ℓ_1-Penalized Likelihood Smoothing and Segmentation of Volatility Processes allowing for Abrupt Changes

David Neto[(a)], Sylvain Sardy[(b)] and Paul Tseng[(c)]

We consider the problem of estimating the volatility of a financial asset from a time series record. We believe the underlying volatility process is smooth, possibly stationary, and with potential small to large jumps due to market news. By drawing parallels between time series and regression models, in particular between stochastic volatility models and Markov random fields smoothers, we propose a semiparametric estimator of volatility. For the selection of the smoothing parameter, we derive a universal rule borrowed from wavelet smoothing. Our Bayesian posterior mode estimate is the solution to an ℓ_1-penalized likelihood optimization that we solve with an interior point algorithm that is efficient since its complexity is bounded by $O(T^{3/2})$, where T is the length of the time series. We apply our volatility estimator to real financial data, diagnose the model and perform back-testing to investigate forecasting power by comparison to standard methods. Supplemental materials (Appendices, `Matlab` code and data) are available online.

Keywords: financial time series, forecasting, GARCH, Markov random field, stochastic volatility model, wavelet.

1 Introduction

Suppose we observe values $\{y_t\}_{t=1,\ldots,T}$ of a financial asset, such as stock log-returns, at regularly spaced times $t = 1, \ldots, T$. Our goal is to estimate an important intrinsic characteristic of the financial asset of interest, the evolution of the conditional variance of the stochastic return process, so as to assess the past, present and future risk of the asset. To that aim, many stochastic models have been proposed. Standard among them is to assume a data generating process for y_t. Most models assume $y_t = \sigma_t \epsilon_t$, where σ_t is a measure of volatility, and ϵ_t is white standard Gaussian noise. Without additional assumption, the maximum likelihood estimate $\hat{\sigma}_t^{\mathrm{MLE}} = |y_t|$ is practically useless due to its nonsmooth nature (i.e., highly variable). A temporal structure for σ_t is assumed to regularize the maximum likelihood estimation, so as to obtain a smoother estimate of σ_t while capturing the stylized features (Rydberg, 2000) observed in financial econometrics, like heavy tails of the marginal distributions, volatility clustering or evolution with possible abrupt changes (e.g., jumps, peaks). The popular GARCH-type models (Engle, 1982; Bollerslev, 1986) are parametric and enjoy good estimation properties. Stochastic volatility models (Taylor, 1986; Taylor, 1994) are powerful semiparametric alternatives which fit more flexibly the stylized features. In particular, the log-Normal stochastic volatility model seems to better capture the leptokurticity of the marginal distribution of the financial data

than the standard GARCH model (Shephard, 1996). Stochastic volatility models are the empirical discrete-time analogs of some continuous-time models in finance theory and, in particular, option pricing (Hull and White, 1987). They can also be viewed as an Euler discretization of a diffusion. The general discrete model (Andersen, 1994) includes GARCH (with $q = 2$, $\varphi(h_t) = h_t$, $\eta_t = \epsilon_{t-1}^2$ and $\gamma = 0$) and stochastic volatility models (with $q = 1$, $\varphi(h_t) = \exp(h_t)$ and $\psi = 0$):

$$y_t = \sigma_t \epsilon_t, \tag{1}$$

$$\sigma_t^q = \varphi(h_t), \tag{2}$$

$$h_t = \omega + \phi h_{t-1} + (\gamma + \psi h_{t-1})\eta_t, \tag{3}$$

where $\varphi(\cdot)$ is positive, continuous and strictly monotone, the autoregressive process h_t has positive coefficients ψ, ϕ, γ such that $\psi + \phi > 0$, the errors η_t and ϵ_t are i.i.d. and mutually independent with mean-variance of $\left(1, \sigma_\eta^2\right)$ and $(0, 1)$. Often η_t and ϵ_t are assumed to be Gaussian.

One challenge for these models has been how to infer the parameters $(\psi, \phi, \gamma, \omega, \sigma_\eta^2)$ and how to estimate the hidden conditional volatility σ_t within the sample (smoothing) and out of sample (forecasting). While parametric GARCH models have been relatively easy to fit owing to the small number of parameters, it is not the case for semi-parametric SVM that represent the returns as noisy observations of volatilities modeled as a nonlinear Markov chain. Several methods have been proposed to estimate the SVM data generating process, many of which are based on the posterior conditional means calculated by an approximate Kalman filter (Harvey, Ruiz, and Shephard, 1994) or by MCMC.

Using GARCH, the estimated persistence of the conditional volatilities of many

financial time series is seemingly high, so that stationarity is questionable. Starica (2003) discusses the danger of assuming global stationarity. Cai (1994) and Hamilton and Susmel (1994) use Markov switching ARCH models, while Starica and Granger (2005) consider a succession of locally stationary processes.

Jumps have been another important modeling issue. For instance, Chernov, Gallant, Ghysels, and Tauchen (1999) add a jump process to capture large jumps. Chan and Maheu (2002) combine GARCH with a jump specification for the returns. Bates (2000) and Pan (2002) highlight the misspecification of diffuse processes for jumps to motivate the use of a continuous SVM model with a pure jump process in the returns y_t. Eraker, Johannes, and Polson (2008) additionally consider jumps in the volatility process h_t to not only capture large jumps (such as big crashes) but also smaller jumps (whose impact are more persistent). Aït-Sahalia and Jacod (2007) and Li, Wells, and Yu (2008), among others, use a Lévy process to model jumps.

The aforementioned results motivate the features we aim to reproduce with our proposed model and estimator, namely, that volatility evolves smoothly except for occasional small or abrupt jumps whose transient effect prevent the market from returning quickly to the level before shock. Our estimator adaptively estimates the segmentation between jumps' occurrences and the amplitude of the jumps by means of an ℓ_1-based regularization of the likelihood and a particular selection of a single smoothing parameter. This paper is organized as follows. In Section 2.1 we define our volatility estimator and discuss its link to wavelet and Markov random field smoothing. In Section 2.2 we use the connection to smoothers to present a selection rule for the smoothing parameter. In Section 2.3 we show that the volatility estimate can

be efficiently computed using an interior-point algorithm. Section 2.4 illustrates the performance of the estimator on a simulated time series. In Section 3 we apply our estimator to the DOW JONES, NASDAQ and S&P500 returns, analyze the results and evaluate its forecasting performance in comparison with existing methods. Section 4 draws some conclusions and points to future extensions.

2 Smoothing the volatility

2.1 ℓ_1 penalized likelihood estimator

To enforce temporal smoothness on the standard deviation σ_t estimates, we consider Markov random field modeling of the underlying volatilities. Originally developed to estimate trends in the mean (Geman and Geman, 1984; Besag, 1986), we employ this Bayesian modeling here for the estimation of trends in the standard deviation. Our model has three components:

$$y_t = \sigma_t \epsilon_t, \tag{4}$$

$$\sigma_t = \varphi(h_t), \tag{5}$$

$$h_t = \mu + \phi(h_{t-1} - \mu) + \eta_t, \tag{6}$$

where we assume ϵ_t's have density f_ϵ and η_t's are Laplace innovations with density $f_\eta(x) = \frac{\lambda}{2}\exp(-\lambda|x|)$. We will consider the Gaussian density for f_ϵ here, although an asymmetric density f_ϵ, such as an asymmetric t-distribution, would better fit the asymmetry of the returns. We discuss this issue later.

