Random projections

Dimensionality reduction of large data sets

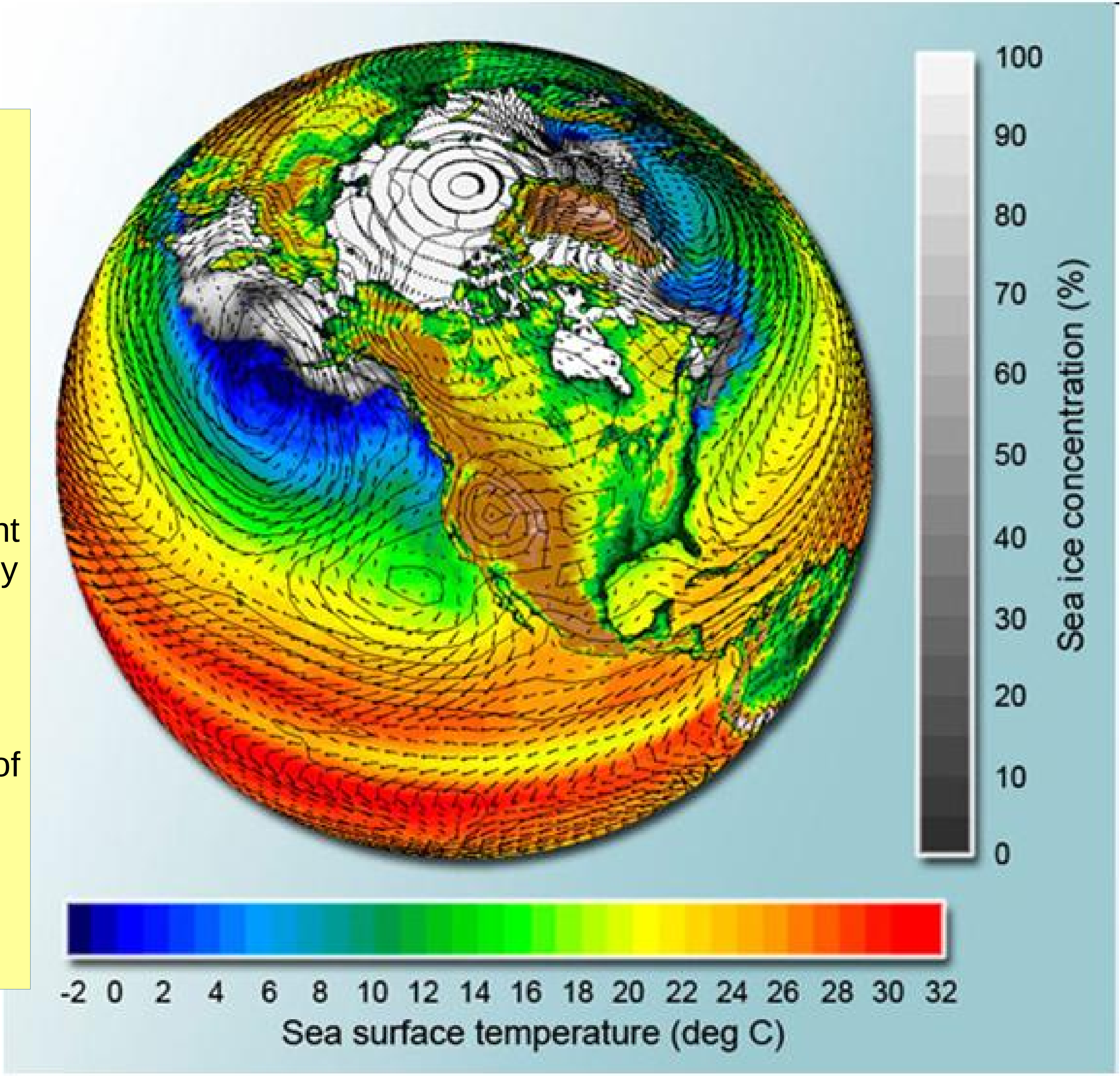
Teija Seitola
FMI
16.11.2015

Contents

The amount of data is too large to treat in any simple manner.

There are several different kinds of information in any given data set.

What is the meaning(!?) of the data?



