# $\ell_1-$Penalized Likelihood Smoothing of Volatility Processes allowing for Abrupt Changes

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

April 2009

**Abstract**

We consider the problem of estimating the volatility of a financial asset from a time series record of length $T$. We believe the underlying volatility process is smooth, possibly stationary, and with potential abrupt changes 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. Our Bayesian posterior mode estimate is the solution to an $\ell_1$-penalized likelihood optimization that we solve with an interior point algorithm that is efficient since its complexity is bounded by $O(T^{3/2})$. We apply our volatility estimator to real financial data, diagnose the model and perform back-testing to investigate to forecasting power of the method by comparison to (I)GARCH.

*Keywords*: $\ell_1$ penalty, Markov random field, stochastic volatility model, smoothing, wavelet, extreme value theory, forecasting, GARCH.

*JEL classification*: C14, C22, C61.

[(a)] Department of Econometrics, University of Geneva, CH-1211 Geneva 4, Switzerland.
E-mail: david.neto@unige.ch).

[(b)] Department of Mathematics, University of Geneva, CH-1211 Geneva 4, Switzerland.
E-mail: Sylvain.Sardy@unige.ch).

[(c)] Department of Mathematics, University of Washington, Seattle, WA.
E-mail: tseng@math.washington.edu).

# 1 Introduction

Suppose we observe values $y_t$ of a financial asset (stock-returns, interest-rates or exchange-rates) 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 $\sigma_t$ is a measure of volatility, and $\epsilon_t$ is white standard Gaussian noise. Without additional assumption, the maximum likelihood estimate is $\hat{\sigma}_t^{\mathrm{MLE}} = |y_t|$, which is practically useless due to its nonsmoothness (i.e., high variability). A temporal structure for $\sigma_t$ is assumed to regularize the maximum likelihood estimation, so as to obtain a smoother estimate of $\sigma_t$ while capturing the stylized features (Rydberg 2000) observed in financial econometric, like heavy tails of the marginal distributions, volatility clustering or evolution with possible abrupt changes (e.g., peaks, discontinuities). 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 distributions of the financial data than the standard GARCH model (Shephard 1996). Stochastic volatility models are the empirical discrete-time analogs of 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 of Andersen (1994) includes GARCH and stochastic volatility models as particular cases. It defines the discrete polynomial stochastic autoregressive volatility process as:

$$
\begin{aligned}
y_t &= \sigma_t \epsilon_t, && (1)\\
\sigma_t^q &= \varphi(h_t), && (2)\\
h_t &= \omega + \phi h_{t-1} + (\gamma + \psi h_{t-1})\, \eta_t, && (3)
\end{aligned}
$$

where $q \in \{1, 2\}$, $\varphi(\cdot)$ is a positive continuous and strictly monotone function, the autoregressive process $h_t$ has positive coefficients $\psi, \phi, \gamma$ such that $\psi + \phi > 0$ and $\psi + \phi > 0$, the errors $\eta_t$ and $\epsilon_t$ are i.i.d. and mutually independent with mean-variance of $\left(1, \sigma_\eta^2\right)$ and $(0, 1)$, respectively. This process includes, for instance, the GARCH process (corresponding to $q = 2$, $\varphi(h_t) = h_t$, $\eta_t = \epsilon_{t-1}^2$, and $\gamma = 0$) and the stochastic volatility model (SVM), corresponding to $\varphi(h_t) = \exp(h_t)$, $q = 1$ and $\psi = 0$. Often $\eta_t$ and $\epsilon_t$ are assumed to be Gaussian.

One challenge for these models has been how to infer the parameters $(\psi, \phi, \gamma, \omega)$ and how to estimate the hidden conditional volatility $\sigma_t$ 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 a non-linear hidden Markov chain. Several methods have been proposed to estimate the latter data generating process, many of which are based on calculating or estimating the conditional expectations

$$(\hat{\psi}, \hat{\phi}, \hat{\gamma}, \hat{\omega}) = \mathbb{E}((\psi, \phi, \gamma, \omega) \mid \mathbf{y}; \sigma_\eta^2) \quad \text{and} \quad \hat{h}_t = \mathbb{E}\left(h_t | \mathbf{y}; \sigma_\eta^2\right),\ t = 1, \ldots, T, \tag{4}$$

where $\mathbf{y} = (y_1, ..., y_T)$ are the returns up to time $T$. For a linear and Gaussian data generating process, the Kalman filter calculates a closed form expression for the posterior means (4). The quasi-maximum likelihood approach of Harvey, Ruiz and Shephard (1994) applies Kalman filtering to the logarithm of the squared observations $\log y_t^2 = 2h_t + \zeta_t$, where $\zeta_t = \log \epsilon_t^2$. However, the data generating process is non-Gaussian, and the poor approximation of $\zeta_t$ by a Gaussian distribution leads to poor estimation of $\sigma_t$, particularly when the variance $\sigma_\eta^2$ of the log-volatility $\eta_t$ in (3) is small (Jacquier, Polson and Rossi 1994). Another possible approach is to generate by Monte-Carlo an ergodic chain from the exact posterior distribution given the observations (Jacquier et al. 1994) and calculate the componentwise averages to estimate the conditional expectations (4). However, this may require generating long sample paths before reaching convergence. The estimate of the variance $\sigma_\eta^2$ of the log-volatility innovations also remains a difficult issue.

Based on a GARCH model, the estimated persistence of the conditional volatilities of many financial time series is seemingly high, which also makes the inference difficult. For example the second moment of an Integrated GARCH (Engle and Bollerslev 1986) does not exist even if the stationarity conditions are met. One also wonders whether the volatility process is really stationary, and in fact, it is now believed that such strong persistence is due to misspecification modeling. For instance Diebold (1986) argued that the persistence may be due to the instability of the unconditional variance. Lastrapes (1989) observed that the persistence of exchange-rate volatility depends on U.S. monetary policy regimes. This statement is confirmed by Lamoureux and Lastrapes (1990) who obtained a substantial lower persistence volatility measure (i.e. the sum of the GARCH parameters, $\psi + \phi$ in equation (3)) when some dummy variables, corresponding to some periods over which the volatility seems to be stationary, are included in the GARCH model. Cai (1994) and Hamilton and Susmel (1994) used Markov switching ARCH models to take into account the regime effect in

the volatility, and in both cases, the persistence is captured by the stochastic process which describes the state of the system, rather than by the GARCH parameters. Another widely used approach (see for instance Chan and Maheu (2002) and Chernov, Gallant, Ghysels and Tauchen (1999)) adds a jump process to capture large discrete changes. However, the empirical evidence suggests that the jump models capture less the persistence than additional extremes in the returns marginal distribution. Starica (2003) shares the concerns early stated by Diebold (1986) about persistence and misspecification, and discusses the danger of assuming global stationarity. Starica and Granger (2005) proposed a nonstationary approach to estimate volatility by considering the log-absolute returns as a locally stationary process as defined in Dahlhaus (1997). Their methodology considers the underlying process as a slowly and smoothly varying process with sharp changes.

The aforementioned results motivate the features we aim to reproduce with our proposed model and estimator, namely, that volatility evolves smoothly except for occasional abrupt changes whose transient effect prevents the market from returning quickly to the level before shock. However, while Starica and Granger's approach is based on successive stationarity tests over intervals of varying width to identify the times of regime switching, our $\ell_1$-based estimator finds the desired segmentation via the selection of a single parameter. Moreover, the solution of our $\ell_1$-penalized likelihood optimization problem is calculated with a fast algorithm. 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 which guarantees smoothness of the volatility estimate. 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 and NASDAQ, analyze the results and evaluate the forecasting performance. Section 4 draws some conclusions and points to future extensions.

