A few meteorological applications of Sparse Principal Component Analysis

Caren Marzban

Dept. of Statistics and Applied Physics Lab
University of Washington
Seattle, WA 98195

Abstract

Principal Component Analysis is generally used either to reduce the dimensionality of data (e.g., for compression, or feature extraction), or to provide an explanation/interpretation of the underlying structure of data. Both of these goals, however, become increasingly problematic as the number of variables increases or exceeds the number of cases. In such situations, it has been proposed that principal components should additionally be constrained to be sparse (i.e., to have zero loadings on many/most of the variables). In this talk, the construction of the Sparse Principal Component Analysis is reviewed, and the method is applied to a number of example data sets from meteorology.

Introduction

Principal Component Analysis (PCA): workhorse of multivariate statistics.

Two main functions:
1) Dimensionality reduction and compression
2) Data interpretation.

The idea:
Given p variables and n observations,
find a linear combination of variables with maximum variance,
subject to some constraints.
The weights are called loadings.

Equivalent to eigen decomposition of $p \times p$ Cov/Cor matrix.
Eigenvalues = variances of PC1, PC2, ...
Eigenvectors = loadings

Equivalent to SVD on $n \times p$ data matrix.

Lots of acronyms:
NLPCA: Nonlinear PCA (use nonlinear combinations of p variables)
RPCA: Rotated PCA (constrain loadings to be small or large)
SCoTLASS: Simplified Component Technique-LASSO (same)
ICA: Independent CA (maximize non-normality)
. . .

Jolliffe and Cadima (2016): Principal Component Analysis: A review and recent developments. *Phil. Trans. R. Soc.* **A 374**.

In Meteorology, Spatial fields pose specific problems ("Buell effect")
Even harder to do inference.

Regression

Multivariate Multiple Regression:

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||^2$$

$X = n \times (p + 1)$ Data matrix
$Y = n \times q$ matrix of Responses
$\beta = (p + 1) \times q$ matrix of coefficients

Ridge Regression:

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

The L_2-norm penalty term shrinks some of the coefficients.
Tames overfitting.

Lasso Regression:

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

The L_1-norm penalty term shrinks some of the coefficients to zero.
I.e., variable selection.
But for $p > n$, at most n variables can be selected.

Elastic Net Regression:

$$\hat{\beta} = \arg\min_{\beta} ||Y - X\beta||^2 + \lambda \sum_{j=1}^{p} \beta_j^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

The L_1-term gives sparsity
The L_2-term allows for more than n selected variables.

Optimization algorithms: not covered here.

PCA and SPCA

PCA (no flavors):

eigen(cov(X))

PCA (Vanilla-flavored):

SVD (X)

PCA (as regression)

$$min_A \sum_{i=1}^{n} ||X - AA^T X||^2 + \lambda \sum_{j=1}^{k} |\alpha_j|^2$$

subject to $A^T A = 1_{k \times k}$ (i.e., orthonormality).
$A = p \times k = \{\alpha_1, \cdots \alpha_k\}$

Essentially, PCA = Regression on X to X.

Intuitively, think of a neural-net with p inputs, 1 hidden node, and p outputs.

Sparse PCA (SPCA):

$$min_{A,B} \sum_{i=1}^{n} ||X - AB^T X||^2 + \lambda \sum_{j=1}^{k} ||\beta_j||^2 + \sum_{j=1}^{k} \lambda_{1j} |\beta_j|_1$$

$B = p \times k = \{\beta_1, \cdots, \beta_k\}$

Essentially, SPCA = Elastic Net Regression on X to X.

In R

package: elasticnet
function: spca()
lambda = λ (above)
para = λ_{1j} (above)

Instead of para, can specify number of nonzero loadings (nz, below).

spca() works on cov(X) or cor(X) - like eigen version of PCA
and on X itself - like SVD version of PCA.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse Principal Component Analysis. *Journal of Computational and Graphical Statistics*, **15** 265-286.

