Kernel ridge regression is a technique to perform ridge regression with a potentially infinite number of nonlinear transformations of the independent variables as regressors. This makes it a powerful forecasting tool, which is applicable in many different contexts. However, it is usually applied only to independent and identically distributed observations.

This paper introduces a variant of kernel ridge regression for time series with stochastic volatility. The conditional mean and volatility are both modelled as nonlinear functions of observed variables. We set up the estimation problem in a Bayesian manner and derive a Gibbs sampler to obtain draws from the predictive distribution. A simulation study and an application to forecasting the distribution of returns on the S&P500 index are presented, and we find that our method outperforms most popular GARCH variants in terms of one-day-ahead predictive ability. Notably, most of this improvement comes from a more adequate approximation to the tails of the distribution.

**Keywords:** Nonlinear forecasting, shrinkage estimation, kernel methods, stochastic volatility.
1 Introduction

The measurement and forecasting of volatilities, especially on financial markets, has attracted much attention over the last decades. After the introduction of ARCH (Engle 1982) and GARCH (Bollerslev 1986), many variants of these methods have been proposed. Hansen and Lunde (2005) give an overview, but they show that in forecasting, none of these variants outperform the simple GARCH(1,1) model. Another strand of literature (Jacquier et al. 2002, among many others) studies stochastic volatility models.

An important drawback of all of these models is the rigid functional form that needs to be assumed for the volatility equation. As the volatility process is by definition unobserved, a parsimonious forecasting model is often desired, so that restrictive assumptions need to be made. For the same reason, it is uncommon to include many lags, or even exogenous predictors for the volatility, in the model.

In this article, we propose a stochastic volatility model that allows for both flexible functional forms and a larger number of predictors, while avoiding overfitting and maintaining computational tractability. Our starting point is the technique of kernel ridge regression. The basic idea is to nonlinearly map the predictors into an infinite-dimensional space, and to perform linear regression with the transformed predictors as regressors. If the mapping from predictors into regressors is chosen carefully, estimation and forecasting can be carried out efficiently, without working in the regressor space explicitly. For a more detailed introduction to kernel ridge regression, see Hofmann et al. (2008).

Kernel ridge regression was developed in the machine learning literature, where independent and identically distributed data are the norm. A recent adaptation to homoscedastic time series, with an economic application, is presented by Exterkate et al. (2013). To our knowledge, the present paper is the first to introduce an extension to time-varying volatility. We modify a standard stochastic volatility model to allow for nonlinear regression equations for both the mean and the log-volatility of the observed process. To estimate this model, we take a Bayesian approach and construct a Gibbs sampler. The steps where regression coefficients are drawn come down to a modified version of kernel ridge regression.

Our Monte Carlo study shows the applicability of the proposed KRR-SV method to a wide range of models, including the GARCH framework. An empirical application to forecasting the returns distribution on the S&P500 index shows that our method outperforms most popular GARCH variants in terms of the Kolmogorov-Smirnov test. Moreover, we find that KRR-SV places more mass in the left tail of the predictive returns distribution before large crashes, which is desirable for risk management.
We introduce our model in Section 2. We show how a linear version of it can be estimated using a Gibbs sampler, and then use kernel ridge regression techniques to adapt the Gibbs sampler to the general nonlinear model. The simulation study and the empirical application are discussed in Sections 3 and 4, respectively, and we provide conclusions and discuss possible extensions in Section 5.

2 Methodology

We describe the model that we use in Section 2.1. As this is a model with a latent volatility process, it lends itself well to estimation using a Gibbs sampler with data augmentation. The required Gibbs steps are derived in Section 2.2. However, this derivation presumes that the model is linear. We handle nonlinearity through the use of kernel ridge regression, which we briefly introduce in Section 2.3. Section 2.4 contains the derivation of an adapted version of the Gibbs sampler for nonlinear models. Finally, in Section 2.5 we discuss the hyperparameters of the model.

2.1 Notation and setup of the model

The goal is to forecast the distribution of a future value of a process, $y_{T+1} \in \mathbb{R}$, and we have historical observations on it in the $T \times 1$ vector $y$. In addition, we are given a $T \times N$ matrix $X$ containing explanatory variables for the mean of $y$, and a $T \times M$ matrix $Z$ with explanatory variables for the volatility of $y$. For forecasting purposes, $x_{T+1} \in \mathbb{R}^N$ and $z_{T+1} \in \mathbb{R}^M$ are also available.

We assume that the conditional model for $y_t$ is normal, for each $t = 1, \ldots, T + 1$:

$$y_t | \eta_t, \beta, \tau \sim N \left( \phi \left( x_t \right)' \beta, \tau^{-1} \exp(2\eta_t) \right).$$ (1)

Here, $x_t'$ is the $t$-th row of $X$ and $\eta_t$ is the unobserved time-varying log-volatility. The mapping $\phi$ transforms the predictors into an infinite-dimensional set of regressors; this is what makes the model nonlinear. For the sake of exposition, we shall first discuss a linear model, with $\phi$ the identity mapping. The nonlinear extension is studied in Sections 2.3 and 2.4.