# 2 Smoothing the volatility

## 2.1 $\ell_1$ penalized likelihood estimator (fSVM)

We consider a discrete time SVM with Laplace innovations $\eta_t$ whose density $f_\eta(x) = \frac{\lambda}{2}\exp(-\lambda|x|)$ has heavier tails than the Gaussian density to better reflect sudden changes of regime due to important market news. More precisely,

this model corresponds to (1)–(3) with $\psi = 0$, $q = 1$, $\gamma = 1$ and $\omega = \mu(1-\phi)$. Hence the variance of the log-volatility innovations is $\sigma_\eta^2 = 2/\lambda^2$ and this SVM is related to variance gamma (VG) models (Madan and Seneta 1990) for modeling the non-Gaussian nature of stock market returns and future index prices. Indeed the VG model can be expressed by equation (1), where $\sigma_t$ 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). Hence, the Laplace motion is defined as a Brownian motion, denoted $B(\tau)$, evaluated at random time distributed as a Gamma process, denoted $\gamma(t)$. This Brownian is said subordinated to the process $\gamma(t)$. Therefore, the Laplace motion is defined as $L(t) \stackrel{d}{=} B(\gamma(t))$. Another appeal of the Laplace motion is it can be written as a compound Poisson process with independent and random jumps. In this sense it is a pure jumps process able then to capture abrupt changes. In addition to VG model, our model considers an extra parameter to model and capture the persistence in volatility.

Using Bayes theorem with (1) for the noise model and (2)–(3) for the prior assumption on the volatility process, we derive the posterior distribution of the volatility given the returns $\mathbf{y}$. More precisely, by considering the negative log-posterior distribution, we define the maximum a posteriori estimate as the solution to

$$\min_{\phi,\mu,\mathbf{h}} \sum_{t=1}^{T} \log\varphi(h_t) - \log\{f_\epsilon(y_t/\varphi(h_t))\} + \lambda \sum_{t=2}^{T} |h_t - (\mu + \phi(h_{t-1} - \mu))|, \quad (5)$$

where $f_\epsilon$ denotes the density function of the error term $\epsilon_t$, $(h_t)_{t=1}^T$ are the $\varphi^{-1}$-volatilities, $\mu$ is an average volatility measure, and $\phi$ is the persistence parameter. The function $\varphi(\cdot)$ is a strictly monotone function that maps the estimand $h_t$ to the volatility $\sigma_t = \varphi(h_t)$. For instance, the exponential function used by the original stochastic volatility model conveniently maps $\mathbb{R}$ into $\mathbb{R}^+$ for a positive volatility. Our estimator has the advantage that the positivity constraint is already active with the first logarithmic term in (5) which acts as a barrier against negativity. Hence we can 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) that includes as special cases the exponential link (take $\delta \to 0$) and the linear link (take $\delta = 1$).

The first sum in (5) is the negative log-likelihood for (1)–(2) and the second sum stems from the autoregressive process prior (3) with Laplace $\eta$-innovations. The estimator is akin to a Tikhonov regularization (Tikhonov 1963) of the erratic maximum likelihood estimate, but using an $\ell_1$-based penalty and sufficiently large penalty parameter $\lambda > 0$ to enforce smoothness. The $\ell_1$ penalized

likelihood formulation of the estimator (5) draws connection to two nonparametric function estimation techniques: One is Markov random field smoothing (Geman and Geman 1984, Besag 1986, Sardy and Tseng 2004), which solves a similar problem

$$\min_{\boldsymbol{h}} -l(\boldsymbol{h};\mathbf{y}) + \lambda \sum_{t=2}^{T} |h_t - h_{t-1}|,$$

where $-l(\cdot;\cdot)$ is the negative log-likelihood of a noisy signal/image and $\boldsymbol{h} = (h_1, \ldots, h_T)$. Another is wavelet smoothing (Donoho and Johnstone 1994). For instance, soft-Waveshrink for a wavelet matrix $\Phi$, 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 that it has near minimax properties for a class of loss functions and smoothness classes (Donoho, Johnstone, Kerkyacharian and Picard 1995) for a simple selection of $\lambda$ 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 $\lambda$ that leads to a smooth estimation of the volatility process.

How does our smoothing approach differ from existing approaches? We aim at estimating a smooth evolution of volatility in time with possible abrupt changes of regime which seems more realistic for finance applications, while the latter aims at estimating the true coefficients of some assumed erratic volatility data generating process. To that aim we propose a selection of $\lambda$ (hence of $\sigma_\eta^2$) borrowing ideas from wavelet smoothing. Our estimator also differs in the way the estimator is computed since we solve a convex optimization problem (5) (see Section 2.3) instead of solving an integration problem (4) approximately by sampling an ergodic Markov chain.

## 2.2 Selection of the smoothing parameter $\lambda$

The $\ell_1$-regularization parameter $\lambda \geq 0$ in (5) controls the smoothing: When $\lambda = 0$, the solution is the wiggly maximum likelihood estimate $\hat{\sigma}_t^{\text{MLE}} = |y_t|$ for $t = 1, \ldots, T$, while when $\lambda$ tends to infinity the estimates $h_t$ tends to either a linear function for $\phi = 1$ or a function that is asymptotically constant for $\phi < 1$. We see that selection of $\lambda$ (or equivalently $\sigma_\eta^2$ in (3)) is crucial. 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 $\ell_1$-penalized likelihood wavelet estimator with near minimax properties for a class of loss functions and smoothness measures

(Donoho and Johnstone 1994, Donoho et al. 1995). Borrowing from wavelet smoothing, we derive a universal penalty $\lambda_T^{\text{fSVM}}$ for the maximum a posteriori estimator (5) such that the estimated volatility is a smooth fit to the data. With $\lambda$ selected by the universal rule, our stochastic volatility model (1)–(3) can be seen as a *functional* data generating process in the sense that it leads to an estimated volatility process which is a smooth rather than erratic *function of time* that fits the volatility of the financial asset. Hence our estimator is derived from a functional stochastic volatility model (fSVM).

In wavelet smoothing, the universal parameter is chosen such that, when the underlying signal is piecewise constant, 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 (i.e., persistence with $\phi = 1$ on each interval) on $K_T$ successive times, the volatility estimate is also piecewise constant with probability tending to one. In Appendix A we derive the universal parameter $\lambda_T^{\text{fSVM}} = \sqrt{K_T \log(n_T \log n_T)}$ with $n_T = T/K_T$ for standard Gaussian $\epsilon$-innovations, $\phi = 1$, $K_T \sim \log T$ and the link $\varphi(\cdot) = \exp(\cdot)$. Deriving the universal parameter for $\phi = 1$ is valid since we expect strong persistence in practice, and this avoids the derivation of a $\phi$-dependent universal parameter.

