Lecture 25: Inference for data visualization

Pratheepa Jeganathan

05/31/2019

Recall

- One sample sign test, Wilcoxon signed rank test, large-sample approximation, median, Hodges-Lehman estimator, distribution-free confidence interval.
- Jackknife for bias and standard error of an estimator.
- Bootstrap samples, bootstrap replicates.
- Bootstrap standard error of an estimator.
- Bootstrap percentile confidence interval.
- Hypothesis testing with the bootstrap (one-sample problem.)
- Assessing the error in bootstrap estimates.
- Example: inference on ratio of heart attack rates in the aspirin-intake group to the placebo group.
- The exhaustive bootstrap distribution.

- Discrete data problems (one-sample, two-sample proportion tests, test of homogeneity, test of independence).
- Two-sample problems (location problem - equal variance, unequal variance, exact test or Monte Carlo, large-sample approximation, H-L estimator, dispersion problem, general distribution).
- Permutation tests (permutation test for continuous data, different test statistic, accuracy of permutation tests).
- Permutation tests (discrete data problems, exchangeability.)
- Rank-based correlation analysis (Kendall and Spearman correlation coefficients.)
- Rank-based regression (straight line, multiple linear regression, statistical inference about the unknown parameters, nonparametric procedures - does not depend on the distribution of error term.)
- Smoothing (density estimation, bias-variance trade-off, curse of dimensionality)
- Nonparametric regression (Local averaging, local regression, kernel smoothing, local polynomial, penalized regression)

- Cross-validation, Variance Estimation, Confidence Bands, Bootstrap Confidence Bands.
- Wavelets (wavelet representation of a function, coefficient estimation using Discrete wavelet transformation, thresholding - VishuShrink and SureShrink).
- One-way layout (general alternative (KW test), ordered alternatives), multiple comparison procedure.
- Two-way layout (complete block design (Friedman test)), multiple comparison procedure, median polish, Tukey additivity plot, profile plots.
- Better bootstrap confidence intervals (bootstrap-t, percentile interval, BCa interval, ABC interval).

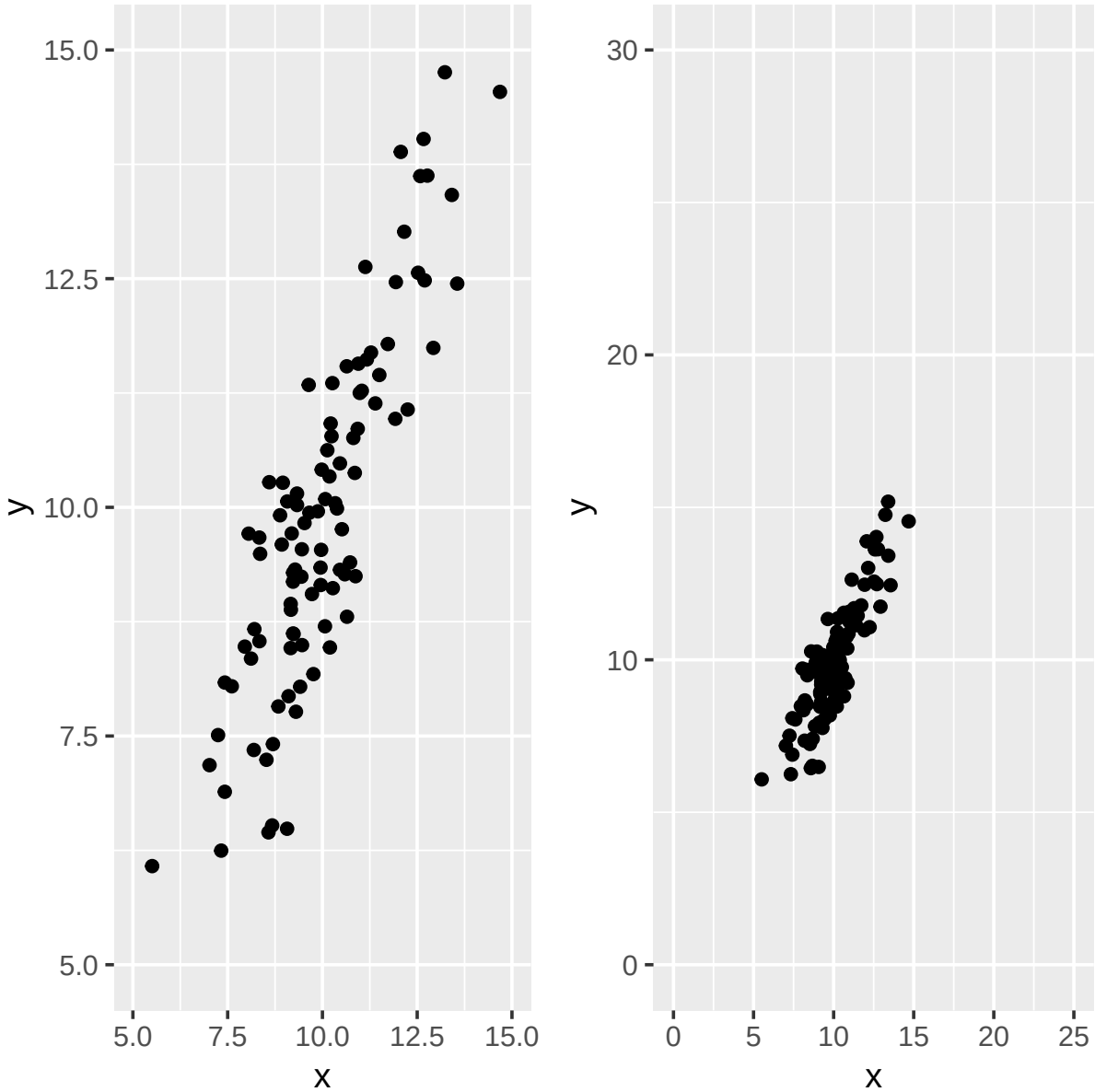
Inference for data visualization

Overview

- Visualization for
  - exploratory data analysis (to search for structure/pattern in the data)
  - model diagnostics (to search for structure not captured by the fitted model)
- `Magical thinking`: natural human inclination to over-interpret connections between random events (Diaconis 1985)

Overview

- Which plot shows the higher degree of association? (left/right)



- Reference (Diaconis 1985)
- Scale does change the perception of a viewer.
- We can use some protocols to check whether a pattern in the data arise by chance.

