Opportunities and Challenges: Lessons from Analyzing Terabytes of Scanner Data

Serena Ng
Columbia University

World Congress 2015

Big Data Frenzy

- 2.7 zetabytes$= 2.7 \times 1024^3$ TB in the digital universe
- Obama's initiative, data centers, machine learning classes.
- Big data pre-Google (Census, PSID, Compustat, etc.).
- Big data post-Google:
  - Social media: job loss index, UNICEF project.
  - Transactions data: (Ebay, Spending pulse)
  - Web search data: Google flu trend, leading indicators.
  - Online prices: BPP, Premise.

Unintentionally collected but cheap.

Unstructured and can be constantly being revised.

3V and Implications

- Volume:
  - Finite sample issues: jacknife, bootstrap: Important?
  - Likelihood dominates: Role of prior?
  - Time intensive estimators: Feasible?
  - Optimal estimation/inference: Practical?
- Variety:
  - One DGP for all data: Realistic?
  - Aggregation: along which dimension?
- Velocity:
  - Noise. more data= more information?
  - Asymptotic framework?
  - Granger (1988): standard errors too small?

Statistics+ Comp. Sci+ Math= Data Science

- Statistics: model based, probabilistic, beautiful theory.
- Data science: no dgp, algorithms, downstream use.
  - Detail vs. Approximate analysis
  - Reasonably homogeneous vs. highly heterogeneous data
  - Likelihoods/moments vs. random forest/boosting trees.
- ASA, IMS: merging of statistics and computing.
- Reality check: I know
  - something about the many predictors problem.
  - some machine learning algorithms: eg. boosting
  - little about the computing part of data science.

Nielsen Scanner Data: 3.6TB

- $i = 1, \ldots, n$ = store-upc pair. $t = 1, \ldots T = 260$ weeks.
- Unique features:
  - variables: p_{ti} and q_{ti}: instead of $p_{ti}q_{ti}$ and p_{ti}.
  - frequency: weekly
  - spatial: many locations, products.
  - span: 2006-2010 (Great Recession)

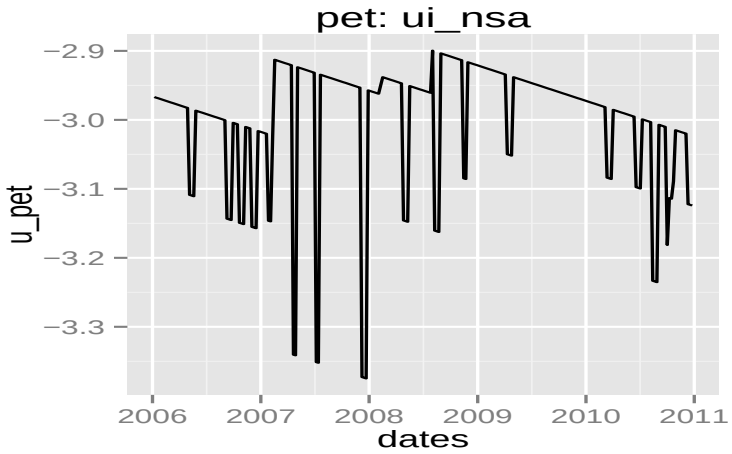
This study: extract cyclical component

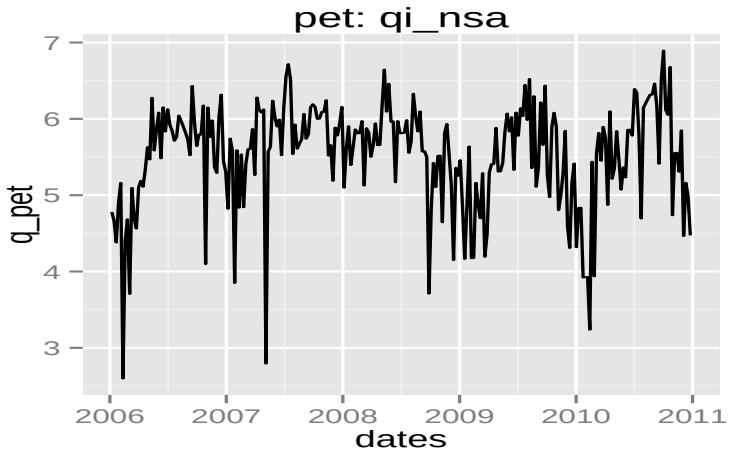
- q_{ti} from 10 product categories
- Weekly, no time aggregation.
- Data issues before economic analysis.
- BIG n, small K (predictors).

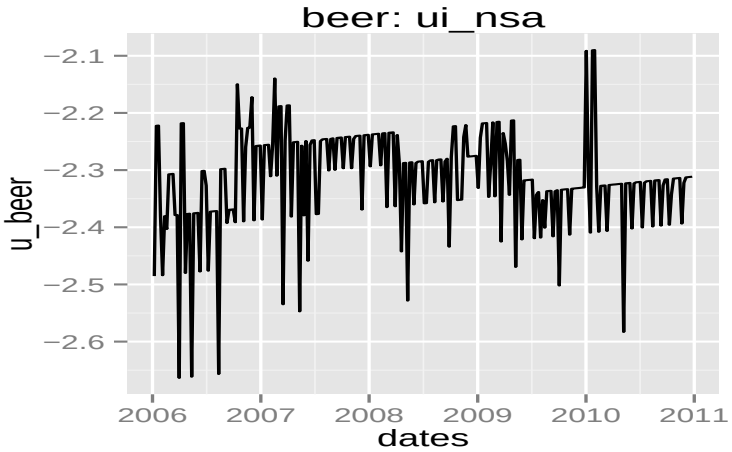
This Paper

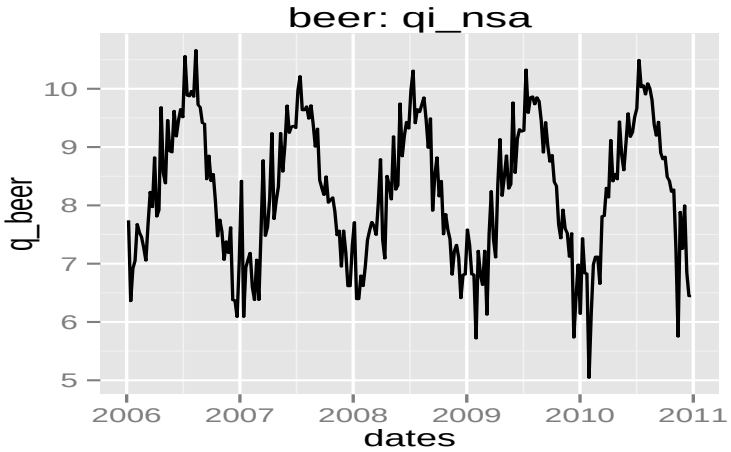
- What are the big data specific issues.
- Econometrics: big data friendly? (seasonal adjustment)
  - goal: cyclical component
  - need to remove seasonal variations for MANY series.
  - weekly data: not exactly periodic.
- Algorithms: useful? (subspace sampling)
  - Sub sampling for computationally efficient estimation.
  - Not subsampling for finite sample inference.