## 2.3 Optimization issues

We study here how to solve (5) to obtain the proposed estimate. To solve (5) in $(\phi, \mu, \boldsymbol{h})$, we use a decomposition approach that alternately solves in $(\mu, \boldsymbol{h})$ with $\phi$ held fixed, and in $\phi$ 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 (5) is convex piecewise-linear in $\phi$ and the minimum can be found by, e.g., sorting the breakpoints. In what follows, we focus on solving in $(\mu, \boldsymbol{h})$, with $\phi$ held fixed. The resulting subproblem can be written compactly as

$$\min_{\boldsymbol{h},\mu} \sum_{t=1}^{T} g_t(h_t) + \pi(B_\phi \boldsymbol{h} + \mu(\phi - 1)\mathbf{1}), \tag{6}$$

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 $\pi(\cdot) = \lambda \| \cdot \|_1$. We will focus on the link $\varphi(\cdot) = \exp(\cdot)$, and it can be seen that $g_t(h_t) = h_t + \frac{1}{2} y_t^2 \exp(-2h_t) +$ constant is strictly convex for a Gaussian $\epsilon$-innovations. Hence (6) is a convex optimization problem.

The case of $\phi = 1$ can be solved efficiently using the 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 $\epsilon$-innovations, so that a dual coordinate descent algorithm can be employed on the dual problem in $\mathbf{u} = (u_1, ..., u_T)$. We consider below the more challenging and interesting case of $\phi \neq 1$, for which the IDM algorithm is impractical. As we show below, in this case (6) can be efficiently solved by a primal-dual interior-point algorithm.

### 2.3.1 Dual formulation

Using $\varphi(\cdot) = \exp(\cdot)$, we derive in Appendix B the dual of the primal subproblem (6). It has the general form

$$\min \sum_{t=1}^{Q} q_t(x_t) \quad \text{s.t.} \quad A\mathbf{x} = b,\ \mathbf{x} \geq \mathbf{0}, \tag{7}$$

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

$$(\xi + \delta)\left(q_t'(\xi + \delta) - q_t'(\xi) - q_t''(\xi)\delta\right) \geq -\kappa q_t''(\xi)\delta^2 \quad \text{whenever} \quad \frac{|\delta|}{\xi} \leq \rho, \tag{8}$$

for some $\kappa > 0$ and $0 < \rho < 1$. In particular, Appendix C shows that (8) 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 $X = \operatorname{diag}(x_1, \ldots, x_Q)$. This can be rewritten as

$$A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} > \mathbf{0}, \quad \mathbf{s} = q'(\mathbf{x}) - A^\top \mathbf{u}, \quad X\mathbf{s} = \varepsilon \mathbf{1}. \tag{9}$$

The exact solution of (9) traces the central path as $\varepsilon$ ranges over $(0, \infty)$. The primal-dual interior-point algorithm solves the equations approximately using damped Newton method and decreases $\varepsilon$ after each iteration. Specifically, $(\mathbf{x}, \mathbf{u})$ is an approximate solution of (9) if it belongs, together with $\varepsilon$, 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

$$\begin{aligned} X d_s + S d_x &= \delta\varepsilon \mathbf{1} - X\mathbf{s}, && (10) \\ A d_x &= 0, && (11) \\ q''(\mathbf{x}) d_x - A^\top d_u &= d_s, && (12) \end{aligned}$$

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 $\nu$ and $\bar{\alpha}$ be given by (26) and (27) in Appendix D. Then, beginning with $\alpha = 1$, it checks if

$$(\mathbf{x}[\alpha], \mathbf{u}[\alpha]) \in \mathcal{N}(\tau), \qquad \varepsilon[\alpha] = \frac{\mathbf{x}[\alpha]^\top \mathbf{s}[\alpha]}{Q} \leq (1 - \bar{\alpha}\nu)\varepsilon, \tag{13}$$

and if not, it decreases $\alpha$ by some factor $0 < \varrho < 1$ and repeat, until (13) 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 code is available from the authors and is fast since its complexity is bounded by $O(Q^{3/2})$, as we now show.

### 2.3.4 Iteration complexity

Appendix D shows that (13) is satisfied when $\alpha = \bar{\alpha}$ given by (27) in Appendix D. Thus $\varepsilon$ 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 (27), $\frac{1}{\bar{\alpha}} = O(\kappa Q^{3/2} + Q)$, 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 (7) and (8) (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 (7) is a linear or a convex quadratic program, i.e., $\kappa = 0$ in (8), 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 from 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 classical functions in wavelet smoothing, the blocks and bumps functions (Donoho and Johnstone 1994), rescaled to have a range of volatility $\sigma_t \in [0.1, 10]$. The log of the volatility function $\sigma_t$ for $t = 1, \ldots, 5000$ is the curve plotted on Figure 1 (b), where the dots are the maximum likelihood estimates $\log \hat{\sigma}_t^{\text{MLE}} = \log |y_t|$. 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 the fSVM penalized likelihood estimated log-volatilities solution to (5) with $\varphi(\cdot) = \exp(\cdot)$, using the universal penalty $\lambda_T$ 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 residual process matches well the i.i.d. standard Gaussian distribution.

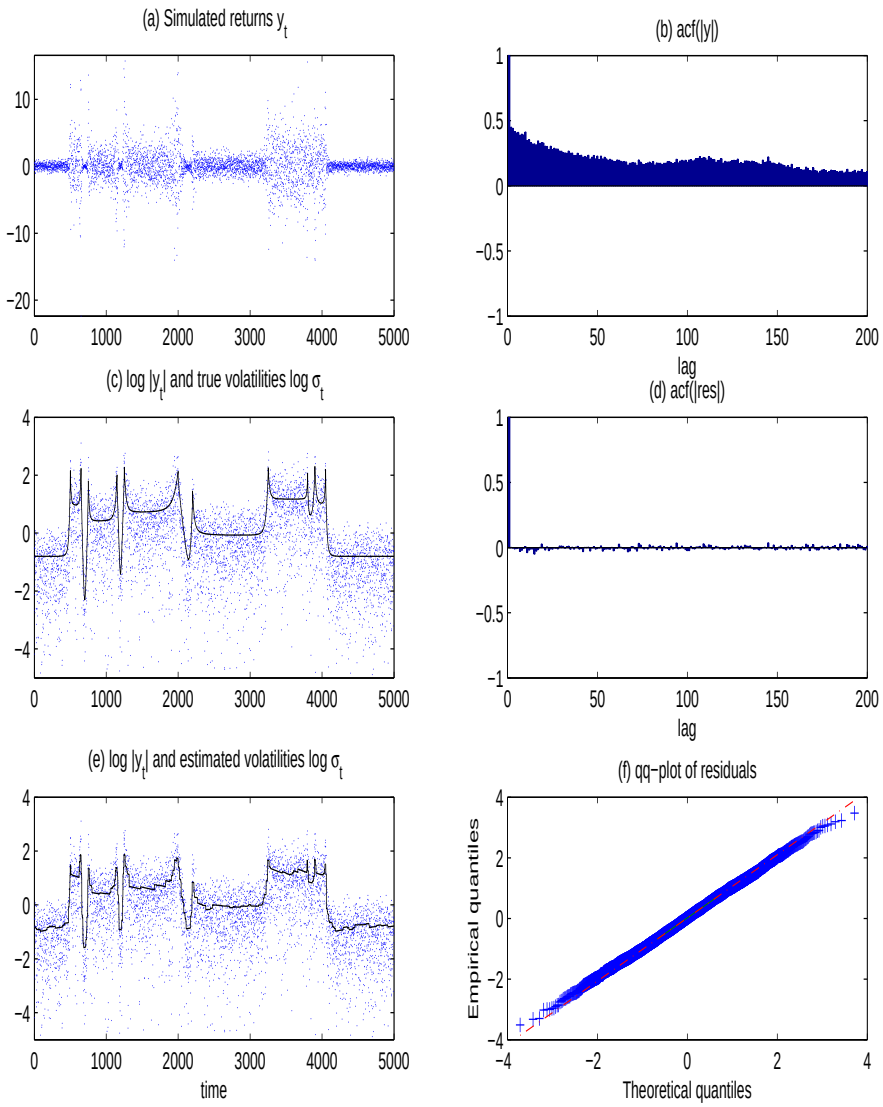


Figure 1: Volatility simulation with $T = 5000$. Top: simulated returns $y_t$ (dots), and empirical correlation function (acf) of $|y_t|$; Middle: true log-volatilities $\sigma_t$ (line) with log-absolute-returns $\log|y_t|$ (dots), and acf of residuals $|y_t/\hat{\sigma}_t|$; Bottom: log-volatilities (line) estimated with the universal penalty $\lambda_T^{\mathrm{fSVM}}$ with $\log|y_t|$ (dots), and Gaussian qq-plot of residuals.

