Dynamical Products of Experts for Modeling Financial Time Series

Yutian Chen YUTIANC@ICS.UCI.EDU
Department of Computer Science, University of California, Irvine, CA 92697, USA

Max Welling WELLING@ICS.UCI.EDU
Department of Computer Science, University of California, Irvine, CA 92697, USA

Abstract

Predicting the "Value at Risk" of a portfolio of stocks is of great significance in quantitative finance. We introduce a new class models, "dynamical products of experts" that treats the latent process over volatilities as an inverse Gamma process. We show that our multivariate volatility models significantly outperform all related Garch and stochastic volatility models which are in popular use in the quantitative finance community.

1. Introduction

Many natural signals such as speech and images exhibit the characteristic heavy tailed distributions over their inputs. For instance, after filtering out the first and second order statistics, the distribution of brightness values of randomly sampled pixels in an image is well described by a Student-t density. The current best explanation for this phenomenon is that the variance (of a normal distribution) is itself subject to random fluctuations. The two-stage process of first sampling the variance from e.g. a gamma distribution and then conditional on that sampling a normal variate results in the heavy tails of a Student-t distribution.

For temporal or spatial processes such as speech and images there are also important interactions between the (centered and sphered) input dimensions. One can distinguish between two dominant effects (Lyu & Simoncelli, 2009): I) The input can often be described as a linear combination of independent basis functions. Finding this basis is known as "independent components analysis" (Bell & Sejnowski, 1995) and has led to interesting applications such as unmixing sound recordings and fMRI images. II) The variances of a group of inputs are correlated. This effect can for instance be observed in filtered images where at the location of edges one can observe a high magnitude of the coefficients but with unpredictable sign. Hence, the phenomenon is best understood as a clustering of the variance of the signal. Many models have been successful in modeling this phenomenon such as Gaussian scale mixtures (GSM) (Wainwright & Simoncelli, 2000) and energy based models such as PoT (Welling et al., 2002; Gehler & Welling, 2006) and FoE (Roth & Black, 2005).

A very similar phenomenon has been observed in the financial domain. Here the returns of stock prices also show a clear clustering or persistence of volatility. This phenomenon is nicely captured by Garch models (Bollerslev, 1986) where the variance at time t is a deterministic function of both the variance and the squared returns at previous time steps. Note that this induces smoothness in the variance but due to the deterministic nature of the regression there are no independent fluctuations in the volatility resulting in too small tails. Another approach to model the persistence of volatility is the stochastic volatility (SV) models (Taylor, 1982) where the variance conditioned on previous time steps is a stochastic variable. This class of models is usually considered to be better fitted to financial data.

The purpose of this paper is to show that we can improve on all these models by extending the PoT model to a temporal variant. PoT has the advantage that it models the variance as a stochastic process like a SV while its variance propagation combines the properties of both Garch and SV models. Also, the conditional distribution of variance is different from the usual log-normal random walk in SV. Moreover PoT also naturally models the independent components in the data covering both types of interactions described above. PoT models have been extended to hierarchical (topographic) models in (Osindero et al., 2006) which makes for a promising direction of future research.

In the following we will describe our temporal extension of PoT (DPoT) (see Figure 1) and show empirically that

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010. Copyright 2010 by the author(s)/owner(s).

it significantly outperforms all relevant multivariate Garch models and SV models on metrics that are of interest to the quantitative finance community (such as Value at Risk).

2. The Product of Student-t model (PoT)

The Product of Student-t model (PoT) was introduced in (Welling et al., 2002) to model the statistics of natural images. Its density function is given as,

$$p(\mathbf{x}) = \frac{1}{Z}\prod_{j=1}^{m}\frac{1}{\left(1+\frac{1}{2}(W_j^T\mathbf{x})^2\right)^{\alpha_j}} \tag{1}$$

where Z is the normalization term, a.k.a. partition function, and the jth row of matrix W, w_j^T, is called a filter. It can be understood as a kind of energy based model by introducing auxiliary variables $\{h_j\}$,

$$p(\mathbf{x},\mathbf{h}) = \frac{1}{Z}e^{-E} \tag{2}$$

$$E = \sum_{j=1}^{m}\left(h_j(1+\frac{1}{2}y_j^2) - (\alpha_j - 1)\log h_j\right) \tag{3}$$

with transformed inputs $y_j \triangleq \sum_{i=1}^{n} W_{ji}x_i$. The conditional distributions are given by $\mathbf{h} \sim \prod_j \Gamma(\alpha_j, 1 + \frac{1}{2}y_j^2)$ and $\mathbf{x} \sim \mathcal{N}(0, WH^{-1}W^T)$ with the second argument the covariance of the normal distribution and $H =$ diag$[h_1, .., h_m]$. From the conditional $p(\mathbf{x}|\mathbf{h})$ one can see that $\mathbf{h}$ acts as a precision variable.

In the special case of a complete model, $n = m$, we can also view the PoT model as a causal (directed) model closely related to the probabilistic formulation of independent components analysis (ICA) (Pearlmutter & Parra, 1996),

$$p(\mathbf{x},\mathbf{h}) = p(\mathbf{x}|\mathbf{h})p(\mathbf{h}) = \left|\frac{d\mathbf{y}}{d\mathbf{x}}\right| p(\mathbf{y}|\mathbf{h})p(\mathbf{h})$$

$$= |\det(W)| \prod_j \left[\mathcal{N}(y_j; 0, h_j^{-1})\Gamma(h_j; \alpha_j - \frac{1}{2}, 1)\right] \tag{4}$$

where y_j is an independent component and h_j its precision. $\mathbf{x}$ is given as a linear combination $\mathbf{x} = A\mathbf{y}$, and $A = W^{-1}$ is called the mixing matrix.

