GAUSSIAN PROCESSES AND NON-PARAMETRIC VOLATILITY FORECASTING

Nicolas CHAPADOS
Computer Science and Operations Research, University of Montreal, Montreal, Canada (chapados@iro.umontreal.ca)

Christian DORION
Department of Finance, HEC Montreal, Montreal, Canada (christian.dorion@hec.ca)

[Preliminary version — July 2010]

Abstract

We provide a formulation of stochastic volatility based on Gaussian processes, a flexible framework for Bayesian nonlinear regression. The advantage of using Gaussian processes in this context is to place volatility forecasting within a regression framework; this allows a large number of explanatory variables to be used for forecasting, a task difficult with standard volatility-forecasting formulations. Our approach builds upon the range-based estimator of Alizadeh, Brandt, and Diebold (2002) to provide much greater accuracy than traditional close-to-close estimators using daily data. Experiments with high-frequency stock market data, using five-minute realized volatility as a benchmark, show that the approach significantly outperforms standard stochastic volatility and GARCH models in one-month-ahead out-of-sample volatility forecasting. It also shows promise for long-range (several months ahead) forecasting, an application that could prove useful in pricing long-dated options.

Keywords Bayesian volatility models; Stochastic volatility; Generalized autoregressive conditional heteroscedasticity models; Long memory in volatility; Multifactor volatility.

1 Introduction

Forecasting volatility at long horizons is a challenging task. While the results of West and Cho (1995) and Christoffersen and Diebold (2000) suggest that volatility predictability is principally a short-horizon phenomenon, recent work by Brandt and Jones (2006) and Engle and Rangel (2008), amongst others, show that respectable long-horizon forecasts can be obtained by using, in the words of Brandt and Jones, "less misspecified volatility models and more informative volatility proxies." This paper aims at pushing this logic further by introducing a new volatility model based on Gaussian processes (GPs), a flexible framework for Bayesian nonlinear regression that naturally allows the use of economic observables in order to better forecast volatility.

Gaussian processes constitute a general and flexible class of models for nonlinear regression and classification.[1] Over the past decade, they have received wide attention in the machine learning community, having originally been introduced in geostatistics, where they are known under the name "kriging".[2] They allow the modeler to engage in a full Bayesian treatment without the complexities of Monte Carlo Markov Chain methods, supplying a complete posterior distribution of forecasts. For regression, they are also computationally relatively simple to implement, the basic model requiring only solving a system of linear equations, albeit one of size equal to the number N of points in the estimation sample (requiring $O(N^3)$ computation).

As we shall observe below, Gaussian processes are closely related to standard stochastic volatility models. In particular, the results of Alizadeh, Brandt, and Diebold (2002) and of Brandt and Jones (2006) can straightforwardly be incorporated within a GP framework. These papers discuss the superior informativeness of the price range[3] over standard volatility proxies such as absolute or squared returns for estimating stochastic volatility. As a result, they demonstrate that range-based estimation of stochastic volatility provides much greater accuracy than traditional estimation based on close-to-close estimators, results that have been known for empirical return variance estimation since at least the work of Parkinson (1980). The advantage of using Gaussian processes in this context is to place stochastic volatility forecasting within a regression framework; this allows a large number

[1] On Gaussian processes, see, amongst many others, O'Hagan (1978), Williams and Rasmussen (1996), and Rasmussen and Williams (2006).

[2] See Matheron (1973) and Cressie (1993).

[3] In other words, the difference between the high and low prices during a time interval such as a day.

of explanatory variables to be used for forecasting, a task difficult with standard volatility-forecasting formulations.

Models used for volatility forecasting abound and how they handle seasonality and explanatory variables differ from one application to the next. Much work has been done on handling seasonality in intraday volatility, starting with Andersen and Bollerslev (1998). At longer horizons, the impact of seasonality on the conditional expectation of returns has received considerable attention as in, for instance, Martens, Van Dijk, and De Pooter (2009). The impact of macroeconomic news on volatility has been studied in a variety of frameworks including that of Flannery and Protopapadakis (2002), who use a number of dummies and macroeconomic observables that impact both on the conditional expected return and variance, or in the MIDAS-GARCH model of Engle, Ghysels, and Sohn (2008), using distributed lags of mixed sampled macroeconomic data. Studies by Engle and Rangel (2008), Corradi, Distaso, and Mele (2009), and Dorion (2010) also suggest that the use of various macroeconomic indicators can improve the forecasting ability of a volatility model.

In all cases however, the authors ultimately blend their dummies and inputs in a linear fashion, all sources of nonlinearity having been chosen by the econometrician. Modifying ABD's linear system to account for seasonal dummies and macro signals would of course be possible, but we would have to select yet another parameterization of an arbitrary linear blend of these observables. Moreover, incorporating a large number of variables directly in a standard stochastic volatility model is likely to prove challenging. Our motivation for using GPs is precisely to let the model learn the linear *and* nonlinear interactions between multiple variables in a non-parameteric framework.

The results in this paper show that GPs can fruitfully exploit the informational content of different economic variables to better forecast volatility. On data simulated using the volatility model discussed in Alizadeh, Brandt, and Diebold (2002), we first establish that a GP model can explain and forecast volatility almost as well as the data-generating model. Yet, as all models are misspecified, the real challenge is to perform accurate out-of-sample forecasts of the volatility observed on actual markets. The Gaussian process model is thus estimated using actual daily range data, and its out-of-sample forecasts are evaluated against the evolution of the five-minute realized volatility, which is derived using high-frequency stock market data. The GP model outperforms standard stochastic volatility and GARCH models in one-month-ahead out-of-sample volatility forecasting. It also shows promise for long-range (several months ahead) forecasting, an application that could prove useful in pricing long-dated options.

This article is organized as follows. We review dynamic volatility models in Section 2 along with related properties of the range. Section 3 provides an overview of the theory behind Gaussian processes, and Section 4 links dynamic volatility processes with the Matérn covariance function, suggesting how they can be employed with GPs. Section 5 presents the results of a simulation study to establish that GPs perform comparably to established filtering techniques when the underlying volatility generating process is known. In Section 6, we turn to the estimation and forecasting of volatility for stock market data and show that, in this context, GPs can significantly outperform established models, particularly if additional input variables are provided. Section 7 concludes and suggests avenues for future research.

2 Dynamic Volatility Models

Volatility of financial asset prices is now widely accepted to be time-varying. While GARCH models (Engle 1982; Bollerslev 1986) acknowledge this issue, they rely on the assumption that shocks to asset prices and asset volatilities arise from a single source, yielding a measurable volatility process. In contrast, stochastic volatility models state that, while they can be correlated, these shocks arise from two different sources. Volatility is thus an unobservable process driven by non-measurable shocks.[4] This paper considers the use of a GP representation of such a volatility process. Before doing so, however, this section discusses recent advances in the estimation of dynamic volatility models, namely the use of ranged-based estimators, and describes two benchmarks, one from the stochastic volatility family and one from the GARCH family.

2.1 Dynamic Volatility and the Daily Price Range

Alizadeh, Brandt, and Diebold (ABD, 2002) demonstrate that the daily (log-) price range, defined as

$$r_t = \log\left(h_{t\tau} - l_{t\tau}\right), \tag{1}$$

where h_t and l_t respectively are the logarithm of the highest and lowest price taken by an asset on day t, is a highly efficient, approximately-normally-distributed proxy of volatility. Consequently, they advocate its use in the estimation of a simple stochastic volatility model. Building on ABD's results, Brandt and Jones (2006) compare the use of absolute or

[4] On stochastic volatility, see Andersen, Bollerslev, Christoffersen, and Diebold (2006) and Shephard (2004).

squared returns to the range-based estimation of a variety of GARCH models. The models estimated using the range dominate their return-based counterparts both in terms of in-sample fit and, most importantly, in terms of their ability to precisely forecast volatility out of sample.

The theory behind these results is the following. Consider a driftless Brownian motion x with constant diffusion coefficient σ and let $r_\tau \equiv \log\left(\sup_{0<t\leq\tau} x_t - \inf_{0<t\leq\tau} x_t\right)$ be the range over a finite interval of length τ.[5] Assuming that $x_0 = 0$, the density of r_τ is given by

$$P(r_\tau|\sigma) = 8\sum_{k=1}^{\infty}(-1)^{k-1}\frac{k^2 e^{r_\tau}}{\sigma\sqrt{\tau}}\phi\Big(\frac{ke^{r_\tau}}{\sigma\sqrt{\tau}}\Big),$$

where $\phi(\cdot)$ is the standard normal density. Under the above process, it can be shown that the first four moments of the range over the time interval τ are given by

$$\begin{aligned}
\mathbb{E}[r_\tau] &= 0.43 + \frac{1}{2}\log\tau + \log\sigma\,, \\
\mathrm{Var}[r_\tau] &= 0.084\,, \\
\mathrm{Skew}(r_\tau) &= 0.17\,, \\
\mathrm{Kurt}(r_\tau) &= 2.80\,.
\end{aligned} \tag{2}$$

Quite remarkably, the variance, skewness and kurtosis of the range distribution are independent of σ. Moreover, both the skewness and kurtosis are very close to those of a normal distribution. Hence, given an observed range r_τ, one may use the quantity $r_\tau - \frac{1}{2}\log\tau - 0.43$ as an unbiased and approximately-normally-distributed estimator of $\log\sigma$.