# 3 Applications

We illustrate our methodology on the DOW JONES and NASDAQ stock index returns. Data are the $T = 4846$ daily closing prices $P_t$ between 29 December 1989 until 23 march 2009. Stock returns $y_t$ are computed as $100 \log(P_t/P_{t-1})$. Figures 2 and 3 display the log-returns for both indices on their top-left graphs. In Section 3.1 we first consider the entire sequence of returns, apply our volatility estimator to it and analyze the result. Note that the estimation for such time-series took less than 2 minutes running a Matlab code on a standard computer. In Section 3.2 we evaluate and compare the forecasting performance of our method to that of 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 and 3 (b). Results of our volatility estimation is presented in graphs (d) on log-scale, and graphs (e)-(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 (e)). And the normality holds quite well, except in the negative extreme tail (see graphs (f)). This asymmetry is certainly due to the fact that our methodology does not yet model the leverage effect on the volatility (Nelson 1991, Glosten, Jagannathan and Runkle 1993), i.e. the responses of the volatility to the negative shocks (bad news) are stronger than the ones to the positive shocks. To capture this asymmetry, model (1)-(3) can be modified as in Omori, Chib, Shephard and Nakajima (2007).

The persistence parameter estimates for both indices are $\hat{\phi} = 1.0003$ and $\hat{\phi} = 1.0001$, for the DOW JONES and the NASDAQ, respectively. These high values of $\phi$ lead to conclude that the volatility may be nonstationary. However in the out-of-sample forecast exercise presented below, the estimates $\hat{\phi}$ are less than 1.00 and quite stable over the period from November 2001 until just before the recent subprime crisis, where the estimate of $\phi$ rises dramatically. Therefore, the fact that the full sample estimate of $\phi$ reaches above unity is probably due to the instability related to the recent crashes (see Figure 4). To derive the variance-covariance matrix of the vector of parameters and in particular of $\phi$, we observe that for given $\lambda$ and $(h_t)_{t=1}^T$, the second term of the objective function (5) is a least absolute error regression problem of $(h_t)_{t=2}^T$ on $(h_t)_{t=1}^{T-1}$ with slope $\phi$ and intercept $\tilde{\mu} = \mu(\phi - 1)$. Bassett and Koenker (1978) provide asymptotic theory for least absolute fit. Hence the asymptotic covariance is given by $\mathbb{V}((\hat{\tilde{\mu}}, \hat{\phi})^\top) = \frac{1}{\lambda^2} Q^{-1}$ with $\lambda = \lambda_T^{\text{fSVM}}$ of Section 2.2, where $Q = X^\top X$

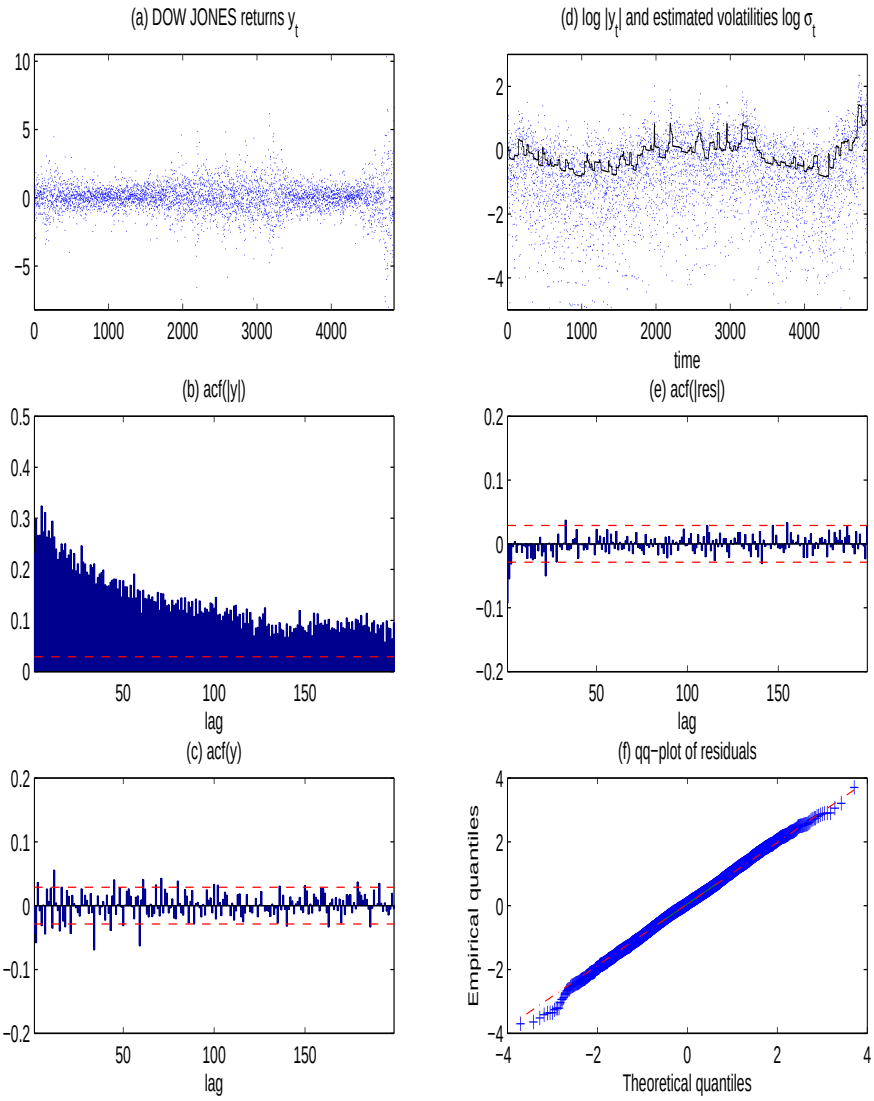


Figure 2: DOW JONES: (a) returns $y_t$ from 29 december 1989 until 23 march 2009; (b) empirical autocorrelation function (acf) of $|y_t|$; (c) acf of $y_t$; (d) log-absolute-returns $\log|y_t|$; (e) acf of standardized absolute residuals $|y_t/\hat{\sigma}_t|$; (f) Gaussian qq-plot of standardized residuals.

