Lectures in AstroStatistics: Topics in Machine Learning for Astronomers

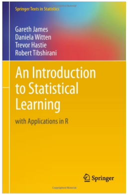
Jessi Cisewski
Yale University

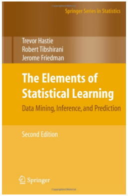
American Astronomical Society Meeting
Wednesday, January 6, 2016

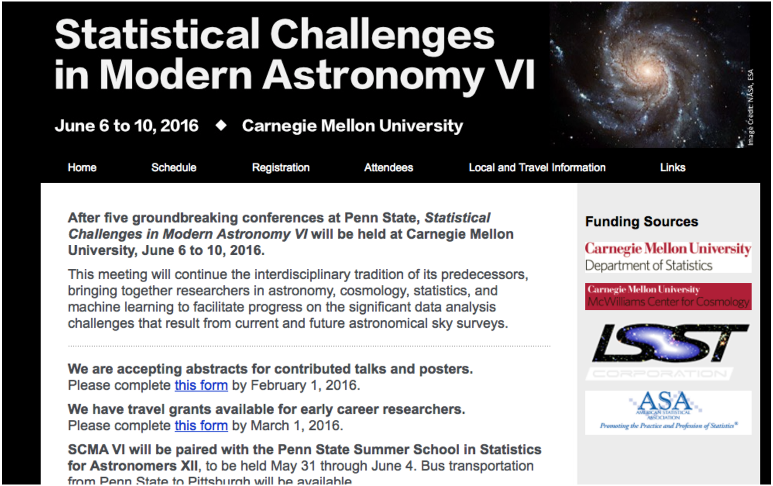
Statistical Learning - learning from data

We'll discuss some methods for **classification** and **clustering** today.

Good references:







Co-Chairs: Shirley Ho (CMU, Cosmology) and Chad Schafer (CMU, Statistics)

More info at `http://www.scma6.org`

Statistical and Applied Mathematical Sciences Institute (SAMSI) 2016-17

Program on Statistical, Mathematical and Computational Methods for Astronomy (ASTRO)

- Opening Workshop: August 22 - 26, 2016
- Current list of proposed Working Groups
    1. Uncertainty Quantification and Reduced Order Modeling in Gravitation, Astrophysics, and Cosmology
    2. Synoptic Time Domain Surveys
    3. Time Series Analysis for Exoplanets & Gravitational Waves: Beyond Stationary Gaussian Processes
    4. Population Modeling & Signal Separation for Exoplanets & Gravitational Waves
    5. Statistics, computation, and modeling in cosmology

More info at `http://www.samsi.info/workshop/2016-17-astronomy-opening-workshop-august-22-26-2016`

Classification

Use *a priori* group labels in analysis to assign new observations to a particular group or class

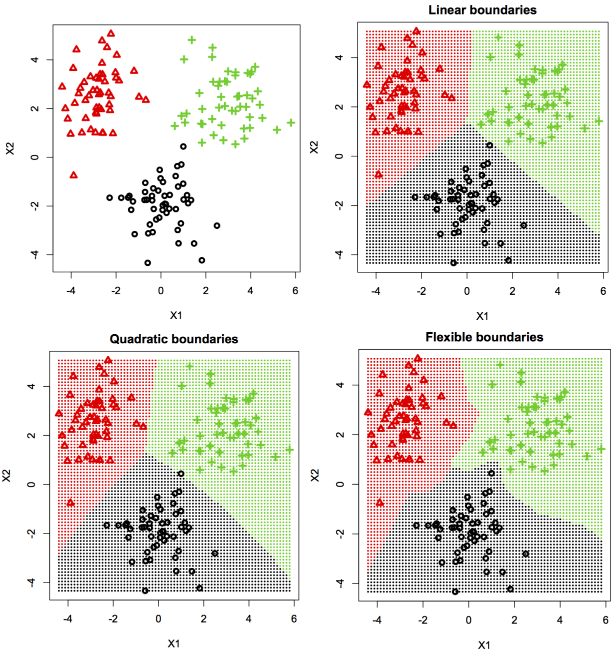
$\longrightarrow$ "Supervised learning" or "Learning with labels"

Data: $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\} \in \mathbb{R}^p$, labels $\mathbf{Y} = \{y_1, y_2, \ldots, y_n\}$

Stars can be classified into labels $\mathbf{Y} = \{O, B, F, G, K, M, L, T, Y\}$

Using features $\mathbf{X} = \{\text{Temperature}, \text{Mass}, \text{Hydrogen lines}, \ldots\}$

Classification rules



Classification: evaluating performance

Training error rate: number of misclassified observations over sample of size n is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(\hat{y}_i \neq y)$$

where $\hat{y}_i$ is the predicted class for observation i, and $\mathbb{I}$ is the indicator function.

