time-varying covariates. Slowly-varying temporal dependency in wind power forecasts is considered in [1] using recursive estimation of the empirical covariance and motivates parametric modelling to capture faster dynamics and avoid the lag associated with recursive estimation.

Covariance estimation plays an important role in many statistical sciences including genetics, ecology, finance and machine learning more generally. The main challenges arise from the constraint that covariance matrices must be positive-definite, and that the number of parameters grows quadratically with dimension p. Ensuring positive definiteness can be achieved either through constrained optimisation or by adopting an unconstrained covariance matrix re-parametrisation [8]. Such re-parametrisations may be based on matrix decomposition [8] or covariance functions [9]–[12]. Additional properties such as sparsity (of the precision matrix) or smoothness may also be desirable and regularisation may be used to encourage these [13], [14], perhaps the most well-known example being the graphical LASSO [15].

In a static covariance context we have single covariance matrix which, in the Gaussian copula case, can be estimated via the empirical covariance, provided we have enough data to do so reliably. If not, it may be advantageous to impose a structure with fewer parameters in order to obtain a more robust estimate. Furthermore, if we want to let the covariance matrix change with one or more covariates capturing weather or date/time effects, then the number of unknown parameter grows by at least one order of magnitude relative to the static case. In particular, non-parametric modelling (e.g., using splines) of each covariate's effect on each unique element of the covariance matrix leads to an overly complex model that will over-fit to the available data. The solution is making the covariance matrix vary with the covariates only in a limited set of 'directions', for example by modelling only some of its elements or by modelling the covariance matrix via a few global parameters, which are then allowed to vary with the co-

variates. In this work we consider some parametrisations that are meant to enable the latter approach. Re-parametrisation using covariance functions can drastically reduce the number of parameters to be estimated from $p(p+1)/2$ to just a few, provided that a suitable function can be found.

Two limitations the Gaussian copula are is symmetric nature, and that joint extreme events are unlikely, which may not be appropriate for some data or applications [16]. Copula vines, which are a series of linked bivariate copula families, offer a more flexible framework for modelling high dimensional dependency [2], [17]. However, estimation and sampling of the vine structure and bivariate families is computationally intensive even in a static context, making them unattractive for developing dynamic correlation models, which is the focus of this work. Further the results in [2] show that the Gaussian approach with an exponential covariance matrix outperformed the vine copula for two out of three wind farms in terms of the energy score and the p-variogram score. Alternatives copulas include forecasting multivariate regions [18], [19], using Stochastic Differential Equations [20], [21], and a rank-reordering method which preserves spatial-temporal characteristics known as the Schaake shuffle [22], [23]. To the best of the author's knowledge, none of these methods have been demonstrated in very high dimensional energy forecasting problems.

In this work we propose a framework for generalising covariance functions by allowing their parameters to take the form of additive functions of explanatory variables. We refer to these as Generalised Additive Covariance (GAC) models and focus on the choice loss function and evaluation model. We begin with the necessary preliminaries in Section II before introducing GAC models and their estimation in Section III. Before applying the proposed methods we describe a suite of evaluation metrics in Section IV. We proceed with two simulation-based examples where the true covariance matrix used to generate synthetic data is known and consider dynamic isotropic covariance in Section V-A and static non-stationary covariance in V-B; followed by an example using real energy forecasts in Section VI. We conclude with a discussion and suggestions for future developments.

II. Covariance Functions

Spatio-temporal datasets are widespread in environmental sciences, climatology and meteorology, and related areas, including the energy sector which is increasingly weather-dependent. Covariance models play an important role in these fields, and here we focus on functional models widely used in geostatistics when considering a random process $Z(\boldsymbol{s}, t)$ observed at location $\boldsymbol{s}$ and time t. A subtlety of the present setting is that the temporal dimension of interest is the forecast lead-time l for a given issue time or origin t, hence we adopt the notation $Z_t(\boldsymbol{s}, l)$.

A covariance function, C_t, is *stationary* if the covariance

$$\text{cov}(Z_t(\boldsymbol{s}, l), Z_t(\boldsymbol{s}+\boldsymbol{h}, l+u)) = C_t(\boldsymbol{h}, u) \tag{1}$$

depends only on separation $(\boldsymbol{h}, u)$, and *isotropic* if a further condition of invariance to the direction of $\boldsymbol{h}$ and u

$$\text{cov}(Z_t(\boldsymbol{s}, l), Z_t(\boldsymbol{s}+\boldsymbol{h}, l+u)) = C_t(\|\boldsymbol{h}\|, |u|) \tag{2}$$

applies. Stationarity and isotropy may apply to only one of the spatial or temporal components. Whether a process is spatially and/or temporally stationary and isotropic must be determined on a case by case basis. However, a wider range of parametric isotropic covariance functions exists and serve as a good starting point for many applications. Some isotropic covariance functions are listed in Table I. Anisotropy accounting for direction may be included by replacing the r with the Mahalanobis distance between locations, and non-stationary extensions to the Powered Exponential and Matérn covariance functions are described in [11] and [24], respectively. We propose a new flexible approach that encompasses non-stationary and anisotropic functions in the next section.