Background

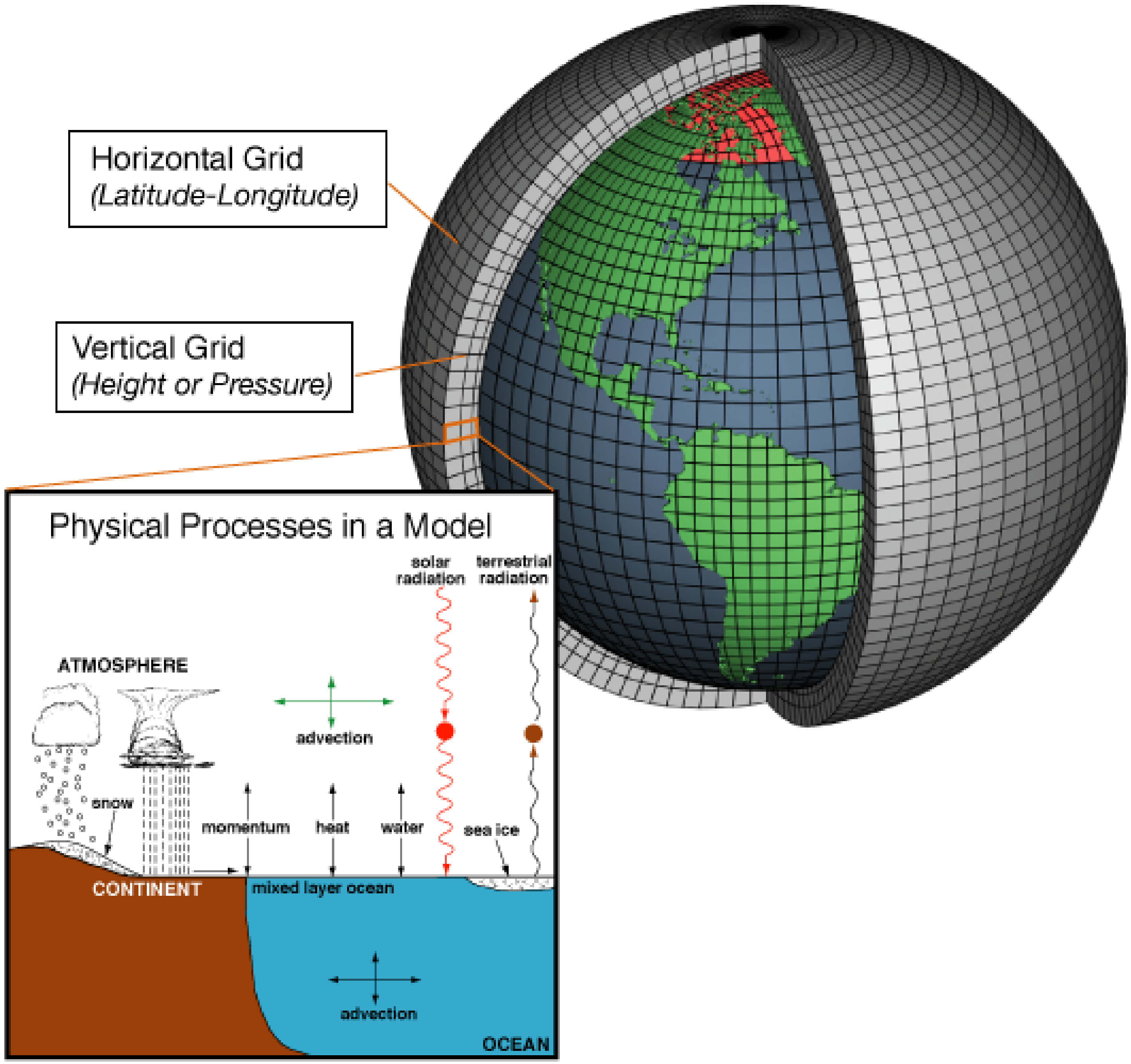
- In many applications the high dimensionality of the data restricts the choice of data processing methods
- A statistically optimal way is to project the data onto a lower-dimensional orthogonal subspace that captures majority of the variance
- Most widely used method is PCA → quite expensive to compute, in very high-dimensional cases not even applicable
- A computationally more simple method would be needed!