This volatility model can be seen as a nonlinear hidden Markov model. The autoregressive process (6) can also be seen as a first order Markov prior conditional on the

time prior to time t,

$$h_t \mid h_{t-1} \sim \text{Laplace}(\mu + \phi(h_{t-1} - \mu), \lambda),$$

with a persistence parameter ϕ, a mean volatility measure μ, and a Laplace distribution with variance $2/\lambda^2$ with tails heavier than the Gaussian to capture jumps. Appendix A derives some moment properties of this Laplace SVM model.

Using Bayes theorem, we derive from (4)–(6) the posterior distribution of the volatilities $\mathbf{h} = (h_1, \ldots, h_T)$ given the returns $\mathbf{y} = (y_1, \ldots, y_T)$. By then considering the negative log-posterior distribution, we define the ℓ_1-SVM maximum a posteriori estimate as the solution to

$$\min_{\phi>0, \mu\in\mathbb{R}, \mathbf{h}\in\mathbb{R}^T} \sum_{t=1}^{T} \log \varphi(h_t) - \log\{f_\epsilon(y_t/\varphi(h_t))\} + \lambda \sum_{t=2}^{T} |h_t - (\mu + \phi\,(h_{t-1} - \mu))| \quad (7)$$

for some $\lambda > 0$ parameter of which we discuss the selection in the next section. The first sum in (7) is the negative log-likelihood for (4)–(5) and the second sum stems from the autoregressive process prior (6) with Laplace η-innovations. The estimator is akin to a Tikhonov regularization (Tikhonov, 1963) of the erratic maximum likelihood estimate, but using an ℓ_1-based penalty and sufficiently large penalty parameter $\lambda > 0$ to enforce smoothness while capturing the jumps.

As opposed to GARCH type models, this estimator is semiparametric, which offers the advantage of being more flexible to fit the data. The parameter λ plays the role of the smoothing parameter and reflects the belief in the Laplace-AR(1) prior: the larger λ the smoother the estimated volatility; we discuss its selection in Section 2.2. The link function $\varphi(\cdot)$ maps the estimand h_t to the volatility $\sigma_t = \varphi(h_t)$. For instance, the exponential function used by the original log-Normal stochastic volatility model

maps $\mathbb{R}$ into $\mathbb{R}^+$ for a positive volatility. A likelihood-based estimate like ℓ_1-SVM already has active positivity constraints with the logarithmic terms in (7) which acts as a barrier against negativity. Hence we could consider a broader class of links, for instance, the power transform $h_t = (\sigma_t^\delta - 1)/\delta = \varphi^{-1}(\sigma_t)$ (Box and Cox, 1982).

The ℓ_1 penalized likelihood formulation of the ℓ_1-SVM estimator (7) draws connection to two nonparametric function estimation techniques. One is Markov random field smoothing (Geman and Geman, 1984; Besag, 1986). In particular, Sardy and Tseng (2004) consider a Laplace Markov random field prior, leading to a negative log-posterior of the form

$$\min_{\boldsymbol{h}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{h}\|_2^2 + \lambda \sum_{t=2}^{T} |h_t - h_{t-1}|,$$

where $\boldsymbol{h} = (h_1, \ldots, h_T)$ is the estimand; their corresponding prior includes a strong persistence fixed to $\phi = 1$. Another is wavelet smoothing (Donoho and Johnstone, 1994). For instance, soft-Waveshrink for a wavelet matrix Φ, solves

$$\min_{\boldsymbol{h}=\Phi\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{y} - \boldsymbol{h}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1,$$

where $\boldsymbol{\alpha}$ are the wavelet coefficients (Donoho, Johnstone, Hoch, and Stern, 1992). The main appeal of soft-Waveshrink is its near minimax properties for a class of loss functions and smoothness classes (Donoho, Johnstone, Kerkyacharian, and Picard, 1995) for a simple selection of the smoothing parameter to the value $\lambda_T^{\text{wave}} = \sqrt{2\log T}$ that only depends on T, the so-called universal rule (Donoho and Johnstone, 1994). We exploit this connection to propose in Section 2.2 a selection rule for λ that leads to a smooth estimation of volatility.

ℓ_1-SVM also draws connection to models used in econometrics and finance. In particular, this model corresponds to (1)–(3) with $\psi = 0$, $q = 1$, $\gamma = 1$, $\omega = \mu(1 - \phi)$

and Laplace innovations for η_t. SVM also has some connection with variance gamma (VG) models (Madan and Seneta, 1990) for modeling the non-Gaussian nature of stock market returns and future index prices. Indeed VG models can be expressed by equation (4), where σ_t follows a Gamma process, which in its continuous version writes as $S(t) = S(0)\exp(L(t))$, where $S(t)$ are the stock prices and $L(t)$ is a Laplace motion (Kotz, Kozubowski, and Podgórski, 2001). Interestingly the Laplace motion can be written as a compound Poisson process with independent and random jumps. In this sense it is a pure jumps process capable of capturing abrupt changes. In addition to VG models, our model considers an additional persistence parameter ϕ.

How does our smoothing approach differ from existing approaches? We aim at estimating a smooth evolution of volatility in time with possible small or large jumps, while existing approaches aim at estimating the true coefficients of some assumed erratic volatility data generating process. Our estimator also differs in the way it is computed since we solve a convex optimization problem (7) instead of solving an integration problem to calculate a posterior mean (see Section 2.3).

2.2 Selection of the smoothing parameter λ

The ℓ_1-regularization parameter $\lambda \geq 0$ in (7) controls the smoothing. Its selection (or equivalently σ_η^2 in (3)) is crucial. Indeed, when $\lambda = 0$, the solution is the wiggly maximum likelihood estimate $\hat{\sigma}_t^{\mathrm{MLE}} = |y_t|$ for $t = 1, \ldots, T$, while when λ tends to infinity the estimates $h_1, h_2, ..., h_T$, ϕ, μ approach a solution of the equations

$$h_t = \phi h_{t-1} + \mu(1-\phi), \quad t = 2, ..., T,$$

with h_1 arbitrary. In turn, if $\phi = 1$, then the solution is the constant vector of value μ; if $\phi < 1$, then the solution is

$$h_t = \phi^{t-1} h_1 + (1 - \phi^{t-1})\mu, \quad t = 2, ..., T, \tag{8}$$

which has an asymptotic limit (as $t \to \infty$) of μ. So h_t tends to either a constant function for $\phi = 1$ or a function that is asymptotically constant for $\phi < 1$ as $\lambda \to \infty$.

Between zero and infinity, an appropriately chosen λ confers "optimal" property to the estimator. If one defines optimality in term of predictive performance at an horizon h, then λ_h can be selected by minimizing a predictive cross-validation criterion (see the prediction formula (18) for SVM) by splitting the data into training and test sets with a sliding window over time. This approach suffers from the well known instability of cross-validation, from the correlation between data in adjacent windows, and from its high computational cost. Another approach consists in taking a single training set (i.e., one window with all the records until today) and in estimating the underlying volatility process optimally by selecting λ with a rule; we then rely on the prediction formula to predict the future. The selection rule can be based on minimizing an information criterion over λ, or by deriving a universal $\lambda = \lambda_T$ that only depends on the length T of records, an idea borrowed from wavelet smoothing. This latter approach requires solving (7) only once for that given λ_T, and, owing to ℓ_1-regularization, this approach detects jumps well, while it smooths quiet periods without erratic behaviors. This rule for λ also confers ℓ_1-SVM excellent predictive performance at various horizons, as we show with several back-tests in Section 3.