In the remainder of this paper we focus on covariance functions of a single distance measure only (i.e. spatial or temporal), although the methods are easily extendable to multiple distances (i.e. spatio-temporal). We consider random vectors $\boldsymbol{z}_t$ with elements $i = 1, ..., p$ being realisations of $Z_t(l_i)$ and define the separation matrix R with elements $R_{i,j,} = |l_2 - l_1|$. The dynamic (time-dependent) covariance matrix Σ_t may be formed as

$$\Sigma_t = \begin{pmatrix} C_t(R_{1,1}) & C_t(R_{1,2}) & \dots & C_t(R_{1,p}) \\ C_t(R_{2,1}) & \ddots & & \vdots \\ \vdots & & & \\ C_t(R_{p,1}) & \dots & & C_t(R_{p,p}) \end{pmatrix} . \tag{3}$$

The time index t is dropped if the covariance function under consideration is *static* and does not depend on t.

III. Generalised Additive Covariance Modelling

The classes of covariance functions discussed above provide a framework which we extend to capture the complex covariance structures observed in practice, with a special focus on energy forecasting applications. In particular, let $C(r; \boldsymbol{\xi})$ be a covariance function parametrised by the m-dimensional parameter vector $\boldsymbol{\xi}$ (e.g., $\boldsymbol{\xi} = \{\theta, \sigma, \gamma\}$ for the powered exponential, see Table I) and let $\boldsymbol{x}_t$ be a d-dimensional vector of explanatory variables, which could include t itself. We propose to let each element of $\boldsymbol{\xi}$ vary with $\boldsymbol{x}_t$ via a semi-parametric generalised additive model, which provides much modelling flexibility while retaining interpretability.

A. Formulation

The elements of $\boldsymbol{\xi}$ are modelled via

$$g_j(\xi_j) = \boldsymbol{A}_{j,t}\boldsymbol{\beta}_j + \sum_i f_{j,i}(\boldsymbol{x}_t^{S_{j,i}}), \quad \text{for} \quad j = 1, \dots, m, \tag{4}$$

where $g_j(\cdot)$ is a monotonic function, $\boldsymbol{A}_{j,t}$ is the t-th row of the design matrix $\boldsymbol{A}_j$, $\boldsymbol{\beta}_j$ is a vector of regression coefficients and $S_{j,i} \subset \{1, \dots, d\}$ such that, for instance, if $S_{j,i} = \{2, 4\}$

TABLE I: Some parametric classes of isotropic covariance functions where $C(\boldsymbol{h})$ takes the form $C(\|\boldsymbol{h}\|;\boldsymbol{\xi})$. The Whittle–Matérn covariance is defined in terms of the modified Bessel function of the second kind K_ν.

Class	Function $C(r;\boldsymbol{\xi})$	Parameters $\boldsymbol{\xi}$
Powered Exponential	$\sigma^2 e^{-(\theta r)^\gamma}$	$0<\gamma\leq 2;\ \theta>0;\ \sigma\geq 0$
Whittle–Matérn	$\sigma^2\frac{2^{1-\nu}}{\Gamma(\nu)}(\theta r)K_\nu(\theta r)$	$\nu>0;\ \theta>0;\ \sigma\geq 0$
Cauchy	$\sigma^2(1+(\theta r)^\gamma)^{-\nu}$	$0<\gamma\leq 2;\ \nu>0;\ \theta>0;\sigma\geq 0$
Spherical	$\sigma^2\left(1-\frac{2}{\pi}(\frac{r}{\theta}\sqrt{1-(\frac{r}{\theta})^2}+\sin^{-1}\frac{r}{\theta})\right)$	$c(r)=0$ if $r>\theta;\ \sigma^2\geq 0;\ \theta>0$
Canonic Periodic	$\sigma^2\exp\left(-\frac{2\sin^2(\omega_0 r/2)}{l^2}\right)$	$\sigma^2\geq 0;\ l>0$

then $\boldsymbol{x}_t^{S_{j,i}}$ is a vector including the second and fourth element of $\boldsymbol{x}_t$. Each $f_{j,i}$ is a smooth function of the form

$$f_{j,i}(\boldsymbol{x}_t^{S_{j,i}})=\sum_{k=1}^{K_{j,i}} b_k^{ji}(\boldsymbol{x}_t^{S_{j,i}})\beta_k^{ji}, \tag{5}$$

where b_k^{ji} are spline basis functions of dimension $|S_{j,i}|$, while β_k^{ji} are regression coefficients. Henceforth, we indicate the vector of regression coefficients used to model all the elements of $\boldsymbol{\xi}$ with $\boldsymbol{\beta}$. Note that, while $\boldsymbol{\xi}$ depends both on $\boldsymbol{\beta}$ and $\boldsymbol{x}_t$, in the following we use the simpler notation $\boldsymbol{\xi}_t$.

B. Estimation

The above models may be fitted to data by minimizing an appropriate loss function with respect to $\boldsymbol{\beta}$. In the static case, model (4) contains only an intercept (i.e., $g_j(\xi_j)=\beta_j$) and one might consider a loss that quantifies the difference between the modelled covariance function $\hat{C}(r;\boldsymbol{\xi})$ and the empirical covariance $\hat{C}_{\text{emp}}(r)$. Ordinary least squares is generally sub-optimal for this purpose though because variogram estimates at different lags are heteroscedastic and correlated. Therefore, following [25], [26], we apply Weighted Least Squares

$$L^S_{\text{WLS}}(\boldsymbol{\beta})=\sum_{i\neq j}\left(\frac{\hat{C}(R_{i,j})-\hat{C}(R_{i,j};\boldsymbol{\xi})}{1-\hat{C}_{\text{cor}}(R_{i,j};\boldsymbol{\xi})}\right)^2 \tag{6}$$