The conditional model for $\eta_t$ is also normal:

$$\eta_t | \gamma \sim N \left( \psi \left( z_t \right)' \gamma, 1 \right),$$ (2)
where again, \( \psi \) is a possibly nonlinear mapping. Note that the model in (1) and (2) is not as restrictive as it may seem. Although the conditional distribution of \( y_t \) is normal, the fact that its volatility is stochastic makes it flexible enough to approximate a wide variety of processes, including the heavy-tailed distributions commonly associated with financial returns; see e.g. Jacquier et al. (2002). Moreover, we will perform Bayesian model averaging over hyperparameters of the model (see Section 2.5), resulting in a predictive distribution that is a mixture of normals.

Although Model (1)-(2) does not explicitly include autoregressive effects, there is nothing that prevents \( x_t \) and \( z_t \) from including lags of the dependent variable, \( y_{t-1}, \ldots, y_{t-P} \). The derivations below remain valid if such lags are present, provided that we condition on pre-sample values \( y_0, y_{-1}, \ldots, y_{1-P} \). The justification for this statement comes from the following decomposition:

\[
\begin{align*}
p(y \mid y_{1-P}, \ldots, y_0) &= p(y_1 \mid y_{1-P}, \ldots, y_0) p(y_2 \mid y_{1-P} \ldots, y_1) \cdots p(y_T \mid y_{1-P}, \ldots, y_{T-1}) \\
&= p(y_1 \mid y_{1-P}, \ldots, y_0) p(y_2 \mid y_{2-P} \ldots, y_1) \cdots p(y_T \mid y_{T-P}, \ldots, y_{T-1}).
\end{align*}
\]

Thus, each row of \( X \) or \( Z \) contains the correct conditioning information.

To complete the model, we use a standard set of priors. The prior on the precision parameter is uninformative, \( p(\tau) \propto \tau^{-1} \), and the priors on the regression coefficients are normal,

\[
\beta \mid \tau \sim N \left( 0, (\lambda \tau)^{-1} I \right) \quad \text{and} \quad \gamma \sim N \left( 0, \mu^{-1} I \right),
\]

where \( \lambda \) and \( \mu \) are hyperparameters, and \( I \) denotes an identity matrix of the appropriate dimensions.

### 2.2 Gibbs sampler for linear models

We estimate the model in Section 2.1 using a Gibbs sampler with data augmentation (Geman and Geman, 1984; Tanner and Wong, 1987). In this section, we abstract from the mappings \( \varphi \) and \( \psi \); thus, writing \( \Sigma = \text{diag} \left( \exp(2\eta) \right) \), equations (1) and (2) can be simplified to

\[
\begin{align*}
p(y \mid \eta, \beta, \tau, \gamma) &= N \left( X \beta, \tau^{-1} \Sigma \right) \quad \propto \quad \tau^{T/2} \exp \left( -\iota' \eta \right) \exp \left( -\frac{\tau}{2} (y - X \beta)' \Sigma^{-1} (y - X \beta) \right) \\
\text{and} \quad p(\eta \mid \beta, \tau, \gamma) &= N \left( Z \gamma, I \right) \quad \propto \quad \exp \left( -\frac{1}{2} (\eta - Z \gamma)' (\eta - Z \gamma) \right).
\end{align*}
\]

Here, \( \iota \) is a \( T \times 1 \) vector of ones, so that \( \iota' \eta = \sum_{t=1}^{T} \eta_t \).

In each step of the Gibbs sampler, we need to draw from the conditional posterior distributions
Thus, for each \( t \) of \( \eta \) where \( \gamma \) from Damien et al. (1999). Introduce a variable \( v \). As we are conditioning on \( \eta \), we are now dealing with two ordinary linear regression models and we can use standard techniques to obtain the correct sampling distributions (see e.g. [Koop, 2003]):

- Draw \( \tau \) from a gamma distribution with shape parameter \( T/2 \) and scale parameter \( 2 \left( y' \left( \Sigma^{-1} - \Sigma^{-1} X (X'\Sigma^{-1}X + \lambda I)^{-1} X'\Sigma^{-1} \right) y \right)^{-1} \).
- Draw \( \beta \) from a normal distribution with mean \( (X'\Sigma^{-1}X + \lambda I)^{-1} X'\Sigma^{-1}y \) and variance matrix \( \tau^{-1} (X'\Sigma^{-1}X + \lambda I)^{-1} \).
- Draw \( \gamma \) from a normal distribution with mean \( (Z'Z + \mu I)^{-1} Z'\eta \) and variance matrix \( (Z'Z + \mu I)^{-1} \).