Inference for plots

Inference for plots

- Reference (Buja et al. 2009)
- 'The lineup' protocol:
  - Generate 19 null plots (assuming that structure is absent.)
  - Arrange all 19 plots and insert the plot from real data at random location.
  - Ask human viewer to single out the real plot.
  - Under the null hypothesis that all plots are the same, there is a one in 20 chance to single out the real one.
  - If the viewer chooses the plot of the real data, then the discovery can be assigned a p-value of $1/20 = 0.05$
  - Larger number of null plots could yield a smaller p-value.
    - But there is a limit of how many plots a human can consider.

Inference for plots

- Repeat the above 'the lineup' protocol with K independently recruited viewers.
  - Let X be the number of viewers choose the real data plot.
  - Suppose $k \leq K$ selected the plot of the real data.
    - p-value $= P(X \geq k)$, where $X \sim$ binomial $(n = K, p = 1/20)$.
    - If all the viewers picked the real data plot, p-value is $.05^K$.

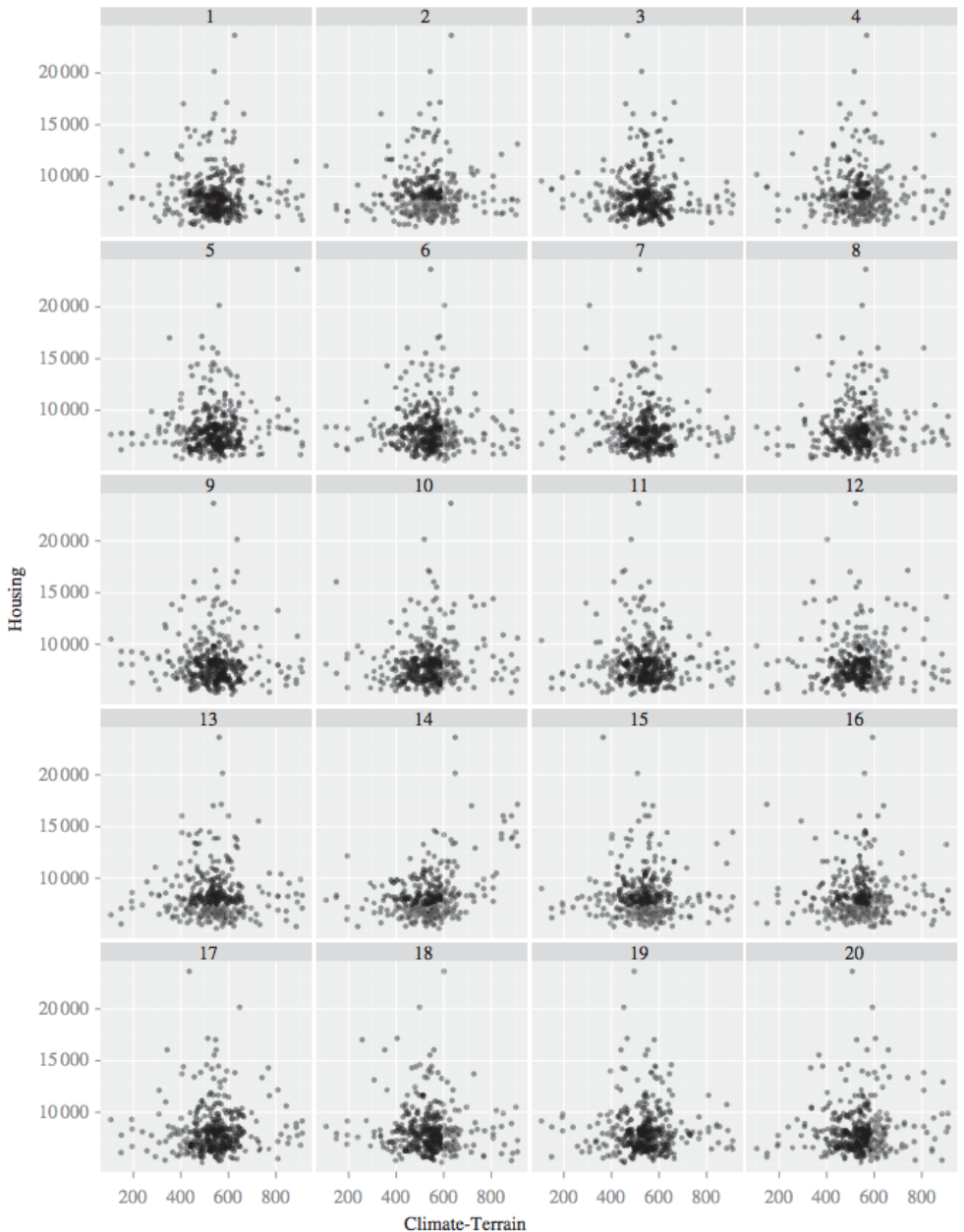
Inference for plots (Example)

- Scatter plot.
- Null hypothesis: the two variables are independent
  - null data sets can be produced by permuting the values of one variable against the other.