3. Dynamical PoT

To turn the PoT into a dynamical Markov process, we need to define the transition from states at time $t-1$ to time t. Because we are interested in the application of this model to financial time series we will be inspired by the interaction structure of a Garch model (see Figure 1). In Garch, the variance at time t is deterministically regressed on the variances σ_{t-1}^2 and the squared returns y_{t-1}^2 at the previous time step, $\sigma_t^2 = \tilde{c} + \tilde{a}y_{t-1}^2 + \tilde{b}\sigma_{t-1}^2$. In particular large values for squared returns and variances at time $t-1$ will result in large values for the variances at time t capturing the desired persistency of volatility. To capture the same intuition in a model with stochastic volatilities we write,

$$h_t \sim \Gamma^{-1}(\alpha - \frac{1}{2}, c + ay_{t-1}^2 + bh_{t-1}); \quad y_t \sim N(0, h_t) \tag{5}$$

where we have replaced $h \rightarrow h^{-1}$ to let h represent variance instead of precision which is therefore described by an *inverse* Gamma distribution. Note the somewhat counter intuitive implication that we had to introduce interaction terms of the form $h_t^{-1}bh_{t-1}$ which is notably different from the interaction type $h_t b h_{t-1}$ used in (Sutskever et al., 2009).

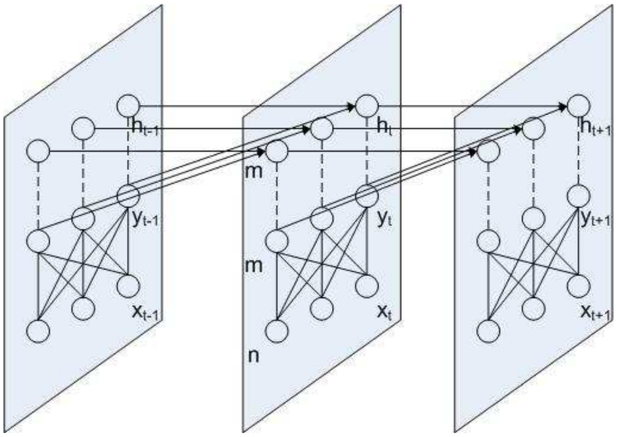

Figure 1. Causal Bayes net structure for the dynamical PoT model

One can show that by taking the following limit: $\alpha, a, b, c \rightarrow \infty$ such that $a/(\alpha - 3/2) = \tilde{a}, b/(\alpha - 3/2) = \tilde{b}, c/(\alpha - 3/2) = \tilde{c}$, the DPoT and the Garch models become equivalent. However, for finite α, a, b, c, DPoT has fatter tails than Garch. This can be seen by integrating out h_t and comparing $p(x_t|h_{t-1}, x_{t-1})$, which is given by a student-t distribution for DPoT:

$$p(x_t|h_{t-1}, x_{t-1}) = $$
$$t_1\left(\frac{x_t}{\sqrt{(c + ax_{t-1}^2 + bh_{t-1})/(\alpha - 1.5)}}, 2\alpha - 1\right) \tag{6}$$

while it is a normal distribution for Garch.

DPoT essentially belongs to the class of stochastic volatility models, see (Shephard et al., 2008) for a review. The difference between DPoT and the regular SV models is that the latter usually model the logarithm of the variance as a normal random walk to ensure the positiveness of variance while our model uses an inverse Gamma distribution to naturally satisfy the restriction. That also gives a slower decay rate in the tail of the conditional distribution of the volatility than the log-normal distribution. Furthermore, in DPoT the mean of the variance at the next time step is affected by the current return through both direct inference (parameter a) like Garch and the posterior distribution of the current variance like SV. This provides a hybrid mechanism of Garch and SV models, and should be more flexible than either model.

The above univariate model is easily extended to the multivariate case by introducing m independent Markov processes on the variables $\{h_j, y_j\}$ (Eqn.5) which we now linearly mix to produce the inputs $\mathbf{x}_t = A\mathbf{y}_t$. Written as an energy-based model this becomes, $p(x,h) = \prod_t \frac{1}{Z_t} e^{-E_t}$ with,

$$E_t = \sum_{j=1}^{m} \left(h_{jt}^{-1} \left[c_j + a_j y_{jt-1}^2 + b_j h_{jt-1} + \frac{1}{2} y_{jt}^2 \right] - (\alpha_j + 1) \log h_{jt}^{-1} \right) \tag{7}$$

$$Z_t = \frac{1}{|\det(W)|} \prod_j Z_{jt} \tag{8}$$

$$Z_{jt} = \frac{\sqrt{2\pi}\,\Gamma(\alpha_j - \frac{1}{2})}{(c_j + a_j y_{jt-1}^2 + b_j h_{jt-1})^{\alpha_j - 1/2}} \tag{9}$$

and $y_{jt} = W_j^T \mathbf{x}_t$.

