

**SV mixture models with application to
S&P 500 index returns**

Online supplement

Appendix A

Conditional densities of some continuous-time models

Since the Euler scheme approximation shown in Eq. (2) is based on a discretization interval equal to the data sampling interval (one day), the conditional returns density, $X_{t+1}|V_t$, is Gaussian. However, if the true continuous-time model (Eq. 1) were used, this would no longer be true (due to variation in the volatility factor over the course of the day). In particular, correlation between the Brownian motions driving the process would induce skewness in the conditional returns distribution. It is possible that this skewness could enable the model to fit the data without needing to resort to the mixture idea.

In consideration of this idea, Figure 23 shows conditional returns densities corresponding to one- and two-factor continuous-time models with several different parameter settings. The parameters are taken from the estimates shown in Table 2 (transformed appropriately to reflect the slightly different parameterizations of the discrete-time and continuous-time models, as described after Eq. (2)), but with several alternative values for ρ_{21} , the correlation parameter. The plots are constructed in a manner analogous to those in Figure 2. The continuous-time models are simulated using the Euler-scheme approximation, but with the time increments equal to one-hundredth of a day. For both the one- and two-factor models, V_t is initialized at zero. For the two-factor model, U_t is initialized by drawing from its marginal distribution. The conditional density corresponding to the MIX3 model is shown for comparison (it is identical to the one plotted in Figure 2).

While the continuous-time models do indeed induce some skewness, the resulting densities fail to capture the shape implied by the mixture density. For the one-factor model, thickening the left tail causes the right tail to be too thin; there does not appear to be much hope of matching the extreme left tail. The two-factor model does somewhat better. With no correlation, the right tail is too thick. Adding in some correlation thins it out, while at the same time thickening the left tail — both steps in the correct direction. However, the shape of the left tail still fails to capture the shape implied by the mixture distribution.

While the preceding simulations are suggestive that taking the continuous-time

nature of the models into account is not sufficient to match the conditional distributions exhibited by the data, a more complete analysis would involve estimating and testing the models using the approach applied to the other models in this paper. Unfortunately, this is not feasible using techniques currently available.

On the other hand, such models have been estimated using the simulated method of moments. For example, Chernov, Gallant, Ghysels and Tauchen (2003) fit a variety of continuous-time models over Dow Jones Industrial Average returns. Their diagnostics reject the standard two-factor model at the 5% (but not the 1% level). Their preferred models incorporate what they refer to as “volatility feedback”, whereby the volatility of volatility increases as the level of volatility does. With the appropriate parameter settings, such models (or variations on them) may be able to generate thickness in the far left tail such as that exhibited by the SV-mix model.

The extensive literature on jump-diffusion models also provides evidence that standard diffusion models are unable to capture important features of the conditional distribution of returns.

Appendix B

Out-of-Sample Performance

The analysis elsewhere in this paper is all based on in-sample fit. The problem with this is that a highly flexible model can fit the data very well in-sample, but perform poorly out-of-sample. At issue is whether the model is just fitting artifacts that appear in the sample but are not robust features of the true data-generating process. To test for this possibility, some experiments to gauge the out-of-sample performance of the MIX3 model were performed. The idea is to fit the model over different subsamples then check how sensitive the parameter estimates are to choice of subsample and examine some diagnostics on the out-of-sample performance of the fitted models. For comparison, results for the SV2 model on the same experiments are also presented.

Parameter estimates obtained by fitting the MIX3 and SV2 models over several subsamples are shown in Table 6. The first subsample corresponds to the first 2100 observations of the full sample (June 23, 1980 - Oct 12, 1988). This subsample

includes the most extreme crash event, October 19, 1987, but not the second most extreme event, October 13, 1989. The second subsample corresponds to the first 2800 observations (June 23, 1980 - July 22, 1991; roughly the first half of the sample), and includes both of the two extreme crash days. The third subsample examined corresponds to observations 2801 through 5616 (and includes neither of the two extreme crash days).

The parameter estimates for the MIX3 models are similar across all subsamples. Among the non-mixture parameters, the main difference is that the leverage parameter, ρ , is lower in the earlier part of the sample than in the latter part (-0.45 for the first half of the sample versus -0.70 for the second half). Figure 24 shows the conditional distributions for $X_{t+1}|V_t = 0$ implied by the estimates from the various subsamples. These densities are roughly similar in shape. The density corresponding to observations 2801-5616 does not extend as far out in the left tail as the others. This is not surprising since this subsample does not include the two extreme crash events. Implied probabilities for less extreme events are similar for all three sets of estimates. For example, the probability of a return less than -0.04 (conditional on $V_t = 0$) is 0.0019 based on estimates from the first subsample, 0.0018 based on estimates from the second subsample, and 0.0011 based on estimates from the third subsample (compared to 0.0015 for the full sample). These probabilities are reasonably consistent considering the rarity of the events being estimated (one would only expect to see about 8 such events in a sample of size $n = 5616$).

