CIMI Machine Learning Workshop 09.11.2015

The Baire Metric and Ultrametric for Linear Computational Time Hierarchical Clustering: Applications and Implementations

Fionn Murtagh

In summary

We use random projections of our data – of our cloud of points. When in high dimensions, such projections can even approximate a projection into a lower dimension orthonormal space. (This is what we want in PCA or CA, etc.)

But we have pursued a new objective: to use random projection in order to approximate our data cloud such that it is rescaled well. What we mean by that: that its clustering properties are well respected – the interrelationships among points in our cloud of points.

In summary

Having rescaled our data, based on random projections, we next show how we can simplify that mapping of our data cloud.

Then we want to read off the clusters. We show that the Baire metric, that is also an ultrametric, is an excellent framework for this. (The Baire metric, as will be shown, is the "longest common prefix metric".)

Applications in Search and Discovery

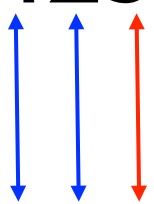
- First, agglomerative hierarchical clustering; or: “hierarchical encoding” of data.
- Ultrametric topology, Baire distance.
- Clustering of large data sets.
- Hierarchical clustering via Baire distance using SDSS (Sloan Digital Sky Survey) spectroscopic data.
- Hierarchical clustering via Baire distance using chemical compounds.

Next: the Baire (ultra)metric

Baire, or longest common prefix distance – and also an ultrametric

An example of Baire distance for two numbers (x and y) using a precision of 3:

$x = 0.425$



$y = 0.427$

Baire distance between x and y:

$d_{\mathcal{B}}(x, y) = 10^{-2}$

Base ($\mathcal{B}$) here is 10 (suitable for real values)

Precision here = |K| = 3

That is:

k=1 -> $x_k = y_k$ -> 4

k=2 -> $x_k = y_k$ -> 2

k=3 -> $x_k \neq y_k$ -> 5≠7

On the Baire (ultra)metric

- Baire space consists of countable infinite sequences with a metric defined in terms of the longest common prefix *[A. Levy. Basic Set Theory, Dover, 1979 (reprinted 2002)]*

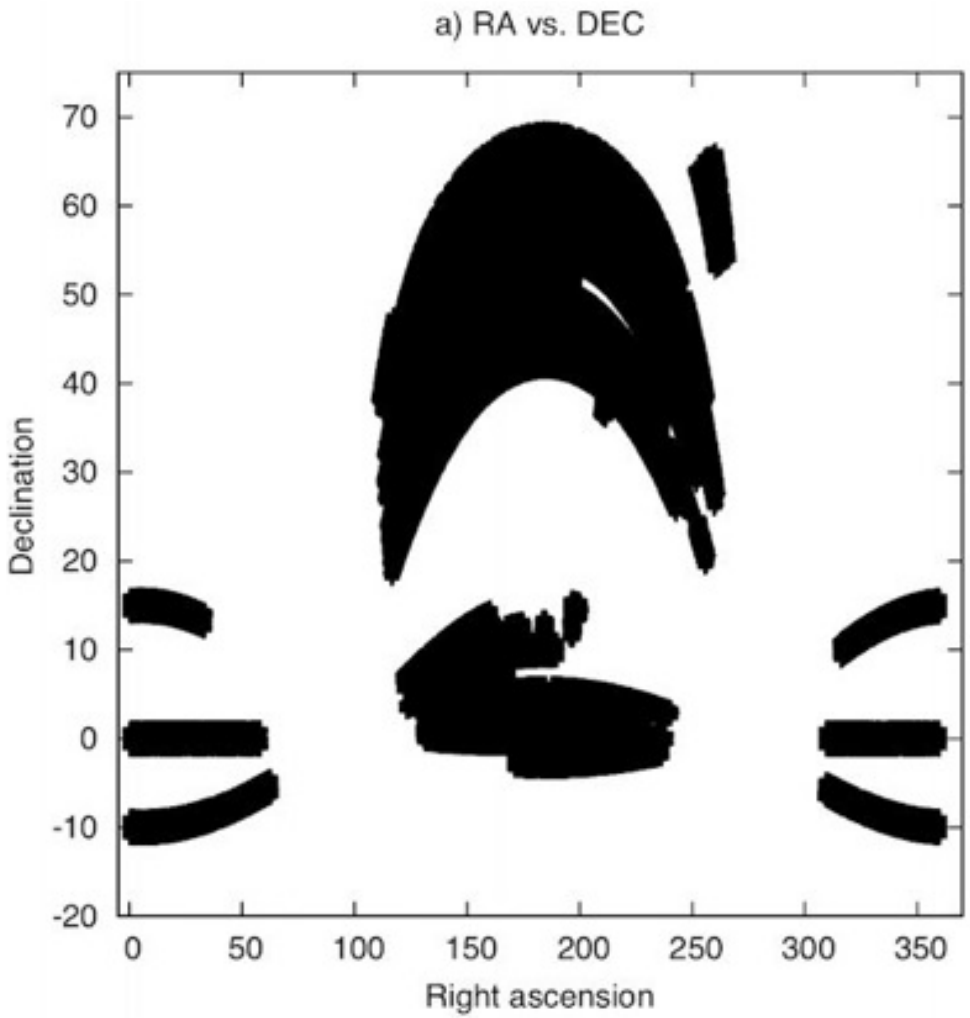
- The longer the common prefix, the closer a pair of sequences.

- The Baire distance is an ultrametric distance. It follows that a hierarchy can be used to represent the relationships associated with it. Furthermore the hierarchy can be directly read from a linear scan of the data. (Hence: hierarchical hashing scheme.)

- We applied the Baire distance to: chemical compounds, spectrometric and photometric redshifts from the Sloan Digital Sky Survey (SDSS), and various other datasets.