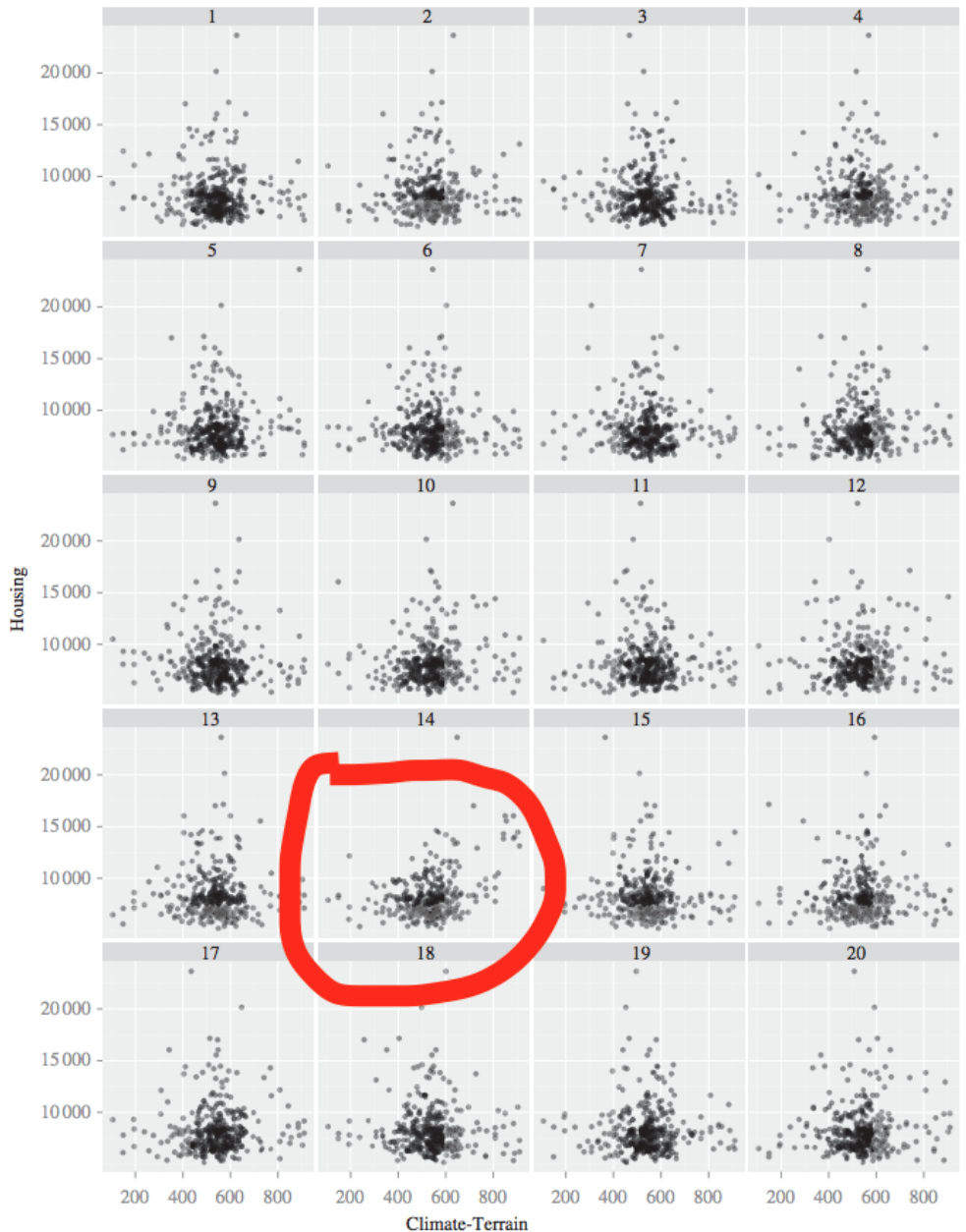
Inference for plots (Scatter plot Example)

- Cities across the USA were rated in 1984 according to many features (Boyer & Savageau (1984)).
  - census data, not a random sample.
- Consider 'Climate-Terrain' and 'Housing' variables.
  - Climate-Terrain: low values - uncomfortable temperatures, higher values - moderate temperatures.
  - Housing: a higher cost of owning a single family residence
- Expected association: more comfortable climates call for higher average housing costs.

Inference for plots (Scatter plot Example)



Inference for plots (Scatter plot Example)



Inference for plots (Scatter plot Example)

```
# number of students
K = 8
# number of correct picks
k = 2
pvalue = sum(dbinom(k:K,K,1/20)); pvalue
```