Why consider seasonal adjustment? pet food and beer

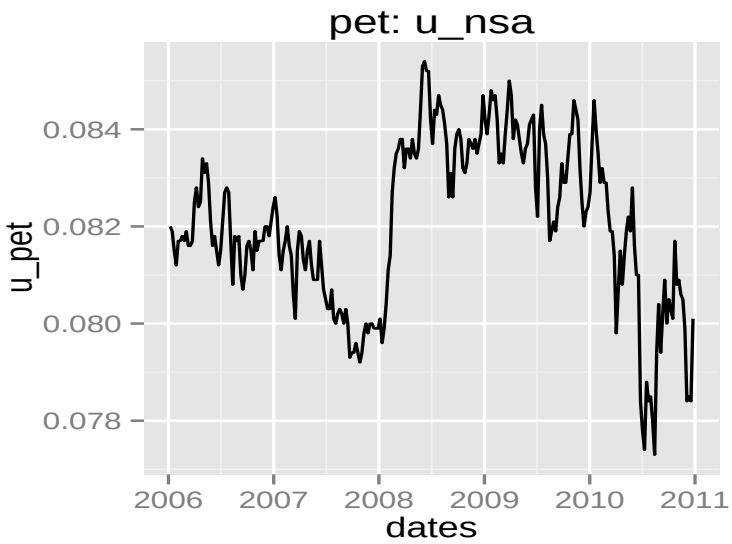


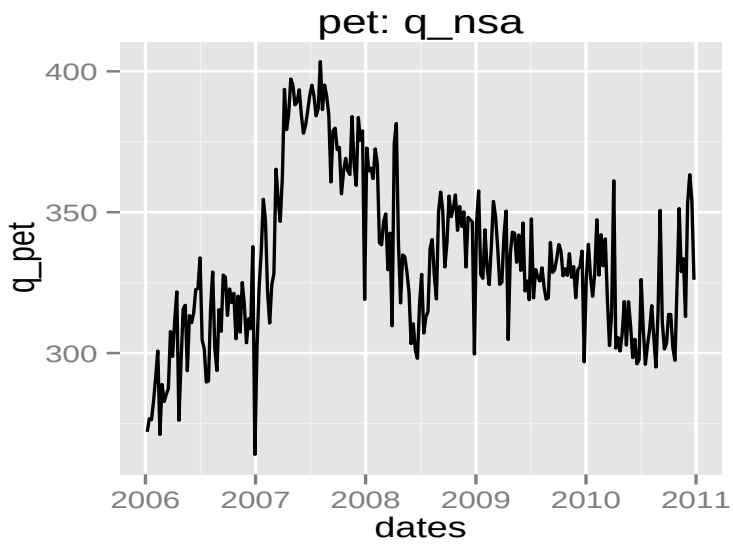


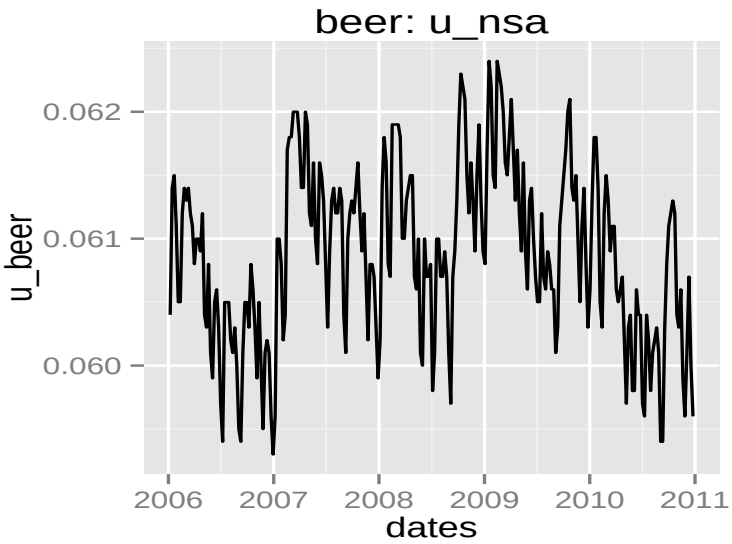


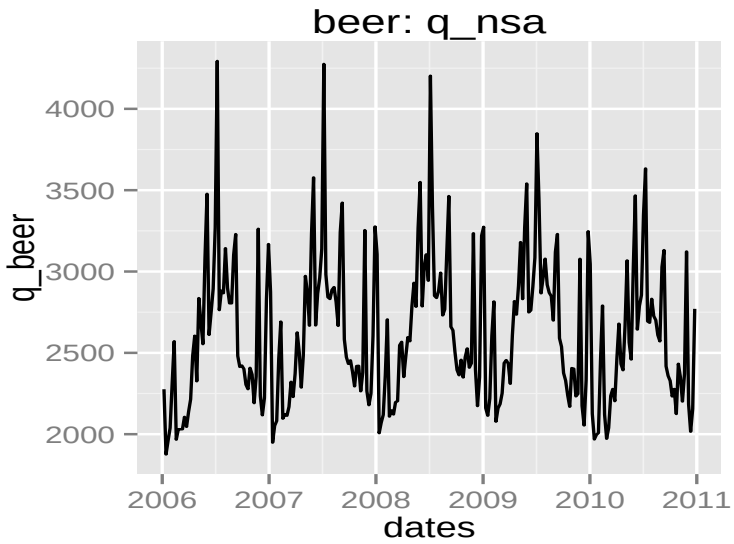


Aggregate data: pet food and beer









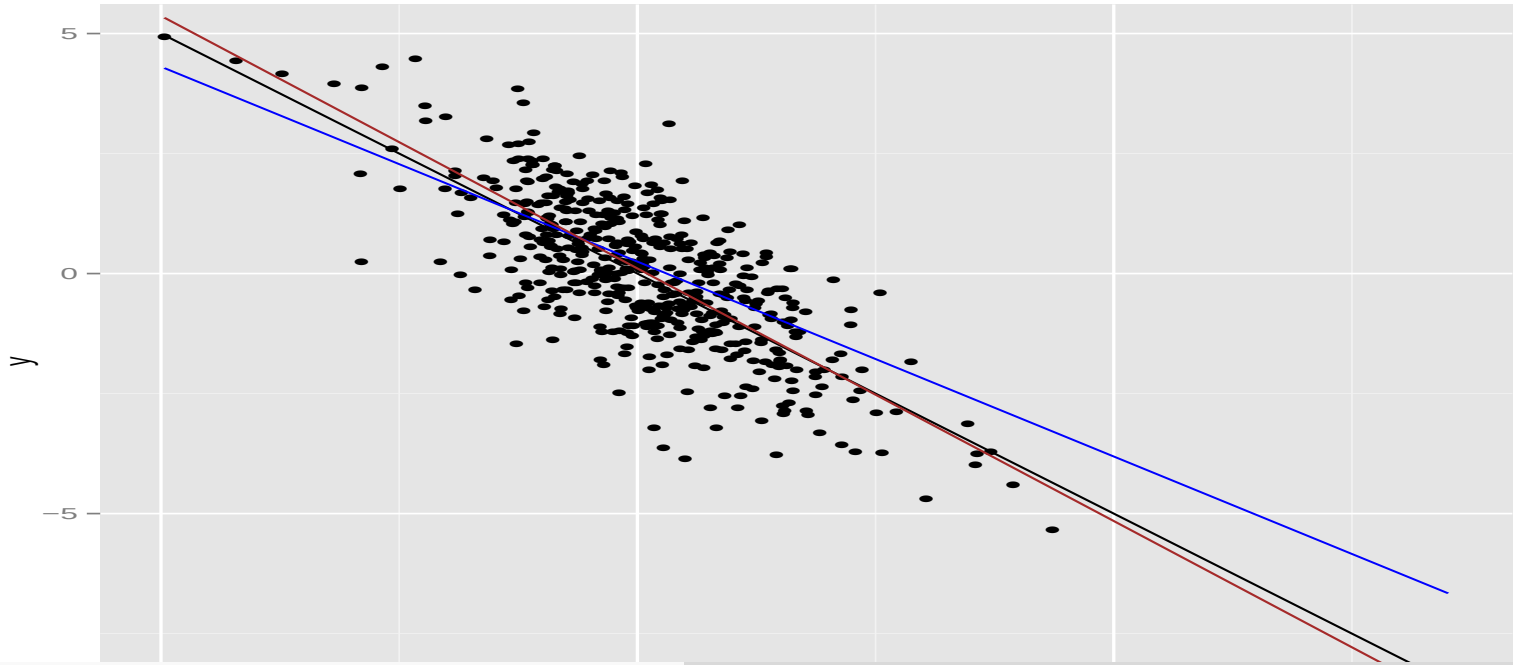
Why Subsampling?

- Too much data: beer data: 20.3 GB, 100E6 data points.
- Balanced panel, medium data, but hands are full!
  - balancing the panel: try different approaches/software!
  - painful to arrange cost-effective computing environment!
- RAM constraint, painfully slow I/O. Need to repeat.
- Do we really so many observations even if we can?
  - Deaton/Ng (JASA 1998): N=9119. Average derivatives.
  - Boivin/Ng (JOE 2006): factor models.

Some Alternatives?

- 'Data squashing':
  - group obs. with similar likelihood profiles.
  - Problem: Parametric analysis not easily upward scalable.