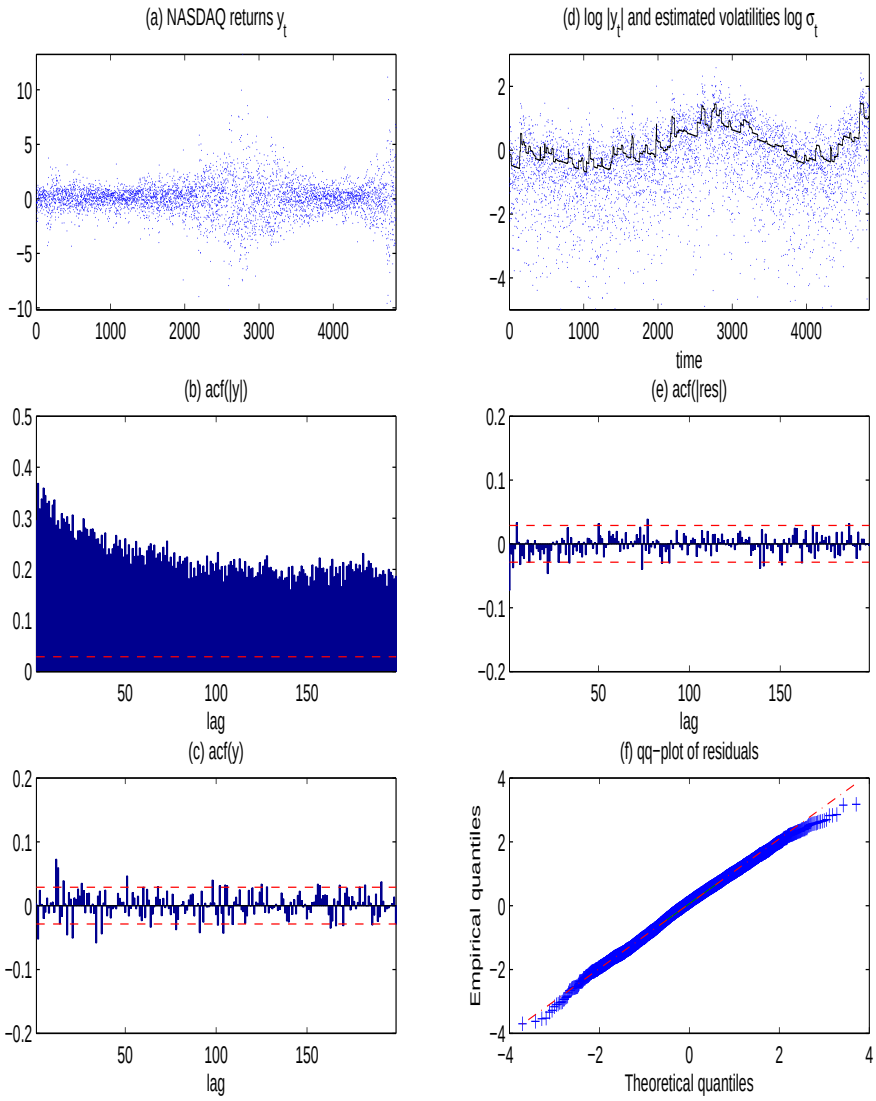


Figure 3: NASDAQ: (a) returns $y_t$ from 29 december 1989 until 23 march 2009; (b) empirical autocorrelation function (acf) of $|y_t|$; (c) acf of $y_t$; (d) log-absolute-returns $\log|y_t|$; (e) acf of standardized absolute residuals $|y_t/\hat{\sigma}_t|$; (f) Gaussian qq-plot of standardized residuals.

with $X = (\mathbf{1}, \mathbf{h}_{-1})$, and $\mathbf{h}_{-1} = (h_1, ..., h_{T-1})\prime$, 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 our estimator. So we compute the matrix $Q$ with the estimated log-volatilities to provide an approximate covariance matrix for the two estimated parameters of the model. For the time series considered the estimated standard deviations of $\hat{\phi}$ are 0.0055 for the DOW JONES and 0.0045 for the NASDAQ.

Figure 4 compares the evolution of both estimated log-volatilities for the DOW JONES and the NASDAQ. 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 the main abrupt changes estimated by fSVM.

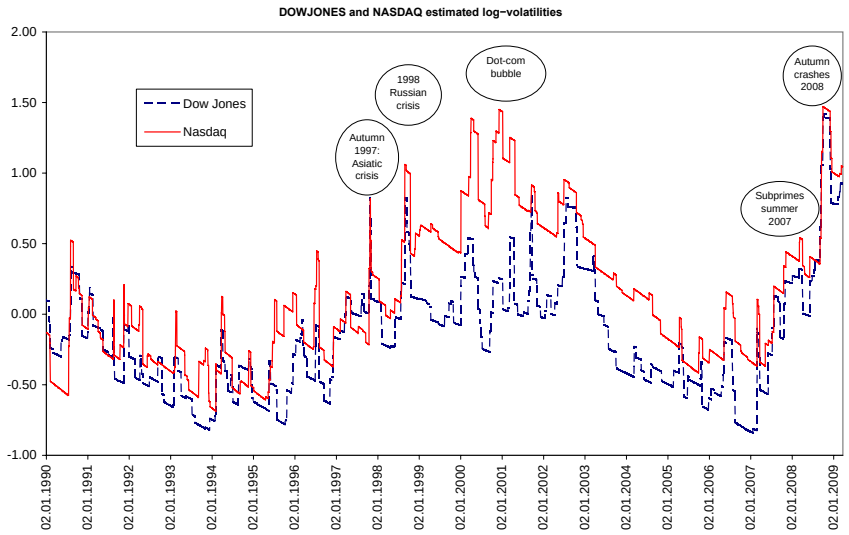


Figure 4: DOW JONES and NASDAQ's estimated log-returns from 29 december 1989 until 23 march 2009.

### 3.2 Volatility forecasting

To evaluate the forecasting performance of various models, 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 $t+H$, for a short horizon $H = 20$ business days (one month) and long horizon $H = 120$ days (six months). The calibration up to time $t$ provides parameter estimates for forecasting the volatility beyond time $t$ using the following formulas for the different models, with the notation $\sigma^2_{t+j} = \mathbb{E}(\varphi(h_{t+j})|y_1, ..., y_t)$:

for fSVM:

$$\sigma^{2,\mathrm{fSVM}}_{t+j} = \exp(2(\mu + \phi^j(h_t - \mu))), \quad j = 1, \ldots, H,$$

for GARCH:

$$\sigma^{2,\mathrm{GARCH}}_{t+j} = (\psi+\phi)^j\sigma^2_t + \omega\sum_{k=0}^{j-1}(\psi+\phi)^k, \quad j = 1, \ldots, H$$

and for IGARCH:

$$\sigma^{2,\mathrm{IGARCH}}_{t+j} = \sigma^2_t + j\omega, \quad j = 1, \ldots, H.$$

For each model, we then calculate the median absolute error (MAE) forecasting measure between the forecasted volatilities and the realized volatilities

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

where "*" stands for fSVM, GARCH or IGARCH. We report in Table 1 the relative MAE with respect to that of fSVM (i.e., $\mathrm{MAE}^{(\mathrm{I})\mathrm{GARCH}}/\mathrm{MAE}^{\mathrm{fSVM}}$), so that a ratio larger than 1 means better forecasting for fSVM. The results show that fSVM outperforms GARCH for forecasting, especially for a long horizon. Consequently our methodology appears better for the purpose of option pricing or portfolio management.

