Unsupervised learning and clustering

Artificial Intelligence and Machine Learning for SupTech – Lecture 6

Iman van Lelyveld – Michiel Nijhuis

VU Amsterdam

Unsupervised learning and clustering

1. Supervised versus unsupervised learning
2. What can we do with unsupervised learners?
   - K-means, t-SNE, DBSCAN, Gaussian mixtures
3. How to open the black box and explain results?

Outline

Unsupervised learning and clustering

- Unsupervised Learning
  - *k*-Means clustering
  - t-SNE
  - DBSCAN
  - Gaussian mixtures

Knowledge clips

- K-means animation (Andrey Shabalin) (link)
- K-means clustering (StatQuest) (link)

Outline

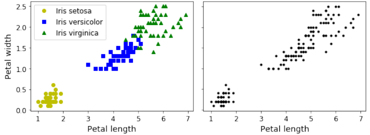
Unsupervised learning

- The vast majority of data is unlabeled → enter unsupervised learning
- Flavours:
  - Clustering: group observations in similar groups for customer segmentation, recommender systems
  - Anomaly detection: what is “normal” so you can detect abnormal observations
  - Density estimation: what is the PDF of a DGP. Anomalies are probably in the low density areas

Unsupervised learning

- The vast majority of data is unlabeled → enter unsupervised learning
- Flavours:
  - Clustering: group observations in similar groups for customer segmentation, recommender systems
  - Anomaly detection: what is "normal" so you can detect abnormal observations
  - Density estimation: what is the PDF of a DGP. Anomalies are probably in the low density areas

Grouping with labels is easy

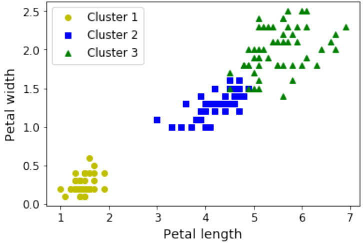


Source: Géron Github

Unsupervised learning

- The vast majority of data is unlabeled → enter unsupervised learning
- Flavours:
  - Clustering: group observations in similar groups for customer segmentation, recommender systems
  - Anomaly detection: what is "normal" so you can detect abnormal observations
  - Density estimation: what is the PDF of a DGP. Anomalies are probably in the low density areas

A Gaussian mixture model (covered later) can separate these clusters pretty well

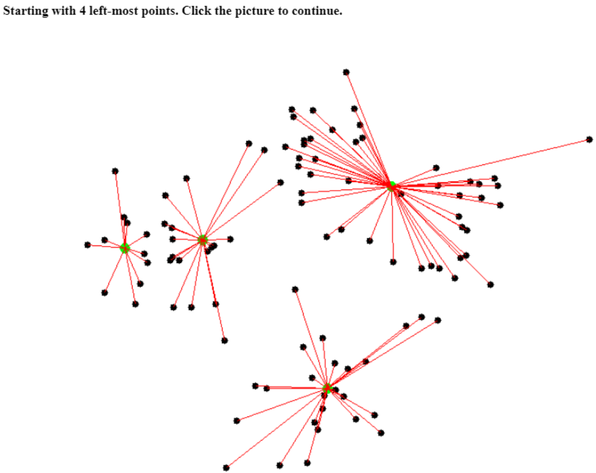


Source: Géron Github

Clustering techniques

- Goal: to find subgroups or clusters by partitioning dataset into distinct groups that are maximally "different" from one another.
  - Requires a definition of what is similar/different. This is often domain-specific
- Types of clustering techniques
  - *k*-means clustering: requires decision for the number of clusters k
  - t-SNE: non-linear PCA
  - DBSCAN: looks for "dense" areas in feature space
  - Gaussian mixtures: data is generated form an unknown mixture of several Gaussian distributions with unknown parameters
  - Agglomerative clustering, BIRCH, mean-shift, affinity propagation, spectral clustering

k-Means clustering



k-Means – considerations

- hard clustering = assigning to a single cluster vs soft clustering = assigning a score
- Inertia is the performance metric: mean squared distance to the closest centroid