Drawing the latent log-volatilities \( \eta \) is somewhat more involved. We have

\[
p(\eta \mid \beta, \tau, \gamma, y) \propto p(y \mid \eta, \beta, \tau, \gamma) p(\eta \mid \beta, \tau, \gamma) \
\propto \tau^{T/2} \exp \left( -\frac{1}{2} \eta' \eta - \frac{1}{2} (y - X\beta)' \Sigma^{-1} (y - X\beta) - \frac{1}{2} (\eta - Z\gamma)' (\eta - Z\gamma) \right) \
\propto \exp \left( -\frac{1}{2} \eta' \eta + (Z\gamma - \tau)' \eta - \frac{1}{2} (y - X\beta)' \Sigma^{-1} (y - X\beta) \right) \
= \prod_{t=1}^{T} \exp \left( -\frac{1}{2} \eta_t^2 + (z_t'\gamma - 1) \eta_t - \frac{1}{2} (y_t - x_t'\beta)^2 \exp (-2\eta_t) \right) \
\propto \prod_{t=1}^{T} \exp \left( -\frac{1}{2} (\eta_t - z_t'\gamma + 1)^2 - \frac{1}{2} (y_t - x_t'\beta)^2 \exp (-2\eta_t) \right). \]

Thus, we see that we can draw each \( \eta_t \) individually. To obtain such a draw, we implement a technique from [Damien et al., 1999]. Introduce a variable \( v_t \) that has the following joint distribution with \( \eta_t \):

\[
p(\eta_t, v_t \mid \beta, \tau, \gamma, y) \propto I \left\{ v_t > \frac{1}{2} (y_t - x_t'\beta)^2 \exp (-2\eta_t) \right\} \exp (-v_t) \exp \left( -\frac{1}{2} (\eta_t - z_t'\gamma + 1)^2 \right). \quad (3)
\]

where \( I \{ \cdot \} \) is the indicator function. It can be checked that (3) leads to the desired marginal distribution of \( \eta_t \), and moreover, the conditional distributions are relatively simple:

\[
p(v_t \mid \eta_t, \beta, \tau, \gamma, y) \propto I \left\{ v_t > \frac{1}{2} (y_t - x_t'\beta)^2 \exp (-2\eta_t) \right\} \exp (-v_t) \\
\text{and} \quad p(\eta_t \mid v_t, \beta, \tau, \gamma, y) \propto I \left\{ \eta_t > \frac{1}{2} \log \left( \frac{y_t - x_t'\beta}{2v_t} \right) \right\} \exp \left( -\frac{1}{2} (\eta_t - z_t'\gamma + 1)^2 \right).
\]

Thus, for each \( t = 1, 2, \ldots, T \), we obtain the sampling distributions as follows:

- Let \( v_t \) be \( \frac{1}{2} (y_t - x_t'\beta)^2 \exp (-2\eta_t) \) plus a draw from the exponential distribution with mean one.
- Draw \( \eta_t \) from the \( N \left( z_t'\gamma - 1, 1 \right) \) distribution, truncated to the interval \( \left( \frac{1}{2} \log \left( \frac{y_t - x_t'\beta}{2v_t} \right), \infty \right) \).
Ultimately, we are interested in draws from the predictive distribution $p(y_{T+1} | y)$. These draws can be obtained within the Gibbs sampler, so that it is sufficient to sample from conditional posteriors:

- Draw $\eta_{T+1}$ from the normal distribution with mean $z_T' \gamma$ and variance 1.
- Draw $y_{T+1}$ from the normal distribution with mean $x_T' \beta$ and variance $\tau^{-1} \exp (2 \eta_{T+1})$.

### 2.3 Kernel ridge regression

The derivations in the preceding section were made under the assumption that both the mean and the log-volatility of the dependent variable are accurately described by linear regression models. That is, $\varphi$ and $\psi$ in equations (1)-(2) were restricted to be identity mappings. In this section, we discuss how this restriction can be relaxed.

Simply replacing $x_t$ by $\varphi(x_t)$ and $z_t$ by $\psi(z_t)$ in the Gibbs sampler in Section 2.2 is not feasible. To ensure that our model is sufficiently flexible to adequately approximate many different nonlinear shapes, both $\varphi(x_t)$ and $\psi(z_t)$ will be specified as infinite-dimensional. Thus, direct sampling of the corresponding coefficient vectors $\beta$ and $\gamma$ is out of the question.

Fortunately, inspection of the Gibbs sampler steps shows that in order to obtain draws from the predictive distribution $p(y_{T+1} | y)$, we do not need to draw $\beta$ and $\gamma$ explicitly, and likewise, we never have to compute the full vectors $\varphi(x_t)$ and $\psi(z_t)$. All we need are inner products of the forms $\varphi(x_s)' \varphi(x_t)$, $\varphi(x_t)' \beta$, $\psi(z_s)' \psi(z_t)$, and $\psi(z_t)' \gamma$; a proof of this assertion is presented in Section 2.4 below. Thus, if we choose the mappings $\varphi$ and $\psi$ in such a way that these inner products can be computed quickly, we can solve the nonlinear regression problem without much more effort than the linear variant. This realization is due to [Boser et al.] (1992), and it is known as the kernel trick.

Let $\Phi$ be the matrix with $t$-th row $\varphi(x_t)'$. The kernel trick revolves around rewriting all expressions from Section 2.2 that involve products like $\Phi' \Phi$ (or $\Phi' \Sigma^{-1} \Phi$) into expressions that involve $\Phi \Phi'$. Assuming that the number of regressors is much larger than the number of observations, such a reformulation leads to considerable computational savings. Denote $\kappa_{\varphi}(x_s, x_t) = \varphi(x_s)' \varphi(x_t)$, the kernel function, which we assume to be easily computable. Define the kernel matrix $K_{\varphi} = \Phi \Phi'$ and observe that its $(s, t)$-th element is $\kappa_{\varphi}(x_s, x_t)$. Moreover, we define the vector $k_{\varphi} = \Phi \varphi(x_{T+1})$, whose $t$-th element is $\kappa_{\varphi}(x_t, x_{T+1})$. For equation (3), the quantities $\Psi$, $\kappa_{\psi}$, $K_{\psi}$, and $k_{\psi}$ are defined analogously.

