Chapter XII: Data Pre and Post Processing

1. Data Normalization
2. Missing Values
3. Curse of Dimensionality
4. Feature Extraction and Selection
   4.1. PCA and SVD
   4.2. Johnson–Lindenstrauss lemma
   4.3. CX and CUR decompositions
5. Visualization and Analysis of the Results
6. Tales from the Wild

Zaki & Meira, Ch. 2.2, 2.4, 6 & 8
XII.5: Visualization and Analysis

1. Visualization techniques
   1.1. Projections onto 2D or 3D
   1.2. Other visualizations

2. Analysis of the Results
   2.1. Significance
   2.2. Stability
   2.3. Leakage
Visualization Techniques

• **Visualization** is an important part of the analysis of the data and the results
  – Good visualization can help us see patterns in the data and verify whether our found results are valid
  – Visualization also helps us to interpret the results

• Visualization can also lead us seeing patterns that are not (significant) in the data
  – Visualization alone can never be the basis of analysis
Projecting multi-dimensional data

• The most common visualization takes $n$-dimensional data and projects it into 2 or 3 dimensions for plotting
  – Different methods retain different type of information
• We’ve already seen few projections
  – SVD/PCA can be used in multiple ways
    • Either project the data in the first singular vectors
    • Or do a singular vector scatter plot
• Creating good projections is an on-going research topic
Example: Cereal data

- Data of 77 different cereals
  - We use only 23 Kellogs manufactured cereals in the examples
Example: Clustering

• We clustered the Cereal data using $k$-means
  – But is the clustering meaningful?
  – How do we plot a clustering?

• One idea: project the data into 2D and mark which point belongs to which cluster
  – Question: will we see the clustering structure?
Cereals in SVD Scatter
Cereals in PCA w/ Gaussian kernel

Zaki & Meira Ch. 7.3
Cereals and multidimensional scaling
Cereals and Isomap

Cereals and Laplacian eigenmaps

Cereals and neighbourhood-preserving embedding
Non-projection visualizations

• Projections are not the only type of visualizations
  – Again, we have seen other visualizations before
  – These are often a bit more specific
    • But not always…
Heat maps

- Original
- Normalized
We further verified the detection of false zeros and ones by preparing two datasets, based on data parametrized by $n_t = 10$ and $n_s = 10$. For the first set, we selected 100 random ones, and flipped them to zero (false zeros). For the second set, we randomly selected 100 zeros, and changed them to ones (false ones). We then performed the analysis, and computed median probability for all zeros to be alive in the first dataset and median probability for all ones to be alive in the second datasets. The probabilities corresponding to 92 of 100 added false zeros was below the median, and 88 of 100 added false ones were above median. The differences are statistically significant when compared to the null hypothesis that the false zeros or ones are equally likely to end up above or below the median (Fisher Sign Test). The median probability that the (site, genus) pair an inserted false one is alive is 0.004, and the median probability that an inserted false zero is alive is 0.92.

We also tested a model where each taxon has its own $c$ and $d$ parameters for false one and false zero. The results of the MCMC runs were almost identical to the ones obtained for the model with one $c$ and one $d$ parameter (unpublished data).

**Discussion**

We have described a probabilistic model for paleontological data and shown that MCMC methods can be used to obtain samples from the posterior distribution of the parameters. The parameters of the model have a natural interpretation, and the hard sites enable us to insert existing prior knowledge of the ordering in a natural way.

The task of finding the optimal ordering, or knowing for certain that a given ordering is optimal, is a very difficult problem. MCMC methods have the advantage of being able to explore various parts of the parameter space, but the issue of guaranteeing convergence of the sampling is always present in these methods. We have solved the problem of convergence by sampling 100 chains in parallel, and taking into account only the chains having the best log-likelihood. We have also checked that the pair-order matrices predicted by these best chains are consistent with each other. This way, we can state with reasonable confidence that our results are indeed an accurate description of the posterior distribution of the model. We also tested the method by adding false zeros and ones to the data randomly, and checking that they were identified correctly.

The results show that for generated data the method is able to reconstruct orderings and locate outliers with excellent accuracy. For the data on large late Cenozoic mammals, the results indicate a high level of agreement with existing orderings and correctly capture the basic feature of paleontological data that false absences are likely to be common and false presences rare.