in the case where $\boldsymbol{x}_t=\boldsymbol{x}$ does not depend on time t, and where $C_{\text{cor}}(\cdot)$ is the correlation function corresponding to $C(\cdot)$, and recall that $\boldsymbol{\xi}=\boldsymbol{\xi}(\boldsymbol{\beta})$ is a function of $\boldsymbol{\beta}$. But, if $\boldsymbol{\xi}$ does depend on t, it is necessary to express the loss in terms of the corresponding covariance matrix $\Sigma_t=\Sigma(\boldsymbol{\xi}_t)$ as

$$L^D_{\text{WLS}}(\boldsymbol{\beta})=\frac{1}{T}\sum_{t=1}^{T}\sum_{i\neq j}\left(\frac{[\boldsymbol{z}_t\otimes\boldsymbol{z}_t-\hat{\Sigma}(\boldsymbol{\xi}_t)]_{i,j}}{1-[\hat{\Sigma}_{\text{cor}}(\boldsymbol{\xi}_t)]_{i,j}}\right)^2 \quad , \tag{7}$$

where $\boldsymbol{z}_t\otimes\boldsymbol{z}_t$ is the Kronecker product of realisations at time t and $\hat{\Sigma}_{\text{cor}}$ is the correlation matrix corresponding to $\hat{\Sigma}$. However, this approach is correlation-focused as it does not evaluate the fit for variances ($i=j$), so in cases where we cannot assume unit or otherwise known variances, we consider the full weighted least squares

$$L^S_{\text{WLSf}}(\boldsymbol{\beta})=\sum_{i,j}|\hat{C}_{\text{cor}}(R_{i,j};\boldsymbol{\xi})|\left(\hat{C}(R_{i,j})-\hat{C}(R_{i,j};\boldsymbol{\xi})\right)^2 \quad , \tag{8}$$

which may be similarly adapted for the dynamic case.

We may estimate $\boldsymbol{\beta}$ by numerically minimising L^S_{WLS} or L^D_{WLS} (or L^S_{WLSf} or L^D_{WLSf}), noting that L^D_{WLS} is equivalent to L^S_{WLS} in the static case but is more computationally demanding to evaluate. However, this would lead to results that strongly depend on the number of spline basis functions used. To address the issue, we use the penalised objective

$$\hat{\boldsymbol{\beta}}=\underset{\boldsymbol{\beta}}{\text{argmin}}\ L^D_{\text{WLS}}(\boldsymbol{\beta})+\sum_{k=1}^{K}\lambda_k J_k(\boldsymbol{\beta}) \quad , \tag{9}$$

where the second term contains penalties of the form $J_k(\boldsymbol{\beta})=\boldsymbol{\beta}^\intercal\boldsymbol{S}_k\boldsymbol{\beta}$, with $\boldsymbol{S}_k$ being a positive semi-definite matrix, while λ_k are positive tuning parameters. The purpose of the penalties is to control the wiggliness of the smooth effects, the strength of the penalties being controlled by the λ_k's, which we select by cross-validation. See [27, Ch. 5] for details.

IV. Evaluation Framework

A range of scoring rules are available for evaluating multivariate probabilistic forecasts in the setting where the 'true' distribution is unknown and only a single realisation is available for a given predictive distribution. However, since here we focus on the dependency structure in a Gaussian copula framework we may evaluate forecast performance in the Gaussian domain following the probability integral transformation of the observations. By applying conventional multivariate scoring rules in this setting we are essentially calculating 'copula scores' as coined by Ziel and Berk in [28]. Data in the original domain $\boldsymbol{y}$ are transformed element-wise through the corresponding predictive distribution $F_i(\cdot)$, which serves as the margin of the Gaussian copula, and standard Gaussian distribution $\Phi(\cdot)$ to yield $z_i=\Phi^{-1}(\hat{F}(y_i))$ with zero mean and unit variance. The vector $\boldsymbol{z}$ of transformed data is considered a sample from a multivariate Gaussian with covariance matrix Σ, the estimation of which is the focus of this work.

The Energy Score (ES) is a multivariate generalisation of the continuous ranked probability score, and a strictly proper scoring rule [29], [30]. The ES for a single forecast-observation pair is given by

$$L_{\text{ES}}=\frac{1}{J}\sum_{j=1}^{J}\|\boldsymbol{z}-\hat{\boldsymbol{z}}^{(j)}\|_2-\frac{1}{2J^2}\sum_{i=1}^{J}\sum_{j=1}^{J}\|\hat{\boldsymbol{z}}^{(i)}-\hat{\boldsymbol{z}}^{(j)}\|_2 \quad , \tag{10}$$

where $\|.\|_2$ represents the ℓ_2 norm, $\boldsymbol{z}$ is the observation, and $\hat{\boldsymbol{z}}^{(j)}$ is the jth scenario or 'trajectory forecast' sampled taken from the predictive multivariate distribution.

The p-Variogram Score (VS-p) [31] is designed to provide greater discrimination between forecasts with different dependency structures than the ES. For a single issue time it is

$$L_{\text{VS}p} = \sum_{i,j=1}^{J} w_{ij} \left(|z_i - z_j|^p - \frac{1}{K} \sum_{k=1}^{K} |\hat{z}_i^{(k)} - \hat{z}_j^{(k)}|^p \right)^2 \quad , \tag{11}$$

where p is the order of the variogram and $\hat{z}^{(k)}$ is the k^{th} forecast scenario. Note that this score is proper, but not strictly proper, so typically both the ES and VS-p are reported for forecast verification. Also the small relative change between ES skill scores may be sufficient to discriminate between dependency models when coupled with significance tests [28].