Random projections - theory

- Data can often be thought of as a point in a high dimensional space
  - e.g. temperature data at a certain time in each gridpoint of a global grid
- Johnson and Lindenstrauss (1984) proved that the data points (*n)* in *d*-dimensional space can be embedded in a k ≥ $O(\log(n/\varepsilon^2))$ -dimensional subspace with only little distortion (the pairwise Euclidean distances are approx. preserved):

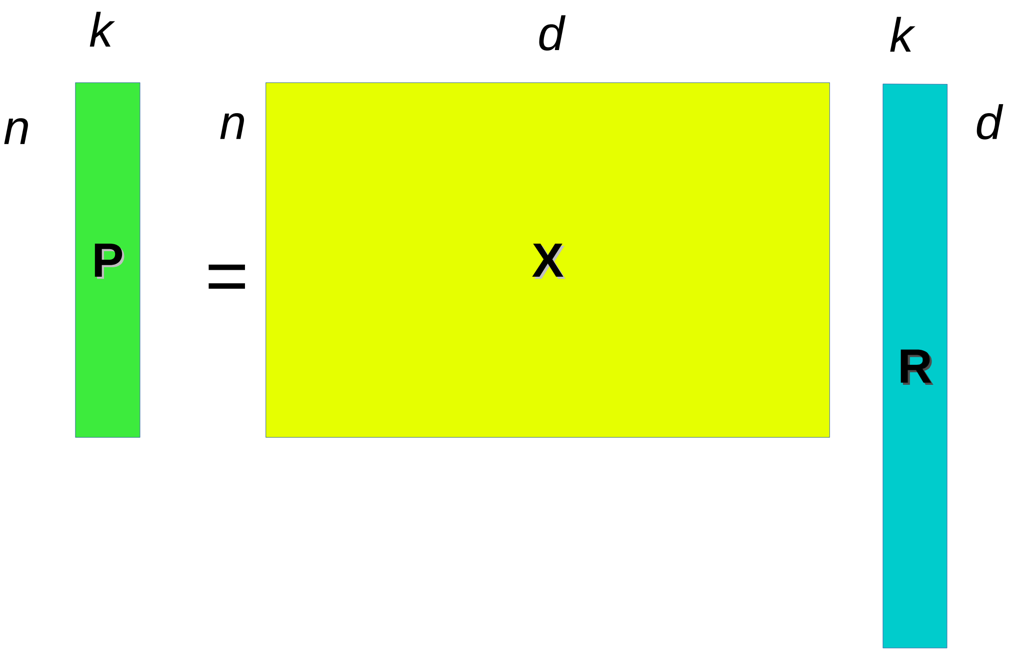
Suppose we have an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$. *Given any* $\varepsilon > 0$, *there is a mapping* f: $\mathbb{R}^d \rightarrow \mathbb{R}^k$, *for any* $k \geq O\frac{\log n}{\varepsilon^2}$, *such that, for any two rows* $\mathbf{x_i}, \mathbf{x_j} \in \mathbf{X}$, *we have*

$$(1-\varepsilon)||\mathbf{x_i} - \mathbf{x_j}||^2 \leq ||f(\mathbf{x_i}) - f(\mathbf{x_j})||^2 \leq (1+\varepsilon)||\mathbf{x_i} - \mathbf{x_j}||^2 \quad (2)$$



Horizontal Grid
(Latitude-Longitude)
Vertical Grid
(Height or Pressure)
Physical Processes in a Model
solar radiation
terrestrial radiation
ATMOSPHERE
advection
snow
momentum
heat
water
sea ice
CONTINENT
mixed layer ocean
advection
OCEAN

Random projection



$$\boldsymbol{P}_{n\times k}=\frac{1}{\sqrt{k}}\boldsymbol{X}_{n\times d}\boldsymbol{R}_{d\times k}$$

- Original data **X** (*n x d*) is projected onto a random matrix **R** (*d x k*) to have a lower dimensional subspace **P** (*n x k)*
- Elements of **R** are ~*N(0,1)*
- Computational complexity is *O(knd)*
  - Compare to PCA: $O(d^2n)$ + $O(d^3)$
- Simple and fast to compute → can be applied to a wide range of data sets

Suitable mappings **R**?