Example 1

Data from *Statistical Methods in the Atmospheric Sciences,* (2nd edition), Daniel S. Wilks.

Daily precip (inches), and min. and max. temperature (F) at Ithaca and Canandaigua, NY, Jan 1987
$p = 6$ variables, $n = 31$ cases

Note: in this example, sparsity is sensitive to choice of params.

That's not a bad thing!
A good technique will have a few knobs.
Allows for data exploration.

Example 1: Continued

PCA:

Variable	PC1	PC2	PC3	PC4	PC5	PC6
ithaca precip	-0.142	0.677	-0.063	-0.149	-0.219	0.668
ithaca maxT	-0.475	-0.203	-0.557	0.093	0.587	0.265
ithaca minT	-0.495	0.041	0.526	0.688	-0.020	0.050
canan precip	-0.144	0.670	-0.245	0.096	0.164	-0.658
canan maxT	-0.486	-0.220	-0.374	-0.060	-0.737	-0.171
canan minT	-0.502	-0.021	0.458	-0.695	0.192	-0.135

SPCA:

nz = c(5,5,5,5,5,5) ; lambda = 10

Variable	PC1	PC2	PC3	PC4	PC5	PC6
ithaca prcp	0.000	0.674	-0.010	0.162	-0.383	0.000
ithaca maxT	-0.494	0.011	-0.775	0.123	-0.521	-0.171
ithaca minT	-0.489	0.249	0.000	-0.231	0.000	-0.009
canan prcp	0.000	0.668	-0.148	0.000	-0.411	-0.944
canan maxT	-0.512	0.000	-0.613	0.032	-0.619	-0.279
canan minT	-0.505	0.193	-0.039	0.951	-0.171	-0.037

Note: SPCA is not same as thresholding the loadings

SPCA:

$\lambda_{1j} = (1, 1, 1, 1, 1, 1), \lambda = 0$

Variable	PC1	PC2	PC3	PC4	PC5	PC6
ithaca prcp	0.000	0.896	0	0	0	0
ithaca maxT	-0.263	0.000	0	0	0	0
ithaca minT	-0.317	0.000	0	0	0	0
canan prcp	0.000	0.445	0	0	0	0
canan maxT	-0.731	0.000	0	0	0	0
canan minT	-0.545	0.000	0	0	0	0

Example 1: Continued

What's the price of this sparsity?
Not much:

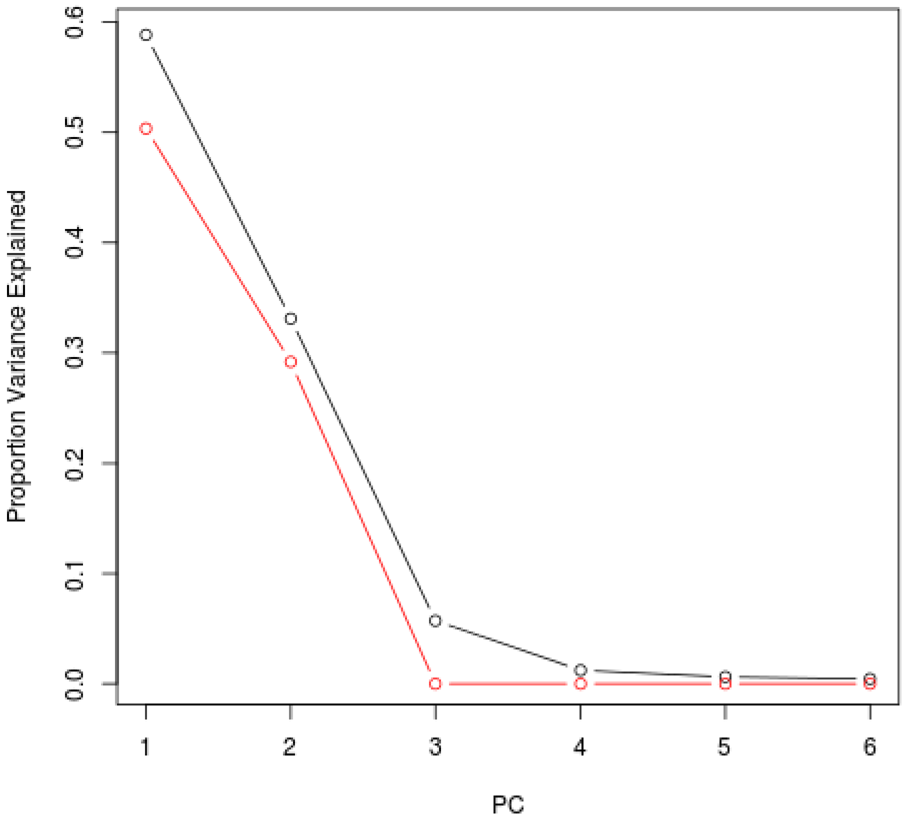


Figure: The proportion of variance explained by each of the 6 PCs for PCA (black) and SPCA (red) for the last SPCA settings.

Example 2

SST data; for details, see *Machine Learning Methods in Environmental Sciences; Neural Networks and Kernels,* William W. Hsieh.

Disclaimer: My PCA results here may be wrong!
Just compare SPCA with PCA.

See Guangoh Jheong and Gyu-Ho Lim in
Parsimonious patterns in sea surface temperature of the tropical Pacific Ocean
(Unpublished, but available upon request from gyuholim@snu.ac.kr)
and
Parsimonious patterns of sea surface temperature in the tropical Pacific Ocean, 2017 AMS Poster.