For the SV2 model, in contrast, there is considerable variation in the parameters corresponding to the second volatility factor across subsamples. The persistence parameter, ϕ_U , is -0.10 on the first half of the data, and 0.73 on the second half. The leverage parameter, ρ_{31} , goes from -0.064 on the first half of the data to -0.594 on the second half.

Table 7 shows log likelihoods over the out-of-sample portion of the data for each experiment. This provides an information theoretic means of assessing out-of-sample model fit. The MIX3 model dominates SV2 in every case. Note that there is no need to apply penalties based on the number of parameters since over-parameterized models would not be expected to perform better out-of-sample.

A better understanding of what is going on can be gotten by looking at the qq-plots in Figure 25. These are obtained by computing generalized residuals for the

out-of-sample observations based on parameter estimates from the various subsamples. When the MIX3 model is estimated based on the first 2100 observations, the generalized residuals corresponding to the remaining observations are almost perfectly normal, consistent with the hypothesis that the model is correctly specified. When the model is estimated on the first 2800 observations, the implied model is slightly fatter in the extreme left tail than the out-of-sample portion of the data. This is not surprising since the in-sample data contains both of the two most extreme crashes. Likewise, when the model is estimated over observations 2801 through 5616, the implied model matches the data well over most of the distribution, but misses the two extreme crash days. Again, this is not surprising since no events of this magnitude occurred in the subsample over which the estimates were based.

In all three out-of-sample experiments, the SV2 model badly misses the left-hand tail of the data. When estimated over data from the first half of the sample, the right tail of the implied model is thicker than the out-of-sample data. When estimated over data from the second half of the sample, the right tail of the implied model is slightly too thin.

The Jarque-Bera test statistic provides a means for quantifying the deviation from normality of the generalized residuals. Table 7 reports Jarque-Bera test statistics computed over the out-of-sample portion of the generalized residuals implied by the SV2 and MIX3 estimates obtained in each experiment. In each case, the statistics support the evidence of the qq-plots that the generalized residuals corresponding to the MIX 3 models deviate substantially less from normality than do those of the SV2 model. The p-values are based on the assumption of an in-sample test, so tests based on out-of-sample residuals would be expected to over-reject. Nonetheless, the test does not reject the MIX3 model at any conventional significance level for either of the first two experiments, and rejects at the 5% level but not the 1% level in the third experiment. Again, it is not surprising that the test indicates rejection in the third experiment since the two most extreme events are included in the subsample over which the statistic is computed, but not the subsample over which the model is estimated.

The in-sample diagnostics reported in Section 4 also included correlograms and Box-Pierce statistics for the squared residuals. These were computed for the out-of-sample experiments but are not reported. The results are about the same as for the

full sample: none of the models are rejected at conventional significance levels by the Box-Pierce test on 20 lags, and all are rejected by the Box-Pierce test on 250 lags. Performance of the SV2 and MIX3 models is about equal on these diagnostics.

Overall, these out-of-sample experiments suggest that the features of the data captured by the MIX3 model are consistent across subsamples. In particular, the evidence does not indicate that the model is over-fitting the data. Although a certain amount of uncertainty regarding tail behavior is unavoidable given the rarity of the events being measured, the problem would be expected to affect commonly used jump-diffusion models to much the same extent.

Additional tables and figures

Table 6. Parameter estimates for various subsamples.

Model	μ	σ_X	ϕ_V	σ_V	ρ_{21}	ϕ_U	σ_U	ρ_{31}
SV2 (full sample)	0.00008 (0.00010)	0.00830 (0.00063)	0.9905 (0.0027)	0.101 (0.014)	-0.459 (0.088)	0.15 (0.27)	0.468 (0.066)	-0.215 (0.112)
SV2 (Obs. 1-2100)	0.00001 (0.00018)	0.00845 (0.00067)	0.9849 (0.0059)	0.106 (0.017)	-0.316 (0.110)	-0.15 (0.14)	0.541 (0.053)	-0.023 (0.073)
SV2 (Obs. 1-2800)	0.00011 (0.00016)	0.00831 (0.00050)	0.9838 (0.0055)	0.099 (0.015)	-0.326 (0.099)	-0.10 (0.12)	0.543 (0.044)	-0.064 (0.064)
SV2 (Obs. 2801-5616)	0.00003 (0.00014)	0.00872 (0.00145)	0.9955 (0.0027)	0.078 (0.019)	-0.421 (0.127)	0.73 (0.09)	0.328 (0.048)	-0.594 (0.078)