The first two scores measure forecast 'errors' while the following two can be motivated by statistical or information theory. In particular, the log or 'ignorance' score [32] for a Gaussian model with zero mean is

$$L_{\text{LS}} = -\log \Phi(\boldsymbol{z}; \hat{\boldsymbol{\mu}}, \hat{\Sigma}) \propto \text{tr}(\hat{\Sigma}^{-1} \boldsymbol{z}\boldsymbol{z}^{\intercal}) + \log\det \hat{\Sigma} \quad . \tag{12}$$

The log score is equivalent to the negative log-likelihood of $\boldsymbol{z}$ under the model, hence it evaluates the ability of the fitted model to generate the observed data. Its expected value is

$$E(L_{\text{LS}}) \propto \text{tr}(\hat{\Sigma}^{-1}\Sigma) + \log\det(\hat{\Sigma}) \quad . \tag{13}$$

By subtracting the constant $\log\det(\Sigma) + p$ from $E(L_{\text{LS}})$ we obtain the Kullback–Leibler divergence $L_{\text{KL}}(\Sigma, \hat{\Sigma})$ or 'relative entropy' between two Gaussians, with mean zero but different covariances. $L_{\text{KL}}(\Sigma, \hat{\Sigma}) \geq 0$ and is a strictly proper scoring rule, that is $L_{\text{KL}}(\Sigma, \hat{\Sigma}) = 0$ if and only if $\hat{\Sigma} = \Sigma$ (and $\hat{\mu} = \mu$, beyond the zero-mean setting). Of course, if z is not Gaussian, then (12) is only proper, because distributional features beyond the mean and covariance are ignored.

To compare model performance, and the significance of any apparent difference in performance, it is useful to define skill scores. Skill scores may be calculated for any metric using

$$\text{Skill} = \frac{M_{\text{ref}} - M}{M_{\text{ref}} - M_{\text{perf}}} \quad , \tag{14}$$

where M is the metric's value for the method being considered, M_{ref} is the metric's value for a reference method, and M_{perf} is the metrics value for the 'perfect' method, usually zero. We will use bootstrap re-sampling of skill scores to determine if apparent differences in forecast performance (i.e. positive or negative skill) are significantly different from the null hypothesis that skill is zero at the 0.05 level.

V. Examples with synthetic data

Here we test the proposed covariance modelling approach using synthetic data generated by sampling from multivariate normal distributions with known covariance matrices. In both cases, we simulate 5000 samples from the 'true' model and attempt to estimate this model using GACs. We then simulate a further 5000 samples to use for out-of-sample testing and evaluation. The log and relative entropy scores are calculated using the estimated covariance matrix $\hat{\Sigma}$ directly, however the energy and variogram scores require trajectory forecasts to be produced. Therefore, for each sample in the test data, we draw trajectories $\hat{z}^{(i)},\ i = 1, ..., 1000$ from $\mathcal{N}(\mathbf{0}, \hat{\Sigma})$ for evaluation.

A. Dynamic isotropic covariance

First we consider the process $Z_t(l)$ with covariance function $C_t(r)$ that generates random vectors $\boldsymbol{z}_t \in \mathbb{R}^6$ with $l = 0, 0.2, ..., 1$. The covariance function is given by

$$C_t(r) = e^{-\theta(x_t)r} \quad , \tag{15}$$

$$\theta(x_t) = \sin(2\pi x_t) + 2, \quad 0 < x < 1 \quad , \tag{16}$$

where the covariate x_t is a realisation of $X_t \sim \mathcal{U}(0, 1)$. We attempt to estimate a covariance model of the form

$$\hat{C}_t(r) = e^{-\hat{\theta}(x_t)r} \quad , \tag{17}$$

$$\hat{\theta}_{\text{GAC-CR}}(x_t) = \beta_0 + f_{\text{cr}}(x) \quad , \tag{18}$$

where $f_{\text{cr}}(\cdot)$ is cubic spline with five basis functions. The model is fitted by minimizing (9), with $\lambda = 5 \times 10^{-5}$.

For comparison, we also consider the empirical covariance matrix of the 5000 training samples, and a GAC model with a simple linear model for θ, specifically

$$\hat{\theta}_{\text{GAC-Linear}}(x_t) = \beta_0 + \beta_1 x_t \quad . \tag{19}$$

The true and GAC covariance structure are dynamic, introduced by their dependency on x_t, but are guaranteed to be positive-definite as the resulting covariance functions are always of the Powered Exponential form for a given realisation of x_t. The estimation of the relationship between θ and x_t is the estimation problem we are addressing. Of course data generated from a real-world process is unlikely to follow an exact, known covariance function, but the purpose of this exercises if to verify that the GAC modelling approach is able to recover something like (16) from samples of $Z_t(l)$.

The true and estimated functions $\theta(x_t)$, $\hat{\theta}_{\text{GAC-Linear}}(x_t)$ and $\hat{\theta}_{\text{GAC-CR}}(x_t)$ are plotted in Fig. 1. The linear fit is unable to replicate the shape of $\theta(x_t)$ but is an improvement on the static empirical covariance, as verified by the evaluation metrics listed in Table II. Visually, GAC-CR is better able to reproduce the shape of $\theta(x_t)$, although even with 5000 samples in the training set, we found a non-zero smoothness penalty to be necessary to achieve a good fit (not shown).

