Examples and Emerging Topics of Exploratory Data Analysis

Tomoko Matsuo

Ann and H.J. Smead
Department of Aerospace Engineering Sciences
UNIVERSITY OF COLORADO **BOULDER**

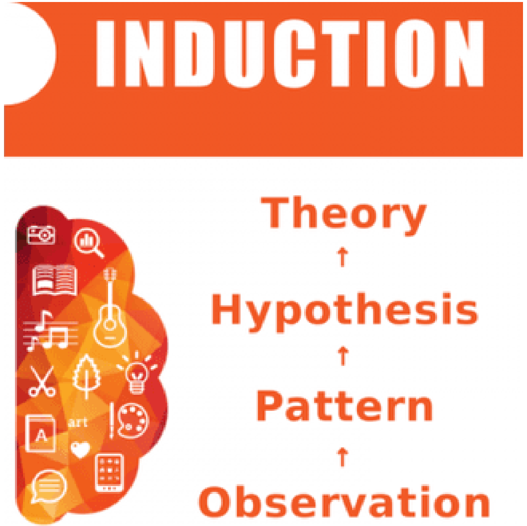
Applied
Mathematics
Department

Transform ***data*** into knowledge

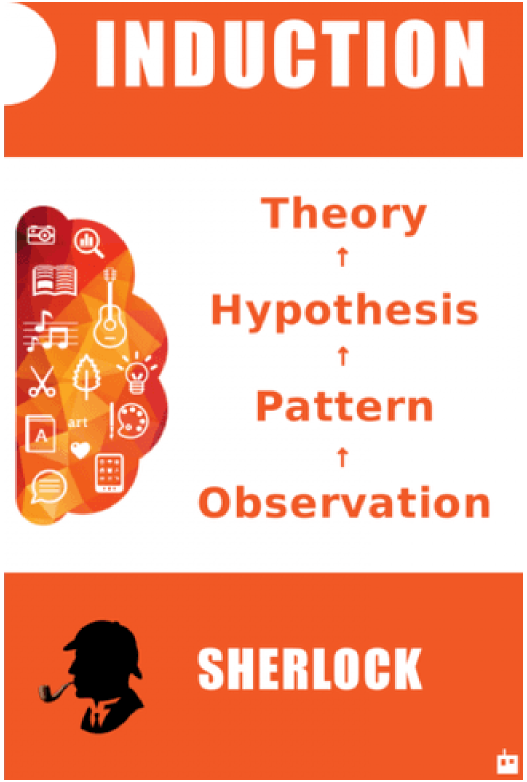
Scientific Reasoning

Theory
↑
Hypothesis
↑
Pattern
↑
Observation

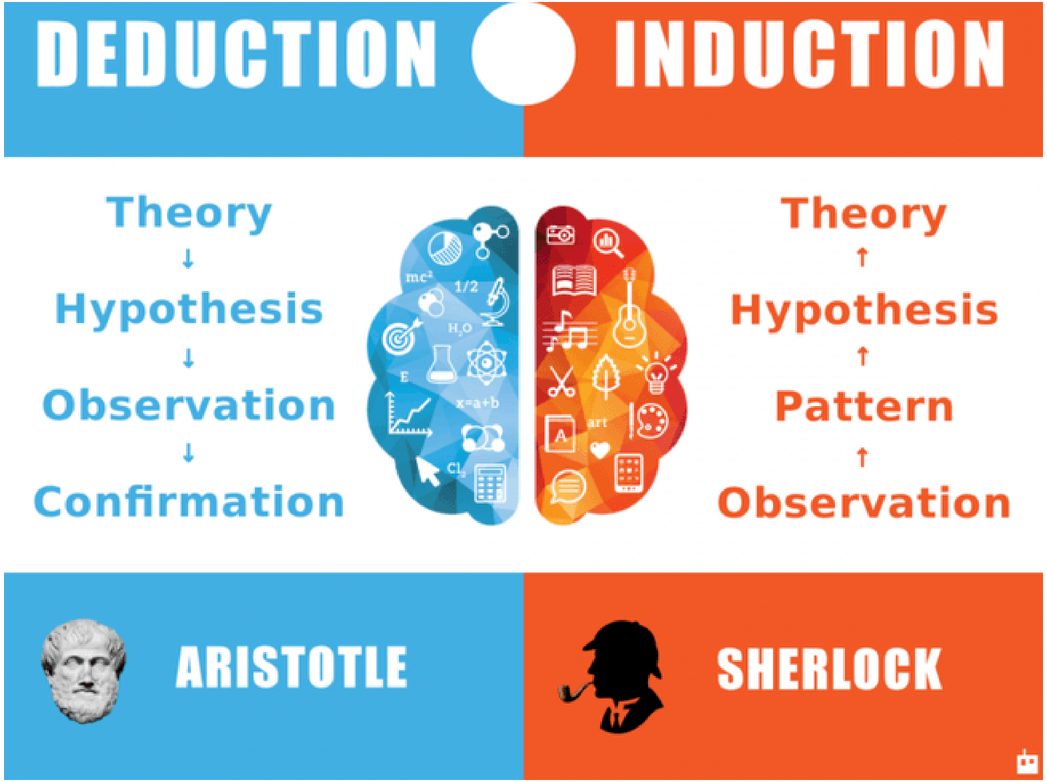
Scientific Reasoning



Scientific Reasoning



Scientific Reasoning



Learning Techniques

In machine learning, **statistical learning** techniques are used to automatically identify patterns in data. Most statistical learning problems fall into one of two categories:

- Supervised Learning

  Learning process is guided by a set of labeled samples $\{x_i, y_i\}$ *(training data), where* x_i *is the predictor measurement and* y_i *is an associated response measurement.*

- Unsupervised Learning

  No labeled data is used to supervise learning process.
  Only $\{x_i\}$ *is known.*

Your First Decision: Which way to go?

- Before anything – you got to know your data
- Remember “Garbage in Garbage out”