Model	μ	σ_X	ϕ_V	σ_V	ρ	skew	kurt
MIX3 (full sample)	-0.00004 (0.00011)	0.00902 (0.00058)	0.9872 (0.0029)	0.118 (0.010)	-0.577 (0.047)	-0.46	6.37
MIX3 (Obs. 1-2100)	-0.00004 (0.00020)	0.00936 (0.00072)	0.9830 (0.0060)	0.105 (0.017)	-0.386 (0.105)	-0.63	8.16
MIX3 (Obs. 1-2800)	0.00000 (0.00017)	0.00924 (0.00055)	0.9807 (0.0060)	0.105 (0.016)	-0.450 (0.087)	-0.75	9.33
MIX3 (Obs. 2801-5616)	-0.00006 (0.00014)	0.00867 (0.00090)	0.9880 (0.0034)	0.137 (0.016)	-0.703 (0.050)	-0.31	4.50

Mixture components

	full sample	1-2100	1-2800	2801-5616
$\log p_1$	-0.176	-0.191	-0.183	-0.108
μ_1	0.015	0.029	0.016	0.035
σ_1	1.029	1.023	1.019	0.993
$\log p_2$	-5.504 (0.411)	-5.539 (0.692)	-5.705 (0.529)	-4.595 (0.316)
μ_2	-3.611 (1.247)	-4.268 (1.906)	-4.745 (1.522)	-2.683 (0.808)
σ_2	2.651 (0.595)	2.807 (0.361)	3.020 (0.396)	1.611 (0.571)
$\log p_3$	-1.848 (0.242)	-1.774 (0.366)	-1.806 (0.260)	-2.388 (0.675)
μ_3	0.015 (0.048)	-0.041 (0.070)	0.016 (0.061)	-0.049 (0.138)
σ_3	0.444 (0.058)	0.436 (0.081)	0.423 (0.067)	0.387 (0.145)

Table 7. Out-of-sample diagnostics.

The models were estimated on the indicated subsample, then the log likelihoods and Jarque-Bera statistics were computed on the out-of-sample portion of the data.

Model	Log L (out-of-sample)	Jarque-Bera (out-of-sample)
MIX3 (1-2100)	11717.76	4.26 (0.12)
SV2 (1-2100)	11707.72	18.50 (0.0000)
MIX3 (1-2800)	9381.84	2.84 (0.24)
SV2 (1-2800)	9375.84	9.72 (0.0077)
MIX3 (2801-5616)	9182.56	9.21 (0.010)
SV2 (2801-5616)	9157.40	53.94 (0.0000)