There is redundancy in the parametrization since multiplying both W_j and c_j by a scalar gives the same probability $p(x)$. We will remove the extra degree of freedom by adding the normalization constraint $||y_j||_2 = ||W_j^T \mathbf{x}||_2 = 1$.

4. Extensions and related work

Many variants of the Garch model have been studied in the literature. For instance, longer range interactions are covered by higher order Garch(p,q) where the variance at time t is now a function of variance and squared returns at longer time lags, σ_{t-i}, y_{t-i}^2. This extension is easily incorporated into DPoT by introducing terms $h_t^{-1} a_i y_{t-i}^2$ and $h_t^{-1} b_i h_{t-i}$ for $i >= 2$. Also, we can add asymmetric terms to model the leverage effect by replacing $a y_{t-1}^2$ with $a^+ y_{t-1}^+ + a^- y_{t-1}^-$ (Threshold-Garch), or we can raise both h and y variables to a power θ (Power-Garch), or we can use different conditional distributions for $p(y|h)$ such as generalized exponential or student-t. Generally the DPoT variant has the form,

$$\begin{aligned} & h_{jt}^{\theta} | \{y_{j,t-p}\}, \{h_{j,t-q}\} \\ & \sim \Gamma^{-1}\left(\alpha_j, (\alpha_j - 1) f(\{y_{j,t-p}\}, \{h_{j,t-q}\})\right) \\ & p = 1, \cdots, P, q = 1, \cdots, Q \end{aligned} \tag{10}$$

$$y_{jt} | h_{jt} \sim g(h_{jt}); \qquad x_{it} = \sum_j A_{ij} y_{jt} \tag{11}$$

where f is any deterministic function, and g is a distribution with variance h_{jt}. For example, if we let

$$f(y_{j,t-p}, h_{j,t-q}) = c + a y_{j,t-1}^2 + b h_{j,t-1} \tag{12}$$

$$y_{jt} | h_{jt} \sim t(\sqrt{h_{jt}}, \nu) \propto \left(1 + \frac{y_{jt}^2}{h_{jt}(\nu - 2)} \right)^{-\frac{\nu+1}{2}} \tag{13}$$

the model reduces to a Garch(1,1) model with Student-t errors if we increase α in Equation 10. This model is also included in the experiments. The estimation and prediction methods described in the next sections can also be applied to this more general family of models with minor modifications.

Multivariate extensions of Garch and SV models also exist such as "Orthogonal Garch" (O-Garch or PCA-Garch) (Alexander, 2001), "GO-GARCH" (van der Weide, 2002), "ICA-Garch" (Wu & Yu, 2005), and factor Multivariate SV models (Chib et al., 2006). Compared to all these models, the multivariate DPoT model is a type of complete factor model with Garch style dynamics and inverse Gamma stochastic volatility.

Our work is also related to the inverse Gamma Markov chain model proposed in (Cemgil & Dikmen, 2007) which has been applied to audio signal processing for denoising and source separation.

In this paper, we only compare the basic form of DPoT to other models and do not take into account the leverage effect and jumps in the volatilities, but as described above, we can import the corresponding methods developed for Garch models into our DPoT without much effort.

5. Model estimation

The parameters of the DPoT model are estimated by maximizing the log-likelihood of the training data through the stochastic EM algorithm. Denote the whole set of parameters to be $\boldsymbol{\theta} = \{W_{ji}, \alpha_j, a_j, b_j, c_j\}$. In the E-step, the posterior distribution of $\mathbf{h}$ is computed by

$$\begin{aligned} & p(\mathbf{h}|\mathbf{x}) \propto p(\mathbf{h}, \mathbf{x}) \\ & = \prod_t \frac{1}{Z_t(\mathbf{h}_{t-1}, \mathbf{x}_{t-1})} e^{-E_t(\mathbf{h}_{t-1}, \mathbf{x}_{t-1}, \mathbf{h}_t, \mathbf{x}_t)} \end{aligned} \tag{14}$$

In the M-step, $\boldsymbol{\theta}$ is updated in the direction of the gradient of the expected log-likelihood:

$$\begin{aligned} & \boldsymbol{\theta} \Leftarrow \boldsymbol{\theta} + \eta \left\langle \nabla_{\boldsymbol{\theta}} L(\mathbf{x}, \mathbf{h}) \right\rangle_{p(\mathbf{h}|\mathbf{x})\tilde{p}(\mathbf{x})} \\ & \text{where} \quad L(\mathbf{x}, \mathbf{h}) = -\sum_t (E_t + \log Z_t) \end{aligned} \tag{15}$$

and η is the step size. Also, $\langle \cdot \rangle_{p(\mathbf{h}|\mathbf{x})\tilde{p}(\mathbf{x})}$ means expectation w.r.t. to the distribution $p(\mathbf{h}|\mathbf{x})\tilde{p}(\mathbf{x})$ with $\tilde{p}$ the empirical distribution.

Since there isn't a closed form expression for $p(\mathbf{h}|\mathbf{x})$, we run Gibbs sampling to draw samples from it and approximate the integration by summation.Conditional on $\mathbf{y}$ (or equivalently $\mathbf{x}$), the Markov processes for $h_{j,t},\ t = 1..T$ are independent over j and thus sampled separately. Also since the Markov processes are first order we can alternatingly block-sample the h_{jt} on even and odd time indices, where in each block the variables are independent with each other.

