Random Projection

and

Projection Pursuit Based PCA

Applied to Chemical Data

Kurt Varmuza[1*], Peter Filzmoser[2], Bettina Liebmann[1]

Vienna University of Technology

[1] **Laboratory for Chemometrics, Institute of Chemical Engineering, Vienna University of Technology**
Getreidemarkt 9/166,
A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at
liebmann@mail.zserv.tuwien.ac.at
www.lcm.tuwien.ac.at

[2] **Institute of Statistics and Probability Theory, Vienna University of Technology**
Wiedner Hauptstrasse 8-10,
A-1040 Vienna, Austria
P.Filzmoser@tuwien.ac.at
www.statistik.tuwien.ac.at/public/filz

Poster Presentation: Conferentia Chemometrica 2009
September 27–30, 2009, Siofok, Hungary

Acknowledgments. We gratefully acknowledge support by the Austrian Research Promotion Agency (FFG), BRIDGE program, project no. 812097/11126. We thank Ulrich Omasits (ETH Zürich) for collaboration in statistics, and Katja Hansen (TU Berlin) for the AMES data.

Abstract

Random projection (RP) is a method for mapping *n* points from a high-dimensional space (with *m* variables) into a low dimensional space (with *m**<<*m* new variables) with the Euclidean distances approximately preserved. In contrary to principal component analysis (PCA) and similar methods, RP uses randomly selected - orthogonal or almost orthogonal - projection axes. RP can be computed independent from the original data, is computationally very simple, fast, and appropriate for large data sets. Successful applications have been reported for clustering and classification of textual documents and image data (for instance face recognition), as well as for protein similarity searches and for data mining.

Basic idea of RP is the fact that high-dimensional vectors with randomly chosen vector components are very often "almost orthogonal". For instance about 95% of 10,000 randomly generated vector pairs (with *m*=1000, vector components uniformly distributed between -1 and 1) have the cosine of the angle between the two vectors in the narrow range of -0.062 to 0.062. In other words, in a high dimensional space much more "almost orthogonal" vectors than orthogonal vectors exist. Published results indicate that RP preserves the similarity of data vectors well, eccentric clusters become more spherical, and classification results (obtained by *k*-nearest neighbor classification) compare favorably with PCA-transformed data. Thus RP is an optional method if the distances in the high-dimensional space are meaningful but may be less useful for highly correlating variables.

We investigate the applicability and limits of RP with some chemical data sets, and compare RP with PCA based on projection pursuit. In the latter case, the principal components are computed sequentially by maximizing the variance or a robust measure of variance of the projected data. Fast algorithms allow a precise estimation of the principal components, even for high-dimensional data.

Introduction

Random projection (RP) is a linear method for a projection from a high-dimensional space into a low-dimensional space, using **projection vectors (loading vectors) with random numbers as vector components**.

RP is based on the fact that high-dimensional vectors with randomly chosen vector components are very frequently "almost orthogonal"*. Some RP methods apply orthogonalization of the projection vectors [1-3].

RP projection generates loading vectors independently from the original data, is simple and fast in computation, and is appropriate for large data sets. Successful applications have been reported for clustering and classification of textual documents and image data [4-5], as well as for protein similarity searches [6].

We report on preliminary experiments with RP

- using artificial data and data from chemistry,
- applying RP for cluster analysis,
- investigating RP as a data reduction method before KNN classification or PLS calibration.

* Example: About 95% of randomly generated vector pairs (with m = 1000 dimensions, and the vector components uniformly distributed between -1 and 1, 10,000 random vector pairs considered) have the cosine of the angle between the two vectors in the range -0.062 to 0.062 (86.4° to 93.6°). For m = 200 the 95%-range is -0.135 to 0.135 (82° to 98°).

Introduction

Generation of Random Projection Vectors

- ■ normally distributed numbers from e.g. N(0, 1)
  - ■ uniformly distributed numbers from e.g. U[-1, +1]
  - ■ fixed values randomly selected from e.g. {-1, 0, +1}
- Optional orthogonalization (Gram-Schmidt or other)
- Normalization to unit length

Projection Pursuit based PCA (PP)

Principal components are extracted sequentially, by maximizing the variance of the projected points on a direction [7]. Two algorithms are considered: CR (the potential directions are determined directly by the data points), and GRID (iterative grid search in planes).

Software

All computations were performed with programs written in R [8].
For PCA and PP the free R-library "chemometrics" [7, 8] has been used.