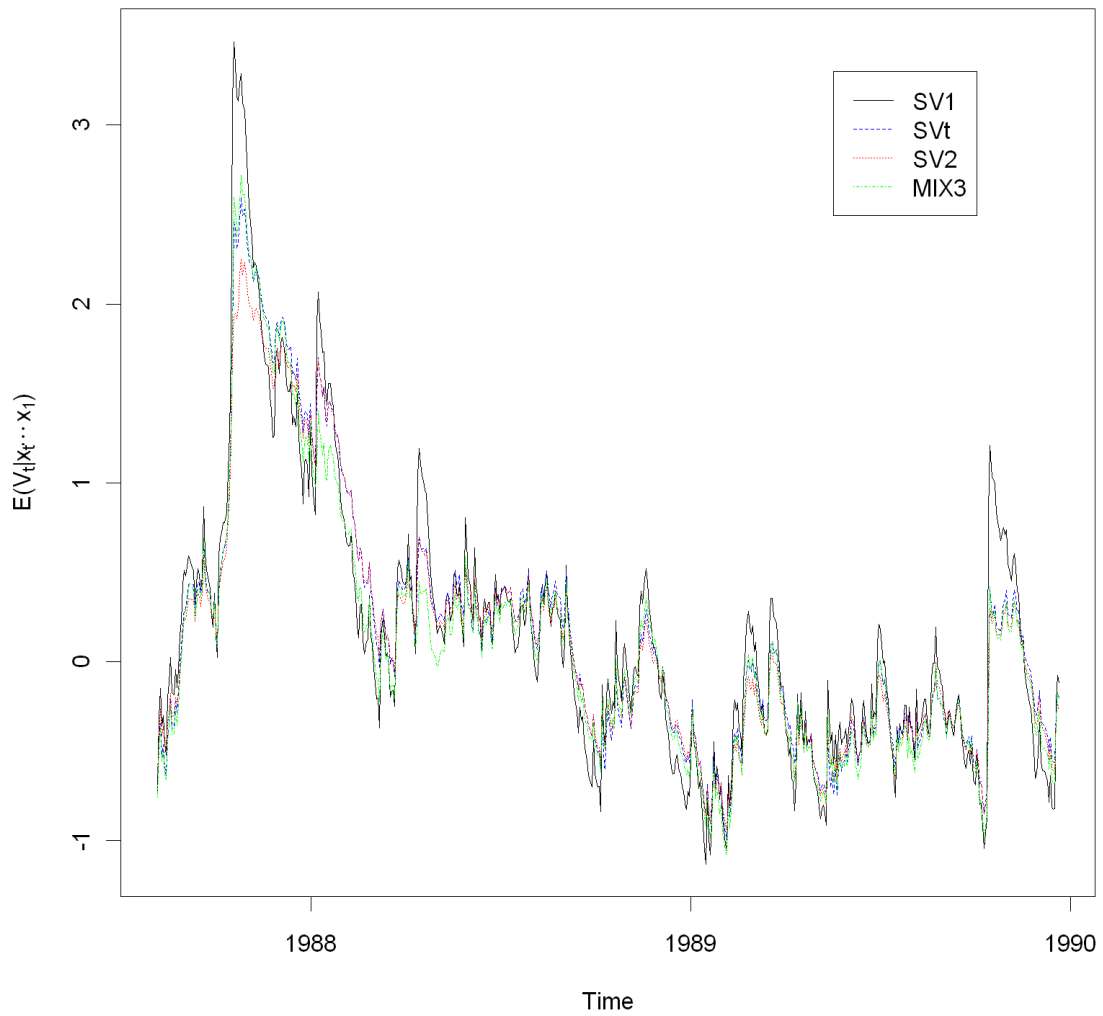


Figure 19. Filtered volatility estimates for various models. (This is identical to Figure 8 except that it uses color).