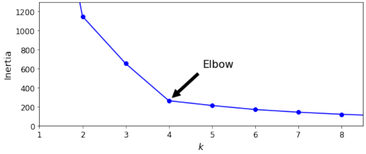
Disadvantages

1. Guaranteed to converge (mostly quickly) but unclear if clustering is optimal
   - See plotting the inertia attribute. Or increase *n_init*
   - Use MiniBatchKMeans estimator in SKLearn, which reduces computation time significantly with only slight worse quality (See comparison)
2. Not easy/impossible to spot visually in more than 3 dimensions
3. Requires choosing the number of clusters.

k-Means – finding the right number of k

- Plot inertia over k and find elbow

Elbow

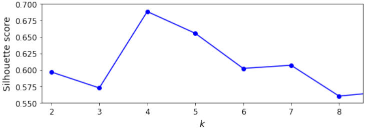


Source: Géron Github

k-Means – finding the right number of *k*

- Plot inertia over *k* and find elbow
- Plot silhouette
  - (b-a) / max(a,b) with a is mean intra-cluster distance and b is mean distance to the next cluster
  - Range: -1 to +1

Silhouette

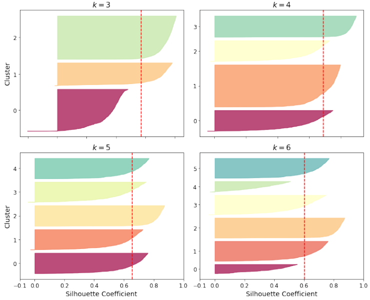


Source: Géron Github

k-Means – finding the right number of k

- Plot inertia over k and find elbow
- Plot silhouette
  - (b-a) / max(a,b) with a is mean intra-cluster distance and b is mean distance to the next cluster
  - Range: -1 to +1
- Plot distribution of silhouette scores

Silhouette diagram

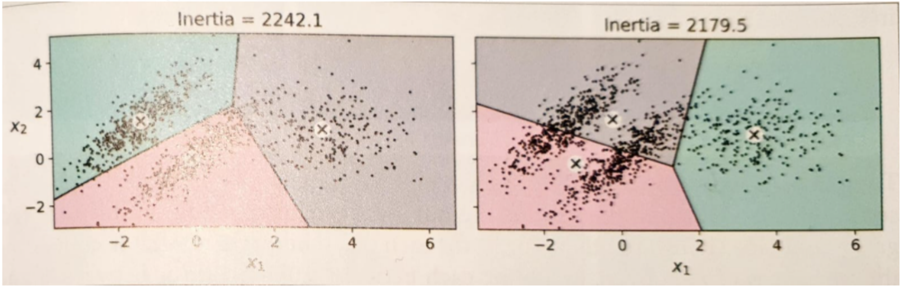


Source: Géron Github

k-Means – limits

K-means does not behave well if:

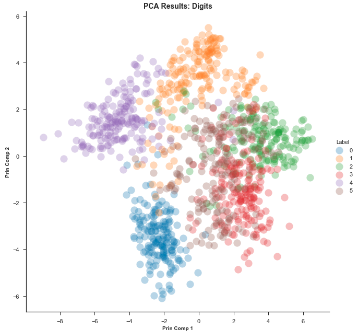
- clusters have varying sizes
- different densities
- nonspherical shapes



Source: Géron (2019)

t-Distributed Stochastic Neighbor Embedding (t-SNE)

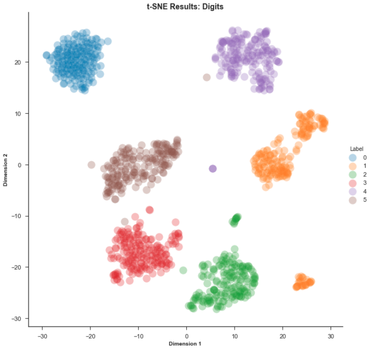
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data
- Difference between PCA and t-SNE: linear vs non-linear
- t-SNE calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space. It then tries to optimize these two similarity measures using a cost function



