Generalised Wishart Processes

Andrew Gordon Wilson*
Department of Engineering
University of Cambridge, UK

Zoubin Ghahramani
Department of Engineering
University of Cambridge, UK

Abstract

We introduce a new stochastic process called the *generalised Wishart process* (GWP). It is a collection of positive semi-definite random matrices indexed by any arbitrary input variable. We use this process as a prior over dynamic (e.g. time varying) covariance matrices $\Sigma(t)$. The GWP captures a diverse class of covariance dynamics, naturally handles missing data, scales nicely with dimension, has easily interpretable parameters, and can use input variables that include covariates other than time. We describe how to construct the GWP, introduce general procedures for inference and prediction, and show that it outperforms its main competitor, multivariate GARCH, even on financial data that especially suits GARCH.

1 INTRODUCTION

Modelling the dependencies between random variables is fundamental in machine learning and statistics. Covariance matrices provide the simplest measure of dependency, and therefore much attention has been placed on modelling covariance matrices. However, the often implausible assumption of constant variances and covariances can have a significant impact on statistical inferences.

In this paper, we are concerned with modelling the dynamic covariance matrix $\Sigma(t) = \mathrm{cov}[\boldsymbol{y}|t]$ (multivariate volatility), for high dimensional vector valued observations $\boldsymbol{y}(t)$. These models are especially important in econometrics. Brownlees et al. (2009) remark that "The price of essentially every derivative security is affected by swings in volatility." Indeed, Robert Engle and Clive Granger won the 2003 Nobel prize in economics "for methods of analysing economic time series with time-varying volatility". The returns on major equity indices and currency exchanges are thought to have a time changing variance and zero mean, and GARCH (Bollerslev, 1986), a generalisation of Engle's ARCH (Engle, 1982), is arguably unsurpassed at predicting the volatilities of returns on these equity indices and currency exchanges (Poon and Granger, 2005; Hansen and Lunde, 2005; Brownlees et al., 2009). Multivariate volatility models can be used to understand the dynamic correlations (or *co-movement*) between equity indices, and can make better univariate predictions than univariate models. A good estimate of the covariance matrix $\Sigma(t)$ is also necessary for portfolio management. An optimal portfolio allocation $\boldsymbol{w}^*$ is said to maximise the Sharpe ratio (Sharpe, 1966):

$$\frac{\text{Portfolio return}}{\text{Portfolio risk}} = \frac{\boldsymbol{w}^\top \boldsymbol{r}(t)}{\sqrt{\boldsymbol{w}^\top \Sigma(t) \boldsymbol{w}}}\,, \qquad (1)$$

where $\boldsymbol{r}(t)$ are expected returns for each asset and $\Sigma(t)$ is the predicted covariance matrix for these returns. One may also wish to maximise the portfolio return $\boldsymbol{w}^\top \boldsymbol{r}(t)$ for a fixed level of risk: $\sqrt{\boldsymbol{w}^\top \Sigma(t) \boldsymbol{w}} = \lambda$. Multivariate volatility models are also used to understand *contagion*: the transmission of a financial shock from one entity to another (Bae et al., 2003). And generally – in econometrics, machine learning, climate science, or otherwise – it is useful to know input dependent uncertainty, and the dynamic correlations between multiple entities.

Despite their importance, existing multivariate volatility models suffer from tractability issues and a lack of generality. For example, multivariate GARCH (MGARCH) has a number of free parameters that scales with dimension to the fourth power, and interpretation and estimation of these parameters is difficult to impossible (Silvennoinen and Teräsvirta, 2009; Gouriéroux, 1997), given the constraint that $\Sigma(t)$ must be positive definite at all points in time.

* http://mlg.eng.cam.ac.uk/andrew

Thus MGARCH, and alternative multivariate stochastic volatility models, are generally limited to studying processes with fewer than 5 components (Gouriéroux et al., 2009). Recent efforts have led to simpler but less general models, which make assumptions such as constant correlations (Bollerslev, 1990).

We hope to unite machine learning and econometrics in an effort to solve these problems. We introduce a stochastic process, the *generalised Wishart process* (GWP), which we use as a prior over covariance matrices $\Sigma(t)$ at all times t. We call it the *generalised* Wishart process, since it is a generalisation of the first Wishart process defined by Bru (1991).[1] To great acclaim, Bru's Wishart process has recently been used (Gouriéroux et al., 2009) in multivariate stochastic volatility models (Philipov and Glickman, 2006; Harvey et al., 1994). This prior work is limited for several reasons: 1) it cannot scale to greater than 5×5 covariance matrices, 2) it assumes the input variable is a scalar, 3) it is restricted to using an Ornstein-Uhlenbeck (Brownian motion) covariance structure (which means $\Sigma(t+a)$ and $\Sigma(t-a)$ are independent given $\Sigma(t)$, and complex interdependencies cannot be captured), 4) it is autoregressive, and 5) there are no general learning and inference procedures. The generalised Wishart process (GWP) addresses all of these issues. Specifically, in our GWP formulation,

- Estimation of $\Sigma(t)$ is tractable in at least 200 dimensions, even without a factor representation.
- The input variable can come from any arbitrary index set $\mathcal{X}$, just as easily as it can represent time. This allows one to condition on covariates like interest rates.
- One can easily handle missing data.
- One can easily specify a vast range of covariance structures (periodic, smooth, Ornstein-Uhlenbeck, ...).
- We develop Bayesian inference procedures to make predictions, and to learn distributions over any relevant parameters. Aspects of the covariance structure are learned from data, rather than being a fixed property of the model.

