



Data Exploration and Visualisation

Part 1: Data Exploration

Dr Tom Diethe

tom.diethe@bristol.ac.uk

University of Bristol

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Recap and Outlook

Lec 1 Intro

Lec 2 Data Ingress

Lec 3 Recommender Systems

Lec 4 Databases

Lec 5 Data Wrangling

Lec 6 Data Fusion

- ▶ Demonstration (continued)
- ▶ Merging Datasets
- ▶ Data Aggregation

Lec 7 Data Exploration

Lec 8 Data Visualisation

Lec 9 Data sharing, privacy and anonymisation

Lec 10 Deploying data science systems

Lec 11 The future of data science

Outline

Recap and Outlook

Descriptive Statistics

Univariate Analysis

Measures of Central Tendency

Moments

Measures of Shape

Transforms

Density Estimation

Multivariate Analysis

Dimensionality Reduction

Linear Dimensionality Reduction

Nonlinear Dimensionality Reduction

Descriptive vs Inferential Statistics

✿ Inferential/Inductive:

- ▶ Summarise *sample*
- ▶ Based on probability theory

✿ Descriptive:

- ▶ Describe or summarise dataset
- ▶ Useful for assessing data quality, developing models, general understanding
- ▶ Often presented alongside conclusions inferential statistics

Univariate Measures

✿ Central tendency

- ▶ Mean, median, mode

✿ Variability

- ▶ Variance, quartiles, min, max, kurtosis, skewness

✿ Graphical/Tabular format

- ▶ Histograms

Types of Variable

🔥 Categorical/Discrete (Qualitative)

- ▶ Nominal
- ▶ Ordinal: ordered
- ▶ Dichotomous: only 2 categories

🔥 Continuous (Quantitative)

- ▶ Interval: measured along a continuum and numerical value
 - ▶ e.g. temperature measured in degrees Celsius or Fahrenheit). So the difference between 20C and 30C is the same as 30C to 40C
- ▶ Ratio: added condition that 0 indicates that there is none of that variable
 - ▶ e.g. temperature measured in Kelvin, as 0 Kelvin indicates that there is “no temperature”
 - ▶ e.g. height, mass, distance etc.
 - ▶ reflects the fact that you can use the ratio of measurements (distance of 10 m is twice the distance of 5 m.

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Central Tendency

Mean (arithmetic) $\mu = \frac{1}{n} \sum_{i=1}^n x_i$

- ▶ can be used with both discrete and continuous values
- ▶ particularly susceptible to the influence of outliers

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

- ▶ mean: 30.7k, most in the 12k to 18k range
- ▶ mean is being skewed by the two large salaries

Median

- ▶ middle score for a set of data after sorting
- ▶ less affected by outliers and skewed data

Mode

- ▶ most frequent value (discrete)
- ▶ highest bar in histogram (continuous)