Source: Violante Medium

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data
- Difference between PCA and t-SNE: linear vs non-linear
- t-SNE calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space. It then tries to optimize these two similarity measures using a cost function



Source: Violante Medium

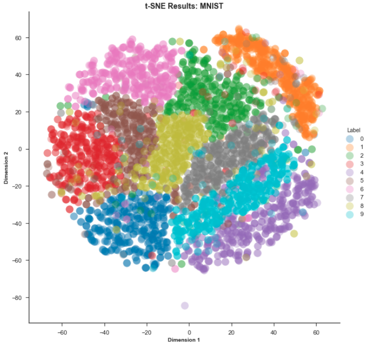
t-Distributed Stochastic Neighbor Embedding (t-SNE)

- t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data
- Difference between PCA and t-SNE: linear vs non-linear
- t-SNE calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space. It then tries to optimize these two similarity measures using a cost function

Source: Violante Medium

t-Distributed Stochastic Neighbor Embedding (t-SNE)

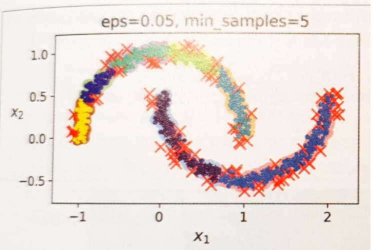
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is an unsupervised, non-linear technique primarily used for data exploration and visualizing high-dimensional data
- Difference between PCA and t-SNE: linear vs non-linear
- t-SNE calculates a similarity measure between pairs of instances in the high dimensional space and in the low dimensional space. It then tries to optimize these two similarity measures using a cost function



Source: Violante Medium

DBSCAN

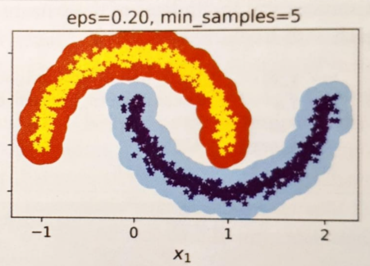
- Looks for "dense" areas in feature space
- Has just 2 hyperparameters: ϵ and *min_samples*
- Works well if dense areas are clearly separated by sparse areas



Source: Géron (2019)

DBSCAN

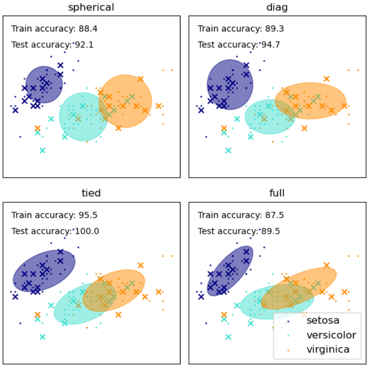
- Looks for "dense" areas in feature space
- Has just 2 hyperparameters: ϵ and *min_samples*
- Works well if dense areas are clearly separated by sparse areas



Source: Géron (2019)

Gaussian mixtures

- Assumes data is generated form an unknown mixture of several Gaussian distributions with unknown parameters
- *Expectation Maximization* (EM)
  - similar to k-means
  - Estimates not only the center but also size, shape, orientation and relative weight with soft assignments
- To reduce computational complexity adjust *covariance_type*: "spherical", "diag", "tied" and "full" (default)