- A subset was taken of approximately 0.5 million data points from the SDSS release 5.
- These were objects with RA and Dec (Right Ascension and Declination, and spectrometric redshift, and photometric redshift). Problem addressed: regress one redshift (spectro.) on the other (photo.).
- Baire approach used, and compared with k-means.
- 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.
- Random projections used on normalized compound/ attribute values.
- Baire approach used; also another approach based on restricting the precision of the normalized compound/ attribute values.

SDSS (Sloan Digital Sky Survey) Data



- We took a subset of approximately 0.5 million data points from the SDSS release 5
  [reference: D'Abrusco et al]
- declination (Dec)
- right ascension (RA)
- spectrometric redshift
- photometric redshift.

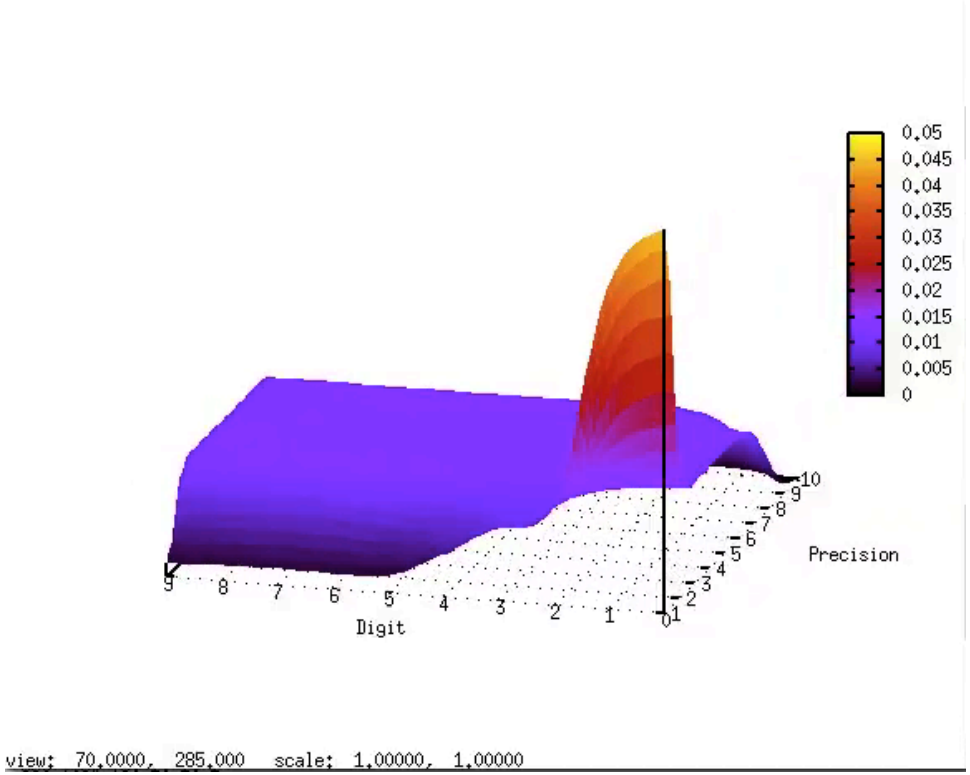
- Dec vs RA are shown

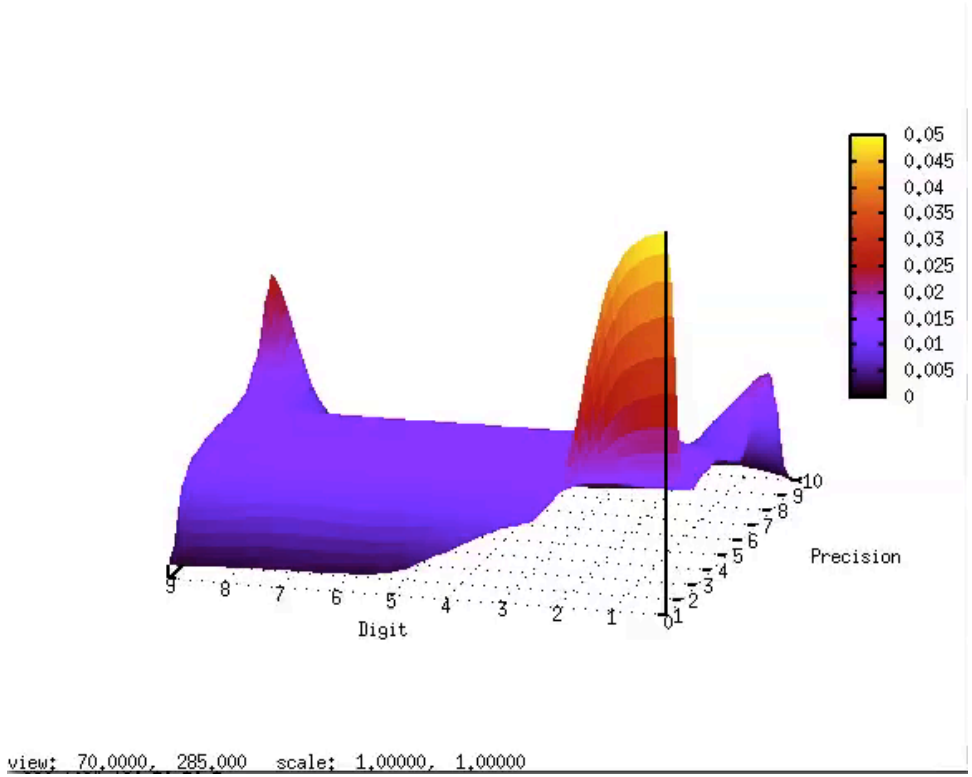
Data – sample

RA	DEC	spec. redshift	phot. redshift
145.4339	0.56416792	0.14611299	0.15175095
145.42139	0.53370196	0.145909	0.17476539
145.6607	0.63385916	0.46691701	0.41157582
145.64568	0.50961215	0.15610801	0.18679948
145.73267	0.53404553	0.16425499	0.19580211
145.72943	0.12690687	0.03660919	0.06343859
145.74324	0.46347806	0.120695	0.13045037