Moreover, the daily range is a much less noisy volatility proxy than absolute or squared returns (Parkinson 1980; Andersen and Bollerslev 1998), and is robust to market microstructure noise (ABD 2002). Given these features, the range proves valuable in the estimation of volatility models, as shown by ABD and Brandt and Jones (2006).

2.2 A Stochastic Volatility Model

As a first benchmark, we consider the stochastic volatility model analyzed in ABD.[6] This model has the asset price S_t evolve according to a geometric Brownian motion with constant drift μ and stochastic volatility σ_t, the logarithm of the latter being modeled as a

[5] In the context of an analysis of daily returns, the assumption of a constant diffusion coefficient, σ, amounts to assuming that volatility is constant throughout a day.

[6] As pointed out by Brandt and Jones (2005), ABD's specification yields a discretized volatility model that is also studied in Duffie and Singleton (1993), Gallant, Hsieh, and Tauchen (1997), Harvey, Ruiz, and Shephard (1994), Jacquier, Polson, and Rossi (1994), Kim, Shephard, and Chib (1998), and Taylor (1994).

mean-reverting Ornstein–Uhlenbeck process,

$$\frac{\mathrm{d}S_t}{S_t} = \mu\,\mathrm{d}t + \sigma_t \mathrm{d}W_{S_t} \tag{3}$$

$$\mathrm{d}\log\sigma_t = \alpha(\log\bar{\sigma} - \log\sigma_t)\mathrm{d}t + \beta\,\mathrm{d}W_{\sigma_t}, \tag{4}$$

where $\mathrm{d}W_{S_t}$ and $\mathrm{d}W_{\sigma_t}$ are two independent Wiener processes, $\log\bar{\sigma}$ is the long-range volatility, α is a constant governing the mean-reversion rate and β is the constant volatility of volatility. In (4), the drift term $\alpha(\log\bar{\sigma} - \log\sigma_t)$ pulls the process back towards its mean $\log\bar{\sigma}$ with a force proportional to the deviation of the process from its mean.[7]

As financial data is observed at discrete intervals, any continuous-time stochastic volatility model must be discretized at some point. Considering daily data, the Ornstein–Uhlenbeck log-volatility process of eq. (4) can be discretized as the following autoregressive process

$$\log\sigma_{(i+1)\tau} = \log\bar{\sigma} + \rho_\tau(\log\sigma_{i\tau} - \log\bar{\sigma}) + \beta\sqrt{\tau}\,\epsilon^{\sigma}_{(i+1)\tau}\,, \tag{5}$$

where $\tau = 1/252$ the fraction of a year taken by a trading day, i is a day index, $\rho_\tau \equiv 1 - \alpha\tau$, and the $\{\epsilon^{\sigma}_{i\tau}\}$ are uncorrelated standard normal variates.

In later sections, we benchmark a GP model's fitting and forecasting performance against that of the ABD model, which considers eq. (5) as a transition equation for the latent log-volatility process and uses daily log-ranges, $r_{i\tau}$, as log-volatility proxies. Hence, this model implies the following observation equation

$$r_{i\tau} = \log\sigma_{i\tau} + \mathbb{E}\left[r_{i\tau} - \log\sigma_{i\tau}\right] + \sqrt{\mathrm{Var}\left[r_{i\tau}\right]}\,\epsilon^{r}_{i\tau}\,, \tag{6}$$

where $\log\sigma_{i\tau}$ is the latent state and $\epsilon^{r}_{i\tau}$ is the observation noise. The bias and variance of $r_{i\tau}$ as a log-volatility proxy (i.e. the expectation and variance terms in (6)) are given analytically below in (2). This observation equation, together with the transition equation (5), forms a linear dynamical system for which the Kalman smoother (Kalman 1960; Hamilton 1994) provides a likelihood function that can be used to obtain maximum-likelihood estimates of $\log\bar{\sigma}, \beta$ and ρ_τ.[8] The Kalman smoother is also used to extract the latent log-volatilities in sample.

[7] ABD also discuss a two-component version of the model in (3)-(4) that we plan to discuss in the next draft of this work.

[8] Given that ϵ^{r}_{i} is a standard normal innovation while we have no such guaranty, we should here talk of quasi-maximum likelihood. Yet, as we be highlighted shortly, the log-range's distribution is very close to normal thus greatly improving the efficiency of the estimation. See (Alizadeh, Brandt, and Diebold 2002) for details.

Given a set of parameters and latent state estimates at the end of an N-observation, in-sample period, this model implies a joint normal predictive distribution for observations beyond the last observed one, with mean vector and covariance matrix elements given by

$$\mathbb{E}\left[\log \sigma_{(N+k)\tau}\right] = \log \bar{\sigma} + \rho_\tau^k \left(\mathbb{E}\left[\log \sigma_{N\tau}\right] - \log \bar{\sigma}\right) \tag{7}$$

$$\mathrm{Cov}\left[\log \sigma_{(N+k)\tau}, \log \sigma_{(N+j)\tau}\right] = \rho_\tau^{k+j}\Big(\mathrm{Var}\left[\log \sigma_{N\tau}\right] + \beta^2 \tau \textstyle\sum_{i=1}^{\min(k,j)} \rho_\tau^{-2i}\Big), \tag{8}$$

with $j, k \geq 1$, and where $\mathbb{E}\left[\log \sigma_{N\tau}\right]$ and $\mathrm{Var}\left[\log \sigma_{N\tau}\right]$ represent the latent log-volatility distribution at the last in-sample observation, as determined by the Kalman smoother.

2.3 A GARCH Volatility Model

The vast majority of GARCH models are specified as

$$R_{t+1} \coloneqq \log \frac{S_{t+1}}{S_t} = \mu_{t+1} + \sigma_{t+1}\epsilon_{t+1}\,, \tag{9}$$

$$\sigma_{t+1} = f\left(\epsilon_t^2;\, \sigma_s, s \leq t\,\right), \tag{10}$$

where S is the stock price, and f is some parametric functions of squared standardized log-returns ϵ^2 and of past filtered volatilities. Following ABD's work, Brandt and Jones (2006) recently developed versions of the EGARCH model building on the superior informativeness of ranges. They found all range-based to dominate their return-based counterparts both in terms of in-sample fit and of their ability to precisely forecast volatility out of sample.

The model we consider here is the REGARCH, a range-based version of the EGARCH model of Nelson (1991).[9] The model specificication is as follows,

$$r_{t+1} \sim \mathcal{N}\left(0.43 + \log \sigma_{t+1}, 0.29^2\right), \tag{11}$$

$$\log \sigma_{t+1} - \log \sigma_t = \kappa(\theta - \log \sigma_t) + \phi\, \frac{r_t - 0.43 - \log \sigma_{t+1}}{0.29} + \delta\, \frac{R_t}{\sigma_t}\,, \tag{12}$$

where R_t is the log-return and r_t the range observed on day t, and where the normal distribution assumption for the range is based on the results of Section 2.1.

[9] In their work, Brandt and Jones (2006) also discuss a component version of the REGARCH model, which they find to dominate its GARCH(1,1) counterpart in and out of sample. However, in our study, preliminary results tend to show that, when using 5 years of data or less to estimate the component version of the model, allowing for a second component only yields a marginal improvement over the performance of the REGARCH model. It might be that 5 years of data are not enough to properly estimate the long-run component of the model.

Estimating the model with maximum likelihood is straigtforward. To forecast volatility out of sample, while an equivalent of (7) can be derived analytically using the work of Brandt and Jones (2005), deriving a covariance between two forecasts, as in (8), is less straigtforward. We thus rely on Monte-Carlo simulations to estimate empirically these two moments of the REGARCH model's log-volatility forecasts. To compute forecasting likelihoods, we rely on the (simplifying) assumption that the volatility forecasts follow a multivariate normal distribution, just as the one implied by the ABD model.

3 Review of Gaussian Processes for Regression

The present section briefly reviews Gaussian processes at a level sufficient for understanding the spread forecasting methodology developed in the next section.

3.1 Basic Concepts for the Regression Case

A Gaussian process is a generalization of the Gaussian distribution: it represents a probability distribution over *functions* which is entirely specified by a mean and covariance *functions*. Borrowing the succinct definition from Rasmussen and Williams (2006), we formally define a Gaussian process (GP) as

Definition 1. *A* Gaussian process *is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

Let $\mathbf{x}$ index into the real process $f(\mathbf{x})$.[10] We write

$$f(\mathbf{x}) \sim \mathcal{GP}\left(M(\cdot), K(\cdot,\cdot)\right), \tag{13}$$

where functions $M(\cdot)$ and $K(\cdot,\cdot)$ are, respectively, the mean and covariance functions:

$$\begin{aligned} M(\mathbf{x}) &= \mathbb{E}\left[f(\mathbf{x})\right], \\ K(\mathbf{x}_1, \mathbf{x}_2) &= \mathbb{E}\left[\left(f(\mathbf{x}_1) - M(\mathbf{x}_1)\right)\left(f(\mathbf{x}_2) - M(\mathbf{x}_2)\right)\right]. \end{aligned}$$

In a regression setting, we shall assume that we are given an estimation sample $\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid i = 1, \ldots, N\}$, with $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. For convenience, let $\mathbf{X}$ be the matrix of all predictor variables (inputs) in the estimation sample, and $\mathbf{y}$ the vector of responses

[10] Contrarily to many treatments of stochastic processes, there is no necessity for $\mathbf{x}$ to represent time; indeed, in our application of Gaussian processes, some of the elements of $\mathbf{x}$ have a temporal interpretation, but most are general macroeconomic variables used to condition the targets.