Sampling even a single h_{jt} variable given its neighbors is nontrivial because its posterior distribution is proportional to a product of an inverse Gamma and a truncated Gamma distribution:

$$\begin{aligned}
&p(h_{jt}|-) \propto \\
&\Gamma^{-1}(h_{jt}; \alpha_j, c_j + a_j y_{j,t-1}^2 + b_j h_{j,t-1} + \frac{1}{2} y_{jt}^2) \\
&\ \times \Gamma(h_{jt} + (c_j + a_j y_{jt}^2)/b_j; \alpha_j - \frac{1}{2}, b_j h_{j,t+1}^{-1}) \qquad (16)
\end{aligned}$$

We use rejection sampling to sample from this product where the upper bound is given by the first term times the maximum of the second term. Occasionally, the modes of these two distributions are so far apart (e.g. when the asset suddenly rises or drops), that the rejection rate becomes too high. For those cases we discretize the domain of h_{jt}. The average acceptance rate for this procedure is about 0.3 in our experiments.

We initialize W by running FastICA (Hyvarinen et al., 1999) over all data collapsed over time and using a small value for α_j. We then train a Garch(1,1)-normal model for each time series $y_{j,1:T} = W_j^T \mathbf{x}_{1:T}$ independently and initialize $a_j \leftarrow (\alpha_j - 3/2)\tilde{a}_j^{Garch}$, $b_j \leftarrow (\alpha_j - 3/2)\tilde{b}_j^{Garch}$, $c_j \leftarrow (\alpha_j - 3/2)\tilde{c}_j^{Garch}$. After this we apply stochastic EM as described above updating the parameters every few iterations of Gibbs sampling. Therefore, Gibbs sampling may not have converged in the E-step. But as long as the learning step size is small enough we expect that the sampler will not be very far from equilibrium.

6. Prediction

In the financial domain, what we really care about is the value of assets in the future. For prediction, we need to sample $h_{j,t}$ from its posterior and then simulate $h_{j,\tau} (\tau > t)$. Although Gibbs sampling runs fast during training, it is not as suitable for prediction, because we have to run Gibbs sampling until convergence every time a new price is observed. In contrast, particle filtering naturally incorporates information during the forward propagation by adjusting the weights of particles.

Auxiliary particle filtering (ASIR) (Pitt & Shephard, 1999) is adopted in this paper for filtering and prediction. We approximate the posterior distribution of h_{jt-1} by a mixture of delta functions concentrated on sample positions $\{h_{jt-1}^{(k)}\}$ with associated weights $\{w_k\}$. The proposal joint distribution of (k, h_{jt}) is exactly the posterior distribution:

$$\begin{aligned}
&g(k, h_{jt}|y_{j,1:t}) = p(k, h_{jt}|y_{j,1:t}) \\
&\ \propto p(k|y_{j,1:t-1}) p(y_{jt}|k, y_{j,1:t-1}) p(h_{jt}|k, y_{j1:t}) \qquad (17)
\end{aligned}$$

Since $p(k|y_{j,1:t-1}) = w_k$, and $p(y_{jt}|k, y_{j,1:t-1})$ is easy to compute (Equation 6), we can draw the mixture index k by $p(k) \propto p(k|y_{j,1:t-1}) p(y_{jt}|k, y_{j,1:t-1})$, and then draw h_{jt} from $p(h_{jt}|k, y_{j1:t})$ which is an inverse Gamma distribution. It's trivial to see that this algorithm is fully adapted in the sense that the second-stage weights are equal ($w_k = 1, \forall k$) and therefore we do not need resampling. For DPoT with student-t errors, ASIR is not fully adapted, but we can still draw samples efficiently from the posterior distribution.

After obtaining samples $h_{jt}^{(k)}$ given $\mathbf{x}_{1:t}$, predictions on statistics of interst can be obtained by first simulating future volatilities from the particles by $p(\mathbf{h}_t|\mathbf{h}_{t-1}^{(k)})$ and then approximating the expectation by summation. Take for instance the estimation of cdf of one day ahead returns (addressed in next section):

$$\begin{aligned}
cdf(x_{it}) &= E_{\mathbf{h}_t|\mathbf{x}_{1:t-1}}[P(x \leq x_{it}|\mathbf{h}_t)] \\
&\approx \sum_k w_{t-1}^{(k)} \Phi\left(x_{it}/\sqrt{\Sigma_{ii}(\tilde{\mathbf{h}}_t^{(k)})}\right) \qquad (18)
\end{aligned}$$

where given $\mathbf{h}_t$, $\mathbf{x}_t$ follows a joint Gaussian distribution with covariance matrix $\Sigma = AH_tA^T$, $H_{jj,t} = h_{j,t}$.

