Challenges in data management and analysis

Heikki Mannila
HIIT Basic Research Unit
Helsinki University of Technology &
University of Helsinki
Heikki.Mannila@cs.helsinki.fi
March 21, 2006

Outline

- Type A and type B areas of research
- Data management
- Data analysis
- An example
- Summary

Computation

- Becoming more and more central in many aspects of science
- Where is the bottleneck?
  - Memory?
  - Availability of processor cycles?
  - Software?
  - Modeling the data?
  - Data?

Computational science

- Two types of areas
  - Type A
  - Type B
- This division has implications for the type of computational tools that are needed

Type A areas

- A (moderately) good understanding of the basic phenomena that create the observational data
  - Meteorology, many parts of physics and chemistry, ...
- The basic laws have lots of predictive power
  - Prediction from first principles, or at least from a lower level
- Computation: solving equations, simulating systems, etc.

Type B areas

- The fundamental laws cannot really be used to predict effectively what the phenomena look like
- Example: ecology
  - Evolution is a fundamental principle, but it is hard to use it to predict what the community structure is like
- Example: linguistics
- Computation in type B: exploratory data analysis; modeling the data

Remarks

- Both types are needed
- The division into type A and type B is not clear-cut
- ”Computational science” has traditionally meant mostly type A areas

Data management challenges

- Amount of data
- Many different types of data
  - Not just observations x variables
  - Several types of entities
  - Numbers, sequences, images, ...
- Multiple sources of data: calibration
- Secondary data: collection for some other use

Amount of data

- Some examples where the sheer amount of data makes storing it difficult

- CERN; some imaging applications; etc.

- In most cases it is the *different types* or *different sources* of data that causes the problems, not the volume of the data

Different types of data

- Example: molecular biology

- DNA sequences, proteins, genes, motifs, pathways, metabolic reactions, expression data, markers, phenotypes, literature, ...

- Huge variety of different types of data

- Challenges: how to cope with this type of data; using different types in analysis

Simpler yet difficult example: geographic data

- GIS data is there, in a multitude of formats, in many different governmental organizations
- Land use, transport information, biodiversity, ...
- Getting access can be a problem
- Technical and organizational problem

Multiple sources of data

- How to be sure that the data coming from different sources really is the same?
- Data measured at different times?
- Calibration: very hard when measurement technology is developing fast
- Going back can be impossible or very costly

Secondary data

- Measured or collected for some other reason than your study

- How to use such data?

Example

- What is a good sample of written Finnish?
- There is a lot of it on the web
- What is a good sample?

Challenges in data analysis

- Large number of observations
- Large number of variables

  à many possible models

Model = probabilistic model

- Efficient algorithms: discrete and continuous techniques
- Robustness of results
- Significance testing

Large number of observations

- How big a problem is this?
- Consider the model estimation problem: given a class of possible models, find the best one
- For i.i.d. observations the loglikelihood is a sum over all observations

- *Linear* in the number of observations
- 10-fold increase in the size of the data can be handled with 10 processors

Large number of variables

- Number of possible explanations for the data grows at least *exponentially* in the number of variables
- Curse of dimensionality
- Are the solutions robust?
- Example: nearest neighbor in high-dimensional spaces

Handling large model spaces

- For simple model classes one can find the optimal model efficiently
  - E.g., linear regression with no interactions between variables
- What if there are very many possible hypotheses?
  - E.g., selecting the variables and the interaction terms
  - Problem becomes NP-complete with respect to the number of variables
  - Equally difficult as a very large class of other problems

Sometimes large model spaces are not that difficult

- Given a timeseries with n points and an integer k
- Find the best piecewise constant approximation of the series using k pieces
- An exponential number of possible models: $\binom{n}{k}$
- The best model can be found in $O(n^2 k)$ operations by using dynamic programming

But in most cases they are

- Approximate methods are needed
- Guarantees on the quality of the result
- Robustness

Example: mixture models

- Modeling data by assuming there are several clusters/groups in it
- Trying to see if the data would stem from a combination of sources
- Powerful (and old) idea
- Can be used for discrete, continuous, and structured data
- Large model space

Expectation-maximization (EM) algorithm

- Trying to see if the data would stem from a combination of sources

- Assume we know the sources: assign each data point to the sources according to the likelihood
- Assume we know which data points stem from the same source: form a new source from them
- Iterate

Expectation-maximization (EM) algorithm

- A beautiful method, widely applicable

- Converges
- ... to a local optimum
- In high-dimensional spaces different initializations can give very different answers

- Characterizing the properties of the method?
- Robustness of the solutions in high dimension?