(targets). We shall further assume that the y_i are noisy measurements from the underlying process $f(\mathbf{x})$:

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \qquad \text{with } \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_n^2).$$

Regression with a GP is achieved by means of Bayesian inference in order to obtain a posterior distribution over functions given a suitable prior and an observed data sample. Then, given new *test inputs*, we can use the posterior to arrive at a *predictive distribution* conditional on the test inputs and the estimation sample. The predictive distribution is normal for a GP. Although the idea of manipulating distributions over functions may appear cumbersome, the *consistency property* of GPs means that any finite number of values sampled from the process f are jointly normal; hence inference over random functions can be shown to be completely equivalent to inference over a finite number of random variables.

It is often convenient, for simplicity, to assume that the GP prior distribution has a mean of zero,[11]

$$f(\mathbf{x}) \sim \mathcal{GP}(0, K(\cdot, \cdot)).$$

Let $\mathbf{f}' = [f(\mathbf{x}_1)', \ldots, f(\mathbf{x})'_N]$ be the vector of (latent) function values evaluated each point of the estimation sample. Their prior distribution is given by

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, K(\mathbf{X}, \mathbf{X})),$$

where $K(\mathbf{X}, \mathbf{X})_{i,j} \triangleq K(\mathbf{x}_i, \mathbf{x}_j)$ is matrix formed by evaluating the covariance function between all pairs $\mathbf{x}_i, \mathbf{x}_j$. (This matrix is also known as the *kernel* or *Gram* matrix.) We shall consider the joint prior distribution between the estimation sample and an additional set of *test points*, with locations given by the matrix $\mathbf{X}_*$, and whose function values $\mathbf{f}_*$ we wish to infer. Under the GP prior, we have[12]

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right). \tag{14}$$

We obtain the predictive distribution at the test points as follows. From Bayes' theorem,

[11] This assumption has no bearing on the ability of the GP to represent distinctly non-zero posterior means, as shall be obvious below.

[12] Note that the following expressions in this section are implicitly conditioned on $\mathbf{X}$ and $\mathbf{X}_*$. The explicit conditioning notation is omitted for brevity.

the joint posterior given the estimation sample is

$$\begin{aligned} P(\mathbf{f}, \mathbf{f}_* \,|\, \mathbf{y}) &= \frac{P(\mathbf{y} \,|\, \mathbf{f}, \mathbf{f}_*) P(\mathbf{f}, \mathbf{f}_*)}{P(\mathbf{y})} \\ &= \frac{P(\mathbf{y} \,|\, \mathbf{f}) P(\mathbf{f}, \mathbf{f}_*)}{P(\mathbf{y})}, \end{aligned} \tag{15}$$

where $P(\mathbf{y} \,|\, \mathbf{f}, \mathbf{f}_*) = P(\mathbf{y} \,|\, \mathbf{f})$ since, by assumption, the likelihood is conditionally independent of $\mathbf{f}_*$ given $\mathbf{f}$, and

$$\mathbf{y} \,|\, \mathbf{f} \;\sim\; \mathcal{N}(\mathbf{f}, \sigma_n^2 I_N) \tag{16}$$

with I_N the $N \times N$ identity matrix. The desired predictive distribution is obtained by marginalizing out the latent variables corresponding to the estimation sample,

$$\begin{aligned} P(\mathbf{f}_* \,|\, \mathbf{y}) &= \int P(\mathbf{f}, \mathbf{f}_* \,|\, \mathbf{y}) \, \mathrm{d}\mathbf{f} \\ &= \frac{1}{P(\mathbf{y})} \int P(\mathbf{y} \,|\, \mathbf{f}) P(\mathbf{f}, \mathbf{f}_*) \, \mathrm{d}\mathbf{f}. \end{aligned}$$

Since these distributions are all normal, the result of the (normalized) marginal is also normal, and can be shown to have mean and covariance given by

$$\mathbb{E}\left[\mathbf{f}_* \,|\, \mathbf{y}\right] = K(\mathbf{X}_*, \mathbf{X}) \mathbf{\Lambda}^{-1} \mathbf{y}, \tag{17}$$

$$\mathrm{Cov}\left[\mathbf{f}_* \,|\, \mathbf{y}\right] = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}) \mathbf{\Lambda}^{-1} K(\mathbf{X}, \mathbf{X}_*), \tag{18}$$

where we have set

$$\mathbf{\Lambda} = K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I_N. \tag{19}$$

Computation of $\mathbf{\Lambda}^{-1}$ is the most computationally expensive step in Gaussian process regression, requiring $O(N^3)$ time and $O(N^2)$ space.

Note that a natural outcome of GP regression is an expression for not only the expected value at the test points (17), but also the full covariance matrix between those points (18). We shall be making use of this covariance matrix later.

3.2 Choice of Covariance Function

We have so far been silent on the form that should take the covariance function $K(\cdot, \cdot)$. A proper choice for this function is important for encoding prior knowledge about the problem; several striking examples are given in Rasmussen and Williams (2006). In order to yield valid covariance matrices, the covariance function should, at the very least, be symmetric and positive semi-definite, which implies that all its eigenvalues are nonnegative,

$$\int K(\mathbf{u}, \mathbf{v}) f(\mathbf{u}) f(\mathbf{v}) \, \mathrm{d}\mu(\mathbf{u}) \, \mathrm{d}\mu(\mathbf{v}) \geq 0,$$

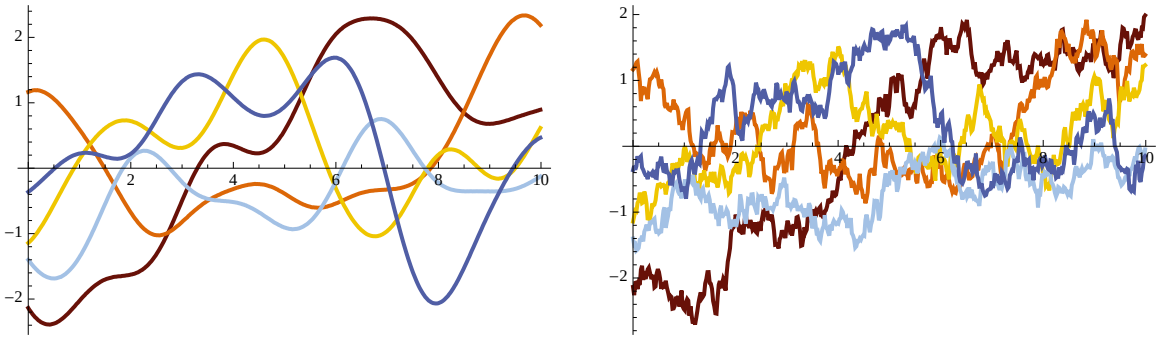

Figure 1: ***Left:*** *Random functions drawn from the GP prior with the* ***squared exponential*** *covariance* ($\sigma_\ell = 1$). ***Right:*** *"Same functions" under the* ***Ornstein–Uhlenbeck*** *covariance* ($\alpha = \frac{1}{2}, \beta = 1$)*, introducted in Section 4.*

for all functions f defined on the appropriate space and measure μ.

One common choice of covariance functions is the *squared exponential*,[13]

$$K_{\mathrm{SE}}(\mathbf{u}, \mathbf{v}; \sigma_\ell) = \exp\left(-\frac{\|\mathbf{u} - \mathbf{v}\|^2}{2\sigma_\ell^2}\right). \tag{20}$$

In the above equation, the hyperparameter σ_ℓ governs the *characteristic length-scale* of the covariance function, indicating the degree of smoothness of the underlying random functions.

Figure 1 illustrates several functions drawn from the GP prior, respectively with the squared exponential and the Ornstein–Uhlenbeck (OU) covariance functions (the latter is introduced below in Section 4). The same random numbers have served for generating both sets, so the functions are "alike" in some sense. Observe that a function under the squared exponential prior is much smoother than the corresponding function under the OU prior.

3.3 Optimization of Hyperparameters

Most covariance functions, including those presented above, have free parameters — termed *hyperparameters* — that control their shape, for instance the characteristic length-scale σ_n^2. It remains to explain how their value should be set. This is an instance of the general problem of *model selection*. For many machine learning algorithms, this problem has often been approached by minimizing a validation error through cross-validation (Stone 1974) and such methods have been proposed for Gaussian processes (Sundararajan and Keerthi 2001).