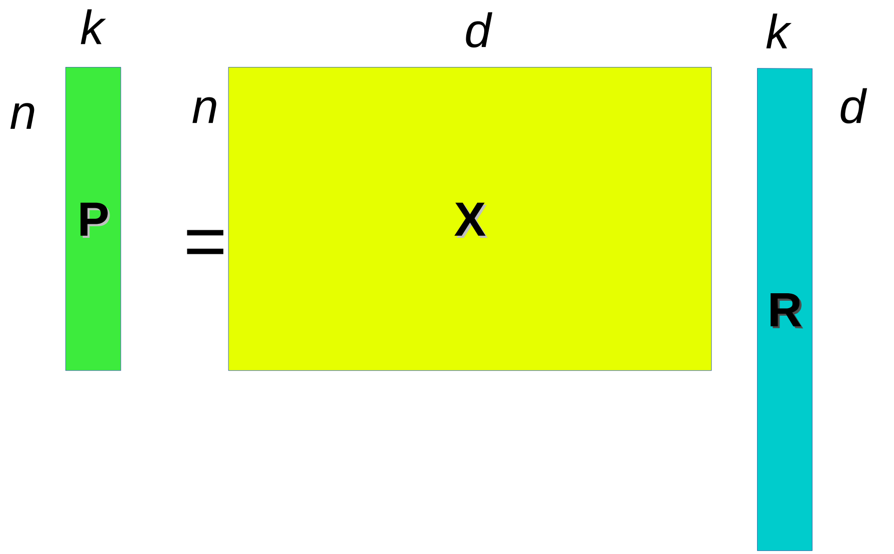
- There are also other random distributions that fullfill the J&L lemma, for example (Achlioptas, 2001)

$$r_{ij} = \sqrt{3} \times \begin{cases} +1 & \textit{with probability } 1/6, \\ 0 & \textit{with probability } 2/3, \\ -1 & \textit{with probability } 1/6. \end{cases}$$

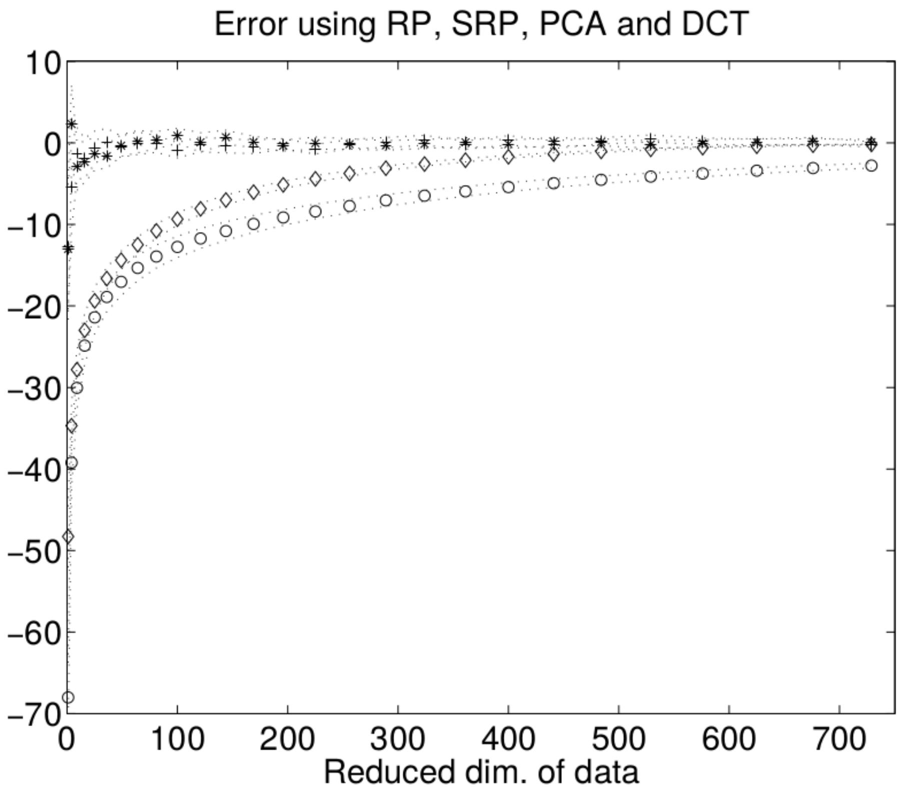
According to Bingham & Mannila (2001) practically all zero mean, unit variance distributions of r_{ij} would give a mapping that satisfies the Johnson-Lindenstrauss lemma.