The GAC model is successfully capturing the dynamics of $Z_t(l)$ and has a clear interpretation though the estimated smooth function $\hat{\theta}_{\text{GAC-CR}}(x_t)$. However, in a real-world process is unlikely to be governed by a known parametric form so model specification may be challenging with a large number of candidate covariance functions and additive models for selected parameters, the latter associated with several hyper-parameters (basis functions, smoothness penalties) to tune.

B. Non-stationary static covariance

Motivated by a structure observed in energy forecasting, we consider a non-stationary (and therefore also anisotropic) covariance function of the following form

$$\text{cov}(Z(l_1), Z(l_2)) = C(l_1, l_2) \tag{20}$$

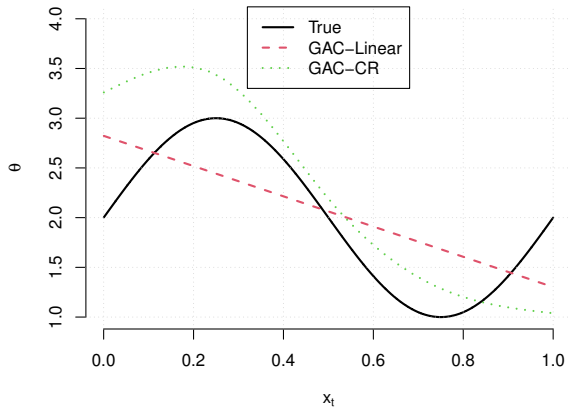


Fig. 1: True function $\theta(x_t)$, simple linear model estimate, and GAC estimate from the dynamic isotropic example in Section V-A.

TABLE II: Results of simulation experiment for example V-A: Isotropic dynamic covariance. Underline indicates that the corresponding skill score relative to the GAC-CR model are not significantly different from zero.

Name	Energy	Log	VS-0.5	VS-1	KL
Static Empirical	1.607	6.993	3.777	11.950	0.292
GAC-Linear	1.606	6.930	3.736	11.810	0.147
GAC-CR	1.605	6.921	3.724	11.770	0.108
True	1.605	6.870	3.697	11.670	0.000

with an exponential covariance function

$$C(l_1, l_2) = \sigma^2 e^{-\theta(s_1,s_2)(|s_1-s_2|)^\gamma}, \quad (21)$$

$$\theta(l_1, l_2) = \frac{5}{1 + l_1 + l_2}, \quad (22)$$

with $\sigma = 1$ and $\gamma = 0.8$. This structure exhibits an larger decay rate for greater values of coordinates l_1 and l_2, and is visualised in Fig. 2. This feature resembles the observations that forecast errors tend to persist for longer the further into the forecast horizon we look.

For the simulation experiment, we consider a 51×51 covariance matrix with l_1 and l_2 taking values $0, 0.02, ..., 1$, illustrated in Fig. 2, and aim to estimate the constant parameters $\sigma = \hat{\sigma}$ and $\gamma = \hat{\gamma}$, and the smooth function $\hat{\theta}(d) = \beta_0 + f_{\text{cr}}(d)$, $d = l_1 + l_2$ with 10 basis functions, using full weighed least squares loss (8) within (9), and with $\lambda = 10^{-4}$. A drawback of this approach is there is no guarantee that the fitted covariance function will produce positive-definite covariance matrices when evaluated at all value of d. Here we apply the algorithm described in [33] to find the nearest covariance matrix, though this is not entirely satisfactory and discussed in Section VII.

For reference, we also evaluate the empirical covariance estimated on the training data, and a stationary exponential covariance function with constant $\theta = \hat{\theta}$.

Evaluation metrics of the resulting models are presented in Table III, and the significance of apparent differences in these

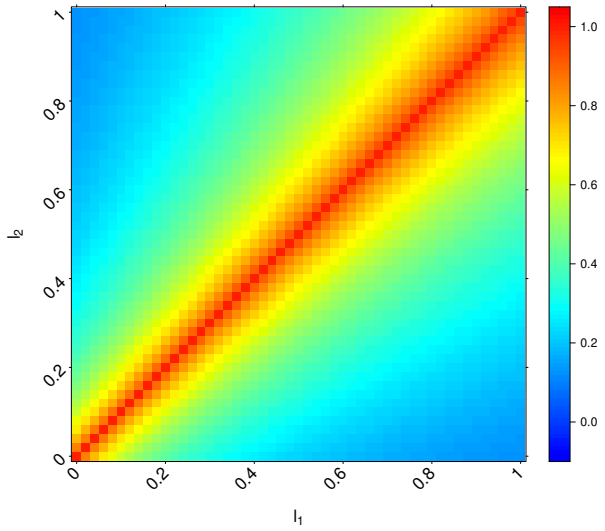


Fig. 2: True covariance matrix considered in the non-stationary example in Section V-B.

TABLE III: Results of simulation experiment for example V-B: Isotropic dynamic covariance. Underline indicates that the corresponding skill score relative to the GAC model are not significantly different from zero.