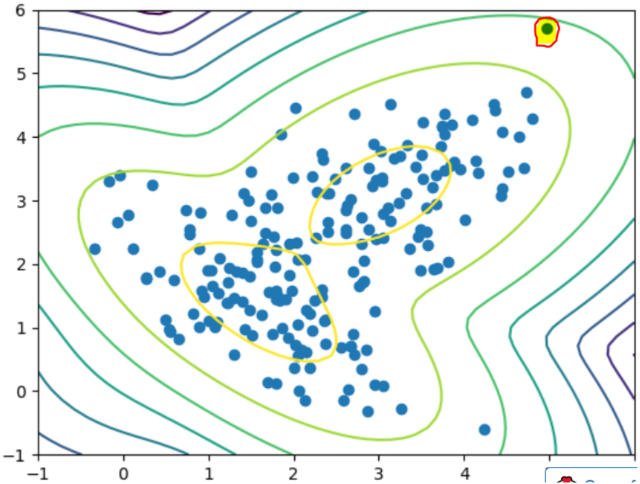


Source: Géron (2019)

Gaussian mixtures – choosing k

- Number of clusters k is a hyperparameter (similar to K-means)
- inertia or silhouette not well defined
- Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC)
- Both BIC and AIC penalize models that have more parameters to learn (== more clusters) and reward models that fit the data well
- Plot BIC/AIC for an elbow plot

Gaussian mixtures for outlier detection



Other algorithms for outlier detection

- Principle Components Analysis (PCA)
- Fast-MCD (minimum covariance determinant): variant of Gaussian mixture with a single distribution
- Isolation forest: Random Forest where each decision tree is grown randomly. At each node a random feature is used to split using a random threshold. Outliers tend to be split of relatively fast.
- Local Outlier Factor (LOF): compares density with its neighbors' density
- One-class SVM: can we split observations from origin? Does not scale

Summary

In this lecture we covered:

1. Some unsupervised learners
   - k-Means clustering, t-SNE, DBSCAN, Gaussian mixtures
2. A first look at explainable AI
   - white box, global and local explainability
3. An Appendix with discussion of how ML approaches can help tame the Fama-French "factor zoo"

Applications

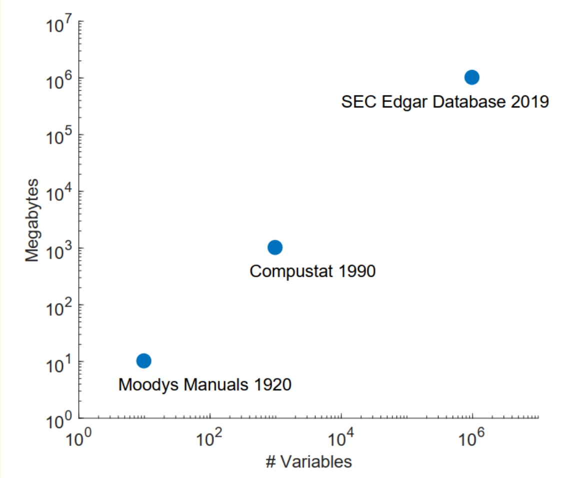
- So far we've talked a lot about techniques and relatively little about applications for finance such as:
  - Classification: robo advice, fraud detection
  - Forecasting: trading bots
  - NLP: compliance
- Here we will look at one example: asset pricing based on Kozak et al. (2019)
- Also see Bianchi, Büchner, Hoogteijling, and Tamoni (2021), Bianchi, Büchner, and Tamoni (2021), Chen (2021), Easley et al. (2021), Erel et al. (2021), Farboodi et al. (2022), Fuster et al. (2021), Goldstein et al. (2021), Leippold et al. (2022), Li et al. (2021), and Obaid and Pukthuanthong (2022)

How would all of this work in asset pricing?

