CFM Challenge

Hamlet J. Medina Ruiz

2019/06/25

Data set

US Market data

Training set: 636313 observations

Testing set: 635397 observations

Per observation:

- date
- product_id
- time series information from 9:30 to 13:55.

Target: average volatility between 14:00 and 16:00

Evaluation metric: MAPE

Data set

TS features: 54 5min time-slots from 9:30 to 13:55

- volatility
- return signs

In a given day:

- between 229 to 318 stocks
- with approx. 300 on average

110 features

Challenges

Missing values (suspended trading)

No order in days.

Short time series (54 samples).

No day's overlap between train and test sets

Basic feature engineering.

- afternoon_vol
- mean_afternoon_vol
- daily_vol
- mean_daily_vol
- missing_features ratio.
- min_daily_vol
- max_daily_vol
- log transformation

Missing features:

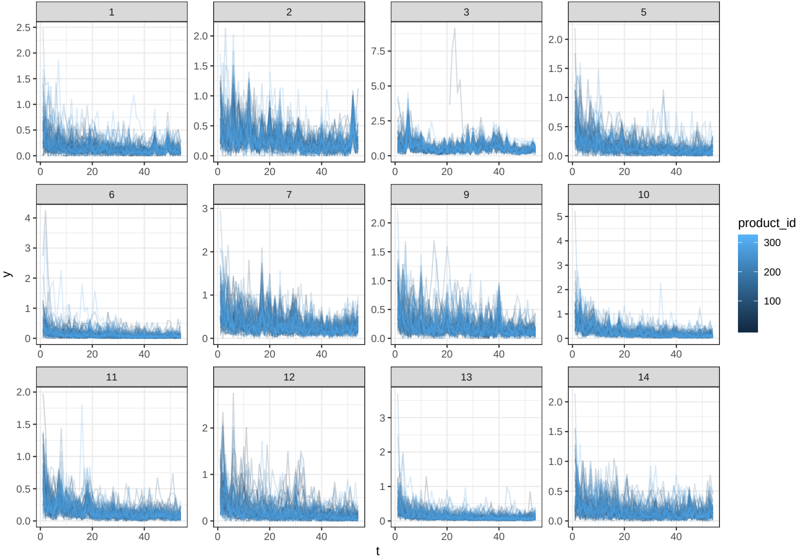
- linear interpolation or daily_vol
- indicator variable

Modelling approaches

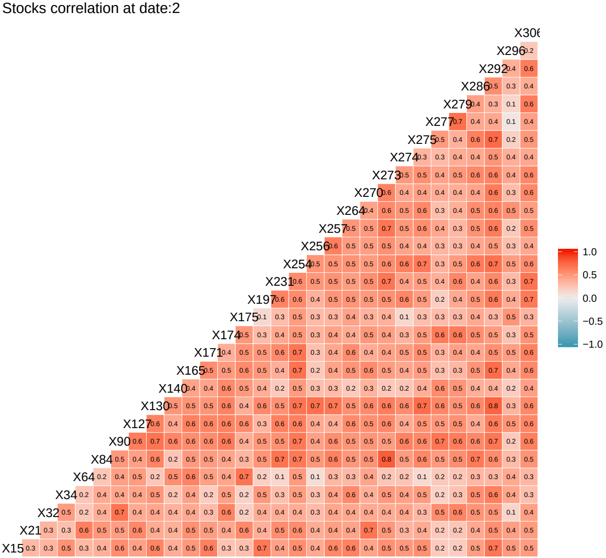
Focus:

- Modelling in a latent space
- Marginal and conditional independence statements

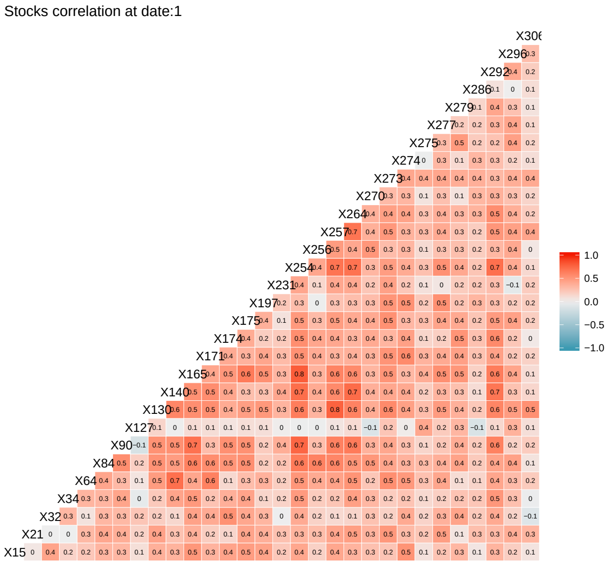
TS evolution



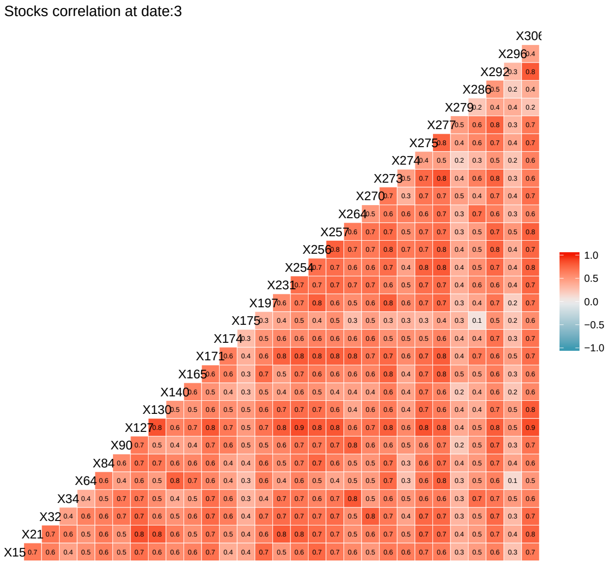
TS correlation.



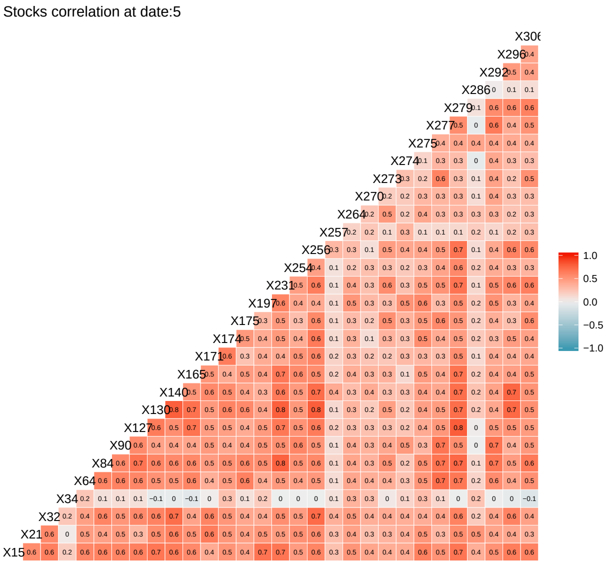
TS correlation.



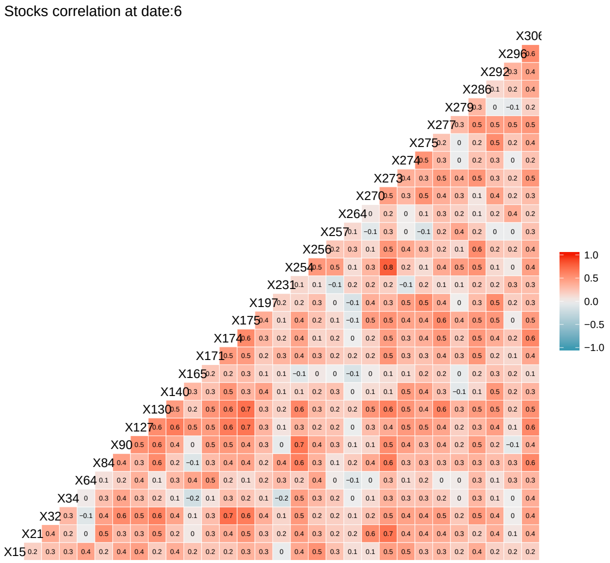
TS correlation.