- Motivation - regress z_spect on z_phot
- Furthermore: determine good quality mappings of z_spect onto z_phot, and less qood quality mappings
- I.e., cluster-wise nearest neighbour regression
- Note: cluster-wise not spatially (RA, Dec) but rather within the data itself

Perspective Plots of Digit Distributions





- On the left we have z_spec where three data peaks can be observed. On the right we have z_phot where only one data peak can be seen.

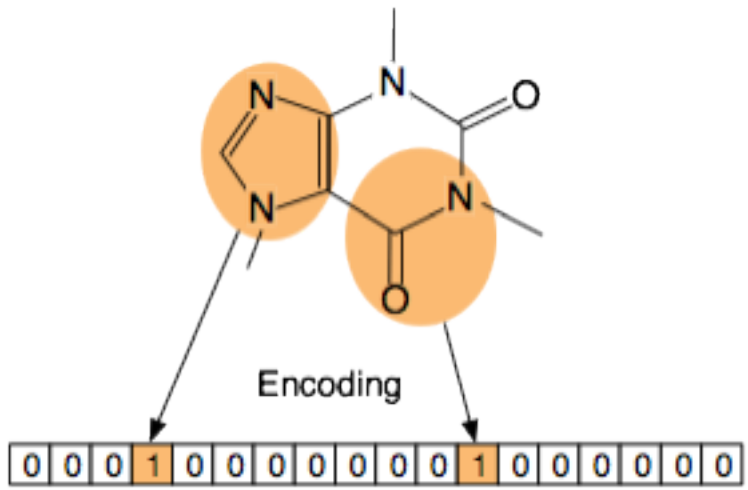
Framework for Fast Clusterwise Regression

- 82.8% of z_spec and z_phot have at least 2 common prefix digits.
- I.e. numbers of observations sharing 6, 5, 4, 3, 2 decimal digits.
- We can find very efficiently where these 82.8% of the astronomical objects are.
- 21.7% of z_spec and z_phot have at least 3 common prefix digits.
- I.e. numbers of observations sharing 6, 5, 4, 3 decimal digits.

- Next - another case study, using chemoinformatics - which is high dimensional.

- Since we are using digits of precision in our data (re)coding, how do we handle high dimensions?

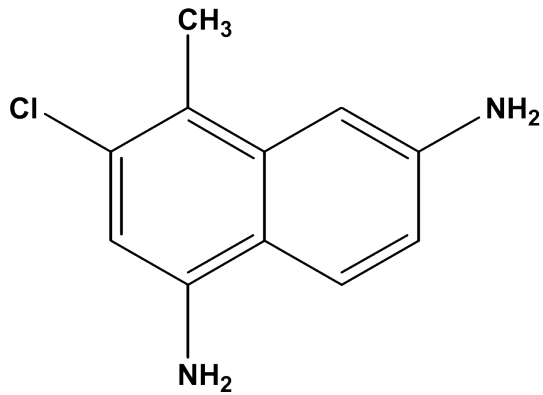
Baire Distance Applied to Chemical Compounds



Matching of Chemical Structures

- - Clustering of compounds based on chemical descriptors or chemical representations, in the pharmaceutical industry.
- - Used for screening large corporate databases.
- - Chemical warehouses are expanding due to mergers, acquisitions, and the synthetic explosion brought about by combinatorial chemistry.