# 4 Conclusion

This paper proposes an original and complete new way to estimate returns volatility. Our approach combines the dynamic proposed by the well known SV models and the Markov random field smoother in order to estimate hidden Markov chains. The proposed estimator is based on Bayesian posterior

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

| | DOW JONES | | NASDAQ | |
|---|---|---|---|---|
| Horizon | $H = 20$ | $H = 120$ | $H = 20$ | $H = 120$ |
| MAE GARCH/fSVM | 1.1 | 1.9 | 0.8 | 1.8 |
| MAE IGARCH/fSVM | 1.2 | 3.0 | 0.8 | 2.7 |

mode estimation. Its performance is illustrated through smulations and empirical applications based on the Dow Jones and Nasdaq. Our volatility forecast outperforms GARCH's forcast, especially on long horizon, and therefore our methodology looks more suitable for option pricing models or portfolio management which require long term forcast. As a major topic for future research, the multivariate extension is probably the one which is most promising, especially given our selection of the smoothing parameter, the speed of our algorithm and the good forecasting performance. Another natural extension would consist in capturing leverage effect or asymmetry in the marginal distribution of the returns.

# APPENDIX

## A Universal penalty

Assuming for simplicity $T$ is a multiple of $K$, let $B_K$ be the matrix operating finite differences skipping every other $K$, i.e., $B_K \boldsymbol{h} = \mathbf{0}$ iff $\boldsymbol{h}$ is piecewise constant taking identical values at $K$ successive points:

$$h_{K(i-1)+1} = \ldots = h_{K(i-1)+K}, \ i = 1, \ldots, n_T = T/K. \tag{14}$$

The Karush-Kuhn-Tucker first-order optimality conditions for (5) with $\phi = 1$ and skipping every other $K$ differences in the penalty are

$$1 - y_t^2 \exp(-2h_t) + (B_K^\top \mathbf{w})_t = 0 \quad t = 1, \ldots, T \tag{15}$$

$$\|\mathbf{w}\|_\infty \leq \lambda, \tag{16}$$

where $\|\mathbf{w}\|_\infty = \max(|w_1|, ..., |w_T|)$. The solution $\mathbf{w}$ to (15)-(16) subject to $K$ successive identical values has entries

$$w_{K(i-1)+k} = k - K \frac{\sum_{j=1}^{k} y_{K(i-1)+j}^2}{\sum_{j=1}^{K} y_{K(i-1)+j}^2}, \quad i = 1, \ldots, n_T, \ k = 1, \ldots, K \tag{17}$$

provided $\lambda$ is large enough. The smallest possible $\lambda$ allowing a solution is $\lambda_{\mathbf{y}} = \|\mathbf{w}\|_\infty$ for $\mathbf{w}$ given in (17). The goal of the universal penalty $\lambda_T$ is to control the extremal behavior of $\lambda_{\mathbf{y}}$ so that $\mathbb{P}(\|\mathbf{w}\|_\infty \leq \lambda_T) \overset{T\to\infty}{\to} 1$ when the true underlying volatility is constant $\sigma_0$, i.e., $y_t$ i.i.d. $N(0, \sigma_0^2)$. The vector $\mathbf{w}$ can be broken into $n_T = T/K$ independent blocks $\mathbf{w}_i = (w_{i1}, \ldots, w_{i(K-1)})$ each of which converging to a Brownian bridge process. To see this, consider the first block for which

$$\mathbb{P}\left(\|\mathbf{w}_1\|_\infty \leq \lambda\right) \;=\; \mathbb{P}\left(\max_{k=1,\ldots,K-1} |k - K\frac{\sum_{j=1}^k Z_j^2}{\sum_{j=1}^K Z_j^2}| \leq \lambda\right)$$

where $Z_j = y_j/\sigma_0$ is i.i.d. $N(0,1)$. Note that $Z_j^2 + Z_{j+1}^2 =_d E_l$, with $j = 2(l-1)+1$ and $l = 1, \ldots, K/2$, where $E_l$ is i.i.d. $Exp(1/2)$, so

$$\frac{\sum_{j=1}^k Z_j^2}{\sum_{j=1}^K Z_j^2} \overset{d}{=} \frac{\sum_{l=1}^{k/2} E_l}{\sum_{l=1}^{K/2} E_l} \overset{d}{=} U_l,$$

where $0 \leq U_1 \leq \ldots \leq U_{K/2} \leq 1$ are distributed as Uniform$(0,1)$ order statistics from a sample of size $K/2$ (Shorack and Wellner 1986, p.496). Moreover the uniform quantile process converges to a Brownian bridge process $W(r)$, so

$$\begin{aligned}
\mathbb{P}(\|\mathbf{w}_1\|_\infty \leq \lambda) &\doteq \mathbb{P}(\sup_{r\in[0,1]} |W(r)| \leq \lambda/\sqrt{2K}) \\
&= 1 - 2\sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2(\lambda/\sqrt{2K})^2) \\
&\geq 1 - 2\exp(-2(\lambda/\sqrt{2K})^2).
\end{aligned}$$

Consequently

$$\mathbb{P}(\|\mathbf{w}\|_\infty \leq \lambda_T) \geq \left(1 - 2\exp(-2(\lambda/\sqrt{2K})^2)\right)^{n_T} \doteq \exp(-2/\log n_T)$$

with $\lambda_T^{fSVM} = \sqrt{2K}\sqrt{\frac{1}{2}\log((n_T \log n_T))}$, which tends to one with a slow rate as $T$ grows. We choose $K = K_T \sim \log T$ for the blocks size to slowly grow with $T$.

# B Derivation of the dual problem

Letting $\theta = (\boldsymbol{h}, \mu)$, the primal problem (6) has the form of

$$\min_\theta g(\theta) + \pi(B\theta),$$

where $g(\theta) = \sum_{t=1}^{T} g_t(h_t) + g_0(\mu)$, $g_0(\mu) = 0$, $B = [B_\phi \ (\phi - 1)\mathbf{1}]$. Thus the Fenchel dual (Rockafellar 1970) has the form

$$\min_w g^*(B^\top \mathbf{w}) + \pi^*(-\mathbf{w}),$$

with $g^*(\eta) = \sum_{t=1}^{T} g_t^*(\eta_t) + g_0^*(\eta_0)$. Here $^*$ denotes the convex conjugate, i.e., $g^*(\eta) = \sup_\theta \theta^T \eta - g(\theta)$. Straightforward calculation yields $g_t^*(u_t) = \frac{1}{2}(1-u_t)(\log(\frac{1-u_t}{y_t^2}) - 1)$ when $g_t(h_t) = h_t + \frac{1}{2} y_t^2 \exp(-2h_t)$ (Gaussian noise) and $g_t^*(u_t) = (1-u_t)(\log(\frac{1-u_t}{y_t}) - 1)$ when $g_t(h_t) = h_t + y_t \exp(-h_t)$ (Laplace noise), for $t = 1, \ldots, T$. Moreover, $g_0^*(\eta_0) = 0$ if $\eta_0 = 0$ and otherwise equals $\infty$. Similarly, $\pi^*(-\mathbf{w}) = 0$ if $\|\mathbf{w}\|_\infty \le \lambda$ and otherwise equals $\infty$. This yields the dual problem

$$\min \sum_{t=1}^{T} g_t^*(u_t) \quad \text{s.t.} \quad \mathbf{u} = B_\phi^\top \mathbf{w}, \ \mathbf{1}^\top \mathbf{w} = 0, \ -\lambda \mathbf{1} \le \mathbf{w} \le \lambda \mathbf{1}.$$