In Gaussian wavelet smoothing, the universal penalty $\lambda_T^{\text{wave}} = \sqrt{2 \log T}$ is a simple but surprisingly efficient choice as it endows the ℓ_1-penalized likelihood wavelet

estimator with near minimax properties for a class of loss functions and smoothness measures (Donoho and Johnstone, 1994; Donoho, Johnstone, Kerkyacharian, and Picard, 1995); in particular, it is striking to see that the universal penalty depends only on T and not on the underlying signal, and yet it provides optimal properties in a minimax sense for a whole range of smoothness (from very smooth function to erratic functions with jumps and peaks).

Borrowing from wavelet smoothing and total variation density estimation (Sardy and Tseng, 2010), we now derive a universal penalty $\lambda_T^{\ell_1-\mathrm{SVM}}$ for the maximum a posteriori estimator (7). In wavelet smoothing, the universal parameter is chosen such that, when the underlying signal is piecewise constant (i.e., $\phi = 1$), the estimate is also piecewise constant with probability tending to one as the sample size tends to infinity. Likewise here we set the universal parameter so that, when the true volatility is piecewise constant on K_T successive times (i.e., persistence with $\phi = 1$ on each interval), the volatility estimate is also piecewise constant with probability tending to one. Appendix B derives the universal parameter $\lambda_T^{\ell_1-\mathrm{SVM}} = \sqrt{K_T \log(n_T \log n_T)}$ with $n_T = T/K_T$ and $K_T \sim \log T$. For financial time series, the persistence parameter is not unity however, but only nearly so. The proposed estimator (7) allows some flexibility with respect to ϕ by estimating it by maximum a posteriori (7) given the selected universal parameter $\lambda_T^{\ell_1-\mathrm{SVM}}$. Pursuing the idea of a universal parameter further, one could derive a prior π_λ for λ based on the universal parameter to derive an information criterion. Indeed one cannot simply minimize (7) over λ as well, because the solution would lead to $\lambda = 0$, but one could add $-\log \pi_\lambda(\lambda)$ to (7) to define an information criterion for λ, and optimize over it as well. This would require

deriving an appropriate prior for λ as in Sardy (2010).

With the choice of the universal penalty $\lambda_T^{\ell_1-\mathrm{SVM}}$, ℓ_1-SVM is able to adaptively alternate between smooth periods separated by jumps, i.e., it is able to find an adaptive segmentation of the volatility process that switches between regimes, and to estimate the location and amplitude of the jumps.

2.3 Optimization issues

We study here how to solve (7) to obtain the proposed estimate. To solve (7) in $(\phi, \mu, \boldsymbol{h})$, we use a decomposition approach that alternately solves in $(\mu, \boldsymbol{h})$ with ϕ held fixed, and in ϕ with $(\mu, \boldsymbol{h})$ held fixed. This alternating minimization approach, though not guaranteed to converge to a global minimum, works well in practice. How to solve each subproblem? For a fixed $(\mu, \boldsymbol{h})$, the objective function of (7) is convex piecewise-linear in ϕ and the minimum can be found by, e.g., sorting the breakpoints. In what follows, we focus on solving in $(\mu, \boldsymbol{h})$, with ϕ held fixed. The resulting subproblem can be written compactly as

$$\min_{\boldsymbol{h},\mu} \sum_{t=1}^{T} g_t(h_t) + \pi(B_\phi \boldsymbol{h} + \mu(\phi - 1)\mathbf{1}), \tag{9}$$

where $\mathbf{1}$ denotes the $T-$vector of ones, $g_t(h_t) = \log \varphi(h_t) - \log\{f_\epsilon(y_t/\varphi(h_t))\}$, $(B_\phi \boldsymbol{h})_t = h_{t+1} - \phi h_t$ for $t = 1, \ldots, T-1$ and $\pi(\cdot) = \lambda \| \cdot \|_1$. Focusing on the link $\varphi(\cdot) = \exp(\cdot)$ and on Gaussian ϵ-innovations, it can be seen that $g_t(h_t) = h_t + \frac{1}{2} y_t^2 \exp(-2h_t) +$ constant, which is strictly convex. Hence (9) is a convex optimization problem.

The case of $\phi = 1$ can be solved using the iterative dual mode (IDM) algorithm. Specifically, Theorem 3 of Sardy and Tseng (2004) applies with $g_t^*(u_t) = \frac{1}{2}(1-u_t)(\log(\frac{1-u_t}{y_t^2}) - 1)$ for Gaussian ϵ-innovations, so that a coordinate descent algo-

rithm can be employed on the dual problem in $\mathbf{u} = (u_1, ..., u_T)$. We consider below the more challenging case of $\phi \neq 1$, for which IDM is impractical. As we show below, in this case (9) can be efficiently solved by a primal-dual interior-point algorithm which can be employed for all positive ϕ except $\phi = 1$, which includes $\phi > 1$.

2.3.1 Dual formulation

Using $\varphi(\cdot) = \exp(\cdot)$, we derive in Appendix C the dual of the primal subproblem (9). It has the general form

$$\min \sum_{t=1}^{Q} q_t(x_t) \quad \text{s.t.} \quad A\mathbf{x} = b,\ \mathbf{x} \geq \mathbf{0}, \tag{10}$$

where $\mathbf{x}$ is the dual vector, the matrix A has $Q = 3T-2$ columns and $q_t(\cdot)$ is a function assumed to be convex, twice differentiable on $(0, \infty)$ with $\lim_{\xi \to 0} q_t(\xi) = q_t(0)$, with $q_t'(\cdot)$ concave and which satisfies

$$(\xi + \delta)\left(q_t'(\xi + \delta) - q_t'(\xi) - q_t''(\xi)\delta\right) \geq -\kappa q_t''(\xi)\delta^2 \quad \text{whenever} \quad \frac{|\delta|}{\xi} \leq \rho, \tag{11}$$

for some $\kappa > 0$ and $0 < \rho < 1$. In particular, Appendix D shows that (11) is satisfied with $\kappa = \frac{1}{2(1-\rho)}$ for Gaussian noise and exponential link. Specifically, we can take $q_t(x_t) = x_t \log(x_t) + c_t x_t$ with $c_t = -(1 + \log(y_t^2))$, $t = 1, \ldots, T$.

2.3.2 Log-barrier problem

The log-barrier problem, parameterized by $\varepsilon > 0$, is

$$\min \sum_{t=1}^{Q} q_t(x_t) - \varepsilon \log(x_t) \quad \text{s.t.} \quad A\mathbf{x} = \mathbf{b},\ \mathbf{x} > \mathbf{0},$$

with Karush-Kuhn-Tucker condition $A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} > \mathbf{0}, \quad q'(\mathbf{x}) - \varepsilon X^{-1}\mathbf{1} - A^\top \mathbf{u} = \mathbf{0}$, where $q'(\mathbf{x}) = (q_t'(x_t))_{t=1}^Q$ and $X = \mathrm{diag}(x_1, \ldots, x_Q)$. This can be rewritten as

$$A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} > \mathbf{0}, \quad \mathbf{s} = q'(\mathbf{x}) - A^\top \mathbf{u}, \quad X\mathbf{s} = \varepsilon \mathbf{1}. \tag{12}$$

The exact solution of (12) traces the central path as ε ranges over $(0, \infty)$. The primal-dual interior-point algorithm solves the equations approximately using damped Newton method and decreases ε after each iteration. Specifically, $(\mathbf{x}, \mathbf{u})$ is an approximate solution of (12) if it belongs, together with ε, to the following so-called "wide neighborhood' of the central path:

$$\mathcal{N}(\tau) = \left\{ (\mathbf{x}, \mathbf{u}, \varepsilon) \mid A\mathbf{x} = b,\ \mathbf{x} > \mathbf{0},\ \mathbf{s} = q'(\mathbf{x}) - A^\top \mathbf{u},\ \min_t x_t s_t \geq \tau\varepsilon,\ \varepsilon = \frac{\mathbf{x}^\top \mathbf{s}}{Q} \right\},$$

with $0 < \tau < 1$; see (Wright, 1997) and references therein.