- The **test error rate** is more important than training error; can estimate using cross-validation
- Class imbalance - strong imbalance in the number of observations in the classes can result in misleading performance measures

Bayes Classification Rule

Test error is minimized by assigning observations with predictors x to the class that has the largest probability:

$$\underset{j}{\operatorname{argmax}} \, P(Y = j \mid X = x)$$

for classes $j = 1, \ldots, J$

- In general, intractable because the distribution of $Y \mid X$ is unknown.

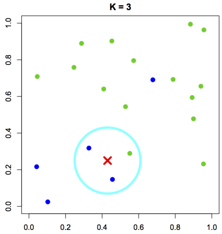
K Nearest Neighbors (KNN)

Main idea: An observation is classified based on the K observations in the training set that are nearest to it

- A probability of each class can be estimated by

$$P(Y = j \mid X = x) = K^{-1} \sum_{i \in N(x)} \mathbf{I}(y_i = j)$$

where $j = 1, \ldots, \#$classes in training set, and $\mathbf{I} =$ indicator function.



- $K = 3$ nearest neighbors to the **X** are within the circle.
- The predicted class of **X** would be blue because there are more blue observations than green among the 3 NN.

Linear Classifiers

- Decision boundary is linear
- If $p = 2$ class boundary is a line
  ($p = 3$ is plane, $p > 3$ is hyperplane)

- Logistic regression
- Linear Discriminant Analysis
  (Quadratic Discriminant Analysis)

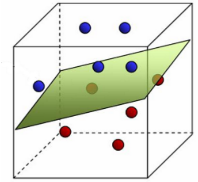
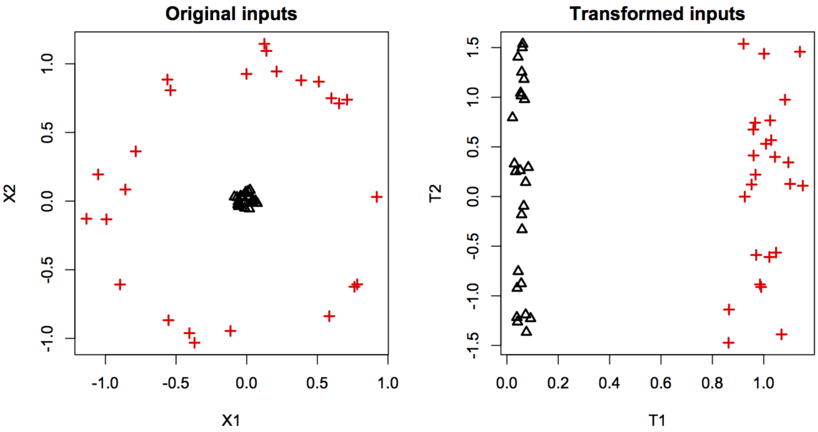


Image: `http://fouryears.eu/2009/02/`


Original inputs
X2
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
X1
Transformed inputs
T2
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
0.0
0.2
0.4
0.6
0.8
1.0
T1

Support Vector Machines

- Goal: Find the hyperplane that "best" separates the two classes (i.e. maximize the margin between the classes)
- If data are not linearly separable, can use the "**Kernel trick**" (transforms data to higher dimensional feature space)

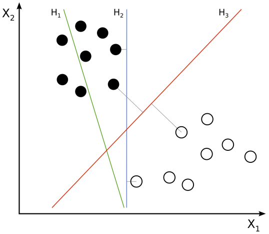
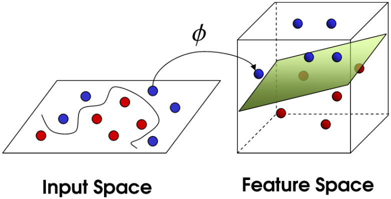