For the past 40 years the main stratigraphic framework for the study of the Cenozoic land mammals from Europe has been the MN system [30–33]. The MN system rests on a complicated base of taxon appearances and associations that has been...
Dendrograms
Heat maps with dendrograms
Radar charts

Figure 408: Temperament cluster centers in the NFBC 66 best model as starplots normalized data. The male and female seem similar to each other despite having been learned independently. The subscales are as follows:

HAz1: anticipatory worry  
HAz2: fear of uncertainty  
HAz3: shyness  
HAz4: fatigability

NSz1: exploratory excitability  
NSz2: impulsiveness  
NSz3: extravaganza  
NSz4: disorderliness

RDz1: sentimentality  
RDz3: attachment  
RDz4: dependence

Stable, persistent, not very impulsive

High socio-economic status and education

Shy, pessimistic, prefer routines and privacy

Low socio-economic status, high levels of depression and schizophrenia
Parallel coordinates
Maps...
Analysis of the results

• Without analysis, there’s not much point in doing data mining

• The analysis should be done by domain experts
  – People who know what the data contains and how to interpret the results

• Data mining is about finding surprising things…
  – … so domain experts are needed to
    • tell if the results really are surprising
    • verify that the surprising results are meaningful in the context
Significance of the results

• Statistical significance tests can be applied to the results
  – But they require forming the null hypothesis
• Too weak null hypothesis $\Rightarrow$ even significant results are not necessarily significant at all
  – But strong null hypotheses are harder to test
• We rarely can use (full-blown) exact tests
• Sometimes we can use asymptotic tests
• In other times we can use permutation tests
Significance testing example (1)

• We want to test the significance of association rule $X \rightarrow Y$ in a data with $n$ rows

• Null hypothesis 1: Itemsets $X$ and $Y$ both appear in the data but their tidsets are independent random variables
  
  – Each transaction contains $X$ with probability $supp(X)/n$

• The probability for $supp(XY)$ is a tail of a binomial distribution for $p = supp(X)supp(Y)/n^2$

\[
\sum_{s=supp(XY)}^{n} \binom{n}{s} p^s (1 - p)^{n-s}
\]
Significance testing example (2)

- Null hypothesis 2: $X \rightarrow Y$ does not add anything over a generalization $W' \rightarrow Y$, where $W \subsetneq X$ assuming the row and column marginals are fixed
- The odds ratio measures the odds of $X$ occurring with $Y$ versus the odds of $W$ (but not other parts of $X$) occurring with $Y$
  - For any $W$, we can consider the null hypothesis that odds ratio $= 1$ ($X \setminus W$ is independent of $Y$ given $W$)
  - We can compute the $p$-value for this hypothesis using hypergeometric distribution
  - We can test null hypothesis 2 by computing the $p$-values for all generalizations of $X$

Z&M Ch. 12.2.1
Significance testing example (3)

• Null hypothesis 3: The confidence of the rule is explained merely by the row and column marginals of the data
  — Confidence can be replaced with any other interest measure

• This we can test by generating new data sets with same row and column marginals
  — If many-enough of them contain rules with higher confidence, we cannot reject our null hypothesis
  — Generating such data can be done e.g. with swap randomization

• This is called permutation test

Z&M Ch. 12.2.2
Stability

- The **stability** of a data mining result refers to its robustness under perturbations
  - E.g. if we change all the numerical values a bit, the clusterings shouldn’t change a lot
  - We can also remove individual rows/columns or make more data unknown

- Stability should be tested after the results have been obtained
  - Run the same analysis with perturbed data
Stability example (1)
Stability example (2)
Leakage

- **Leakage** in data mining refers to the case when prediction algorithm learns from data it should not have access to
  - Problem as the quality is assessed using already-historical test data
  - E.g. INFORMS’10 challenge: predict the value of a stock
    - Exact stock was not revealed
    - But “future” general stock data was available! ⇒ 99% AUC (almost perfect prediction!)
  - More subtle one’s exist
    - E.g. removing a crucial feature creates a new type of correlation
1. Working with non-CS folks
Talk their language!

Archeotype

Red queen's problem

Voronoï tessellation

NP-hard
Data is dirty

Figure 2.1: Matrix of the data used in the migraine study described in Chapter 4.3, observed values in white and missing values in black. X-axis, the 194 original variables. Y-axis, 6283 individuals in the dataset, ordered by ID. Note the obvious non-randomness of missing data. This picture was originally not drawn; only one with rows of the data file was used. Strong correlations of clusters and other variables to running ID very soon called for this one, too.
Not all data is BIG
It’s all just constants

Figure 2. The Data Matrix for the Dataset with $n_t = 10$ and $n_s = 10$.

The sites have been ordered by $E(f_p(n))$ and the genera by $E(f_a(m))$ (top).

Probability that genus $m$ is alive on site $n$ in the dataset specified by $n_t = 10$ and $n_s = 10$ (middle). Probability that one is false (bottom). Black color denotes probability of one, and white probability of zero.

