Density Forecasting in Nonlinear Models with Stochastic Volatility

Peter Exterkate[*]

School of Economics, University of Sydney

Preliminary draft, 15 February 2016

Abstract

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.

This is a preliminary draft. Significant extensions to both the methodology and its empirical evaluation are currently work in progress. These extensions are discussed in paragraphs in italics, similar to this one, throughout the paper.

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

[*] Address for correspondence: School of Economics, Merewether Building H04, The University of Sydney, NSW 2006, Australia; email: peter.exterkate@sydney.edu.au; phone: +61 2 9351 8532.

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. (2015). 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 $t = 1, \ldots, T+1$:

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

Here, x_t' is the t-th row of X and η_t is the unobserved time-varying log-volatility. The mapping φ 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 φ the identity mapping. The nonlinear extension is studied in Sections 2.3 and 2.4.

The conditional model for η_t is also normal:

$$\eta_t \,|\gamma \sim N\left(\psi\left(z_t\right)' \gamma,\, 1\right), \tag{2}$$

where again, ψ 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:

$$\begin{aligned} p\left(y\left|y_{1-P}, \ldots, y_0\right.\right) &= p\left(y_1\left|y_{1-P}, \ldots, y_0\right.\right) p\left(y_2\left|y_{1-P} \ldots, y_1\right.\right) \cdots p\left(y_T\left|y_{1-P}, \ldots, y_{T-1}\right.\right) \\ &= p\left(y_1\left|y_{1-P}, \ldots, y_0\right.\right) p\left(y_2\left|y_{2-P}, \ldots, y_1\right.\right) \cdots p\left(y_T\left|y_{T-P}, \ldots, y_{T-1}\right.\right). \end{aligned}$$

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,

$$\beta\left|\tau \sim N\left(0,\, (\lambda\tau)^{-1} I\right)\right. \quad \text{and} \quad \gamma \sim N\left(0,\, \mu^{-1} I\right),$$

where λ and μ are hyperparameters, and I denotes an identity matrix of the appropriate dimensions.

The most obvious extension of this model would be to allow for autoregressive effects in the volatility equation (2)*. I am currently working out the details, but the most promising setup appears to be one where lagged log-volatilities enter linearly,*

$$\eta_t\left|h_t, \gamma, \delta \sim N\left(\psi\left(z_t\right)' \gamma + h_t' \delta,\, 1\right)\right. \quad \text{where} \quad h_t = \left(\eta_{t-1}, \eta_{t-2}, \ldots, \eta_{t-Q}\right)'.$$

This extension is nontrivial because it introduces dependence between η_t and its lags, which requires a substantial modification of the Gibbs sampler in Sections 2.2 and 2.4. Nevertheless, such an extension appears to be feasible as long as this dependence is restricted to be linear. Assuming a diffuse prior $p(\delta) \propto 1$ essentially leads to the model introduced in Exterkate et al. (2015), where the posterior mode of the predictive density of η_{T+1} is derived. Priors for the pre-sample log-volatilities can also be diffuse.

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 φ and ψ; thus, writing $\Sigma = \text{diag}\left(\exp\left(2\eta\right)\right)$, equations (1) and (2) can be simplified to

$$\begin{array}{rrcll} & p\left(y\,|\eta,\beta,\tau,\gamma\right) & = & N\left(X\beta,\,\tau^{-1}\Sigma\right) & \propto\ \tau^{T/2}\exp\left(-\iota'\eta\right)\exp\left(-\frac{\tau}{2}\left(y-X\beta\right)'\Sigma^{-1}\left(y-X\beta\right)\right) \\ \text{and} & p\left(\eta\,|\beta,\tau,\gamma\right) & = & N\left(Z\gamma,\,I\right) & \propto\ \exp\left(-\frac{1}{2}\left(\eta-Z\gamma\right)'\left(\eta-Z\gamma\right)\right). \end{array}$$