Mode examples

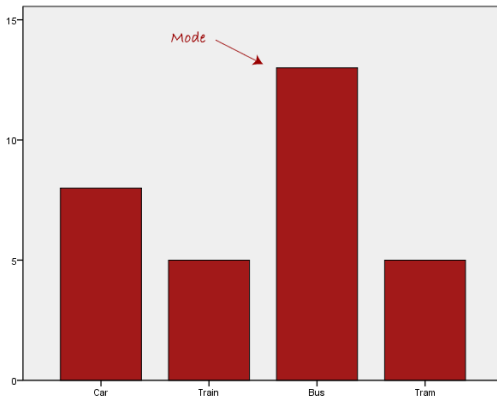


Figure: Discrete

Mode examples

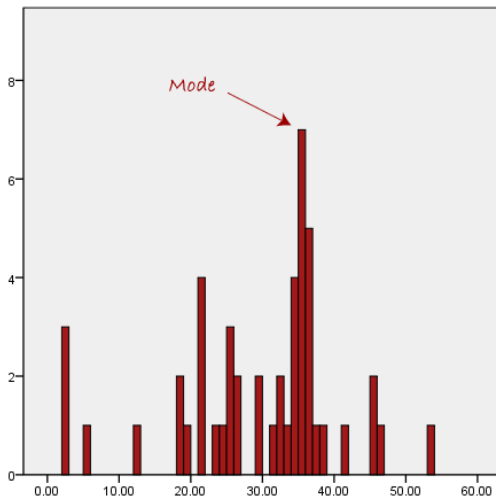


Figure: Continuous

Mode examples

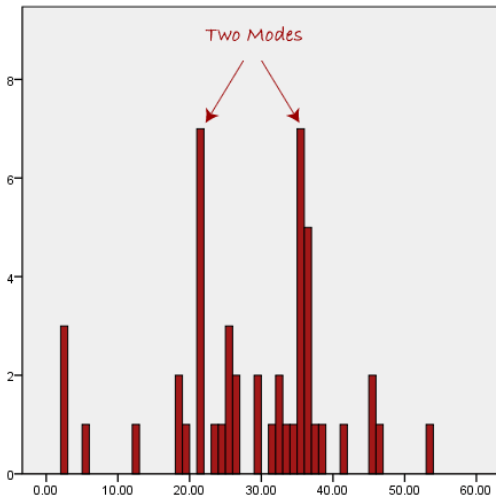


Figure: Two modes

Mode examples

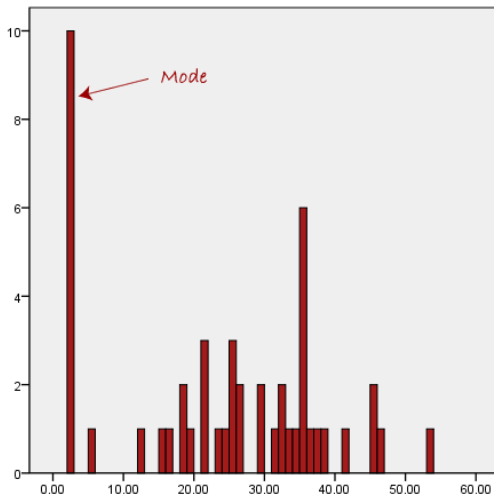


Figure: Outliers

Skewed Distributions and the Mean and Median

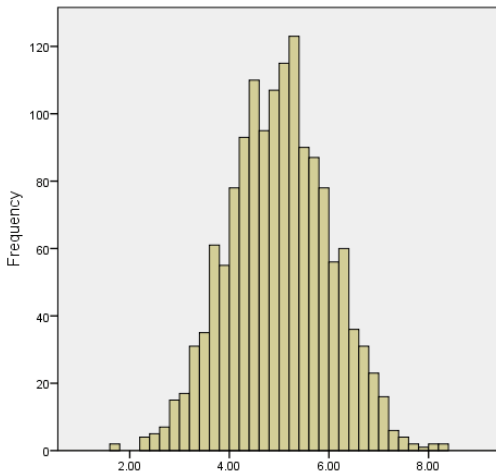


Figure: Normal distribution

Skewed Distributions and the Mean and Median

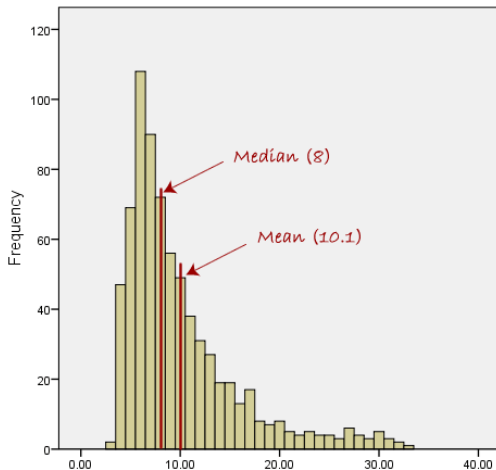


Figure: Skewed Distribution

When to use the mean, median and mode

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Measures of Dispersion

✿ Range: $\max() - \min()$

- ▶ Only depends on 2 values
- ▶ Most useful in representing the dispersion of small data sets

✿ Variance $\text{Var}(X) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2$

✿ Unbiased variance: $\text{Var}(X) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu)^2$

✿ Standard deviation $\sigma = \sqrt{\text{Var}(X)}$ (positive square root)

✿ Quantiles

- ▶ Cut-points dividing the range of a probability distribution into contiguous intervals with equal probabilities
- ▶ or dividing the observations in a sample in the same way

Unbiased Variance: Intuition

The degree to which x_i varies from \bar{X} + The degree to which \bar{X} varies from μ = The degree to which x_i varies from μ .