[13] Also known as the Gaussian or radial basis function kernel.

An alternative approach, quite efficient for Gaussian processes, consists in maximizing the *marginal likelihood*[14] of the observed data with respect to the hyperparameters. This function can be computed by introducing latent function values that are immediately marginalized over. Let θ the set of hyperparameters that are to be optimized; and let $K_{\mathbf{X}}(\theta)$ the covariance matrix computed by a covariance function whose hyperparameters are θ,

$$K_{\mathbf{X}}(\theta)_{i,j} = K(\mathbf{x}_i, \mathbf{x}_j; \theta).$$

The marginal likelihood (here making explicit, for clarity, the dependence on the training inputs and hyperparameters) can be written

$$p(\mathbf{y} \mid \mathbf{X}, \theta) = \int p(\mathbf{y} \mid \mathbf{f}, \mathbf{X}) p(\mathbf{f} \mid \mathbf{X}, \theta) \, \mathrm{d}\mathbf{f}, \tag{21}$$

where the distribution of observations $p(\mathbf{y} \mid \mathbf{f}, \mathbf{X})$, given by eq. (16), is conditionally independent of the hyperparameters given the latent function $\mathbf{f}$. Under the Gaussian process prior (see eq. (14)), we have $\mathbf{f} \mid \mathbf{X}, \theta \sim \mathcal{N}(0, K_{\mathbf{X}}(\theta))$, or in terms of log-likelihood,

$$\log p(\mathbf{f} \mid \mathbf{X}, \theta) = -\frac{1}{2} \mathbf{f}' K_{\mathbf{X}}^{-1}(\theta) \mathbf{f} - \frac{1}{2} \log |K_{\mathbf{X}}(\theta)| - \frac{N}{2} \log 2\pi. \tag{22}$$

Since both distributions in eq. (21) are normal, the marginalization can be carried out analytically to yield

$$\log p(\mathbf{y} \mid \mathbf{X}, \theta) = -\frac{1}{2} \mathbf{y}' (K_{\mathbf{X}}(\theta) + \sigma_n^2 I_N)^{-1} \mathbf{y} - \frac{1}{2} \log \left| K_{\mathbf{X}}(\theta) + \sigma_n^2 I_N \right| - \frac{N}{2} \log 2\pi. \tag{23}$$

This expression can be maximized numerically, for instance by a conjugate gradient algorithm (e.g. Bertsekas 2000) to yield the selected hyperparameters:

$$\theta^* = \arg\max_{\theta} \log p(\mathbf{y} \mid \mathbf{X}, \theta).$$

The likelihood function is in general non-convex and this maximization only finds a local maximum in parameter space; however, empirically it usually works very well for a large class of covariance functions, including those covered in Section 3.2.

It should be mentioned that completely a Bayesian treatment would not be satisfied with such an optimization, since it only picks a single value for each hyperparameter. Instead, one would define a prior distribution on hyperparameters (an *hyperprior*), $p(\theta)$, and marginalize with respect to this distribution. However, for many "reasonable" hyperprior distributions, this is computationally expensive, and often yields no marked improvement over simple optimization of the marginal likelihood (MacKay 1999).

[14] Also called the *integrated likelihood* or (Bayesian) *evidence*.

The gradient of the marginal log-likelihood with respect to the hyperparameters — necessary for numerical optimization algorithms — can be expressed as

$$\frac{\partial \log p(\mathbf{y} \mid \mathbf{X}, \theta)}{\partial \theta_i} = \frac{1}{2}\mathbf{y}' K_{\mathbf{X}}^{-1}(\theta) \frac{\partial K_{\mathbf{X}}(\theta)}{\partial \theta_i} K_{\mathbf{X}}^{-1}(\theta)\mathbf{y} - \frac{1}{2}\mathrm{Tr}\left(K_{\mathbf{X}}^{-1}(\theta) \frac{\partial K_{\mathbf{X}}(\theta)}{\partial \theta_i} \right). \tag{24}$$

See Rasmussen and Williams (2006) for details on the derivation of this equation.

4 Relationship Between GPs and Stochastic Volatility

Process (4) for the log-volatility can be expressed in terms of a stationary covariance function, a consequence of the Wiener-Khintchine theorem (Stein 1999). Without loss of generality, assume that $\log \bar{\sigma} = 0$ (which amounts to shifting the process by a constant). Standard manipulations to (4) (e.g. Rasmussen and Williams 2006) allow to express the prior covariance between $\log \sigma_{t_1}$ and $\log \sigma_{t_2}$ as

$$K_{\mathrm{OU}}(t_1, t_2) = \frac{\beta^2}{2\alpha} e^{-\alpha|t|}, \tag{25}$$

where $t \equiv t_1 - t_2$ is the time interval. This is the covariance function for the Ornstein–Uhlenbeck (OU) process, a special case of the Matérn class of covariance functions (Stein 1999), from which random draws are illustrated in Figure 1. As outlined previously, the hyperparameters α and β do not need to be specified *a priori* and can be estimated by numerical optimization.

Putting together these results with those of Section 2.1, a non-parametric model of the log-volatility can be obtained through GP regression with the K_{OU} covariance function, using as target the quantity $r_\tau - \frac{1}{2}\log \tau - 0.43$ measured for each interval over which the volatility can be assumed constant (e.g. one day), and taking as input a numerical time measurement (e.g. increasing the input by one for daily increments) at which each observation is recorded. Forecasting is carried out by querying the model at time inputs that extend beyond the end of estimation sample data.

5 Simulation analysis

To evaluate the effectiveness of the GP formulation to recover a known underlying stochastic volatility process, given only daily price highs and lows, and perform out-of-sample forecasting, we carry out a simulation analysis in the framework of Alizadeh, Brandt, and Diebold (2002). We first sample daily volatilities from eq. (5). Then, in order to obtain daily

log-range values, we sample the log-price process s intraday using a driftless discretization of (3) , i.e. on day i

$$s_t = s_{t-\tilde{\tau}} + \sigma_{i\tau}\sqrt{\tilde{\tau}}\,\epsilon_t^s \tag{26}$$

where $i\tau < t \leq (i+1)\tau$, $\{\epsilon_t^s\}$ are standard normal variates (independent of $\{\epsilon_{i\tau}^\sigma\}$), and $\tilde{\tau} = \tau/M$. This yields M log-prices per day, whose maximum and minimum values are used to compute $r_{i\tau}$.[15]

5.1 Results

5.1.1 Performance Measures

All performance measures introduced next are computed in the log-volatility domain. A measure of forecasting up to horizon H includes the *entire subtrajectory* for $h = 1, \ldots, H$. In other words, we are interested in modeling the complete future volatility trajectory rather than a point estimate at some future horizon. The mean-squared error (MSE) is computed between the true log-volatility (which is known from the simulated trajectories) and the model forecast. Let t be the time at which the forecast is made. Let $\log \sigma_{t+h}^j$ the true underlying process log-volatility for the j-th Monte Carlo draw, and $\log \hat{\sigma}_{t+h|t}^j$ the corresponding conditional expectation given by some model. The MSE at horizon H is defined as

$$\mathrm{MSE}_H^j = \frac{1}{H}(\Delta_{H|t}^j)'\Delta_{H|t}^j,$$

where $\Delta_{H|t}^j$ is a vector whose h-th element is given by $\left(\Delta_{H|t}^j\right)_h = \log \hat{\sigma}_{t+h|t}^j - \log \sigma_{t+h}^j$. The overall reported MSE at horizon H is the arithmetic average of all MSE_H^j across Monte Carlo draws. Whereas the MSE captures the ability of a model to correctly represent the conditional expectation, the *negative log-likelihood* (NLL) captures the ability of the model to assign greater probability density to likely events. Both the ABD and GP models considered here give rise to a normal joint predictive distribution of future log volatility. The NLL at horizon H for a trajectory j is

$$\mathrm{NLL}_H^j = \frac{1}{2}\left[(\Delta_{H|t}^j)'(\Sigma_{H|t}^j)^{-1}\Delta_{H|t}^j + \log\left|\Sigma_{H|t}^j\right| + H\log 2\pi\right], \tag{27}$$

where $\Delta_{H|t}^j$ is as previously, $\Sigma_{H|t}^j$ is the covariance matrix of the model forecasts up to horizon H. For the REGARCH model, we assume that (27) also holds, with the appropriate mo-

[15] As parameters for the log-volatility process, we took $\alpha = 3.855$, $\log\bar{\sigma} = -2.5$ and $\beta = 0.75$, yielding realistic volatility dynamics, largely consistent, for instance, with those observed on foreign exchange markets (Alizadeh, Brandt, and Diebold 2002). For intraday prices, we used $M = 5184$; see (Alizadeh, Brandt, and Diebold 2002) for a discussion on the choice of M.

ments estimated using Monte-Carlo integration. Finally, the *relative bias* evaluates whether forecasts are more likely to be wrong in one direction than another. The bias at horizon H for a trajectory j is

$$\text{Rel Bias}_H^j = \frac{1}{H}\sum_{h=1}^{H} \frac{\log \hat{\sigma}_{t+h|t}^j - \log \sigma_{t+h}^j}{|\log \sigma_{t+h}^j|}.$$

5.1.2 Recovering Underlying Volatility

An illustration of how well the models are capable of recovering an underlying known log-volatility process appears in Figure 2. Both the ABD and GP models exhibit nearly identical solutions, both on the inference of the underlying in-sample volatility and over the forecasting period. It should be emphasized that neither model has direct access to the underlying volatility process: on each day, they only "see" the high and low prices.

