To Reviewer #1:

On the efficiency/consistency of the parameter learning procedure. Please kindly see line 43-51.

How many probability levels used? All experiments use 99 levels of τ in the quantile regression: $\Psi=\{0.01,0.02,...,0.98,0.99\}$ so that the interval $(0,1)$ is covered sufficiently. One can use a smaller set to reduce the computing cost, e.g., $\Psi=\{0.01,0.05,0.1,...,0.9,0.95,0.99\}$. We find no significant performance difference using either the set of 99 levels or the 21 levels because the degree of freedom of the 4-parameter quantile function we use is 4. Both are sufficient.

The error compounding issue. The fitting of our one-factor model doesn't rely on variable ordering. Its estimation is quite reliable and efficient (line 43-50 has more details). For the lower-triangular model, more data would reduce the potential error compounding in y_n as n becomes large. We are trying to solve this issue by fine-tuning all parameters after the current learning procedure, with an overall objective instead of the step-by-step setting.

The constant A in Eqn.2 and Eqn.3. We require A be bounded below by a real constant such that Eqn.2 &3 are strictly increasing, which ensures Eqn.2 & 3 are indeed quantile functions. The constant A determines the threshold starting from which our Q-Q plot starts to curve or say, be notably different from that of a Gaussian distribution.

Why are $u \geq 1$ and $v \geq 1$ required in Eqn.3? If $u < 1$, then $u^x = (1/u)^{-x}$ and $1/u > 1$ will play the same role as v does, which is redundant. The same reason applies to v. We require the Q-Q plot exhibits both up and down tails.

The ideal number of violations. For the i.i.d. bernoulli distributed violation sequence with parameter τ (e.g., 0.01), the ideal number of violations is the number of observations N times τ. In Table 1, $N = 2075$ or 2407, $\tau = 0.01$.

Cross-sectional variance explained. Our model is designed to capture tail dependence, which measures joint extreme events. They happened very rarely in markets, thus capturing them (fourth-order moment) will contribute little to the total variance explained (second-order moment). We have checked that our model does very slightly better than traditional one-factor statistical models in variance explained. Table 2 gives similar β_i as CAPM, but with left/right tail sensitivity u_i^M and v_i^M added. The residual $\gamma_i g(z_i|u_i, v_i)$ is also heavy-tailed and asymmetric (described by u_i and v_i).

To Reviewer #2:

The advantage of Eqn.3 compared to Eqn.2. We replace Eqn.2 by Eqn.3 for two reasons. First, in terms of controlling the shape of the left & right tail, the cross term $e^{(u-v)Z_\tau}/A^2$ in Eqn.2 is redundant and not "clean". The additive form of Eqn.3 avoids these undesirable properties. Second, changing from e^{ux} to u^x reduces the sensitivity of tail heaviness to u while still allowing a wide range of tail heaviness. We found the u^x form suits the experiments in Table 2 better.

Does v_{ik}, $k < j$ have similar interpretation as v_{ij}? Yes. For example, v_{31} and v_{32} both contribute to the tail dependence between y_3 and y_2. But from the fitting procedure, v_{31} mainly determines tail dependence between y_3 and y_1, we believe v_{32} is the most free parameter that determines tail dependence between y_3 and y_2. The full relationship between tail dependence and parameters is complicated. But in the one-factor model, it is clear and easy to interpret.

How was Ψ chosen in the experiments? please kindly see line 3-5.

How to solve Eqn.13? The left side of Eqn.13 is an increasing function of τ^*. We use the bisection method on 10^7 samples of (X, Y) to solve it. The computing cost is acceptable. We will introduce these in the next version.

Could we also model negative tail dependence? We appreciate the reviewer makes this point. We neglected to discuss the negative tail dependence $\lim_{t\to 0^+} \mathbb{P}\{X<Q_X(t), Y>Q_Y(1-t)\}/t$. Actually, our model does cover this case because it is symmetric, i.e., when $\sigma_{ij} < 0$, $i > j$, what we are modeling is exactly the negative correlation as well as negative tail dependence between y_i and y_j. Now u_{ij} and v_{ij} are interpreted as the parameters controlling negative tail dependence.

On the theoretical analysis of our model. We share the same view that theoretical analysis of either the model or the estimator is important. We are actively developing it. For the consistency/efficiency matter of parameter learning, please kindly see 43-50.

To Reviewer #3:

On the consistency/efficiency of the parameter learning procedure. Developing a theory showing these properties is not easy, given the complexity of our learning procedure. But at this moment we can address this problem numerically. We randomly set the model parameters, simulate some points, and apply the learning procedure on them. The learned parameters of our one-factor model are shown in the following table (i is any one in $1, 2, ..., n$). We also list the number of data points N and the norm of learned parameters minus true ones. We can see the learning for one-factor model

Trial\Parameter	α_M	β_M	u_M	v_M	β_i	u_i^M	v_i^M	α_i	γ_i	u_i	v_i	N	Norm
True Parameters	-0.33	0.60	2.08	2.34	-0.71	1.84	1.83	0.88	0.18	2.37	2.12	—	0.00
1	-0.33	0.56	2.36	2.67	-0.64	1.75	2.57	0.82	0.23	1.84	2.26	1000	1.02
2	-0.34	0.60	2.12	2.40	-0.71	1.88	1.85	0.90	0.19	2.35	2.07	10000	0.10
3	-0.33	0.60	2.10	2.36	-0.71	1.85	1.85	0.88	0.19	2.36	2.10	100000	0.04

is fairly consistent and efficient as N increases. For our lower-triangular model, the learning is less efficient but still consistent enough. The convergence becomes slower and we need more data points to obtain a reliable estimate. This is due to the more complex structure and more parameters of the model. For space limit we do not show its results here.