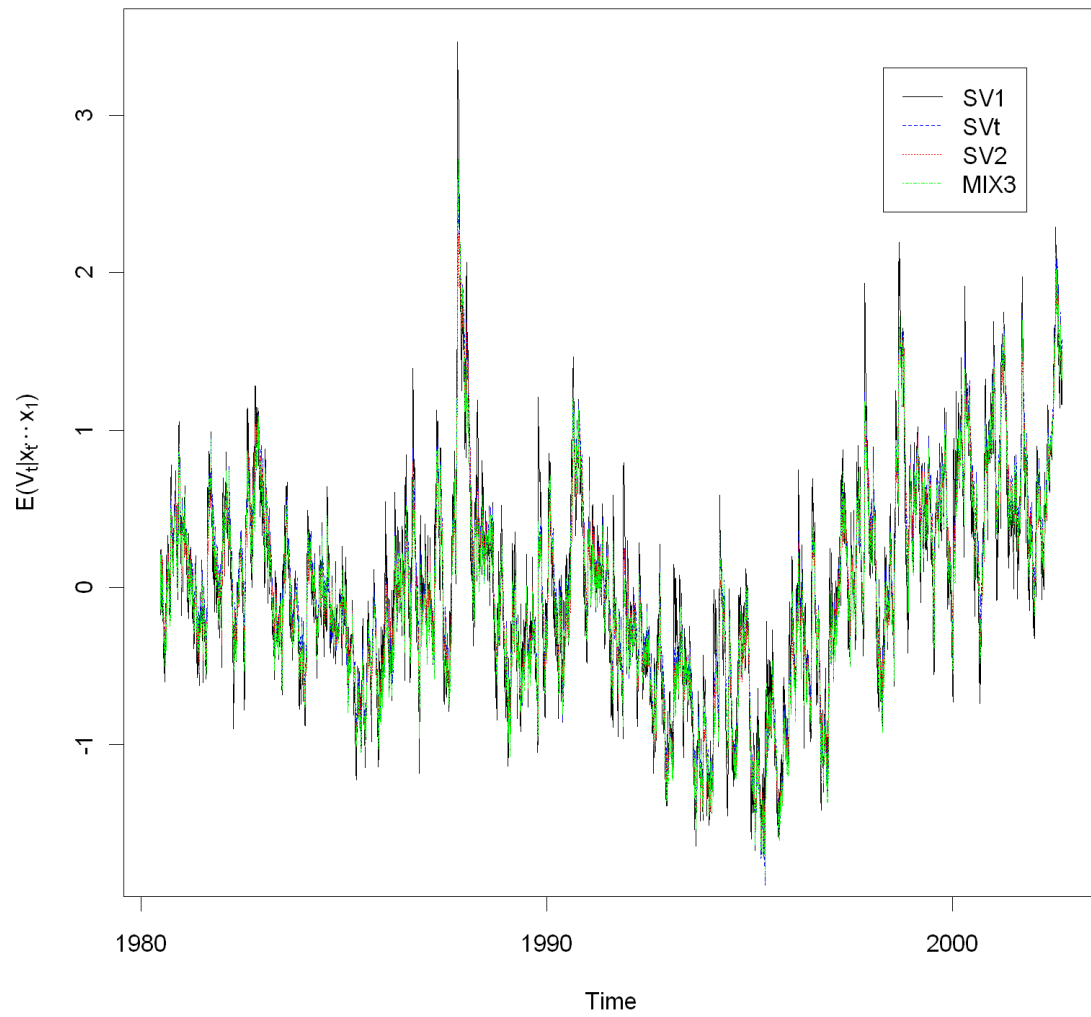


Figure 20. Filtered volatility estimates for various models.

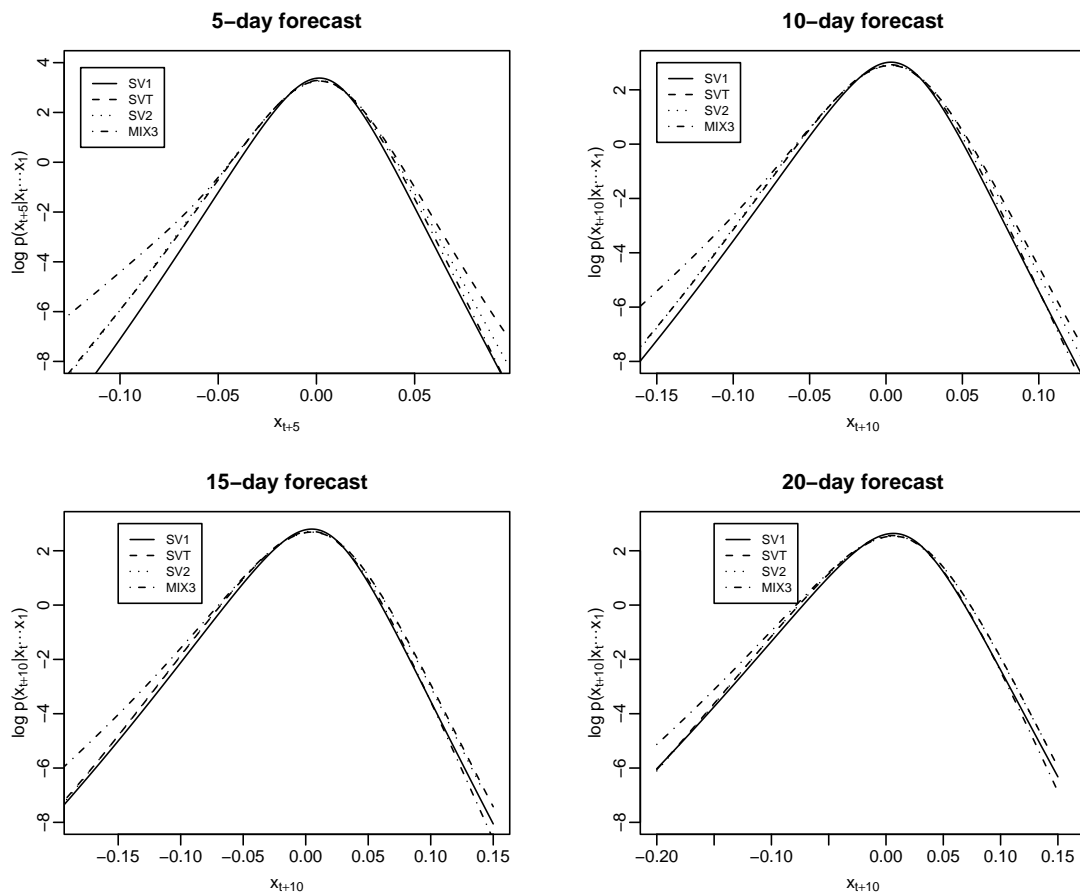


Figure 21. Forecast densities of cumulative returns at 5- and 10-, 15- and 20-day horizons for August 1, 1991.