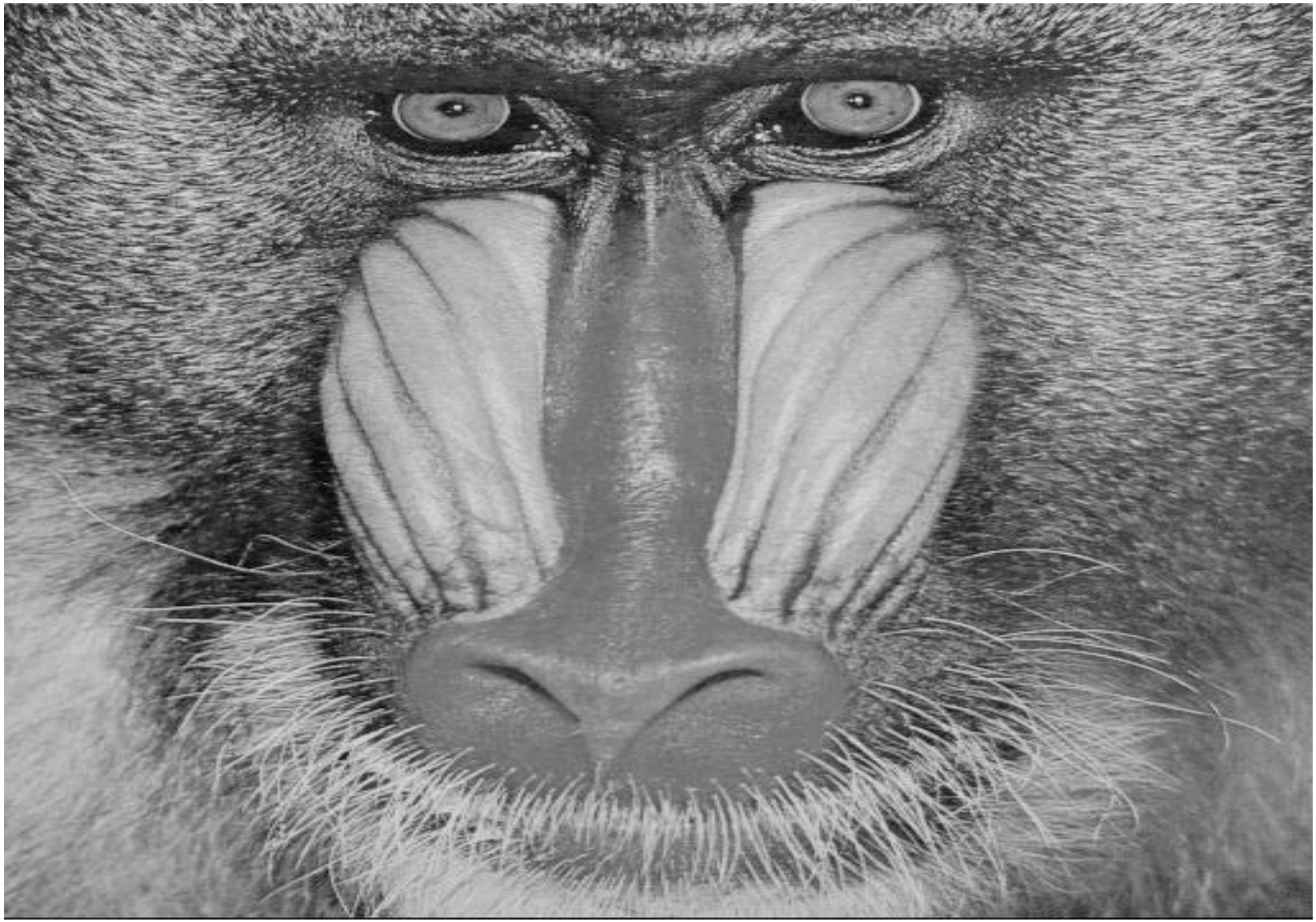
- Uniform sampling? bad if data have non-uniform features.



- Color image: $512 \times 512 \times 3$ matrix of density of RGB pixels.

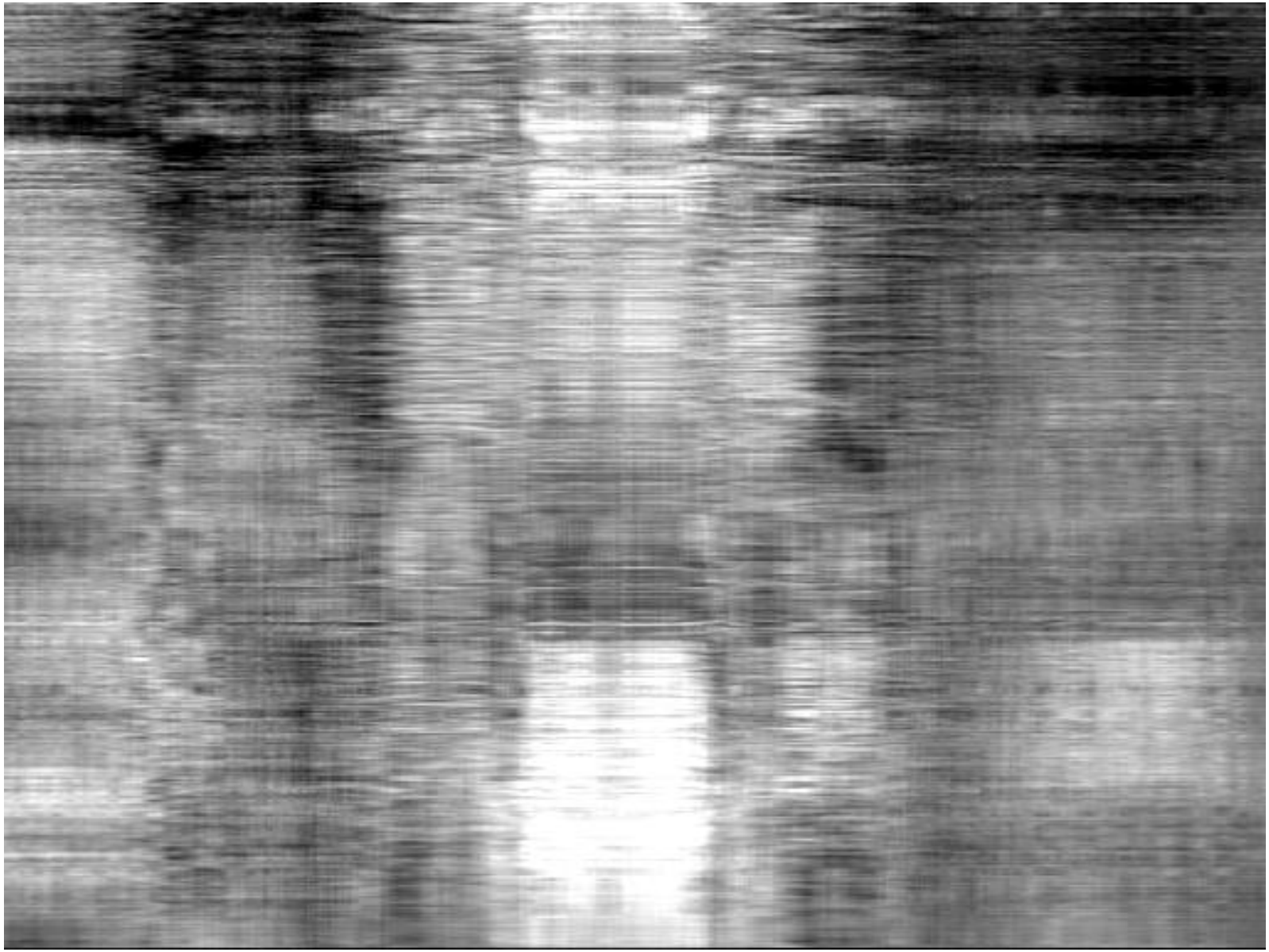
MATLAB: C=im2double(imread('baboon.jpg'));

- BW image: 512×512 matrix storing grey scales.

