BIG DATA'S DIRTY SECRET

Harvey J. Stein

Head, Quantitative Risk Analytics
Bloomberg

Joint work with Yan Zhang

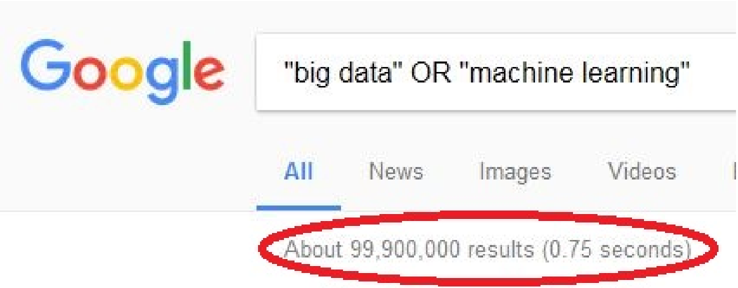
The 2nd Machine Learning & AI in Quantitative Finance Conference USA
November 13, 2018

OUTLINE

Introduction

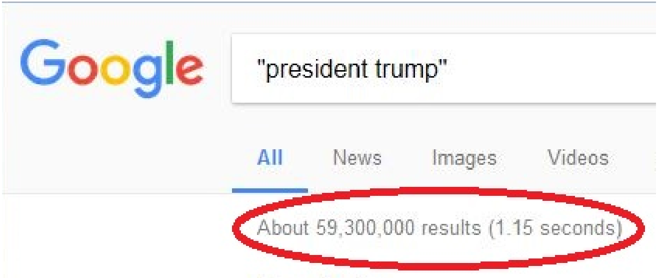
LOTS OF BIG DATA

Big data is big news!



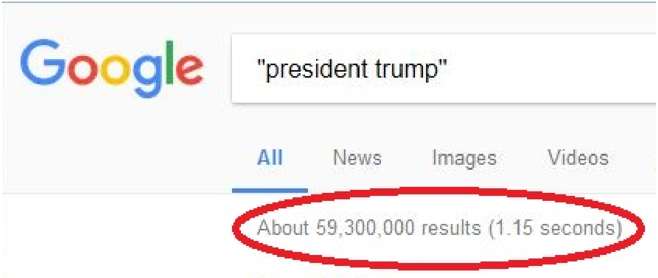
TRUMPS TRUMP

Almost twice as popular as "President Trump"!



TRUMPS TRUMP

Almost twice as popular as "President Trump"!



Although I guess that's not so surprising...

FAKE NEWS

But big data analysis doesn't mean better data analysis

- More variables
- More outliers
- More noise
- More spurious results

Conclusion?

- Data needs to be cleaned

We will discuss data anomalies and methods for cleaning data

ACKNOWLEDGEMENTS

Joint work with Yan Zhang

Additional contributors:

- Mario Bondioli
- Jan Dash
- Xipei Yang

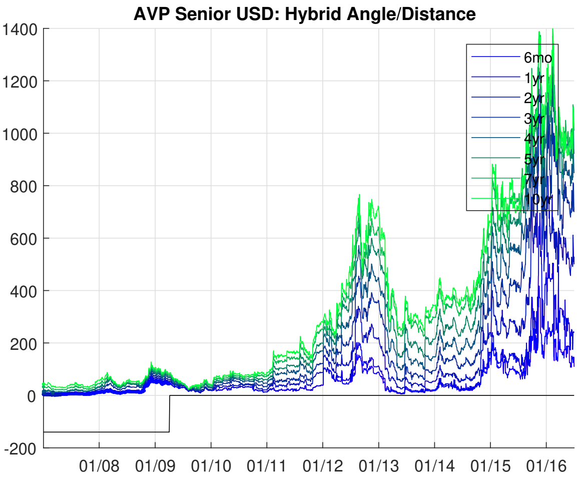
Symptoms

THE DATA

We worked with credit default swap (CDS) spread data

- Spread = cost (in bp) of insuring against default of a given company for a given time period
- Quoted for 6 month, 1 year, 2 year, 3 year, 5 year, 7 year and 10 year horizons
- Quoted for 1,000s of different individual companies
- Quoted both for senior and subordinated debt
- Consider market close data