Largely the same remark applies for the REGARCH model. The volatility filtered by this model is not as smooth as the latent states extracted by the ABD and GP models. This is a natural consequence of the specification of all GARCH models, which implements filtering rather than smoothing of estimated volatilities over the sample period: since the one-step-ahead volatility, σ_{t+1}, is considered to be known at time t, the filtered state is not revised using the information arriving after time t, *even though this information is part of the estimation sample* and thus available to the econometrician.[16] In Figure 2, we see that the filtered nature of GARCH volatility yields a systematic lag with respect to the true underlying log-volatility.

Since the models cannot represent additional structure, the out-of-sample log-volatility trajectory has the shape of an exponential reversion to the long-run mean log-volatility. This limitation is addressed in Section 6, where a more structured covariance function is introduced to enable the GP to represent richer out-of-sample dynamics.

5.1.3 Detailed Performance Comparison

Table 1 shows detailed performance results, comparing the benchmark ABD and REGARCH models to the GP using the K_{OU} covariance function, averaged on 10,000 Monte Carlo draws. Results using both two and five years of in-sample data are given, and out-of-sample forecasting results at up to a one-month (21 trading days) horizon. We note that on

[16] In this context, the conditional variance of σ_{t+1} at time t is zero; in Figure 2, the uncertainty on the filtered log-volatility at time t is based on its standard deviation conditional on the information available at time $t-2$ as obtained by Monte Carlo integration.

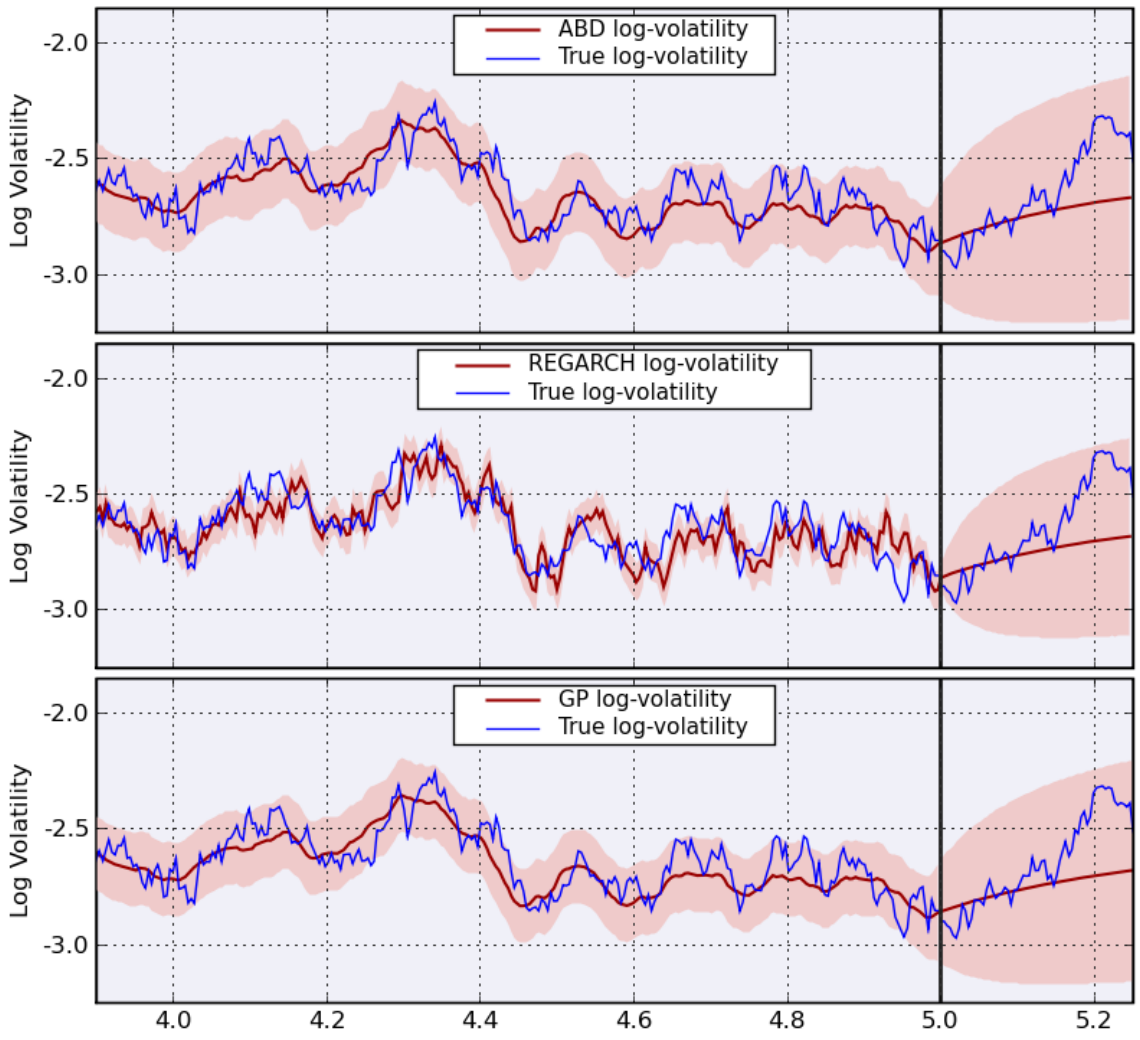

Figure 2: *Recovery of a known latent volatility process by the ABD, EGARCH and GP models. The last one of 5 in-sample years is shown, along with a 3-month forecast (beyond the thick vertical lines). Shaded regions enclose a 95% confidence interval.*

both the MSE and NLL criteria, the ABD and GP models perform comparably, although ABD appears a bit less biased at long horizons. On the other hand, ABD does win out on the out-of-sample NLL criterion, a fact that should not be too surprising in light of the fact that its parametric form corresponds exactly to the data generating process. However, overall, we can be satisfied that the accuracy of the GP compares well to that of the best possible parametric specification under the current simulation setting, even at horizons as long as one month ahead.

As for the REGARCH, the aforementioned lag between the filtered log-volatility and the underlying process leads to poor in-sample performance. Interestingly, this translates

in an out-of-sample MSE performance that is only moderately worse than those of the ABD and GP models. In terms of the magnitude of its forecasting bias, the REGARCH even compares favorably to the ABD and GP at short horizons. On the out-of-sample NLL criterion, the REGARCH is consistently and significantly outperformed by the (data-generating) ABD, but performs better than the GP model. It thus seems that the multivariate normal assumption used to compute the NLLis not too severely violated.

6 Equity modeling

Obviously, the true test of a forecasting methodology is whether it can capture and exploit empirical regularities found in real datasets. We now turn to forecasting the volatility of the S&P 500, a major U.S. stock index. Contrarily to the previous simulation results, a practical hurdle for comparing alternative models is that the "true" volatility of the index is unobserved; hence, it appears that there would be no ground truth forming a basis of comparison. To sidestep this issue, we use targets based on *realized volatility*, computed from high-frequency data, which have been recognized over the past decade as providing a statistically consistent estimate of the underlying volatility (Andersen and Benzoni 2008).

Realized volatility over a day is computed as the sum of high-frequency returns during the day. We used returns spaced at five-minute intervals and implement an adjustment for intraday autocorrelation given in Liu and Maheu (2008).[17] The realized volatility on day t is computed as

$$\mathrm{RV}_t = \sum_{i=1}^{M} r_{t,i}^2 + 2\sum_{h=1}^{q}\left(1 - \frac{h}{q+1}\right)\sum_{i=1}^{M-h} r_{t,i}\, r_{t,i+h}, \tag{28}$$

where $r_{t,i}$ is the i-th logarithmic return during day t, M is the number of returns during the day, and q is the maximum autocorrelation order to consider (we used $q = 1$ in our experiments). We use data from the SPDR S&P 500 ETF (ticker symbol SPY), an Exchange-Traded Fund that replicates the behavior of the S&P 500 index. Our raw data comes from NYSE Trades and Quotes (TAQ) database (New York Stock Exchange) and contains all SPY transactions spanning 2002 to 2006 (over 67M trades). The data was processed to extract prices at five-minute intervals, from which intraday returns are computed.[18]

[17] Due to market microstructure effects, such as price discretization and bid-ask bounces, it is not advisable to consider higher sampling frequencies unless further adjustments are applied; e.g. (Alizadeh, Brandt, and Diebold 2002).