Unsupervised

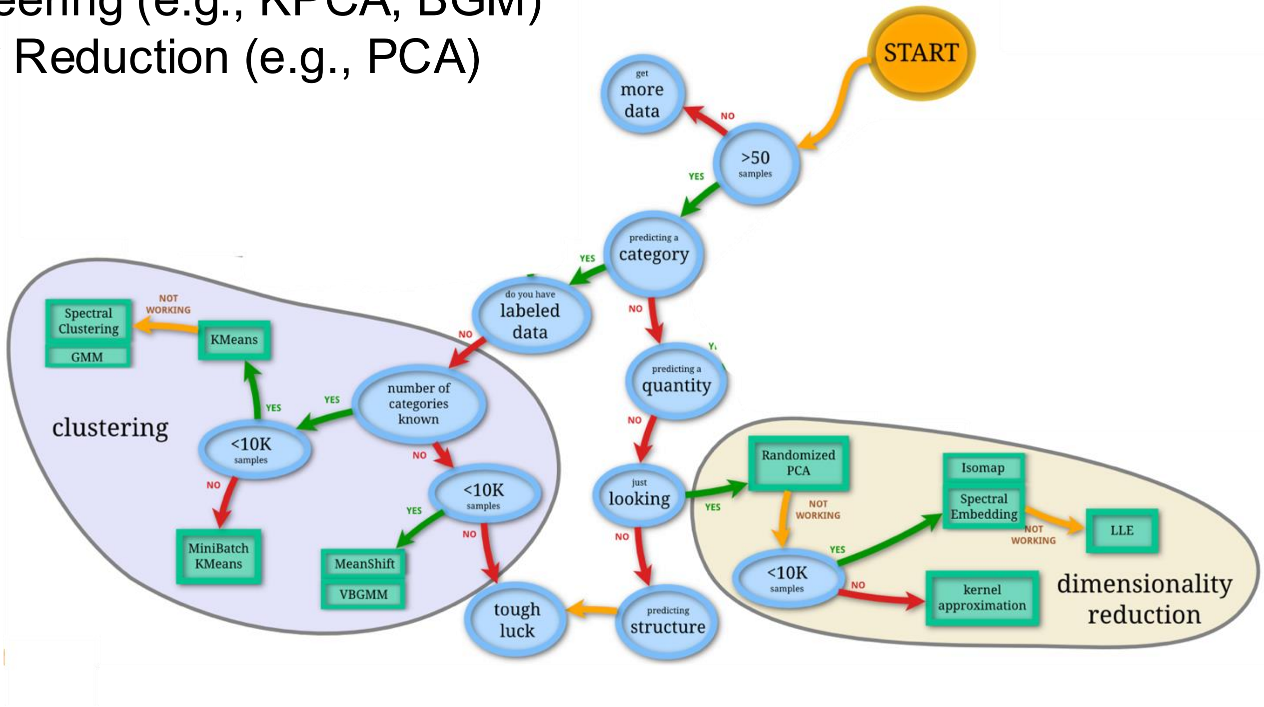
Supervised

Generative

Discriminative

Unsupervised Learning Techniques

- Clustering (e.g., K-means)
- Feature Engineering (e.g., KPCA, BGM)
- Dimensionality Reduction (e.g., PCA)



http://scikit-learn.org/stable/tutorial/machine_learning_map/

Clustering

Clustering refers to a very broad set of techniques for finding subgroups, or clusters, in a data set, such that those within each cluster are more closely related to one another than objects assigned to different clusters.

$K = 19$ clustering of RGB image of aurora ($D = 3$)

K-means Clustering

- The objective is to assign each observation, uniquely labeled by an integer $n \in \{1, \cdots, N\}$, to one and one only cluster $\{\mathcal{C}_1, \cdots, \mathcal{C}_K\}$.
  - The total number of clusters is fixed ($K < N$).
  - The number of data points in the k-th cluster is N_k
- The K-means algorithm is an iterative method that minimizes is the within-cluster variation $W(\mathcal{C}_k)$:

$$\mathcal{C}_1 \cdots \mathcal{C}_K = \operatorname{argmin} \sum_{k=1}^{K} W(\mathcal{C}_k)$$

K-means Clustering Algorithm

- The K-means clustering algorithm can be implemented as the minimization of the objective function

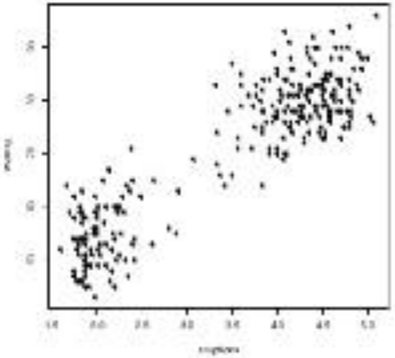
$$\mathcal{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \left\| \mathbf{x}_n - \boldsymbol{\mu}_k \right\|^2$$

where $r_{nk} \in \{0, 1\}$ is a binary indicator variable that describes which of the K clusters the data point $\mathbf{x}_n$ is assigned to.

1. First we choose some initial values for cluster centroids $\boldsymbol{\mu}_k$.
2. Iterate until the cluster assignments stop changing:
   - Minimize $\mathcal{J}$ with respect to r_{nk}, keeping $\boldsymbol{\mu}_k$ fixed.
   - Minimize $\mathcal{J}$ with respect to $\boldsymbol{\mu}_k$, keeping r_{nk} fixed.

Example: Old Faithful Data: K-means

$N = 272$ and $D = 2$



Old Faithful - Yellowstone National Park

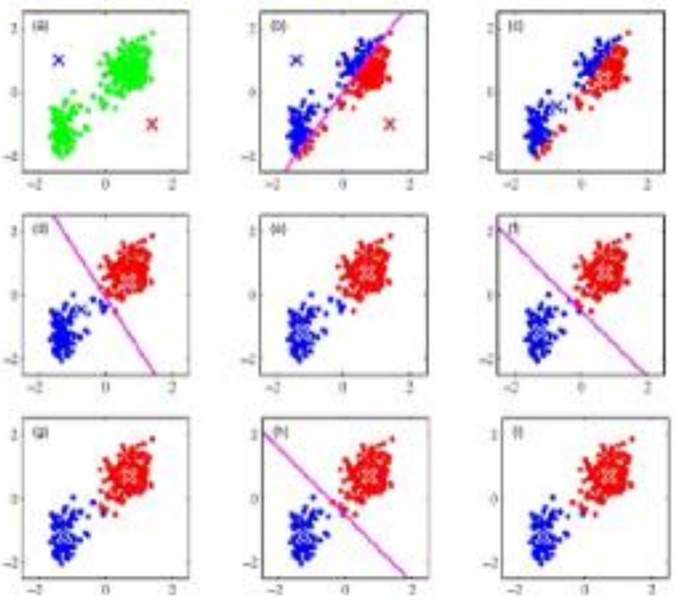


Figure 9.1 from Bishop

Example: Old Faithful Data: K-means

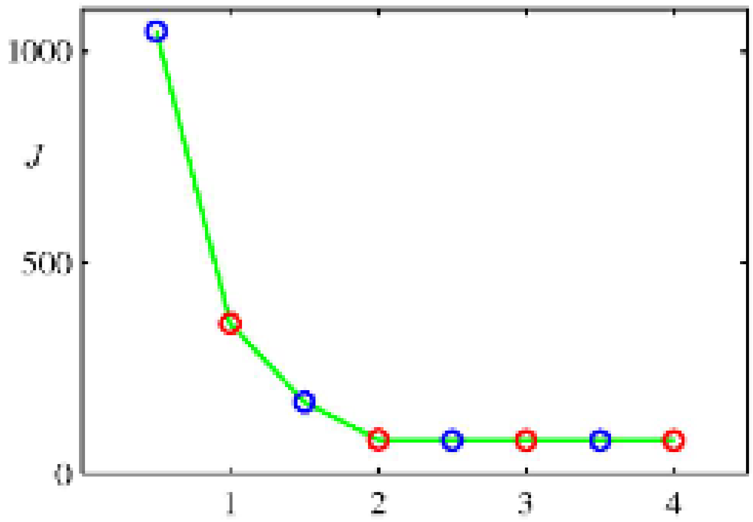


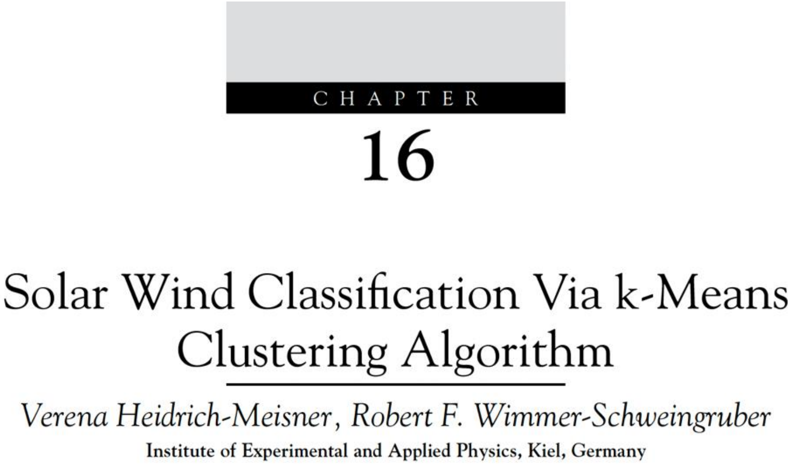
Figure 9.2 from Bishop

K-means Clustering Recap

- The within-cluster variation decreases with each iteration of the algorithm.
- The K-means algorithm finds a local rather than a global optimum, the results obtained will depend on the initial (random) cluster assignment.
- The problem of selecting K is far from simple.