Binary Fingerprints



Encode

1 0 0 0 1 0 0 0 1 ...1

Fixed length bit strings with encoding schemes

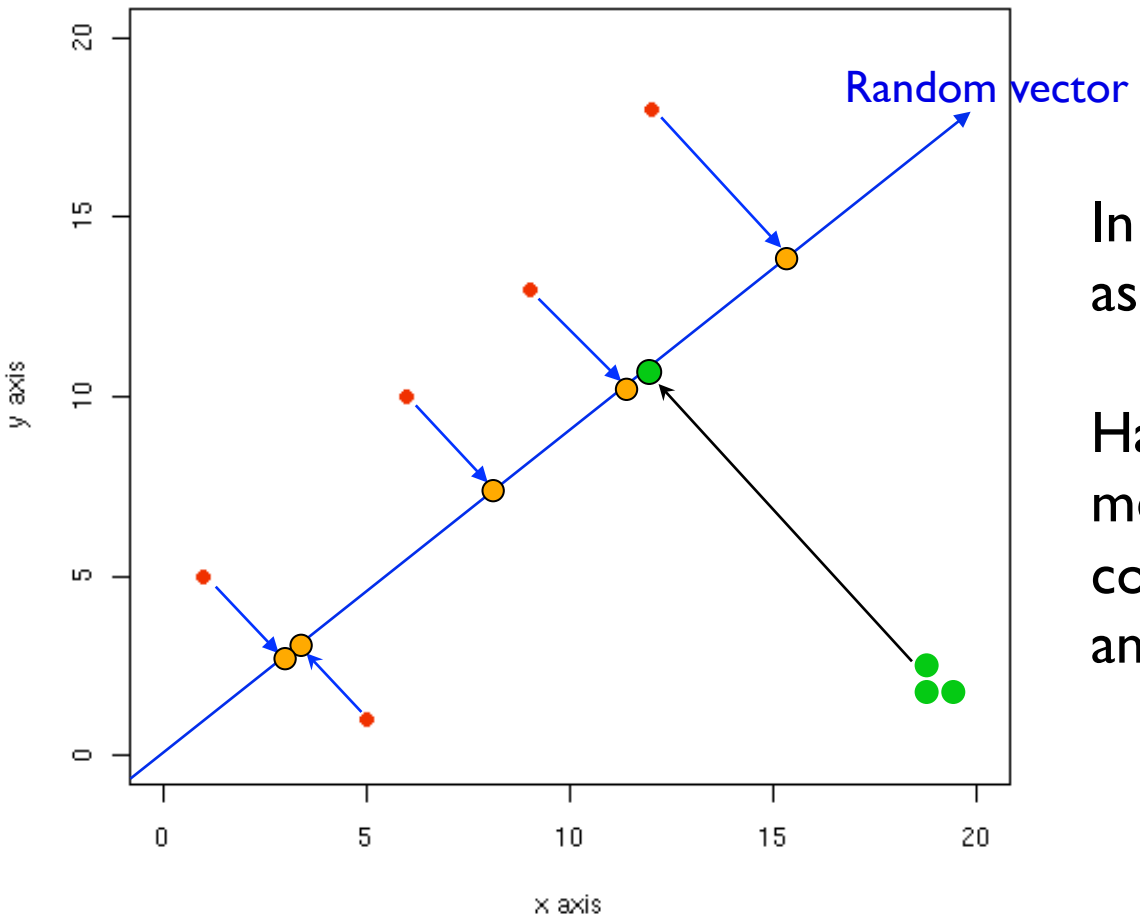
Daylight, MDL

BCI *(We will be using this)*

Chemoinformatics clustering

- 1.2 million chemical compounds, each characterized by 1052 boolean presence/absence values.
- Firstly we note that precision of measurement leads to greater ultrametricity (i.e. the data are more hierarchical).
- From this we develop an algorithm for finding equivalence classes of specified precision chemicals. We call this: data "condensation".
- Secondly, we use random projections of the 1052-dimensional space in order to find the Baire hierarchy. We find that clusters derived from this hierarchy are quite similar to k-means clustering outcomes.

Random projection and hashing



In fact random projection here works as a class of hashing function.

Hashing is much faster than alternative methods because it avoids the pairwise comparisons required for partitioning and classification.

If two points (p , q) are close, they will have a very small |p-q| (Euclidean metric) value; and they will hash to the same value with high probability; if they are distant, they should collide with small probability.

- Normalize chemical compounds by dividing each row by row sum (hence “profile” in Correspondence Analysis terms).
- Two clustering approaches studied:
- Limit precision of compound / attribute values. This has the effect of more compound values becoming the same for a given attribute. Through a heuristic (e.g. interval of row sum values), read off equivalence classes of 0-distance compounds, with restricted precision. Follow up if required with further analysis of these crude clusters. We call this “data condensation”. For 20000 compounds, 1052 attributes, a few minutes needed in R.
- Second approach: use random projections of the high dimensional data, and then use the Baire distance.

Summary Remarks on Search and Discovery

- We have a new way of inducing a hierarchy on data
- First viewpoint: encode the data hierarchically and essentially read off the clusters
- Alternative viewpoint: we can cluster information based on the longest common prefix
- We obtain a hierarchy that can be visualized as a tree
- We are hashing, in a hierarchical or multiscale way, our data
- We are targeting clustering in massive data sets
- The Baire method - we find - offers a fast alternative to k-means and a fortiori to traditional agglomerative hierarchical clustering
- At issue throughout this work: embedding of our data in an ultrametric topology

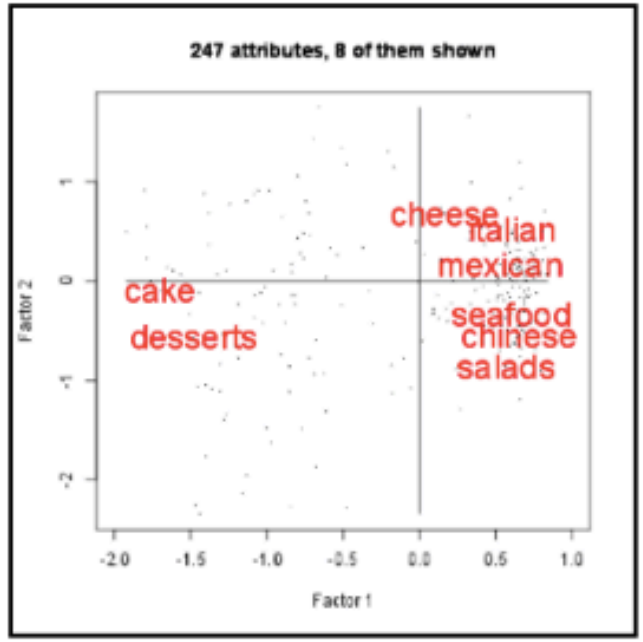
- Quite a different starting point:
- Using Apache Lucene and Solr for indexing, storage, and query support
- The following slide is showing where we used 152,998 cooking recipes, with 101,060 unique words in them.

Apache Solr

Navigation aid:
Cursor over cell below for Factor1, Factor 2 coordinates.