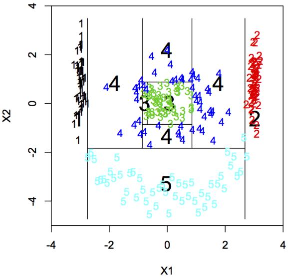
Image: `http://en.wikipedia.org`

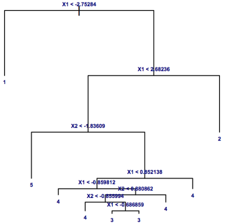


`http://stackoverflow.com/questions/9480605/`

Classification Trees

- CART = "Classification and Regression Trees"
  1. Predictor space is partitioned into hyper-rectangles
  2. Any observations in the hyper-rectangle would be predicted to have the same label
  3. Splits chosen to maximize "purity" of hyper-rectangles





Classification Trees - remarks

- Tree-based methods are not typically the best classification methods based on prediction accuracy, but they are often more easily interpreted (James et al. 2013)

- **Tree pruning** - the classification tree may be over fit, or too complex; pruning removes portions of the tree that are not useful for the classification goals of the tree.

- **Bootstrap aggregation** (aka "bagging") - there is a high variance in classification trees, and bagging (averaging over many trees) provides a means for variance reduction.

- **Random forest** - similar idea to bagging except it incorporates a step that helps to decorrelate the trees.

Clustering

Find subtypes or groups that are <u>not</u> defined *a priori* based on measurements

$\longrightarrow$ "Unsupervised learning" or "Learning without labels"

Data: $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n\} \in \mathbb{R}^p$

- Galaxy clustering
- Bump-hunting (e.g. statistically significant excess of gamma-rays emissions compared to background (Geringer-Sameth et al., 2015))

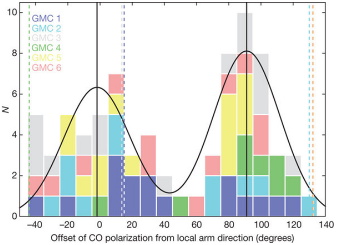
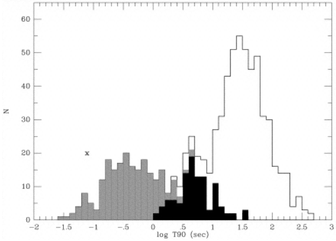


Image: Li and Henning (2011)



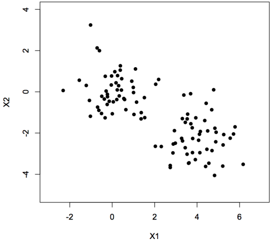
THREE TYPES OF GAMMA-RAY BURSTS

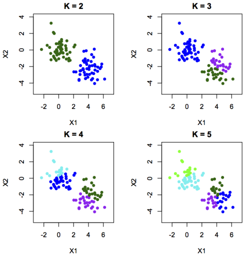
Soma Mukherjee,[1,2,3] Eric D. Feigelson,[4] Gutti Jogesh Babu,[5] Fionn Murtagh,[6,7] Chris Fraley,[8] and Adrian Raftery[8]

Received 1998 February 9; accepted 1998 June 25

K-means clustering

Main idea: partition observations into K separate clusters that do not overlap

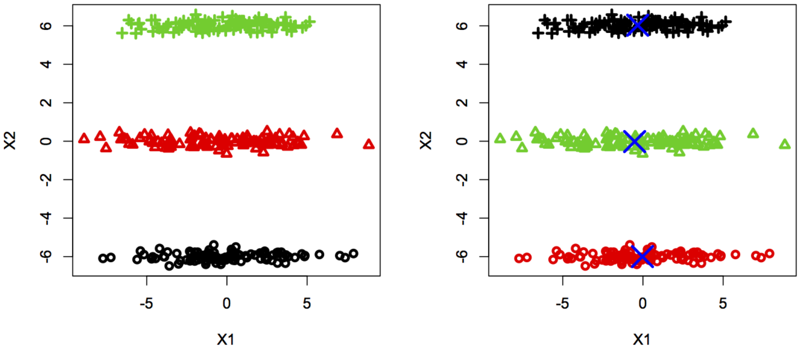


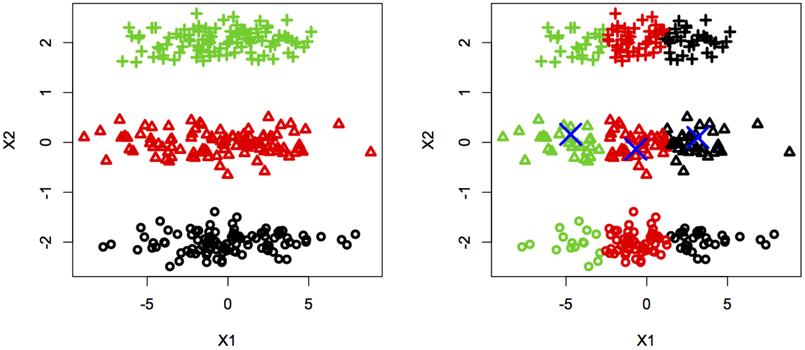


Goal: minimize total within-cluster scatter:

$$\sum_{k=1}^{K} |C_k| \sum_{C(i)=k} ||\mathbf{X}_i - \bar{\mathbf{X}}_k||^2$$

$|C_k|$ = number of observations in cluster C_k, $\bar{\mathbf{X}}_k = (\bar{X}_1^k, \ldots, \bar{X}_p^k)$


X2
6
4
2
0
-2
-4
-6
-5
0
5
X1
X2
6
4
2
0
-2
-4
-6
-5
0
5
X1


X2
2
1
0
-1
-2
-5
0
5
X1
X2
2
1
0
-1
-2
-5
0
5
X1

K-means clustering - comments

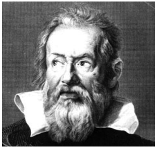
- Cluster assignments are strict $\longrightarrow$ no notion of degree or strength of cluster membership

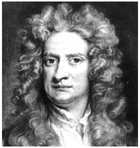
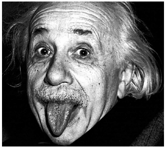
- Not robust to outliers

- Possible lack of interpretability of centers

  $\longrightarrow$ centers are averages:

  - what if observations are images of faces?

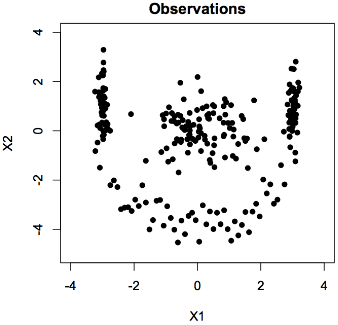


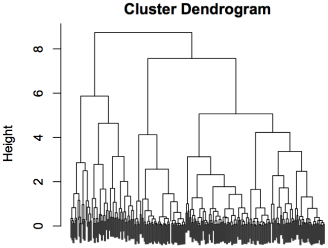



Images: http://cdn1.thefamouspeople.com,http://www.notablebiographies.com,http://mrnussbaum.com,http://3.bp.blogspot.com

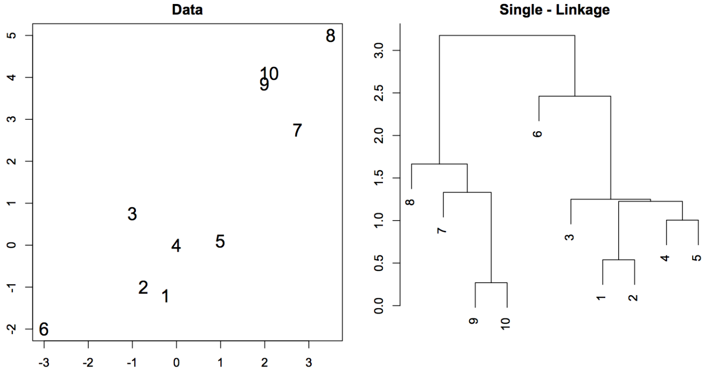
Hierarchical clustering

- Generates a hierarchy of partitions; user selects the partition
- $P_1 = 1$ cluster, ..., $P_n = n$ clusters (agglomerative clustering)
- Partition P_i is the union of one or more clusters from Partition P_{i+1}