Overall, the GWP is versatile and simple. It does not require any free parameters, and any optional parameters are easy to interpret. For this reason, it also scales well with dimension. Yet, the GWP provides an especially general description of multivariate volatility – more so than the most general MGARCH specifications. In the next section, we review Gaussian processes (GPs), which are used to construct the GWP we use in this paper. In the following sections we then review the Wishart distribution, present a GWP construction (which we use as a prior over $\Sigma(t)$ for all t), introduce procedures to sample from the posterior over $\Sigma(t)$, review the main competitor, MGARCH, and present experiments comparing the GWP to MGARCH on simulated and financial data. These experiments include a 5 dimensional data set, based on returns for NASDAQ, FTSE, NIKKEI, TSE, and the Dow Jones Composite, and a set of returns for 3 foreign currency exchanges. We also have a 200 dimensional experiment to show how the GWP can be used to study high dimensional problems.

Also, although it is not the focus of this paper, we show in the inference section how the GWP can be used as part of a new GP based regression model that accounts for *changing* correlations. In other words, it can be used to predict the mean $\boldsymbol{\mu}(t)$ together with the covariance matrix $\Sigma(t)$ of a multivariate process. Alternative GP based multivariate regression models for $\boldsymbol{\mu}(t)$, which account for fixed correlations, were recently introduced by Bonilla et al. (2008), Teh et al. (2005), and Boyle and Frean (2004). We develop this extension, and many others, in a forthcoming paper.

2 GAUSSIAN PROCESSES

We briefly review Gaussian processes, since the generalised Wishart process is constructed from GPs. For more detail, see Rasmussen and Williams (2006).

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. Using a Gaussian process, we can define a distribution over functions $u(x)$:

$$u(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{2}$$

where x is an arbitrary (potentially vector valued) input variable, and the mean $m(x)$ and kernel function $k(x, x')$ are respectively defined as

$$m(x) = \mathbb{E}[u(x)], \tag{3}$$

$$k(x, x') = \mathrm{cov}(u(x), u(x')). \tag{4}$$

This means that any collection of function values has a joint Gaussian distribution:

$$(u(x_1), u(x_2), \ldots, u(x_N))^\top \sim \mathcal{N}(\boldsymbol{\mu}, K), \tag{5}$$

where the $N \times N$ covariance matrix K has entries $K_{ij} = k(x_i, x_j)$, and the mean $\boldsymbol{\mu}$ has entries $\boldsymbol{\mu}_i = m(x_i)$. The properties of these functions (smoothness, periodicity, etc.) are determined by the kernel function. The squared exponential kernel is popular:

$$k(x, x') = \exp(-0.5||x - x'||^2/l^2). \tag{6}$$

[1] Our model is also related to Gelfand et al. (2004)'s coregionalisation model, which we discuss in section 5.

Functions drawn from a Gaussian process with this kernel function are smooth, and can display long range trends. The length-scale *hyperparameter* l is easy to interpret: it determines how much the function values $u(x)$ and $u(x + \boldsymbol{a})$ depend on one another, for some constant $\boldsymbol{a}$.

Autoregressive processes such as

$$u(t+1) = u(t) + \epsilon(t)\,, \tag{7}$$

$$\epsilon(t) \sim \mathcal{N}(0,1)\,, \tag{8}$$

are widely used in time series modelling and are a particularly simple special case of Gaussian processes.

3 WISHART DISTRIBUTION

The Wishart distribution defines a probability density function over positive definite matrices S:

$$p(S|V,\nu) = \frac{|S|^{(\nu-D-1)/2}}{2^{\nu D/2}|V|^{\nu/2}\Gamma_D(\nu/2)} \exp(-\frac{1}{2}\mathrm{tr}(V^{-1}S))\,, \tag{9}$$

where V is a $D \times D$ positive definite scale matrix, and $\nu > D$ is the number of degrees of freedom. This distribution has mean νV and mode $(D-\nu-1)V$ for $\nu \geq D+1$. $\Gamma_D(\cdot)$ is the multivariate gamma function:

$$\Gamma_D(\nu/2) = \pi^{D(D-1)/4} \prod_{j=1}^{D} \Gamma(\nu/2 + (1-j)/2)\,. \tag{10}$$

The Wishart distribution is a multivariate generalisation of the Gamma distribution when ν is real valued, and the chi-square (χ^2) distribution when ν is integer valued. The sum of squares of univariate Gaussian random variables is chi-squared distributed. Likewise, the sum of outer products of multivariate Gaussian random variables is Wishart distributed:

$$S = \sum_{i=1}^{\nu} \boldsymbol{u}_i \boldsymbol{u}_i^\top \sim \mathcal{W}_D(V,\nu)\,, \tag{11}$$

where the $\boldsymbol{u}_i$ are i.i.d. $\mathcal{N}(\boldsymbol{0}, V)$ D-dimensional random variables, and $\mathcal{W}_D(V,\nu)$ is a Wishart distribution with $D \times D$ scale matrix V, and ν degrees of freedom. S is a $D \times D$ positive definite matrix. If $D = V = 1$ then $\mathcal{W}$ is a chi-square distribution with ν degrees of freedom. S^{-1} has the inverse Wishart distribution, $\mathcal{W}_D^{-1}(V^{-1},\nu)$, which is a conjugate prior for covariance matrices of zero mean Gaussian distributions.

4 A GENERALISED WISHART PROCESS CONSTRUCTION

We saw that the Wishart distribution is constructed from multivariate Gaussian distributions. Essentially, by replacing these Gaussian distributions with Gaussian processes, we define a process with Wishart marginals – an example of a *generalised Wishart process*. It is a collection of positive semi-definite random matrices indexed by any arbitrary (potentially high dimensional) variable x. For clarity, we assume that time is the input variable, even though it takes no more effort to use a vector-valued variable x from any arbitrary set. Everything still applies if we replace t with x. In an upcoming journal submission (Wilson and Ghahramani, 2011b), we introduce several new constructions, some of which do not have Wishart marginals.