EXAMPLE



AVP Senior USD: Hybrid Angle/Distance
1400
1200
1000
800
600
400
200
0
-200
01/08
01/09
01/10
01/11
01/12
01/13
01/14
01/15
01/16
6mo
1yr
2yr
3yr
4yr
5yr
7yr
10yr

DATA ISSUES

General data quality issues

- Missing values
- Bad values

Clean for a purpose

- Relative valuation
- Mark to market
- Trading strategy development
- Risk analysis

Risk

- Missing data points
  - Problematic return calculations
  - Problematic covariance calculations
- Bad values
  - Bad returns
  - Bad variances

CDS DATA ISSUES

CDS data specific characteristics:

- 6 month point missing for first 2.5 years
- Often large range of values
- High volatility makes detecting bad values difficult
- Data used for risk analysis
  - Deleting outliers reduces risk measures
  - Leaving anomalies inflates risk measures

TYPICAL APPROACHES

Hole filling

- Regression
- Interpolation
- Flat filling

Anomaly detection

- Comparison to trailing volatility
- Cluster analysis
- Neural networks
- Statistics-sensitive Non-linear Iterative Peak (SNIP) clipping algorithm

Hole filling

OVERVIEW

Hole filling Overview

- Use Multi-channel Singular Spectrum Analysis (MSSA) hole filling algorithm
  - Variant of Singular Spectrum Analysis (SSA) used simultaneously on multiple time series
  - Decomposes each time series into a sum of components, one for each principal component
- Borrowed from geophysical data analysis
- Makes use of both space relationships (covariance) and time relationships (autocovariance and cross-autocovariance)
  - Eigenvector decomposition of the auto-cross covariance matrix

SSA

Uses:

- Inspect eigenvectors and components to extract specific features of data
- Smooth data by throwing away small eigenvalues
- Helpful for stabilizing correlation calculations (smooth data then compute)

References:

- **A beginner's guide to SSA**, Claessen and Groth, [CG]
- **Singular spectrum analysis**, Wikipedia, [Wik16]
- **Analysis of Time Series Structure: SSA and Related Techniques**, Golyandina, Nekrutkin, and Zhigljavsky, [GNZ01]
- **A review on singular spectrum analysis for economic and financial time series**, Hassani and Thomakos, [HT10]
- **SSA, Random Matrix Theory, and Noise-Reduced Correlations**, Dash et al., [Das+16a]
- **Stable Reduced-Noise 'Macro' SSA–Based Correlations for Long-Term Counterparty Risk Management**, Dash et al., [Das+16b]

MSSA

Multi-channel Singular Spectrum Analysis (MSSA):

- Applies SSA algorithm to a set of time series simultaneously

Uses:

- Same as SSA, but takes relationships between different time series into account
- Used for forecasting

References:

- **Multivariate singular spectrum analysis for forecasting revisions to real-time data**, Patterson et al., [Pat+11]
- **Multivariate singular spectrum analysis: A general view and new vector forecasting approach**, Hassani and Mahmoudvand, [HM13]
- **Advanced spectral methods for climatic time series**, Ghil et al., [Ghi+02]

MSSA BASED HOLE FILLING

MSSA hole filling algorithm:

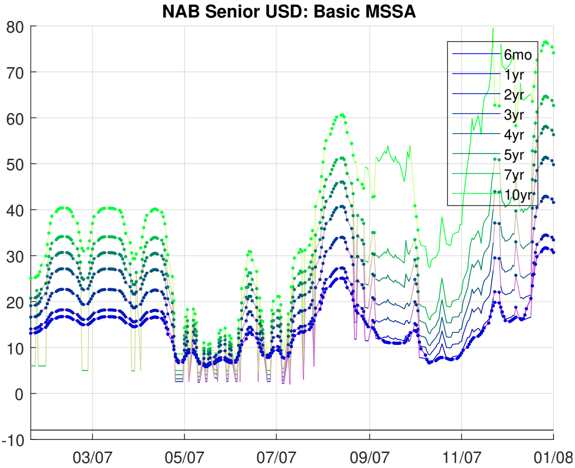
- Nominally fill holes (e.g. via interpolation):
- Iteratively refine hole filling approximation
  - Run MSSA algorithm
  - Replace holes with MSSA reconstruction using l biggest singular values
  - Repeat until convergence
- Increment l by one and repeat until adding singular values doesn't have much impact and used enough singular values

References:

- **Spatio-temporal filling of missing points in geophysical data sets**, Kondrashov and Ghil, [KG06]

MIXED RESULTS

Unfortunately, it doesn't always work:



OBSERVATIONS

Observations:

- Sometimes MSSA doesn't line up with actual data
- Sometimes MSSA bottoms out
- Using too few singular values will smooth the data

Solutions:

- Anchoring – patch in data in a more consistent fashion
- Reparameterization – working in log space
- Adjusting MSSA parameters
- Avoid filling large gaps

ANCHORING

Holes are replaced with MSSA partial reconstruction

- Can yield bias if remaining components shift results

Instead

- Patch in differences relative to endpoints
- Can be additive or multiplicative
- One-sided holes need special treatment

REPARAMETERIZATION

MSSA hole filling is like a fixed point algorithm

- Trying to find points which match reconstruction
- Similar to constrained optimization

Apply classic optimization techniques

- Transform problem to eliminate constraints
- Work in log space if values must be positive
- Log space also helps to handle changes in magnitude

Fast drop-off of eigenvalues is evidence that working in log space is the right thing

ADJUSTING MSSA PARAMETERS

Many parameters to adjust

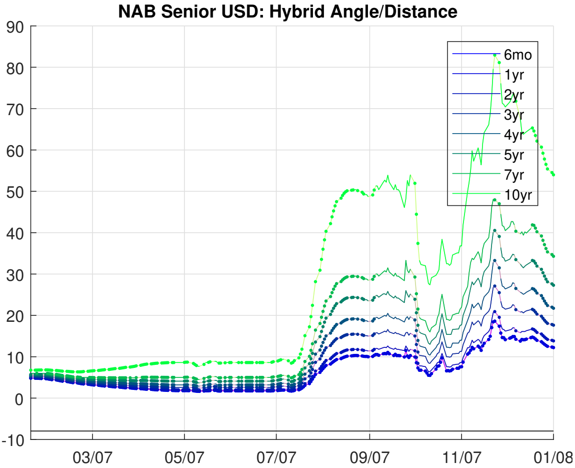
- Lag
- Max/Min number of EVs
- Max/Min percentage of sum of EVs
- Measure of convergence

Smoothing caused by fast drop-off of EVs

- Max/Min percentage ineffective
- Can add more EVs, but leads to instability

NEW RESULTS

After adjustments NAB:



Bad data detection

BAD DATA

How to handle bad data?

- Detect it
- Remove it
- In our case, replace it

BAD DATA DETECTION

Many algorithms

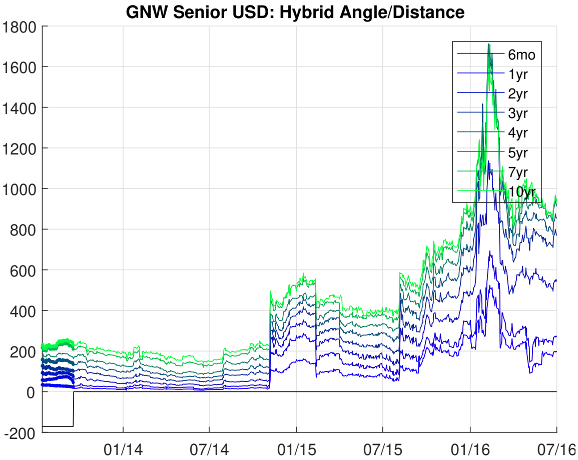
- Statistical – compare to statistical properties (like trailing SD)
- Data science – clustering
- Neural networks