2.3.3 Primal-dual interior-point algorithm

The algorithm begins with any $(\mathbf{x}, \mathbf{u}, \varepsilon) \in \mathcal{N}(\tau)$. Then it solves the Newton equation

$$\begin{aligned} X d_s + S d_x &= \delta\varepsilon \mathbf{1} - X\mathbf{s}, & (13) \\ A d_x &= 0, & (14) \\ q''(\mathbf{x}) d_x - A^\top d_u &= d_s, & (15) \end{aligned}$$

for (d_x, d_s, d_u), where $\mathbf{s} = q'(\mathbf{x}) - A^\top \mathbf{u}$, and $0 < \delta < 1$. Let

$$\mathbf{x}[\alpha] = \mathbf{x} + \alpha d_x, \quad \mathbf{u}[\alpha] = \mathbf{u} + \alpha d_u, \quad \mathbf{s}[\alpha] = q'(\mathbf{x}[\alpha]) - A^\top \mathbf{u}[\alpha] \quad \forall \alpha > 0.$$

Let ν and $\bar{\alpha}$ be given by (E.9) and (E.10) in Appendix E. Then, beginning with $\alpha = 1$, it checks if

$$(\mathbf{x}[\alpha], \mathbf{u}[\alpha]) \in \mathcal{N}(\tau), \qquad \varepsilon[\alpha] = \frac{\mathbf{x}[\alpha]^\top \mathbf{s}[\alpha]}{Q} \leq (1 - \bar{\alpha}\nu)\varepsilon, \tag{16}$$

and if not, it decreases α by some factor $0 < \varrho < 1$ and repeat, until (16) is satisfied. Then we update

$$(\mathbf{x}^{\text{new}}, \mathbf{u}^{\text{new}}, \varepsilon^{\text{new}}) \longleftarrow (\mathbf{x}[\alpha], \mathbf{u}[\alpha], \varepsilon[\alpha]),$$

and re-iterate, until $\varepsilon \leq \varepsilon^{\text{final}}$. In our implementation, we use $\tau = 10^{-4}$, $\delta = 0.5$, $\varrho = 0.7$, $\rho = 0.99$, we initialize by $\mathbf{h} = \frac{1}{\alpha}\mathbf{1} - \mathbf{c}$, $\mathbf{w} = \alpha\lambda\mathbf{1}$, where $\mathbf{c} = (c_1, ..., c_T)$. (which uniquely determine $\mathbf{u}$ and $\mathbf{x}$), $\mathbf{s} = q'(\mathbf{x}) - A^\top \mathbf{u}$, and $\varepsilon = \frac{\mathbf{x}^\top \mathbf{s}}{Q}$, with $0 < \alpha < 1$ chosen so that $(\mathbf{x}, \mathbf{u}, \varepsilon) \in \mathcal{N}(\tau)$.

Our `Matlab` code is part of the online supplemental materials and is fast since its complexity is bounded by $O(Q^{3/2})$, as we now show.

2.3.4 Iteration complexity

Appendix E shows that (16) is satisfied when $\alpha = \bar{\alpha}$ given by (E.10). Thus ε decreases by a factor of at most $1 - \bar{\alpha}\nu$ after each iteration so that, after k iterations, $\varepsilon \leq (1-\bar{\alpha}\nu)^k \varepsilon^{\text{init}}$. Thus $\varepsilon \leq \varepsilon^{\text{final}}$ whenever $k \geq \log\left(\frac{\varepsilon^{\text{init}}}{\varepsilon^{\text{final}}}\right) \frac{1}{-\log(1-\bar{\alpha}\nu)}$. Since $\log(1-\bar{\alpha}\nu) \leq -\bar{\alpha}\nu$ and, by (E.10), $\frac{1}{\bar{\alpha}} = O(\kappa Q^{3/2} + Q)$, this shows that the number of iterations until termination is at most

$$\log\left(\frac{\varepsilon^{\text{init}}}{\varepsilon^{\text{final}}}\right) \frac{1}{\bar{\alpha}\nu} = O\left((\kappa Q^{3/2} + Q) \log\left(\frac{\varepsilon^{\text{init}}}{\varepsilon^{\text{final}}}\right)\right),$$

where $Q = 3T - 2$ and T is the length of the time series. While there have been previous studies of path-following algorithms for entropic optimization of the form (10) and (11) (Potra and Ye, 1993; Tseng, 1992), these algorithms use the so-called "narrow neighborhood', which is not practically efficient. To our knowledge, this is the first study of a path-following algorithm for entropic optimization that uses the wide neighborhood and is practically and theoretically efficient. Specifically, when (10) is a

linear or a convex quadratic program, i.e., $\kappa = 0$ in (11), the above complexity result is the best known for an algorithm using the wide neighborhood; see Wright (1997, Theorem 5.11).

2.4 Simulated time series

We simulate data $y_t = \sigma_t \epsilon_t$, where $\epsilon_t \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$ and $\sigma_t = \sigma(t)$ is a smooth volatility function with periods of abrupt changes of regime and volatility peaks, as one may expect in the financial markets. To that aim we take the sum of two standard benchmark functions, the `blocks` and `bumps` functions, defined on $[0,1]$ by Donoho and Johnstone (1994) for wavelet smoothing:

$$\sigma(t) = \alpha + \beta\{\text{blocks}((t-1)/(T-1)) + \text{bumps}((t-1)/(T-1))\},\ t = 1, \ldots, T$$

where α and β rescale the sum of `blocks` and `bumps` so as to have a range of volatility with a minimum value of exactly 0.1 and a maximum value of exactly 10. The log of the volatility function σ_t for $t = 1, \ldots, 5000$ is the curve plotted on Figure 1 (c), where the dots are the maximum likelihood estimates $\log \hat{\sigma}_t^{\mathrm{MLE}} = \log |y_t|$. Figure 1 (a) shows the simulated returns y_t, and (b) shows the empirical autocorrelation function (acf) of the absolute returns, which reflects potential phenomena observed on real financial time series, such as volatility clustering, nonstationarity or long memory. Figure 1 (e) shows ℓ_1-SVM log-volatility estimates solution to (7) with $\varphi(\cdot) = \exp(\cdot)$, using the universal penalty λ_T derived in Section 2.2, while (d) shows the acf of the fitted absolute residuals $|y_t|/\hat{\sigma}_t$ and (f) shows the quantile-quantile plot of the fitted residuals. We see with these simulated data that the estimation of the underlying volatility captures the important features of the true volatility, and that the fitted