Suppose we have νD independent Gaussian process functions, $u_{id}(t) \sim \mathcal{GP}(0,k)$, where $i = 1,\ldots,\nu$ and $d = 1,\ldots,D$. This means $\mathrm{cov}(u_{id}(t), u_{i'd'}(t')) = k(t,t')\delta_{ii'}\delta_{dd'}$, and $(u_{id}(t_1), u_{id}(t_2), \ldots, u_{id}(t_N))^\top \sim \mathcal{N}(0,K)$, where δ_{ij} is the Kronecker delta, and K is an $N \times N$ covariance matrix with elements $K_{ij} = k(t_i,t_j)$. Let $\hat{\boldsymbol{u}}_i(t) = (u_{i1}(t),\ldots,u_{iD}(t))^\top$, and let L be the lower Cholesky decomposition of a $D \times D$ scale matrix V, such that $LL^\top = V$. Then at each t the covariance matrix $\Sigma(t)$ has a Wishart marginal distribution,

$$\Sigma(t) = \sum_{i=1}^{\nu} L\hat{\boldsymbol{u}}_i(t)\hat{\boldsymbol{u}}_i^\top(t)L^\top \sim \mathcal{W}_D(V,\nu)\,, \tag{12}$$

subject to the constraint that the kernel function $k(t,t) = 1$.

We can understand (12) as follows. Each element of the vector $\hat{\boldsymbol{u}}_i(t)$ is a univariate Gaussian with zero mean and variance $k(t,t) = 1$. Since these elements are uncorrelated, $\hat{\boldsymbol{u}}_i(t) \sim \mathcal{N}(\boldsymbol{0}, I)$. Therefore $L\hat{\boldsymbol{u}}_i(t) \sim \mathcal{N}(0,V)$, since $\mathbb{E}[L\hat{\boldsymbol{u}}_i(t)\hat{\boldsymbol{u}}_i(t)^\top L^\top] = LIL^\top = LL^\top = V$. We are summing the outer products of $\mathcal{N}(0,V)$ random variables, and there are ν terms in the sum, so by definition this has a Wishart distribution $\mathcal{W}_D(V,\nu)$. It was not a restriction to assume $k(t,t) = 1$, since any scaling of k can be absorbed into L. In a forthcoming journal paper (Wilson and Ghahramani, 2011b), we use the following formal and general definition, for a GWP indexed by a variable x in an arbitrary set $\mathcal{X}$.

Definition 4.1. *A Generalised Wishart Process is a collection of positive semi-definite random matrices indexed by $x \in \mathcal{X}$ and constructed from outer products of points from collections of stochastic processes like in* (12)*. If the random matrices have 1) Wishart marginal distributions, meaning that $\Sigma(x) \sim W_p(V,\nu)$ at every $x \in \mathcal{X}$, and 2) dependence on x as defined by a kernel $k(x,x')$, then we write*

$$\Sigma(x) \sim \mathcal{GWP}(V,\nu,k(x,x'))\,. \tag{13}$$

The $\mathcal{GWP}$ notation of Definition 4.1 is just like the notation for a Wishart distribution, but includes a kernel

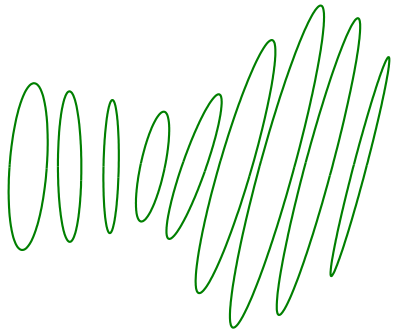

Figure 1: A draw from a generalised Wishart process (GWP). Each ellipse is a 2×2 covariance matrix indexed by time, which increases from left to right. The rotation indicates the correlation between the two variables, and the major and minor axes scale with the eigenvalues of the matrix. Like a draw from a Gaussian process is a collection of function values indexed by time, a draw from a GWP is a collection of matrices indexed by time.

which controls the dynamics of how $\Sigma(t)$ varies with t. This pleasingly compartmentalises the GWP, since there is separation between the shape parameters and temporal (or spatial) dynamics parameters. We show an example of these dynamics in Figure 1 with a draw from a GWP.

Using this construction, we can also define a generalised *inverse* Wishart process (GIWP). If $\Sigma(t) \sim \mathcal{GWP}$, then inversion at each value of t defines a draw $R(t) = \Sigma(t)^{-1}$ from the GIWP. The conjugacy of the GIWP with a Gaussian likelihood could be useful when doing Bayesian inference.

We can further extend this construction by replacing the Gaussian processes with *copula processes* (Wilson and Ghahramani, 2010). For example, as part of Bayesian inference we could learn a mapping that would transform the Gaussian processes u_{id} to Gaussian copula processes with marginals that better suit the covariance structure of our data set. In a forthcoming journal paper (Wilson and Ghahramani, 2011b), we elaborate on such a construction, which allows the marginals to be Wishart distributed with *real valued* degrees of freedom ν. In this journal paper we also introduce efficient representations, and "heavy-tailed" representations.

The formulation we outlined in this section is different from other multivariate volatility models in that one can specify a kernel function $k(t, t')$ that controls how $\Sigma(t)$ varies with t – for example, $k(t, t')$ could be periodic – and t need not be time: it can be an arbitrary input variable, including covariates like interest rates, and does not need to be represented on an evenly spaced grid. In the next section we introduce, for the first time, general inference procedures for making predictions when using a Wishart process prior, allowing 1) the model to scale to hundreds of dimensions without a factor representation, and 2) for aspects of the covariance structure to be learned from data. These are based on recently developed Markov chain Monte Carlo techniques (Murray et al., 2010). We also introduce a new method for doing multivariate GP based regression with *dynamic* correlations.