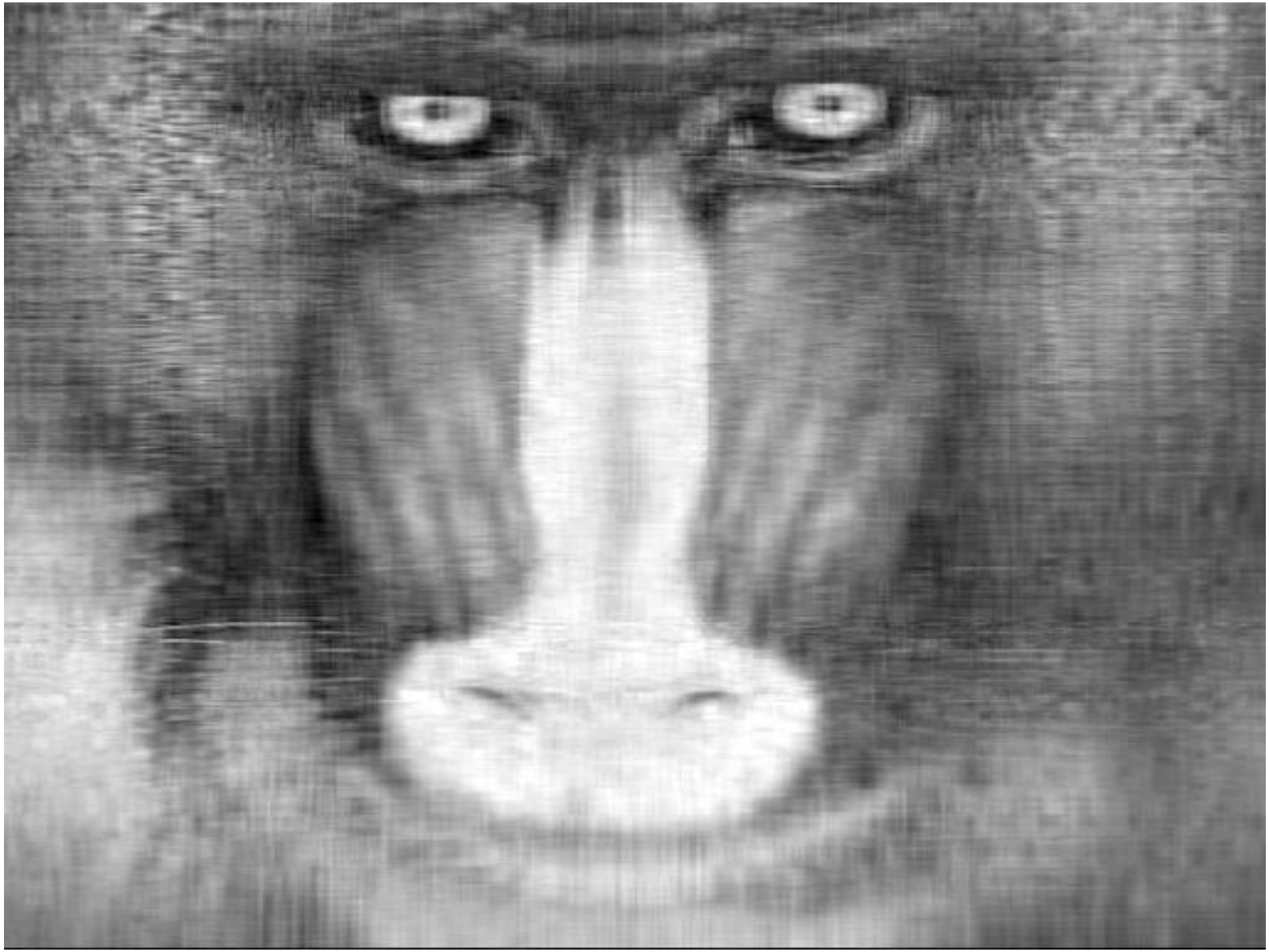


MATLAB A=C(:,:,1);

- BW image: Uniform Sampling: 512×20 matrix.



- BW image: 512×20 matrix.



- Coreset: a smaller set of data to produce a Sketch of A:

$$\underbrace{A_k}_{n\times k} = \underbrace{A}_{n\times d}\underbrace{R}_{d\times k} \approx \underbrace{A}_{n\times d}.$$

- Two algorithms: probability structure unspecified:

  1. Random projections: R dense, removes non-uniformity, then sample uniformly.

  2. Leverage sampling/CSSP: R is indicator matrix, sparse. Takes the non-uniformity into account during sampling.

Best Low Rank Approximation

- SVD: $A = U\Sigma V^T$.
- Best rank k approximation,

$$A_k = U_k\Sigma_k V_k^T = P_{U_k}A = U_k U_k^T A = AV_k V_k^T.$$

  $\min\|A - A_k\|_\xi$, ξ= Frobenius or, 2 (spectral).

- Columns of U_k : linear combinations of all columns of A
- Rows of V_k^T : linear combinations of all rows of A
- Problem: Takes $O(nd^2)$ operations.

Random Projections

- Embed (project, or map) a set of points $(u_1, \ldots, u_n)$ in $\mathbb{R}^d$ down to $(v_1, \ldots, v_n)$ in $\mathbb{R}^k$, $k << d$

- Johnson-Lindenstrauss Lemma: random projections yield small distortions in pairwise difference between points

  For $k \geq k_0$, any $\epsilon \in (0, 1/2)$, $k \geq k_0 = O(\log n/\epsilon^2)$,

$$(1-\epsilon)\|u_i - u_j\|^2 \leq \|v_i - v_j\|^2 \leq (1+\epsilon)\|u_i - u_j\|^2.$$

Proof: Bound eigenvalues of randomly perturbed matrices. The map can be found quickly, ie. in polynomial time.

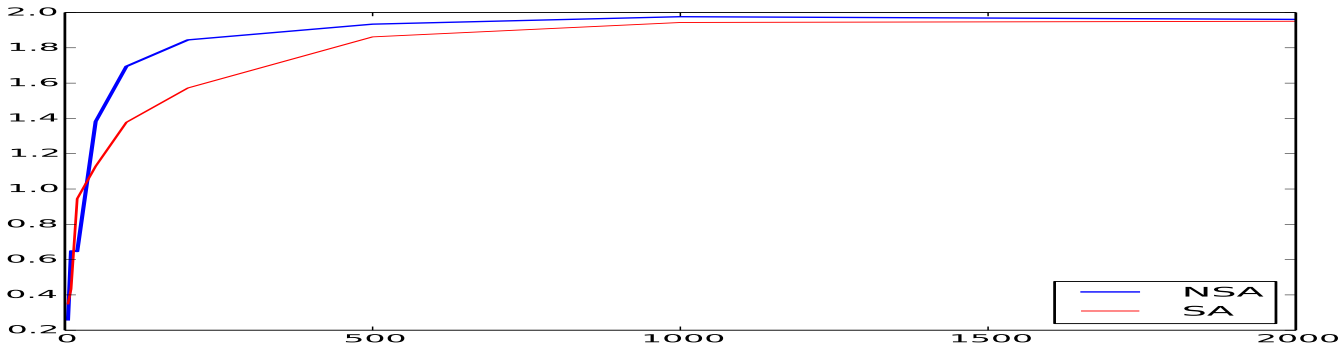
JL Transform: $v = \frac{1}{\sqrt{k}} Ru$

- (dense) R: $R_{ij} \sim N(0, 1)$.
- (sparse) R: $R_{ij} = \{1, 0, -1\}$ with probability $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$.
- Fast JL transform: $R = DHS$
  - S $(d \times k)$: samples columns uniformly, no replacement.
  - D $(d \times d)$: diagonal, $D_{ii} = \{+1, -1\}$ with probability $\frac{1}{2}$.
  - H: $(n \times n)$ Hadamard matrix, destroys non-uniformity.
- Error bound: $B_k = A \cdot R$, $n \times k$ s.t. if $k \geq r$:

$$\|A - P_{B_k} A\|_F \leq (1 + \epsilon)\|A - A_k\|_F.$$

- But I don't care about all the components!

Figure: Corr($PCA_j(A)$, $PCA_j(B_k)$): Beer



- From factor analysis: focus on the common variations.
- Idea: Only compare the largest few components,eg. 2.

$$R^2(B_k) = R_1^2(B_k) + R_2^2(B_k)$$

- good approximation when $k \approx 1K < 5K < 65K$.
- Error analysis of B_k should depend on object of interest?

The Column Subset Selection Problem (CSSP)

- Random projections: columns of B_k are combinations of columns of A. Not often meaningful (eg. barcodes)

- CSSP:

  1 (random) select k_1 columns from A.

  probability of selecting j : $\min(1, c \cdot p_j)$, $\quad p_j = \frac{1}{k}\|(V_k^T)^{(j)}\|_2^2$