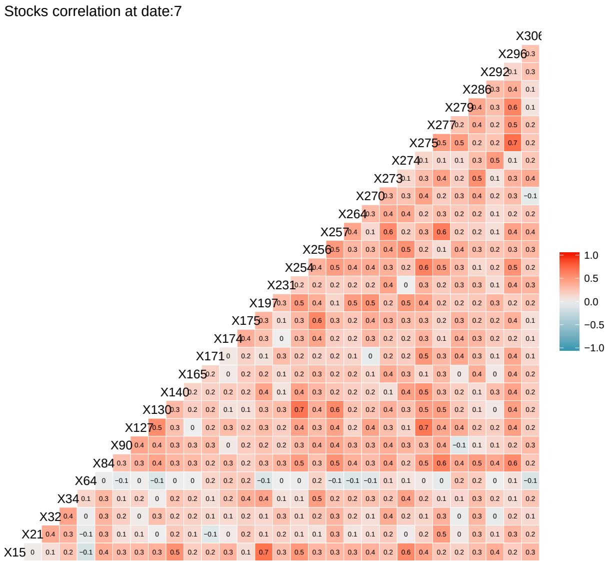
TS correlation.



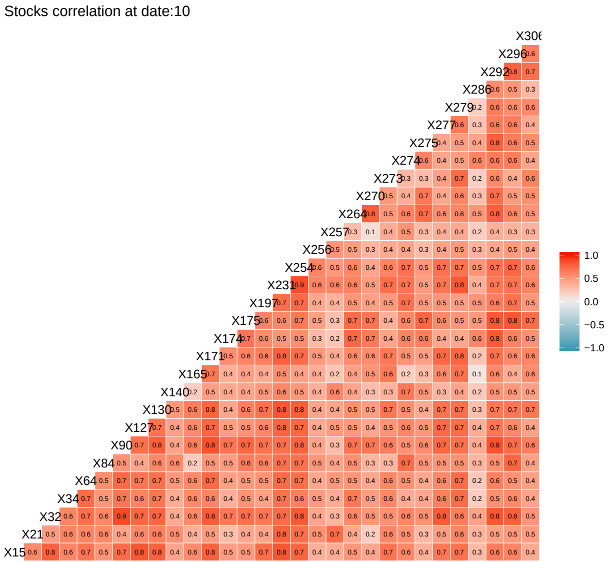
TS correlation.



TS correlation.



TS correlation.



Approach 1: Latent factor models

Embedding observations in a lower dimensional subspace

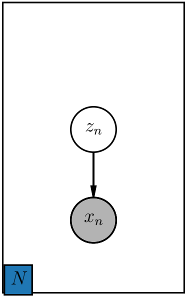


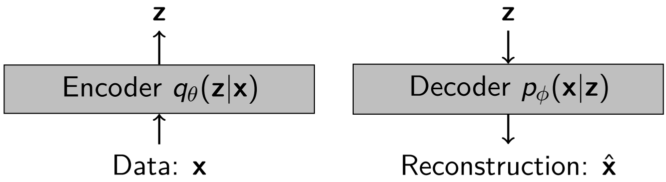
Figure 1: probabilistic view

$$x_n | z_n \sim \mathcal{N}(\Lambda z_n + \mu, \Sigma_0)$$

$$\Sigma_x = \Lambda\Lambda^T + \Sigma_0$$

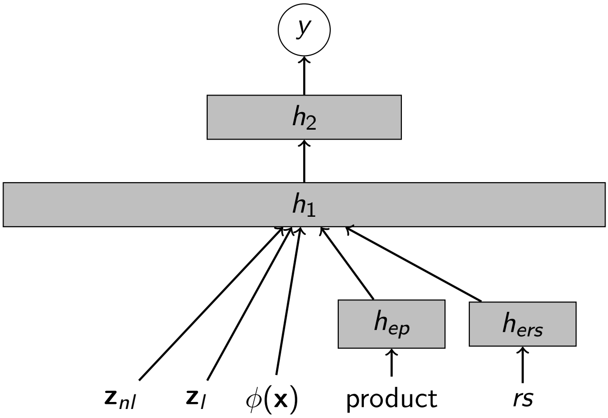
Approach 1: Latent factor models

Based on NN: auto-encoders



- tanh activations
- L1 + L2 Regularization

Model 1.



- $h = [500, 200]$ units
- $h_{ep} = 50$ units, $h_{ers} = 10$ units
- ReLU + Dropout

Score: 21.38

Approach 2.

Date is key

How to extract signal about dates and stocks?

Build model in a similar task.