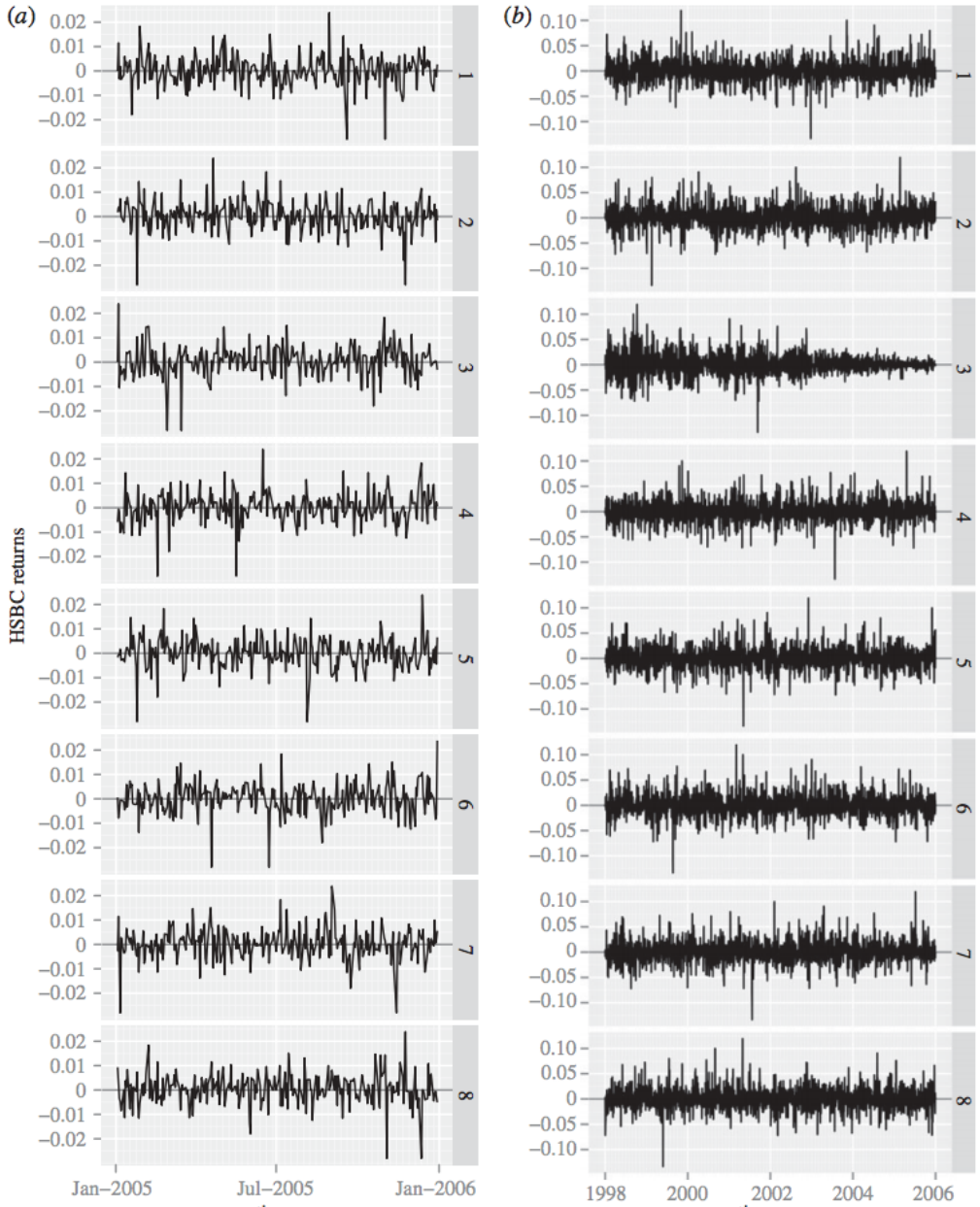
```
## [1] 0.05724465
```

- We reject the null hypothesis that Housing is independent of Climate-Terrain.

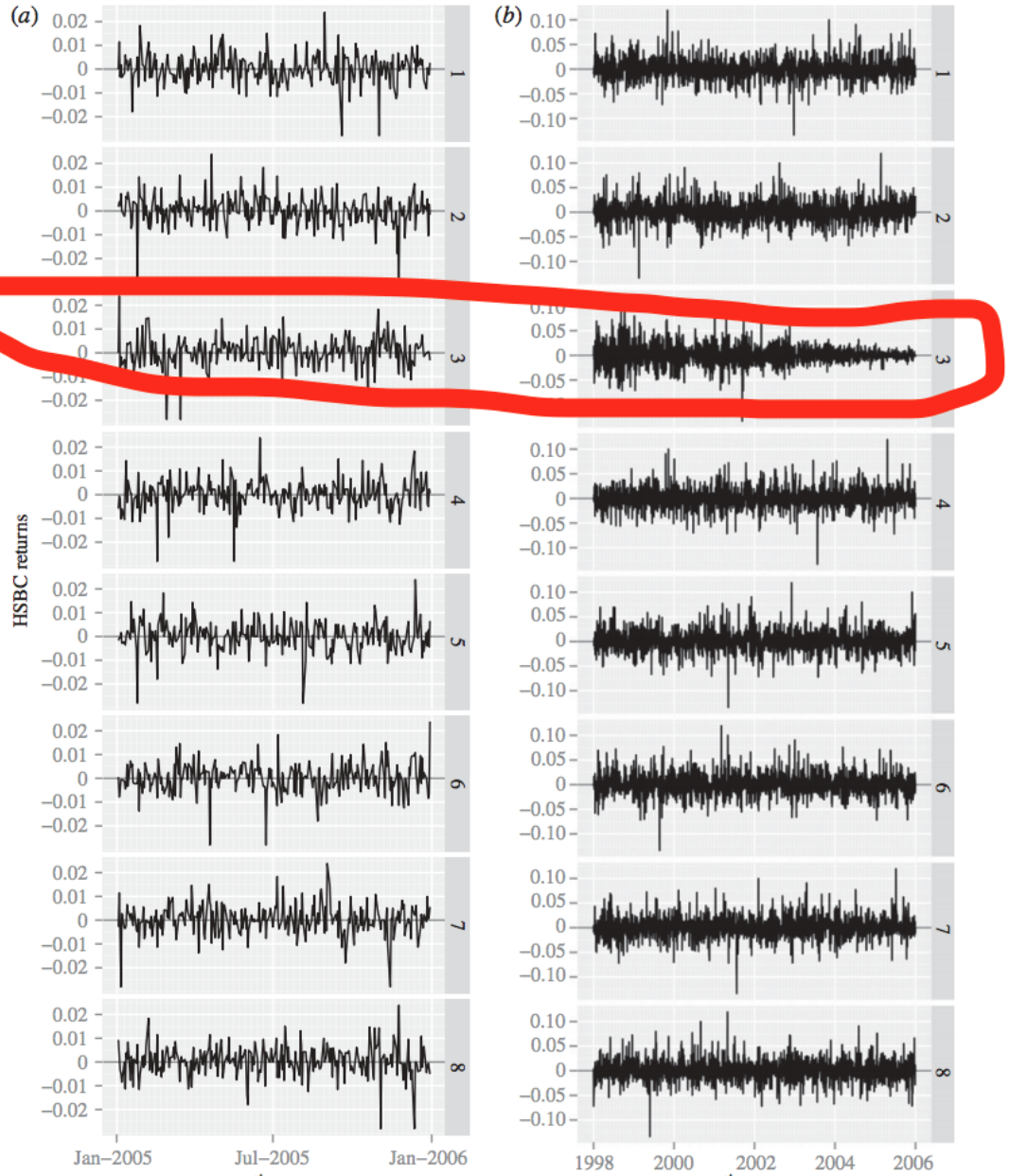
Inference for plots (time series plot example)

- HSBC ('The Hongkong and Shanghai Banking Corporation') daily stock returns (two panel data)
  - 2005 return only.
  - 1998–2005 return.
- In each panel, select which plot is the most different and explain why.

Inference for plots (time series plot example)



Inference for plots (time series plot example)



Inference for plots (time series plot example)

```r
# number of students
K = 8
# number of correct picks
k = 4
pvalue = sum(dbinom(k:K,K,1/8)); pvalue
```

```
## [1] 0.01124781
```

Inference for plots (time series plot example)

- In 2005 return data, the viewer should have had difficulty selecting the real data plot.
  - This is the year of low and stable volatility in return.
- In 1998–2005 return data, it should be easy.
  - There are two volatility bursts.
    - In 1998 due to the Russian bond default and the LTCM (long-term capital management) collapse.
    - In 2001 due to the 9/11 event.
    - Later after 2001, volatility stabilizes at a low level.