Single-linkage clustering



Hierarchical clustering - distances

1. Single-linkage clustering: intergroup distance is smallest possible distance

$$d(C_k, C_{k'}) = \min_{x \in C_k, y \in C_{k'}} d(x, y)$$

2. Complete-linkage clustering: intergroup distance is largest possible distance

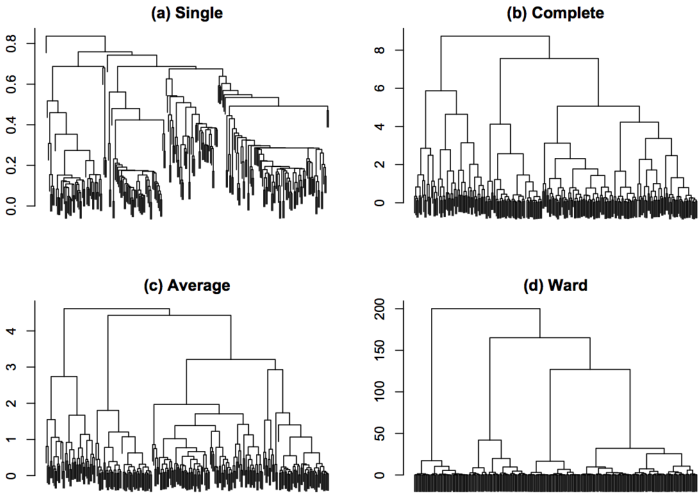
$$d(C_k, C_{k'}) = \max_{x \in C_k, y \in C_{k'}} d(x, y)$$

3. Average-linkage clustering: average intergroup distance

$$d(C_k, C_{k'}) = Ave_{x \in C_k, y \in C_{k'}} d(x, y)$$

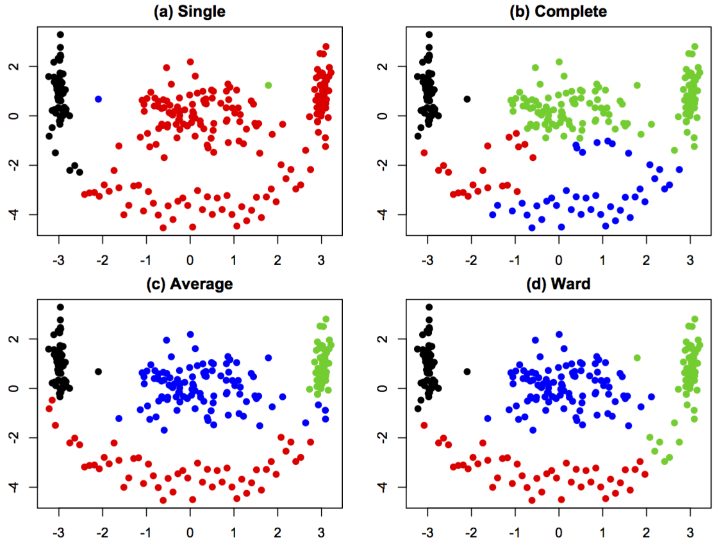
4. Ward's clustering

$$d(C_k, C_{k'}) = \frac{2\left(|C_k| \cdot |C_{k'}|\right)}{|C_k| + |C_{k'}|} ||\bar{X}_{C_k} - \bar{X}_{C_{k'}}||^2$$



(a) Single
0.0
0.2
0.4
0.6
0.8
(b) Complete
0
2
4
6
8
(c) Average
0
1
2
3
4
(d) Ward
0
50
100
150
200

K $= 4$ clusters



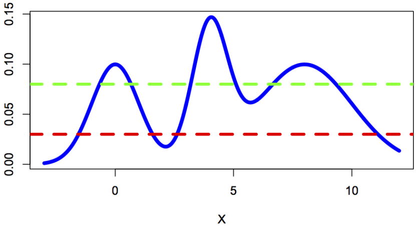
Statistical clustering

1. Parametric - associates a specific model with the density (e.g. Gaussian, Poisson)

   $\longrightarrow$ dataset is modeled by a mixture of these distributions

   $\longrightarrow$ parameters associated with each cluster

2. Nonparametric - looks at contours of the density to find cluster information (e.g. kernel density estimate)



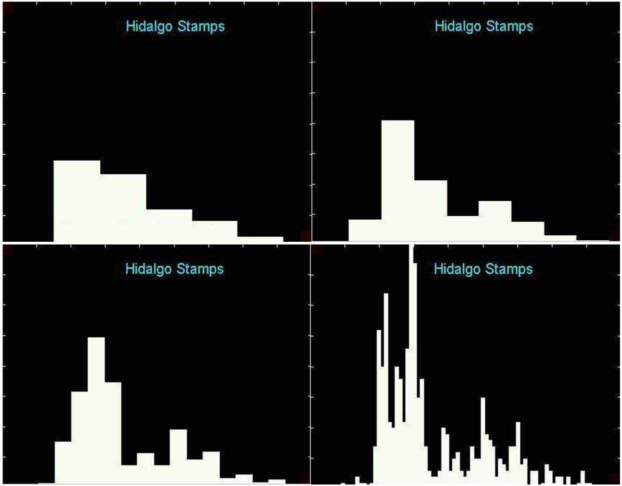
How many clusters are there?