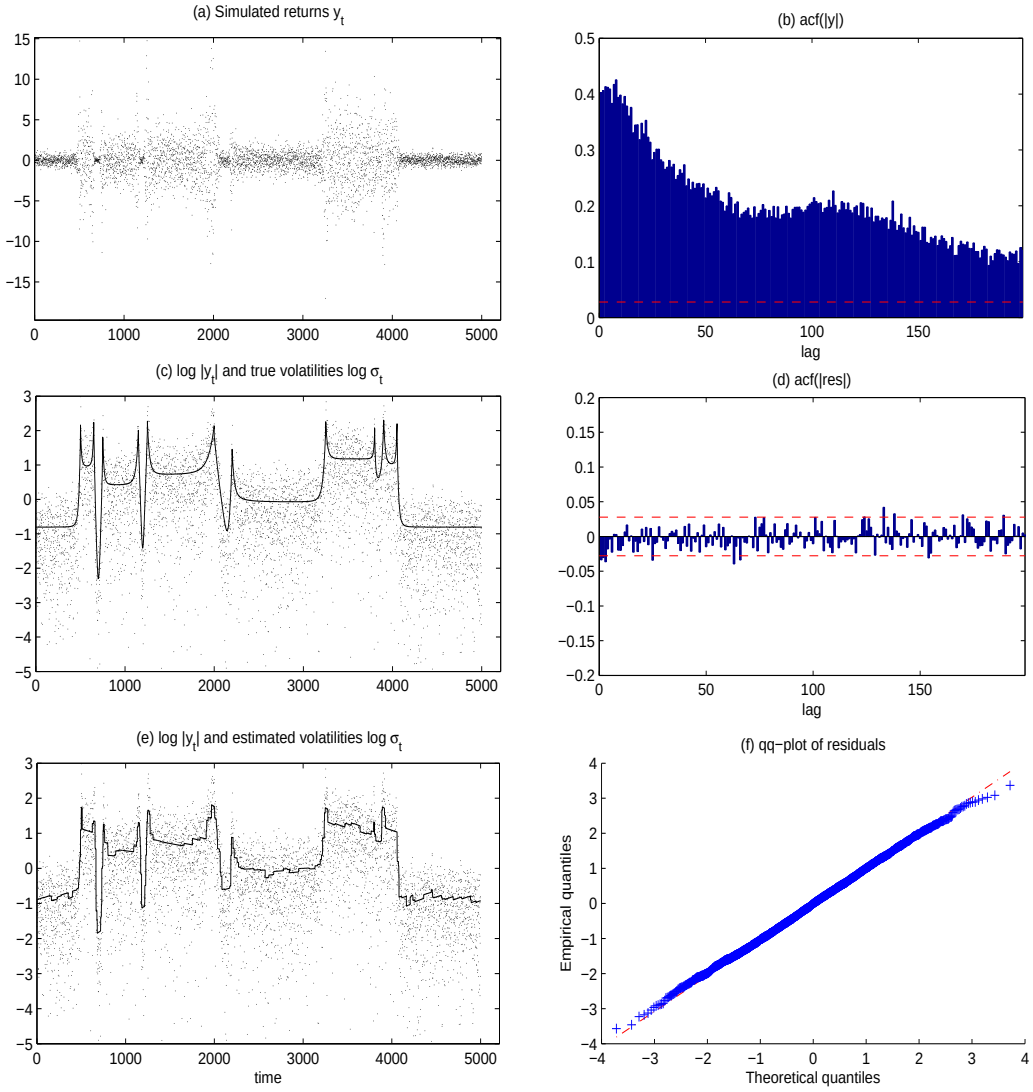

Figure 1: Simulation. Top: (a) simulated log-returns y_t and (b) empirical acf of $|y_t|$; Middle: (c) true log-volatilities σ_t (line) and log-absolute-returns $\log|y_t|$ (dots), and (d) acf of residuals $|y_t|/\hat{\sigma}_t$; Bottom: (e) log-volatilities (line) estimated with the universal penalty $\lambda_T^{\ell_1-\mathrm{SVM}}$ and $\log|y_t|$ (dots), and (f) Gaussian qq-plot of residuals.

residual process matches well the i.i.d. Gaussian assumption.

3 Application

We illustrate our methodology on the DOW JONES, NASDAQ and S&P500 stock index returns. Data are the $T = 5212$ daily closing prices P_t between 2 January 1990 until 2 September 2010 data; they are part of the online supplemental materials. Stock returns y_t are computed as $100 \log (P_t/P_{t-1})$. In Section 3.1 we first consider the entire sequence of returns, apply our volatility estimator to it and analyze the results. Note that the estimation for such time-series took less than 2 minutes running a `Matlab` code (available in online supplemental materials) on a standard laptop computer. In Section 3.2 we evaluate and compare the forecasting performance of our estimate to that of log-Normal SVM, GARCH and IGARCH for short and long horizons.

3.1 Volatility smoothing

The volatility clustering feature is reflected by the strong autocorrelation in the absolute values of the returns on Figures 2 (b) for the NASDAQ. Similar behaviors are observed with the DOW JONES and S&P500, so that we only report on the NASDAQ. Our volatility estimation is jointly plotted with the realized volatility estimator

$$\mathrm{RV}_{t,\tau} = \sqrt{\frac{1}{\tau} \sum_{i=0}^{\tau-1} y_{t-i}^2} \tag{17}$$

with $\tau = 10$ of Andersen, Bollerslev, and Labys (2001) in graphs (c) on a log-scale, and graphs (d) and (f) are diagnostics plots on the residuals. As expected the autocorrelation in the absolute rescaled residuals, $|y_t|/\hat{\sigma}_t$, has been removed (see graphs

(d)). And the normality holds quite well, except in the negative extreme tail (see graphs (f)). We indeed observe some asymmetry, a phenomenon known as *leverage effect* (Nelson, 1991; Glosten, Jagannathan, and Runkle, 1993): the responses of the volatility to the negative shocks (bad news) are stronger than the ones to the positive shocks. One avenue to model the leverage effect with ℓ_1-SVM would consist in introducing a correlation between ϵ- and η-innovations as in Omori, Chib, Shephard, and Nakajima (2007), or in splitting the joint distribution of η and ϵ in terms of marginal and dependence structure, which could be parametrized as a Gaussian copula. Or one could fit the asymmetry using an asymmetric density f_ϵ in (7). We do not pursue these ideas further here.

For the DOW JONES, NASDAQ and S&P500, the estimated persistence parameter are $\hat{\phi} = 0.9986$, $\hat{\phi} = 0.9995$ and $\hat{\phi} = 0.9986$, respectively. Note that these estimates were not constrained to be less than one, so that stationarity seems plausible here. To derive the variance-covariance matrix of the parameter estimates, and in particular of $\hat{\phi}$, we observe that for given λ and $(h_t)_{t=1}^T$, the second term of the objective function (7) is a least absolute error regression problem of $(h_t)_{t=2}^T$ on $(h_t)_{t=1}^{T-1}$ with slope ϕ and intercept $\tilde{\mu} = \mu(\phi - 1)$. Bassett and Koenker (1978) provide asymptotic theory for least absolute fit. In particular the asymptotic covariance is given by $\mathbb{V}((\hat{\tilde{\mu}}, \hat{\phi})^\top) = \frac{1}{\lambda^2} Q^{-1}$ with $\lambda = \lambda_T^{\ell_1\text{-SVM}}$ of Section 2.2, where $Q = X^\top X$ with $X = (\mathbf{1}, \mathbf{h}_{-1})$, and $\mathbf{h}_{-1} = (h_1, ..., h_{T-1})$, and with the standard assumption $\lim_{T\to\infty} \frac{1}{(T-1)} X^\top X$ is positive definite. Here the vector $\mathbf{h}_{-1}$ is not observable, but is estimated by ℓ_1-SVM. Computing the matrix Q with the estimated log-volatilities $\hat{\mathbf{h}}_{-1}$ for $\mathbf{h}_{-1}$ provides an approximate covariance matrix. For the time series considered