7. Value at Risk

Value at Risk (VaR) is a widely accepted measure of the risk of loss on a portfolio of financial assets (Kuester et al., 2006; So & Yu, 2006). Given a loss level l and a time horizon τ, we can compute the probability that our real loss will exceed that level l at time $t + \tau$. Say, that we do not want that probability to be more than λ. The smallest level l that still satisfies that constraint is called the Value at Risk $VaR_{t+\tau,\lambda}$. Mathematically, the VaR is expressed as:

$$\begin{aligned}
&VaR_{t+\tau,\lambda} \triangleq \inf_l \{P(L_{t+\tau} > l) \leq \lambda\} \\
&= -\sup_x \{cdf_{x_{t+\tau}}(x) \leq \lambda\} \triangleq -Q_{t+\tau}(\lambda) \qquad (19)
\end{aligned}$$

where $L_{t+\tau}$ is the loss, $x_{t+\tau} (= -L_{t+\tau})$ the return of a portfolio, and $Q_{t+\tau}(\lambda)$ the λ-quantile of the return $x_{t+\tau}$. For example, as shown in Figure 2 given a probability level $\lambda = 5\%$, $\tau = 2$ days, a stock's $VaR = 1.64$ means there's a chance of 5% that the price of this stock will drop by at least 1.64 in 2 days. A time horizon of one day is used

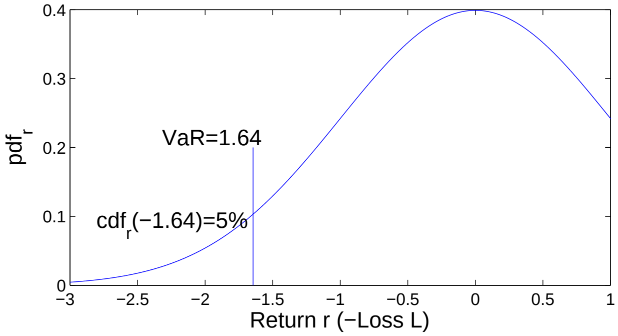

Figure 2. An example of VaR at level λ=5%, the return follows a stardard Gaussian distribution.

in our experiments. For a longer horizon, we can use our Monte Carlo method to simulate VaRs τ days ahead.

Backtesting VaR is a statistical technique to verify if the failure rate, that is the proportion of days an actual loss exceeds the predicted VaR, is in agreement with the risk level λ. Ideally, the failure rate should be an unbiased and consistent estimator of λ. In practice, small deviation of this quantity from its corresponding risk level suggests good predicting performance. Equivalently, we compute this value as the proportion of days that the cdf of an actual return in the predicted distribution falls below λ. Denote with $F(\lambda)$ the failure rate at level λ. Figures 7 and 8 illustrate the deviation $F(\lambda) - \lambda$ against the VaR level λ. A good model should give a line that stays close to the x-axis (dashed line). Usually, the deviation at a specific level (e.g. 1%) is quite noisy. In the experiments we have used more stable criteria such as mean absolute error (MAE) and root mean squared error (RMSE) from 0 to λ, see (Kuester et al., 2006).

8. Experiments

We will apply the DPoT model with normal conditional distribution (DPoT(N)) and one of the generalized DPoT models with a Student-t conditional distribution (DPoT(T)) to stock prices and test their performance on VaR prediction. For comparison, we browse the literature and select a few representative models including PCA-Garch, ICA-Garch and a multivariate SV model discribed in (Chib et al., 2006) with 2 factors and without jumps [1]. We consider both normal and student-t errors for each model. The dataset for the experiment of the univariate model is the closing daily prices of the S&P 500 index from Jan 5, 1960 to Dec 30, 2005, totalling 11579 trading days. The price (p_t) and percentage log-returns defined by

$$x_t = 100 \log \left(p_t / p_{t-1} \right) \tag{20}$$

are plotted in Figure 3. The dataset used for the multivariate models is from a set of 10 stocks: AAPL, HPQ, MSFT, ADI, INTC, TXN, C, JPM, WFC, GE, from the period July 10, 1986 to Dec 30, 2005, totalling 4917 trading days.

We use a rolling window of 2000 days to account for the change of parameters over time. We estimate a separate model on the returns in each window, compute the cdfs of one day ahead returns for the next 50 days and then we move the window forward by 50 days and repeat the process. In each window, the mean is subtracted from the returns before it is fed to the training algorithm. This procedure is standard practice for training the other two classes of volatility models as well.

8.1. Univariate model

Univarite DPoT(N/T) models are trained on the S&P 500 index. The estimated parameter α is shown in Figure 4. Larger values of $\hat{\alpha}$ imply less kurtosis in the marginal distribution of the returns which means that DPoT and Garch are expected to behave similarly. We can find that $\hat{\alpha}$ for DPoT(N) is large from the late 70's to the early 80's corresponding to a stable period of volatilities. However, it decreases fast in the late 80's where we find a large spike in the percentage log-returns (see Figure 3). The value of $\hat{\alpha}$ for DPoT(T) is larger than 100 everywhere which means that we expect similar performance of DPoT(T) and Garch(T) on univariate data.

The deviation of VaR prediction is shown in Figure 7. The plot for our DPoT(N) model is much more stable and consistently closer to 0 than that of Garch(N) (except for a small region where Garch(N) crosses the horizontal line). Results for DPoT(T) and Garch(T) are similar as expected, and both better than the models with Normal errors for small VaR levels (which are usually of more practical interest than the larger levels) but increase fast afterward. Plots of DPoT and SV models are close to each other, and the former usually performs better than the latter with the same type of error at $\lambda > 4\%$. The MAE and RMSE of deviations for four levels 1%, 2.5%, 5%, 10% are plotted in Figure 5 and 6. DPoT(T) is among the best models in either figure.