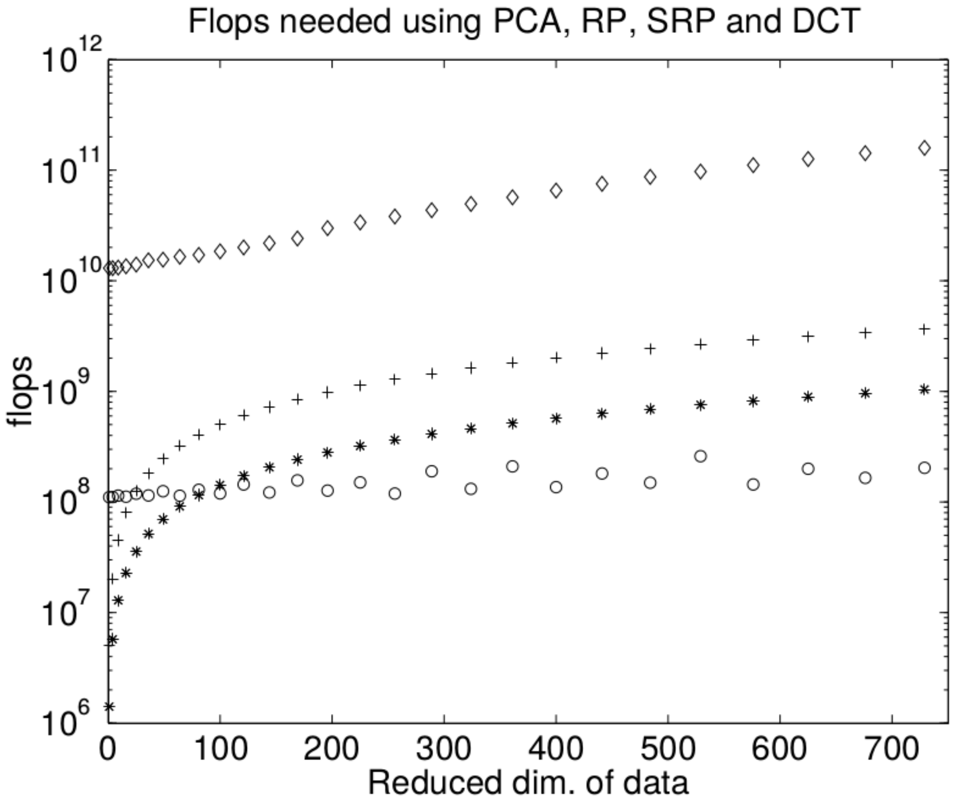
Other considerations

- Orthogonal projection is assumed $\rightarrow$ as the dimension increases, the number of almost orthogonal vectors increases
- Vectors of **R** can be orthogonalized, but often this is not required



- Appropriate value for *k*?
  - There are some explicit formulas, e.g. $k \geq 4(\varepsilon^2/2 - \varepsilon^3/3)^{-1} \log(n)$
  - But this is a worst case estimate and often much lower values for *k* give good results





The error produced by RP (+), SRP (⋆), PCA (◊) and DCT (◦) on image data, averaged over 100 pairs of data vectors

The error is measured by comparing the Euclidean distance between two compressed data vectors to their distance in the original high-dimensional space

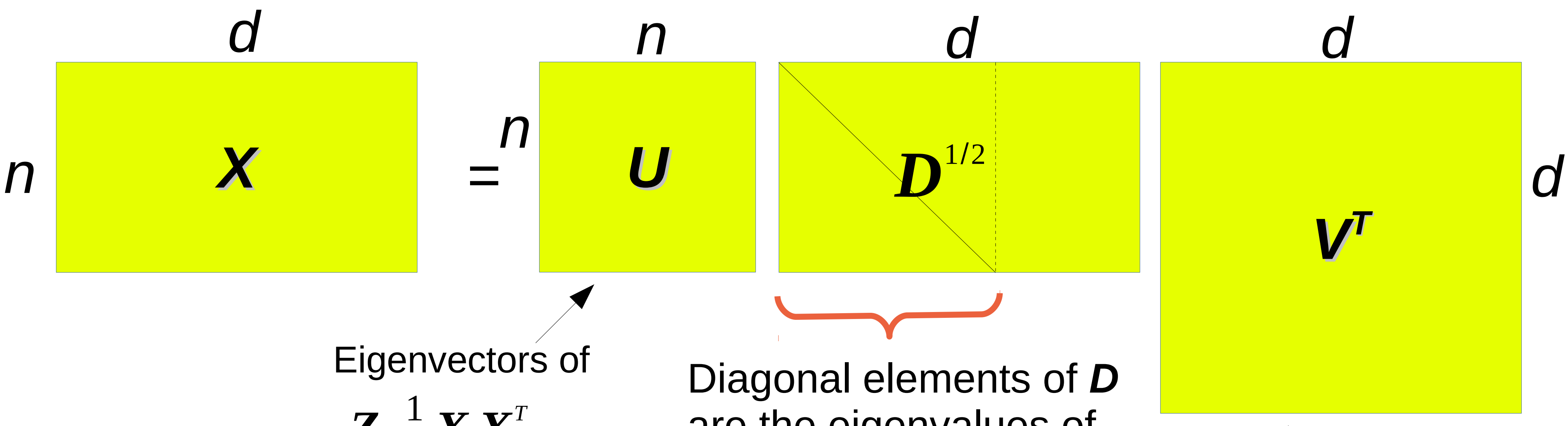
Number of Matlab's floating point operations needed when reducing the dimensionality of image data using RP (+), SRP (⋆), PCA (◊) and DCT (◦)