Then letting $\mathbf{z} = \mathbf{1} - \mathbf{u}$ and $\mathbf{w}_1 = \lambda \mathbf{1} - \mathbf{w}$, $\mathbf{w}_2 = \lambda \mathbf{1} + \mathbf{w}$, and $c_t = -(1 + \log(y_t^2))$ or $c_t = -(1 + \log(y_t))$, and upon eliminating $\mathbf{w}$, the above dual problem is equivalent to

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{w}_2, \mathbf{z}} \quad & \sum_{t=1}^{T} z_t \log(z_t) + c_t z_t \\ \text{s.t.} \quad & \begin{cases} B_\phi^\top \mathbf{w}_2 + \mathbf{z} = \mathbf{1} + \lambda B_\phi^\top \mathbf{1}, \ \mathbf{1}^\top \mathbf{w}_2 = \lambda(T-1), \ \mathbf{w}_1 + \mathbf{w}_2 = 2\lambda \mathbf{1}, \\ \mathbf{z} \ge \mathbf{0}, \ \mathbf{w}_1 \ge \mathbf{0}, \mathbf{w}_2 \ge \mathbf{0}. \end{cases} \end{aligned}$$

where $\mathbf{x} = (\mathbf{z}, \mathbf{w}_1, \mathbf{w}_2)$ are the dual variables. This has the form (7) with

$$A = \begin{bmatrix} I & 0 & B_\phi^\top \\ 0 & 0 & \mathbf{1}^\top \\ 0 & I & I \end{bmatrix}, \ b = \begin{bmatrix} \mathbf{1} + \lambda B_\phi^\top \mathbf{1} \\ \lambda(T-1) \\ 2\lambda \mathbf{1} \end{bmatrix}, \ q_t(x_t) = \begin{cases} z_t \log(z_t) + c_t z_t & \text{if } t \le T; \\ 0 & \text{else.} \end{cases}$$

# C Checking the condition (8)

It is easily seen that the condition (8) with $\kappa = \frac{1}{2(1-\rho)}$ is satisfied by $q_t(x_t) = 0$ for any $0 < \rho < 1$. Below we show that it is also satisfied by $q_t(x_t) = x_t \log(x_t) + c_t x_t$. We have $q_t'(x_t) = \log(x_t) + 1 + c_t$ and $q_t''(x_t) = 1/x_t$. Since $|\delta|/\xi \le \rho < 1$, we have the series expansion

$$\log(\xi + \delta) - \log \xi - \frac{\delta}{\xi} = \sum_{k=2}^{\infty} (-1)^{k+1} \frac{\delta^k}{k \xi^k}.$$

When $\delta < 0$, we have $0 < \xi + \delta \leq \xi$. Moreover, all terms in the series are negative and the series can be bounded from below by $\sum_{k=2}^{\infty} -\frac{\delta^2}{\xi^2}\frac{\rho^{k-2}}{2} = -\frac{1}{2(1-\rho)}\frac{\delta^2}{\xi^2}$. Thus

$$(\xi+\delta)\left(\log(\xi+\delta) - \log\xi - \frac{\delta}{\xi}\right) \geq -\frac{\xi+\delta}{2(1-\rho)}\frac{\delta^2}{\xi^2} \geq -\frac{\delta^2}{2(1-\rho)\xi},$$

implying (8) with $\kappa = \frac{1}{2(1-\rho)}$. When $\delta \geq 0$, we have $0 < \xi + \delta \leq \xi(1+\rho) \leq \frac{\xi}{1-\rho}$. Moreover, the series alternates in sign and can be written as $-\frac{\delta^2}{2\xi^2} + \sum_{k=3,5,\ldots}\left(\frac{1}{k} - \frac{\delta}{(k+1)\xi}\right)\frac{\delta^k}{\xi^k} \geq -\frac{\delta^2}{2\xi^2}$. Then we again obtain (8) with $\kappa = \frac{1}{2(1-\rho)}$.

# D Iteration complexity analysis

Fix any $(\mathbf{x}, \mathbf{u}, \varepsilon) \in \mathcal{N}(\tau)$. Let $(d_x, d_s, d_u)$ be the solution of (10)–(12), where $\mathbf{s} = q'(\mathbf{x}) - A^\top \mathbf{u}$, and $0 < \delta < 1$. We show that (13) is satisfied when $\alpha = \bar{\alpha}$, where $\bar{\alpha}$ and $\nu$ are given by (27) and (26).

As in the proof of Wright (1997, Lemma 5.10), we first bound $\|D_x d_s\|$. Letting $D = X^{-1/2}S^{1/2}$, we rewrite (10) as

$$D^{-1}d_x + Dd_s = (XS)^{-1/2}r,$$

where for simplicity we denote $r = \delta\varepsilon\mathbf{1} - X\mathbf{s}$. Left multiplying (12) by $d_x^\top$ and using $q_t''(x_t) \geq 0$ (since $q_t(\cdot)$ is convex) and (11) yields $0 \leq d_x^\top q''(x) d_x = d_x^\top d_s$. Then, by Wright (1997, Lemma 5.3), we have

$$\begin{aligned}
\|D_x d_s\| &= \|(D^{-1}D_x)(Dd_s)\| \\
&\leq 2^{-3/2}\|D^{-1}d_x + Dd_s\|^2 \\
&= 2^{-3/2}\|(XS)^{-1/2}r\|^2 \\
&= 2^{-3/2}\|(XS)^{-1/2}\delta\varepsilon\mathbf{1} - (XS)^{1/2}\mathbf{1}\|^2 \\
&= 2^{-3/2}\left(\sum_t \frac{\delta^2\varepsilon^2}{x_t s_t} - 2\sigma\varepsilon Q + \mathbf{x}^\top\mathbf{s}\right) \\
&\leq 2^{-3/2}\left(Q\frac{\delta^2\varepsilon^2}{\tau\varepsilon} - 2\delta\varepsilon Q + \varepsilon Q\right) \\
&= \vartheta Q\varepsilon, \qquad (18)
\end{aligned}$$

where we let $\vartheta = 2^{-3/2}\left(\frac{\delta^2}{\tau} - 2\delta + 1\right)$.

Let $\bar{d}_x = X^{-1}d_x$. We next bound $\|\bar{d}_x\|$. We have from left multiplying (10) by $\bar{d}_x^\top$ and using $d_x^\top d_s = d_x^\top q''(\mathbf{x}) d_x$ that

$$d_x^\top q''(\mathbf{x}) d_x + \bar{d}_x^\top SX\bar{d}_x = \bar{d}_x^\top r,$$

where for simplicity we denote $\mathbf{r} = \delta\varepsilon\mathbf{1} - X\mathbf{s}$. Since $(\mathbf{x}, \mathbf{u}, \varepsilon) \in \mathcal{N}(\tau)$ so that $x_t s_t \geq \tau\varepsilon$ for all $t$, this and the Cauchy-Schwarz inequality yields $\tau\varepsilon\|\bar{d}_x\|^2 \leq \|\bar{d}_x\|\|r\|$ so that

$$\|\bar{d}_x\| \leq \frac{\|\mathbf{r}\|}{\tau\varepsilon}. \tag{19}$$

Since $\mathbf{z} = \frac{X\mathbf{s}}{\varepsilon}$ satisfies $\mathbf{z} \geq \mathbf{0}$ and $\mathbf{1}^\top\mathbf{z} = Q$, we have

$$\|r\| \leq \max_{\mathbf{z}\geq\mathbf{0}}\{\varepsilon\|\delta\mathbf{1} - \mathbf{z}\| \mid \mathbf{1}^\top\mathbf{z} = Q\} = \varepsilon\sqrt{(Q-\delta)^2 + (Q-1)\delta^2} \leq \varepsilon Q, \tag{20}$$

where the equality is due to the maximum of a convex function being attained at an extreme point, which in this case is a $Q$ multiple of the unit coordinate vector.