- Prediction is central to ML and also essential to asset pricing (AP)
  - Forecasting returns
  - Forecasting cash-flows
  - Forecasting default
  - Forecasting risk exposures
- Fundamental asset pricing equation for asset with excess return R and Stochastic Discount Factor (SDF) M:
  $\mathbb{E}[R_{t+1}M_{t+1}|x_t] = 0$
- Empirical implementation involves function approximation $x_t \rightarrow$ (Co-)moments of $R_{t+1}; M_{t+1}$
- This is a supervised learning problem + maybe dimension reduction in joint distribution of $(R_{t+1}; M_{t+1})$: unsupervised learning
- Pre-ML literature: x_t typically low-dimensional but little real-world justification

This section is based on Kozak et al. (2019)

More data available to extend the set of x_t in asset pricing



Source: Nagel (2019)

Overfitting! – bring in prior knowledge through regularization

- Consider supervised learning problem: find $y_i = f(x_i)$ where $i = 1, 2, \ldots, N$ and x_i has dimension $J \times 1$.
- When x_i high-dimensional (e.g., $J > N$), standard methods (e.g., OLS) would horribly overfit in-sample $\rightarrow$ bad out-of-sample (OOS) prediction performance
- Regularization: Penalize estimation results that are regarded as implausible based on prior knowledge
  - Example: if big magnitudes of regression coefficient on Sharpe ratio are a priori unlikely, penalize big coefficient estimates
- Remember: many ML methods can be derived as penalized estimators

Many ML methods can be derived as penalized estimators

$$\hat{\theta} = arg\, \underset{\theta}{min} \Sigma_i L\{y_i - f(x_i, \theta)\} + \lambda R(\theta)$$

for loss function $L(.)$ and penalty function $R(.)$.

$R(\theta) = \|\|\theta\|\|_1$:	Lasso
$R(\theta) = \|\|\theta\|\|_2^2$:	Ridge regression
$R(\theta) = \alpha\|\|\theta\|\|_1 + (1-\alpha)\|\|\theta\|\|_2^2$:	Elastic net

- Penalty forces regularization: Well-behaved estimates, useful for prediction, even if $J > N$
- Regularization crucial for prediction performance

Predict monthly return of individual U.S. stocks with past returns

- Cross-section of $i = 1, \ldots, N$, with $Jx1$ characteristics vector (observable predictors) x_{it}.

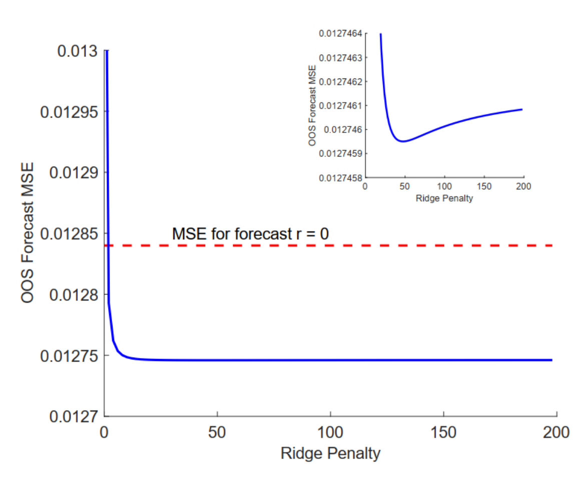
$$\mathbb{E}[r_{i,t+1}|x_{it}] = f(x_{it}, \theta)$$

- Observations $r_t = (r_{1t}, \ldots, r_{N,t})$ for $t = 1, \ldots, T$.
- x_{it} contains:
  - 120 lags of monthly returns, $r_{it}, r_{i,t-1}, r_{i,t-2}, \ldots r_{i,t-120}$
  - 120 lags of monthly squared returns $r_{it}^2, r_{i,t-1}^2, r_{i,t-2}^2, \ldots r_{i,t-120}^2$

  where all returns are cross-sectionally demeaned each month (i.e., cross-sectional focus) and x_{it} is standardized.
- Estimate during 1980-2000. Evaluate forecasts out-of-sample during 2001-2019.
- Ridge regression (where $\lambda = 0$ implements OLS)

$$\hat{\theta} = arg\,\min_{\theta} \Sigma_i (r_{i,t+1} - \theta' x_{i,t})^2 + \lambda \theta' \theta$$

OLS is really bad



Source: Nagel (2019)