Examples are from Bingham & Mannila (2001)

Application to climate data

- Climate simulation data is high dimensional
  - several variables, thousands of time steps and grid points
- High dimensionality causes problems:
  - input/output and post-processing expensive and time consuming
  - excludes use of some analysis methods
- **Objective**:
  - To introduce random projection (RP) as a dimensionality reduction method applied on climate data
  - To show how the projected data preserves the essential structure of the original data (by means of SVD/PCA)

Principal component analysis by SVD



Eigenvectors of

$$Z = \frac{1}{d} X X^T$$

Diagonal elements of ***D*** are the eigenvalues of ***C*** or ***Z***

Eigenvectors of

$$C = \frac{1}{n} X^T X$$

Often called PC loadings

- SVD can be used to decompose the original and randomly projected data

We get principal componets (PC scores, **S**) by

$$\mathbf{S}_{n \times n} = \mathbf{U}_{n \times n} \mathbf{D}_{n \times n}$$

or $$\mathbf{S}_{n \times n} = \mathbf{X}_{n \times d} \mathbf{V}_{d \times d}$$

Experiment 1

- **Data**: monthly surface temperature data set from a millennial full-forcing Earth system model simulation
- Dimensions of original data set n=4608 (time steps) d=4608 (grid points)
- The dimensions of random projections were 10% (k≈460) and 1% (k≈46) of the original dimensions d
- PCA was applied to original and low-dimensional data to study how the structure is preserved
- SVD of original data: $\mathbf{X} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^{\mathbf{T}}$
- SVD of low-dimensional data: $\mathbf{P} = \mathbf{U}_{rp}\,\mathbf{D}_{rp}\,\mathbf{V}^{\mathbf{T}}_{rp}$
  - PC loadings in the original high-dimensional space:

    $\mathbf{V} \approx \mathbf{X}^{\mathbf{T}}\,\mathbf{U}_{rp}\,\mathbf{D}^{-1}_{rp}$

Comparison of original and projected data – PC loadings (eigenvectors)