JS Marron (UNC) **Hidalgo Stamps Data** to illustrate why **histograms** should not be used:

The main points are illustrated by the Hidalgo Stamps Data, brought to the statistical literature by Izenman and Sommer, (1988), Journal of the American Statistical Association, 83, 941-953. They are thicknesses of a type of postage stamp that was printed over a long period of time in Mexico during the 19th century. The thicknesses are quite variable, and the idea is to gain insights about the number of different factories that were producing the paper for this stamp over time, by finding clusters in the thicknesses.

`http://www.stat.unc.edu/faculty/marron/DataAnalyses/SiZer/SiZer_Basics.html`

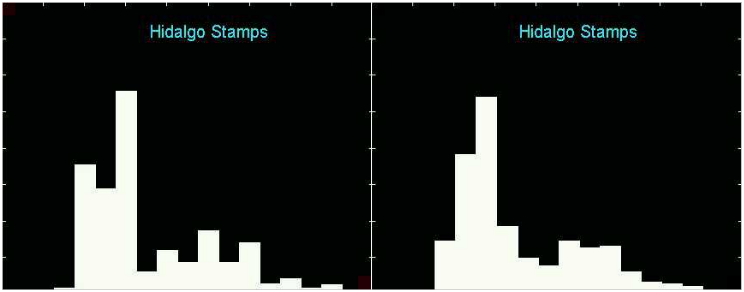
Changing the bin width dramatically alters the number of peaks



Images: JS Marron

These two histograms use the **same bin width**, but the second is **slightly right-shifted**.

Are there 7 modes (left) or two modes (right)?



See **movie version** of shifting issue here:

`http://www.stat.unc.edu/faculty/marron/DataAnalyses/SiZer/StampsHistLoc.mpg`

Images: JS Marron

Clustering - some final comments

- SiZer (Significance of Zero Crossings of the Derivative) - find statistically significant peaks

  `http://www.unc.edu/~marron/DataAnalyses/SiZer/SiZer_Basics.html`

- Nonparametric Inference For Density Modes (Genovese et al., 2015)

- Density ridges/filament finder (Chen et al., 2015b,a)

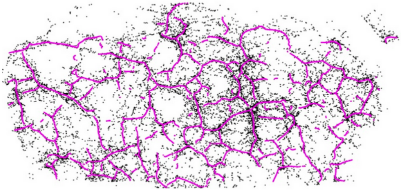


Image: Yen-Chi Chen (`http://www.stat.cmu.edu/~yenchic/research.html`)

Concluding Remarks

- **Classification - supervised/labels $\rightarrow$ predict classes**
  1. KNN
  2. Logistic regression
  3. LDA/QDA
  4. Support Vector Machines
  5. Tree classifiers

- **Clustering - unsupervised/no labels $\rightarrow$ find structure**
  1. K - means
  2. Hierarchical clustering
  3. Parametric/Non-parametric

- Clustering and classification are useful tools, but should be familiar with assumptions associated with the method selected

Bibliography

Chen, Y.-C., Ho, S., Brinkmann, J., Freeman, P. E., Genovese, C. R., Schneider, D. P., and Wasserman, L. (2015a), "Cosmic Web Reconstruction through Density Ridges: Catalogue," *arXiv preprint arXiv:1509.06443*.

Chen, Y.-C., Ho, S., Freeman, P. E., Genovese, C. R., and Wasserman, L. (2015b), "Cosmic Web Reconstruction through Density Ridges: Method and Algorithm," *arXiv preprint arXiv:1501.05303*.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2015), "Non-parametric inference for density modes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Geringer-Sameth, A., Walker, M. G., Koushiappas, S. M., Koposov, S. E., Belokurov, V., Torrealba, G., and Evans, N. W. (2015), "Indication of Gamma-ray Emission from the Newly Discovered Dwarf Galaxy Reticulum II," *Physical review letters*, 115, 081101.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning with Applications in R*, vol. 1 of *Springer Texts in Statistics*, Springer.

Li, H.-b. and Henning, T. (2011), "The alignment of molecular cloud magnetic fields with the spiral arms in M33," *Nature*, 479, 499–501.