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

In each step s of the Gibbs sampler, we need to draw from the conditional posterior distributions $p\left(\beta^{(s)},\tau^{(s)},\gamma^{(s)}\,|\eta^{(s-1)},y\right)$ and $p\left(\eta^{(s)}\,|\beta^{(s)},\tau^{(s)},\gamma^{(s)},y\right)$. We first consider the problem of drawing the parameters β, τ, and γ. As we are conditioning on η, 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 $\tau^{(s)}$ from a gamma distribution with shape parameter $T/2$ and scale parameter given by $2\left(y'\left(\Sigma^{(s-1)\,-1}-\Sigma^{(s-1)\,-1}X\left(X'\Sigma^{(s-1)\,-1}X+\lambda I\right)^{-1}X'\Sigma^{(s-1)\,-1}\right)y\right)^{-1}$. Here and henceforth, the notation $\Sigma^{(s-1)\,-1}$ is to be read as "the inverse of Σ as it was drawn in step $s-1$".
- Draw $\beta^{(s)}$ from a normal distribution with mean $\left(X'\Sigma^{(s-1)\,-1}X+\lambda I\right)^{-1}X'\Sigma^{(s-1)\,-1}y$ and variance matrix $\tau^{(s)\,-1}\left(X'\Sigma^{(s-1)\,-1}X+\lambda I\right)^{-1}$.
- Draw $\gamma^{(s)}$ from a normal distribution with mean $\left(Z'Z+\mu I\right)^{-1}Z'\eta^{(s-1)}$ and variance matrix $\left(Z'Z+\mu I\right)^{-1}$.

Drawing the latent log-volatilities η is somewhat more involved. We have

$$\begin{aligned} p\left(\eta\,|\beta,\tau,\gamma,y\right) &\propto p\left(y\,|\eta,\beta,\tau,\gamma\right)p\left(\eta\,|\beta,\tau,\gamma\right) \\ &\propto \tau^{T/2}\exp\left(-\iota'\eta-\tfrac{\tau}{2}\left(y-X\beta\right)'\Sigma^{-1}\left(y-X\beta\right)-\tfrac{1}{2}\left(\eta-Z\gamma\right)'\left(\eta-Z\gamma\right)\right) \\ &\propto \exp\left(-\tfrac{1}{2}\eta'\eta+\left(Z\gamma-\iota\right)'\eta-\tfrac{\tau}{2}\left(y-X\beta\right)'\Sigma^{-1}\left(y-X\beta\right)\right) \\ &= \textstyle\prod_{t=1}^{T}\exp\left(-\tfrac{1}{2}\eta_t^2+\left(z_t'\gamma-1\right)\eta_t-\tfrac{\tau}{2}\left(y_t-x_t'\beta\right)^2\exp\left(-2\eta_t\right)\right) \\ &\propto \textstyle\prod_{t=1}^{T}\exp\left(-\tfrac{1}{2}\left(\eta_t-z_t'\gamma+1\right)^2-\tfrac{\tau}{2}\left(y_t-x_t'\beta\right)^2\exp\left(-2\eta_t\right)\right). \end{aligned}$$

Thus, we see that we can draw each η_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 η_t:

$$p\left(\eta_t, v_t \,|\beta,\tau,\gamma,y\,\right) \;\propto\; I\Big\{v_t > \tfrac{\tau}{2}\left(y_t - x_t'\beta\right)^2 \exp\left(-2\eta_t\right)\Big\} \exp\left(-v_t\right) \exp\left(-\tfrac{1}{2}\left(\eta_t - z_t'\gamma + 1\right)^2\right), \quad (3)$$

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

$$\begin{aligned} p\left(v_t \,|\eta_t,\beta,\tau,\gamma,y\,\right) &\;\propto\; I\left\{v_t > \tfrac{\tau}{2}\left(y_t - x_t'\beta\right)^2 \exp\left(-2\eta_t\right)\right\} \exp\left(-v_t\right) \\ \text{and} \quad p\left(\eta_t \,|v_t,\beta,\tau,\gamma,y\,\right) &\;\propto\; I\left\{\eta_t > \tfrac{1}{2}\log\left(\tfrac{\tau\left(y_t - x_t'\beta\right)^2}{2v_t}\right)\right\} \exp\left(-\tfrac{1}{2}\left(\eta_t - z_t'\gamma + 1\right)^2\right). \end{aligned}$$