Example 2: Continued

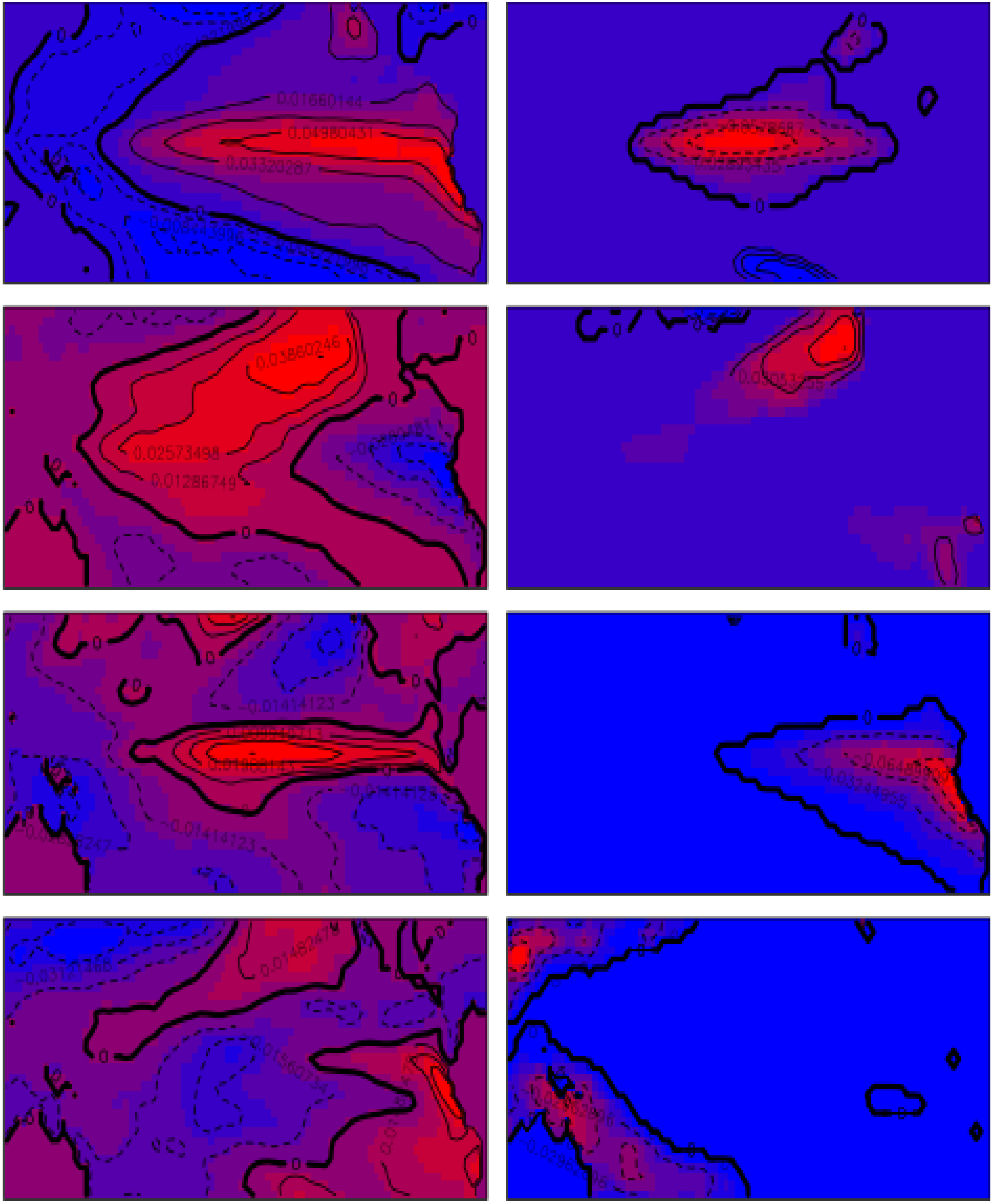
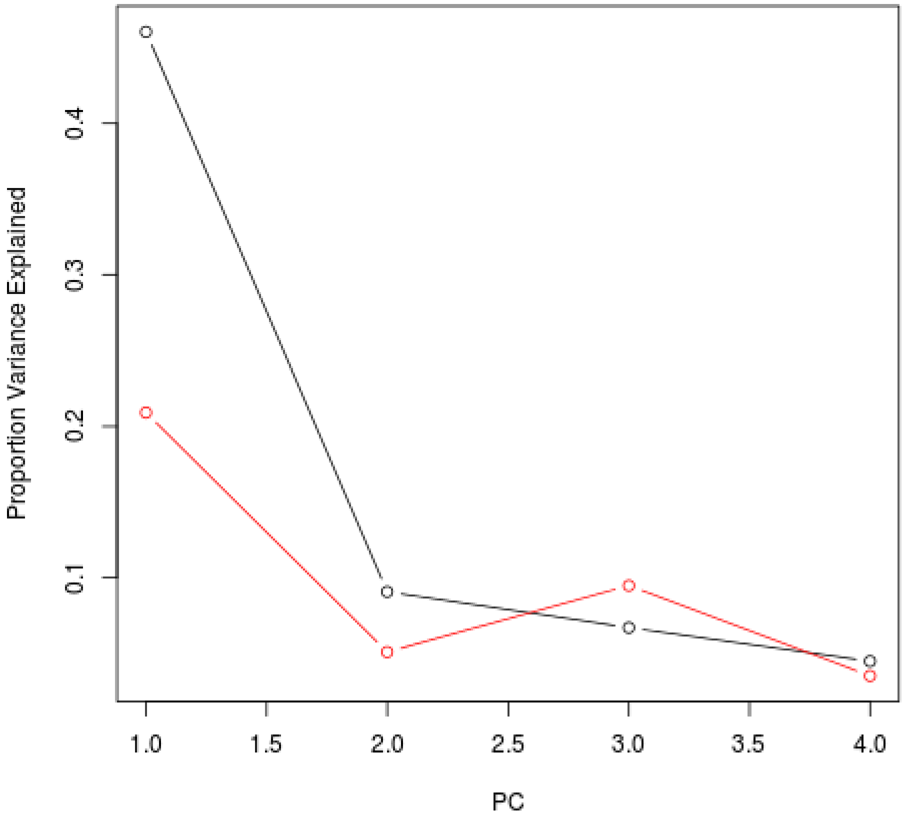


Figure: PCA (left) and SPCA (right), of PC1 - PC4 (from top to bottom).