Principal component analysis (PCA)

PCA

- PCA is primarily an exploratory tool.
- PCA finds a low-dimensional subspace that minimizes the distances between projected points and subspace.
- Consider observations $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ - each $\boldsymbol{x}_i$ is a p column vector.
- Combine $\boldsymbol{x}_i$s in a matrix $\mathbf{X}$ with dimension $n \times p$.
- Start by recentering X, from now on consider X centered i.e. $\mathbf{1}_n X = 0$.
- PCA finds U, S, and V matrices such that $X = \mathbf{USV}^T$ (Singular value decomposition - SVD).
  - Columns of $\mathbf{V}$ are new variables.
  - Principal components $\mathbf{C} = \mathbf{US}$.

PCA

- PCA as an optimization problem.
  - The first column vector $\boldsymbol{v}_1$ of $\mathbf{V}$ is such that $\langle \boldsymbol{v}_1 , \boldsymbol{v}_1 \rangle = 1$ and

$$\hat{\boldsymbol{v}}_1 = \underset{\boldsymbol{v}_1}{\text{maximize}} \left\{ \mathbb{V} \left(\mathbf{X} \boldsymbol{v}_1 \right) \right\}.$$

  - Find $\boldsymbol{v}_2$ such that $\langle \boldsymbol{v}_1 , \boldsymbol{v}_2 \rangle = 0$, $\langle \boldsymbol{v}_2 , \boldsymbol{v}_2 \rangle = 1$, and

$$\hat{\boldsymbol{v}}_2 = \underset{\boldsymbol{v}_2}{\text{maximize}} \left\{ \mathbb{V} \left(\mathbf{X} \boldsymbol{v}_2 \right) \right\}.$$

- Keep going the same way until $\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_q$ have been collected and put them in $\hat{\mathbf{V}}$ of dimensions $p \times q$.

Bootstrap PCA

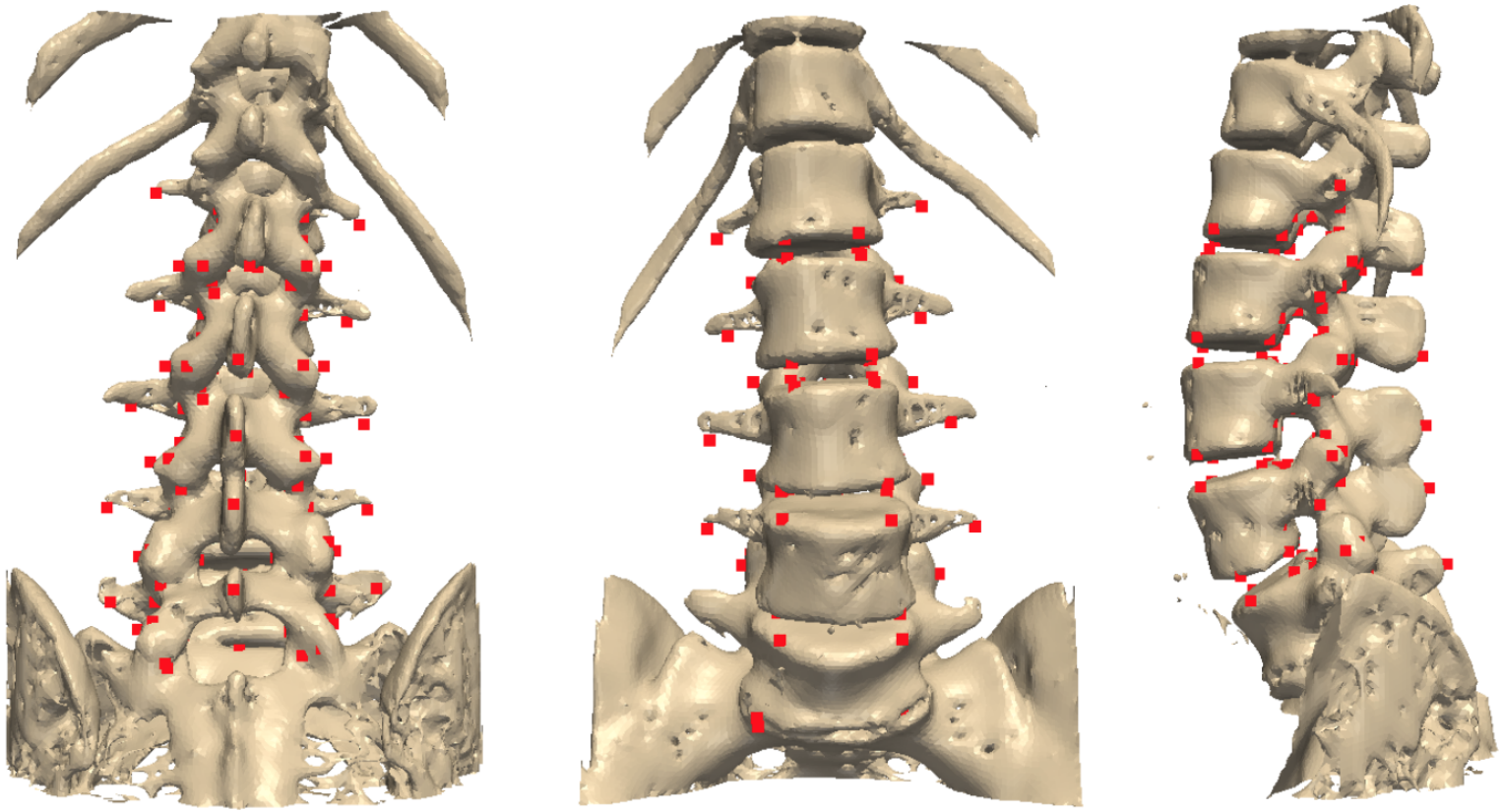
- Two ways to bootstrap PCA in case of random rows $\mathbf{X}$
  - Partial bootstrap.
  - Total bootstrap.
- Partial bootstrap:
  - Project B replications onto initial subspace.
  - Initial subspace is obtained by PCA on original $\mathbf{X}$.
  - Underestimates variation of parameters (Milan and Whittaker 1995).
- Total bootstrap:
  - Perform new PCA on each replication.
  - Nuisance variations in PCA on bootstrap samples: reflections and rotations
  - Align PCAs on bootstrap samples.