5 BAYESIAN INFERENCE

Assume we have a generalised Wishart process prior on a dynamic $D \times D$ covariance matrix:

$$\Sigma(t) \sim \mathcal{GWP}(V, \nu, k)\,. \tag{14}$$

We want to sample from the posterior $\Sigma(t)$ given a D-dimensional data set $\mathcal{D} = \{\boldsymbol{y}(t_n) : n = 1, \ldots, N\}$. We explain how to do this for a general likelihood function, $p(\mathcal{D}|\Sigma(t))$, by finding the posterior distributions over the parameters in the model, given the data $\mathcal{D}$. These parameters are: a vector of all relevant GP function values $\boldsymbol{u}$, the hyperparameters of the GP kernel function $\boldsymbol{\theta}$, the degrees of freedom ν, and L, the lower cholesky decomposition of the scale matrix V ($LL^\top = V$). The graphical model in Figure 1*s* (*of supplementary material* [2]) shows all the relevant parameters and conditional dependence relationships. The free parameters L and $\boldsymbol{\theta}$ have clear interpretations: L gives a prior on the expectation of $\Sigma(t)$ for all t, and $\boldsymbol{\theta}$ describes how this structure changes with time – how much past data, for instance, one would need to make an accurate forecast. The degrees of freedom ν control how concentrated our prior is around our expected value of $\Sigma(t)$; the smaller ν, the more broad our prior is on $\Sigma(t)$. Learning the values of these parameters provides useful information. In the supplementary material, we show explicitly how the kernel function k controls the autocovariances for entries of $\Sigma(t)$ at different times.

We can sample from these posterior distributions using Gibbs sampling (Geman and Geman, 1984), a Markov chain Monte Carlo algorithm where initialising $\{\boldsymbol{u}, \boldsymbol{\theta}, L, \nu\}$ and then sampling in cycles from

$$p(\boldsymbol{u}|\boldsymbol{\theta}, L, \nu, \mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{u}, L, \nu)p(\boldsymbol{u}|\boldsymbol{\theta})\,, \tag{15}$$
$$p(\boldsymbol{\theta}|\boldsymbol{u}, L, \nu, \mathcal{D}) \propto p(\boldsymbol{u}|\boldsymbol{\theta})p(\boldsymbol{\theta})\,, \tag{16}$$
$$p(L|\boldsymbol{\theta}, \boldsymbol{u}, \nu, \mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{u}, L, \nu)p(L)\,, \tag{17}$$
$$p(\nu|\boldsymbol{\theta}, \boldsymbol{u}, L, \mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{u}, L, \nu)p(\nu)\,, \tag{18}$$

will converge to samples from $p(\boldsymbol{u}, \boldsymbol{\theta}, L, \nu|\mathcal{D})$. We will successively describe how to sample from the posterior distributions (15), (16), (17), and (18). In our

[2] http://mlg.eng.cam.ac.uk/andrew/gwpsupp.pdf

discussion we assume there are N data points (one at each time step or input), and D dimensions. We then explain how to make predictions of $\Sigma(t_*)$ at some test input t_*. Finally, we discuss a potential likelihood function, and how the GWP could also be used as part of a new GP based model for multivariate regression with outputs that have changing correlations.

5.1 SAMPLING THE GP FUNCTIONS

In this section we describe how to sample from the posterior distribution (15) over the Gaussian process function values $\boldsymbol{u}$. We order the entries of $\boldsymbol{u}$ by fixing the degrees of freedom and dimension, and running the time steps from $n = 1, \ldots, N$. We then increment dimensions, and finally, degrees of freedom. So $\boldsymbol{u}$ is a vector of length $ND\nu$. As before, let K be an $N \times N$ covariance matrix, formed by evaluating the kernel function at all pairs of training inputs. Then the prior $p(\boldsymbol{u}|\boldsymbol{\theta})$ is a Gaussian distribution with $ND\nu \times ND\nu$ block diagonal covariance matrix K_B, formed using $D\nu$ of the K matrices; if the hyperparameters of the kernel function change depending on dimension or degrees of freedom, then these K matrices will be different from one another. In short,

$$p(\boldsymbol{u}|\boldsymbol{\theta}) = \mathcal{N}(0, K_B)\,. \tag{19}$$

With this prior, and the likelihood formulated in terms of the other parameters, we can sample from the posterior (15). Sampling from this posterior is difficult, because the Gaussian process function values are highly correlated by the K_B matrix. We use Elliptical Slice Sampling (Murray et al., 2010): it has no free parameters, jointly updates every element of $\boldsymbol{u}$, and was especially designed to sample from posteriors with correlated Gaussian priors. We found it effective.

5.2 SAMPLING OTHER PARAMETERS

We can similarly obtain distributions over the other parameters. The priors we use will depend on the data we are modelling. We placed a vague lognormal prior on $\boldsymbol{\theta}$ and sampled from the posterior (16) using axis aligned slice sampling if $\boldsymbol{\theta}$ was one dimensional, and Metropolis Hastings otherwise. We also used Metropolis Hastings to sample from (17), with a spherical Gaussian prior on the elements of L. To sample (18), one can use reversible jump MCMC (Green, 1995; Robert and Casella, 2004). But in our experiments we set $\nu = D+1$, letting the prior be as flexible as possible, and focus on other aspects of our model. In an upcoming journal paper (Wilson and Ghahramani, 2011b), we introduce a GWP construction with real valued degrees of freedom, where it is easy to sample from $p(\nu|\mathcal{D})$. Although learning L is not expensive, one might simply wish to set it by taking the empirical covariance of any data not used for predictions, dividing by the degrees of freedom, and then taking the lower cholesky decomposition.