Example of query around "cake":
xcoord:[-1.6 TO -1.55] &&
ycoord:[-0.11 TO -0.09]

cake	cheese	Mexican	Italian
desserts	seafood	Chinese	salads



Examples: Simple Spatial Group By

Find: Submit Reset

152998 results found in 9 ms Page 1 of 15300

"21" Club Rice Pudding More Like This

Id: mm000001102.txt

F1 (xcoord): -0.7341409 **F2 (ycoord):** -0.09961348

Recipe: Categories: Dessert Yield: 10 Servings 1 qt Milk 1 pt Heavy cream 1/2 ts Salt 1 Vanilla bean 3/4 c Long-grained rice 1 c Granulated sugar 1 Egg yolk 1 1/2 c Whipped cream Raisins (optional) From: Bobb1744 at aol.com Date: Tue, 23 Apr 1996 13:28:27 -0400 Recipe By: Aunt Salli's In a heavy saucepan, combine the milk, cream, salt, vanilla bean and 3/4 cup of the sugar and bring to a boil. Stirring well, add the rice. Allow the mixture to simmer gently, covered, for 1 3/4 hours over a very low flame, until rice is soft. Remove from the heat and cool slightly. Remove the vanilla bean. Blending well, stir in the remaining 1/4 cup of sugar and the egg yolk. Allow to cool a bit more. Preheat the broiler. Stir in all but 2 tablespoons of the whipped cream; pour the mixture into individual crocks or a souffle dish. (Raisins my be placed in the bottom of the dishes, if desired.) After spreading the remaining whipped cream in a thin layer over the top, place the crocks or dish under the broiler until the pudding is lightly browned. Chill before serving. MC-RECIPE at MASTERCOOK.COM MASTERCOOK RECIPES LIST SERVER MC-RECIPE DIGEST V1 No.55 From the MasterCook recipe list. Downloaded from Glen's MM Recipe Archive, http://www.erols.com/hosey. ----- ----------

"A Greeting" More Like This

Id: mm000001103.txt

F1 (xcoord): 0.6679912 **F2 (ycoord):** -0.01193456

Recipe: Categories: None Yield: 1 Servings Recipe by: steelman at execpc.com VLF Vegetarian Recipes This is a kind of very-low-fat vegetarian starter's kit, containing recipes and other information (category "text") collected from a variety of computer sources, and our own experiences. The category marked "family approved" are recipes we use regularly, that we all like. The others were collected, but for the most part have not been tried. For a great collection of recipes, we recommend michelle dick's www.fatfree.com, and the Usenet newsgroup alt.food.fat-free. This type of diet was recommended by our cardiologist to help reverse Merle's heart disease. Whether it is working for that is not yet known, but we both DO feel healthier eating this way. Also, we find the food very satisfying and filling, but low-calorie enough to allow for significant weightloss . And eating this way is getting easier all the time. If you walk SLOWLY through the grocery store each week, you will see new "FAT-FREE" food items each week, increasing your choices. Merle and Deb Steelman steelman at execpc.com 10/2/95 File ftp://ftp.idiscover.co.uk/pub/food/mealmaster/recipes/mmdia006.zip ----- ----------

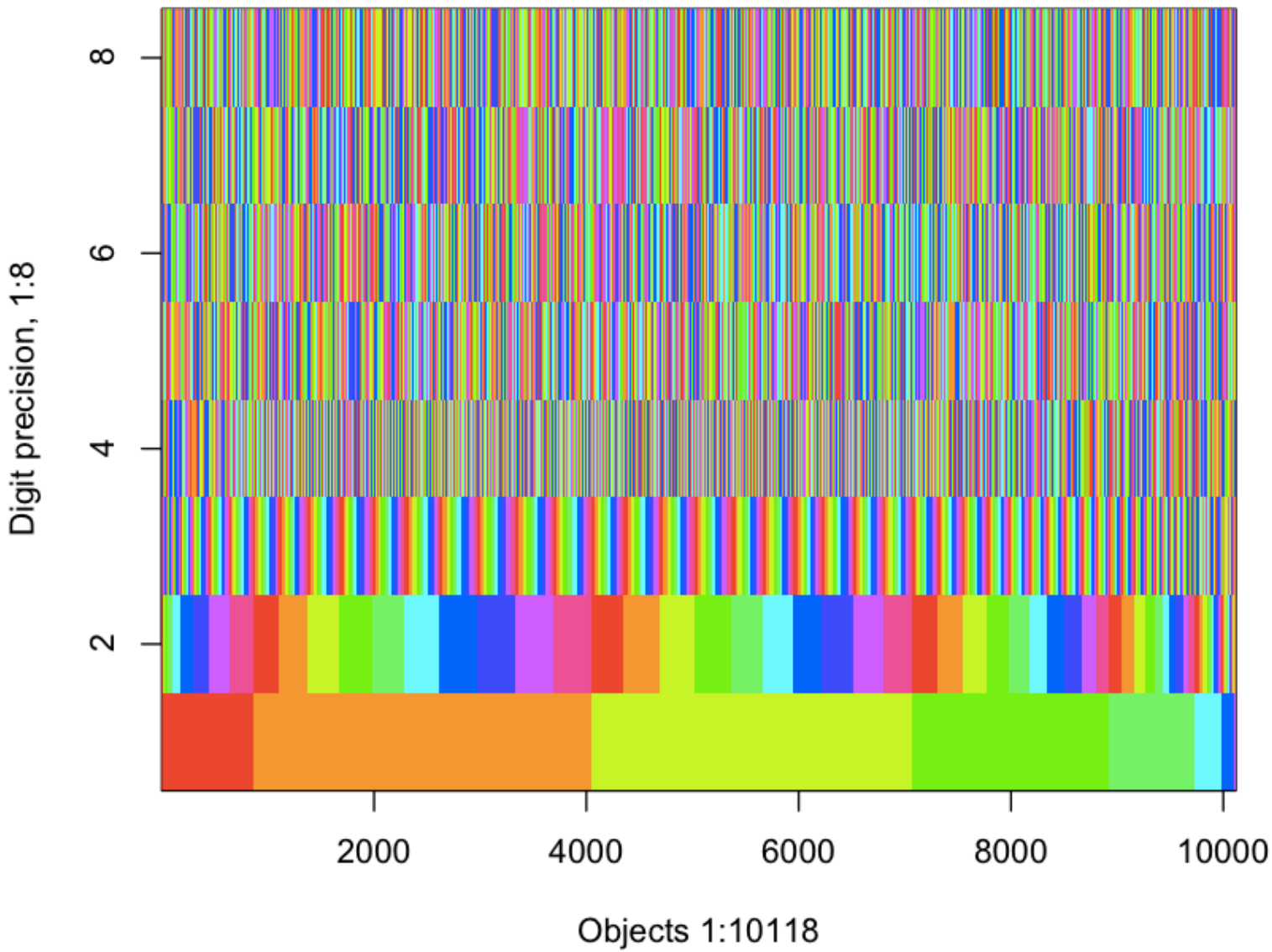
From random projection to the Baire hierarchical clustering