Example 2: Continued

Not much is lost in terms of variance explained:



Note: In this example, results are insensitive to choice of params.

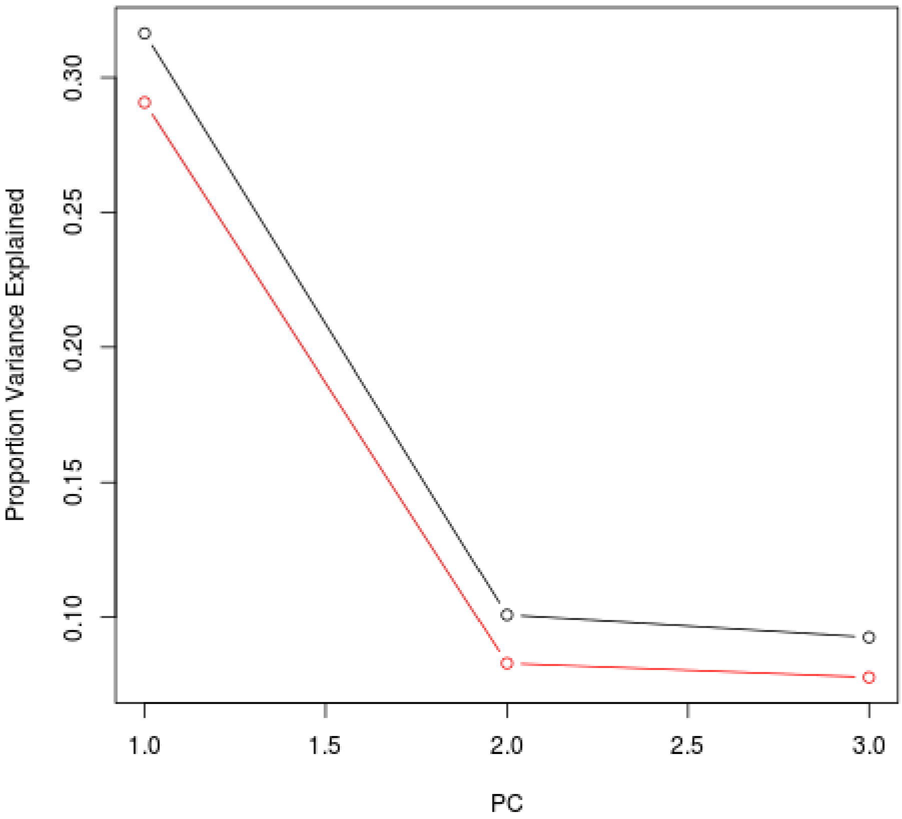
I recommend to standardize data, i.e., do spca on cor(X), to get sense of what params do across different data sets.

Example 3

500hPa geopotential height

Same disclaimer here.

Not much lost:



Example 3: Continued

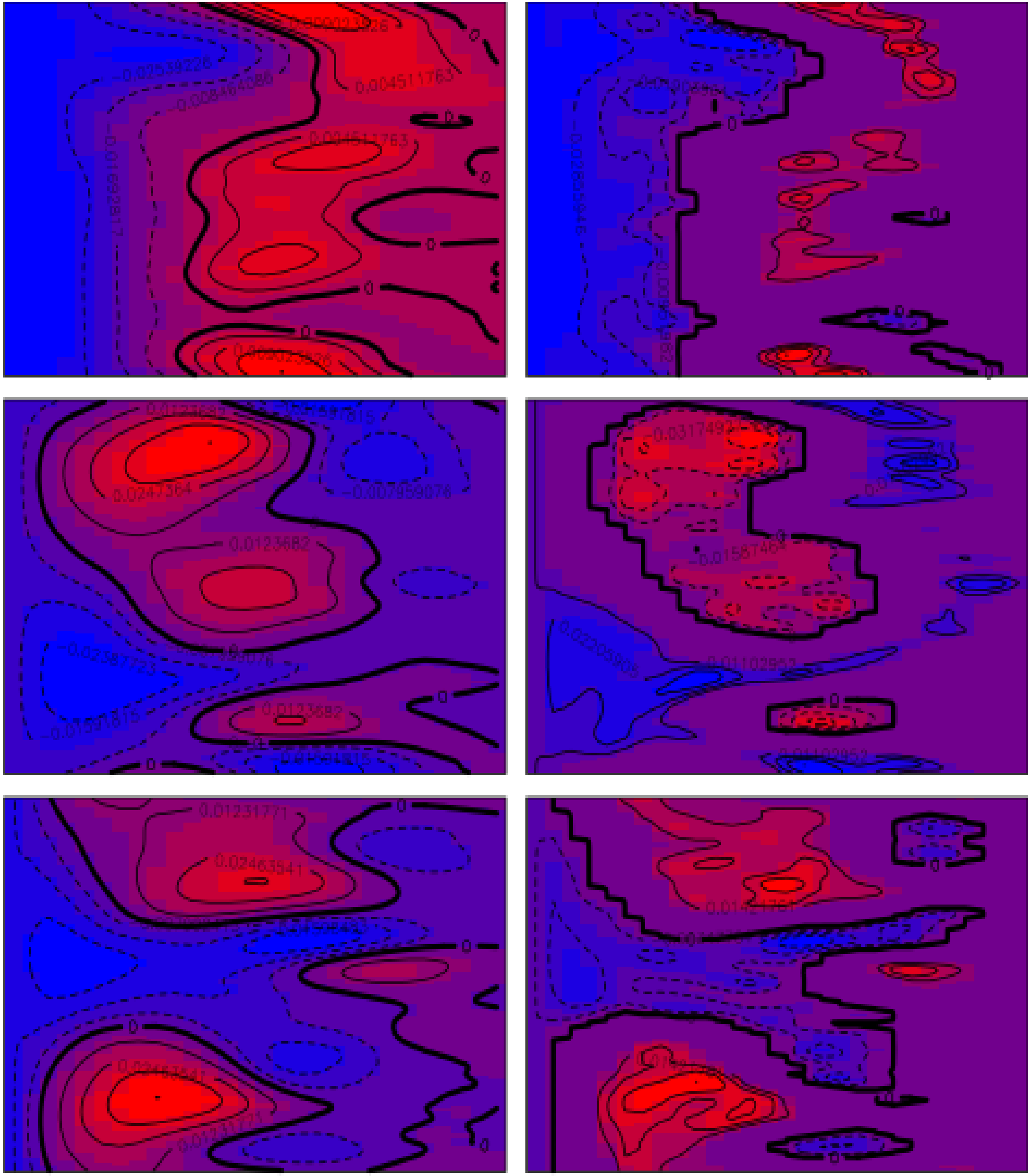


Figure: PCA (left) and SPCA (right), of PC1 - PC3 (from top to bottom).