5.3 MAKING PREDICTIONS

Once we have learned the parameters $\{\boldsymbol{u}, \boldsymbol{\theta}, L, \nu\}$, we can find a distribution over $\Sigma(t_*)$ at a test input t_*. To do this, we must infer the distribution over $\boldsymbol{u}_*$ – all the relevant GP function values at t_*:

$$\begin{aligned}\boldsymbol{u}_* = [&u_{11}(t_*), \ldots, u_{1D}(t_*), u_{21}(t_*), \ldots, u_{2D}(t_*), \\ &\ldots, u_{\nu 1}(t_*), \ldots, u_{\nu D}(t_*)]^\top .\end{aligned} \tag{20}$$

Consider the joint distribution over $\boldsymbol{u}$ and $\boldsymbol{u}_*$:

$$\begin{bmatrix}\boldsymbol{u} \\ \boldsymbol{u}_*\end{bmatrix} \sim \mathcal{N}(\boldsymbol{0}, \begin{bmatrix} K_B & A^\top \\ A & I_p \end{bmatrix}). \tag{21}$$

Supposing that $\boldsymbol{u}_*$ and $\boldsymbol{u}$ respectively have p and q elements, then A is a $p \times q$ matrix of covariances between the GP function values $\boldsymbol{u}_*$ and $\boldsymbol{u}$ at all pairs of the training and test inputs: $A_{ij} = k_i(t_*, t_{\mathrm{mod}(N+1,j)})$ if $1 + (i-1)N \leq j \leq iN$, and 0 otherwise. The kernel function k_i may differ from row to row, if it changes depending on the degree of freedom or dimension; for instance, we could have a different length-scale for each new dimension. I_p is a $p \times p$ identity matrix representing the prior independence between the GP function values in $\boldsymbol{u}_*$. Conditioning on $\boldsymbol{u}$, we find

$$\boldsymbol{u}_*|\boldsymbol{u} \sim \mathcal{N}(AK_B^{-1}\boldsymbol{u}, I_p - AK_B^{-1}A^\top)\,. \tag{22}$$

We can then construct $\Sigma(t_*)$ using equation (12) and the elements of $\boldsymbol{u}_*$.

5.4 LIKELIHOOD FUNCTION

So far we have avoided making the likelihood explicit; the inference procedure we described will work with a variety of likelihoods parametrized through a matrix $\Sigma(t)$, such as the multivariate t distribution. However, assuming for simplicity that each of the variables $\boldsymbol{y}(t_n)$ has a Gaussian distribution,

$$\boldsymbol{y}(t) \sim \mathcal{N}(\boldsymbol{\mu}(t), \Sigma(t)), \tag{23}$$

then the likelihood is

$$\begin{aligned}p(\mathcal{D}|\boldsymbol{\mu}(t), \Sigma(t)) &= \prod_{n=1}^{N} p(\boldsymbol{y}(t_n)|\boldsymbol{\mu}(t_n), \Sigma(t_n)) \\ = \prod_{n=1}^{N} |2\pi\Sigma(t_n)|^{-1/2} &\exp[-\frac{1}{2}\boldsymbol{w}(t_n)^\top \Sigma(t_n)^{-1}\boldsymbol{w}(t_n)],\end{aligned} \tag{24}$$

where $\boldsymbol{w}(t_n) = \boldsymbol{y}(t_n) - \boldsymbol{\mu}(t_n)$.

We can learn a distribution over $\boldsymbol{\mu}(t)$, in addition to $\Sigma(t)$. One possible specification would be to let $\boldsymbol{\mu}(t)$ be a separate vector of Gaussian processes:

$$\boldsymbol{\mu}(t) = \hat{\boldsymbol{u}}_{\nu+1}(t)\,. \tag{25}$$

We discuss this further in an upcoming paper where we develop a multivariate Gaussian process regression model which accounts for dynamic correlations between the outputs.

Alternative models, which account for fixed correlations, have recently been introduced by Bonilla et al. (2008), Teh et al. (2005), and Boyle and Frean (2004). Rather than use a GP based regression as in (25), Gelfand et al. (2004) combine a spatial Wishart process with a parametric linear regression on the mean, to make correlated mean predictions in a 2D spatial setting. There are some important differences in our methodology: 1) the correlation structure is a fixed property of their model (e.g. they do not learn the parameters of a kernel function and so the autocorrelations are not learned from data), 2) they are not interested in developing a model of multivariate volatility; they do not explicitly evaluate or provide a means to forecast dynamic correlations, 3) the prior expectation $\mathbb{E}[\Sigma(t)]$ at each t is diagonal (which is not what we expect when applying a multivariate volatility model), 4) the regression on $\boldsymbol{\mu}(x)$ is linear, 5) the observations must be on a regularly spaced grid (no missing observations), and perhaps most importantly, 6) the inference relies solely on Metropolis Hastings with Gaussian proposals, which is not tractable for $p > 3$, and will not mix efficiently as the strong GP prior correlations are not accounted for (Murray et al., 2010). In this paper we focus on making predictions of $\Sigma(t)$, setting $\boldsymbol{\mu} = 0$. In an upcoming paper we develop and implement a multivariate GP regression model which accounts for dynamic correlations between the outputs, like we have briefly described in this section.

5.5 COMPUTATIONAL COMPLEXITY

Our method is mainly limited by taking the cholesky decomposition of the block diagonal K_B, a $ND\nu \times ND\nu$ matrix. However, $\mathrm{chol}(\mathrm{blkdiag}(A, B, \dots)) = \mathrm{blkdiag}(\mathrm{chol}(A), \mathrm{chol}(B), \dots)$. So in the case with equal length-scales for each dimension, we only need to take the cholesky of an $N \times N$ matrix K, an $\mathcal{O}(N^3)$ operation, independent of dimension! In the more general case with D different length-scales, it is an $\mathcal{O}(DN^3)$ operation. The total "training" complexity therefore amounts to one $\mathcal{O}(N^3)$ or one $\mathcal{O}(DN^3)$ operation. Sampling then requires likelihood evaluations, which cost $\mathcal{O}(N\nu D^2)$ operations. Thus we could in principle go to about 1000 dimensions, assuming ν is $\mathcal{O}(D)$, and for instance, a couple years worth of financial training data, which is typical for GARCH (Brownlees et al., 2009). In practice, MCMC may be infeasible for very high D, but we have found Elliptical Slice Sampling incredibly robust. Overall, this is impressive scaling – without further assumptions in our model, we can go well beyond 5 dimensions with full generality. In the supplementary material we have a 200 dimensional experiment as an empirical accompaniment to this discussion.