That is,

$$\mathbb{E} \left[(X - \bar{X})^2 \right] + \mathbb{E} \left[(\bar{X} - \mu)^2 \right] = \mathbb{E} \left[(X - \mu)^2 \right].$$

Proof requires a bit of algebra. But assuming it is true, we can rearrange to get:

$$\mathbb{E} \left[(X - \bar{X})^2 \right] = \underbrace{\mathbb{E} \left[(X - \mu)^2 \right]}_{\sigma^2} - \underbrace{\mathbb{E} \left[(\bar{X} - \mu)^2 \right]}_{\frac{\sigma^2}{n}} = \frac{n-1}{n} \sigma^2.$$

See https://en.wikipedia.org/wiki/Bessel's_correction

Quantiles

✿ k^{th} q -quantile is where the cumulative distribution function crosses k/q :

$$Pr[X \geq x] \geq 1 - \frac{k}{q}$$

- ✿ The only 2-quantile is called the median
- ✿ The 4-quantiles are called quartiles Q ; the difference between upper and lower quartiles is also called the interquartile range, mid-spread or middle fifty $IQR = Q_3 - Q_1$
- ✿ The 100-quantiles are called percentiles P

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments**

- Measures of Shape

- Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Raw Moments

Let X be a random variable with a probability distribution P and mean value $\mu = \mathbb{E}[X]$ (i.e. the first raw moment or moment about zero), the operator \mathbb{E} denoting the expected value of X .

Central Moments


 n^{th} central moment:

$$\mu_n = \mathbb{E} [(X - \mathbb{E}[X])^n]$$

$$\mu_0 \mathbb{E} [(X - \mathbb{E}[X])^0] = 1$$

$$\mu_1 \mathbb{E} [(X - \mathbb{E}[X])^1] = 0$$

$$\mu_2 \mathbb{E} [(X - \mathbb{E}[X])^2] = \sigma^2 \text{ (variance)}$$

 The third and fourth central moments are used to define the standardised moments which are used to define skewness and kurtosis, respectively.

Standardised Moments

- ✿ Ratio of the k^{th} moment about the mean and the standard deviation to the power of k

$$\bar{\mu}_k = \frac{\mu_k}{\sigma^k} = \frac{\mathbb{E}[(X - \mu)^k]}{\left(\sqrt{\mathbb{E}[(X - \mu)^2]}\right)^k}$$

- ✿ power of k is because moments scale as $x^k \rightarrow$ **scale invariant**

Cumulants

Alternative to moments, some useful properties

Relation to moments:

Cumulant	Raw moment	Central	Standardised
κ_1	μ_1	0	0
κ_2	$\mu_2 - \mu_1^2$	μ_2	1
κ_3	$\mu_3 - 3\mu_2\mu_1 + 2\mu_1^3$	μ_3	μ_3
κ_4	$\mu_4 - 4\mu_3\mu_1 - 3\mu_2^2 + 12\mu_2\mu_1^2 - 6\mu_1^4$	$\mu_4 - 3\mu_2^2$	$\mu_4 - 3$

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape**

- Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Measures of Shape: Skewness

- “lopsidedness” of the distribution
- symmetric distributions (if defined) = 0

$$\text{Skew}[X] = \mathbb{E} \left[\left(\frac{X - \mu}{\sigma} \right)^3 \right] = \frac{\mu_3}{\sigma^3} = \frac{\mathbb{E} [(X - \mu)^3]}{(\mathbb{E} [(X - \mu)^2])^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}}$$

where:

- ▶ μ_3 : third central moment
- ▶ κ_t : t^{th} cumulants

Sample skewness:

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \right]^{3/2}}$$

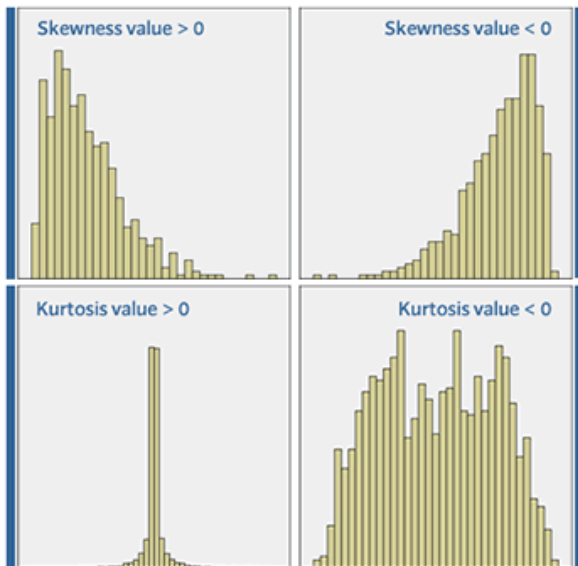
Measures of Shape: Kurtosis

- heaviness of the tail of the distribution, compared to the normal distribution of the same variance
- always positive
- kurtosis of a normal is $3\sigma^4 \implies$ **excess kurtosis** = $\text{Kurt}[X] - 3$

$$\text{Kurt}[X] = \mathbb{E} \left[\left(\frac{X - \mu}{\sigma} \right)^4 \right] = \frac{\mu_4}{\sigma^4} = \frac{\mathbb{E}[(X - \mu)^4]}{(\mathbb{E}[(X - \mu)^2])^2}$$

- Sample excess kurtosis:

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4}{\left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^2} - 3$$



Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Variance Stabilising Transforms

- ✿ Usual assumption in regression is that the variance of each observation is the same
- ✿ Problem: In many cases, the variance is not constant, but is related to the mean
 - ▶ Poisson Data (Counts of events): $\mathbb{E}(X) = \text{Var}(X) = \mu$
 - ▶ Binomial Data (and Percents): $\mathbb{E}(X) = m\pi$, $\text{Var}(X) = m\pi(1 - \pi)$
 - ▶ General Case: $\mathbb{E}(X) = \mu$, $\text{Var}(X) = f(\mu)$
 - ▶ Power relationship: $\text{Var}(X) = \sigma^2 = \alpha^2 \mu^{2\beta}$

$$\sigma = \alpha \mu^\beta \implies \log(\sigma) = \log(\alpha) + \beta \log(\mu)$$

- ✿ **Box-Cox transformation** (Sakia, 1992) can be used to diagnose and transform

De-correlating

Transform data so that it has diagonal covariance matrix $\Sigma = \mathbf{X}\mathbf{X}^T$. This transform can be found by solving an eigenvalue problem:

$$\Sigma\Phi = \Phi\Lambda$$

where Λ is a diagonal matrix of eigenvalues, and the columns of Φ are the eigenvectors of the covariance matrix.

$\therefore \Phi$ diagonalises Σ .

We can also write the diagonalised covariance as:

$$\Phi^T \Sigma \Phi = \Lambda \quad (1)$$

So to de-correlate a single vector \mathbf{x}_i (e.g. at test time), we do:

$$\hat{\mathbf{x}}_i = \Phi^T \mathbf{x}_i \quad (2)$$

Whitening

- ✿ The diagonal elements (eigenvalues) in Λ may be the same or different
- ✿ If we make them all the same, then this is called **whitening** the data
- ✿ Each eigenvalue determines the length of its associated eigenvector
- ✿ Not whitened \implies elliptical Σ . Whitened \implies spherical Σ

$$\Lambda^{-1/2} \Lambda \Lambda^{-1/2} = \mathbf{I}$$
$$\Lambda^{-1/2} \Phi^T \Sigma \Phi \Lambda^{-1/2} = \mathbf{I} \quad \text{substituting in (1)}$$