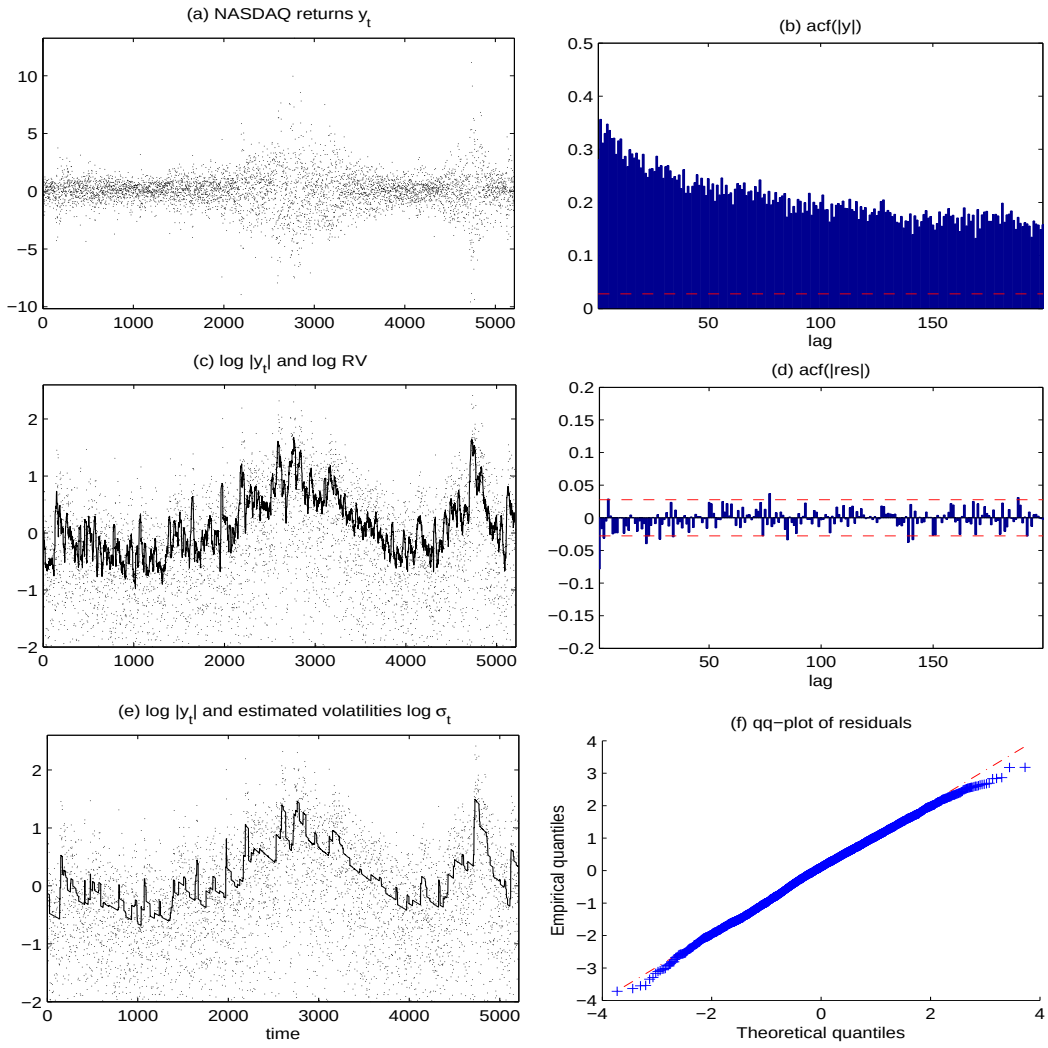

Figure 2: NASDAQ. Top: (a) log-returns y_t from 2 January 1990 until 2 September 2010 and (b) empirical acf of $|y_t|$; Middle: (c) realized volatilities from (17) (line) and log-absolute-returns $\log|y_t|$ (dots), and (d) acf of residuals $|y_t|/\hat{\sigma}_t$; Bottom: (e) log-volatilities (line) estimated with the universal penalty $\lambda_T^{\ell_1-\mathrm{SVM}}$ and $\log|y_t|$ (dots), and (f) Gaussian qq-plot of residuals.

the estimated standard deviations of $\hat{\phi}$ for the DOW JONES, NASDAQ and S&P500 are 0.0053, 0.0044 and 0.0049.

Figure 3 plots some interesting comparisons of our estimate (middle graph) for the S&P500 returns with the VIX (a popular measure of the implied volatility of S&P500 index options plotted in the top graph), with the realized volatility (17) (second from top), with log-Normal SVM (Liesenfeld and Richard, 2003) (second graph from bottom), and with GARCH(1,1) (bottom graph). As a whole, Figures 2 and 3 show how well ℓ_1-SVM captures small and large jumps. In particular, ℓ_1-SVM can on the one hand fit abrupt jumps, and on the other hand be smooth in a quiet volatility period. On the contrary, the other estimators considered seem to either oversmooth jumps or undersmooth quiet periods. In the following section, we perform some backtesting to quantify and compare the predicting power of these estimators.

Finally Figure 4 compares the evolution of estimated log-volatilities for both DOW JONES and NASDAQ with ℓ_1-SVM. As expected the NASDAQ is more volatile than the DOW JONES, except for the recent crashes of Autumn 2008 which affected all economic sectors. Some important market turbulence periods have been identified, and correspond to some of the largest abrupt changes estimated by ℓ_1-SVM.

3.2 Volatility forecasting

To evaluate the forecasting performance of various models, we do a back-test, that is, we calibrate on time series running from day one until day $t = 3000 + (k-1)H/2,\ k = 1, 2, \ldots$ and forecast the volatility out-of-sample from day $t + 1$ until day $t + H$. We choose two horizons: a short horizon $H = 20$ business days (one month) and