Fix any $\alpha > 0$ with $\alpha\|\bar{d}_x\|_\infty \leq \rho$. By (19) and (20), this occurs whenever

$$\alpha \leq \frac{\rho\tau}{Q}. \tag{21}$$

We have $A\mathbf{x}[\alpha] = A\mathbf{x} + \alpha A d_x = b$ and

$$\begin{aligned}
& X[\alpha]\mathbf{s}[\alpha] \\
= \;& X[\alpha](q'(\mathbf{x}+\alpha d_x) - A^\top(\mathbf{u}+\alpha d_u)) \\
= \;& X[\alpha](q'(\mathbf{x}+\alpha d_x) - q'(\mathbf{x}) - \alpha q''(\mathbf{x})d_x) + X[\alpha](\mathbf{s} + \alpha q''(\mathbf{x})d_x - \alpha A^\top d_u) \\
= \;& X[\alpha](q'(\mathbf{x}+\alpha d_x) - q'(\mathbf{x}) - \alpha q''(\mathbf{x})d_x) + X\mathbf{s} + \alpha S d_x + \alpha X d_s + \alpha^2 D_x d_s \\
= \;& X[\alpha](q'(\mathbf{x}+\alpha d_x) - q'(\mathbf{x}) - \alpha q''(\mathbf{x})d_x) + (1-\alpha)X\mathbf{s} + \alpha\delta\varepsilon\mathbf{1} + \alpha^2 D_x d_s \\
\geq \;& -\kappa\alpha^2 q''(\mathbf{x})D_x d_x + (1-\alpha)\tau\varepsilon\mathbf{1} + \alpha\delta\varepsilon\mathbf{1} + \alpha^2 D_x d_s \\
\geq \;& -\kappa\alpha^2\|q''(\mathbf{x})D_x d_x\|_1\mathbf{1} + ((1-\alpha)\tau + \alpha\delta)\varepsilon\mathbf{1} - \alpha^2\|D_x d_s\|\mathbf{1} \\
= \;& -\kappa\alpha^2 d_x^\top d_s\mathbf{1} + ((1-\alpha)\tau + \alpha\delta)\varepsilon\mathbf{1} - \alpha^2\|D_x d_s\|\mathbf{1} \\
\geq \;& -\kappa\alpha^2\vartheta Q^{3/2}\varepsilon\mathbf{1} + ((1-\alpha)\tau + \alpha\delta)\varepsilon\mathbf{1} - \alpha^2\vartheta Q\varepsilon\mathbf{1},
\end{aligned} \tag{22}$$

where the third equality uses (12); the fourth equality uses (10); the first inequality uses $\mathbf{x}[\alpha] > 0$, (8), and $x_t s_t \geq \tau\varepsilon$ for all $t$ (since $(\mathbf{x}, \mathbf{u}, \varepsilon) \in \mathcal{N}(\tau)$); the last inequality uses $d_x^\top d_s \leq \|D_x d_s\|_1 \leq \sqrt{Q}\|D_x d_s\|$ and (18). Similarly, we have

$$\begin{aligned}
\mathbf{x}[\alpha]^\top\mathbf{s}[\alpha] &= \mathbf{x}[\alpha]^\top(q'(\mathbf{x}+\alpha d_x) - q'(\mathbf{x}) - \alpha q''(\mathbf{x})d_x) \\
&\quad + (1-\alpha)\mathbf{x}^\top\mathbf{s} + \alpha\delta\varepsilon Q + \alpha^2 d_x^\top d_s \\
&\leq (1-\alpha)\varepsilon Q + \alpha\delta\varepsilon Q + \alpha^2\vartheta Q^{3/2}\varepsilon,
\end{aligned} \tag{23}$$

where the inequality uses the concavity of $q_t'(\cdot)$ as well as $d_x^\top d_s \leq \sqrt{Q}\|D_x d_s\|$ and (18). Comparing the two bounds (22) and (23), we see that $X[\alpha]\mathbf{s}[\alpha] \geq \frac{\tau}{Q}\mathbf{x}[\alpha]^\top\mathbf{s}[\alpha]\mathbf{1}$ whenever

$$-\kappa\alpha^2\vartheta Q^{3/2} + ((1-\alpha)\tau + \alpha\delta) - \alpha^2\vartheta Q \geq \tau(1 - \alpha + \alpha\delta + \alpha^2\vartheta Q^{1/2})$$

or, equivalently,

$$\alpha \leq \frac{\delta(1-\tau)}{\vartheta(\kappa Q^{3/2} + Q + \tau Q^{1/2})}. \tag{24}$$

Moreover, by (23), we have

$$\varepsilon[\alpha] = \frac{\mathbf{x}[\alpha]^\top \mathbf{s}[\alpha]}{Q} \leq (1 - \alpha(1 - \delta - \alpha\vartheta Q^{1/2}))\varepsilon \leq (1 - \alpha\nu)\varepsilon$$

whenever

$$\alpha \leq \frac{\rho(1-\delta)}{\vartheta Q^{1/2}}, \tag{25}$$

where

$$\nu = (1-\delta)(1-\rho). \tag{26}$$

The minimum of the three bounds (21), (24), and (25) is

$$\bar{\alpha} = \min\left\{\frac{\rho\tau}{Q}, \frac{\delta(1-\tau)}{\vartheta(\kappa Q^{3/2} + Q + \tau Q^{1/2})}, \frac{\rho(1-\delta)}{\vartheta Q^{1/2}}\right\}. \tag{27}$$

# References


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

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

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.

Dahlhaus, R. (1997), "Fitting time series models to nonstationary processes," *Annals of Statistics,* 25, 1–37.

Diebold, F.X. (1986), "Modeling the persistence of conditional variances: A comment," *Econometric Reviews,* 5, 51–56.

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.

Engle, R.F., and Bollerslev, T. (1986), "Modelling the persistence of conditional variance," *Econometric Reviews,* 5, 1–50.

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

Harvey, A.C., 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.

Jacquier, E., Polson, N. G., and Rossi, P. E. (1994), "Bayesian analysis of stochastic volatility models (Disc: P389-417)," *Journal of Business and Economic Statistics,* 12, 371–389.

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.

Lamoureux, C. G., and Lastrapes, W. D. (1990), "Persistence in variance, structural change, and the GARCH model," *Journal of Business and Economic Statistics,* 8, 225–234.

Lastrapes, W.D. (1989), "Exchange rate volatility and u.s. monetary policy: An arch application," *Journal of Money, Credit, and Banking,* 21, 66–77.

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.B. (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.

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.

Rockafellar, R. (1970) *Convex Analysis,* Princeton: Princeton University Press.

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

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.

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

Shorack, G.R., and Wellner, J. (1986) *Empirical Processes with Applications to Statistics,* Wiley.

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.J. (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.