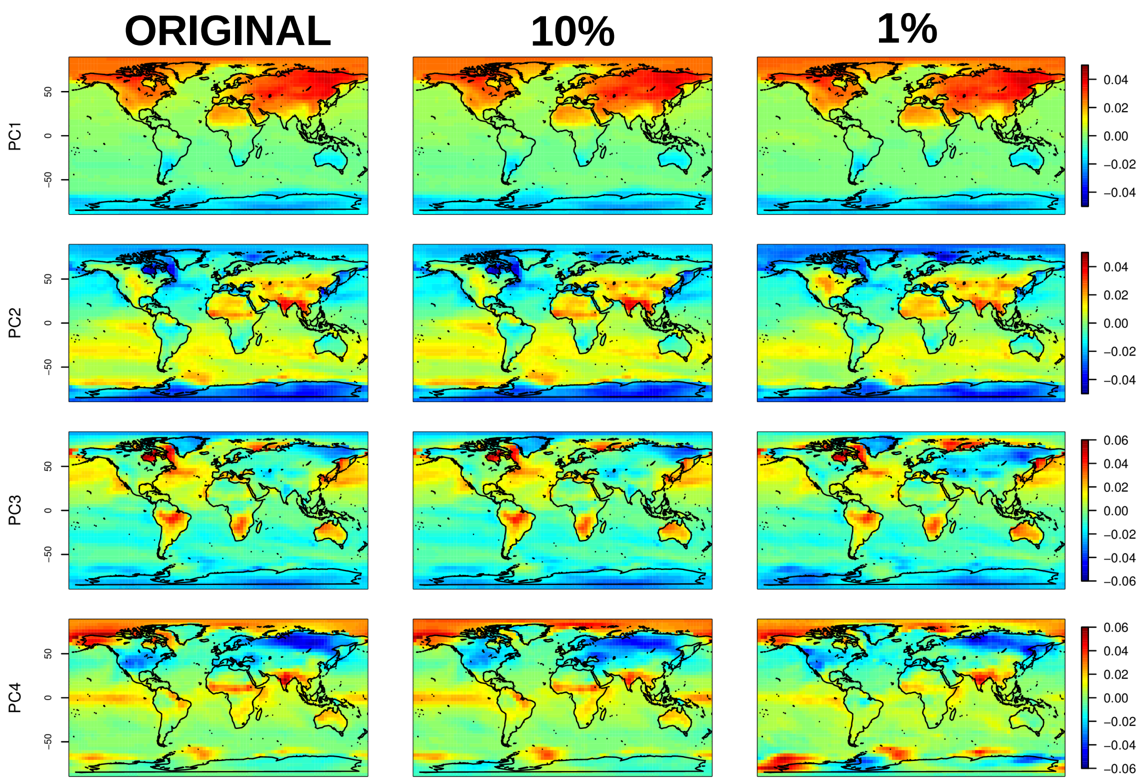


Fig.1: Spatial patterns of PC1-PC4 (loading vectors), original and projected (10%, 1%) data.

PCs 1-12 explain 96% of variance
PCs 1-5 explain 94% of variance

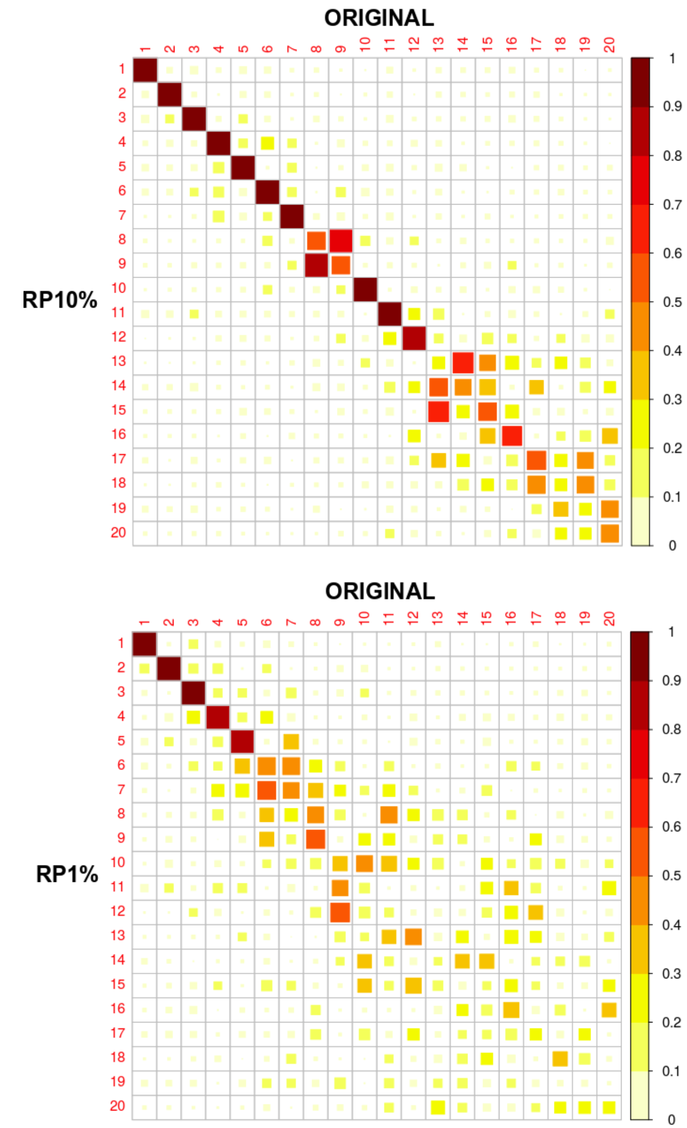
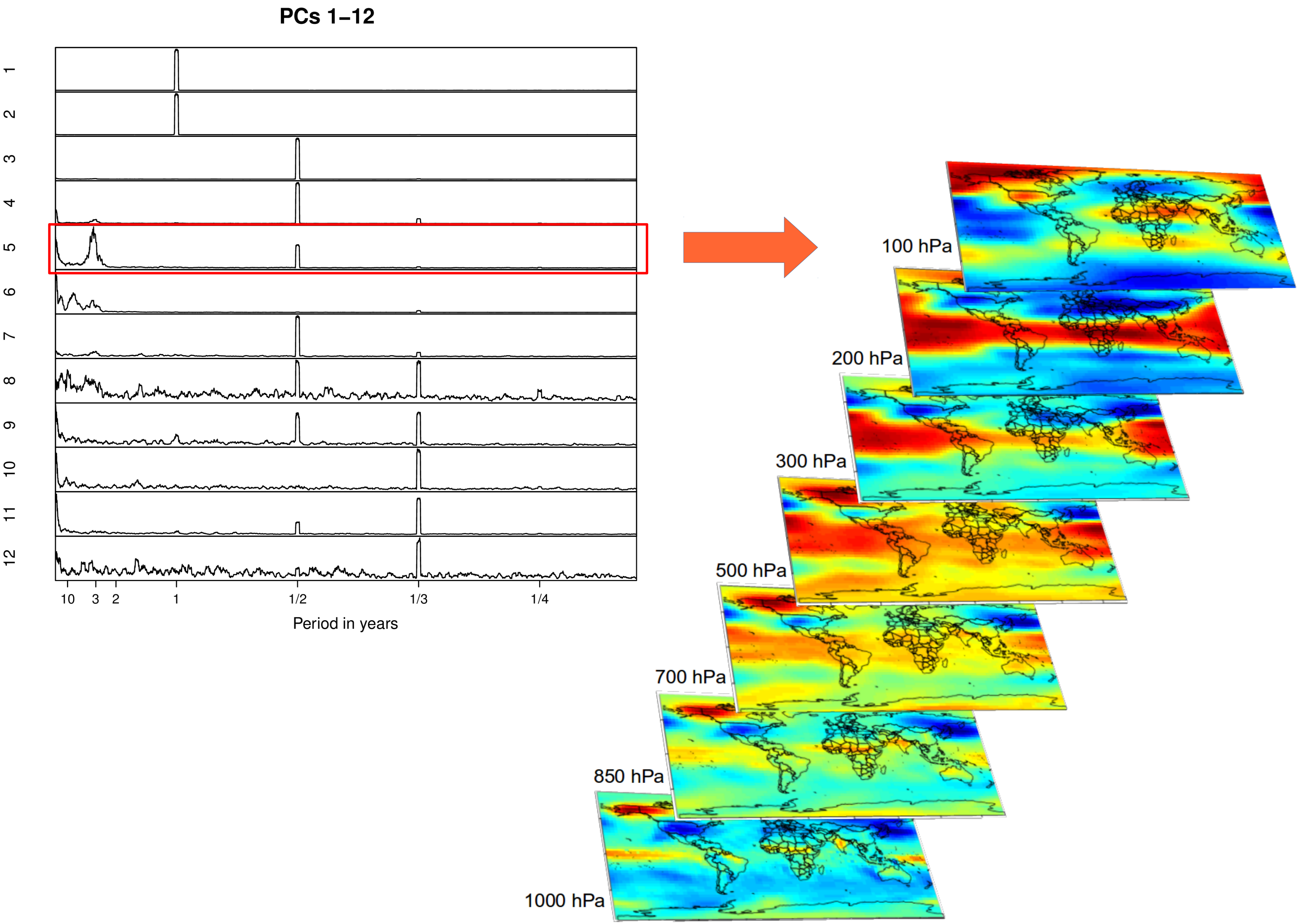