Bootstrap PCA (need of Procrustes analysis)

- For the total bootstrap, need to align PCAs of bootstrap samples.
- This is usually done using Procrustes analysis.
- Procrustes refers to a bandit from Greek mythology who made his victims fit his bed by stretching their limbs (or cutting them off)
- Procrustes analysis is used in statistical shape analysis to align shapes and minimize difference between shapes to retain real shape (by removing nuisance parameters):
  - translation in space
  - rotation in space
  - sometimes scaling of the objects

Bootstrap PCA (need of Procrustes analysis)

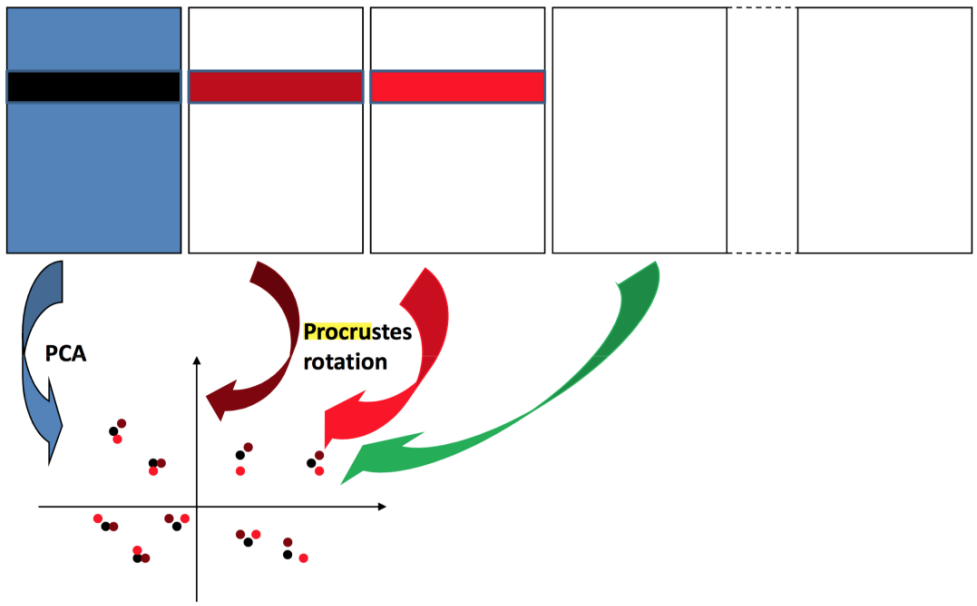
- Shapes are different by rotation, translation, and scaling.
  - Shape example: landmarks for the human spine



- We need to remove translation, rotation, and scaling effects before comparing shapes of the spinal code defined by the landmarks (red points).

Bootstrap PCA (need of Procrustes analysis)

- In PCA, shapes are the projected points onto the lower dimensional subspace spanned by say PC1 and PC2.



Source: Josse, Wager, and Husson (2014)

- Procrustes rotations of the rows of the PCA estimators onto the initial configuration. The first table corresponds to $\mathbf{X}$ and the other to $\hat{\mathbf{X}}^{*b}, b = 1, \cdots, B$.

Bootstrap PCA

- Collecting B bootstrap sampled PCAs by resampling rows of data matrix $\mathbf{X}$

$$\mathbf{V}_q^{*1}, \mathbf{V}_q^{*2}, \ldots \mathbf{V}_q^{*B}.$$

- Align all the projected point set using Procrustes alignment.
  - Find the rotation for each bootstrap replicates

$$\hat{R}^b = \underset{R}{\text{minimize}} \left\{ \left|\left| \mathbf{X}^{*1}\mathbf{V}_q^{*1} - \mathbf{X}^{*b}\mathbf{V}_q^{*b}R \right|\right| \right\}.$$

  - Apply the rotation to projected data point for each bootstrap sample

$$X^{*b}\mathbf{V}_q^{*b}\hat{R}^b.$$

- Overlay points and draw contours around it.
- This nonparamteric bootstrap approach modify the structure of the rows of $\mathbf{X}$ for each bootstrap replication.

Parametric bootstrap PCA

- In case of fixed rows and columns X, we can use parametric bootstrap.
  - Perform PCA on $\mathbf{X}$ to estimate $\hat{\mathbf{V}}_q$.
  - Estimate residual σ^2 from the residual matrix $\mathbf{E} = X - \hat{\mathbf{V}}_q \hat{\mathbf{V}}_q^T \mathbf{X}$
  - Draw $\epsilon_{ij} \sim \mathrm{N}\left(0, \hat{\sigma}^2\right)$.
  - Generate new matrix $X^{*b} = \hat{\mathbf{V}}_q \hat{\mathbf{V}}_q^T \mathbf{X} + \mathbf{E}^{*b}$.
  - Perform PCA on X^{*b}.

Parametric bootstrap PCA (Example)