Several possible choices for the kernel functions $\kappa_{\varphi}$ and $\kappa_{\psi}$ have been proposed in the literature. In
this paper, we limit ourselves to Gaussian kernels \cite{Broomhead and Lowe, 1988},

\[
\kappa(x_s, x_t) = \exp\left(-\frac{(x_s - x_t)'(x_s - x_t)}{2\sigma^2}\right),
\]

where \( \sigma \) is a hyperparameter. The mapping \( \varphi \) that is associated with this kernel function maps the predictors into an infinite-dimensional regressor space, which is parameterized such that the prior on the regression coefficients \( \beta \) implies a smooth, but otherwise unrestricted regression function \( \varphi(x)'\beta \). For a full discussion, we refer to Exterkate (2013).

2.4 Gibbs sampler for general nonlinear models

Looking over the Gibbs sampler in Section 2.2, we observe that the only expressions in which the draws of \( \beta \) and \( \gamma \) are used are of the forms \( X\beta, x_{T+1}'\beta, Z\gamma, \) and \( z_{T+1}'\gamma \). Thus, as we move to nonlinear models, for which the dimensions of the coefficient vectors are infinite, it is sufficient to only draw the linear combinations \( \begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix} \beta \) and \( \begin{pmatrix} \Psi \\ \psi(z_{T+1})' \end{pmatrix} \gamma \), using the notation introduced in Section 2.3. Linear combinations of normally distributed variables also follow normal distributions, and as the sampling distribution of \( \beta \) is \( N\left(\left(\Phi'\Sigma^{-1}\Phi + \lambda I\right)^{-1}\Phi'\Sigma^{-1}y, \tau^{-1}\left(\Phi'\Sigma^{-1}\Phi + \lambda I\right)^{-1}\right)\), we have

\[
\begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix} \beta \sim N\left(\begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix} \left(\Phi'\Sigma^{-1}\Phi + \lambda I\right)^{-1} \Phi'\Sigma^{-1}y, \tau^{-1}\left(\Phi'\Sigma^{-1}\Phi + \lambda I\right)^{-1}\left(\begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix}' \Phi \varphi(x_{T+1})\right)\right).
\]

Straightforward but tedious algebra, detailed in Exterkate et al. (2013), can be used to rewrite the mean and variance of this distribution entirely in terms of the kernel function: we have

\[
\begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix} (\Phi'\Sigma^{-1}\Phi + \lambda I)^{-1} \Phi'\Sigma^{-1}y = \begin{pmatrix} \Phi' \varphi(x_{T+1})' \end{pmatrix} \left(\Phi'\varphi(x_{T+1}) + \lambda \Sigma\right)^{-1} y = \begin{pmatrix} K_{\varphi} \\ k_{\varphi}' \end{pmatrix} (K_{\varphi} + \lambda \Sigma)^{-1} y,
\]
and $\tau^{-1} \begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix} (\Phi'\Sigma^{-1}\Phi + \lambda I)^{-1} \begin{pmatrix} \Phi' \\ \varphi(x_{T+1}) \end{pmatrix} = \tau^{-1} \begin{pmatrix} \Phi \Phi' (\Phi\Phi' + \lambda \Sigma)^{-1} \Sigma \\ \varphi(x_{T+1})' (\Phi\Phi' + \lambda \Sigma)^{-1} \Sigma \\ \lambda^{-1} \left( \varphi(x_{T+1})' \varphi(x_{T+1}) - \varphi(x_{T+1})' \Phi' (\Phi\Phi' + \lambda \Sigma)^{-1} \Phi \varphi(x_{T+1}) \right) \end{pmatrix}$

Finally, we rewrite the scale parameter of the sampling distribution for $\tau$; we find

$$2 \left( y' \left( \Sigma^{-1} - \Sigma^{-1} \Phi' \Sigma^{-1} \Phi + \lambda I \right)^{-1} \Phi' \Sigma^{-1} \right) y^{-1} = 2 \lambda^{-1} \left( y' (K_\varphi + \lambda \Sigma)^{-1} y \right)^{-1}.$$ 

To summarize, the Gibbs sampler steps for Model (1)-(2) are as follows:

- Draw $\tau$ from a gamma distribution with shape parameter $T/2$ and scale parameter $2\lambda^{-1} \left( y' (K_\varphi + \lambda \Sigma)^{-1} y \right)^{-1}$.

- Draw $\begin{pmatrix} \Phi \beta \\ \varphi(x_{T+1})' \beta \end{pmatrix}$ from a normal distribution with mean $\begin{pmatrix} K_\varphi \\ K_\psi' \end{pmatrix} (K_\varphi + \lambda \Sigma)^{-1} y$ and variance $\tau^{-1} \begin{pmatrix} K_\varphi (K_\varphi + \lambda \Sigma)^{-1} \Sigma \\ K_\psi' (K_\varphi + \lambda \Sigma)^{-1} \Sigma \\ \lambda^{-1} \left( K_\varphi (x_{T+1}) - K_\psi' (K_\varphi + \lambda \Sigma)^{-1} \right) \end{pmatrix}$.