Name	Energy	Log	VS-0.5	VS-1	KL
Stationary	4.811	28.98	314.7	995.1	3.073
Empirical	4.813	27.97	312.8	989.4	0.269
GAC	4.811	27.90	312.9	989.6	0.140
True	4.811	27.83	312.8	989.1	0.000

scores has been tested using bootstrap re-sampling of skill scores. The resulting GAC model, with constant parameter estimates $\hat{\sigma} = 0.992$ and $\hat{\gamma} = 0.803$ and smooth estimate $\hat{\theta}(d)$, outperforms both the empirical and stationary (constant θ) references models in terms of the Log and relative entropy (KL) scores. Energy and Variogram skill scores, are however not significantly different from zero when comparing the GAC and Empirical covariance models. In fact, the Energy score is not able to discriminate between the performance of any of the models.

VI. Wind Power Case Study

Multivariate wind power 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. Furthermore, with regions now containing 10s or 100s of wind farms, and lead-times from 0 to 120 hours ahead required for operational planning, multiple years of historical forecast data are needed to estimate a positive definite (PD) empirical covariance matrix. That said, practically speaking older years are less relevant as wind farm development is

ongoing, and missing data due to curtailments and so on can pollute historic datasets. In this high dimensional setting, parametrisation of the covariance matrix is essential.

Here we consider the temporal dependency structure of short-term wind power forecasts for the total of Scotland's approximately 10 GW of wind capacity. The dependency is modelled in a Gaussian copula framework where the marginals of the copula are density forecasts of wind power and the temporal dependencies are described by a covariance matrix. We use a non-stationary GAC parametrisation to capture the temporal covariance structures observed in the data.

The case study is based on short-term (0–48h ahead) forecasting of the metered Scottish wind fleet during 2018–2019. Periods where curtailment is over 10% of the estimated total capacity are excluded. Density forecasts for each half-hour period are generated using multiple quantile regression with inputs based on 10m and 100m wind speed forecasts from ECMWF, with parametric estimates for the tails of the distributions [6]. The first 18 months of the dataset are used for model training and tuning via cross-validation, and the last 6 months are used for out-of-sample evaluation.

Fig. 3a shows the empirical temporal dependency structure for the forecasts estimated on the training data, which is non-singular. The correlation matrix has a 'funnel' structure along the diagonal describing how errors tend to persist for longer further into the forecast horizon. Furthermore, the funnel is not monotonically increasing but exhibits some smooth, perhaps periodic, variation along its length and there is also the suggestion of additional off-diagonal structures.

We model the funnel structure using the Powered Exponential correlation function ($\sigma = 1$) initially with constant parameters to serve as a reference, and then using the proposed GAC approach. In the latter, we allow θ to be a smooth function of the distance $d = s_1 + s_2$ along the diagonal

$$\theta = \hat{\theta}_{\text{cr}}(d) = \beta_0 + f_{\text{cr}}(d) \quad , \tag{23}$$

and with a constant parameter $\gamma = \hat{\gamma}$ to be estimated. We model θ flexibly rather than γ in the first instance as it is the simpler effect in the model. Since this is a correlation matrix, the model is estimated using $L^S_{\text{WLS}}(\boldsymbol{\beta})$, and a smoothing parameter of $\lambda = 0.1$ is chosen. As in the example in Section V-B, this non-stationary structure does not guarantee positive definiteness so as before we find the nearest PD matrix to the fitted GAC model where necessary.

The resulting stationary and GAC correlation matrices are plotted in Fig. 3b and 3c, respectively, and clearly highlights how the stationary model is unable to capture the structure observed in the empirical correlation matrix. The GAC model successfully captures the 'wavy funnel' diagonal structure observed in the empirical correlation matrix. Visually, the GAC matrix resembles a smoothed representation of the Empirical, which is desirable, while the stationary matrix appears deficient in comparison. This is verified by the evaluation scores in Table IV.

The three matrices are evaluated using the forecast metrics introduced previously, however, with six months of testing data, only the large improvement in log score is significant at the 0.05 level. The log score for the Empirical correlation matrix is returned as infinity due to the precision limits of the computation, and highlights the challenge of working with high-dimensional probabilistic forecasts.

TABLE IV: Results for temporal wind power forecasting. Underline indicates that the corresponding skill score relative to the GAC model are not significantly different from zero.

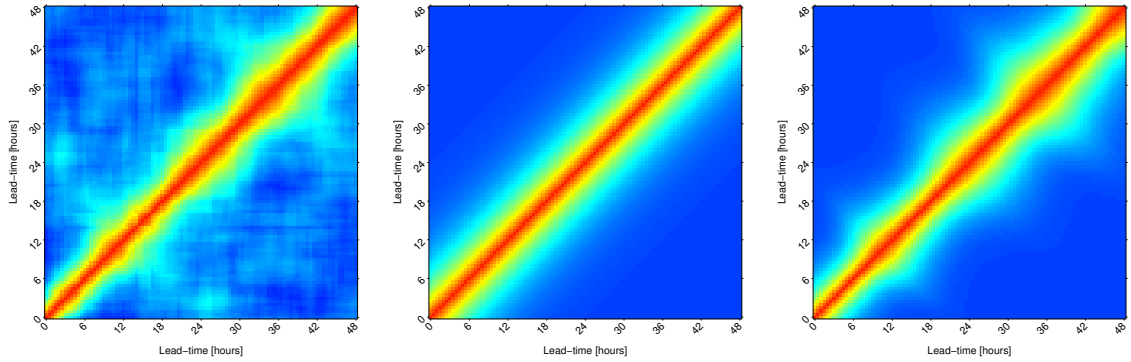
Name	Energy	Log	VS-0.5	VS-1
Empirical	7.139	Inf	1409	5444
Constant	7.142	19.86	1409	5439
GAC	7.137	15.46	1406	5433