References

- **Outlier Detection Techniques**, Kriegel, Kröger, and Zimek, [KKZ10]
- **Detecting Local Outliers in Financial Time Series**, Verhoevena and McAleer, [VM]
- **Outlier analysis**, Aggarwal, [Agg13]
- **Algorithms for Mining Distance-Based Outliers in Large Datasets**, Knorr and Ng, [KN98]
- **Outlier detection**, Ben-Gal, [BG05]
- **An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes**, Franke et al., [Fra+10]
- **A survey of outlier detection methodologies**, Hodge and Austin, [HA04]

DIFFICULTIES

Regime changes and changing volatility



HYBRID APPROACH

Data science approach – Cluster analysis

- Angle-based
- Distance-based

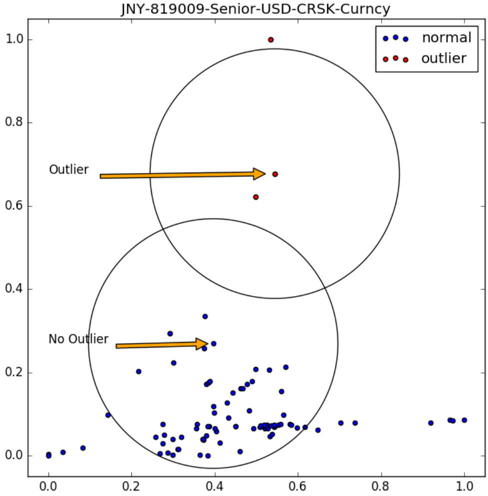
Hybrid approach

- Run clustering on a windowed basis (in a neighborhood of each point)
- Combine MSSA with clustering
- Remove points using analysis, then put them back if MSSA reconstructs them close enough

Conservative approach

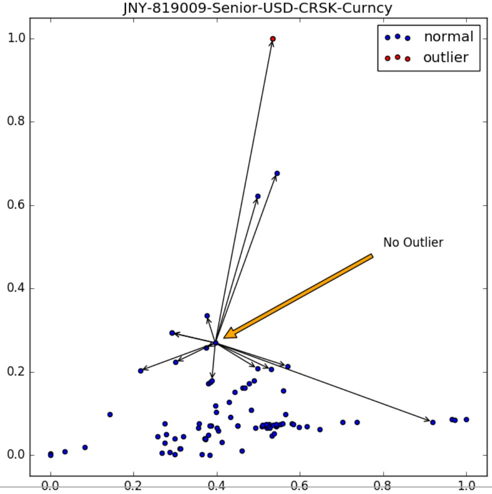
- Do both angle and distance-based combined with MSSA
- If both algorithms agree, then it's really an anomaly

DISTANCE-BASED EXAMPLE



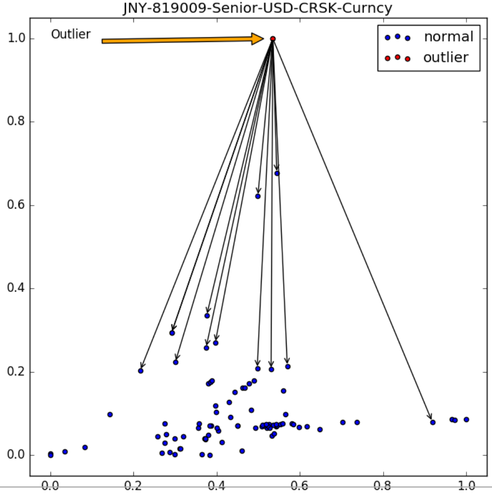
ANGLE-BASED EXAMPLE

Angle-based, no outlier:



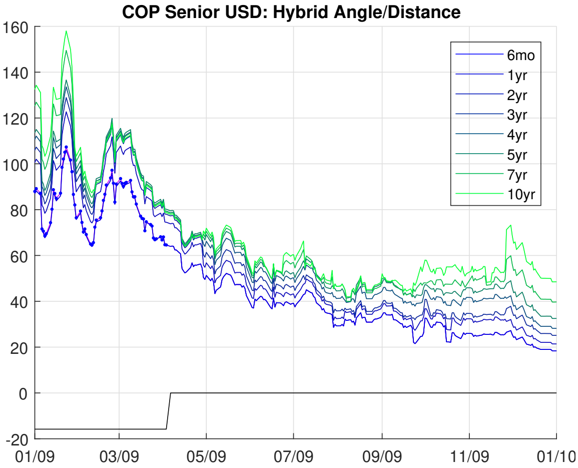
ANGLE-BASED EXAMPLE

Angle-based outlier:



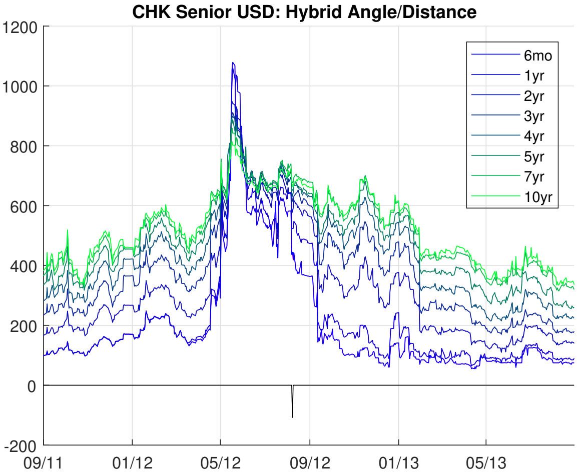
RESULTS

Filling of large holes



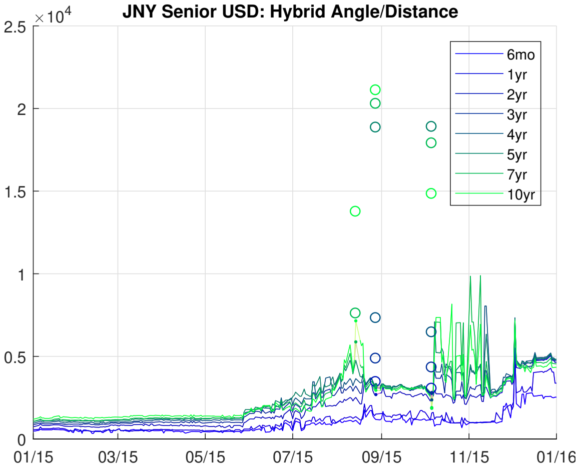
RESULTS

Ignoring regime changes



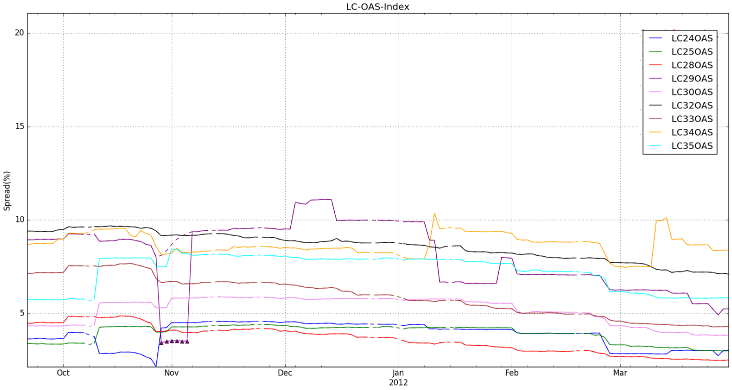
RESULTS

Detecting and correcting bad data



RESULTS

Even works on CMO OASs!