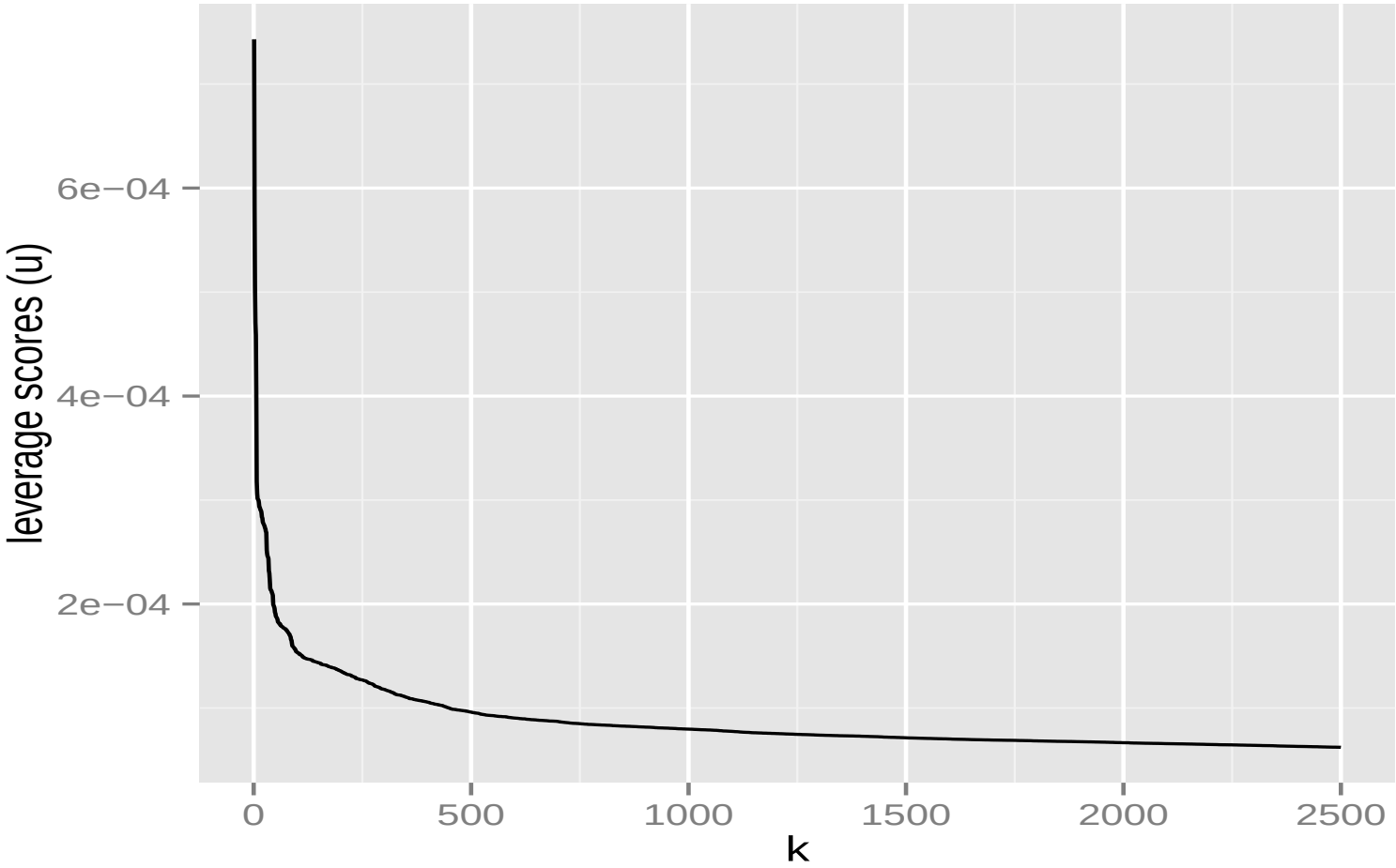
  2 (deterministic) use RRQR to pick k columns.

Leverage Scores

- what is $p_j = \frac{1}{k}\sum_{i=1}^{k}(V_k^T)^{(j)})^2$?
  - Hat matrix $H = X(X^TX)^{-1}X^T = UU^T$.
  - H_{ii}: influence of i. High influence = high leverage.
  - $(V_k^T)^{(j)}$ is (column) leverage score: favors columns that exert more influence on A.
  - Normalization by k ensures that p_j sums to one. p_j defines an importance sampling distribution.
- Error analysis: put k columns of A into C_k s.t.

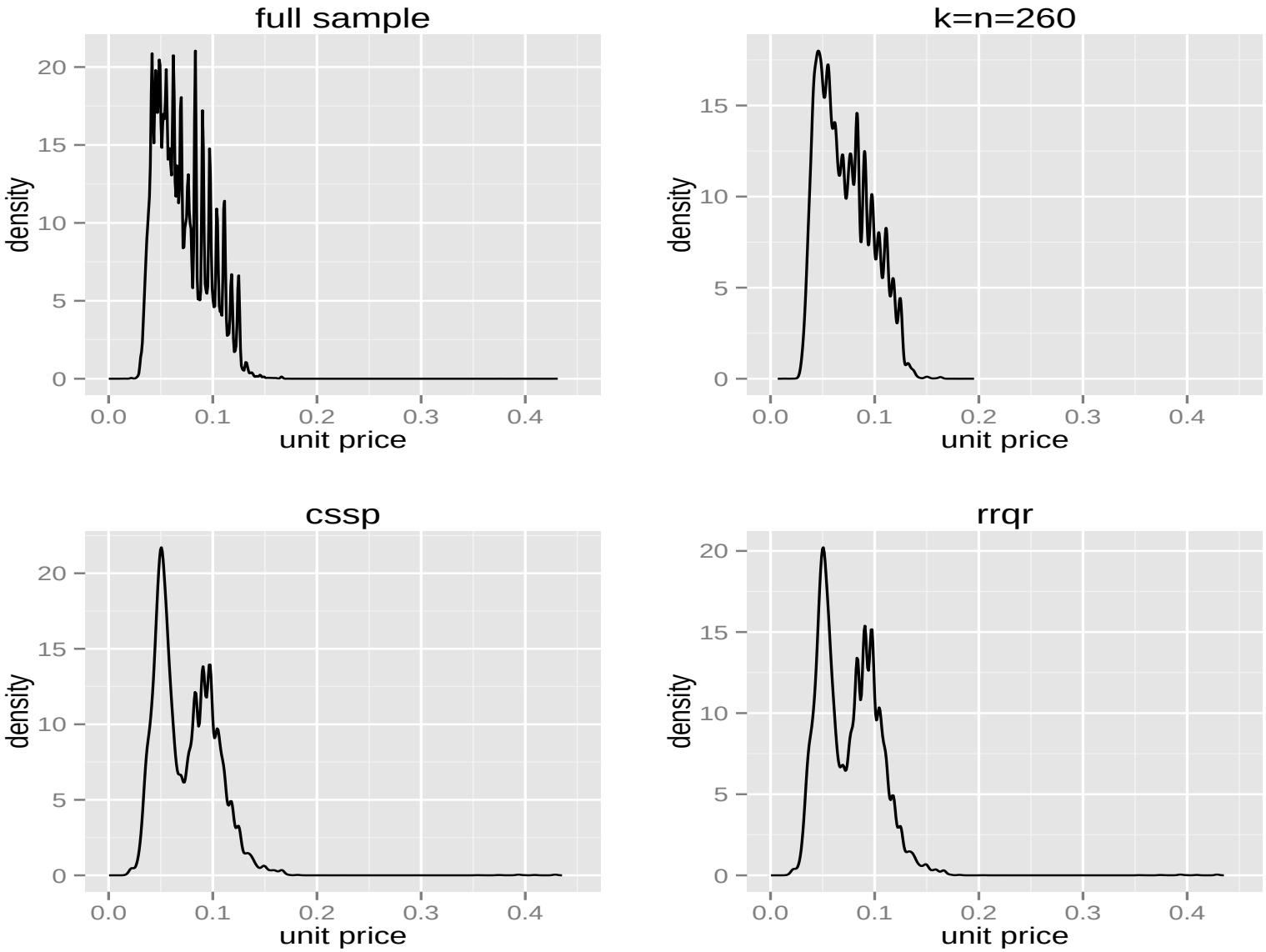
$$\|A - P_{C_k}A\|_\xi \leq k\sqrt{\log k}\|A - A_k\|_\xi.$$