[1] S.S. Vempala. The random projection method. *Series in Discrete Mathematics and Theoretical Computer Science*, vol. 65, American Mathematical Society, Providence, RI (2004).
[2] S. Dasgupta. Experiments with random projection. In *Proceedings of the Sixteenth Conference* on *Uncertainty in Artificial Intelligence*, p. 143-151. Morgan Kaufmann Publishers Inc., San Francisco, CA (2000).
[3] X.Z. Fern, C.E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Machine Learning - International Workshop* (2003).
[4] E. Bingham, H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2001).
[5] N. Goel, G. Bebis, A. Nefian. Face recognition experiments with random projection. In *Proceedings SPIE* (2005).
[6] J. Buhler, M. Tompa. Finding motifs using random projections. *Journal of Computational Biology*, **9** [2], 225 (2002).
[7] K. Varmuza, P. Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*, CRC Press, Boca Raton, FL (2009). Info: www.lcm.tuwien.ac.at/cm-book.pdf.
[8] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, 2008. www.r-project.org.

Projection Methods and Cluster Analysis

Data simulation

2 classes of objects; n_1 = 800; n_2 = 200

m = 100 (or 1000) variables, normally distributed, centered ("Gaussians")

The Gaussians are defined by a covariance matrix Σ, orthog. randomly rotated

Separation *c* of Gaussians: $c = ||\mu_1 - \mu_2|| / (\max(\text{trace } \Sigma_1, \text{trace } \Sigma_2))^{0.5}$ [2]

$||\mu_1 - \mu_2||$ is Euclidean distance of centers

Projection methods

PCA

PP applied in two versions ("CR" and "GRID")

RP: ten orthogonalized random projection vectors from U[-1, 1] or N(0, 1)

Efficiency of variance preservation = $\dfrac{\text{sum of first } q \text{ variances from projection method}}{\text{sum of first } q \text{ variances from PCA}}$

Clustering

k-means clustering, 2 clusters

Misclassification rate = ratio of objects assigned to the wrong cluster

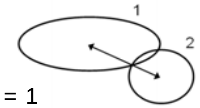
Simulations

50 repetitions; results presented in box plots

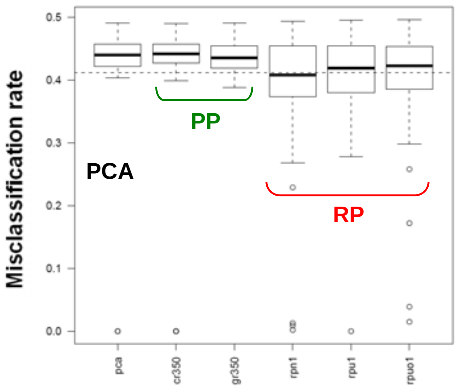
Results for data set 1

m = 100, Σ_1, with $s(j, j) = 1/j^2$ (j = 1, ..., m)

Σ_2, with $s(j, j) = 1/m^2$ (constant)



c = 1



Conclusions

PCA, PP and RP yield appr. the same misclassification rates, however, RP results show a much larger variability.

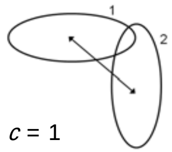
The separation, *c* (not shown), is conserved in all methods, again with a high variability for RP.

Projection Methods and Cluster Analysis

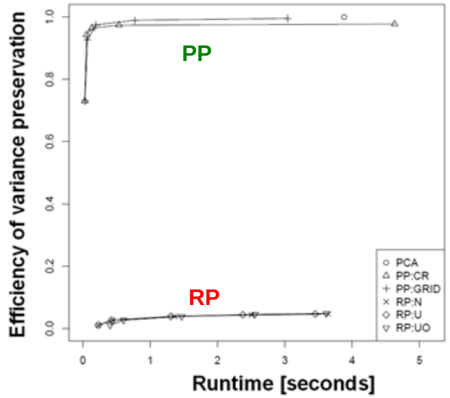
Results for data set 2

m = 1000, Σ_1, with $s(j, j) = 1/j^2$ (j = 1, ..., m)

Σ_2, with $s(j, j) = 1/j^2$ (j = m, ..., 1)



c = 1



Conclusions

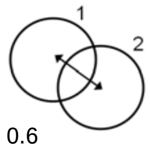
Efficiency of variance preservation is very high for PP, but very low for RP (at equal runtime).

Misclassification rates (not shown) are near zero for PCA and PP; also low for RP but with a high variability.

Results for data set 3

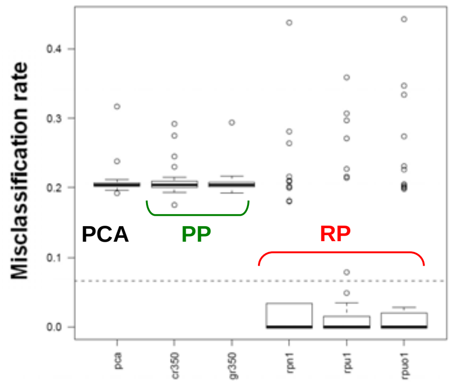
m = 100, Σ_1, and Σ_2, with $s(1,1), \ldots, s(10,10) = 100$

$s(11,11), \ldots, s(100,100)$ = U[0,1]



c = 0.6

High (and equal) variances in 10 components,
very low (and varying) variances in other 90 components.
c = 0.6 makes the two Gaussians not separable with 10 PCA components.