Selection of 10,317 funding proposals, out of set of 34,352, were indexed in Apache Solr. Their similarities were determined, using Solr's MLT ("more like this") score. (This uses weights for fields in the proposal documents, and is analogous to a chi squared, or tf-idf-based similarity.) *(tf-idf: term frequency – inverse document frequency)*

We used a very sparse similarity matrix of dimensions 10317 x 34252.
Through random projection, we obtained a unidimensional scaling of the 10317 proposals.

In the following the mean of 99 random projections was used.
The projection values were rescaled to the interval 0,1.
Layer 1 clusters: the same first digit.
Layer 2 clusters: given the same first digit, having the same second digit.
Layer 3 clusters: given the same first two digits, having the same third digit.
And so on.
This is a regular 10-way tree.

Abscissa: 10118 documents sorted by random projection value. Ordinate: 8 digits comprising random projection value.
Layer 1: 8 clusters are very evident. Layer 2: there are 87 clusters (maximum is 100). Layer 3: here 671 clusters (maximum is 1000).



Low dimensional goodness of fit to our data, versus linear rescaling

Conventional use of random projections:

Project data into lower dimension subspace, of dimension > 1.

Aim is to have proximity relations respected in the low dimensional fit to the high dimensional cloud of points.

In the work presented here, we seek a consensus one-dimensional mapping of the data, that represents relative proximity.

Two following slides: Our aim is relative clustering properties. Cf. the now conventional use of the Johnson-Lindenstrauss lemma.

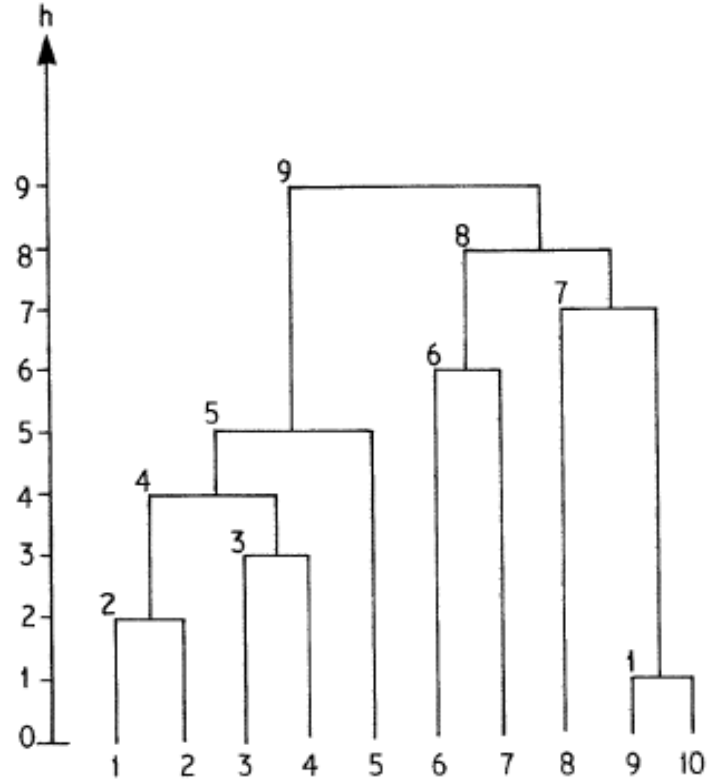


Figure 1a. Ultrametric tree, with heights.

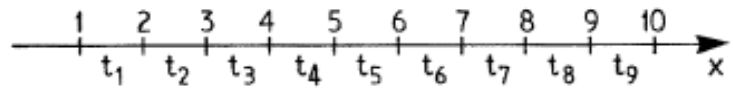


Figure 1b. Euclidean representation, with nine intervals to be determined.

F. Critchley and W. Heiser, “Hierarchical trees can be perfectly scaled in one dimension", Journal of Classification, 5, 5-20, 1988.

Reduced dimensionality: k << d
Below: Johnson-Lindenstrauss Lemma
Distance changes by a fraction 1+ε

$$F(x) : \mathbb{R}^d \to \mathbb{R}^k$$

Lemma 1. *For any $0 < \varepsilon < 1$ and any integer n, let k be a positive integer such that*

$$k \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \ln n. \tag{2}$$

Then for any set V of any points in $\mathbb{R}^d$, there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in V$,

$$(1-\varepsilon) \parallel u - v \parallel^2 \;\leq\; \parallel f(u) - f(v) \parallel^2 \;\leq\; (1+\varepsilon) \parallel u - v \parallel^2.$$

Furthermore, this map can be found in randomized polynomial time.

S. Kaski, "Dimensionality reduction by random mapping: fast similarity computation for clustering", Proceedings of The 1998 IEEE International Joint Conference on Neural Networks, pp. 413-418, 1998.

In random projection matrix, each column is of unit norm. Values are 0-mean Gaussian. So – random Gaussian vectors.