Summary

SUMMARY

Moral of the story

1. Know your data!
   - Bad data = bad results
   - Big data increases need for data cleaning
   - Look at your data!
2. Know its usage!
   - Cleaning must respect usage of data
3. Algorithms will often not work as advertised!
   - Your data can be different
   - Your data usage can be different
4. Expect substantial work modifying and adjusting algorithms
   - Tuning
   - Modifying algorithms
   - Combining algorithms
   - Performance must be inspected

Thank you!

Harvey J. Stein

`hjstein@bloomberg.net`

References

REFERENCES

[Agg13] Charu C. Aggarwal. *Outlier analysis*. Springer, 2013.

[BG05] Irad Ben-Gal. "Outlier detection." In: *Data mining and knowledge discovery handbook*. Ed. by Oded Maimon and Rokach Lior. Springer, 2005, pp. 131–146.

[CG] David Claessen and Andreas Groth. *A beginner's guide to SSA*. CERES-ERTI, Ecole Normale Supérieure. URL: http://environnement.ens.fr/IMG/file/DavidPDF/SSA_beginners_guide_v9.pdf.

[Das+16a] Jan W. Dash et al. *SSA, Random Matrix Theory, and Noise-Reduced Correlations*. Tech. rep. Bloomberg LP, Sept. 2016. URL: https://ssrn.com/abstract=2808027.

REFERENCES

[Das+16b] Jan W. Dash et al. *Stable Reduced-Noise 'Macro' SSA–Based Correlations for Long-Term Counterparty Risk Management*. Tech. rep. Bloomberg LP, May 2016. URL: https://ssrn.com/abstract=2808015.

[Fra+10] Felix Franke et al. "An online spike detection and spike classification algorithm capable of instantaneous resolution of overlapping spikes." In: *Journal of computational neuroscience* 29.1-2 (2010), pp. 127–148.

[Ghi+02] M. Ghil et al. "Advanced spectral methods for climatic time series." In: *Reviews of Geophysics* 40.1 (2002).

[GNZ01] Nina Golyandina, Vladimir Nekrutkin, and Anatoly A. Zhigljavsky. *Analysis of Time Series Structure: SSA and Related Techniques*. CRC Monographs on Statistics & Applied Probability. Chapman & Hall, 2001.

REFERENCES

[HA04] Victoria Hodge and Jim Austin. "A survey of outlier detection methodologies." In: *Artificial intelligence review* 22.2 (2004), pp. 85–126.

[HM13] Hossein Hassani and Rahim Mahmoudvand. "Multivariate singular spectrum analysis: A general view and new vector forecasting approach." In: *International Journal of Energy and Statistics* 1.01 (2013), pp. 55–83.

[HT10] Hossein Hassani and Dimitrios Thomakos. "A review on singular spectrum analysis for economic and financial time series." In: *Statistics and Its Interface* 3.3 (2010), pp. 377–397. ISSN: 1938-7989.

[KG06] Dmitri Kondrashov and Michael Ghil. "Spatio-temporal filling of missing points in geophysical data sets." In: *Nonlinear Processes in Geophysics* 13.2 (2006), pp. 151–159.

REFERENCES

[KKZ10] Hans-Peter Kriegel, Peer Kröger, and Arthur Zimek. "Outlier Detection Techniques." In: The 2010 SIAM International Conference on Data Mining, 2010. URL: http://www.imada.sdu.dk/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf.

[KN98] Edwin M. Knorr and Raymond T. Ng. "Algorithms for Mining Distance-Based Outliers in Large Datasets." In: Proceedings of the 24th VLDB Conference New York, 1998.

[Pat+11] Kerry Patterson et al. "Multivariate singular spectrum analysis for forecasting revisions to real-time data." In: *Journal of Applied Statistics* 38.10 (2011), pp. 2183–2211.

[VM] Peter Verhoevena and Michael McAleer. *Detecting Local Outliers in Financial Time Series*. Department of Economics, University of Western Australia.

REFERENCES

[Wik16] Wikipedia. *Singular spectrum analysis*. [Online; accessed 9-December-2016]. 2016. URL: https://en.wikipedia.org/w/index.php?title=Singular_spectrum_analysis.