To apply to $\hat{\mathbf{x}}_i$, multiply by this scale factor \rightarrow whitened data point $\tilde{\mathbf{x}}_i$:

$$\tilde{\mathbf{x}}_i = \Lambda^{-1/2} \hat{\mathbf{x}}_i = \Lambda^{-1/2} \Phi^T \mathbf{x}_i \quad (3)$$

Now Σ is not only diagonal, but also uniform (white), since $\mathbb{E}(\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T) = \mathbf{I}$

When this might not be useful

1. The scaling of data examples is somehow important in the inference problem you are looking at
 - ▶ Could use the eigenvalues as an additional set of features to get around this
2. Computation:
 - ▶ you have to compute the covariance matrix Σ , which may be too large to fit in memory (if you have thousands of features) or take too long to compute;
 - ▶ secondly the eigenvalue decomposition is $\mathcal{O}(n^3)$ (see <http://mathoverflow.net/questions/62904/complexity-of-eigenvalue-decomposition>)
3. Common ML “gotcha”: *calculate the scaling factors on the training data*, and then you use equations (2) and (3) to apply the same scaling factors to the test data, otherwise you are at risk of over-fitting (you would be using information from the test set in the training process).

Source: <http://courses.media.mit.edu/2010fall/mas622j/whiten.pdf>

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation**

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Density Estimation

✿ Non-parametric: histogram, kernel density estimate

- + Very flexible
- + Little or no prior knowledge required
- + Inference is easy
- Expensive in memory and CPU (must store all data)
- Not much opportunity to incorporate prior knowledge

✿ Parametric: Gaussian mixture model

- + Restricted family of functions
- + Encode assumptions
- + Compact representation
- Inflexible; model might be wrong!
- Appropriate model might be obscure, complicated

✿ Semi-parametric: Dirichlet process mixture model (infinite limit of GMM)

<http://scikit-learn.org/stable/modules/density.html>

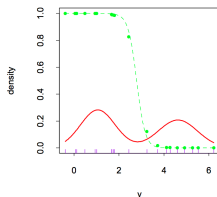
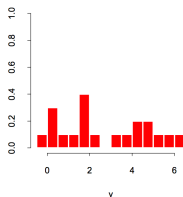
Gaussian Mixture Model

✿ Mixture Model: $f(x) = \sum_j \pi_j g_j(x)$ s.t. $0 \leq \pi_j \leq 1, \sum_j \pi_j = 1$

✿ Gaussian mixture: $g_j(x) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(x-\mu_j)^2/2\sigma_j^2}$

EM algorithm

- ✿ Take initial guesses for the parameters
- ✿ E-step: compute responsibilities
- ✿ M-step: compute weighted Gaussians
- ✿ Iterate until convergence



Kernel Density Estimation

🔥 Non-parametric way to estimate the probability density function

$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^n K_h(x - x_i) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - x_i}{h}\right)$$

where $K(\cdot)$ is a **kernel** or smoothing function - see [https://en.wikipedia.org/wiki/Kernel_\(statistics\)#In_non-parametric_statistics](https://en.wikipedia.org/wiki/Kernel_(statistics)#In_non-parametric_statistics)

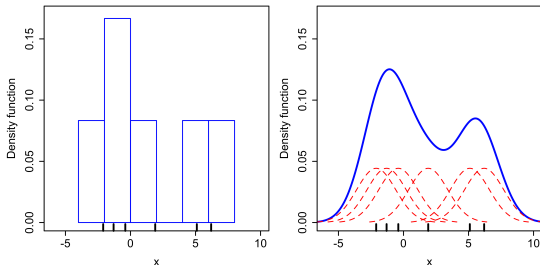


Figure: Comparison of 1D histogram and KDE

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation

- Multivariate Analysis**

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Bivariate and Multivariate Analysis