[18] It must be emphasized that realized volatility is used in this work only to *evaluate performance*; it is not required to estimate the models. High-frequency data of the kind needed to compute realized volatility does

Table 1: *Simulation results comparing the baseline ABD model to the GP representation.*

		Estimation Sample: 2 years						Estimation Sample: 5 years					
		ABD		**REGARCH**		**GP**		**ABD**		**REGARCH**		**GP**	
In-sample MSE	**($\times 10^{-3}$)**	7.15	(0.11)	13.45	(0.20)	7.21	(0.11)	7.06	(0.06)	13.56	(0.14)	7.10	(0.06)
In-sample Bias	**($\times 10^{-4}$)**	-67.93	(5.29)	-80.76	(5.23)	-68.85	(5.24)	-66.71	(3.03)	-81.17	(3.11)	-67.73	(3.02)
Out-of-sample NLL	**H 1**	-0.75	(0.01)	2.52	(0.07)	-0.19	(0.00)	-0.78	(0.01)	1.68	(0.05)	-0.19	(0.00)
	H 5	-7.06	(0.02)	1.43	(0.15)	-0.97	(0.00)	-7.24	(0.02)	-0.65	(0.11)	-0.97	(0.00)
	H 10	-14.94	(0.03)	-5.97	(0.16)	-1.97	(0.01)	-15.31	(0.02)	-8.51	(0.11)	-1.97	(0.01)
	H 21	-32.33	(0.05)	-22.36	(0.18)	-4.19	(0.01)	-33.09	(0.03)	-25.81	(0.12)	-4.20	(0.01)
Out-of-sample MSE	**H 1**	12.52	(0.15)	13.08	(0.16)	12.41	(0.15)	12.17	(0.15)	12.36	(0.15)	12.17	(0.15)
($\times 10^{-3}$)	**H 5**	16.54	(0.19)	17.15	(0.20)	16.40	(0.19)	16.00	(0.18)	16.22	(0.19)	15.94	(0.18)
	H 10	21.35	(0.25)	21.99	(0.26)	21.21	(0.25)	20.53	(0.23)	20.78	(0.24)	20.44	(0.23)
	H 21	30.26	(0.34)	30.99	(0.35)	30.15	(0.34)	28.91	(0.32)	29.23	(0.33)	28.78	(0.32)
Out-of-sample Bias	**H 1**	13.93	(4.40)	0.48	(4.49)	15.11	(4.37)	17.31	(4.32)	9.60	(4.35)	18.45	(4.32)
($\times 10^{-4}$)	**H 5**	7.89	(4.78)	-4.23	(4.87)	10.11	(4.75)	11.01	(4.69)	4.19	(4.72)	11.81	(4.67)
	H 10	-2.22	(5.25)	-15.67	(5.34)	1.11	(5.21)	1.01	(5.14)	-7.65	(5.17)	1.29	(5.11)
	H 21	-18.59	(6.02)	-35.51	(6.11)	-13.38	(5.99)	-14.02	(5.87)	-27.05	(5.91)	-14.93	(5.83)

A total of 10,000 Monte Carlo draws are used. Model estimation is performed using both 2 and 5 years of past observations. Out-of-sample forecasting performance is evaluated at various horizons, given in days. Standard errors are given in parentheses. Both models perform comparably on the MSE and relative bias criteria, although ABD shows an advantage in out-of-sample NLL.

6.1 Adding explanatory variables

By treating volatility as a regression problem, the Gaussian process formulation allows the incorporation of a number of explanatory variables, a task difficult with the Kalman filter used by traditional stochastic volatility models (including our benchmark ABD model). These variables are important because they allow for exogenous factors driving volatility, including seasonalities and macroeconomic conditions. *Seasonality variables* are computed deterministically from a day t (assumed to be numbered sequentially, with one calendar day separating t and $t+1$). They are included in the model denoted "GP (+seasonalities)". The following variables are considered:

- Day of week, represented as a discrete integer: $\text{DOW}(t) = t \mod 7$.[19]
- Relative position within the current month, represented as a pair on the unit circle.[20] Given a day t, let t_{bm} and t_{nm} respectively denote the start of the current month and the start of the next month. Let $f_{t,m} = (t - t_{bm})/(t_{nm} - t_{bm})$ represent the relative position (between 0 and 1) since the start of the month. The monthly seasonal variable is the pair $M(t) = \Big(\cos(2\pi f_{t,m}), \sin(2\pi f_{t,m})\Big)$.
- Relative positions within the current quarter and the current year, denoted respectively by $Q(t)$ and $Y(t)$; they are computed exactly in the same manner as for $M(t)$.

In addition to seasonalities, it is plausible to anticipate *market expectations and business cycle conditions* to be predictive of realized volatility. We consider the following variables: the CBOE VIX index of implied volatility (Exchange 2009), including both the current level of the VIX and its logarithmic return over the previous month. We also incorporate the recently-proposed ADS real-time indicator of business conditions (Aruoba, Diebold, and Scotti 2009), a scalar variable summarizing the state of a large number of macroeconomic time series. For out-of-sample forecasting purposes, these variables are held at their last-known value at the time of making the forecast.[21] These variables, in addition to the seasonals described above, are included in the model denoted "GP (+macro/fin)".

---

not constitute a panacea: it remains available only for a few highly-developed markets, does not have a long history, and requires a substantial data management infrastructure.

[19] Note that a "one-hot" encoding is not used here; instead, as explained below, a special term in the covariance function handles this variable.

[20] The motivation for this representation is given in Section 6.2.

[21] A more sophisticated approach could attempt to independently model these variables, or use a time-representation split analogous to that used by Chapados and Bengio (2008).

6.2 Covariance function engineering

One of the most attractive features of GPs is their ability to incorporate domain knowledge in a principled manner via the specification of a problem-specific covariance function. This is how we incorporate the supplementary variables described above. Assuming that the input is given by a day t and macro/financial variables $\mathbf{z}$, the covariance function used for the GP (+macro/fin) model is

$$K_{\text{MF}}\Big((t,\mathbf{z}),(t',\mathbf{z}')\Big) = \sigma_t^2 K_{\text{OU}}(t,t') + \sigma_{\text{DOW}}^2\delta(\text{DOW}(t),\text{DOW}(t')) + \sigma_M^2 K_{\text{SE}}(M(t),M(t')) \\ + \sigma_Q^2 K_{\text{SE}}(Q(t),Q(t')) + \sigma_Y^2 K_{\text{SE}}(Y(t),Y(t')) + \sigma_{\mathbf{z}}^2 K_{\text{SE}}(\mathbf{z},\mathbf{z}'), \tag{29}$$

where the squared-exponential covariance is given by $K_{\text{SE}}(\mathbf{x},\mathbf{x}') = \exp -\frac{1}{2}\|(\mathbf{x}-\mathbf{x}')\|^2/\sigma_\ell^2$ with σ_ℓ^2 a length-scale hyperparameter,[22] all $\sigma_\bullet^2$ are hyperparameters to be estimated by maximizing marginal likelihood (*cf.* (21)), and $\delta(\cdot,\cdot)$ is the Kronecker delta; the impact of this term, for days-of-week, is to introduce an additional discrete similarity between two elements that fall on the same day of the week. This construction of the covariance function draws from a time-series modeling analysis in (Rasmussen and Williams 2006) and allows different problem dimensions to compose independently. For the GP (+seasonalities) model, the covariance function is identical except that the last term is omitted.

The unit-circle representation for seasonalities suppresses disruptions due to end-of-period effects and ensures that there is a constant distance between consecutive days. To see how this arises in conjunction with the K_{SE} covariance, consider the expansion for the monthly seasonalities of two dates t and t' (assuming, for simplicity, unit length-scale, and where $\theta = 2\pi f_{t,m}$),

$$K_{\text{SE}}(M(t),M(t')) = \exp\Big[-\frac{1}{2}\Big((\cos(\theta)-\cos(\theta'))^2 + (\sin(\theta)-\sin(\theta'))^2\Big)\Big] = \exp\Big[\cos(\theta-\theta')-1\Big],$$

where the second equality follows from basic trigonometric identities. From this, it is clear that the induced similarity measure is periodic and depends only on the angular difference between relative locations within a month.

6.3 Results

The same performance metrics as those of Section 5.1 are employed. In the present case, the target is the realized volatility computed as per (28), with out-of-sample data extending from January 2002 to December 2006. The models are first estimated using five years

[22] Each K_{SE} term in (29) has a separate length-scale hyperparameter.

of SPY data from December 27, 1996, to December 31, 2001. The models' forecasting performance is assessed on a 21-day out-of-sample horizon extending beyond the end of the in-sample data. Then, the in-sample period is rolled one week forward, models are re-estimated, and new out-of-sample performance measures are computed. The results in Table 2 report the average of these measures for the 257 (weekly) iterations of this procedure.

Table 2 compares the ABD and REGARCH models to GPs utilizing progressively richer inputs; Figure 3 provides a more detailed picture of MSE and relative bias results. The GP (time-only) model, which is provided with exactly the same information as the ABD model, outperforms the latter on both the MSE and NLL criteria, especially at long horizons, indicating that the GP's non-parametric representation captures significant features that are missed by the ABD functional form.