Reduced dimensionality space is not guaranteed to be in an orthonormal coordinate system.

Distortion of the variances/covariances relative to orthogonality of the random projections has approximate variance 2/m where m is low dimensionality.

For sufficient m, orthonormal system is mapped into a near-orthonormal system.

Kaski cites Hecht-Nielsen: the number of almost orthogonal directions in a coordinate system, that is determined at random in a high dimensional space, is very much greater than the number of orthogonal directions.

Conventional random projections: random vectors that are iid 0-mean Gaussian. This is only necessary condition for preserving pairwise distances (Li, Hastie, Church, Proc. 12th ACM SIGKDD, 2006).

Other work has used 0 mean, 1 variance, 4th moment =3.

Also elements of random projection matrix from {-1,0,1} with different (symmetric in sign) probabilities.

It is acknowledged that: "a uniform distribution is easier to generate than normals, but the analysis is more difficult".

The non-conventional approach to random projections that is at issue in the case studies described here

Uniform [0,1) valued vectors in the random projection matrix.

Projections are rescaled to be in [0,1), i.e. closed/open interval.

Take mean (over random projections) of projected values.

It is known from the central limit theorem, and the concentration, or data piling, effect of high dimensional data clouds, that: pairwise distances become equidistant, and orientation tends to be uniformly distributed.

We find also: norms of the target space axes are Gaussian. (That is, before taking the mean of the projections.) As typifies sparsified data, the norms of the points themselves are negative exponential, or power law, distributed.

Scaling followed by clustering

Correlation between most projection vectors > 0.99. We also found very high correlation between first principal component loadings and the mean random projection (> 0.999999).

Our objective is less to determine or model cluster properties as they are in very high dimensions, than it is to extract useful analytics by “re-representing" the data. That is to say, we are having our data coded (or encoded) in a different way.

Summary Remarks on Reading Baire Distance Properties from the (Mean) Random Projected Values

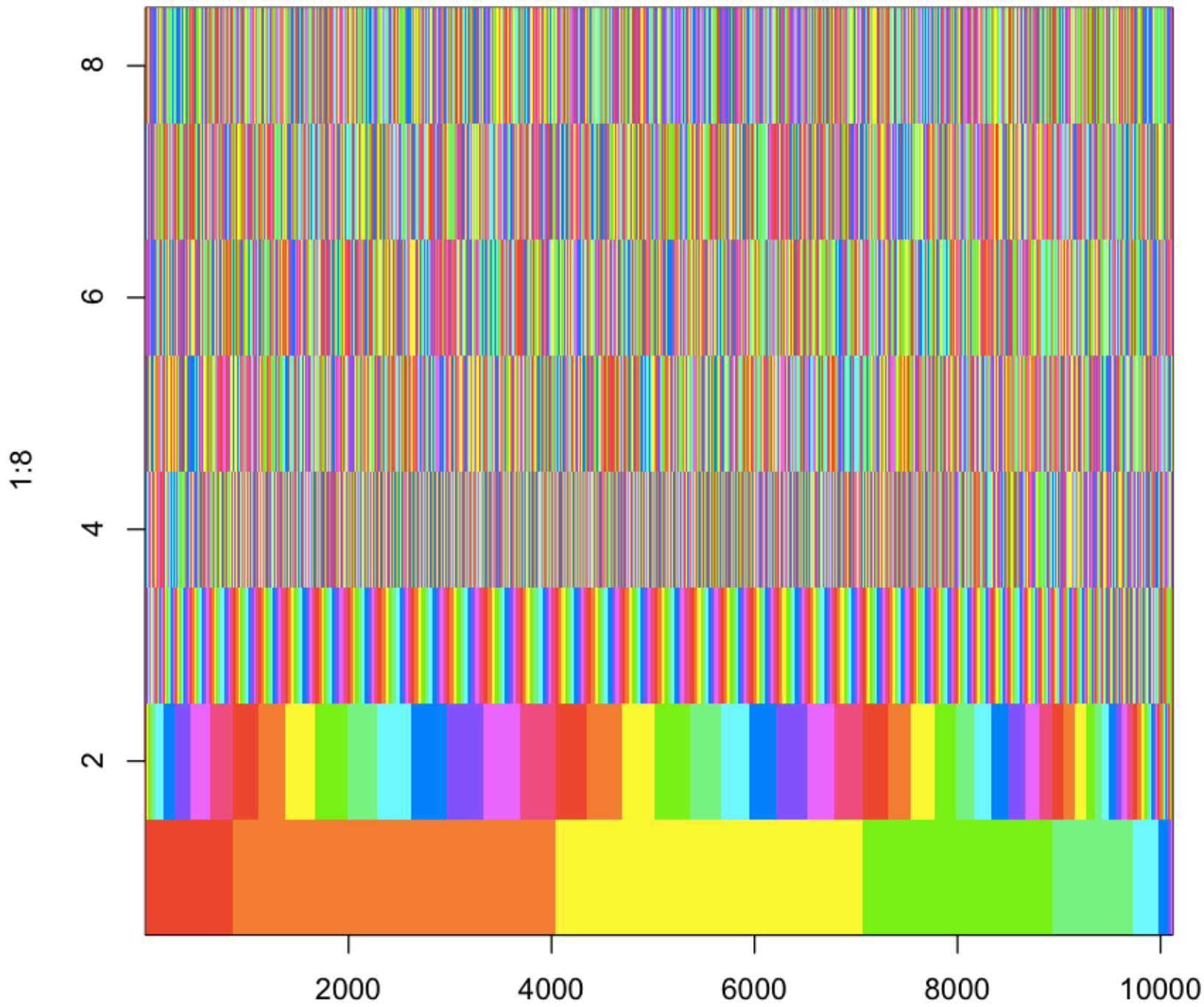
- We have a new way of inducing a hierarchy on data
- First viewpoint: encode the data hierarchically and essentially read off the clusters
- Alternative viewpoint: we can cluster information based on the longest common prefix
- We obtain a hierarchy that can be visualized as a tree
- We are hashing, in a hierarchical or multiscale way, our data
- We are targeting clustering in massive data sets
- The Baire method - we find - offers a fast alternative to k-means and a fortiori to traditional agglomerative hierarchical clustering
- At issue throughout this work: embedding of our data in an ultrametric topology

Visualization of Baire hierarchy.