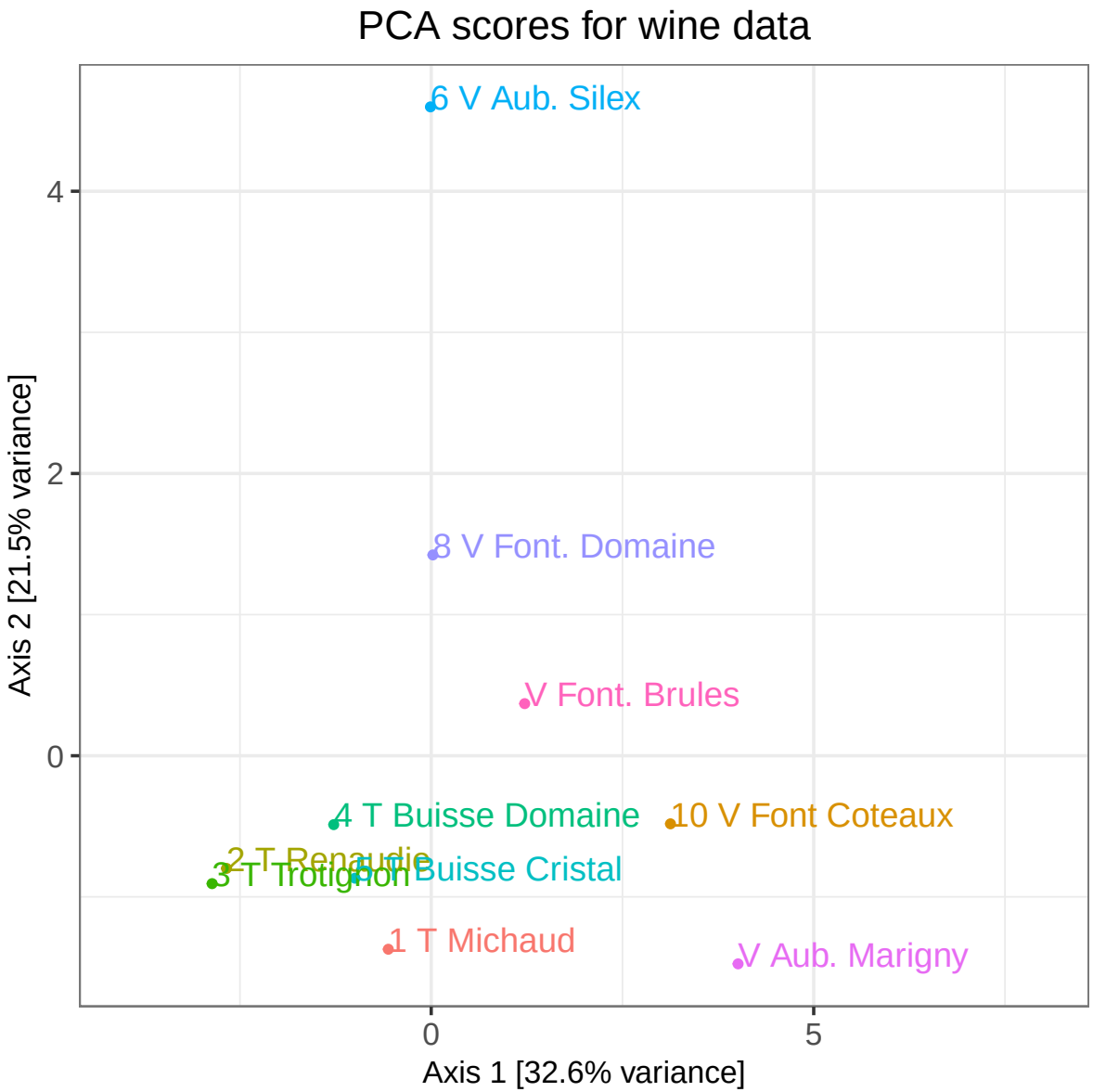
- Consumers describe 10 white wines with 14 sensory attributes.
- Consumers score wines between 1 and 10 for each attribute.
- Collect averages across consumers in 10×14 matrix $\boldsymbol{X}$.

```
library(SensoMineR)
data(napping)
data(napping.words)
library(ade4)
library(magrittr)
library(dplyr)
df = napping.words
wine.pca = dudi.pca(df, scannf = F, nf = 3)
row.scores = data.frame(li = wine.pca$li,
  SampleCode = rownames(df))
row.scores = row.scores %>%
  left_join(data.frame(SampleCode = rownames(df)))
evals.prop = 100 * (wine.pca$eig / sum(wine.pca$eig))
```

Parametric bootstrap PCA (Example)

```
evals = wine.pca$eig
p = ggplot(data = row.scores, aes(x = li.Axis1,
  y = li.Axis2, label = SampleCode, col = SampleCode)) +
  geom_point(size = 1)  +
  geom_text(aes(label = SampleCode),
    hjust=0, vjust=0.1) +
  labs(x = sprintf("Axis 1 [%s%% variance]",
    round(evals.prop[1], 1)),
       y = sprintf("Axis 2 [%s%% variance]",
         round(evals.prop[2], 1))) +
  ggtitle("PCA scores for wine data")
```

Parametric bootstrap PCA (Example)



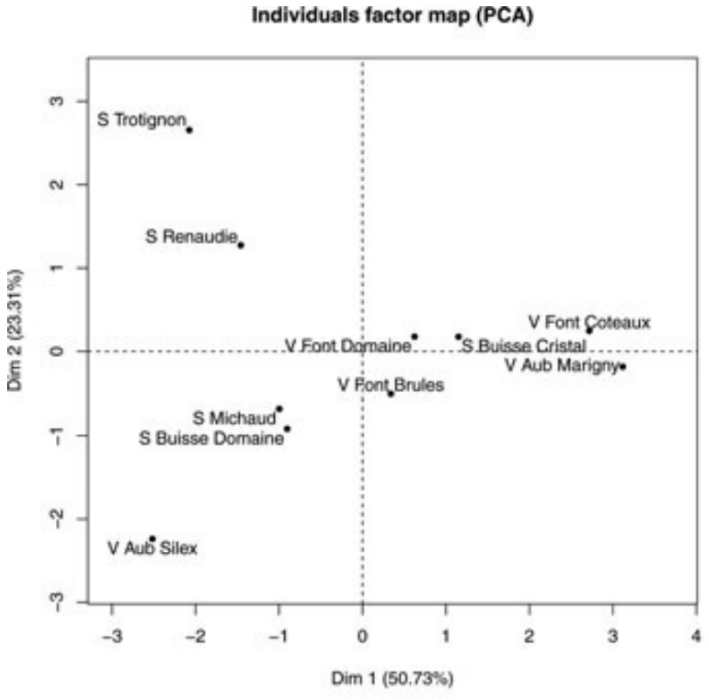
- The first two PCs explained almost 54% of total variability.