6 MULTIVARIATE GARCH

We compare predictions of $\Sigma(t)$ made by the generalised Wishart process to those made by multivariate GARCH (MGARCH), since GARCH (Bollerslev, 1986) is extremely popular and arguably unsurpassed at predicting the volatility of returns on equity indices and currency exchanges (Poon and Granger, 2005; Hansen and Lunde, 2005; Brownlees et al., 2009).

Consider a zero mean D dimensional vector stochastic process $\boldsymbol{y}(t)$ with a time changing covariance matrix $\Sigma(t)$ as in (23). In the general MGARCH framework,

$$\boldsymbol{y}(t) = \Sigma(t)^{1/2}\boldsymbol{\eta}(t)\,, \tag{26}$$

where $\boldsymbol{\eta}(t)$ is an i.i.d. vector white noise process with $\mathbb{E}[\boldsymbol{\eta}(t)\boldsymbol{\eta}(t)^\top] = I$, and $\Sigma(t)$ is the covariance matrix of $\boldsymbol{y}(t)$ conditioned on all information up until time $t-1$.

The first and most general MGARCH model, the VEC model of Bollerslev et al. (1988), specifies Σ_t as

$$\mathrm{vech}(\Sigma_t) = \boldsymbol{a}_0 + \sum_{i=1}^{q} A_i \mathrm{vech}(\boldsymbol{y}_{t-i}\boldsymbol{y}_{t-i}^\top) + \sum_{j=1}^{p} B_j \mathrm{vech}(\Sigma_{t-j})\,. \tag{27}$$

A_i and B_j are $D(D+1)/2 \times D(D+1)/2$ matrices of parameters, and $\boldsymbol{a}_0$ is a $D(D+1)/2 \times 1$ vector of parameters.[3] The vech operator stacks the columns of the lower triangular part of a $D \times D$ matrix into a vector of length $D(D+1)/2$. For example, $\mathrm{vech}(\Sigma) = (\Sigma_{11}, \Sigma_{21}, \dots, \Sigma_{D1}, \Sigma_{22}, \dots, \Sigma_{D2}, \dots, \Sigma_{DD})^\top$. This model is general, but difficult to use. There are $(p+q)(D(D+1)/2)^2 + D(D+1)/2$ parameters! These parameters are hard to interpret, and there are no conditions under which Σ_t is positive definite for all t. Gouriéroux (1997) discusses the challenging (and sometimes impossible) problem of keeping Σ_t positive definite. Training is done by a constrained maximum likelihood, where the log likelihood is given by

$$\mathcal{L} = -\frac{ND}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{N}[\log|\Sigma_t| + \boldsymbol{y}_t^\top \Sigma_t^{-1}\boldsymbol{y}_t]\,, \tag{28}$$

[3] We use $\Sigma(t)$ and Σ_t interchangeably.

supposing that $\boldsymbol{\eta}_t \sim \mathcal{N}(0, I)$, and that there are N training points.

Subsequent efforts have led to simpler but less general models. We can let A_j and B_j be diagonal matrices. This model has notably fewer (though still $(p+q+1)D(D+1)/2$) parameters, and there are conditions under which Σ_t is positive definite for all t (Engle et al., 1994). But now there are no interactions between the different conditional variances and covariances. A popular variant assumes *constant* correlations between the D components of $\boldsymbol{y}$, and only lets the marginal variances – the diagonal entries of $\Sigma(t)$ – vary (Bollerslev, 1990).

We compare to the 'full' BEKK variant of Engle and Kroner (1995), as implemented by Kevin Shepphard in the UCSD GARCH Toolbox.[4] We chose BEKK because it is the most general MGARCH variant in widespread use. We use the first order model:

$$\Sigma_t = CC^\top + A^\top \boldsymbol{y}_{t-1}\boldsymbol{y}_{t-1}^\top A + B^\top \Sigma_{t-1} B\,, \qquad (29)$$

where A, B and C are $D \times D$ matrices of parameters. C is lower triangular to ensure that Σ_t is positive definite during maximum likelihood training. For a full review of MGARCH, see Silvennoinen and Teräsvirta (2009).

7 EXPERIMENTS

In our experiments, we predict the covariance matrix for multivariate observations $\boldsymbol{y}(t)$ as $\hat{\Sigma}(t) = \mathbb{E}[\Sigma(t)|\mathcal{D}]$. These experiments closely follow Brownlees et al. (2009), a rigorous empirical comparison of GARCH models. We use a Gaussian likelihood, as in (24), except with a zero mean function. We make historical predictions, and one step ahead forecasts. Historical predictions are made at observed time points, or between these points. The one step ahead forecasts are predictions of $\Sigma(t+1)$ taking into account all observations until time t. Historical predictions can be used, for example, to understand the nature of covariances between equity indices during a past financial crisis.

To make these predictions we learn distributions over the GWP parameters through the Gibbs sampling procedure outlined in section 5. The kernel functions we use are solely parametrized by a one dimensional length-scale l, which indicates how dependent $\Sigma(t)$ and $\Sigma(t+a)$ are on one another. We place a lognormal prior on the length-scale, and sample from the posterior with axis-aligned slice sampling.

For each experiment, we choose a kernel function we want to use with the GWP. We then compare to a GWP that uses an Ornstein-Uhlenbeck (OU) kernel function, $k(t,t') = \exp(-|t-t'|/l)$. Even though we are still taking advantage of the inference procedures in the GWP formulation, we refer to this variant of GWP as a simple Wishart process (WP), since the classic Bru (1991) construction is like a special case of a generalised Wishart process restricted to using a one dimensional GP with an OU covariance structure.