- Draw $\begin{pmatrix} \Psi \gamma \\ \psi(z_{T+1})' \gamma \end{pmatrix}$ from a normal distribution with mean $\begin{pmatrix} K_\psi \\ K_\psi' \end{pmatrix} (K_\psi + \mu I)^{-1} y$ and variance $\tau^{-1} \begin{pmatrix} K_\psi (K_\psi + \mu I)^{-1} \\ K_\psi' (K_\psi + \mu I)^{-1} \Sigma \\ \lambda^{-1} \left( K_\psi (z_{T+1}) - K_\psi' (K_\psi + \mu I)^{-1} \right) \end{pmatrix}$. 

Here, we see the kernel trick "in action": instead of the infinite-dimensional $\beta$, we only draw the required linear combinations, which can be done in finite time. Likewise, the sampling distribution for

$\begin{pmatrix} \Psi \\ \psi(z_{T+1})' \end{pmatrix} \gamma$ is normal with mean $\begin{pmatrix} K_\psi \\ K_\psi' \end{pmatrix} (K_\psi + \mu I)^{-1} \eta$ and variance

$\begin{pmatrix} K_\psi (K_\psi + \mu I)^{-1} \\ K_\psi' (K_\psi + \mu I)^{-1} \Sigma \\ K_\psi' (K_\psi + \mu I)^{-1} \mu \left( \kappa_\psi (z_{T+1}) - K_\psi' (K_\psi + \mu I)^{-1} \right) \end{pmatrix}$.
For \( t = 1, \ldots, T \), let \( \nu_t \) be \( \frac{\tau}{2} \left( y_t - \varphi (x_t)' \beta \right)^2 \exp (-2\eta_t) \) plus a draw from the exponential distribution with mean one.

For \( t = 1, \ldots, T \), draw \( \eta_t \) from the \( N \left( \psi (z_t)' \gamma - 1, 1 \right) \) distribution, truncated to the interval \( \left( \frac{1}{2} \log \left( \frac{\tau (y_t - \varphi (x_t)' \beta)^2}{2\nu_t} \right), \infty \right) \).

Draw \( \eta_{T+1} \) from the normal distribution with mean \( \psi (z_{T+1})' \gamma \) and variance 1.

Draw \( y_{T+1} \) from the normal distribution with mean \( \varphi (x_{T+1})' \beta \) and variance \( \tau^{-1} \exp (2\eta_{T+1}) \).

We refer to this procedure as kernel ridge regression with stochastic volatility, or KRR-SV for short.

### 2.5 Hyperparameters

Four hyperparameters are present in the description of the Gibbs sampler in the previous sections: the ridge parameters \( \lambda \) and \( \mu \), and the kernel parameters \( \sigma_\varphi \) and \( \sigma_\psi \). Exterkate (2013) constructs a five-point grid for each of these parameters, and proposes to select their values from these grids using cross-validation. For \( \sigma_\varphi \) and \( \sigma_\psi \), these grids are centered on \( 2\sqrt{c_K}N/\pi \) and \( 2\sqrt{c_M}/\pi \), respectively, where \( c_K \) is the 95% quantile of the \( \chi^2 \) distribution with \( K \) degrees of freedom. The grid values for \( \lambda \) and \( \mu \) are based on estimates of the signal-to-noise ratios in equations (1) and (2).

In this study, we impose mutually independent priors on these four parameters, namely improper uniform priors over \( \{ \ldots, 2^{-2}, 2^{-1}, 2^0, 2^1, 2^2, \ldots \} \) times the values at the centers of the grids discussed above. We then implement Bayesian model averaging within the Gibbs sampler, using Markov Chain Monte Carlo Model Composition, or MC3 (Madigan and York, 1995).

### 3 Monte Carlo simulation

As we are concerned with the forecasting of entire distributions, point forecast evaluation tools such as mean squared or absolute errors are not suitable. Instead, let \( U_i \) be the quantile of the observed \( y_{T+1} \) relative to the sampled predictive distribution, in the \( i \)-th simulation run (this section) or for the \( i \)-th day (next section). If a forecasting method works well, \( U_i \) should be close to uniformly distributed. Thus, ordering the \( U_i \) as \( U_{(1)} \leq U_{(2)} \leq \ldots \leq U_{(S)} \), we compute the Kolmogorov-Smirnov statistic

\[
D_S = \sup_{1 \leq i \leq S} \left| U_{(i)} - \frac{i}{S + 1} \right|
\]
Table 1: Data-generating processes used in the static simulation study.

<table>
<thead>
<tr>
<th>DGP #</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10 explanatory variables in both $x_t$ and $z_t$, with $x_t = z_t$</td>
</tr>
<tr>
<td>2</td>
<td>as #1, but adding 20 redundant predictors to both $x_t$ and $z_t$</td>
</tr>
<tr>
<td>3</td>
<td>10 explanatory variables in both $x_t$ and $z_t$, including 7 which are common to both</td>
</tr>
<tr>
<td>4</td>
<td>as #3, but adding 20 redundant predictors to $z_t$</td>
</tr>
<tr>
<td>5</td>
<td>as #3, but adding 20 redundant predictors to $x_t$</td>
</tr>
<tr>
<td>6</td>
<td>as #3, but adding 20 redundant predictors to both $x_t$ and $z_t$</td>
</tr>
</tbody>
</table>

Its asymptotic distribution is given by $\sqrt{S} D_S \to \sup_{0 \leq t \leq 1} |B(t)|$, where $B(\cdot)$ is a Brownian bridge.