8.2. Multivariate model

The multivariate DPoT(N/T) models for 10 stocks are compared to the PCA-Garch, ICA-Garch, MSV models with normal and student-t errors. Following the recommended training procedure for the PCA/ICA-Garch models (Alexander, 2001; Wu & Yu, 2005), we first estimate the demixing matrix using PCA/ICA on all the data col-

[1] We estimate Garch models with the UCSD Garch matlab toolbox and the MSV model with code from the authors of (Chib et al., 2006)

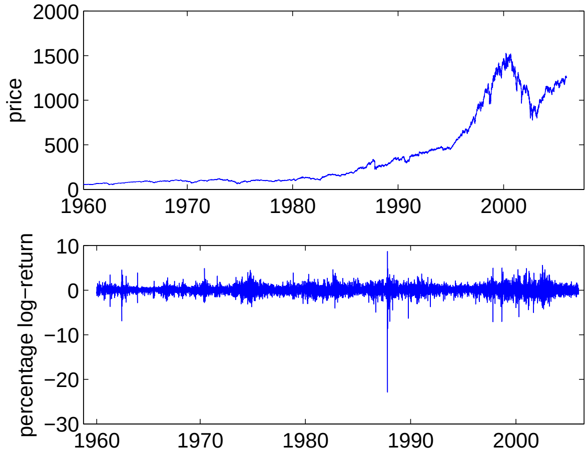

Figure 3. Prices and Percentage log-returns of S&P 500 index from Jan 5, 1960 to Dec 30, 2005.

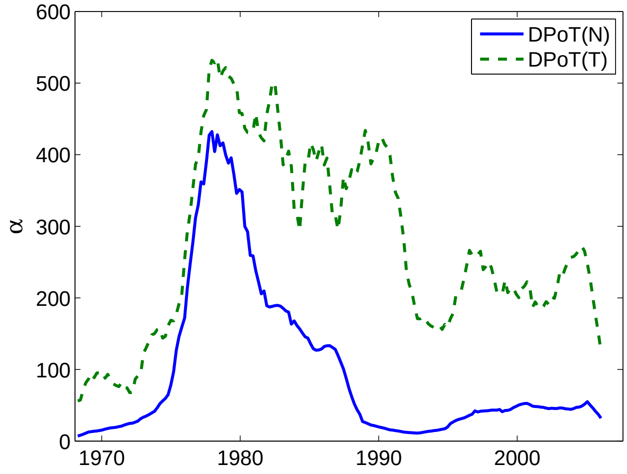

Figure 4. Estimated α for S&P 500 index from 1960$\sim$2005. The time axis corresponds to the end of each sliding window.

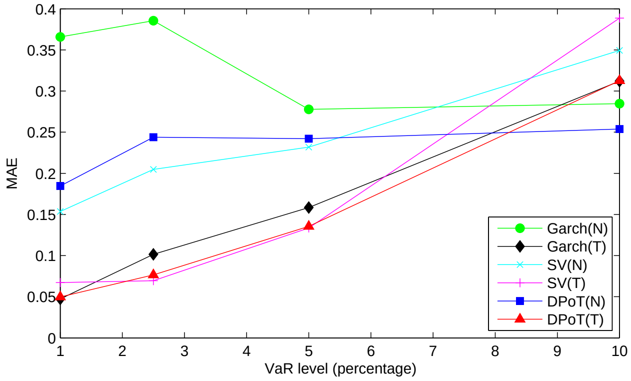

Figure 5. MAE of VaR deviation on the S&P 500 index from 1960$\sim$2005 for Garch(N) (circle), Garch(T) (diamond), SV(N) (x), SV(T) (plus), DPoT(N) (square) and DPoT(T) (triangle) model. The x axis is the VaR level in percentage.

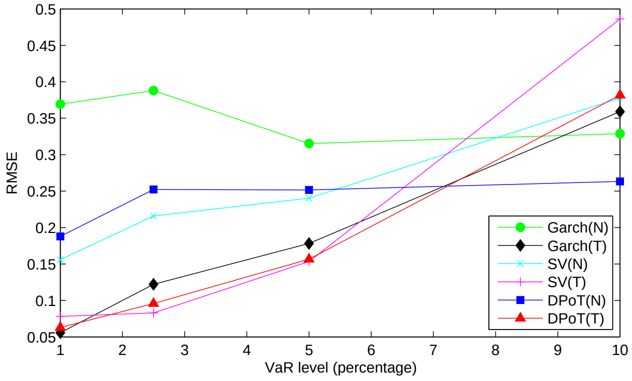

Figure 6. RMSE of VaR deviation on the S&P 500 index from 1960$\sim$2005 for Garch(N) (circle), Garch(T) (diamond), SV(N) (x), SV(T) (plus), DPoT(N) (square) and DPoT(T) (triangle) model. The x axis is the VaR level in percentage.

lapsed over time. After that, the percentage log-returns are linearly transformed into their latent factor spaces and the corresponding univariate Garch(N/T) models are independently fit across time.

Due to space limitations, we only show the results of deviations against VaR levels in Figure 8 on the whole set of 10 stocks. The advantage of our models is clear for the multivariate case. Both PCA- and ICA-Garch(N) models decrease fast due to their thin tails in the conditional distribution. PCA-Garch(T) also behaves badly, presumably because PCA cannot find independent factors that control the volatility over time. ICA-Garch(T) performs well for very small values of $\lambda < 0.5\%$ but at larger values it quickly becomes inferior to both types of DPoT models. Both MSV(N) and MSV(T) perform worse than DPoT(N/T), indicating the advantage of the DPoT type of dynamics over the regular SV models. If we look at $b/(\alpha - 1)$ (direct propagation of volatility from previous returns which doesn't exist in MSV) in the fitted DPoT models, it has a significant positive value in over half of the sliding windows.