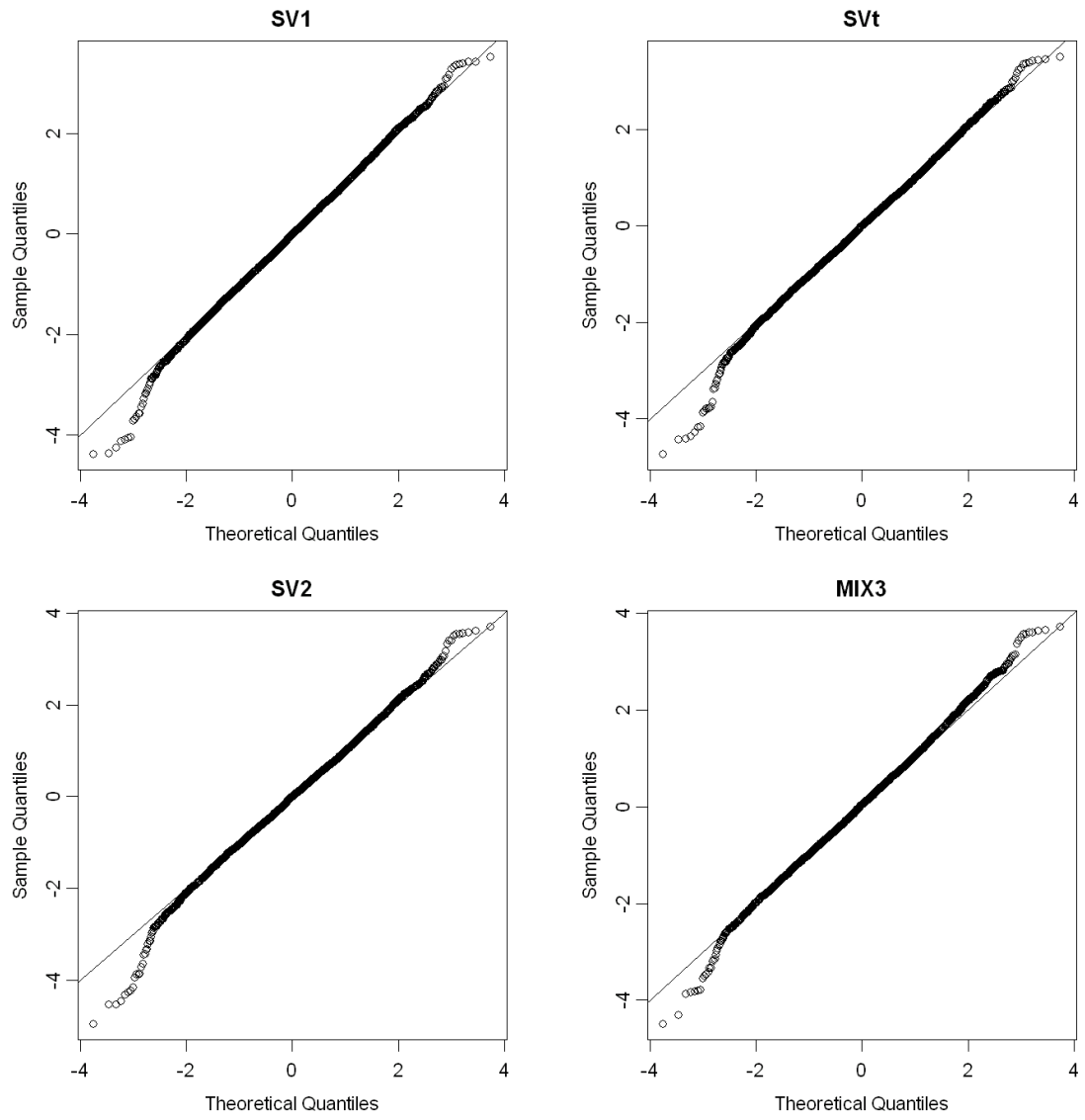


Figure 22. QQ-plots for generalized residuals of 20-day cumulative returns.