To assess the validity of our procedure, we start with a static model, for which KRR-SV is correctly specified. The data-generating processes are summarized in Table 1. The variables $x_t$ and $z_t$ are drawn from the standard normal distribution, and we scale $\beta$ and $\gamma$ such that the signal-to-noise ratios are as desired: either 0.1 or 0.5 ($R^2 = 0.09$, 0.33) in the mean equation, and either 0.2 or 1.0 ($R^2 = 0.17$, 0.50) in the volatility equation. We perform $S = 100$ replications for each DGP, with time series length $T = 100$ or 250, and we obtain 1000 Gibbs draws per replication, after discarding 200 burn-in draws.

Results are summarized in Table 2. Reassuringly, all of the Kolmogorov-Smirnov tests fail to reject at 5% significance. This result implies that the draws obtained using KRR-SV fit the true distribution well, and that our estimation procedure is not misled by redundant predictors, nor by overlap between predictors in the mean and volatility equations, nor by relatively low signal-to-noise ratios.

We now move to a more realistic set of examples, in which dynamics are present. Let $x_t = (y_{t-1}, \ldots, y_{t-\ell_m})'$ and $z_t = (y_{t-1}, \ldots, y_{t-\ell_v})'$; that is, both the mean and the volatility of $y_t$ have a nonlinear autoregressive structure. We set $\ell_m = 1$ and $\ell_v = 5$, and we estimate the model in three ways: underspecified, with $\hat{\ell}_m = \hat{\ell}_v = 1$; correctly specified, with $\hat{\ell}_m = 1$ and $\hat{\ell}_v = 5$; and overspecified, with $\hat{\ell}_m = \hat{\ell}_v = 10$. To control the signal-to-noise ratio, we set $\tau$ in Equation (1) to either 0.1 or 0.5.

Table 2: Results of Kolmogorov-Smirnov tests in the static simulation study.

<table>
<thead>
<tr>
<th>DGP #</th>
<th>Signal-to-noise ratios, $T = 100$</th>
<th>Signal-to-noise ratios, $T = 250$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>mean: 0.1 vol.: 0.2</td>
<td>mean: 0.1 vol.: 1.0</td>
</tr>
<tr>
<td>1</td>
<td>0.094 (0.320) 0.099 (0.263) 0.081 (0.502) 0.085 (0.441)</td>
<td>0.070 (0.685) 0.094 (0.320) 0.074 (0.617) 0.083 (0.471)</td>
</tr>
<tr>
<td>2</td>
<td>0.077 (0.567) 0.079 (0.534) 0.111 (0.158) 0.086 (0.426)</td>
<td>0.084 (0.456) 0.094 (0.320) 0.127 (0.073) 0.114 (0.137)</td>
</tr>
<tr>
<td>3</td>
<td>0.066 (0.751) 0.077 (0.567) 0.073 (0.634) 0.081 (0.502)</td>
<td>0.081 (0.502) 0.091 (0.357) 0.106 (0.197) 0.075 (0.600)</td>
</tr>
<tr>
<td>4</td>
<td>0.100 (0.253) 0.083 (0.471) 0.092 (0.345) 0.075 (0.600)</td>
<td>0.077 (0.567) 0.089 (0.384) 0.120 (0.103) 0.107 (0.188)</td>
</tr>
<tr>
<td>5</td>
<td>0.101 (0.243) 0.068 (0.718) 0.089 (0.384) 0.079 (0.534)</td>
<td>0.093 (0.332) 0.076 (0.384) 0.104 (0.214) 0.103 (0.223)</td>
</tr>
<tr>
<td>6</td>
<td>0.092 (0.345) 0.105 (0.205) 0.082 (0.487) 0.095 (0.308)</td>
<td>0.106 (0.197) 0.086 (0.426) 0.087 (0.412) 0.075 (0.600)</td>
</tr>
</tbody>
</table>

Notes: This table reports the Kolmogorov-Smirnov test statistics for all data-generating processes in the static simulation study. Asymptotic p-values follow in parentheses. The DGP numbering corresponds to Table 1.
Table 3: Results of Kolmogorov-Smirnov tests in the dynamic simulation study.

<table>
<thead>
<tr>
<th>$\tau$</th>
<th>$T$</th>
<th>underspecified</th>
<th>correctly specified</th>
<th>overspecified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
<td>100</td>
<td>0.161 (0.010)</td>
<td>0.104 (0.214)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>0.165 (0.008)</td>
<td>0.063 (0.799)</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>100</td>
<td>0.156 (0.014)</td>
<td>0.131 (0.059)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>250</td>
<td>0.128 (0.069)</td>
<td>0.067 (0.735)</td>
</tr>
</tbody>
</table>

Notes: This table reports the Kolmogorov-Smirnov test statistics for all data-generating processes in the dynamic simulation study, as described in the text. Asymptotic p-values follow in parentheses.