VII. Discussion and Conclusions

We have proposed a modelling framework called Generalised Additive Covariance for covariance (or correlation) functions and matrices that depend on explanatory variables, as is the case in energy forecasting and other applications. By modelling the parameters of covariance functions using additive models, it is possible to describe high dimensional covariance matrices with a small number of parameters in an interpretable way. The proposed method has been verified in two synthetic examples of time varying isotropic covariance and static non-stationary covariance, and on a further example using real wind power forecast data. In all cases the proposed method out performs benchmarks including the empirical covariance and conventional parametric covariance functions in terms of the ignorance score, the difference is present but not always statistically significant in terms of other evaluation metrics. However, much work remains to determine the theoretical properties of these models and to improve their estimation.

The use of a parsimonious parametrisation is key to enable modelling of the covariance matrix as a function of explanatory variables such as date/time or meteorological conditions, which are common in energy forecasting. In particular, this approach limits that number of parameters that need to be estimated and allows the user to focus on modelling few, interpretable, parameters as functions of the covariates. For example, if we consider the powered exponential function, it is quite simple to understand the effect of varying each parameter on the resulting covariance structure. Hence, a modeller should be able to use their expertise to develop a sensible additive model for each parameter, and methods for automatic feature selection could be explored in future works.

However, the resulting covariance function is generally not guaranteed to be positive definite. The lack of such a guarantee is problematic from a model-fitting perspective. For instance, likelihood-based fitting of a Gaussian copula requires a PD covariance at each optimization step, otherwise the likelihood corresponding to the proposed parameters is not defined. Assuming that the current parameter vector lead to a PD matrix, the simplest solution is to backtrack from the proposed toward the current parameter until the resulting matrix is



(a) Empirical (b) Constant $\hat{\theta}$ (Stationary) (c) GAC $\hat{\theta}_{\text{cr}}(d)$

Fig. 3: Correlation matrices describing temporal dependency structure of wind power forecasts from 0 to 48 hours-ahead issues at midnight, using the same colour scale as Fig. 2. The forecasts have a visible non-stationary structure. The width of the diagonal ridge indicates how long forecast errors are likely to persist for in time, which grows with lead-time but also appears to depend on the time of day.

positive definite, as done by [4] and [12]. But such a "brute-force" approach is inelegant and might not scale well with the number of parameters. Other fitting criteria, such the variations on least squares used here, do not rely on the proposed matrix being PD at each optimization step, but the final matrix might not be positive definite. Hence, it must be perturbed to produce a PD matrix, although this might disrupt the interpretation of the underlying matrix parametrisation.

A possible solution to this problem is to penalise the smallest eigenvalue in the optimisation routine to maintain positive definiteness, as in [13], although the implications of this on the quality of the resulting fit are unclear. Alternatively, the positive definiteness problem can be solved by adopting a parametrisation under which the resulting matrix is PD for any parameter value. For example, positive definiteness can be guaranteed by modelling the elements of the Cholesky decomposition of a covariance matrix, but see [8] for other unconstrained parametrisations. The issues now are the lack of an intuitive understanding of the covariance matrix's Cholesky decomposition, and that modelling all the element of the decomposition would likely lead to an over-parametrised model.

In summary, parsimonious covariance modelling requires the use of an interpretable parametrisation which enables users to focus on modelling few parameters as functions of the covariates, but fulfilling the PD constraint requires the adoption of an unconstrained, less interpretable, parametrisation. How to fulfill the positive definiteness constraint while retaining interpretability is an open research question, see for example [14] for an overview.

Acknowledgements

Jethro Browell is supported by EPSRC Innovation Fellowship (EP/R023484/1 and EP/R023484/2). The authors thank National Grid ESO, TNEI Services Ltd, and Graeme Hawker for their roles supporting the examples presented in this paper.

Data statement: The code and data required to reproduce the methods and examples presented here are available in the form of an R package, DOI: `10.5281/zenodo.5541782`.

References

[1] P. Pinson, H. Madsen, H. A. Nielsen, G. Papaefthymiou, and B. Klöckl, "From probabilistic forecasts to statistical scenarios of short-term wind power production," *Wind Energy*, vol. 12, no. 1, pp. 51–62, 1 2009.

[2] R. J. Bessa, "On the quality of the Gaussian copula for multi-temporal decision-making problems," *19th Power Systems Computation Conference, PSCC 2016*, 8 2016.

[3] A. Pircalabu, "A Regime-Switching Copula Approach to Modeling Day-Ahead Prices in Coupled Electricity Markets," *SSRN Electronic Journal*, vol. 68, no. Supplement C, pp. 302–283, 2017.

[4] J. Tastu, P. Pinson, and H. Madsen, "Space-Time Trajectories of Wind Power Generation: Parametrized Precision Matrices Under a Gaussian Copula Approach," pp. 296–267, 2015.

[5] J. E. B. Iversen and P. Pinson, "RESGen: Renewable Energy Scenario Generation Platform," in *Proceedings of IEEE PES General Meeting*. IEEE, 2016.

[6] C. Gilbert, "Topics in High-dimensional Energy Forecasting," Ph.D. dissertation, University of Strathclyde, Glasgow, 2021.