- ✦ Samples consists of more than one variable
- ✦ Descriptive statistics may be used to describe the relationship between pairs of variables
 - ▶ Cross-tabulations and contingency tables
 - ▶ Graphical representation via scatter-plots
 - ▶ Quantitative measures of dependence
 - ▶ Correlation (Pearson's r if both continuous else Spearman's ρ)
 - ▶ Covariance (reflects scale)
 - ▶ Descriptions of conditional distributions

Dimensionality Reduction

- Assume we are given a data set of (high-dimensional) input objects

$$\mathbf{X} = \{\mathbf{x}_i\}_{i=1,\dots,m}, \quad \mathbf{x}_i \in \mathcal{R}^n$$

- Aim is to learn a k -dimensional embedding in which each object is represented by a point

$$\mathbf{P} = \{\mathbf{p}_i\}_{i=1,\dots,m}, \quad \mathbf{p}_i \in \mathcal{R}^k$$

- typical values for k are 2 or 3 (for visualisation), in general $k \ll n$

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Principal Component Analysis (PCA)

- Standardise the data
- Obtain the eigenvectors and eigenvalues from the covariance matrix or correlation matrix, or perform SVD.
- Sort eigenvalues (descending) and choose the k eigenvectors that correspond to the k largest eigenvalues ($k \leq n$)
- Construct the projection matrix \mathbf{R} from the selected k eigenvectors
- Transform the original dataset \mathbf{X} via \mathbf{R} to obtain a k -dimensional feature subspace \mathbf{P}

Random Projections

$$\mathbf{P} = \mathbf{X}\mathbf{R}$$

where:

$$\mathbf{X} \in \mathcal{R}^{m \times n} \quad \text{(data matrix)}$$

$$\mathbf{R} \in \mathcal{R}^{k \times m} \quad \text{(projection matrix)}$$

$$\mathbf{P} \in \mathcal{R}^{m \times k} \quad \text{(lower dimensional representation)}$$

Gaussian Random Projections

- ✿ First row is a random unit vector uniformly chosen from S^{n-1} .
- ✿ Second row is a r.u.v from the space orthogonal to the first row
- ✿ Third row is a r.u.v. from the space orthogonal to the first two rows, etc.
- ✿ In this way of choosing \mathbf{R} , the following properties are satisfied:
 - Spherical symmetry** : For $\mathbf{A} \in \mathcal{O}(n)$, i.e. $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}$, $\mathbf{A}\mathbf{R}$ and \mathbf{R} have the same distribution.
 - Orthogonality** : The rows of \mathbf{R} are orthogonal to each other.
 - Normality** : The rows of \mathbf{R} are unit-length vectors.

Johnson-Lindenstrauss Lemma

J-L Lemma

Given $0 < \varepsilon < 1$, a set X of m points in \mathcal{R}^n , and a number $k > 8 \frac{\log(m)}{\varepsilon^2}$, there is a linear map $f : \mathcal{R}^n \mapsto \mathcal{R}^k$ such that:


$$(1 - \varepsilon)\|\mathbf{u} - \mathbf{v}\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{v})\|^2 \leq (1 + \varepsilon)\|\mathbf{u} - \mathbf{v}\|^2$$

for all $\mathbf{u}, \mathbf{v} \in X$

- 🔥 Proof: see Dasgupta and Gupta (2003)
- 🔥 An orthogonal projection will, in general, reduce the average distance between points
 - ▶ Roll the dice \rightarrow random projection
 - ▶ Scale up the distances so that the average distance is the same
- 🔥 Compatible with **approximate nearest neighbours**

Database friendly random projections (Achlioptas, 2001)

$$\mathbf{R}_{i,j} = \sqrt{3} \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

 http://scikit-learn.org/stable/modules/random_projection.html

Outline

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation

- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

(Van Der Maaten et al., 2009)