Solar Wind Classification Example I

Learn solar wind classification rules from 10 years of ACE data from 2001 to 2010 by using K-means clustering [Heidrich-Meisner and Wimmer-Schweingruber, 2018]

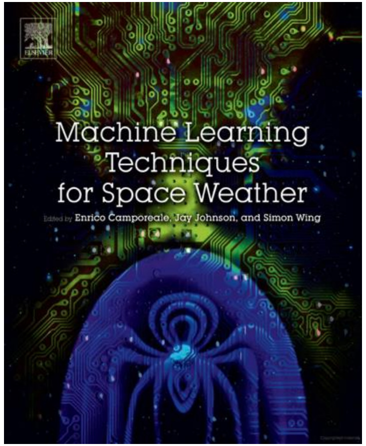

CHAPTER

16

Solar Wind Classification Via k-Means Clustering Algorithm

Verena Heidrich-Meisner, Robert F. Wimmer-Schweingruber

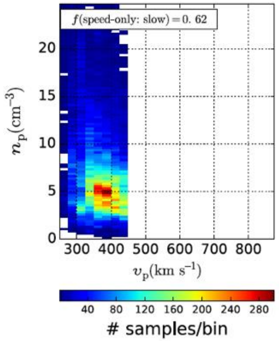
Institute of Experimental and Applied Physics, Kiel, Germany



Solar Wind Classification Example I

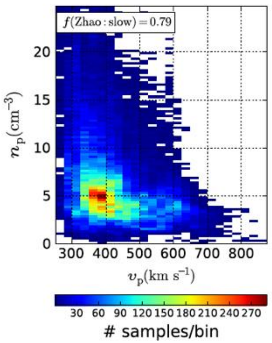
Learn solar wind classification rules from 10 years of ACE data from 2001 to 2010 by using K-means clustering [Heidrich-Meisner and Wimmer-Schweingruber, 2018]

"Slow" solar wind in n_p-v_p plane based on "Heuristic" domain-knowledge based decision rules



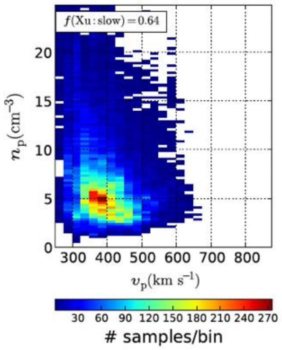
Speed-Only

Proton solar wind speed: v_p



Zhao and Fisk (2010)

O and C charge state ratios: $\frac{n_{O7+}}{n_{O6+}}, \frac{n_{C6+}}{n_{C5+}}$



Xu and Borovsky (2015)

Alfvén speed: v_A, proton-specific entropy: S_p, ratio of expected proton temperature T_{exp}, and observed proton temperature T_p

Heidrich-Meisner and Wimmer-Schweingruber, 2018

Feature Selection

In machine learning, features can be directly observed parameters (raw features) and/or engineered features

raw features

Feature Space	Dim.	Solar Wind Parameters
F_{basic}	7	$v_{\mathrm{p}}, n_{\mathrm{p}}, T_{\mathrm{p}}, B, a_{\mathrm{c}}, n_{\mathrm{O}^{7+}}/n_{\mathrm{O}^{6+}}, n_{\mathrm{C}^{6+}}/n_{\mathrm{C}^{5+}}$

Heidrich-Meisner and Wimmer-Schweingruber, 2018

Select K Using Cross Validation

- In m-fold cross validation, the data are randomly divided into m groups (or folds) of approximately equal size

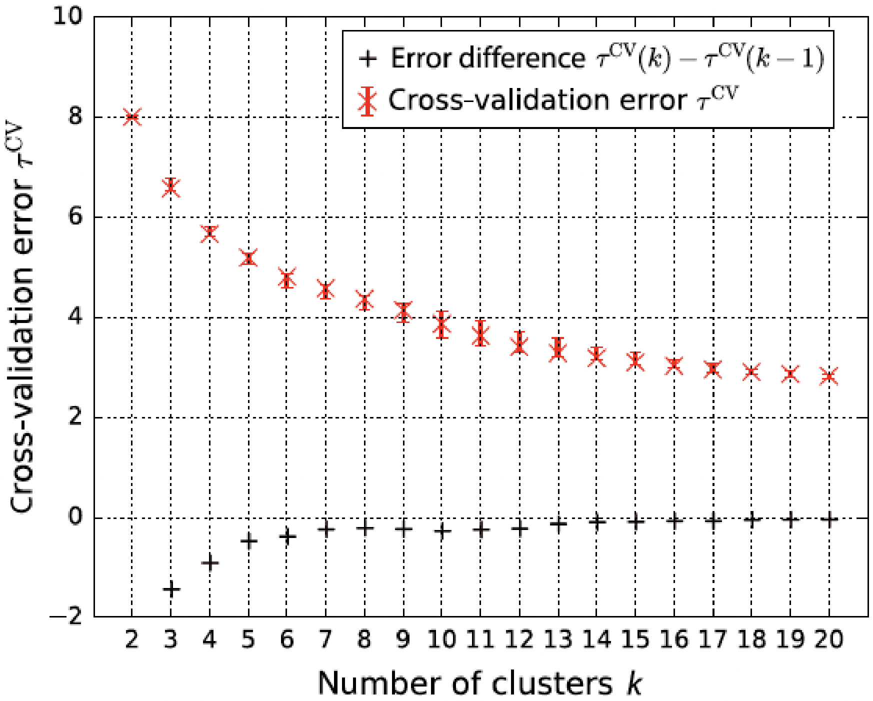
5-fold example

James, 2021

- The one fold is treated as a validation set, and the clustering rule (e.g., $\hat{C}_k$, $\hat{\mu}_k$) is learned from the remaining $m-1$ folds.
- The cross validation estimate error is

$$\tau^{CV}(K) = \frac{1}{M} \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{n_m \in \hat{C}_k} \|\mathbf{x}_{n_m} - \hat{\mu}_k\|^2$$

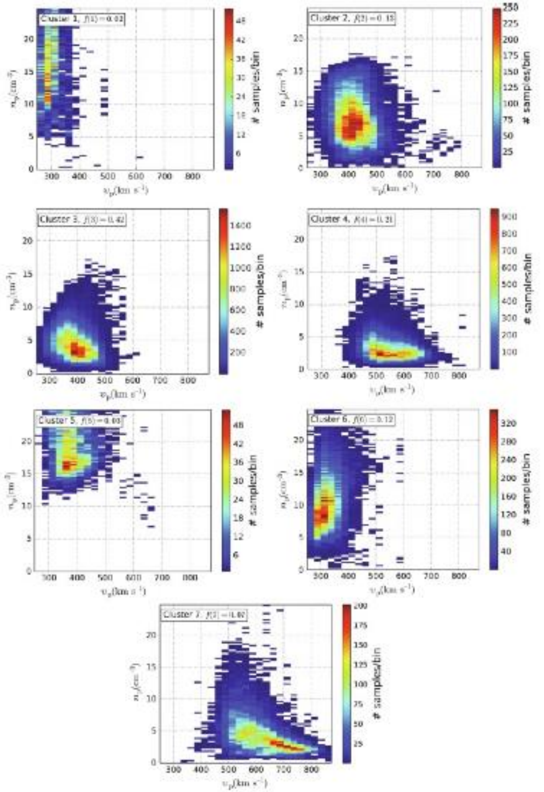
10-fold Cross Validation of K-means Clustering