Figure: Leverage Scores: Unit Price, Beer

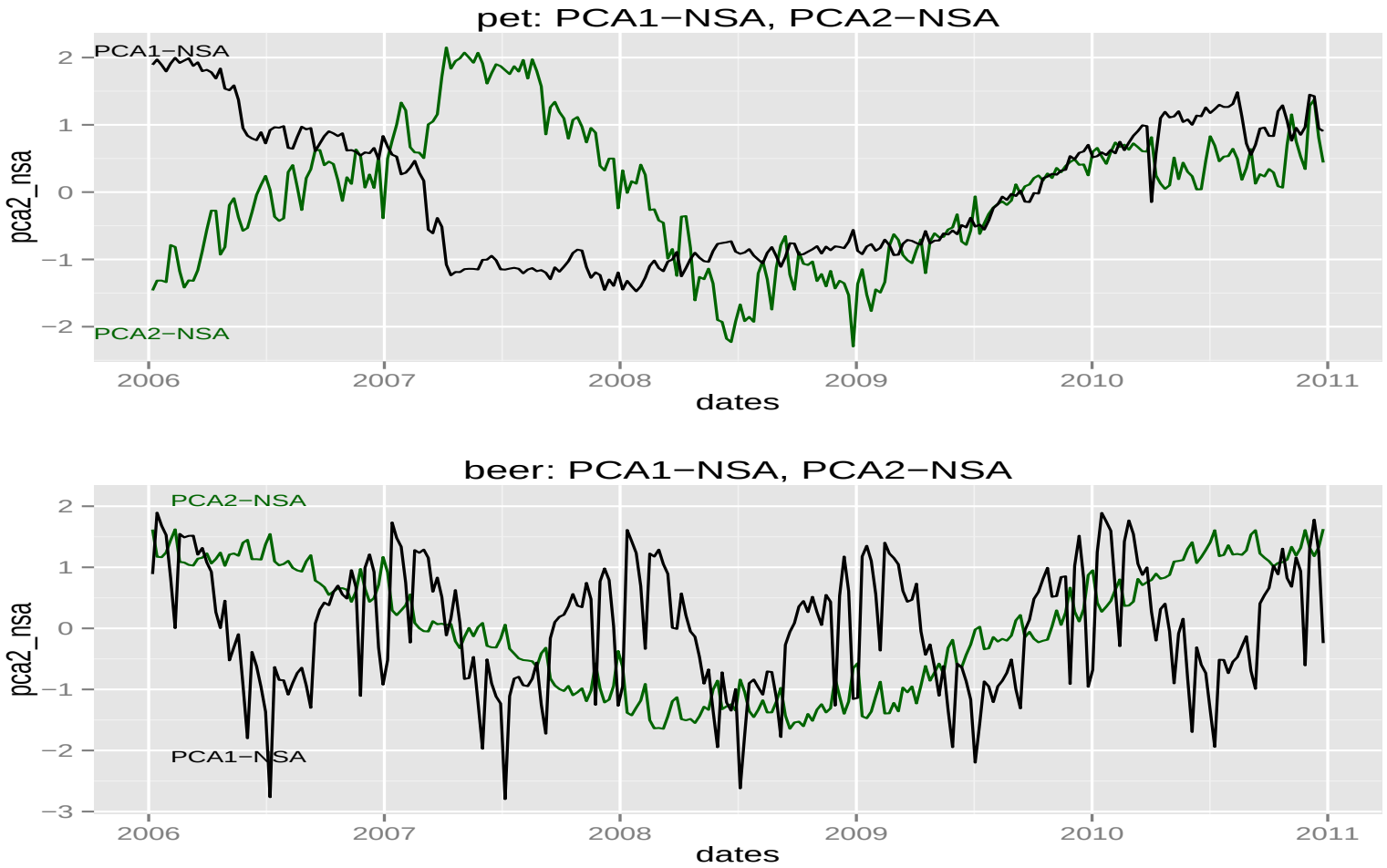


Non-uniform data structure. What are the implications?

Figure: Density: Unit Price, Beer



PCA1, PCA2 (seasonally unadjusted data)



Trends, Cycles and Seasonality

$$z_t = \underbrace{d_t}_{trend} + \underbrace{c_t}_{cycle} + \underbrace{s_t}_{seasonal} + \underbrace{h_t}_{holiday} + \underbrace{e_t}_{irregular}$$

- No more BLS/StatsCan.
- Gregorian calendar: 400 year (20871 week) cycle.
  - Christmas and new year can be any day of the week.
  - variations not exactly periodic.
  - X-13/structural modeling not practical when we have millions of series.

Seasonal Component

- Linear regression: based on CATS-D

$$
\begin{aligned}
s_{it} = \quad = & \sum_{v=1}^{k_y}\left[a_{iyv}\sin(2\pi v\cdot y_t)+b_{iyv}\cos(2\pi\cdot vy_\tau)\right] \\
& +\sum_{v=1}^{k_m}\left[a_{mi}\sin(2\pi v\cdot m_\tau)+b_{mv}\cos(2\pi\cdot vm_\tau)\right] \\
& +\vartheta_1 \mathrm{TEMPMAX}_{it}+\vartheta_2 t.
\end{aligned}
$$

where $y_t = \frac{\text{day of year}_t}{\text{days in year}_t}$, $m_t = \frac{\text{day of month}_t}{\text{days in month}_t}$.

- stochastic variations: climate data at fips county level merged with scanner data.

Holiday Component

- Idea: dates of holidays are common across units, even if effect might differ.
- Let the data determine the dummies.
  - $\mathbb{H}_\tau$: weeks of top sales in year τ at unit level
  - $\mathbb{A}$: weeks of top aggregate quantities sold over five years.
  - make sure $\mathbb{H}_\tau$ and $\mathbb{A}$ do not overlap.

Beer

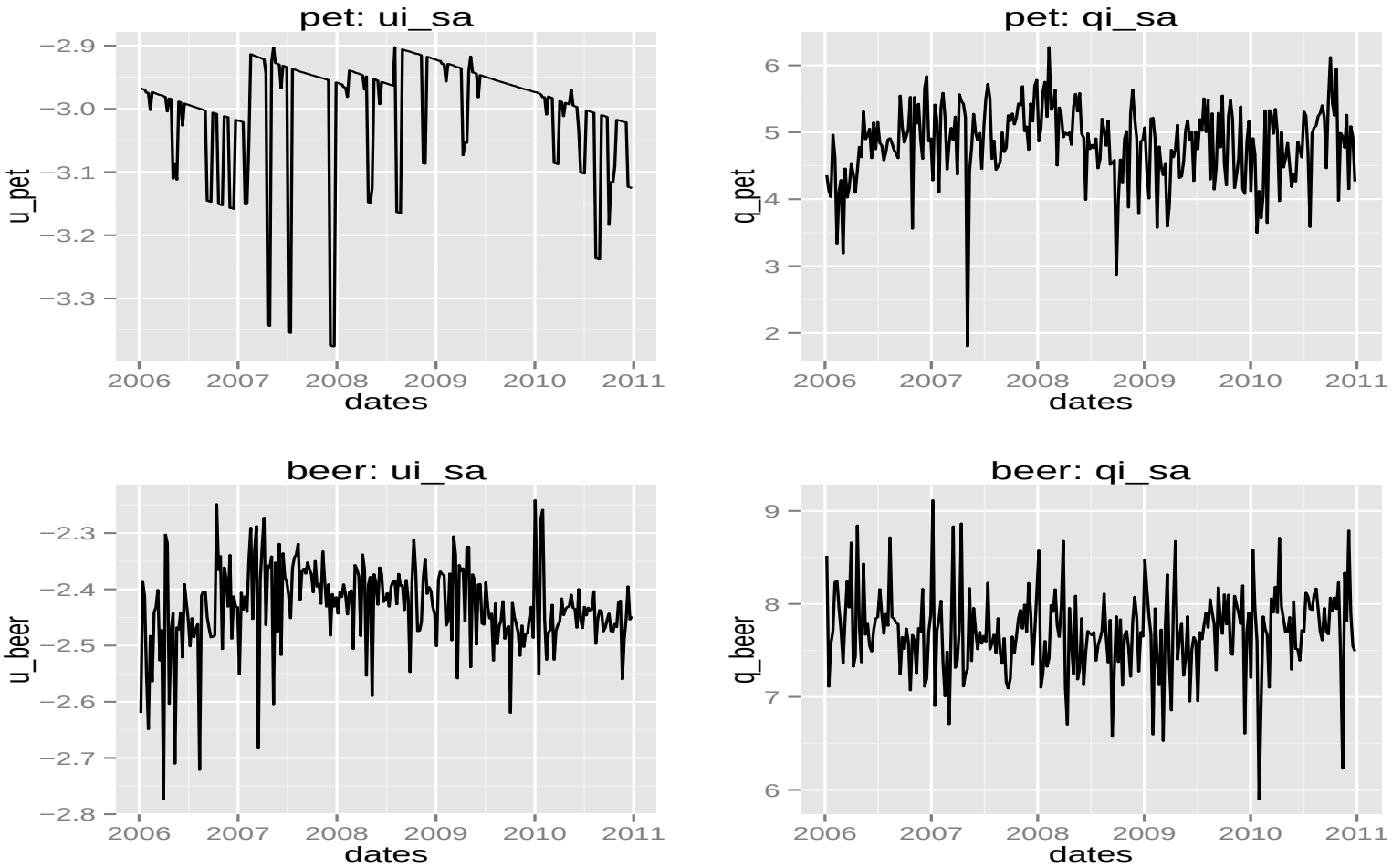
- $\mathbb{A} = (27, 22, 21, 26, 25, 36, \ldots)$
- $\mathbb{H}_\tau$:

Year	Top Weeks: Units of Beer Sold					
2006	27	51	22	47	18	26
2007	27	51	22	47	18	36
2008	27	51	22	28	36	1
2008	27	51	22	28	1	21
2010	27	22	26	18	47	6

- Series by series instead of pooled regression.
- Too much heterogeneity.