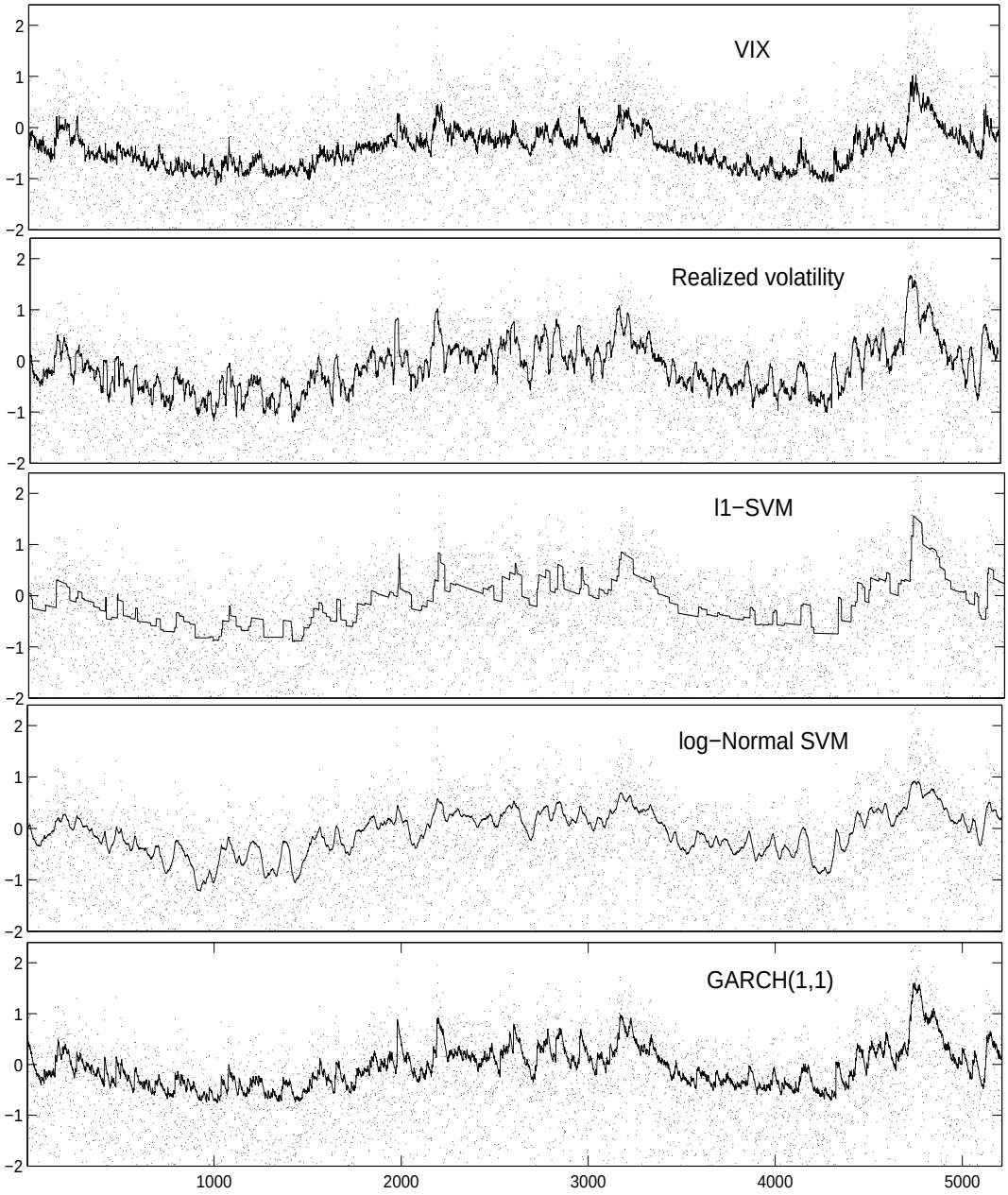

Figure 3: S&P500. (on a log-scale truncated below) $\hat{\sigma}_t^{\mathrm{MLE}} = |y_t|$ (dots) versus five other volatility estimates (lines).

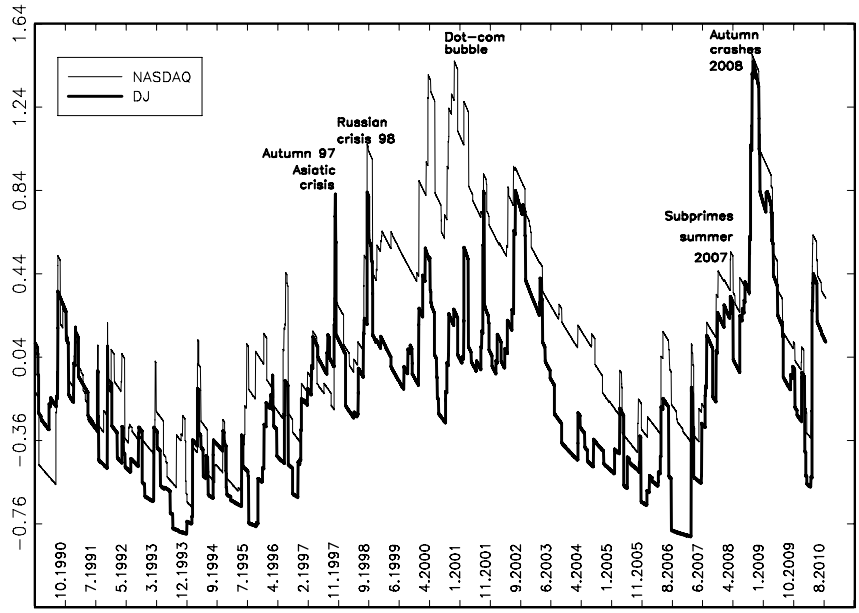

Figure 4: DOW JONES and NASDAQ's estimated log-returns from 29 December 1989 until 2 September 2010.

long horizon $H = 120$ days (six months). The calibration up to time t provides parameter estimates for forecasting the volatility beyond time t. With the notation $\sigma^2_{t+j} = \mathbb{E}(\varphi(h_{t+j})|y_1, ..., y_t)$, the forecasted volatilities are given by:

$$\sigma^{2,\ell_1/\ell_2-\mathrm{SVM}}_{t+j} = \exp(2(\mu + \phi^j(h_t - \mu))), \quad j = 1, \ldots, H \tag{18}$$

for SVM (ℓ_1 or log-Normal ℓ_2), by

$$\sigma_{t+j}^{2,\mathrm{GARCH}} = (\psi+\phi)^j \sigma_t^2 + \omega \sum_{k=0}^{j-1} (\psi+\phi)^k, \quad j = 1, \ldots, H \tag{19}$$

for GARCH, and by

$$\sigma_{t+j}^{2,\mathrm{IGARCH}} = \sigma_t^2 + j\omega, \quad j = 1, \ldots, H \tag{20}$$

for IGARCH. We consider the median absolute error (MAE) forecasting measure between the forecasted volatilities $\hat{\sigma}_{t+h}$ and the realized volatilities (RV_{t+h} in (17))

$$\mathrm{MAE}^*(H) = \mathrm{median}\{|\sum_{h=1}^{H} \hat{\sigma}_{t+h}^{2,*} - \sum_{h=1}^{H} \mathrm{RV}_{t+h}^2|,\ t = 3001 + (k-1)H/2,\ k = 1, 2, \ldots\},$$

where "*" stands for ℓ_1-SVM, log-Normal SVM, GARCH or IGARCH.

Table 1: Volatility forecast based on relative median absolute errors (MAE) for a short ($H = 20$) and long ($H = 120$) horizons.

	DOW JONES		NASDAQ		S&P500	
Horizon $H =$	20	120	20	120	20	120
log-Normal SVM/ℓ_1-SVM	0.92	1.16	1.22	1.25	0.89	1.22
GARCH/ℓ_1-SVM	1.14	1.54	1.06	1.78	1.07	1.79
IGARCH/ℓ_1-SVM	1.15	2.18	1.10	2.33	1.10	2.35

We report in Table 1 the relative MAE with respect to that of ℓ_1-SVM (i.e., $\mathrm{MAE}^*/\mathrm{MAE}^{\ell_1-\mathrm{SVM}}$). A ratio larger than 1 means better forecasting for ℓ_1-SVM. The results show that ℓ_1-SVM outperforms log-Normal SVM, GARCH and IGARCH, especially for a long horizon. Hence our methodology appears better for the purpose of option pricing and portfolio management, which require long term forcast.