DOI: 10.1371/journal.pcbi.0020006.g002
The best algorithm is the algorithm you have with you.
Wear the analysis possibly because herbivore distributions are most directly influenced by the maritime–continental climate gradient. The species with the highest grid cell incidence give more coherent clusters than other groups (Fig. 1). Those with an incidence of 10–20% give coherence values approaching those of all species and small mammals, but higher incidence values give lower coherence, perhaps because the species with the highest incidence are few and widespread. The subset of species 'at risk' gives spatially the least coherent clusters found in this study, even less coherent than seen for large mammals (Fig. 1). The regional divisions identified by the clusterings show significant differences in the values of basic climate variables and elevation (Table 4). All cluster pairs in the 'all species' clustering seen in Fig. 3 differ significantly in at least two environmental variables, and most cluster pairs differ in all of the variables (Table 4a,b). For almost all groupings temperature is the variable for which the cluster pairs have the most significant differences (Table 4c). For precipitation, the number of significant differences is also high. For all environmental variables the set 'species at risk' has the smallest number of significantly different cluster pairs, while the species set with the largest number of significant differences is different for each considered variable. However, more important than these relatively minor differences is the high overall percentage of significant differences. The results of the ANOVA tests complete with P-values for all of the species groupings are provided as Table S3 in the supplementary material.

DISCUSSION
We find that Europe can be divided into coherent subregions based on the distributions of mammal species. We also find a high degree of geographical coherence displayed by the clusters, and consistency in the basic spatial pattern among non-overlapping subsets of the data and despite changes in the number of clusters. These observations, in combination with the environmental contrast observed between the clusters and the concordance of the geographical cluster pattern with the EnS environmental stratification strongly suggest that the clusters represent real biological units rather than arbitrary constructs generated by the clustering algorithms. We take this to indicate that, even in present-day Europe with its long history of intensive human presence, the main controls on mammalian metacommunity distributions remain
4.8 Experimental Evaluation

The 20Newsgroups data.
This data is a bag-of-words representation of almost 20,000 Usenet news posts spread (almost) evenly in 20 newsgroups. As with the Abstracts data, the data was first stemmed using Porter Stemmer, and all words appearing in less than 36 or in more than 999 posts were removed. This data set was only used in this chapter; for subsequent chapters, a different preprocessing was applied to yield a smaller data set.

The NOW data.
This data contains information about species’ fossils found in specific palaeontological sites in Europe [F+03]. The data is preprocessed following Fortelius et al. [FGJM06].

4.8.5 Randomization of the Data
A simple reconstruction error, even if it is small, does not necessarily guarantee that the data contains the latent structure we were trying to find. It is possible that the results were obtained by chance. The standard method to distinguish these cases is to measure the statistical significance of the results against the null hypothesis, which – broadly speaking – is ‘results were obtained by chance’.

5 http://people.csail.mit.edu/jrennie/20Newsgroups/

Itkonen: Proto-Finnic Final Consonants: Their history in the Finnic languages with particular reference to the Finnish dialects, part I: 1, Introduction and The History of -k in Finnish, 1965
Know the math of the domain
The species' set. The clustering is the best out of 100 clustering runs in terms of squared error. The results of the Euclidean distance (a) and the Hellinger distance (b). The differences in the delineation of geographical regions considerably different from those presented by Heikinheimo et al. (2007) (Fig. 1). Differences were most significant in the values of basic climate variables. The subset of species with the highest incidence are few and widespread. The subset of species with the incidence of 10–20% give coherence values approaching those based on the distributions of mammal species. We also find a high degree of geographical coherence displayed by the mammal data cells in 12 clusters with the 'all significantly different cluster pairs, while the species set with the variables the set 'species at risk' has the smallest number of controls on mammalian metacommunity distributions remain for almost all groupings environmental variables, and most cluster pairs differ in all of the variables (Table 4a,b). For almost all groupings temperature is the variable for which the cluster pairs have the most significant differences. The results of the anova tests complete distance among sites (Legendre & Gallagher, 1992)

REFERENCES

European land mammal biogeographical re-

Correspondence...
Data mining = voodoo science
The response from several social scientists has been rather unappreciative along the following lines: “Where is your hypothesis? What you’re doing isn’t science! You’re doing DATA MINING!”

http://andrewgelman.com/2007/08/a_rant_on_the_v/
The clash of paradigms

- Form a hypothesis
- Design a test
- Collect the data
- Test hypothesis
- Rinse and repeat

- Take somebody else’s data
- Pick an algorithm
- Run the algorithm
- Analyse the results
- Rinse and repeat
Summary

• Think before you do
• Think while you do
• Think what you just did

• Real-world data analysis requires care and expertise
• Visualizations are powerful tools in data analysts' toolbox
  – With great power comes great responsibility
• Data mining might be voodoo science
  – But who wouldn’t want to know the voodoo?