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

- Let $v_t^{(s)}$ be $\frac{\tau^{(s)}}{2}\left(y_t - x_t'\beta^{(s)}\right)^2 \exp\left(-2\eta_t^{(s-1)}\right)$ plus a draw from the exponential distribution with mean one.
- Draw $\eta_t^{(s)}$ from the $N\left(z_t'\gamma^{(s)} - 1,\ 1\right)$ distribution, truncated to the interval $\left(\frac{1}{2}\log\left(\frac{\tau\left(y_t - x_t'\beta^{(s)}\right)^2}{2v_t^{(s)}}\right), \infty\right)$.

As mentioned, this part of the Gibbs sampler will be more involved if autocorrelation between the η_t is introduced. However, as long as we keep this dependence linear, the approach outlined above can be salvaged; the variables will just need to be drawn in the order $\eta_{1-Q}, \ldots, \eta_0, v_1, \eta_1, v_2, \eta_2, \ldots, v_T, \eta_T$. Drawing the autoregressive coefficients δ along with γ should not pose any problems, since we are still operating within the normal linear regression framework.

Ultimately, we are interested in draws from the predictive distribution $p\left(y_{T+1} \,|y\,\right)$. These draws can be obtained within the Gibbs sampler, so that it is sufficient to sample from conditional posteriors:

- Draw $\eta_{T+1}^{(s)}$ from the normal distribution with mean $z_{T+1}'\gamma^{(s)}$ and variance 1.
- Draw $y_{T+1}^{(s)}$ from the normal distribution with mean $x_{T+1}'\beta^{(s)}$ and variance $\tau^{(s)\,-1}\exp\left(2\eta_{T+1}^{(s)}\right)$.

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, φ and ψ 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 β and γ 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 β and γ 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 φ and ψ 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 Φ 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 (2), the quantities Ψ, κ_ψ, K_ψ, and k_ψ are defined analogously.

Several possible choices for the kernel functions κ_φ and κ_ψ have been proposed in the literature. In this paper, we limit ourselves to Gaussian kernels (Broomhead and Lowe, 1988),

$$\kappa(x_s, x_t) = \exp\left(-\frac{(x_s - x_t)'(x_s - x_t)}{2\sigma^2}\right), \tag{4}$$

where σ is a hyperparameter. The mapping φ 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 β 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 β and γ 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 β 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

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

Straightforward but tedious algebra, detailed in Exterkate et al. (2015), can be used to rewrite the mean and variance of this distribution entirely in terms of the kernel function: we have

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

and $\tau^{-1}\begin{pmatrix} \Phi \\ \varphi(x_{T+1})' \end{pmatrix}\left(\Phi'\Sigma^{-1}\Phi+\lambda I\right)^{-1}\begin{pmatrix} \Phi' & \varphi(x_{T+1}) \end{pmatrix}$

$$\begin{aligned}&= \tau^{-1}\begin{pmatrix} \Phi\Phi'\left(\Phi\Phi'+\lambda\Sigma\right)^{-1}\Sigma & \Sigma\left(\Phi\Phi'+\lambda\Sigma\right)^{-1}\Phi\varphi(x_{T+1}) \\ \varphi(x_{T+1})'\Phi'\left(\Phi\Phi'+\lambda\Sigma\right)^{-1}\Sigma & \lambda^{-1}\left(\varphi(x_{T+1})'\varphi(x_{T+1})-\varphi(x_{T+1})'\Phi'\left(\Phi\Phi'+\lambda\Sigma\right)^{-1}\Phi\varphi(x_{T+1})\right) \end{pmatrix} \\ &= \tau^{-1}\begin{pmatrix} K_\varphi\left(K_\varphi+\lambda\Sigma\right)^{-1}\Sigma & \Sigma\left(K_\varphi+\lambda\Sigma\right)^{-1}k_\varphi \\ k_\varphi'\left(K_\varphi+\lambda\Sigma\right)^{-1}\Sigma & \lambda^{-1}\left(\kappa_\varphi(x_{T+1},x_{T+1})-k_\varphi'\left(K_\varphi+\lambda\Sigma\right)^{-1}k_\varphi\right) \end{pmatrix}.\end{aligned}$$