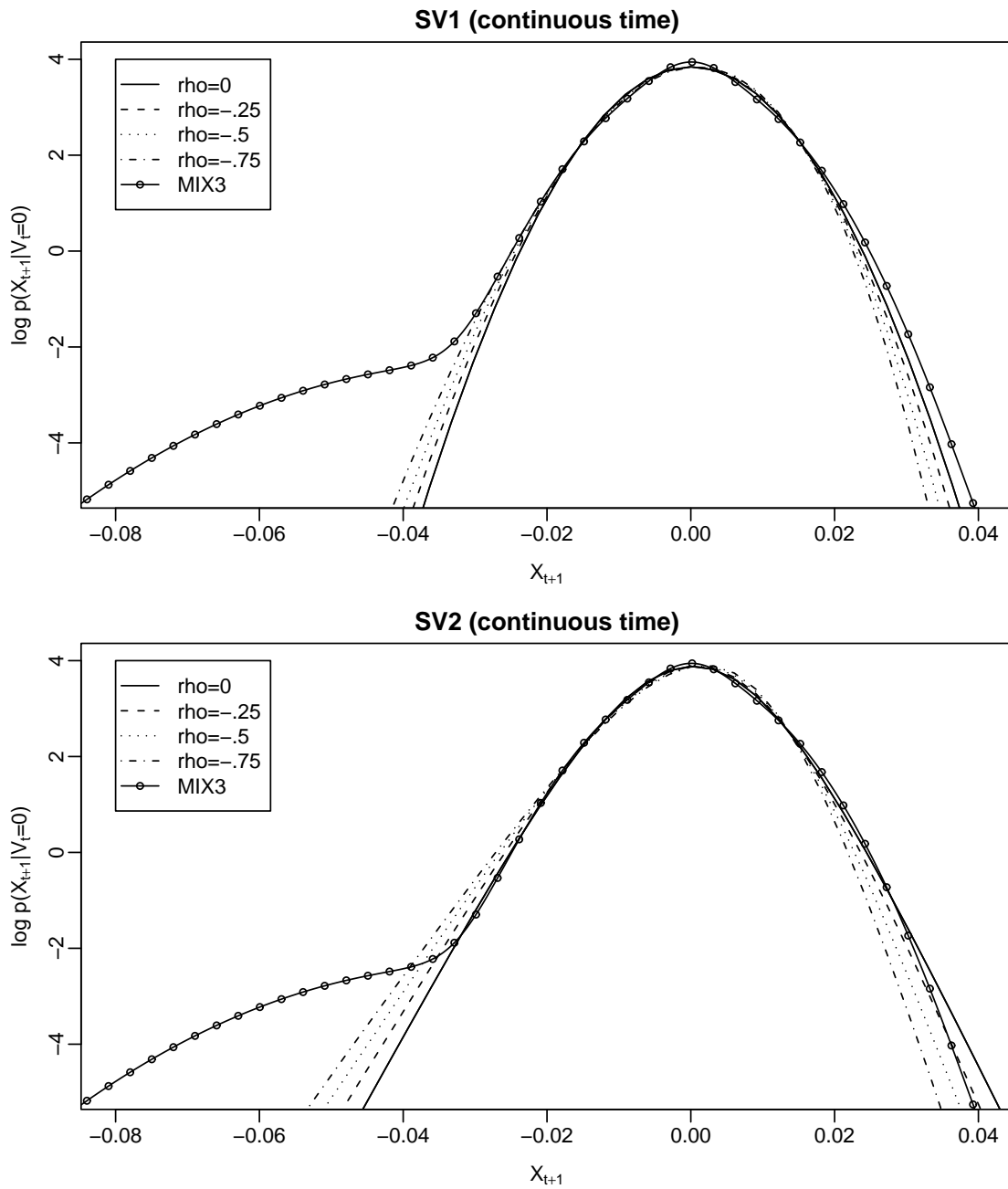


Figure 23. Log density of $X_{t+1} = \log(S_{t+1}/S_t)$ conditional on $V_t = 0$ for one- and two-factor continuous-time models with various settings for ρ_{21} . Other parameters calibrated based on estimates taken from Table 2. MIX3 shown for comparison.

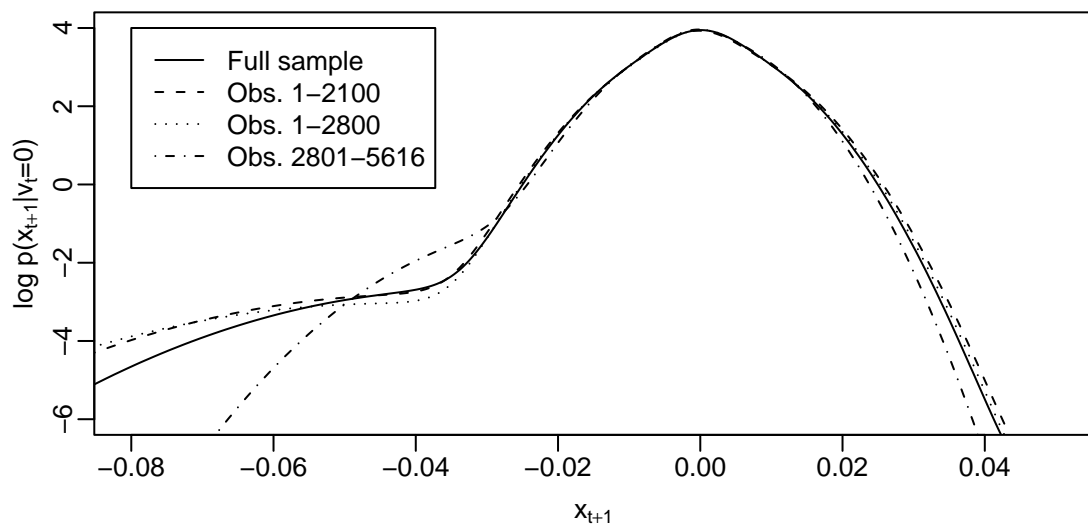


Figure 24. Conditional densities implied by MIX3 model estimated on various sub-samples of the data.