Table 3 lists the results for these dynamic models. It is apparent that KRR-SV also works well in this dynamic specification, as expected. Difficulties arise if the model is underspecified, whereas overspecification does not lead to any problems. In realistic applications, where the correct lag length is unknown, it is therefore advisable to err on the side of overspecification rather than underspecification.

Finally, we turn to a set of examples for which KRR-SV is misspecified, namely GARCH models. Data is generated by $y_t = \sqrt{h_t} \varepsilon_t$, where $\varepsilon_t \sim N(0, 1)$. We consider three specifications for $h_t$:

- **ARCH(1)**: $h_t = 0.15 + 0.85y_{t-1}^2$,
- **GARCH(1,1)**: $h_t = 0.05 + 0.85h_{t-1} + 0.10y_{t-1}^2$,
- **GARCH(2,1)**: $h_t = 0.05 + 0.65h_{t-1} + 0.20h_{t-2} + 0.10y_{t-1}^2$.

We estimate all models with one lag in the mean equation, and either a “reasonable” lag length ($p+q$ lags of $y_t$ for GARCH($p,q$)) or an unnecessarily long lag length (10 lags) in the volatility equation. Observe that there is no need to additionally include such transformations as $y_{t-j}^2, |y_{t-j}|, y_{t-j}^2 \cdot I\{y_{t-j} > 0\}$, etc., in $x_t$ or $z_t$. They are already modelled implicitly through the mappings $\varphi$ and $\psi$ in Model (1)-(2).

Table 4: Results of Kolmogorov-Smirnov tests in the GARCH simulation study.

<table>
<thead>
<tr>
<th>Model</th>
<th>$T$</th>
<th>$p + q$</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARCH(1)</td>
<td>100</td>
<td>0.054 (0.917)</td>
<td>0.078 (0.551)</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.071 (0.668)</td>
<td>0.111 (0.158)</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>100</td>
<td>0.074 (0.617)</td>
<td>0.077 (0.567)</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.101 (0.243)</td>
<td>0.086 (0.426)</td>
</tr>
<tr>
<td>GARCH(2,1)</td>
<td>100</td>
<td>0.110 (0.165)</td>
<td>0.083 (0.471)</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>0.129 (0.065)</td>
<td>0.183 (0.002)</td>
</tr>
</tbody>
</table>

Notes: This table reports the Kolmogorov-Smirnov test statistics for all data-generating processes in the GARCH simulation study, as described in the text. Asymptotic p-values follow in parentheses.
The results can be found in Table 4. As this table makes clear, the KRR-SV procedure performs well even in this misspecified case. If a reasonable lag length is specified, the performance is very good; if we estimate a model with many unnecessary lags, the predictive distribution is still quite a good fit. The exception is the GARCH(2,1) model, where it appears necessary not to overspecify the number of lags.

We conclude that KRR-SV performs well in a variety of static and dynamic models with time-varying volatility, including GARCH models, for which KRR-SV is misspecified. The great flexibility offered by kernel ridge regression accommodates the approximation of many different functional forms, without being misled by overspecification of the model.

4 Empirical application

We evaluate the forecast performance of KRR-SV in an empirical application to the distribution of daily returns on the S&P500 index. A time-series plot of these returns over the period 2007–2012 is shown in the left panel of Figure 1. The volatility clustering is obvious from this graph. We additionally present kernel density estimates of the returns distributions in the turbulent year from mid-2008 to mid-2009, and in the relatively quiet year following it. Aside from the obvious difference in volatilities between the two years, this plot also provides some visual evidence that a normal or Student’s t assumption may be insufficient to describe the tails of these distributions.

We use KRR-SV to obtain one-day-ahead forecasts of the returns distribution, for each trading day in 2008–2012. As in the simulation study, performance is assessed using the Kolmogorov-Smirnov
Table 5: Results of Kolmogorov-Smirnov tests in the S&P500 forecasting application.

<table>
<thead>
<tr>
<th>Model</th>
<th>Statistic</th>
<th>Model</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRR-SV(5)</td>
<td>0.029 (0.250)</td>
<td>EGARCH(1,1)</td>
<td>0.034 (0.107)</td>
</tr>
<tr>
<td>ARCH(1)</td>
<td>0.039 (0.041)</td>
<td>EGARCH(1,2)</td>
<td>0.031 (0.175)</td>
</tr>
<tr>
<td>ARCH(2)</td>
<td>0.041 (0.028)</td>
<td>EGARCH(2,1)</td>
<td>0.032 (0.153)</td>
</tr>
<tr>
<td>GARCH(1,1)</td>
<td>0.035 (0.093)</td>
<td>GJR(1,1)</td>
<td>0.030 (0.210)</td>
</tr>
<tr>
<td>GARCH(1,2)</td>
<td>0.031 (0.169)</td>
<td>GJR(1,2)</td>
<td>0.026 (0.357)</td>
</tr>
<tr>
<td>GARCH(2,1)</td>
<td>0.035 (0.086)</td>
<td>GJR(2,1)</td>
<td>0.030 (0.218)</td>
</tr>
<tr>
<td>GARCH(2,2)</td>
<td>0.032 (0.148)</td>
<td>GJR(2,2)</td>
<td>0.029 (0.252)</td>
</tr>
</tbody>
</table>