Here, we see the kernel trick "in action": instead of the infinite-dimensional β, we only draw the required linear combinations, which can be done in finite time. Likewise, the sampling distribution for $\begin{pmatrix} \Psi \\ \psi\left(z_{T+1}\right)' \end{pmatrix} \gamma$ is normal with mean $\begin{pmatrix} K_\psi \\ k'_\psi \end{pmatrix} \left(K_\psi + \mu I\right)^{-1} \eta$ and variance

$$\begin{pmatrix} K_\psi \left(K_\psi + \mu I\right)^{-1} & \left(K_\psi + \mu I\right)^{-1} k_\psi \\ k'_\psi \left(K_\psi + \mu I\right)^{-1} & \mu^{-1} \left(\kappa_\psi\left(z_{T+1}, z_{T+1}\right) - k'_\psi \left(K_\psi + \mu I\right)^{-1} k_\psi\right) \end{pmatrix}.$$

Finally, we rewrite the scale parameter of the sampling distribution for τ; we find

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

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

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

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

- Draw $\begin{pmatrix} \Psi\gamma^{(s)} \\ \psi\left(z_{T+1}\right)'\gamma^{(s)} \end{pmatrix}$ from a normal distribution with mean $\begin{pmatrix} K_\psi \\ k'_\psi \end{pmatrix} \left(K_\psi + \mu I\right)^{-1} \eta^{(s-1)}$ and variance $\begin{pmatrix} K_\psi \left(K_\psi + \mu I\right)^{-1} & \left(K_\psi + \mu I\right)^{-1} k_\psi \\ k'_\psi \left(K_\psi + \mu I\right)^{-1} & \mu^{-1} \left(\kappa_\psi\left(z_{T+1}, z_{T+1}\right) - k'_\psi \left(K_\psi + \mu I\right)^{-1} k_\psi\right) \end{pmatrix}$.

- For $t = 1, \ldots, T$, let $v_t^{(s)}$ be $\frac{\tau^{(s)}}{2}\left(y_t - \varphi\left(x_t\right)'\beta^{(s)}\right)^2 \exp\left(-2\eta_t^{(s-1)}\right)$ plus a draw from the exponential distribution with mean one.

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

- Draw $\eta_{T+1}^{(s)}$ from the normal distribution with mean $\psi\left(z_{T+1}\right)'\gamma^{(s)}$ and variance 1.

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

We refer to this procedure as *kernel ridge regression with stochastic volatility*, or KRR-SV for short. *Introducing autocorrelation in the volatility process has the same consequences outlined in Section 2.2.*

2.5 Hyperparameters

Four hyperparameters are present in the description of the Gibbs sampler in the previous sections: the ridge parameters λ and μ, and the kernel parameters σ_φ and σ_ψ. 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 σ_φ and σ_ψ, these grids are centered on $2\sqrt{c_N}/\pi$ and $2\sqrt{c_M}/\pi$, respectively, where c_K is the 95% quantile of the χ^2 distribution with K degrees of freedom. The grid values for λ and μ 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 MC^3 (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

$$D_S = \sup_{1 \leq i \leq S} \left| U_{(i)} - \frac{i}{S+1} \right|.$$

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.

Of course, different evaluation metrics are possible. In the Bayesian context of this paper, I am planning to compute and compare log predictive scores across models. Another classical candidate would be the Anderson-Darling test, which is known to be more sensitive to lack of fit in the tails of the distribution than Kolmogorov-Smirnov.

Table 1: Data-generating processes used in the static simulation study.

DGP #	Description
1	10 explanatory variables in both x_t and z_t, with $x_t = z_t$
2	as #1, but adding 20 redundant predictors to both x_t and z_t
3	10 explanatory variables in both x_t and z_t, including 7 which are common to both
4	as #3, but adding 20 redundant predictors to z_t
5	as #3, but adding 20 redundant predictors to x_t
6	as #3, but adding 20 redundant predictors to both x_t and z_t

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 β and γ 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 $x_t = (y_{t-1}, \dots, y_{t-\ell_m})'$ and $z_t = (y_{t-1}, \dots, y_{t-\ell_v})'$; 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 τ in Equation (1) to either 0.1 or 0.5.