4 Conclusion

The semi-parametric ℓ_1-SVM estimator estimates the volatility process of a financial asset as a smooth curve between occurrences of jumps of varying amplitude. Our approach combines the dynamic of the well known stochastic volatility models and the properties of ℓ_1 Markov random field smoothing in order to estimate the second moment of hidden Markov chains while fitting the stylized features observed in finance. The performance of the proposed estimator is illustrated through simulations and empirical applications based on the DOW JONES, NASDAQ and S&P500. Our volatility forecast outperforms log-Normal SVM and GARCH's forecasts, especially for a long horizon. The following multivariate extension is promising, especially given our selection of the smoothing parameter, the speed of our algorithm and the good forecasting performance. For a portfolio made of Q financial assets, a possible multivariate ℓ_1-SVM model is

$$y_t^{(q)} = \sigma_t^{(q)} \epsilon_t^{(q)}, \tag{21}$$

$$\sigma_t^{(q)} = \varphi\left(\mu_q + \rho_q h_t\right), \tag{22}$$

$$h_t = \mu + \phi(h_{t-1} - \mu) + \eta_t, \tag{23}$$

for each asset $q = 1, \ldots, Q$ in the portfolio. Hence each asset has its own average volatility μ_q and constant of proportionality ρ_q with respect to a volatility h_t and a persistence parameter ϕ common to all assets.

5 Supplemental materials

`Matlab` **code:** the main code `stationMRFest_ip.m` is the interior point algorithm. The estimation of ϕ requires `stationMRFest_ip.m`. The simulation requires `wavebumps.m` and `waveblocks.m`. The code `Mrf_L1L2_Dual_1D.m` calculates the estimator when $\phi = 1$. The codes `fig1papersimu.m`, `fig2paperappli.m` and `fig3paperappli.m` reproduce Figures 1, 2 and 3.

Data: (ASCII files) the file "rdt_sp_nq_dj.txt" contains the S&P500, NASDAQ and DOW JONES log-returns columnwise; likewise, "volreal_sp_nq_dj.txt" contains the realized volatilities of the same corresponding indices. The files "VIX.txt" contains the VIX data, "GARCH_Ht.txt" the GARCH estimate and "volL2svm.txt" the log-Normal SVM estimate needed for Figure 3.

Appendix: the file "L1SVM-appendix.pdf" contains Appendices A to E.

6 Acknowledgements

We would like to thank three anonymous referees, the Associate Editor and the Editor for a careful review of the article, who helped improve its quality. We thank Roman Liesenfeld for sending us his code for estimating log-Normal SVM models. The first two authors would like to pay tribute to the third author, who was last seen in 2009 kayaking in the Yangtze River. More than a colleague, Paul was a friend.

References

Aït-Sahalia, Y. and Jacod, J. (2007). Volatility estimators for discretaly sampled lévy processes'. *The Annals of Statistics* **35**, 355–392.

Andersen, T., Bollerslev, T. Diebold, F., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association* **96**, 42–55.

Andersen, T. G. (1994). Stochastic autoregressive volatility: A framework for volatility modeling. *Mathematical Finance* **4**, 75–102.

Bassett, G. J. and Koenker, R. (1978). Asymptotic theory of least absolute error regression. *Journal of the American Statistical Association* **73**, 618–622.

Bates, D. (2000). Post-'87 crash fears in the s&p 500 futures option market. *Journal of Econometrics* **94**, 181–238.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society, Series B* **48**, 192–236.

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

Box, G. E. P. and Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association* **77**, 209–210.

Cai, J. (1994). A Markov model of switching-regime ARCH. *Journal of Business and Economic Statistics* **12**, 309–316.

Chan, H. W. and Maheu, J. M. (2002). Conditional jump dynamics in stock market returns. *Journal of Business and Enonomic Statistics* **20**, 377–389.

Chernov, M., Gallant, A. R., Ghysels, E., and Tauchen, G. (1999). A new class of stochastic volatility models with jumps: Theory and estimation. Technical report, CIRANO, Montréal.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**, 425–455.

Donoho, D. L., Johnstone, I. M., Hoch, J. C., and Stern, A. S. (1992). Maximum entropy and the nearly black object (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* **54**, 41–81.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptopia? (with discussion). *Journal of the Royal Statistical Society, Series B* **57**, 301–369.

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

Eraker, B., Johannes, M., and Polson, N. (2008). The impact of jumps in volatility and returns. *The Journal of Finance* **58**, 1018–1031.

Geman, S. and Geman, D. (1984). Stochastic relaxation. Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **61**, 721–741.

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

Hamilton, J. D. and Susmel, R. (1994). Autoregressive conditional heteroskedastic-

ity and changes in regime. *Journal of Econometrics* **64**, 307–333.

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

Hull, J. and White, A. (1987). The pricing of options on assets with stochastic volatilities. *Journal of Finance* **42**, 281–300.

Kotz, S., Kozubowski, T. J., and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: a Revisit with Applications to Communications, Economics, Engineering, and Finance*. Birkhaeuser Verlag.

Li, H., Wells, M., and Yu, C. (2008). A bayesian analysis of return dynamics with lvy jumps. *Review of Financial Studies* **21**, 2345–2378.

Liesenfeld, R. and Richard, J. (2003). Univariate and multivariate stochastic volatility models: Estimation and diagnostics. *Journal of Empirical Finance* **10**, 505–531.

Madan, D. B. and Seneta, E. (1990). The variance gamma (V.G.) model for share market returns. *The Journal of Business* **63**, 511–524.

Nelson, D. (1991). Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* **59**, 397–370.

Omori, Y., Chib, S., Shephard, N., and Nakajima, J. (2007). Stochastic volatility with leverage: fast and efficient likelihood inference. *Journal of Econometrics* **140**, 425–449.

Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics* **63**, 3–50.

Potra, F. and Ye, Y. (1993). A quadratically convergent polynomial interior-point algorithm for solving entropy optimization problems. *SIAM Journal on Optimization* **3**, 843–860.

Rydberg, T. H. (2000). Realistic statistical modelling of financial data. *International Statistical Review 68*(3), 233–258.

Sardy, S. (2010). Adaptive posterior mode estimation of a sparse sequence for model selection. *Scandinavian Journal of Statistics* **36**, 577–601.

Sardy, S. and Tseng, P. (2004). On the statistical analysis of smoothing by maximizing dirty markov random field posterior distributions. *Journal of the American Statistical Association* **99**, 191–204.

Sardy, S. and Tseng, P. (2010). Density estimation by total variation penalized likelihood driven by the sparsity ℓ_1 information criterion. *Scandinavian Journal of Statistics* **37**, 321–337.

Shephard, N. (1996). Statistical aspects of arch and stochastic volatility. In *Time Series Models in Econometrics, Finance and Other Fields*, pp. 1–67. London: Chapman and Hall: Eds. D.R. Cox, D. V. Hinkley and O. E. Barndorff-Nielsen.

Starica, C. (2003). Is garch(1,1) as good a model as the nobel prize accolades would imply? *Working paper*.

Starica, C. and Granger, C. (2005). Nonstationarities in stock returns. *The Review of Economics and Statistics* **87**, 503–522.

Taylor, S. (1986). *Modelling Financial Time Series*. John Wiley.

Taylor, S. (1994). Modelling stochastic volatility: a review and comparative study.

Mathematical Finance **4**, 183–204.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **4**, 1035–1038.

Tseng, P. (1992). Global linear convergence of a path-following algorithm for some monotone variational inequality problems. *Journal of Optimization Theory and Applications* **75**, 265–279.

Wright, S. (1997). *Primal-Dual Interior-Point Methods*. Philadelphia: SIAM.