[7] D. v. d. Meer, D. Yang, J. Widén, and J. Munkhammar, "Clear-sky index space-time trajectories from probabilistic solar forecasts: Comparing promising copulas," *Journal of Renewable and Sustainable Energy*, vol. 12, no. 2, p. 026102, 4 2020.

[8] J. C. Pinheiro and D. M. Bates, "Unconstrained parametrizations for variance-covariance matrices," *Statistics and Computing 1996 6:3*, vol. 6, no. 3, pp. 289–296, 1996.

[9] D. J. C. MacKay, "Introduction to Gaussian Processes," in *Neural Networks and Machine Learning*, C. M. Bishop, Ed., vol. 168. Springer Verlag, 1998, pp. 166–133.

[10] T. Gneiting, M. G. Genton, and P. Guttorp, "Geostatistical space-time models, stationarity, separability and full symmetry," in *Statistical Methods for Spatio-Temporal Systems*, B. Finkenstaedt, L. Held, and V. Isham, Eds. Statistical Methods for Spatio-Temporal Systems, 2007, pp. 151–175.

[11] D. Higdon, J. Swall, and J. Kern, "Non-stationary spatial modeling," in *Bayesian Statistics 6*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, Eds., 1999, pp. 761–768.

[12] 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, pp. 649–675, 11 2016.

[13] K. Meyer, "Simple Penalties on Maximum-Likelihood Estimates of Genetic Parameters to Reduce Sampling Variation," *Genetics*, vol. 203, no. 4, pp. 1885–1900, 8 2016.

[14] M. Pourahmadi, "Covariance Estimation: The GLM and Regularization Perspectives," *Statistical Science*, vol. 26, no. 3, pp. 369–387, 8 2011.
[15] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 441–432, 2008.
[16] E. Kole, K. Koedijk, and M. Verbeek, "Selecting copulas for risk management," *Journal of Banking & Finance*, vol. 31, no. 8, pp. 2405–2423, 8 2007.
[17] Z. Wang, W. Wang, C. Liu, Z. Wang, and Y. Hou, "Probabilistic Forecast for Multiple Wind Farms Based on Regular Vine Copulas," *IEEE Transactions on Power Systems*, vol. 33, no. 1, pp. 578–589, 1 2018.
[18] F. Golestaneh, P. Pinson, R. Azizipanah-Abarghooee, and H. B. Gooi, "Ellipsoidal Prediction Regions for Multivariate Uncertainty Characterization," *IEEE Transactions on Power Systems*, vol. 33, no. 4, pp. 4519–4530, 7 2018.
[19] F. Golestaneh, P. Pinson, and H. B. Gooi, "Polyhedral Predictive Regions for Power System Applications," *IEEE Transactions on Power Systems*, vol. 34, no. 1, pp. 693–704, 1 2019.
[20] J. K. Møller, M. Zugno, and H. Madsen, "Probabilistic Forecasts of Wind Power Generation by Stochastic Differential Equation Models," *Journal of Forecasting*, vol. 35, no. 3, pp. 189–205, 4 2016.
[21] E. B. Iversen, J. M. Morales, J. K. Møller, P.-J. Trombe, and H. Madsen, "Leveraging stochastic differential equations for probabilistic forecasting of wind power using a dynamic power curve," *Wind Energy*, vol. 20, no. 1, pp. 33–44, 1 2017.
[22] R. P. Worsnop, M. Scheuerer, T. M. Hamill, and J. K. Lundquist, "Generating wind power scenarios for probabilistic ramp event prediction using multivariate statistical post-processing," *Wind Energy Science*, vol. 3, no. 1, pp. 393–371, 2018.
[23] M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, "The Schaake Shuffle: A Method for Reconstructing Space–Time Variability in Forecasted Precipitation and Temperature Fields," *Journal of Hydrometeorology*, vol. 5, no. 1, pp. 243–262, 2 2004.
[24] C. J. Paciorek and M. J. Schervish, "Spatial modelling using a new class of nonstationary covariance functions," *Environmetrics*, vol. 17, no. 5, pp. 483–506, 8 2006.
[25] N. Cressie, "Fitting variogram models by weighted least squares," *Journal of the International Association for Mathematical Geology 1985 17:5*, vol. 17, no. 5, pp. 563–586, 7 1985.
[26] N. A. Cressie, "Statistics for spatial data revised edition," *Statistics for Spatial Data*, pp. 1–900, 4 2015.
[27] S. N. Wood, *Generalized Additive Models: An Introduction with R (2nd edition)*. Chapman and Hall/CRC, 2017.
[28] F. Ziel and K. Berk, "Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules," *arXiv:1910.07325*, 10 2019.
[29] T. Gneiting and A. E. Raftery, "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, vol. 102, no. 477, pp. 378–359, 2007.
[30] T. L. Thorarinsdottir and N. Schuhen, "Verification: Assessment of Calibration and Accuracy," *Statistical Postprocessing of Ensemble Forecasts*, pp. 155–186, 1 2018.
[31] M. Scheuerer and T. M. Hamill, "Variogram-Based Proper Scoring Rules for Probabilistic Forecasts of Multivariate Quantities," *Monthly Weather Review*, vol. 143, no. 4, pp. 1334–1321, 2015.
[32] M. S. Roulston and L. A. Smith, "Evaluating Probabilistic Forecasts Using Information Theory," *Monthly Weather Review*, vol. 130, no. 6, 6 2002.
[33] N. J. Higham, "Computing the nearest correlation matrixa problem from finance," *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 343–329, 2002.