Issues in applying ML to Asset Pricing

	Typical ML application	Asset pricing
Signal-to-noise	**Outcome observable** e.g. { hotdog, not hotdog }	**Very noisy observation** of outcome e.g. {high $\mathbb{E}[r]$, low $\mathbb{E}[r]$}
Big Data dimensions	**N and J big**	**J big, N not so much**
Sparsity	**Often sparse** e.g., some regions of image irrelevant	**Unclear**
Lucas critique	**Often not an issue** e.g. hotdogs don't change shape in response to image classification	**Investors learn from data and adapt**

Source: Nagel (2019)

Quest: describe cross-section with a few factors

- Multi-decade quest: Describe cross-section of N excess stock returns, $\mathbb{E}[r]$, with small number (K) of factor excess returns where factors are returns on portfolios constructed based on firm characteristics (size, momentum, ...).
- Popular factor models are sparse in characteristics, e.g: Fama-French 3-, 4-, 5-factor models
- But can a characteristics-sparse representation of the SDF be adequate?
  - Taking into account all anomalies that have been discovered
  - Plus potentially hundreds or thousands of additional stock characteristics, including interactions
  - High-dimensional problem!

It used to be so simple . . .

THE JOURNAL OF FINANCE • VOL. LI, NO. 1 • MARCH 1996

Multifactor Explanations of Asset Pricing Anomalies

EUGENE F. FAMA and KENNETH R. FRENCH*

ABSTRACT

Previous work shows that average returns on common stocks are related to firm characteristics like size, earnings/price, cash flow/price, book-to-market equity, past sales growth, long-term past return, and short-term past return. Because these

. . . but now the list of factors is endless

Risk type		Description	Examples
Common (113)	**Financial** (46)	Proxy for aggregate financial market movement, including market portfolio returns, volatility, squared market returns, among others	Sharpe (1964): market returns; Kraus and Litzenberger (1976): squared market returns
	Macro (40)	Proxy for movement in macroeconomic fundamentals, including consumption, investment, inflation, among others	Breeden (1979): consumption growth; Cochrane (1991): investment returns
	Microstructure (11)	Proxy for aggregate movements in market microstructure or financial market frictions, including liquidity, transaction costs, among others	Pastor and Stambaugh (2003): market liquidity; Lo and Wang (2006): market trading volume
	Behavioral (3)	Proxy for aggregate movements in investor behavior, sentiment or behavior-driven systematic mispricing	Baker and Wurgler (2006): investor sentiment; Hirshleifer and Jiang (2010): market mispricing
	Accounting (8)	Proxy for aggregate movement in firm-level accounting variables, including payout yield, cash flow, among others	Fama and French (1992): size and book-to-market; Da and Warachka (2009): cash flow
	Other (5)	Proxy for aggregate movements that do not fall into the above categories, including momentum, investors' beliefs, among others	Carhart (1997): return momentum; Ozoguz (2009): investors' beliefs
Characteristics (202)	**Financial** (61)	Proxy for firm-level idiosyncratic financial risks, including volatility, extreme returns, among others	Ang et al. (2006): idiosyncratic volatility; Bali, Cakici, and Whitelaw (2011): extreme stock returns
	Microstructure (28)	Proxy for firm-level financial market frictions, including short sale restrictions, transaction costs, among others	Jarrow (1980): short sale restrictions; Mayshar (1981): transaction costs
	Behavioral (3)	Proxy for firm-level behavioral biases, including analyst dispersion, media coverage, among others	Diether, Malloy, and Scherbina (2002): analyst dispersion; Fang and Peress (2009): media coverage
	Accounting (87)	Proxy for firm-level accounting variables, including PE ratio, debt-to-equity ratio, among others	Basu (1977): PE ratio; Bhandari (1988): debt-to-equity ratio
	Other (24)	Proxy for firm-level variables that do not fall into the above categories, including political campaign contributions, ranking-related firm intangibles, among others	Cooper, Gulen, and Ovtchinnikov (2010): political campaign contributions; Edmans (2011): intangibles

Source: Factor List from Harvey et al. (2016)