Due to space limitations we are not able to show MAE/RMSE values for all the 10 stocks on 4 levels for all 6 models. Alternatively, we use a measurement called mean rank (So & Yu, 2006) to compare the average performance across stocks (see Table 1). For each λ level, it first ranks models in each stock. Smaller ranks are assigned to smaller deviations. Then the mean of ranks across stocks is computed. The DPoT(T) model has the highest mean rank for almost all the λ levels followed by DPoT(N) and then MSV(N/T). MSV(T) works the best at level $\lambda = 1\%$, but degenerates fast at larger levels. Among the Garch models, only ICA-Garch(T) works comparably to DPoT and MSV, and the other 3 models rank quite low, consistent with the plots in Figure 8

Table 1. Mean ranks of models over 10 stocks and their final ranking (in parentheses) across models at each VaR level according to MAE and RMSE. A smaller rank is better. "G" means Garch model.

	MAE							
VaR Level	DPoT(N)	DPoT(T)	ICA-G(N)	PCA-G(N)	ICA-G(T)	PCA-G(T)	MSV(N)	MSV(T)
1%	4.0(2)	4.1(3)	5.9(8)	5.8(7)	4.3(5)	4.2(4)	4.4(6)	**3.3(1)**
2.5%	3.6(4)	**3.0(1)**	6.9(7)	7.1(8)	3.7(5)	4.9(6)	3.4(2)	3.4(3)
5%	3.1(2)	**2.8(1)**	6.7(7)	7.8(8)	3.2(3)	5.2(6)	3.4(4)	3.8(5)
10%	**2.7(1)**	**2.7(1)**	7.0(7)	8.0(8)	3.0(4)	5.5(6)	2.9(3)	4.2(5)
	RMSE							
VaR Level	DPoT(N)	DPoT(T)	ICA-G(N)	PCA-G(N)	ICA-G(T)	PCA-G(T)	MSV(N)	MSV(T)
1%	4.0(3)	3.8(2)	5.9(7)	6.1(8)	4.4(5)	4.4(5)	4.0(3)	**3.4(1)**
2.5%	3.4(2)	**3.0(1)**	6.8(7)	7.3(8)	3.7(5)	4.9(6)	3.4(2)	3.5(4)
5%	3.0(2)	**2.8(1)**	6.8(7)	8.0(8)	3.2(3)	5.1(6)	3.2(3)	3.9(5)
10%	2.7(2)	**2.6(1)**	7.0(7)	8.0(8)	3.0(4)	5.6(6)	2.8(3)	4.3(5)

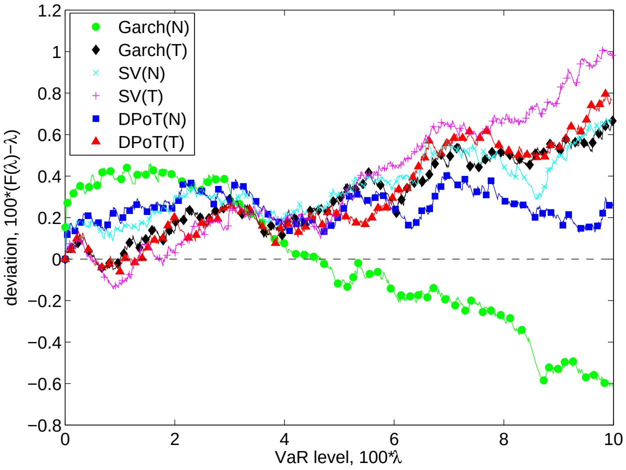

Figure 7. Deviation plots on the S&P 500 index from 1960∼2005 for Garch(N) (circle), Garch(T) (diamond), SV(N) (x), SV(T) (plus), DPoT(N) (square) and DPoT(T) (triangle) model. The x axis is the VaR level in percentage and the y axis is the deviation of the failure rate from the given level in percentage. The closer to the horizontal line, the better prediction a model makes.

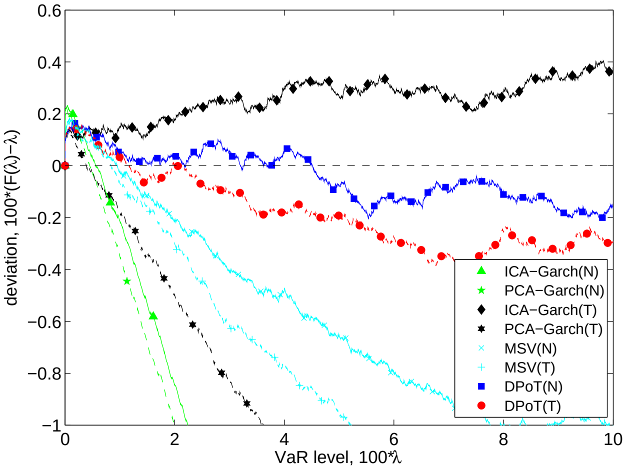

Figure 8. Deviation plots on the entire dataset of 10 stocks for ICA-Garch(N) (triangle), PCA-Garch(N) (pentagram), ICA-Garch(T) (diamond) and PCA-Garch(T) (hexagram), MSV(N) (x), MSV(T) (plus), DPoT(N) (square), DPoT(T) (circle) model. $F(\lambda)$ is the proportion of cdfs below λ in all the 10 stocks. The closer to the horizontal line, the better prediction a model makes.