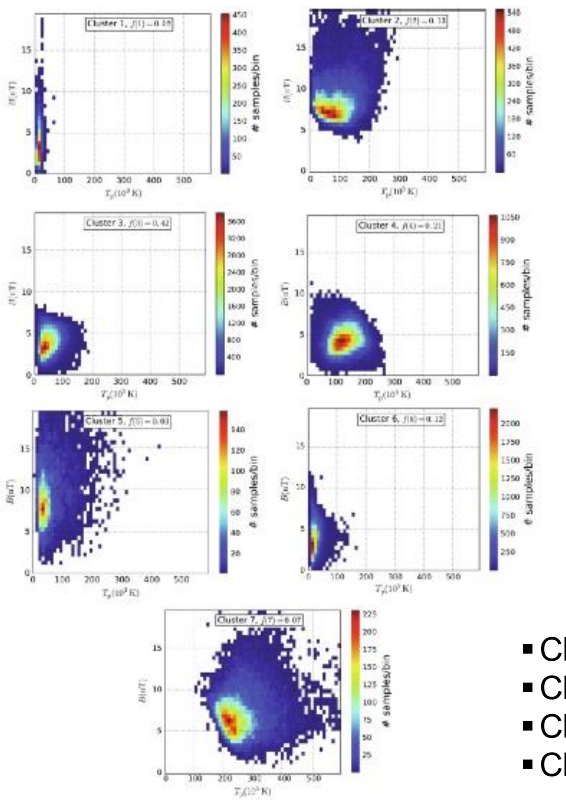
Heidrich-Meisner and Wimmer-Schweingruber, 2018

K=7 clustering of 10-year solar wind data (D=7)

n_p-v_p plane



B-T_p plane

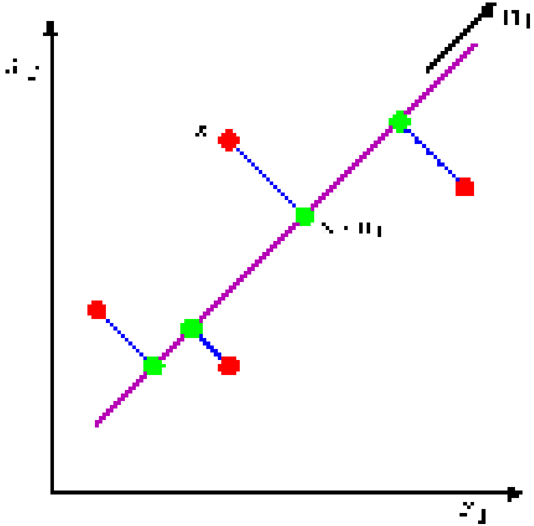


- Cluster 3 - the bulk of slow solar wind
- Clusters 4 & 7 - coronal hole wind
- Clusters 2 & 6 - interstream solar wind
- Clusters 1 & 5 - ?

Heidrich-Meisner and Wimmer-Schweingruber, 2018

Principal Component Analysis

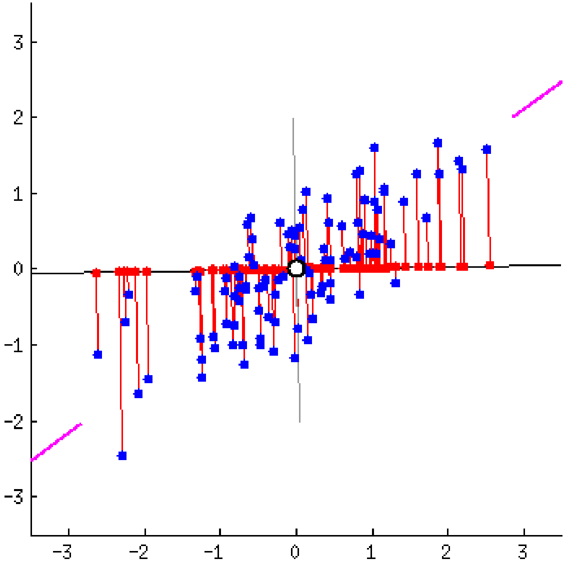
- PCA is defined as the orthogonal projection of the data onto a lower dimensional linear space such that the variance of the projected data is maximized.



Bishop, 2006

Principal Component Analysis

- Learning *principal components* $\mathbf{u}$ from the data is a constrained optimization problem.



Stack Exchange

PCA in Practice: Eigenvalue Problem

- Find projections $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_D]$ that maximizes the variance of the data:

$$\begin{aligned} \texttt{maximize}: &\quad \mathrm{var}(\mathbf{XU}) = \mathrm{tr}(\mathbf{U}^T\mathbf{X}^T\mathbf{XU}) = \mathrm{tr}(\mathbf{U}^T\mathbf{C}_{xx}\mathbf{U}) \\ \texttt{subject to}: &\quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \end{aligned}$$

where $\mathbf{X} \in \mathbb{R}^{N\times D}$, $\mathbf{U} \in \mathbb{R}^{D\times D}$, and $\mathbf{C}_{xx} \in \mathbb{R}^{D\times D}$

- This problem is effectively solved by finding eigenvectors of a covariance matrix $\mathbf{C}_{xx}$

$$\mathbf{C}_{xx}\mathbf{u} = \lambda\mathbf{u}$$

Learning Principal Components From Data

- Compute $\mathbf{C}_{xx}$ from the data

$$\mathbf{C}_{xx} \approx \frac{1}{N-1} \sum_{n=1}^{N} (\boldsymbol{x}_n - \overline{\boldsymbol{x}})^T (\boldsymbol{x}_n - \overline{\boldsymbol{x}}) \quad \overline{\boldsymbol{x}} = \frac{1}{N} \sum \boldsymbol{x}_n$$

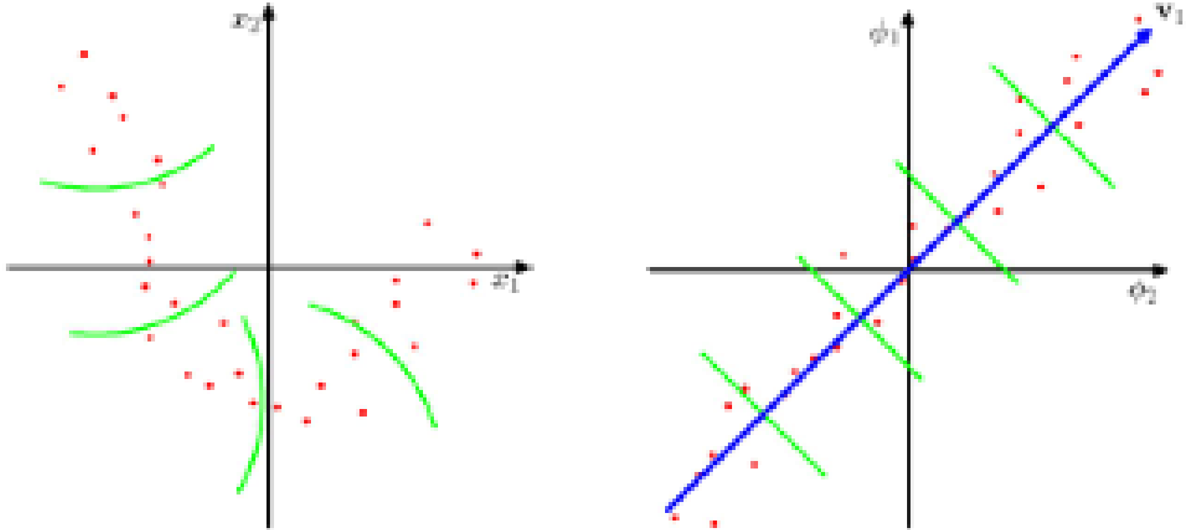
- Note that for $D \ll N$ case, $\mathbf{C}_{xx}$ is not uniquely determined
- Apply the eigenvalue decomposition to $\mathbf{C}_{xx}$

The Matlab PCA code:

```
>> C = cov(X);
>> [U L] = eigs(C,d);
>> Xtest_projected = Xtest*U;
```

Kernel PCA: Nonlinear Version of PCA

- In KPCA, the data are mapped to a higher-dimensional *feature space* where data can be linearly separable.