Recent contributions

	Regularization	Assets	Nonlinearity
SDF models			
Kozak, Nagel, Santosh (2019)	elastic net	char. portfolios PC portfolios	interactions
Kozak (2019)	elastic net	char. portfolios PC portfolios	kernels
Giglio, Feng, and Xiu (2019)	Lasso	char. portfolios	-
DeMiguel et al. (2019)	Lasso	char. portfolios	-
Beta models			
Kelly, Pruitt, Su (2018)	PCA cutoff	indiv. stocks	-
Gu, Kelly and Xiu (2019)	Lasso	char. portfolios	autoencoder neural nets
Return prediction models			
Freyberger, Neuhierl, Weber (2018)	Group lasso	indiv. stocks	splines
Moritz and Zimmerman (2016)	Random forest	indiv. stocks	interactions
Gu, Kelly, Xiu (2018)	many	indiv. stocks	many

Source: Nagel (2019)

VU VRIJE UNIVERSITEIT AMSTERDAM

How to tackle this with ML?

- penalize based on economic theory to reduce overfitting

$$\hat{b} = \arg\min_b (\hat{f} - \Sigma b)' \Sigma^{-1} (\hat{f} - \Sigma b) + \underbrace{\gamma_1 b' b}_{L2} + \underbrace{\gamma_2 \Sigma_{i=1}^{H} |b_i|}_{L1}$$

- L_1 en L_2 are regularization penalties and are based on economic theory
  - Sharpe ratio's can't be too big
  - Many of the covariates will be uninformative
- Summary of key results
  1. Shrinkage is extremely important
  2. Very little redundancy in original characteristics space: Characteristics-sparse SDF not achievable
  3. But PC-sparse SDF based on a few (high-variance) PCs prices well
- Result (2) could be partly a consequence of looking at a set of data-mined anomalies
- Could there be more characteristics-sparsity if we include some unexplored factors, or factors that are not known to be associated with return premia?

How does Big Data affect investors?

- See Martin and Nagel (2022, JFE) for an excellent discussion
- Modern investors face a high-dimensional prediction problem: thousands of observable variables are potentially relevant for forecasting
- Framed as an ML problem, N assets have cash flows that are a (linear) function of J firm characteristics, but with uncertain coefficients
- Risk-neutral Bayesian investors impose shrinkage (Ridge regression) or sparsity (Lasso) when they estimate the J coefficients of the model and use them to price assets.
- When J is comparable in size to N, returns appear cross-sectionally predictable using firm characteristics to an econometrician who analyzes data from the economy ex post. A factor zoo emerges even without p-hacking and data-mining.
- Standard in-sample tests of market efficiency reject the no-predictability null with high probability, despite the fact that investors optimally use the information available to them in real time.
- In contrast, out-of-sample tests keep their economic meaning

Big Data and Investors *cont.*

- The economic content of the (semi-strong) market efficiency notion that prices "fully reflect" all public information is not clear in a high-dimensional setting
  - Abstracting from joint hypothesis problem Fama (1970, JoF): the econometrician studying asset prices does not know the model that determines risk premia required by risk-averse investors
- Does "fully reflect" mean:
  1. investors know the parameters of the cash-flow prediction model $\rightarrow$ typical RE notion?
  2. investors employ Bayesian updating when they learn from data about the parameters of the cash-flow prediction model?
- The null hypothesis in a vast empirical literature in asset pricing is 1)
  - Literature on return predictability regressions, event studies, and asset pricing model estimation based on orthogonality conditions
- An apparent rejection of market efficiency == unsurprising consequence of investors not having precise knowledge of the parameters of a DGP that involves thousands of predictor variables

Will your new approach be implemented?