Seas. Adj: pet food and beer, selected series

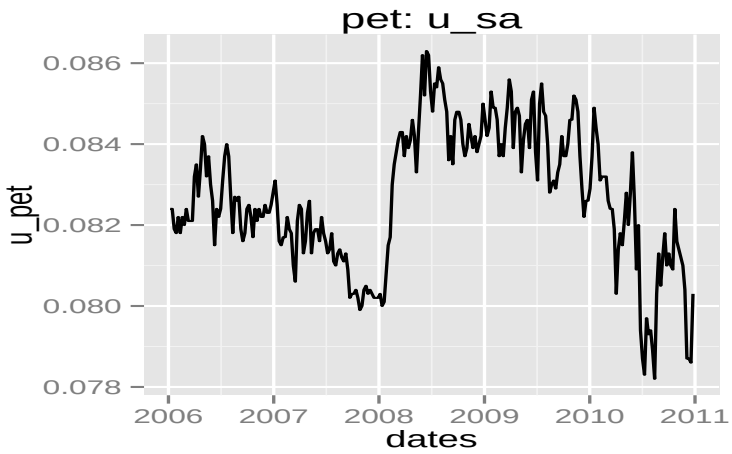
unit price volume

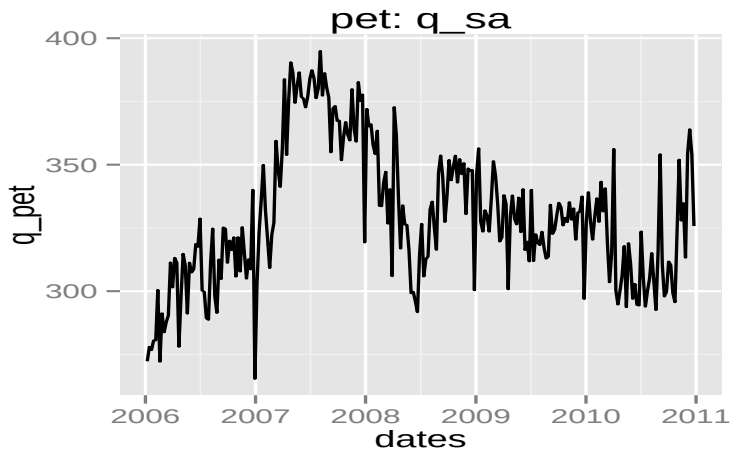


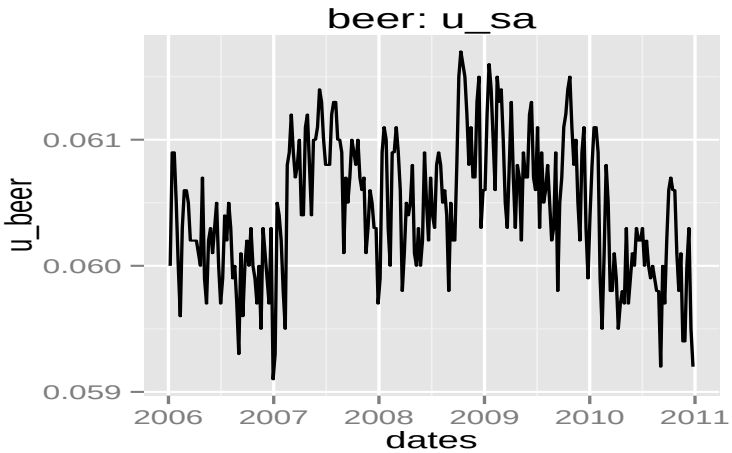
No more seasonal variations in individual series.

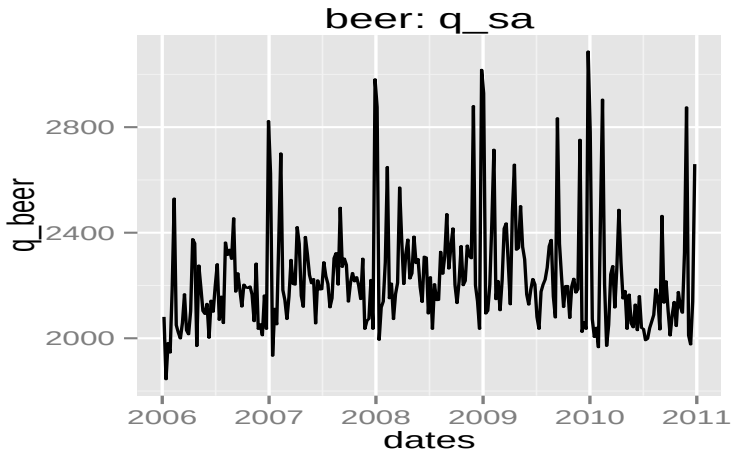
Seas. Adj: pet food and beer, aggregate

unit price volume





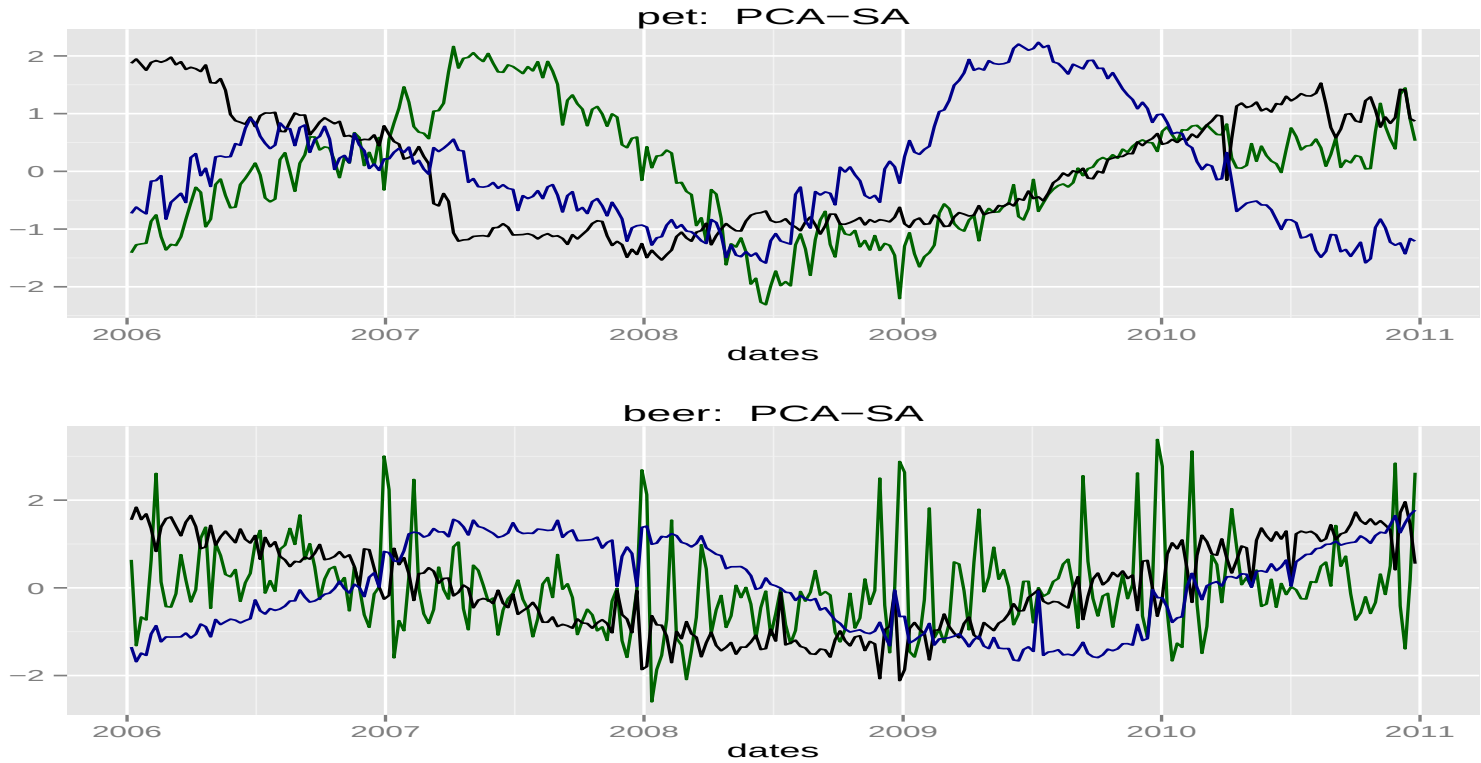




Adjust then aggregate does not always work!