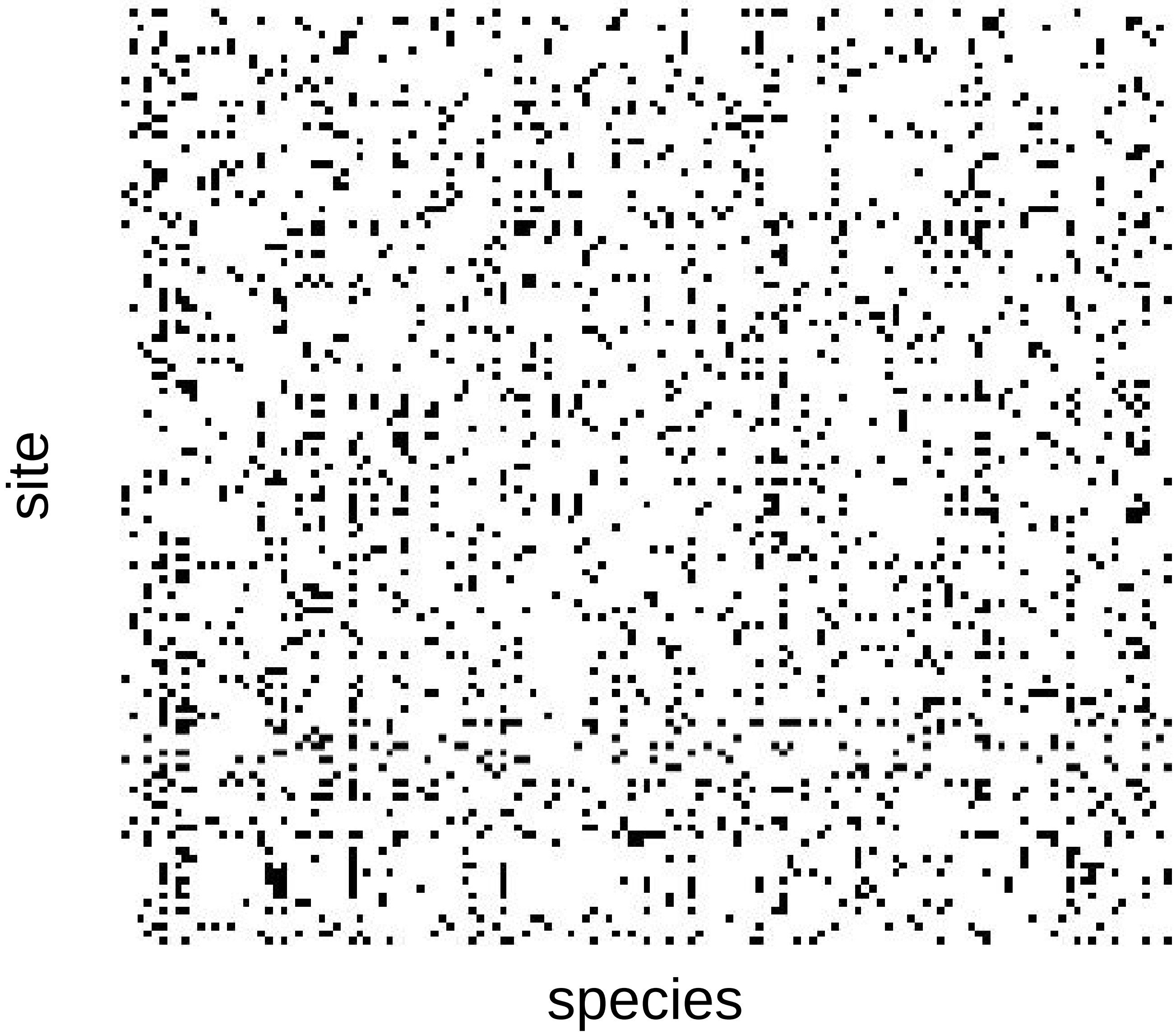
Significance testing

- A test score S(D) computed from a data set D containing many different types of data
- How do we know whether S(D) is in any way significant?
- Analytical results are typically unavailable
- *Randomization*
  - But how to randomize?
  - What is the null hypothesis
- Lots of interesting computational problems

Example: different models for the same problem

- Paleontological data (M. Fortelius)
  - Fossil sites, species
- The amount of data is not that great
- Lots of data management problems
  - Many different sources of the data
- Many different computational problems
- Many approaches are possible

Example



Black:
species
found

White:
not found

Example problem: seriation

- What is the correct order for the sites?
- A difficult problem (layer information is not available, neither are radiometric etc. dates)
- Background information:

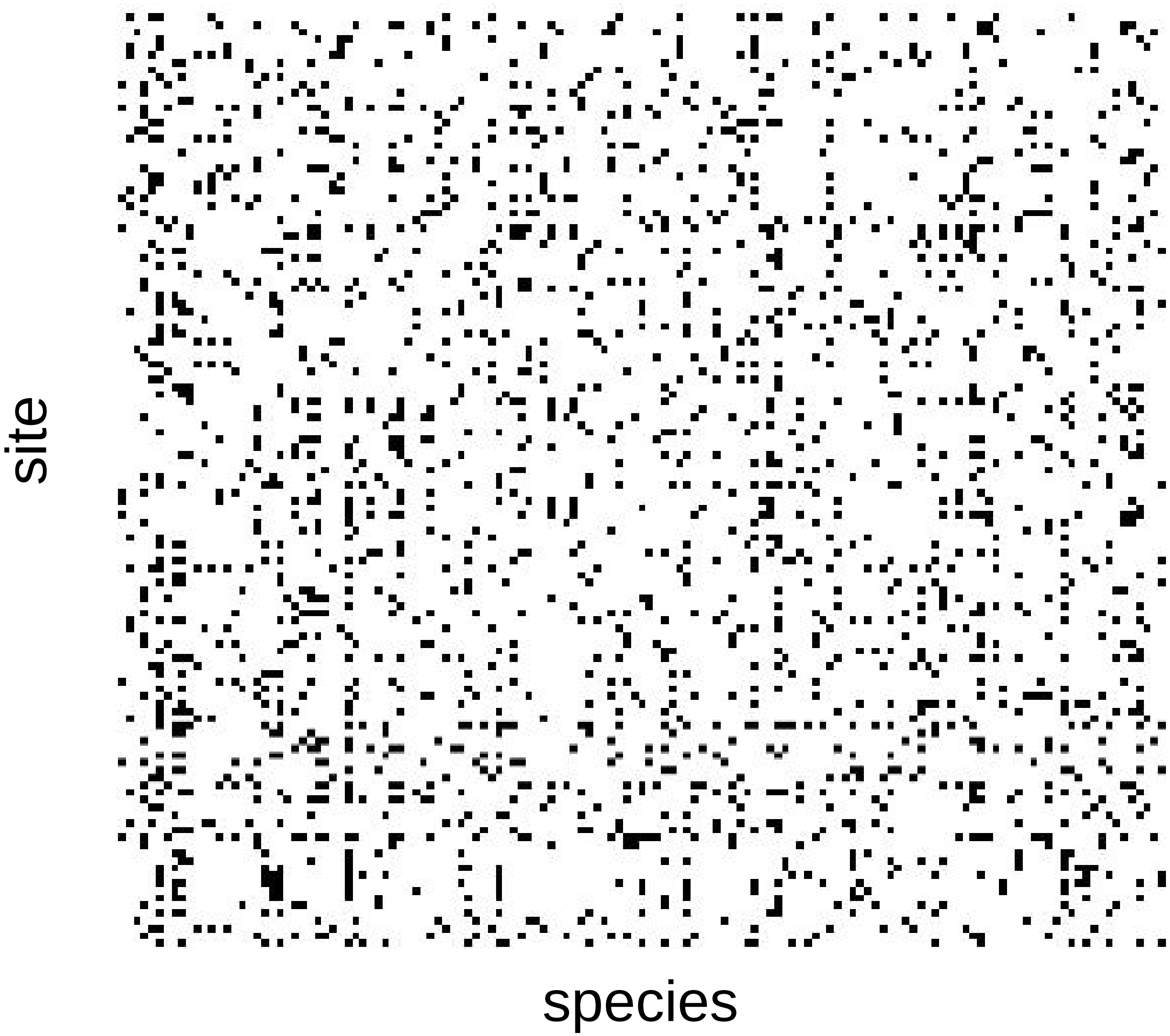
  species do not vanish and reoccur

  – No Lazarus events

Seriation problem

- Model space: all possible orderings or all possible partial orders of the sites
- Score function: how many times a 0 is between two 1s (number of Lazarus events)
- Approaches:
  - Discrete: TSP-type of approaches
  - Spectral (eigenvalue method)
  - Bayesian probabilistic model + MCMC

Site-species -matrix



Discrete approach

- Search for an ordering that minimizes the number of 0s between 1s
- A TSP-like problem

- Local search algorithms work OK

Eigenvalue approach

- Spectral ordering

- Compute a similarity measure *s(i,j)* between sites (e.g., dot product)

- Laplacian L(i,j)

$$L(i,j) = \begin{cases} -s(i,j), & i \neq j \\ \sum_k s(i,k), & i = j \end{cases}$$

- The eigenvector *v* corresponding to the second smallest eigenvalue of L satisfies

$$\sum_i v_i = 0,\ \ \sum_i v_i^2 = 1,\ \text{and}\ \sum_i s(i,j)(v_i - v_j)^2 = 1\ \text{is minimized.}$$

- Maps the points to 1-d, keeping similar points close to each other
- The values v_i can be used to order the points

MCMC approach

- Build a full probabilistic model
- Use MCMC method to sample parameter values from the posterior distribution

Partial orders

- Ordering all sites does not perhaps make sense
- Search for a partial order among the sites
- A combinatorial optimization problem

Differences between the methods

- Accuracy?
- Robustness?
- Model class?
  - Is a total order really what we would like to get?
  - Additional (useful) parameters in MCMC

- Speed (is the method feasible)?

Properties of the methods

- Local search
  - We know what it does
  - Quite slow
- Spectral technique
  - Very fast
  - Robustness?
- MCMC
  - Detailed information
  - Slow
- Partial orders
  - Good but large model class

Conclusions

- Where is the bottleneck?

- Type A and type B areas
- Data management and curation
- Different approaches to the same problem

Some of the challenges

- Handling *many different types* of data
- *Large model spaces*: efficient methods
- Robustness
- *Significance testing*
- ...