- Is there potential "Alpha content"?
  - Does the new data or method give rise to sufficient risk-adjusted return to merit implementation of a stand-alone strategy or as a component of a portfolio strategy (cf Kolanovic and Krishnamachari (2017))
- Markets already digest a lot of information .... so the room for improvement is small

"The flat maximum effect states that for most problems there is not a single best model that is substantially better than all others." *(Finlay (2014) , page 105)*

The markets' view: J.P. Morgan

- Kaggle suggests that structured data is best analyzed by tools like XGBoost and Random Forests
- Use of Deep Learning is limited to analysis of images or text
  - Deep Learning tools still require a substantial amount of data to train. Training on small sample sizes (e.g. generative-adversarial models) is still at an early stage
  - Large sample data required implies that first applications of Deep Learning will be in intraday or high-frequency trading before we see its application in lower frequencies (See Algorithmic Trading course!!!)
- Deep Learning finds immediate use for portfolio managers in an indirect manner. Parking lot images are analyzed using Deep Learning architectures (like CNN) to count cars. Text in social media is analyzed using Deep Learning architectures (like LSTM) to detect sentiment
- Such traffic and sentiment signals can be integrated directly into quantitative strategies (See Kolanovic and Krishnamachari (2017))
- Calculation of signals often outsourced to specialized firms

References I

- Bianchi, D., Büchner, M., Hoogteijling, T., & Tamoni, A. (2021). Corrigendum: Bond Risk Premiums with Machine Learning. The Review of Financial Studies, 34, 1090–1103.
- Bianchi, D., Büchner, M., & Tamoni, A. (2021). Bond Risk Premiums with Machine Learning [Publisher: Oxford Academic]. The Review of Financial Studies, 34(2), 1046–1089.
- Chen, A. Y. (2021). The Limits of p-Hacking: Some Thought Experiments [Publisher: John Wiley & Sons, Ltd]. The Journal of Finance, 76(5), 2447–2480.
- Easley, D., López De Prado, M., O'hara, M., & Zhang, Z. (2021). Microstructure in the Machine Age. The Review of Financial Studies, 34, 3316–3363.
- Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2021). Selecting Directors Using Machine Learning. Review of Financial Studies, 34, 3226–3264.
- Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. The Journal of Finance, 25(2), 383–417.

References II

- Farboodi, M., Matray, A., Veldkamp, L., & Venkateswaran, V. (2022). Where Has All the Data Gone? Review of Financial Studies, Forthcomin.
- Finlay, S. (2014). Predictive Analytics, Data Mining and Big Data. Palgrave MacMillan.
- Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2021). Predictably Unequal? The Effects of Machine Learning on Credit Markets. Journal of Finance, (0), 1–43.
- Géron, A. (2019). Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. O'Reilly. Retrieved April 21, 2019, from https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/
- Goldstein, I., Spatt, C. S., & Ye, M. (2021). Big Data in Finance. Review of Financial Studies, 34, 3213–3225.
- Harvey, C. R., Liu, Y., & Zhu, H. (2016). … and the Cross-Section of Expected Returns [Publisher: Narnia]. Review of Financial Studies, 29(1), 5–68.

References III

- Kolanovic, M., & Krishnamachari, R. T. (2017). Big Data and AI Strategies - Machine Learning and Alternative Data Approach to Investing Quantitative and Derivatives Strategy. J.P. Morgan report.
- Kozak, S., Nagel, S., & Santosh, S. (2019). Shrinking the Cross Section. Journal of Financial Economics.
- Leippold, M., Wang, Q., & Zhou, W. (2022). Machine learning in the Chinese stock market. Journal of Financial Economics.
- Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. The Review of Financial Studies, 34, 3265–3315.
- Martin, I., & Nagel, S. (2022). Market Efficiency in the Age of Big Data. Journal of Financial Economics, 145(1), 154–177. http://www.nber.org/papers/w26586
- Nagel, S. (2019). Asset Pricing and Machine Learning - Lecture 1. Princeton Lectures in Finance.

References IV

Obaid, K., & Pukthuanthong, K. (2022). A picture is worth a thousand words: Measuring investor sentiment by combining machine learning and photos from news. Journal of Financial Economics.