Means of 99 random projections.

Abscissa: the 10118 (non-empty) documents are sorted (by random projection value).

Ordinate: each of 8 digits comprising random projection values.


1:8
2
4
6
8
2000
4000
6000
8000
10000

Traditional clustering: use pairwise distances, determine clustering structure (hierarchy or optimization of criterion). Often: then a partition is determined.

Here we build a series of partitions. Then the hierarchy is determined from them.

To be followed by – a look at accompanying web site content

- Data sets used in this presentation.
- Software used – mainly in R.
- The processing carried out.
- Samples of the output produced.

http://www.multiresolutions.com/Courses/CIMI

Presentation
File with Listing of all Processing
Files Used
References: Applications in Search and Discovery

User name: **Courses**
Password: **WelICome** [Wel figure I Come]

CIMI Machine Learning Workshop 09.11.2015

Presentation

- CIMI-MLWorkshop-TLS-9Nov2015.pdf

File with Listing of all Processing

- All background and R code, text file. File to read for all software, data, processing.

Files Used

- Datasets.
  - rBCI1052encoded.scn.zip (All chemicals data, 53 MB, unzipped > 500 MB)
  - testdata1k2k.txt.zip (1000 chemicals)
  - BCI1052enc7500.dat.zip (7500 chemicals)
- Software.
  - Program to extract data from the BCI-encoded chemical dataset: ExtractBCI.c
    Compile and link using: gcc -lm ExtractBCI.c -o ExtractBCI
  - Used in processing in R, repeatedly carry out random projections and determine their mean: RepeatRanProj.r
  - Slimmed down Correspondence Analysis program: ca-projonly.r
  - Just drawing in a factor plane projection: plaxes.r
- Output plots. (See full description in the background description above.)
  - sample-output-hist1.pdf
  - sample-output-hist2.pdf
  - sample-output-CA-1.pdf
  - sample-output-convergence-mean-ranproj.pdf
  - sample-output-convengence-mean-ranproj-2.pdf
  - sample-output-histograms.pdf

File with Listing of all Processing

- All background and R code, text file. File to read for all software, data, processing.

Files Used

- Datasets.
  - rBCI1052encoded.scn.zip (All chemicals data, 53 MB, unzipped > 500 MB)
  - testdata1k2k.txt.zip (1000 chemicals)
  - BCI1052enc7500.dat.zip (7500 chemicals)
- Software.
  - Program to extract data from the BCI-encoded chemical dataset: ExtractBCI.c
    Compile and link using: gcc -lm ExtractBCI.c -o ExtractBCI
  - Used in processing in R, repeatedly carry out random projections and determine their mean: RepeatRanProj.r
  - Slimmed down Correspondence Analysis program: ca-projonly.r
  - Just drawing in a factor plane projection: plaxes.r
- Output plots. (See full description in the background description above.)
  - sample-output-hist1.pdf
  - sample-output-hist2.pdf
  - sample-output-CA-1.pdf
  - sample-output-convergence-mean-ranproj.pdf
  - sample-output-convengence-mean-ranproj-2.pdf
  - sample-output-histograms.pdf
  - sample-output-display.pdf
  - sample-output-CA1000.pdf
  - sample-output-PCA-9cluster-centres.pdf
  - sample-output-kmeans-PCA-9cluster-centres.pdf
  - sample-output-loglogplot.pdf
- Some Further Data Sets.
  - doc_similarities.mm.zip (document similarities, > 22 MB).
  - redshifts5000.dat.zip (5000 astronomical redshifts).

References: Applications in Search and Discovery

1. F. Murtagh and P. Contreras, "Random projection towards the Baire metric for high dimensional clustering", A. Gammerman, V. Vovk and H. Papadopoulos, Eds, Statistical Learning and Data Sciences, Springer Lecture Notes in Artificial Intelligence (LNAI) Volume 9047, 424-431, 2015. Paper.
2. F. Murtagh and P. Contreras, "Linear storage and potentially constant time hierarchical clustering using the Baire metric and random spanning paths", Proc. ECDA 2014, European Conference on Data Analysis, Springer, 2015, in press.
3. P. Contreras and F. Murtagh, "Linear time Baire hierarchical clustering for enterprise information retrieval", International Journal of Software and Informatics, 6 (3), 363-380, 2012. Paper.
4. F. Murtagh and P. Contreras, "Fast, linear time, m-adic hierarchical clustering for search and retrieval using the Baire metric, with linkages to generalized ultrametrics, hashing, Formal Concept Analysis, and precision of data measurement", p-Adic Numbers, Ultrametric Analysis and Applications, 4, 45-56, 2012. Paper.
5. P. Contreras and F. Murtagh, "Fast, linear time hierarchical clustering using the Baire metric", Journal of Classification, 29, 118--143, 2012. Paper.
6. F. Murtagh, G. Downs and P. Contreras, Hierarchical clustering of massive, high dimensional data sets by exploiting ultrametric embedding, SIAM Journal on Scientific Computing, 30, 707-730, 2008. Paper.


ARABIA
The End
A
WARNER BROS.-FIRST NATIONAL
PICTURE