To assess predictions we use the Mean Squared Error (MSE) between the predicted and true covariance matrices, which is always safe since we never observe the true $\Sigma(t)$. When the truth is not known, we use the proxy $S_{ij}(t) = y_i(t)y_j(t)$, to harmonize with the econometrics literature. y_i is the i^{th} component of the multivariate observation $\boldsymbol{y}(t)$. This is intuitive because $\mathbb{E}[y_i(t)y_j(t)] = \Sigma_{ij}(t)$, assuming $\boldsymbol{y}(t)$ has a zero mean. In a thorough empirical study, Brownlees et al. (2009) use the univariate analogue of this proxy. We do not use likelihood for assessing historical predictions, since that is a training error (for MGARCH), but we do use log likelihood ($\mathcal{L}$) for forecasts.

We begin by generating a 2×2 time varying covariance matrix $\Sigma_p(t)$ with periodic components, and simulating data at 291 time steps from a Gaussian:

$$\boldsymbol{y}(t) \sim \mathcal{N}(\mathbf{0}, \Sigma_p(t))\,. \qquad (30)$$

Periodicity is especially common to financial and climate data, where daily trends repeat themselves. For example, the intraday volatility on equity indices and currency exchanges has a periodic covariance structure. Andersen and Bollerslev (1997) discuss the lack of – and critical need for – models that account for this periodicity. In the GWP formulation, we can easily account for this by using a periodic kernel function, whereas in previous Wishart process volatility models, we are stuck with an OU covariance structure. We reconstruct $\Sigma_p(t)$ using the kernel $k(t,t') = \exp(-2\sin((t-t')^2)/l^2)$. We reconstructed the historical Σ_p at all 291 data points, and after having learned the parameters for each of the models from the first 200 data points, made one step forecasts for the last 91 points. Table 1 and Figure 2 show the results. We call this data set `PERIODIC`. The GWP outperforms the competition on all error measures. It identifies the periodicity and underlying smoothness of Σ_p that neither the WP nor MGARCH accurately discern: both are too erratic. MGARCH is especially poor at learning the time changing covariance in this data set.

For our next experiment, we predict $\Sigma(t)$ for the returns on three currency exchanges – the Canadian to US Dollar, the Euro to USD, and the USD to the Great Britain Pound – in the period 15/7/2008-15/2/2010; this encompasses the recent financial crisis and so is of particular interest to economists.[5] We call this data

[4] http://www.kevinsheppard.com/wiki/UCSD_GARCH

[5] We define a *return* as $r_t = \log(P_{t+1}/P_t)$, where P_t is

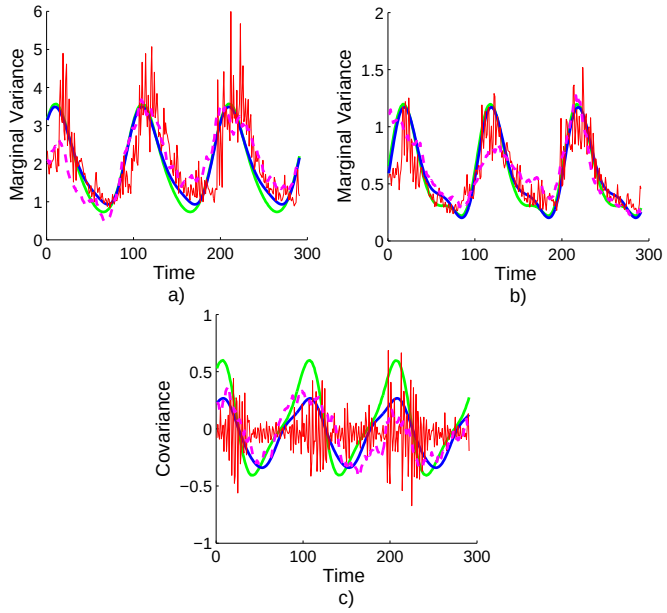

Figure 2: Reconstructing the historical $\Sigma_p(t)$ for the `PERIODIC` data set. We show the truth (green), and GWP (blue), WP (dashed magenta), and MGARCH (thin red) predictions. a) and b) are the marginal variances (diagonal elements of $\Sigma_p(t)$), and c) is the covariance.

set `EXCHANGE`. We use the proxy $S_{ij}(t) = y_i(t)y_j(t)$. With the GWP, we use the squared exponential kernel $k(t,t') = \exp(-0.5(t-t')^2/l^2)$. We make 200 one step ahead forecasts, having learned the parameters for each of the models on the previous 200 data points. We also make 200 historical predictions for the same data points as the forecasts. Results are in Table 1.

Unfortunately, we cannot properly assess predictions on natural data, because we do not know the true $\Sigma(t)$. For example, whether we use MSE with a proxy, or likelihood, historical predictions would be assessed with the same data used for training. In consideration of this problem, we generated a time varying covariance matrix $\tilde{\Sigma}(t)$ based on the empirical time varying covariance of the daily returns on five equity indices – NASDAQ, FTSE, TSE, NIKKEI, and the Dow Jones Composite – over the period from 15/2/1990-15/2/2010. We then generated a return series by sampling from a multivariate Gaussian at each time step using $\tilde{\Sigma}(t)$. As seen in Figure 2*s* (*supplementary*), the generated return series behaves like equity index returns. This method is not faultless; for example, we assume that the returns are normally distributed. However, the models we compare between also make this assumption, and so no model is given an unfair advantage. And there is a critical benefit: we compare predictions with the true underlying $\tilde{\Sigma}(t)$.

the price on day t.

Table 1: Error for predicting multivariate volatility.