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

	Signal-to-noise ratios, $T = 100$				Signal-to-noise ratios, $T = 250$			
	mean: 0.1	mean: 0.1	mean: 0.5	mean: 0.5	mean: 0.1	mean: 0.1	mean: 0.5	mean: 0.5
DGP #	vol.: 0.2	vol.: 1.0	vol.: 0.2	vol.: 1.0	vol.: 0.2	vol.: 1.0	vol.: 0.2	vol.: 1.0
1	0.094 (0.320)	0.099 (0.263)	0.081 (0.502)	0.085 (0.441)	0.070 (0.685)	0.094 (0.320)	0.074 (0.617)	0.083 (0.471)
2	0.077 (0.567)	0.079 (0.534)	0.111 (0.158)	0.086 (0.426)	0.084 (0.456)	0.094 (0.320)	0.127 (0.073)	0.114 (0.137)
3	0.066 (0.751)	0.077 (0.567)	0.073 (0.634)	0.081 (0.502)	0.081 (0.502)	0.091 (0.357)	0.106 (0.197)	0.075 (0.600)
4	0.100 (0.253)	0.083 (0.471)	0.092 (0.345)	0.075 (0.600)	0.077 (0.567)	0.089 (0.384)	0.120 (0.103)	0.107 (0.188)
5	0.101 (0.243)	0.068 (0.718)	0.089 (0.384)	0.079 (0.534)	0.093 (0.332)	0.076 (0.584)	0.104 (0.214)	0.103 (0.223)
6	0.092 (0.345)	0.105 (0.205)	0.082 (0.487)	0.095 (0.308)	0.106 (0.197)	0.086 (0.426)	0.087 (0.412)	0.075 (0.600)

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.

τ	T	underspecified	correctly specified	overspecified
0.1	100	0.161 (0.010)	0.104 (0.214)	0.100 (0.253)
	250	0.165 (0.008)	0.063 (0.799)	0.056 (0.895)
0.5	100	0.156 (0.014)	0.131 (0.059)	0.089 (0.384)
	250	0.128 (0.069)	0.067 (0.735)	0.075 (0.600)

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:

$$\begin{aligned}
\text{ARCH(1)}:&\quad h_t \;=\; 0.15 \,+\, 0.85 y_{t-1}^2,\\
\text{GARCH(1,1)}:&\quad h_t \;=\; 0.05 \,+\, 0.85 h_{t-1} \,+\, 0.10 y_{t-1}^2,\\
\text{GARCH(2,1)}:&\quad h_t \;=\; 0.05 \,+\, 0.65 h_{t-1} \,+\, 0.20 h_{t-2} \,+\, 0.10 y_{t-1}^2.
\end{aligned}$$

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 φ and ψ in Model (1)-(2).

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

		Number of lags	
Model	T	$p+q$	10
ARCH(1)	100	0.054 (0.917)	0.078 (0.551)
	250	0.071 (0.668)	0.111 (0.158)
GARCH(1,1)	100	0.074 (0.617)	0.077 (0.567)
	250	0.101 (0.243)	0.086 (0.426)
GARCH(2,1)	100	0.110 (0.165)	0.083 (0.471)
	250	0.129 (0.065)	0.183 (0.002)

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.