Fig.2: Correlations of original and projected (10%, 1%) PC loadings 1-20.

Experiment 2

- **Data**: monthly air temperature data set from an ESM simulation
- Resolution: 96 x 48 x 17 (lon x lat x lev)
- Original data matrix: 3600 x 78336 (*n x d*)
- The dimensionality of the original data $\mathbf{X}_{n \times d}$ was reduced by RP to 1% (k≈783) of original dimensions


PCs 1–12
1
2
3
4
5
6
7
8
9
10
11
12
10
3
2
1
1/2
1/3
1/4
Period in years
100 hPa
200 hPa
300 hPa
500 hPa
700 hPa
850 hPa
1000 hPa

Summary

- Some information is naturally lost in RP → essential structure can still be recovered from the randomly projected low-dimensional subspaces
- Random projection is computationally fast involving only matrix multiplication → can be applied to very high dimensional data sets
- RP can be used with many standard matrix factorizations, e.g. SVD, PCA, QR

References

- Ella Bingham and Heikki Mannila: "Random projection in dimensionality reduction: applications to image and text data", Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), August 26-29, 2001, San Francisco, CA, USA, pp. 245-250.
- W.B. Johnson and J.Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space. Contemporary Mathematics, vol. 26. 1989-206, (1984).
- Halko, N., Martinsson, P. G., & Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review, 53(2), 217-288.
- Seitola, T., Mikkola, V., Silen, J. and Järvinen, H. 2014. Random projections in reducing the dimensionality of climate simulation data. Tellus A, 66.