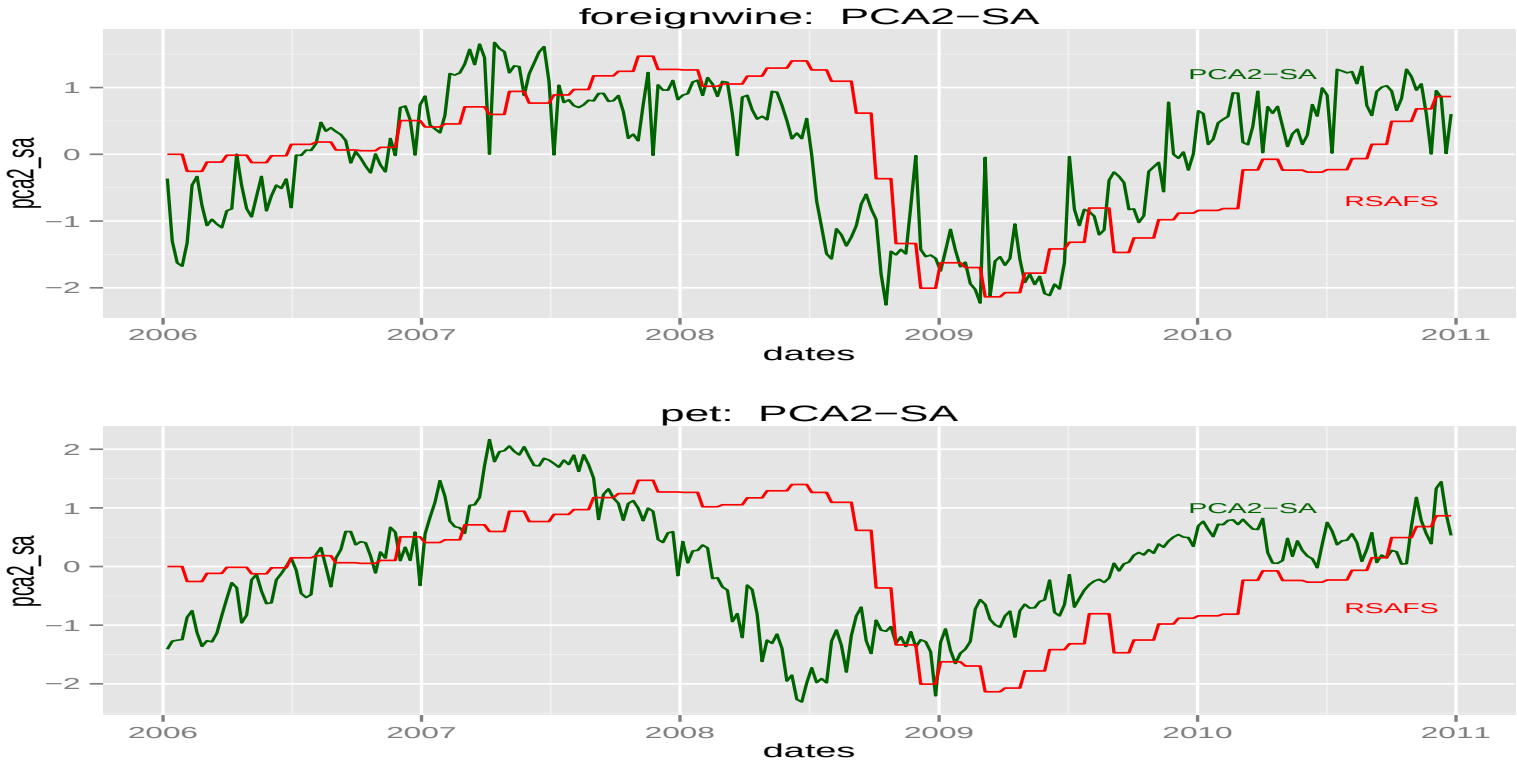
But can I find the cyclical component?

PCA1, PCA2, PCA3: pet food and beer



pca	1 (black)	2 (green)	3 (blue)
beer	trend	seasonal	cycle
pet food	trend	cycle	cycle

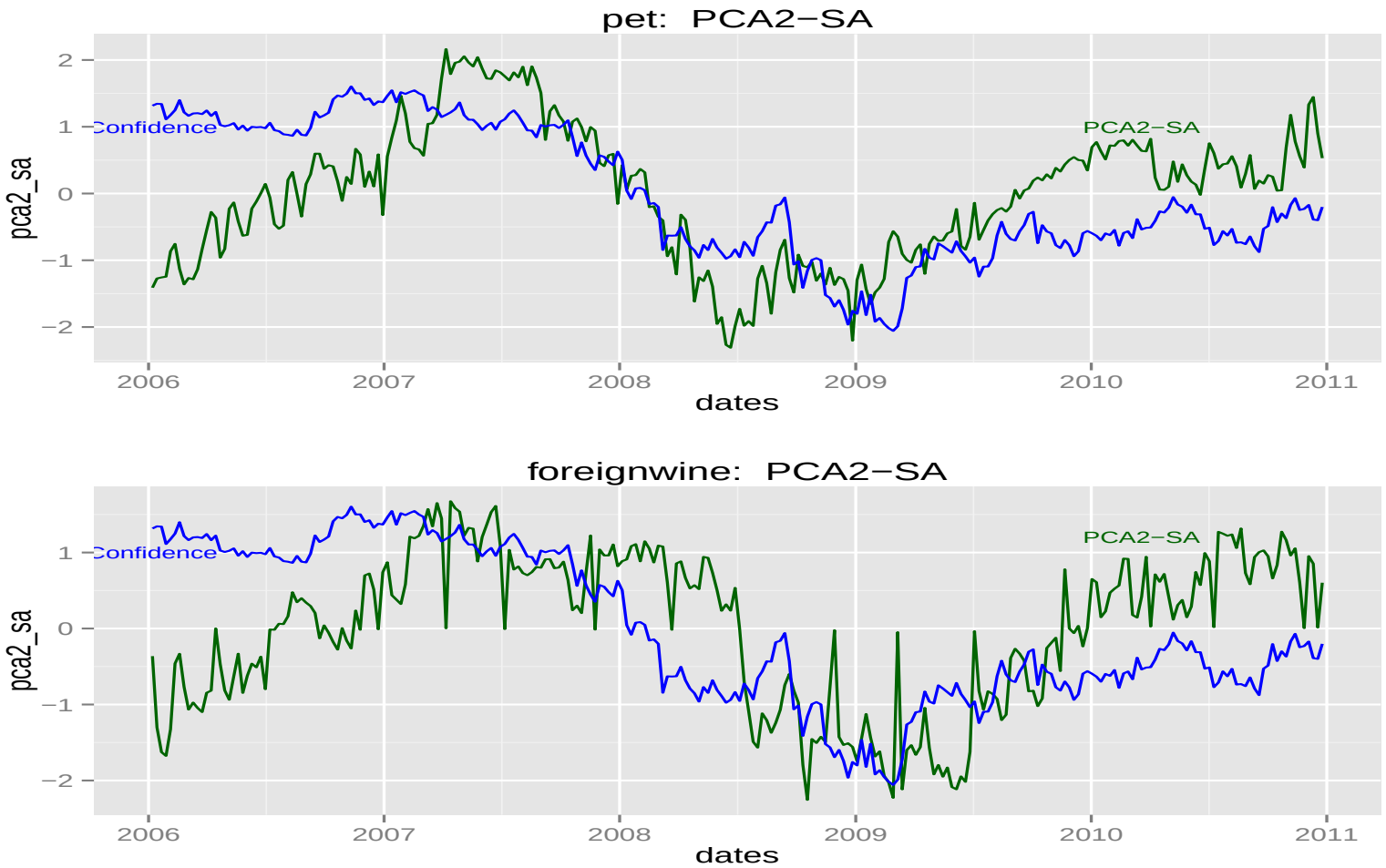
Retail Sales, FRED vs. PCA2-SA



PCA2	peak	trough
Pet food	2007-05-12	2008-06-21
Foreign wine	2007-03-24	2008-12-27

Leads RSAFS. Useful for monitoring?

Rasmussen Confidence vs. PCA2-SA



Actions and intent are in sync.

Some thoughts: Econometrics

- 3Vs: methods we know well may not work off the shelf.
  - Often need subjective tinkering.
  - Replication of results can be a real challenge!
- Variety: Pool or not pool? Bottom up or top down?
- DGP: strong assumptions, better than no structure?
- Easy implementation vs. Optimality?

Some Thoughts: Algorithms

- Scalable and efficient. But how to evaluate?
  - in terms of what we are interested in?
  - factor models, density: back to probabilistic structure.
- Coresets, active area of research:
  - L_p regressions, graph, network analysis: solve overdetermined systems. What about bias, mse?
- Computational efficiency $\Rightarrow$ statistical efficiency?
  - Not necessarily: Bin Yu, Trevor Hastie eg.
- Customize algorithms for economic data? select multiple matrices not mutually independent, prior information.

Conclusion

- Big data: not substitute for conventional data.
- Statistical models and Algorithms: co-exist.
- Big data here to stay: need scalable methods.
  - Statistical foundations of algorithms
  - Statistical machine learning.
- 75% data cleaning, 25% analysis. Big data, big hay stack. Is there is a needle?

THANK YOU FOR COMING TO THE TALK

Acknowledgments:
1. Organizers and members of program committee
2. My grad school bound exceptional students:
   - Rishab Guha and Evan Munro