9. Discussion

Predicting value at risk is of great importance to the financial community because it allows traders to assess the risk for their portfolio of stocks. Inspired by the similarities between natural signals such as images and sound, we have introduced an extension to the popular Garch models that treats the latent process over volatilities as a stochastic process rather than a deterministic regression. Experimentally we have shown that this results in a significant improvement to both regular Garch and SV models for multivariate models over multiple stocks in terms of Value at Risk.

Besides the extensions mentioned in section 4 an exciting direction for improving the DPoT model is by exploiting its close relationship to the HPoT models studied in (Osindero et al., 2006). In that paper over-complete models ($m > n$) and their hierarchical extensions were studied and the close relation to simple and complex cells in V1 was established. Moreover, there is mounting evidence that these type of architectures can be stacked into deep hierarchies (MarcAurelio Ranzato et al.; Hinton et al., 2006). Multi-layer DPoT models can potentially be very powerful in modeling the long range interactions in variance in financial time series. Learning these models from data will be more challenging than what was needed for the DPoT, but new algorithms have become available recently in machine learning that could make this possible (Hinton, 2002). We believe that the close connections between the statistics of sound and images on the one hand and financial time series on the other make the latter an exciting playing field to test new ideas in “deep learning”.

Acknowledgments

This work is supported in part by NSF grants IIS- 0447903 and IIS-0535278 as well as ONR/MURI grant 00014-06-1-073.

References

Alexander, C. Orthogonal garch. *Mastering risk*, 2:21–38, 2001.

Bell, A.J. and Sejnowski, T.J. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

Bollerslev, T. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics*, 31(3):307–327, 1986.

Cemgil, A. T. and Dikmen, O. Conjugate gamma Markov random fields for modelling nonstationary sources. In *ICA 2007, 7th International Conference on Independent Component Analysis and Signal Separation*, pp. 697–705, September 2007.

Chib, S., Nardari, F., and Shephard, N. Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2):341–371, 2006.

Gehler, P. and Welling, M. Products of“Edge-perts”. *Advances in Neural Information Processing Systems*, 18: 419, 2006.

Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.

Hinton, G.E., Osindero, S., and Teh, Y.W. A fast learning algorithm for deep belief nets. *Neural Computation*, 18 (7):1527–1554, 2006.

Hyvarinen, A. et al. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

Kuester, K., Mittnik, S., and Paolella, M.S. Value-at-risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics*, 4(1):53–89, 2006.

Lyu, S and Simoncelli, E P. Reducing statistical dependencies in natural signals using radial Gaussianization. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L. (eds.), *Adv. Neural Information Processing Systems 21*, volume 21, pp. 1009–1016, Cambridge, MA, May 2009. MIT Press.

MarcAurelio Ranzato, Y., Boureau, L., and LeCun, Y. Sparse feature learning for deep belief networks. *Advances in neural information processing systems*, 20: 1185–1192.

Osindero, S., Welling, M., and Hinton, G.E. Topographic product models applied to natural scene statistics. *Neural Computation*, 18(2):381–414, 2006.

Pearlmutter, B. and Parra, L. A context sensitive generalization of ICA. *Proc. of the Int’l Conf. on Neural Information Processing*, 9:151–157, 1996.

Pitt, M.K. and Shephard, N. Filtering via simulation: Auxiliary particle filters. *Journal of the American Statistical Association*, 94(446):590–599, 1999.

Roth, S. and Black, MJ. Fields of experts: A framework for learning image priors. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, pp. 860–867, 2005.

Shephard, N., Andersen, T.G., and NBER, C. Stochastic volatility: origins and overview. *Handbook of Financial Time Series*, pp. 233–254, 2008.

So, M.K.P. and Yu, P.L.H. Empirical analysis of garch models in value at risk estimation. *Journal of International Financial Markets, Institutions & Money*, 16(2): 180–197, 2006.

Sutskever, I., Hinton, G., and Taylor, G. The recurrent temporal restricted boltzmann machine. *NIPS*, pp. 1601–1608, 2009.

Taylor, S.J. Financial returns modelled by the product of two stochastic processesa study of the daily sugar prices 1961-75. *Time Series Analysis: Theory and Practice*, 1: 203–226, 1982.

van der Weide, R. Go-garch: A multivariate generalized orthogonal garch model. *Journal of Applied Econometrics*, pp. 549–564, 2002.

Wainwright, M.J. and Simoncelli, E.P. Scale mixtures of Gaussians and the statistics of natural images. In *Advances Neural Information Processing Systems*, volume 12, pp. 855–861, 2000.

Welling, M., Hinton, G.E., and Osindero, S. Learning sparse topographic representations with products of student-t distributions. In *Neural Information Processing Systems*, volume 15, pp. 1359–1366, Vancouver, Canada, 2002.

Wu, E.H.C. and Yu, P.L.H. Volatility modelling of multivariate financial time series by using ica-garch models. *Intelligent Data Engineering and Automated Learning-IDEAL 2005, Lecture Notes in Computer Science*, 3578: 571–579, 2005.