Bishop, 2006

PCA vs. Kernel PCA

- PCA finds projections maximizing the variance of the data

$$\begin{aligned} \texttt{maximize}: &\quad \mathrm{var}(\mathbf{XU}) = \mathrm{tr}(\mathbf{U}^T\mathbf{X}^T\mathbf{XU}) \\ \texttt{subject to}: &\quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \end{aligned}$$

- KPCA finds projections maximizing the variance of the *mapped* data

$$\begin{aligned} \texttt{maximize}: &\quad \mathrm{var}(\mathbf{\Phi U}) = \mathrm{tr}(\mathbf{U}^T\mathbf{\Phi}^T\mathbf{\Phi U}) \\ \texttt{subject to}: &\quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \end{aligned}$$

where data are mapped to a higher-dimensional *feature space*:

$$\mathbf{\Phi} = [\phi(\boldsymbol{x}_1), \cdots, \phi(\boldsymbol{x}_N)]$$

Kernel PCA Solved with the Kernel Trick

- KPCA finds projections maximizing the variance of the *mapped* data

$$\begin{aligned} \texttt{maximize}: &\quad \mathrm{var}(\mathbf{\Phi U}) = \mathrm{tr}(\mathbf{U}^T \mathbf{\Phi}^T \mathbf{\Phi U}) \\ \texttt{subject to}: &\quad \mathbf{U}^T\mathbf{U} = \mathbf{I} \end{aligned}$$

- Use the projection relationship $\mathbf{U} = \mathbf{\Phi}^T\mathbf{A}$ and a N-by-N kernel matrix $\mathbf{K} = \mathbf{\Phi\Phi}^T$
- The above KPCA problem becomes

$$\begin{aligned} \texttt{maximize}: &\quad \mathrm{tr}(\mathbf{A}^T\mathbf{KKA}) \\ \texttt{subject to}: &\quad \mathbf{A}^T\mathbf{KA} = \mathbf{I} \end{aligned}$$

- This circumvents explicitly mapping data into a feature space, which is computationally expensive and impractical (e.g., kernel trick).

Solar Wind Classification Example II

Learn solar wind classification rules from 14 years of ACE data from 1998 to 2011 by using unsupervised methods (e.g., PCA, clustering, BGM, more) [Amaya et al. 2020]

frontiers in Astronomy and Space Sciences

ORIGINAL RESEARCH
published: 25 September 2020
doi: 10.3389/fspas.2020.553207

Visualizing and Interpreting Unsupervised Solar Wind Classifications

Jorge Amaya, Romain Dupuis, Maria Elena Innocenti and Giovanni Lapenta*

Mathematics Department, Centre for Mathematical Plasma-Astrophysics, KU Leuven, Leuven, Belgium

Solar Wind Classification Example II

Learn solar wind classification rules from 14 years of ACE data from 1998 to 2011 by using unsupervised methods (e.g., PCA, clustering, BGM, more) [Amaya et al. 2020]

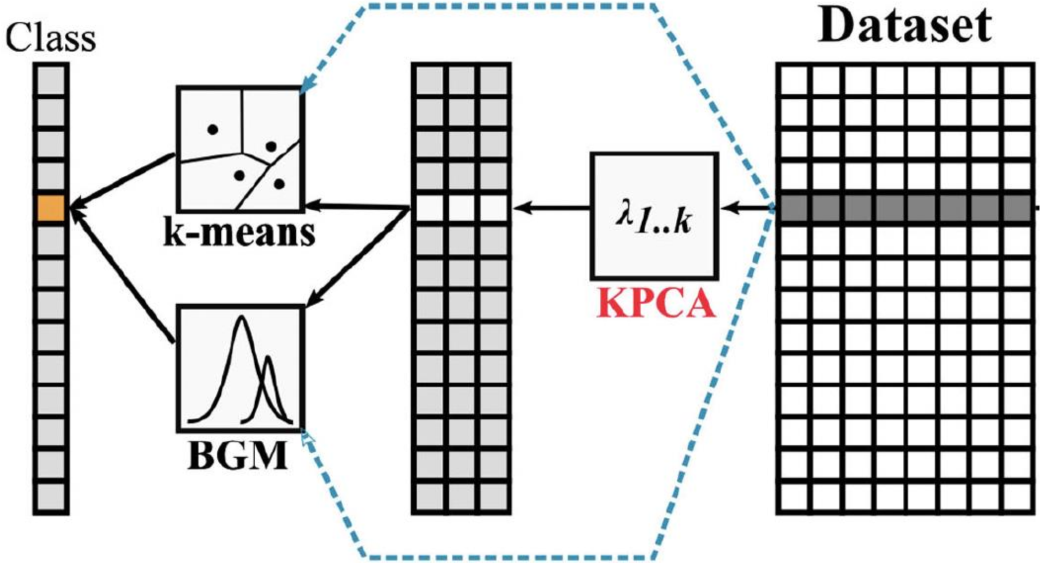
raw features

Dim.	Solar Wind Parameters	
15	proton_speed	Sp
	proton_density	Va
	O7to6	Tratio
	C6to5	proton_speed_range
	FetoO	Bn_range
	avqFe	FetoO_range
	proton_temp	O7to6_range
	sigmac(*)	