Parametric bootstrap PCA (Example)

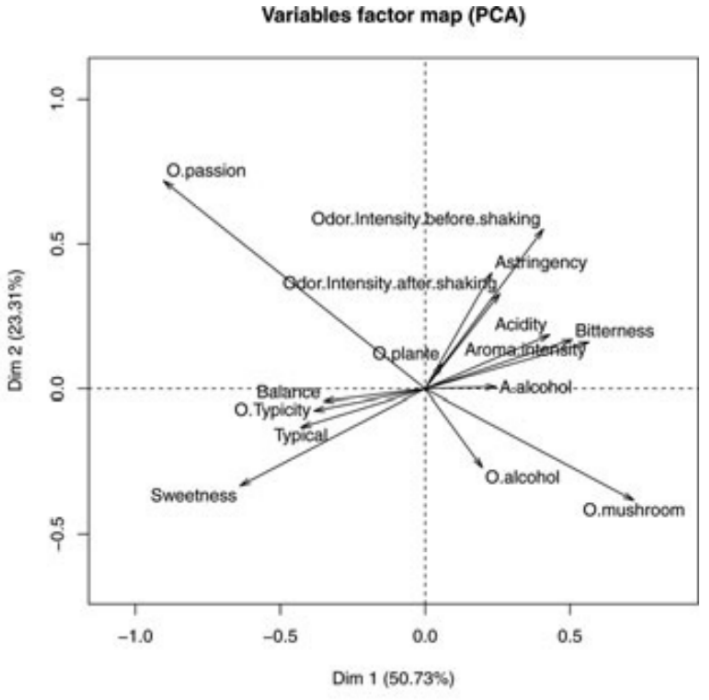
- With the bootstrap confidence ellipses

Parametric bootstrap PCA (Example)

- PCA on the wine data set (Josse, Wager, and Husson 2016).



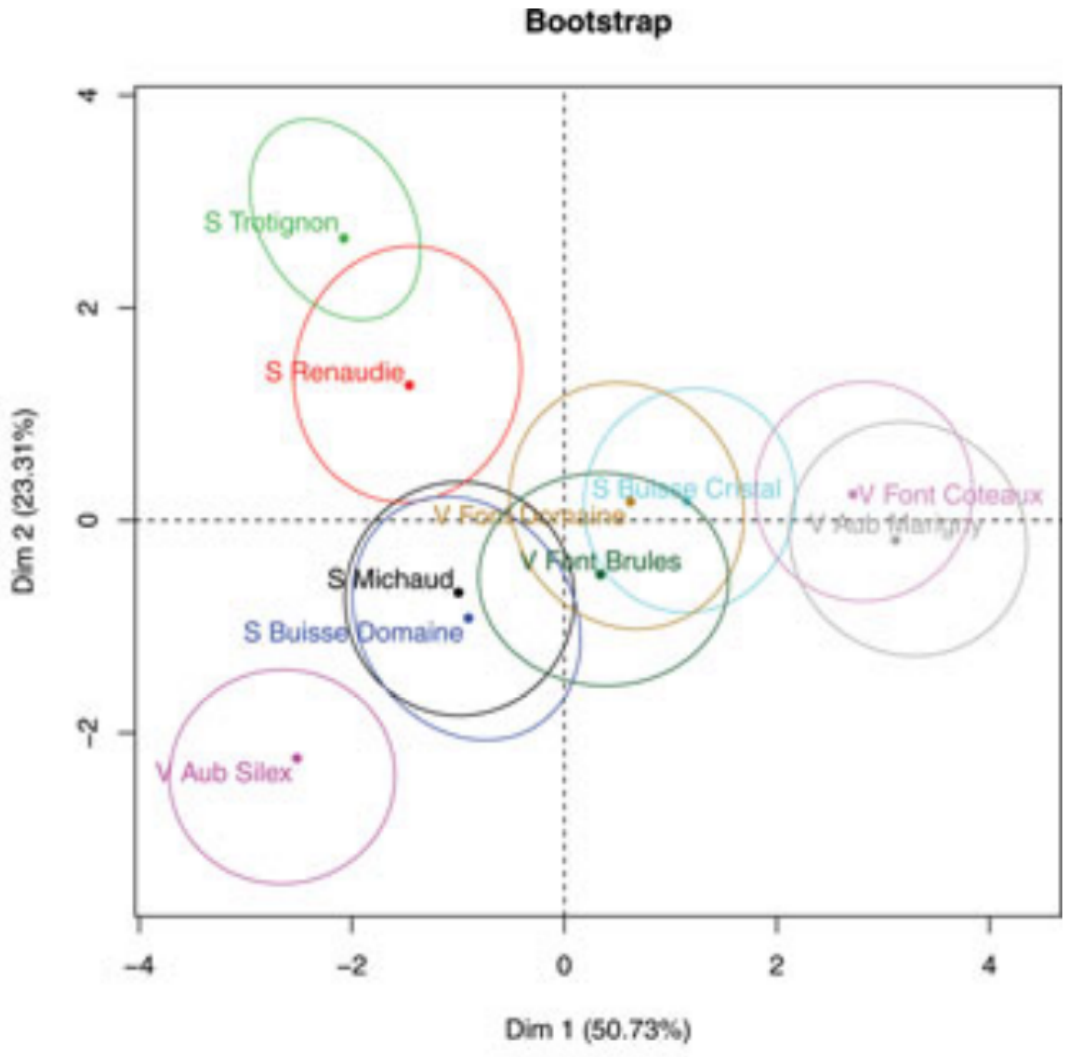
(a) 2 dimensional wines representation.



(b) PCA variables.

Parametric bootstrap PCA (Example)

- Bootstrap for PCA scores. (Josse, Wager, and Husson 2016).



Parametric bootstrap PCA (Notes)

- Read (Josse, Wager, and Husson 2016) for confidence areas using jackknife for PCA.

References for this lecture

- **HW2018** Modern statistics for modern biology.
- **Christof 2016:** Lecture notes on inference for visualization

References for this lecture

Buja, Andreas, Dianne Cook, Heike Hofmann, Michael Lawrence, Eun-Kyung Lee, Deborah F Swayne, and Hadley Wickham. 2009. "Statistical Inference for Exploratory Data Analysis and Model Diagnostics." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367 (1906). The Royal Society Publishing: 4361–83.

Diaconis, Persi. 1985. "Theories of Data Analysis: From Magical Thinking Through Classical Statistics." *Exploring Data Tables, Trends, and Shapes*. Wiley Online Library, 1.

Josse, Julie, Stefan Wager, and François Husson. 2016. "Confidence Areas for Fixed-Effects Pca." *Journal of Computational and Graphical Statistics* 25 (1). Taylor & Francis: 28–48.

Milan, Luis, and Joe Whittaker. 1995. "Application of the Parametric Bootstrap to Models That Incorporate a Singular Value Decomposition." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 44 (1). Wiley Online Library: 31–49.