On the MSE criterion, the REGARCH model outperforms the GP (time-only) model. It is worth mentioning that the REGARCH model allows for leverage effect and is provided information about log-returns, in addition to the ranges provided to the other models. The MSE results obtained here thus confirm the importance of allowing for leverage, strongly suggesting that a future version of the GP model should incorporate this feature. Considering out-of-sample NLL performance, the REGARCH model is dramatically outperformed by both the ABD and the simplest GP model. These results could seen as a consequence of a severe violation of the assumed multivariate normal distribution for the log-volatilities forecasted by the REGARCH; however, this interpretation seems to be dismissed by the relatively good performance of the REGARCH model in terms of the out-of-sample NLL obtained on simulated data. Another interpretation is that, while the REGARCH properly forecast trends in volatility (as per the MSE criterion), it improperly describes the covariance structure of potential future volatility paths.

The use of seasonal (GP+seasonalities) and macroeconomic/financial (GP+macro/fin) variables make the difference between the GP and the ABD models even more dramatic; both GP models improving upon ABD and GP (time-only) at all horizons and all performance measures. These latter flavors of the GP model also outperform the REGARCH model in all regards. This strongly suggests that equity market volatility is characterized by non-trivial seasonalities and co-dependencies on exogenous economic and financial drivers.

Table 2: *Volatility forecasting of the S&P 500 index, using the daily realized volatility as a benchmark.*

		ABD		REGARCH		GP (time-only)		GP (+season.)		GP (+macro/fin)	
Out-of-sample NLL	**H 1**	1.18	(0.19)	19.68	(2.16)	0.34	(0.04)	0.27	(0.03)	0.18	(0.03)
	H 5	20.59	(1.79)	97.36	(6.25)	1.14	(0.07)	1.06	(0.07)	0.90	(0.08)
	H 10	44.96	(3.18)	203.54	(10.42)	2.16	(0.11)	2.04	(0.11)	1.87	(0.13)
	H 21	100.18	(5.82)	443.09	(17.95)	4.42	(0.15)	4.21	(0.16)	4.02	(0.20)
Out-of-sample MSE	**H 1**	112.56	(10.27)	98.07	(8.54)	111.61	(10.28)	92.68	(9.40)	75.09	(7.29)
($\times 10^{-3}$)	**H 5**	104.60	(5.75)	92.89	(4.98)	99.27	(4.51)	92.07	(4.23)	82.12	(3.88)
	H 10	119.44	(5.60)	104.26	(4.90)	109.75	(3.50)	100.74	(3.23)	90.52	(2.98)
	H 21	147.66	(6.00)	122.91	(5.04)	129.90	(2.80)	116.39	(2.54)	101.87	(2.30)
Out-of-sample Bias	**H 1**	614.02	(92.07)	632.07	(77.84)	618.68	(9.23)	366.75	(9.32)	231.34	(8.74)
($\times 10^{-4}$)	**H 5**	414.92	(73.62)	406.86	(58.42)	409.23	(4.16)	311.23	(4.06)	250.69	(3.90)
	H 10	442.54	(77.79)	400.87	(61.73)	421.45	(3.24)	297.11	(3.15)	237.35	(2.95)
	H 21	527.49	(84.14)	415.99	(69.22)	465.58	(2.52)	279.60	(2.47)	217.98	(2.26)

This table compares, at various out-of-sample forecasting horizons (in days), the baseline ABD model against (i) the GP using a single time input, (ii) the GP adding seasonality-dependent inputs, (iii) the GP adding macroeconomic and financial inputs to the seasonals. Standard errors are in parentheses. We note that the GP is capable of exploiting additional inputs to yield significantly improved performance against the baseline ABD model.

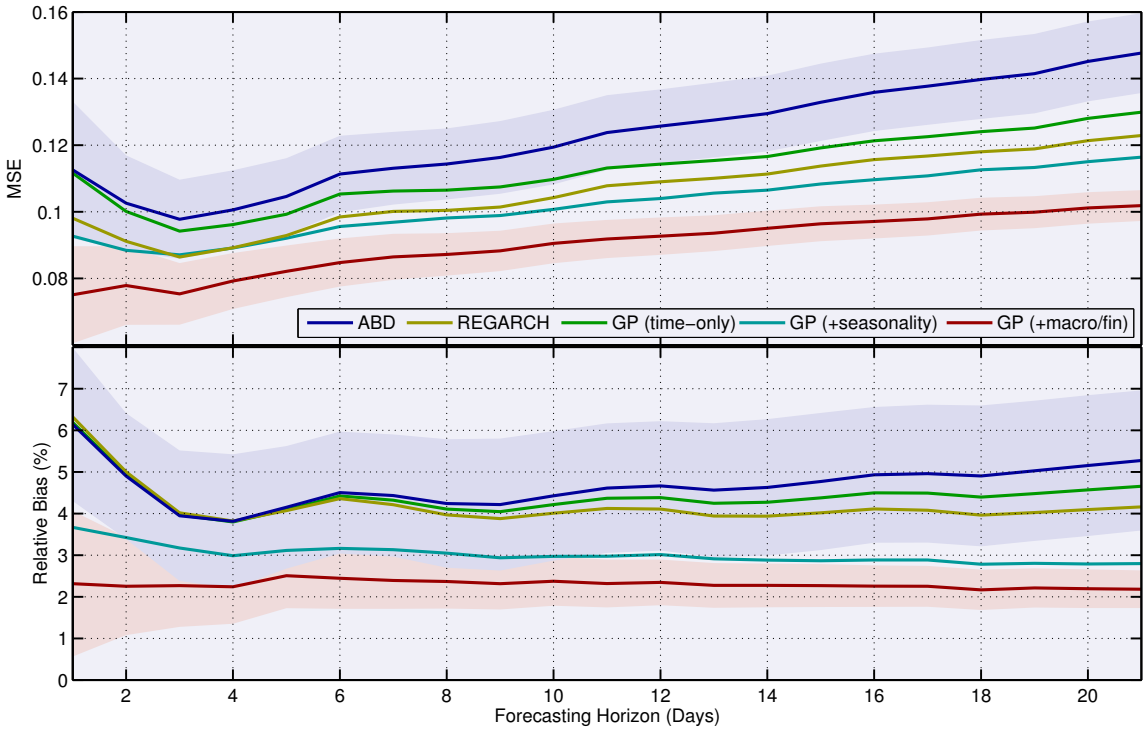

Figure 3: *MSE and relative bias for realized volatility forecasting as a function of the horizon. The shaded regions (shown only for the benchmark ABD and the GP (+macro/fin) models) enclose a 95% confidence interval. The GP with the full set of inputs significantly outperforms the benchmark model at most horizons, on both MSE and bias criteria.*

6.4 Long-horizon forecasting with seasonalities

A tantalizing application of the methodology proposed herein lies in long-horizon volatility forecasting, extending one year ahead or further, taking advantage of seasonal modeling to improve performance. An illustration of this possibility is shown in Figure 4, which plots a one-year out-of-sample forecast of SPY volatility performed from January 2005 (beyond the thick vertical bar) under the GP+seasonalities model. The realized volatility over the out-of-sample period is also plotted (in magenta). Visual inspection shows that periods of increased volatility during spring and fall are captured by the model, although a false positive was also raised during the summer. The high-frequency "spikes" are day-of-week effects. For reference, the estimated volatility during the in-sample period is also plotted, as well as the daily SPY closing prices and returns (top and middle panes). It must be noted that standard volatility forecasting models in finance would quickly revert to an unconditional volatility, lacking the ability to handle seasonalities. It is obvious that a more comprehensive investigation is required to qualify the approach, although the ease with which the GP can be applied to this task makes this a promising research avenue.

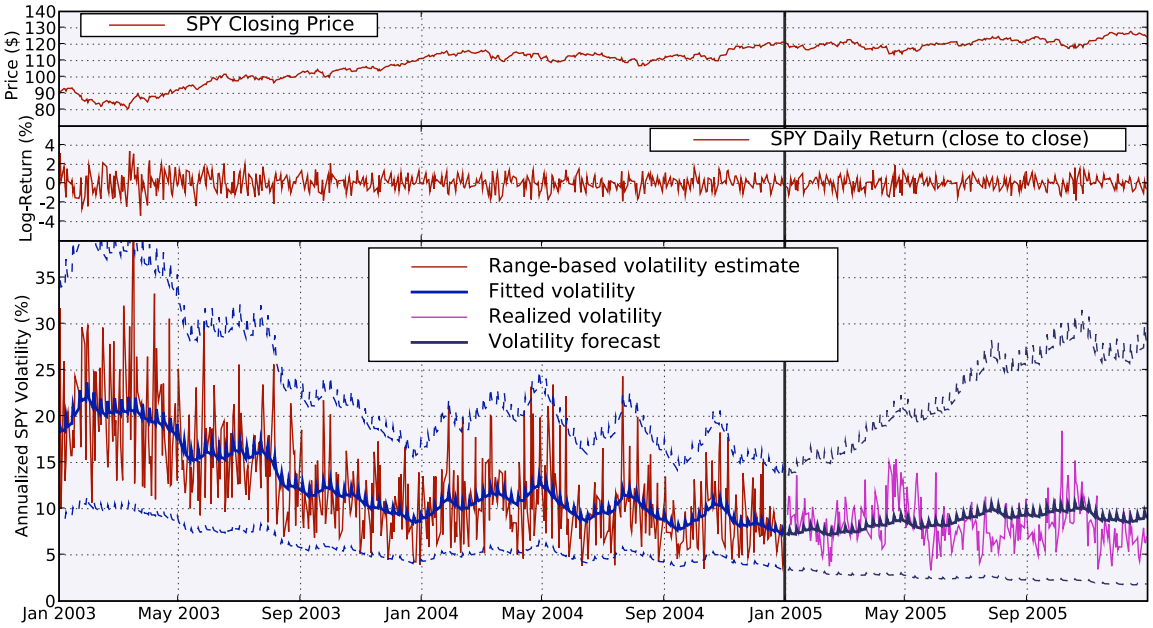

Figure 4: *Impact of seasonalities in long-range forecasting of volatility for a stock price index. In this illustration, a one-year out-of-sample forecast is performed, starting in January 2005. Seasonal effects are clearly visible, and approximately match periods of greater realized volatility during the out-of-sample period except in summer 2005. Confidence bands enclose a 95% region around the observed range-based (for the in-sample period) and realized (for the out-of-sample period) volatilities.*