Amaya et al. 2020

PCA Feature Engineering Before K-means/BGM

Heidrich-Meisner and Wimmer-Schweingruber 2018
Roberts et al., 2020

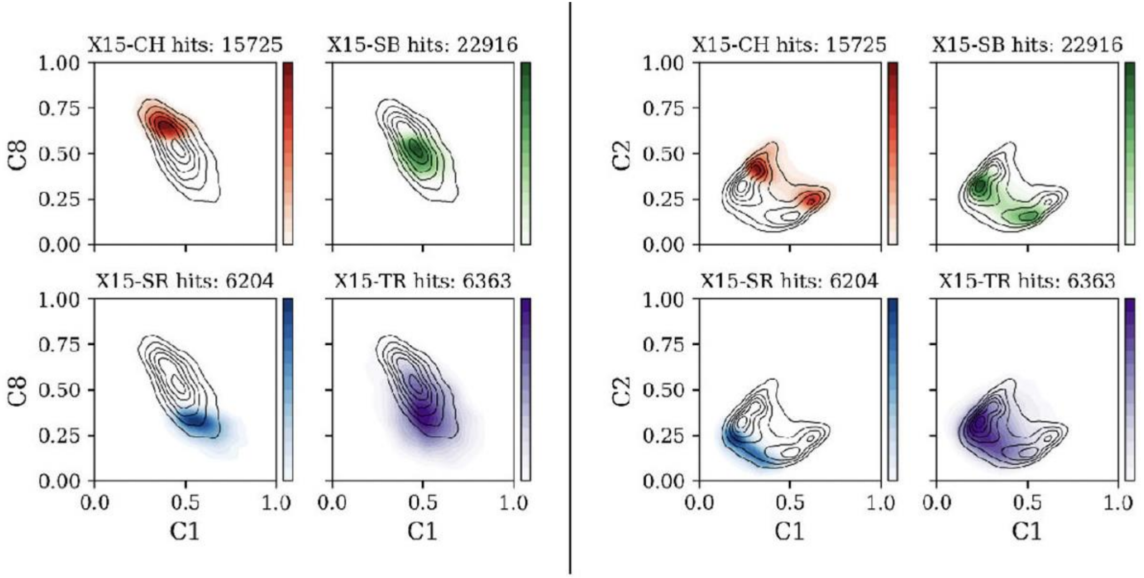


Bloch et al., 2020

Amaya et al., 2020

Data Points Projected on Principal Components

PCA KPCA



Color coded by 4 categories

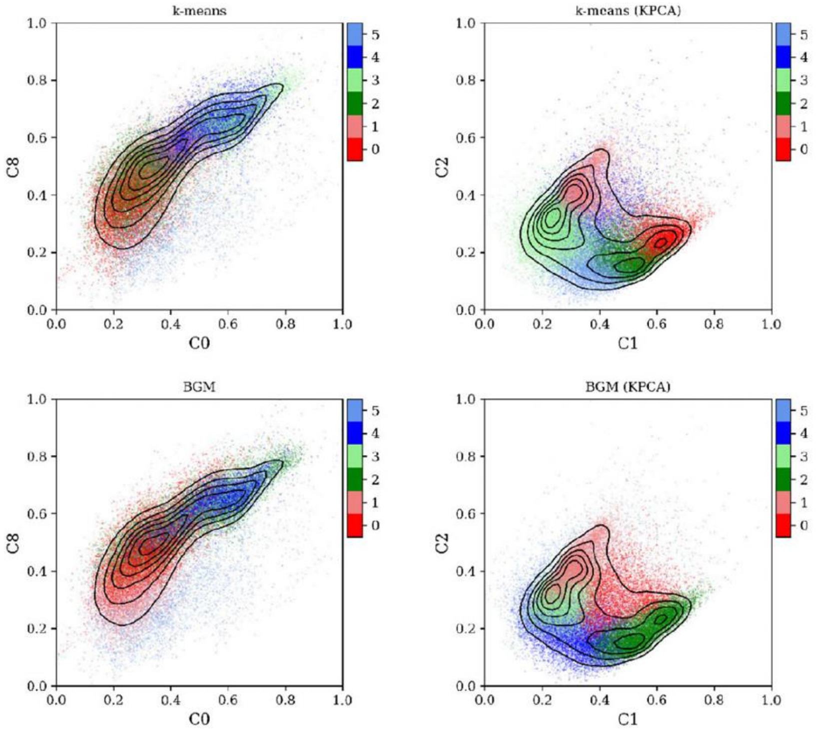
- CH: fast solar wind of coronal hole origin
- TR: Transients (e.g., ejecta, ICMEs, CIRs, MCs)
- SB: Solar wind originated in the streamer belt
- SR: Solar wind of sector reversal origins

Xu and Borovsky (2015)

Alfvén speed: v_A, proton-specific entropy: S_p, ratio of expected proton temperature T_{exp}, and observed proton temperature T_p

Amaya et al., 2020

K=5 clustering of 14-year solar wind data (D=15)

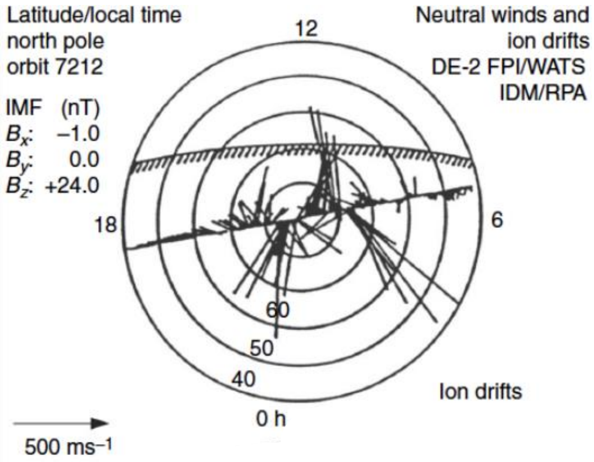


Amaya et al., 2020

Ionospheric Convection PCA Example

Learn principal components from DE-2 plasma drift data from 1981 to 1983 to characterize the spatiotemporal variability of convection patterns [Matsuo et al. 2002]

E x **B** plasma drift vector measurements from 650 passes



Matsuo, Richmond, and Nychka , 2002

Karhunen-Loéve Expansion

By Karhunen–Loéve expansion, a stochastic process $x(\mathbf{r})$ can be expanded as a sequence of independent random variables as

$$x(\mathbf{r}) = \sum_{i=1}^{\infty} \alpha_i \Psi_i(\mathbf{r})$$

- $\Psi_i(\mathbf{r})$ is a fixed orthonormal basis (eigenfunction of $\mathrm{cov}(x, x)$).
- α_i is an independent random variable $\alpha_i \sim (0, \lambda_i)$.
- λ_i is the eigenvalue of $\mathrm{cov}(x, x)$.

Random Fields Model /w Karhunen-Loéve Basis

$x(\mathbf{r}, t)$ can then be expanded on principal components as

$$x(\mathbf{r}, t) = \alpha_1(t)\Psi_1(\mathbf{r}) + \alpha_2(t)\Psi_2(\mathbf{r}) + \alpha_3(t)\Psi_3(\mathbf{r}) + \cdots$$

so that $\mathrm{cov}(x, x) = \mathbf{\Psi}\mathbb{E}[\alpha\alpha^T]\mathbf{\Psi}^T$ where $\mathbb{E}[\alpha\alpha^T] = \mathrm{diag}(\lambda)$

PCA Solved as Sequential Nonlinear Regression

$x(\mathbf{r}, t)$ can then be expanded on principal components as

$$x(\mathbf{r}, t) = \alpha_1(t)\Psi_1(\mathbf{r}) + \alpha_2(t)\Psi_2(\mathbf{r}) + \alpha_3(t)\Psi_3(\mathbf{r}) + \cdots$$

so that $\mathrm{cov}(x, x) = \mathbf{\Psi}\mathbb{E}[\alpha\alpha^T]\mathbf{\Psi}^T$ where $\mathbb{E}[\alpha\alpha^T] = \mathrm{diag}(\lambda)$

- Estimation of $\mathbf{\Psi}$ and λ directly from *sparse* data is not feasible
- Projecting $\mathbf{\Psi}$ onto predefined orthonormal basis,

$$x(\mathbf{r}, t) \approx \alpha_1(t)\mathbf{\Phi}(\mathbf{r})\beta^{(1)} + \alpha_2(t)\mathbf{\Phi}(\mathbf{r})\beta^{(2)} + \cdots$$

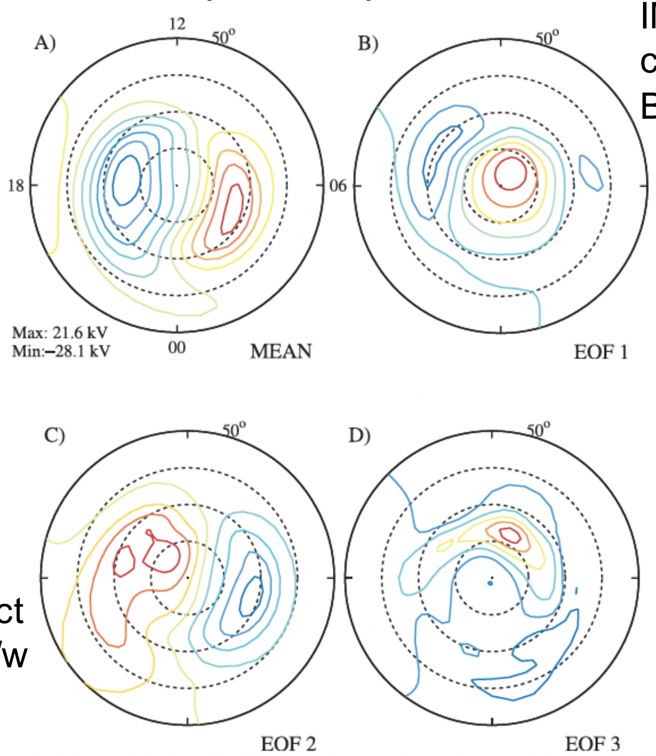
$$\Psi_i(\mathbf{r}) = \sum_{j=1}^{J} \beta_j^{(i)}\Phi_j(\mathbf{r})$$

- Learn $\alpha_i(t)$ and $\beta^{(i)}$ from the data by backfitting (e.g., nonlinear sequential regression, iterative algorithm used in statistics to fit additive models)