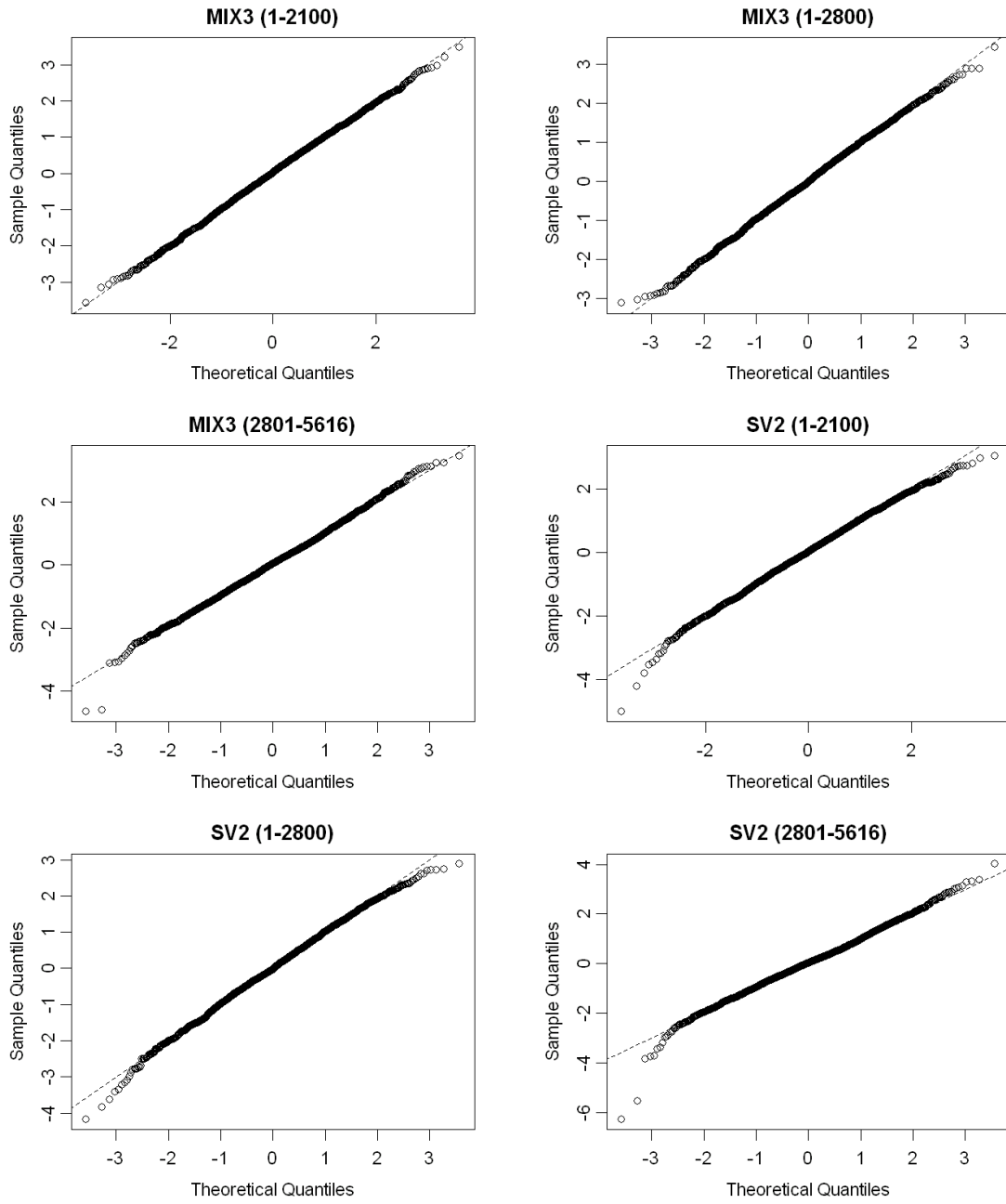


Figure 25. QQ-plots of out-of-sample generalized residuals against the standard normal for various sub-samples of the data.