Notes: This table reports the Kolmogorov-Smirnov test statistics in the S&P500 forecasting study, over all trading days in 2008–2012. Asymptotic p-values follow in parentheses.

goodness-of-fit test. Note that this test can be interpreted as quantifying the validity of value-at-risk measures based on each model. We compare KRR-SV to standard ARCH and GARCH models, and to two popular GARCH-type models that allow for leverage effects: EGARCH (Nelson [1991]), where

$$\log (h_t) = \omega + \sum_{i=1}^{p} \beta_i \log (h_{t-i}) + \sum_{i=1}^{q} \left[ \alpha_i \varepsilon_{t-i} + \gamma_i \left( \varepsilon_{t-i} - E[\varepsilon_{t-i}] \right) \right],$$

and GJR (Glosten et al. [1993]), where

$$h_t = \omega + \sum_{i=1}^{p} \beta_i h_{t-i} + \sum_{i=1}^{q} \left[ \alpha_i + \gamma_i I\{y_{t-i} > 0\} \right] y_{t-i}^2.$$

All models are estimated on a rolling window of length $T = 250$ trading days. For KRR-SV, we use one lag in the mean equation and five in the log-volatility equation. To enable a fair comparison, the other models are also estimated with one lag in the mean equation. The lag lengths $p$ and $q$ are set to 1 or 2, and a Student’s t distribution is assumed for the innovations $\varepsilon_t$.

The results of the Kolmogorov-Smirnov tests are reported in Table 5. At the 5% significance level, we find that the predictive distributions implied by the ARCH models can be rejected. The KRR-SV model provides as good a fit as the GJR models, outperforming both GARCH and EGARCH. Recall the interpretation of these statistics: they imply that value-at-risk measures calculated using KRR-SV are as good as those from GJR models, and more accurate than those from any of the other models.

1Allowing for a mean equation with only a constant term, for longer lag lengths in the volatility equations, or for a normal innovation distribution does not improve the performance of the GARCH-type models. These results are not reported here, but are available from the author upon request.
Thus, we find that KRR-SV is capable of predicting the distribution of returns at least as well as popular GARCH-type models. An advantage of our model, however, is that it is not bound to restrictive functional forms or distributional assumptions. This benefit reveals itself particularly when major crashes occur. The predictive distribution implied by KRR-SV is sufficiently flexible to account for the possibility of such crashes, while remaining a good approximation on “regular” days as well.

To illustrate this point, we show the predictive distributions implied by various models for Tuesday, January 20, 2009, in the left panel of Figure 2. The S&P500 index lost 5.4% on that day, following two announcements from Europe – the European Central Bank stated that “2009 growth will be substantially below December 2008 forecasts”, and the Royal Bank of Scotland announced a record loss. The plots clearly indicate that KRR-SV is the only model to attach a sizeable probability to the occurrence of such a loss. In fact, the realized loss corresponds to the 5% quantile of the predictive distribution calculated on the day before. This figure is around 1% for both GJR and EGARCH, and around 0.1% for both ARCH and GARCH. From a risk-management perspective, the KRR-SV forecast seems preferable.

For comparison, the predictive distributions for Tuesday, January 19, 2010 are also shown in the right panel of Figure 2. This is the corresponding day one year later. It was a “regular”, quiet day, on which the index gained 1.3%. This realized return corresponds to the 78% quantile of the distribution predicted by KRR-SV. Yet, the predictive distributions from GARCH, EGARCH, and GJR models classify it in their 5% right tails. Recalling the right panel of Figure 1, all GARCH-type models seem to have underestimated the kurtosis here.
5 Conclusion

We have extended kernel ridge regression, a method for estimating very flexible nonlinear models, to processes with stochastic volatility. Both the mean and the volatility are modelled as nonlinear functions of observed variables. We have shown that draws from the predictive distribution can be obtained in a relatively straightforward way, using a Gibbs sampler that does not require much more computational effort than if the model were linear.

Evidence from a Monte Carlo study shows that kernel ridge regression with stochastic volatility (KRR-SV) delivers highly competitive forecasts in a wide variety of data-generating processes. In an empirical example to forecasting the S&P500 index return distribution one day ahead, KRR-SV outperforms most popular GARCH-type models. It places more mass in the left tail of the predictive returns distribution before large crashes, while retaining a good fit to the distribution as a whole.

So far, we have only used lags of the dependent variable as predictors, in order to enable a fair comparison to the GARCH framework. As our simulation study shows, the KRR-SV approach could easily deal with more explanatory variables, such as realized volatilities, liquidity measures, or even macroeconomic quantities. Another interesting potential extension of this model is to allow for a multivariate dependent variable, such as a panel of stock returns rather than the index. Extending the Gibbs sampler in this paper to a model with \( y_t \mid \eta_t, B, \Upsilon \sim N \left( B' \varphi (x_t), \exp (2\eta_t) \Upsilon^{-1} \right) \), where \( B \) and \( \Upsilon \) are now matrices, is straightforward. An extension that allows the conditional variance matrix of \( y_t \) to depend on a multivariate latent process could be an interesting avenue for further research.

References