	MSE Historical	MSE Forecast	$\mathcal{L}$ Forecast
`PERIODIC`:			
GWP	**0.0210**	**0.0295**	**−257**
WP	0.115	0.760	−286
MGARCH	0.228	0.488	−270
`EXCHANGE`:			
GWP	$\mathbf{3.88 \times 10^{-9}}$	$\mathbf{4.80 \times 10^{-9}}$	2020
WP	$\mathbf{3.88 \times 10^{-9}}$	6.98×10^{-9}	1950
MGARCH	3.96×10^{-9}	4.94×10^{-9}	**2050**
`EQUITY`:			
GWP	$\mathbf{2.80 \times 10^{-9}}$	$\mathbf{5.84 \times 10^{-9}}$	**2930**
WP	3.96×10^{-9}	8.92×10^{-9}	1710
MGARCH	6.68×10^{-9}	29.4×10^{-9}	2760

To make forecasts and historical predictions on this data set (`EQUITY`), we used a GWP with a squared exponential kernel, $k(t,t') = \exp(-0.5(t-t')^2/l^2)$. We follow the same procedure as before and make 200 forecasts and historical predictions; results are in Table 1.

Both the `EXCHANGE` and `EQUITY` data sets are especially suited to GARCH (Poon and Granger, 2005; Hansen and Lunde, 2005; Brownlees et al., 2009; McCullough and Renfro, 1998; Brooks et al., 2001). However, the generalised Wishart process outperforms GARCH on both of these data sets. Based on our experiments, there is evidence that the GWP is particularly good at capturing the co-variances (off-diagonal elements of $\Sigma(t)$) as compared to GARCH. The GWP also outperforms the WP, which has a fixed OU covariance structure, even though in our experiments the WP takes advantage of the new inference procedures we have derived. Thus the difference in performance is likely because the GWP is capable of capturing complex interdependencies, whereas the WP is not.

8 DISCUSSION

We introduced a stochastic process – the generalised Wishart process (GWP) – which we used to model time-varying covariance matrices $\Sigma(t)$. In the future, the GWP could be applied to study how Σ depends on covariates like interest rates, in addition to time. In a forthcoming journal paper we introduce several new GWP constructions, with benefits in expressivity and efficiency (Wilson and Ghahramani, 2011b).

We hope to unify efforts in machine learning and econometrics to inspire new multivariate volatility models that are simultaneously general, easy to interpret, and tractable in high dimensions.

References

Andersen, T. G. and Bollerslev, T. (1997). Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4(2-3):115–158.

Bae, K., Karolyi, G., and Stulz, R. (2003). A new approach to measuring financial contagion. *Review of Financial Studies*, 16(3):717.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. (1990). Modeling the coherence in short-term nominal exchange rates: A multivariate generalized arch approach. *Review of Economics and Statistics*, 72:498–505.

Bollerslev, T., Engle, R. F., and Wooldridge, J. M. (1988). A capital asset pricing model with time-varying covariances. *The Journal of Political Economy*, 96(1):116–131.

Bonilla, E., Chai, K., and Williams, C. (2008). Multi-task Gaussian process prediction. In *NIPS*.

Boyle, P. and Frean, M. (2004). Dependent Gaussian processes. In *NIPS*.

Brooks, C., Burke, S., and Persand, G. (2001). Benchmarks and the accuracy of GARCH model estimation. *International Journal of Forecasting*, 17:45–56.

Brownlees, C. T., Engle, R. F., and Kelly, B. T. (2009). A practical guide to volatility forecasting through calm and storm. Available at SSRN: http://ssrn.com/abstract=1502915.

Bru, M. (1991). Wishart processes. *Journal of Theoretical Probability*, 4(4):725–751.

Chen, Y. and Welling, M. (2010). Dynamical products of experts for modeling financial time series. In *ICML*, pages 207–214.

Engle, R. and Kroner, K. (1995). Multivariate simultaneous generalized ARCH. *Econometric theory*, 11(01):122–150.

Engle, R., Nelson, D., and Bollerslev, T. (1994). Arch models. *Handbook of Econometrics*, 4:2959–3038.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica*, 50(4):987–1007.

Gelfand, A., Schmidt, A., Banerjee, S., and Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, 13(2):263–312.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):721–741.

Gouriéroux, C. (1997). *ARCH models and financial applications*. Springer Verlag.

Gouriéroux, C., Jasiak, J., and Sufana, R. (2009). The Wishart autoregressive process of multivariate stochastic volatility. *Journal of Econometrics*, 150(2):167–181.

Green, P. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711.

Hansen, P. R. and Lunde, A. (2005). A forecast comparison of volatility models: Does anything beat a GARCH(1,1). *Journal of Applied Econometrics*, 20(7):873–889.

Harvey, A., Ruiz, E., and Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2):247–264.

McCullough, B. and Renfro, C. (1998). Benchmarks and software standards: A case study of GARCH procedures. *Journal of Economic and Social Measurement*, 25:59–71.

Murray, I., Adams, R. P., and MacKay, D. J. (2010). Elliptical Slice Sampling. *JMLR: W&CP*, 9:541–548.

Philipov, A. and Glickman, M. (2006). Multivariate stochastic volatility via Wishart processes. *Journal of Business and Economic Statistics*, 24(3):313–328.

Poon, S.-H. and Granger, C. W. (2005). Practical issues in forecasting volatility. *Financial Analysts Journal*, 61(1):45–56.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for Machine Learning*. The MIT Press.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods*. Springer Verlag.

Sharpe, W. (1966). Mutual fund performance. *Journal of business*, 39(1):119–138.

Silvennoinen, A. and Teräsvirta, T. (2009). Multivariate GARCH models. *Handbook of Financial Time Series*, pages 201–229.

Teh, Y., Seeger, M., and Jordan, M. (2005). Semiparametric latent factor models. In *Workshop on Artificial Intelligence and Statistics*, volume 10.

Wilson, A. G. and Ghahramani, Z. (2010). Copula processes. In *NIPS*.

Wilson, A. G. and Ghahramani, Z. (2011a). Generalised Wishart Processes Supplementary Material. `mlg.eng.cam.ac.uk/andrew/gwpsupp.pdf`.

Wilson, A. G. and Ghahramani, Z. (2011b). In preparation.