Matsuo, Richmond and Nychka, 2002

Leading Principal Components and Variance

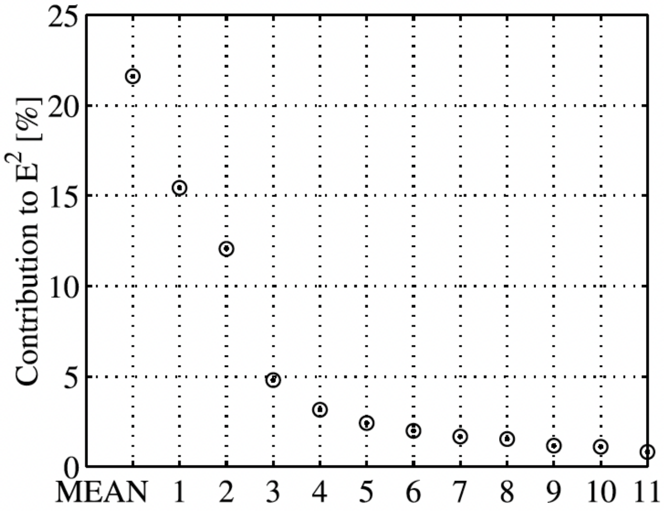
Principal Components



IMF B_Y effect correlation /w B_Y -0.35

IMF B_Z effect correlation /w B_Z 0.54

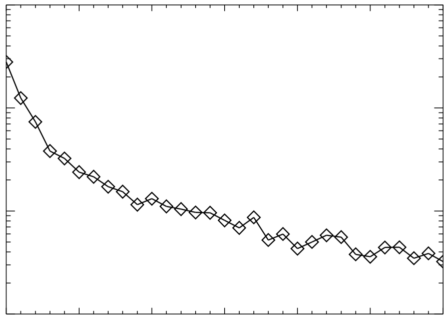
Variance

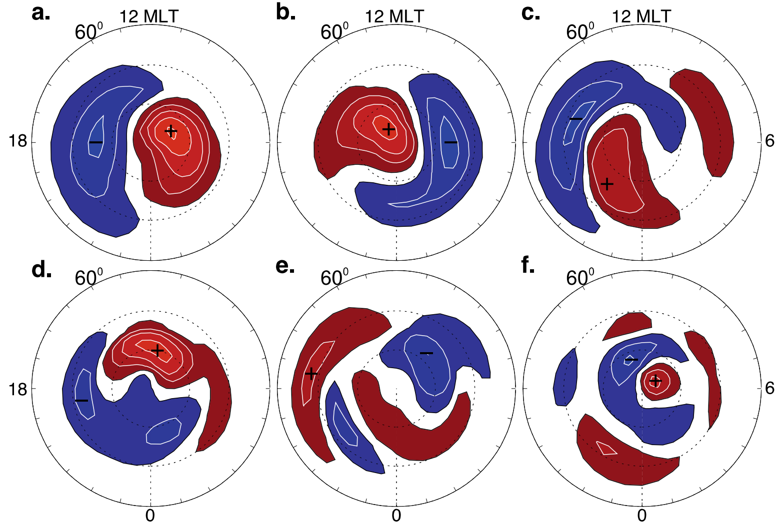


Matsuo, Richmond, and Nychka , 2002

Ionospheric Convection PCA Example II

Learn principal components from SuperDARN plasma drift data from 2011 to 2012 to characterize the spatiotemporal variability of convection patterns [Cousins et al. 2013]





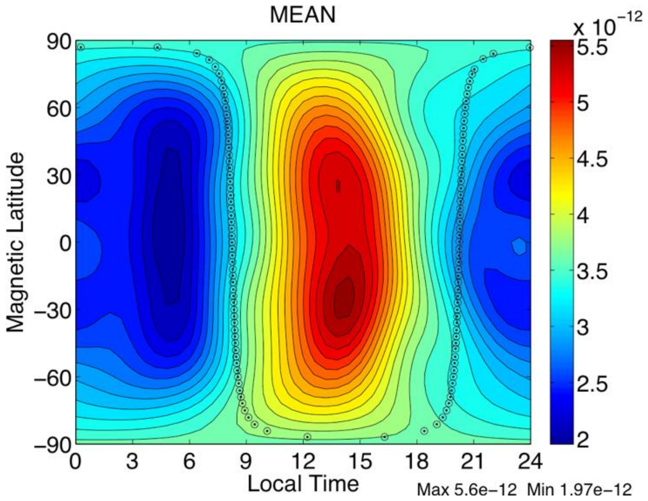
EOF	1	2	3	4	5	6
Mode	Strengthening	Shaping	Expanding	Rotating	—	—
Contribution to E^2	12%	7%	4%	3%	2%	2%
Top correlation	AE: 0.44	B_y: –0.38	AE: 0.42	B_z: –0.20	AE: –0.14	Tilt: 0.08
Second correlation	B_z: –0.42	B_z: –0.36	Kp: 0.35	AE: 0.16	Kp: –0.14	B_x: 0.03
Third correlation	B_y: –0.41	AE: –0.35	$\|V_x\|$: 0.27	Kp: 0.13	$\|V_x\|$: –0.08	σB_x : 0.03[b]

Cousins et al., 2013

Thermospheric Mass Density PCA Example

Learn principal components from CHAMP data from 2001 to 2008 to characterize the spatiotemporal variability of thermospheric mass density [Matsuo and Forbes, 2008]

Mass density measurements normalized to 400km

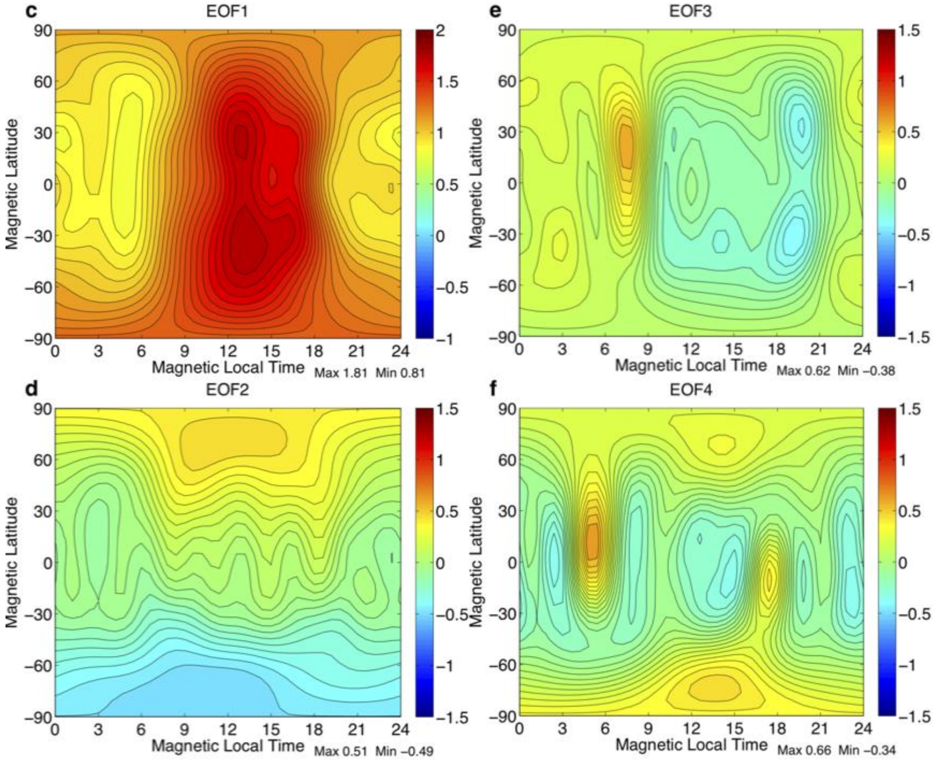


Matsuo and Forbes, 2010

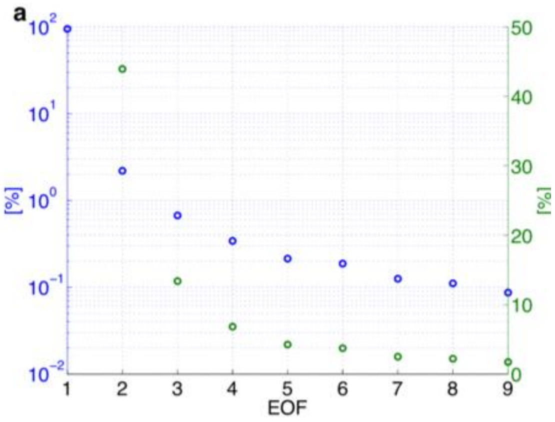
Leading Principal Components and Variance