Example 3: Continued

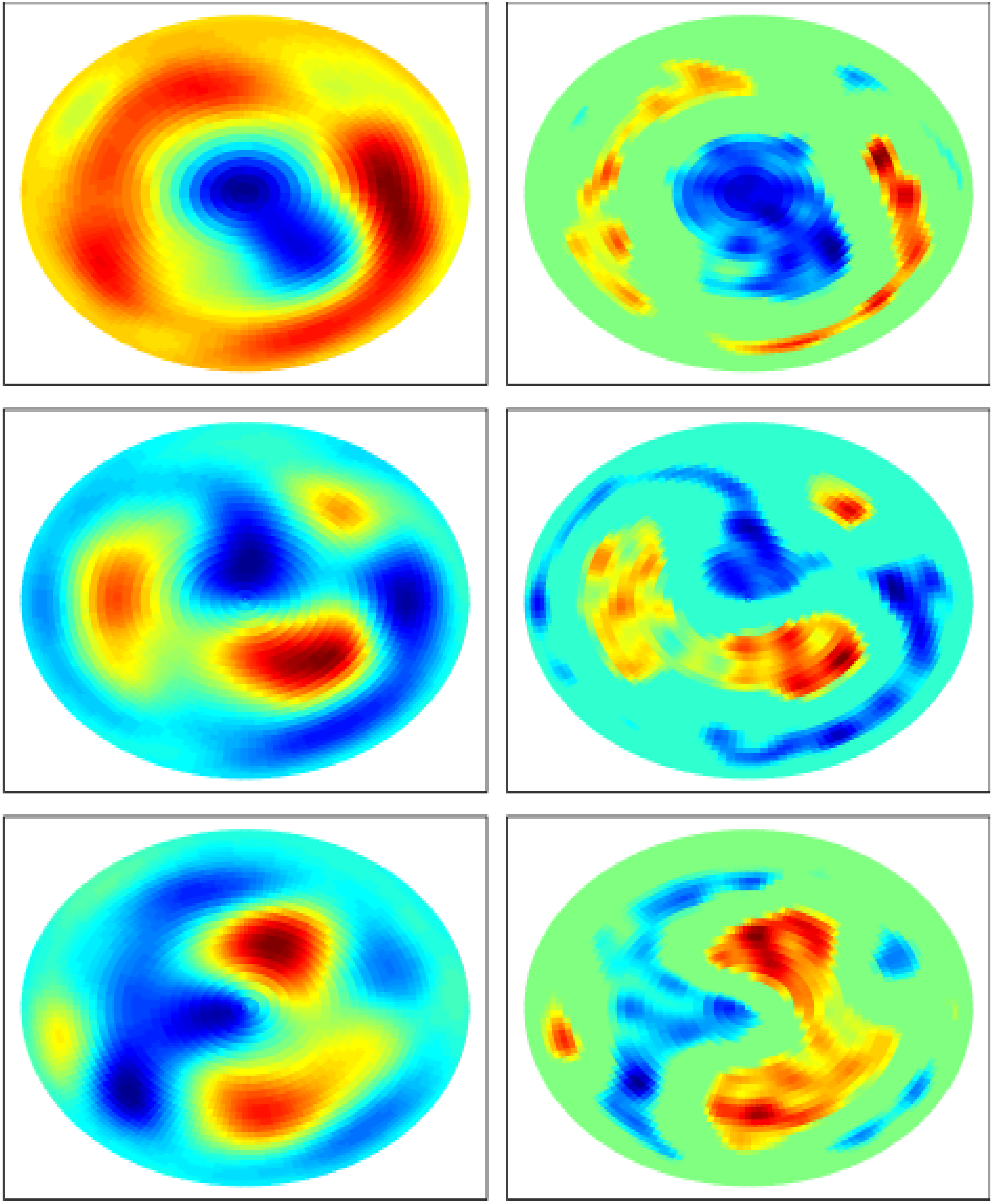


Figure: In polar coordinates.

Interpretation?

Conclusion

SPCA appears promising
But when applied to spatial fields, the spatial structure may render the results "obvious."

It has been compared with other sparse PCA method (e.g., SCoTLASS) but not on spatial fields.
It ought to be compared with RPCA as well.

Acknowledgements

Thanks to William Hsieh and Mike Richman.