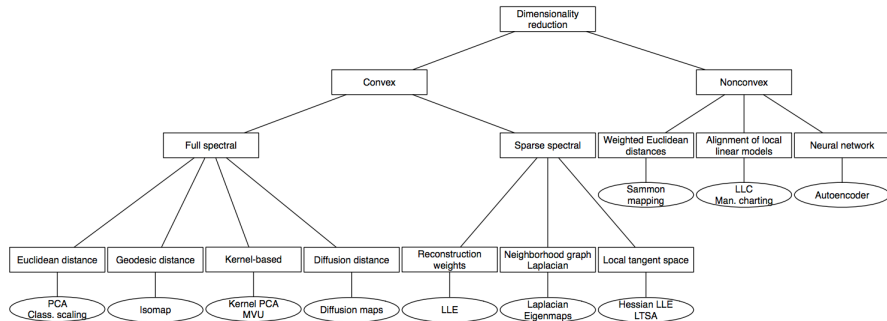
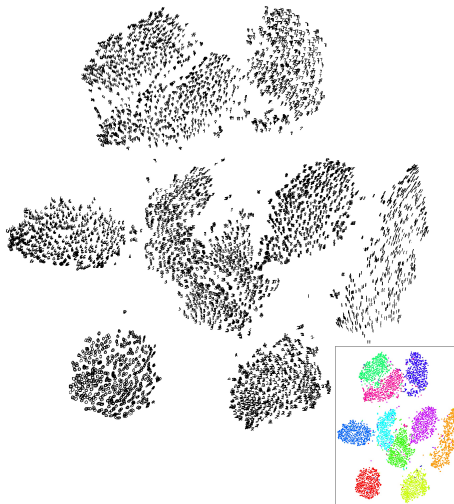


Figure: Taxonomy of nonlinear dimensionality reduction techniques

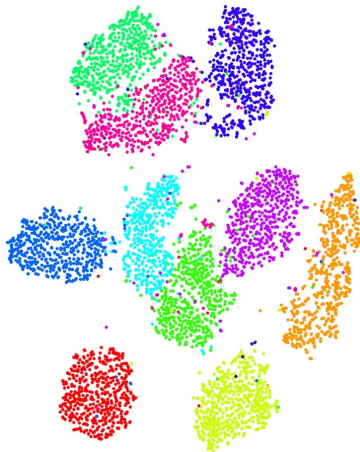
t-Distributed stochastic neighbour embedding (Maaten and Hinton, 2008)

- ✿ t-SNE minimises divergence of two distributions over pairwise similarities of:
 - ▶ input objects (P_i)
 - ▶ corresponding low-dimensional points in the embedding (Q_i)
- ✿ Student-t distribution rather than a Gaussian to compute the similarity between two points in the *low-dimensional space*
- ✿ Assume a function that computes a distance between a pair of objects, e.g. Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|$
- ✿ Minimise cost function (KL-divergence)
$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log p_{j|i} q_{j|i}$$
- + t-SNE compares favourably to other techniques for data visualisation
 - unclear how t-SNE performs on general dimensionality reduction tasks
 - relatively local nature of t-SNE makes it sensitive to the curse of the intrinsic dimensionality of the data
 - not guaranteed to converge to a global optimum of its cost function

t-SNE on MNIST



t-SNE on MNIST



🔥 Dimensionality reduction example:

```
https://github.com/njtwomey/ADS/blob/master/04\_data\_exploration\_and\_visualisation/01\_mnist.ipynb
```

Summary

Recap and Outlook

Descriptive Statistics

- Univariate Analysis

- Measures of Central Tendency

- Moments

- Measures of Shape

- Transforms

- Density Estimation




- Multivariate Analysis

Dimensionality Reduction

- Linear Dimensionality Reduction

- Nonlinear Dimensionality Reduction

Selected References

-  https://en.wikipedia.org/wiki/Bessel's_correction
-  <http://mathoverflow.net/questions/62904/complexity-of-eigenvalue-decomposition>
-  http://scikit-learn.org/stable/modules/random_projection.html

Dimitris Achlioptas. Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 274–281. ACM, 2001.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1): 60–65, 2003.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

RM Sakia. The box-cox transformation technique: a review. *The statistician*, pages 169–178, 1992.

Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10:66–71, 2009.