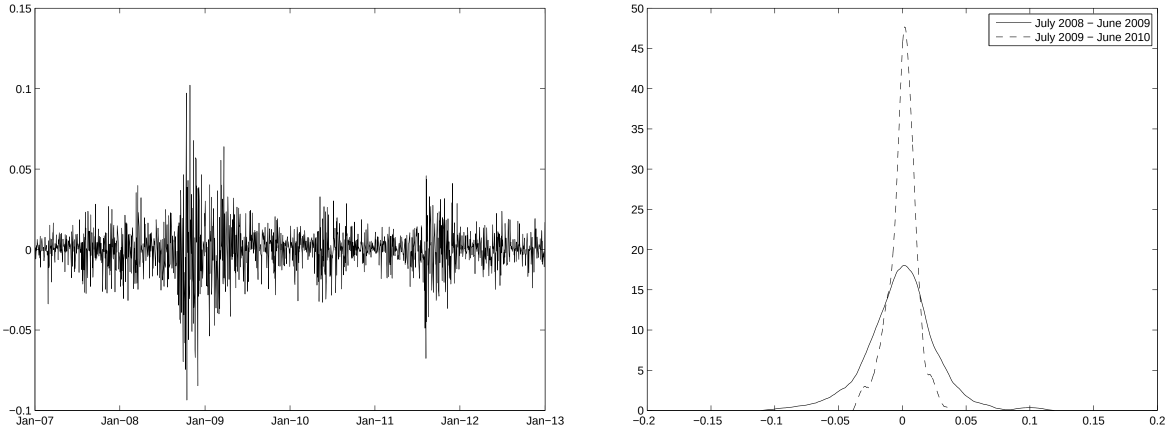
The results can be found in Table 4. As this table makes clear, the KRR-SV procedure performs well even in this misspecified case. If a reasonable lag length is specified, the performance is very good; if we estimate a model with many unnecessary lags, the predictive distribution is still quite a good 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.

Figure 1: Daily returns on the S&P500 index. Left panel: time-series plot, 2007–2012. Right panel: Kernel density estimates for two selected years.



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 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 ε_t. Allowing 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.

The setup described above provides for a preliminary empirical comparison of KRR-SV to several established methods. After working out the details of the extension of the proposed model as described in the earlier notes in this paper, more competing models will of course be considered. These additional benchmark models will definitely include linear stochastic volatility models, as well as GARCH-type models estimated in a Bayesian framework. A skew-t rather than standard t distribution will also be included for all of these models. Finally, more extensive model comparison tools will be considered, and the data set will be extended to include more recent observations.

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.

Table 5: Results of Kolmogorov-Smirnov tests in the S&P500 forecasting application.

Model	Statistic	Model	Statistic
KRR-SV(5)	0.029 (0.250)	EGARCH(1,1)	0.034 (0.107)
		EGARCH(1,2)	0.031 (0.175)
ARCH(1)	0.039 (0.041)	EGARCH(2,1)	0.032 (0.153)
ARCH(2)	0.041 (0.028)	EGARCH(2,2)	0.032 (0.154)
GARCH(1,1)	0.035 (0.093)	GJR(1,1)	0.030 (0.210)
GARCH(1,2)	0.031 (0.169)	GJR(1,2)	0.026 (0.357)
GARCH(2,1)	0.035 (0.086)	GJR(2,1)	0.030 (0.218)
GARCH(2,2)	0.032 (0.148)	GJR(2,2)	0.029 (0.252)

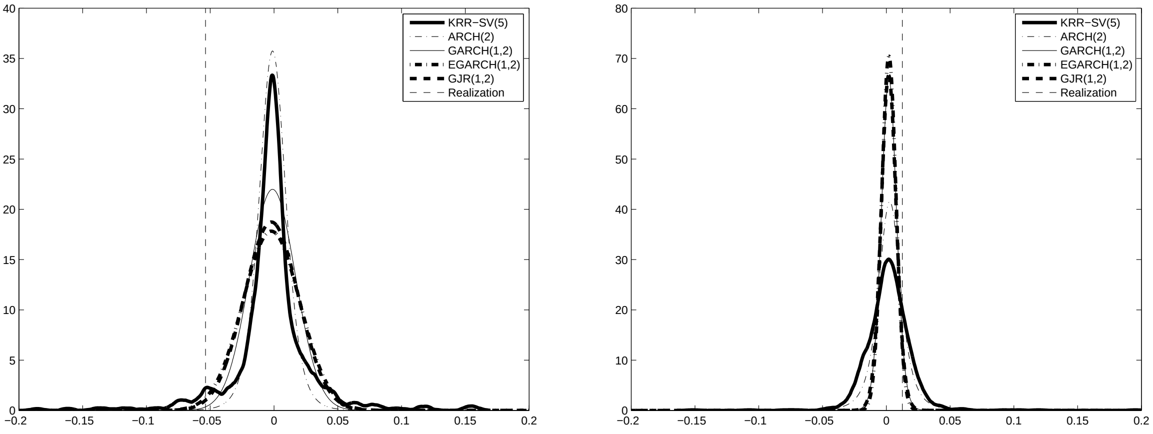
Notes: This table reports the Kolmogorov-Smirnov test statistics in the S&P500 forecasting study, over all trading days in 2008–2012. Asymptotic p-values follow in parentheses.