7 Conclusion

In this paper, we established a link between Gaussian processes and standard stochastic volatility models. A simulation analysis showed Gaussian processes to be comparable to benchmark models at capturing a known latent volatility process and at forecasting its behavior out of sample. On actual stock market data, Gaussian processes improved on the performance of these benchmarks and were able to draw on seasonality and macroeconomic and financial inputs to further improve their forecasts through the use of an application-specific covariance function.

Gaussian processes show promise for long-range volatility forecasting, an ability that could prove highly relevant in pricing long-term derivative contracts and assessing the long-term risk exposures of complex portfolios. A more finely-engineered covariance function could prove successful at better modeling long-run volatility, an issue of interest given the growing consensus in the financial literature that volatility dynamics are driven by both transitory and high-persistence shocks (Andersen, Bollerslev, Diebold, and Ebens 2001; Engle and Lee 1999; Engle and Rangel 2008).

References

Alizadeh, S., M. W. Brandt, and F. X. Diebold (2002, June). Range-Based Estimation of Stochastic Volatility Models. *Journal of Finance 57*(3), 1047–1091.

Andersen, T. G. and L. Benzoni (2008, July). Realized Volatility. FRB of Chicago Working Paper No. 2008-14. Available at `http://ssrn.com/abstract=1092203`.

Andersen, T. G. and T. Bollerslev (1998). Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. *The Journal of Finance 53*(1), 219–265.

Andersen, T. G., T. Bollerslev, P. F. Christoffersen, and F. X. Diebold (2006). Volatility and Correlation Forecasting. In G. ELLIOTT, C. W. J. GRANGER, and A. TIMMERMANN (Eds.), *Handbook of Economic Forecasting*, Volume 1, pp. 777–878. North-Holland.

Andersen, T. G., T. Bollerslev, F. X. Diebold, and H. Ebens (2001). The Distribution of Realized Stock Return Volatility. *Journal of Financial Economics 61*, 43–76.

Aruoba, S., F. Diebold, and C. Scotti (2009). Real-Time Measurement of Business Conditions. *Journal of Business and Economic Statistics (Forthcoming)*.

Bertsekas, D. P. (2000). *Nonlinear Programming* (Second ed.). Belmont, MA: Athena Scientific.

Bollerslev, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics 31*, 307–327.

Brandt, M. and C. Jones (2006). Volatility Forecasting With Range-Based EGARCH Models. *Journal of Business and Economic Statistics 24*(4), 470–486.

Brandt, M. W. and C. S. Jones (2005). Bayesian Range-Based Estimation of Stochastic Volatility Models. *Finance Research Letters 2*, 201–209.

Chapados, N. and Y. Bengio (2008). Augmented Functional Time Series Representation and Forecasting with Gaussian Processes. In J. C. PLATT, D. KOLLER, Y. SINGER, and S. ROWEIS (Eds.), *Advances in Neural Information Processing Systems 20*, Cambridge, MA, pp. 265–272. MIT Press.

Christoffersen, P. and F. Diebold (2000). How Relevant Is Volatility Forecasting for Financial Risk Management? *Review of Economics and Statistics 82*(1), 12–22.

Corradi, V., W. Distaso, and A. Mele (2009). Macroeconomic Determinants of Stock Market Returns, Volatility and Volatility Risk-Premia. Working Paper.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. John Wiley & Sons.

Dorion, C. (2010). Business Conditions, Market Volatility and Options Prices. HEC Montreal Working Paper.

Duffie, D. and K. Singleton (1993). Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica 61*, 929–952.

Engle, R., E. Ghysels, and B. Sohn (2008). On the Economic Sources of Stock Market Volatility. Working Paper.

Engle, R. F. (1982). Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of UK Inflation. *Econometrica 50*, 987–1008.

Engle, R. F. and G. Lee (1999). A Permanent and Transitory Component Model of Stock Return Volatility. In R. ENGLE and H. WHITE (Eds.), *Cointegration, Causality, and Forecasting: a Festschrift in Honor of Clive W.J. Granger*, New York, pp. 475–497. Oxford University Press.

Engle, R. F. and J. G. Rangel (2008, February). The Spline-GARCH Model for Low-Frequency Volatility and Its Global Macroeconomic Causes. *Review of Financial Studies (Forthcoming)*.

Exchange, C. B. O. (2009). The CBOE Volatility Index – VIX white paper. Available at `http://www.cboe.com`.

Flannery, M. J. and A. A. Protopapadakis (2002). Macroeconomic Factors *Do* Influence Aggregate Stock Returns. *Review of Financial Studies 15*, 751–782.

Gallant, A. R., D. A. Hsieh, and G. E. Tauchen (1997). Estimation of Stochastic Volatility Models with Diagnostics. *Journal of Econometrics 81*, 159–192.

Goldberg, P. W., C. K. I. Williams, and C. M. Bishop (1998). Regression with Input-Dependent Noise: a Gaussian Process Treatment. In *Advances in Neural Information Processing systems 10*, Cambridge, MA, pp. 493–499. MIT Press.

Hamilton (1994). *Time Series Analysis*. Princeton University Press.

Harvey, A., E. Ruiz, and N. Shephard (1994). Multivariate Stochastic Variance Models. *Review of Economic Studies 61*, 247–264.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994, October). Bayesian Analysis of Stochastic Volatility Models. *Journal of Business and Economic Statistics 12*(4), 69–87.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering 82*(Series D), 35–45.

Kersting, K., C. Plagemann, P. Pfaff, and W. Burgard (2007). Most Likely Heteroscedastic Gaussian Process Regression. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, New York, NY, pp. 393–400. ACM.

Kim, S., N. Shephard, and S. Chib (1998). Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Review of Economic Studies 65*, 361–393.

Le, Q. V., A. J. Smola, and S. Canu (2005). Heteroscedastic Gaussian process regression. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, New York, NY, pp. 489–496. ACM.

Liu, C. and J. M. Maheu (2008). Are There Structural Breaks in Realized Volatility? *Journal of Financial Econometrics 6*(3), 326–360.

MacKay, D. J. C. (1999). Comparison of Approximate Methods for Handling Hyperparameters. *Neural Computation 11*(5), 1035–1068.

Martens, M., D. Van Dijk, and M. De Pooter (2009). Forecasting S&P 500 volatility: Long memory, level shifts, leverage effects, day-of-the-week seasonality, and macroeconomic announcements. *International Journal of Forecasting 25*(2), 282–303.

Matheron, G. (1973). The Intrinsic Random Functions and Their Applications. *Advances in Applied Probability 5*, 439–468.

New York Stock Exchange. Trades and Quotes Database. Available at `http://www.nyxdata.com`.

Nelson, D. B. (1991). Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica 59*, 347–370.

O'Hagan, A. (1978). Curve Fitting and Optimal Design for Prediction. *Journal of the Royal Statistical Society B 40*, 1–42. (With discussion).

Parkinson, M. (1980). The Extreme Value Method for Estimating the Variance of the Rate of Return. *Journal of Business 53*(1), 61–65.

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

Shephard, N. (Ed.) (2004). *Stochastic Volatility: Selected Readings*. Oxford, UK: Oxford University Press.

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. New York, NY: Springer.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, B 36*(1), 111–147.

Sundararajan, S. and S. S. Keerthi (2001). Predictive Approaches for Choosing Hyperparameters in Gaussian Processes. *Neural Computation 13*(5), 1103–1118.

Taylor, S. J. (1994). Modelling Stochastic Volatility. *Mathematical Finance 4*, 183–204.

West, K. D. and D. Cho (1995). The Predictive Ability of Several Models of Exchange Rate Volatility. *Journal of Econometrics 69*, 367–391.

Williams, C. K. I. and C. E. Rasmussen (1996). Gaussian Processes for Regression. In D. S. TOURETZKY, M. C. MOZER, and M. E. HASSELMO (Eds.), *Advances in Neural Information Processing Systems 8*, pp. 514–520. MIT Press.