Approach 2.

	s_1	s_2	s_3	$\cdots$	s_S
d_1	■	■			■
d_2	■		■		■
d_3		■	■		■
$\vdots$					
d_T	■	■	■		

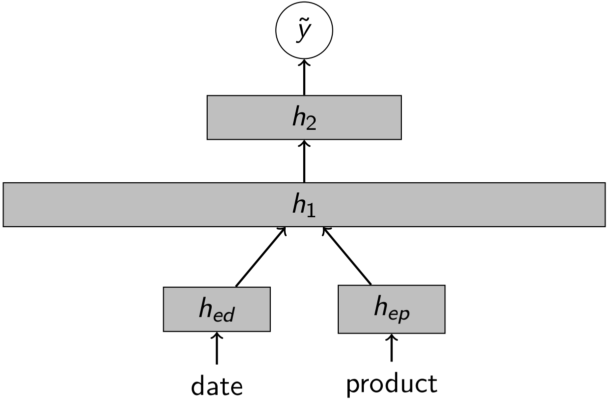
Approach 2.

	s_1	s_2	s_3	$\cdots$	s_S
d_1	■	■			■
d_2	■		■		■
d_3		■	■		■
$\vdots$					
d_T	■	■	■		

Standard technique in recommender systems (SVD)

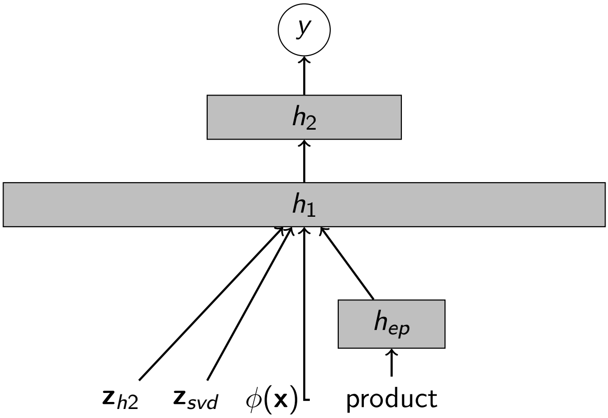
Approach 2.

NonLinear matrix completion based on NN



- $h = [200, 100]$ units
- $h_{ep} = h_{ed} = 50$ units
- ReLU + Dropout

Model 2.



- $h = [500, 200]$ units
- $h_{ep} = 50$ units
- ReLU + Dropout

Score: 21.30

Approach 3.

Modelling across stocks (strong hypothesis)

Hypothesis:

Given the market -> conditional independent statements.

How to estimate "the market" volatility $\psi(t)$?

Approach 3.

Build a naive mean-market per day j

$$\psi^{(j)}(t) = \frac{1}{N}\sum_s x_s^{(j)}(t)$$

- Build $\hat{\psi}^{(j)}$ for the U-shape volatility profile.
- Forecast $\psi^{(j)}(t+h)\ \forall h \in [14{:}00, 16{:}00]$
- Condition on $\hat{\psi}^{(j)}(t+h)$
- Build standard conditional Gaussian Linear models

Approach 3.

Model:

- Basic features
- Market predictions

Linear Regression + L1 Regularisation

Sampling weight $\propto \frac{1}{y}$

Score: 21.87

Approach 4.

Modelling across stocks

- Basic Features.
- **Linear and non Linear embeddings on afternoon Volatility**

Linear Regression + L1

Sampling weight $\propto \frac{1}{y}$

Score: 21.71

Model 5

- Basic Features.
- U shape market_features
- Linear/Non Linear embeddings on afternoon Volatility.
- **Predictions from linear model**
- **Clustering on the latent space**
- **Clustering using the Power Spectral Density**

Random Forest: 300 trees, min samples leaf $= 50$

Sampling weight $\propto \frac{1}{y}$

Score: 21.61

Final Solution.

Final solution: ensembling models

	PublicScore	PrivateScore
MAPE (%)	21.0174	20.9396

0.2077% from 1st place.

Technical details

- Optimiser: Adam.
- Cosine annealing with restarts.
- Batch size: 512
- Dropout.
- Log transformation.
- Approx 15min to train NN models

Conclusion

No order in days makes the problem harder

Volatility is clustered across-day

Some perspectives:

- Dynamic latent factors: time varying covariance
- Switching Dynamic Linear/NonLinear Models
- Hierarchical Bayesian time series models

References.

- Graphical Models for Time-Series, David Barber, A. Taylan Cemgil
- Nonlinear Time Series: Theory, Methods and Applications with R Examples, Eric Moulines, David S Stoffer, Randal Douc.
- Bayesian Nonparametric Inference of Switching Dynamic Linear Models, Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, & Alan S. Willsky
- Pattern Recognition And Machine Learning by Christopher M. Bishop.
- Machine learning a probabilistic perspective by Kevin P. Murphy.