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.

Figure 2: Predicted return distributions for the S&P500 index on two selected dates. Left panel: January 20, 2009. Right panel: January 19, 2010.



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 $y_t \,|\eta_t, B, \Upsilon \sim N\left(B'\varphi\left(x_t\right), \exp\left(2\eta_t\right)\Upsilon^{-1}\right)$, where B and Υ 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.

Acknowledgements

The author acknowledges support from CREATES, Center for Research in Econometric Analysis of Time Series, funded by the Danish National Research Foundation (DNRF78), and additional financial support from the Danish Council for Independent Research (Grant #12-125914).

I wish to thank Stanislav Anatolyev, Dick van Dijk, David T. Frazier, Gael M. Martin, Chris F. Parmeter, Artem B. Prokhorov, Annastiina Silvennoinen, and Rami V. Tabri for useful comments and suggestions. Helpful input from conference participants at the Conference on Forecasting Structure and Time-Varying Patterns in Economics and Finance (Rotterdam, May 2013), the Conference on Computational and Financial Econometrics (London, December 2013), the Rimini Bayesian Econometrics Workshop (Rimini, June 2014), the International Symposium on Forecasting (Rotterdam, June 2014), the Conference on Recent Developments in Financial Econometrics and Applications (Geelong, December 2014), the New Zealand Econometric Study Group (Brisbane, February 2015), the Melbourne Bayesian Econometrics Workshop (Melbourne, June 2015), the Society for Financial Econometrics (Aarhus, June 2015), the Canadian Econometric Study Group (Guelph, September 2015), and the Sydney Econometrics Research Group (Sydney, October 2015), as well as seminar participants at the Universities of Aarhus, Bonn, Monash, New South Wales, Queensland, and Sydney, is also gratefully acknowledged. Any remaining errors are my own.

References

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

B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *Proceedings of the Annual Conference on Computational Learning Theory*, pages 144–152. ACM Press, Pittsburgh, Pennsylvania, 1992.

D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

P. Damien, J. Wakefield, and S. Walker. Gibbs sampling for Bayesian non-conjugate and hierarchical models by using auxiliary variables. *Journal of the Royal Statistical Society: Series B*, 61:331–344, 1999.

R.F. Engle. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50:987–1007, 1982.

P. Exterkate. Model selection in kernel ridge regression. *Computational Statistics and Data Analysis*, 68:1–16, 2013.

P. Exterkate, P.J.F. Groenen, C. Heij, and D. van Dijk. Nonlinear forecasting with many predictors using kernel ridge regression. *International Journal of Forecasting*, accepted for publication, 2015.

S. Geman and D. Geman. Stochastic relaxations, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.

L.R. Glosten, R. Jagannathan, and D. Runkle. On the relation between the expected value and the volatility of the nominal excess return on stocks. *Journal of Finance*, 48:1779–1801, 1993.

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

T. Hofmann, B. Schölkopf, and A.J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36:1171–1220, 2008.

E. Jacquier, N.G. Polson, and P.E. Rossi. Bayesian analysis of stochastic volatility models. *Journal of Business and Economic Statistics*, 20:69–87, 2002.

G. Koop. *Bayesian Econometrics*. Wiley, Chichester, 2003.

D. Madigan and J. York. Bayesian graphical models for discrete data. *International Statistical Review*, 63:215–232, 1995.

D.B. Nelson. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica*, 59:347–370, 1991.

M.A. Tanner and W. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–550, 1987.