Solar EUV effect correlation /w F10.7 0.91

Principal Components

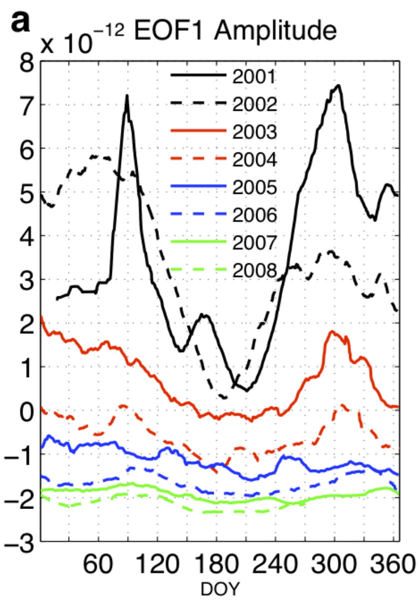


Variance

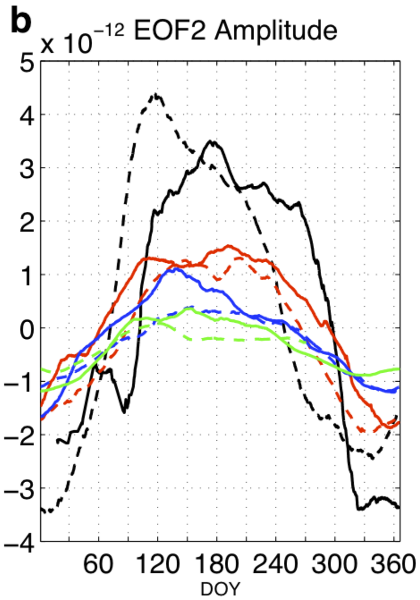


Matsuo and Forbes, 2010

Temporal Variation of Leading Modes

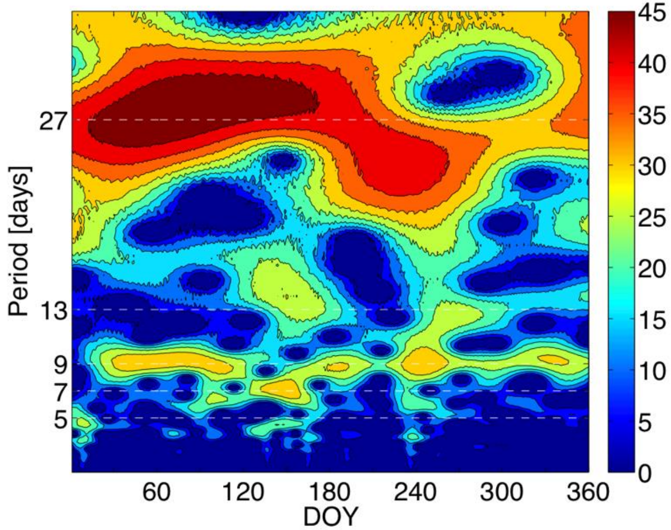


semiannual variation



annual variation

periodgram for 2005



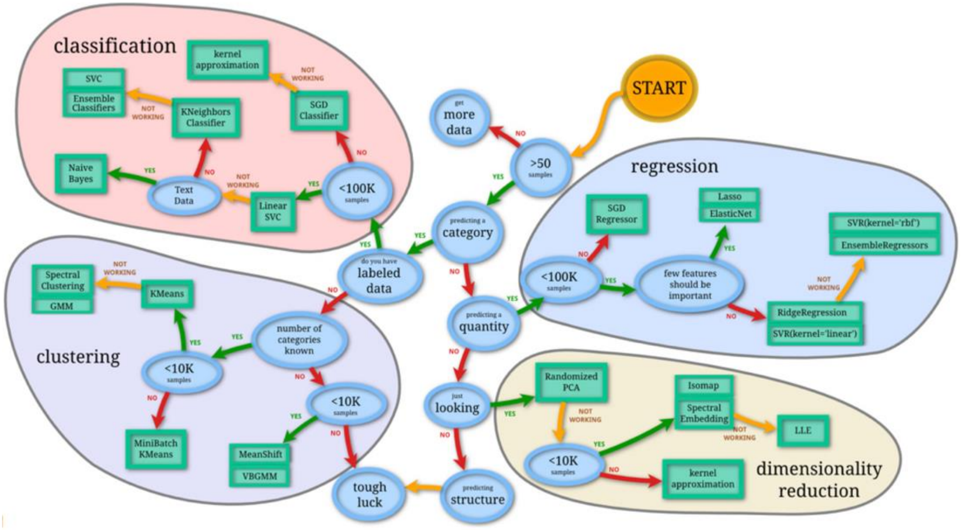
9-day oscillations due to recurrent geomagnetic activity driven by corotating high-speed streams

Matsuo and Forbes, 2010

Summary

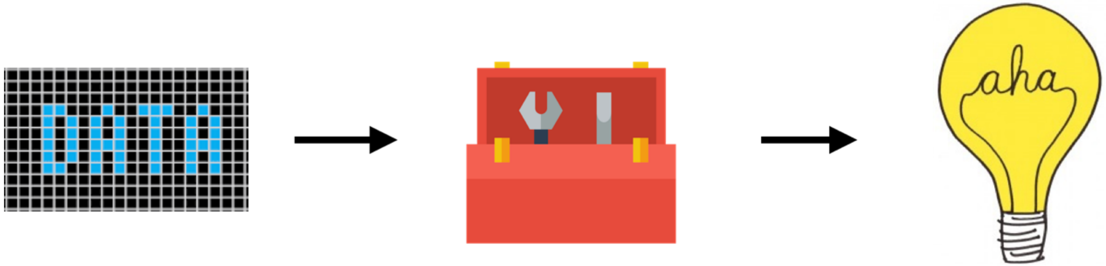
Unsupervised Machine Learning Data Analysis Tools help *you*

- Examine and understand a dataset
- Manage large volumes of data by dimensionality reduction
- Extract characteristic patterns and trends from data
- Identify the way data are naturally grouped together



http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

Ready for *your* data-driven Heliophysics science discovery?



References

Heidrich-Meisner, V., and Wimmer-Schweingruber, R. F. (2018). “Chapter 16 solar wind classification via k-means clustering algorithm,” in Machine Learning Techniques for Space Weather, eds E. Camporeale, S. Wing, and J. R. Johnson (Elsevier), 397–424.

Amaya J, Dupuis R, Innocenti ME and Lapenta G (2020) Visualizing and Interpreting Unsupervised Solar Wind Classifications. Front. Astron. Space Sci. 7:553207. doi: 10.3389/fspas.2020.553207.

Matsuo, T., A. D. Richmond, and D. W. Nychka (2002), Modes of high-latitude electric field variability derived from DE-2 measurements: Emperical Orthogonal Function (EOF) analysis, Geophys. Res. Lett., 29(7), 1107, doi:10.1029/2001GL014077.

Matsuo, T., and J. M. Forbes (2010), Principal modes of thermospheric density variability: Empirical orthogonal function analysis of CHAMP 2001–2008 data, J. Geophys. Res., 115, A07309, doi:10.1029/2009JA015109

Cousins, E. D. P., T. Matsuo, and A. D. Richmond (2013), Mesoscale and large-scale variability in high-latitude ionospheric convection: Dominant modes and spatial/temporal coherence, J. Geophys. Res. Space Physics, 118, 7895–7904, doi:10.1002/2013JA019319.