Conclusions

RP is able to ignore the high variability directions - *just by chance* - and yields better results than PCA or PP, and better results than obtained with the original data (dashed line).

RP shows advantages only for *specially designed* data sets.

Random Projection and KNN

Data

AMES n = 6506 compounds used in AMES tests for mutagenicity; n_0 = 3004 inactive (class 0), n_1 = 3502 active (class 1); m = 1455 molecular descriptors (Dragon software), autoscaled

KNN

Leave-1-out, Euclidean distance, number of neighbors (k) varied; performance criterion $P_{MEAN} = (P_0 + P_1)/2$, with P_0 and P_1 the fraction correctly classified compounds in class 0 and 1, resp.

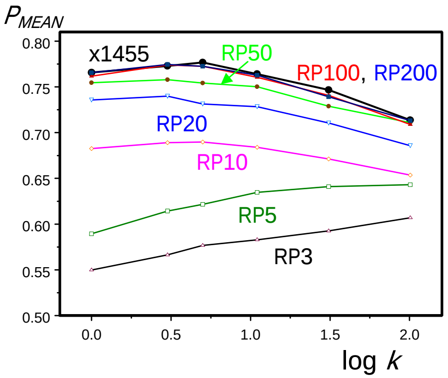
Results

KNN classification has been performed for k = 1, 3, 5, 11, 31, 101, with

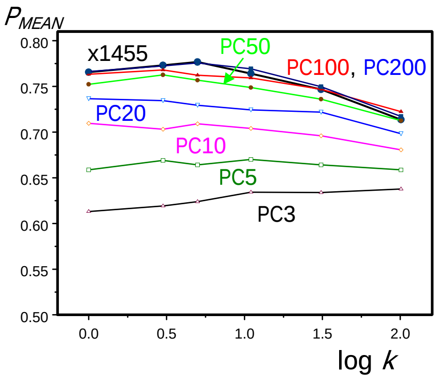
- all original variables (x1455)
- 3, 5, 10, 20, 50, 100, 200 random vector scores (**RP**3, ...)
- 3, 5, 10, 20, 50, 100, 200 principal component scores (**PC**3, ...)

Computing time: 100 RP scores, 3s; 100 PC scores, 200s; KNN with 100 variables (RP/PC scores), 30s; KNN with 1455 variables, 2000s.

Random Projection (RP) - KNN



PCA Projection (PC) - KNN



Conclusion

About 100 random projection scores* yield about the same prediction performance as 100 principal component scores or all 1455 original variables.

Optimum number of neighbors is 3 to 5.

*** RP100 (k = 3): P_1 = 0.740, P_2 = 0.808, P_{MEAN} = 0.774**

Random Projection and PLS

Data

ET n = 166 fermentation mashes, m = 235 NIR absorptions, y is the glucose content in g/L [9]

BIO n = 35 biomass samples (wood, cereals), m = 435 IR absorptions, y is the heating value in kJ/kg [10]

TOX n = 846 compounds from toxicology, m = 529 molecular descriptors (Dragon software), y is the Kovats GC retention index [11]

Results

PLS models have been created with the original variables and with scores from Random Projection (RP). For evaluation the rdCV strategy (repeated double cross validation) was applied [12].

ET

Variables for PLS	#	*SEP*	*a*
Original ***x***	235	7.0	9
RP scores	3	12.2	1
RP scores	20	9.4	4
RP scores	100	7.9	8

BIO

Variables for PLS	#	*SEP*	*a*
Original ***x***	435	143	1
RP scores	3	145	1
RP scores	20	127	2
RP scores	100	138	3

TOX

Variables for PLS	#	*SEP*	*a*
Original ***x***	529	86	15
RP scores	3	270	1
RP scores	20	195	5
RP scores	100	126	13

SEP standard deviation of 20*n prediction errors from test-set objects

a optimum number of PLS components

Conclusion

A rather small number of RP scores (ca 10 % of the number of variables) give a similar - usually a somewhat worse - prediction performance as all original variables.

[9] Liebmann B., Friedl A., Varmuza K.: Anal. Chim. Acta, **642**, 171-178 (2009).
[10] Varmuza K., Liebmann B., Friedl A.: University of Plovdiv, Bulgaria, Scientific Papers - Chemistry **35** [5], 5-10 (2007) ISSN 0204-5346.
[11] Garkani-Nejad Z., Karlovits M., Demuth W., Stimpfl T., Vycudilik W., Jalali-Heravi M., Varmuza K.: J. Chromatogr. A, **1028**, 287-295 (2004).
[12] Filzmoser P., Liebmann B., Varmuza K.: J. Chemometrics, **23**, 160-171 (2009).