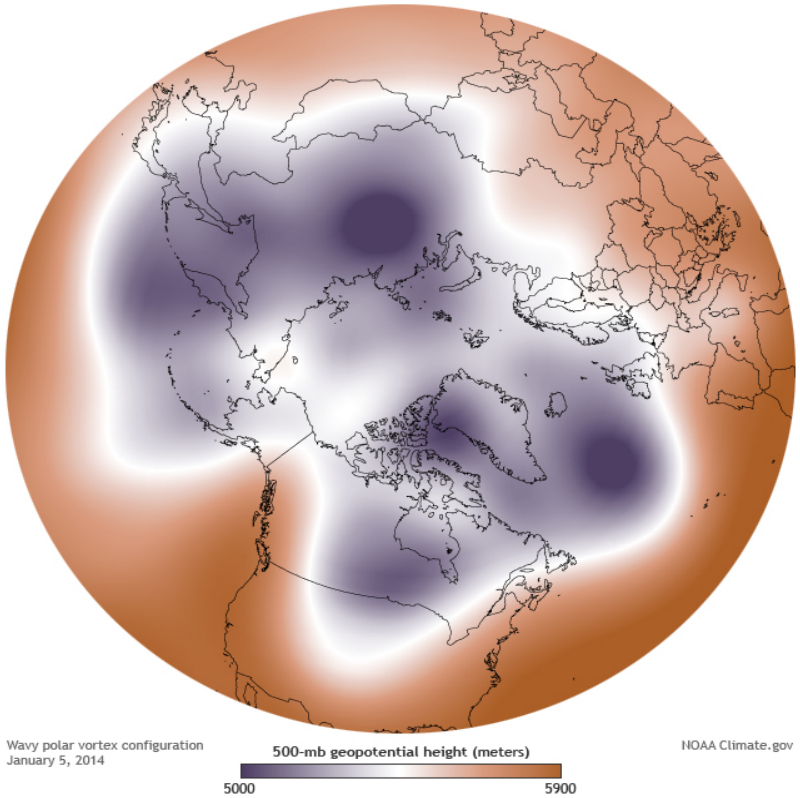
# Interpreting and Extending Principal Components

Naomi Altman
**Penn State University**




**Dept. of Meteorology**

February, 2015

# The Many Roles of Principal Component Analysis

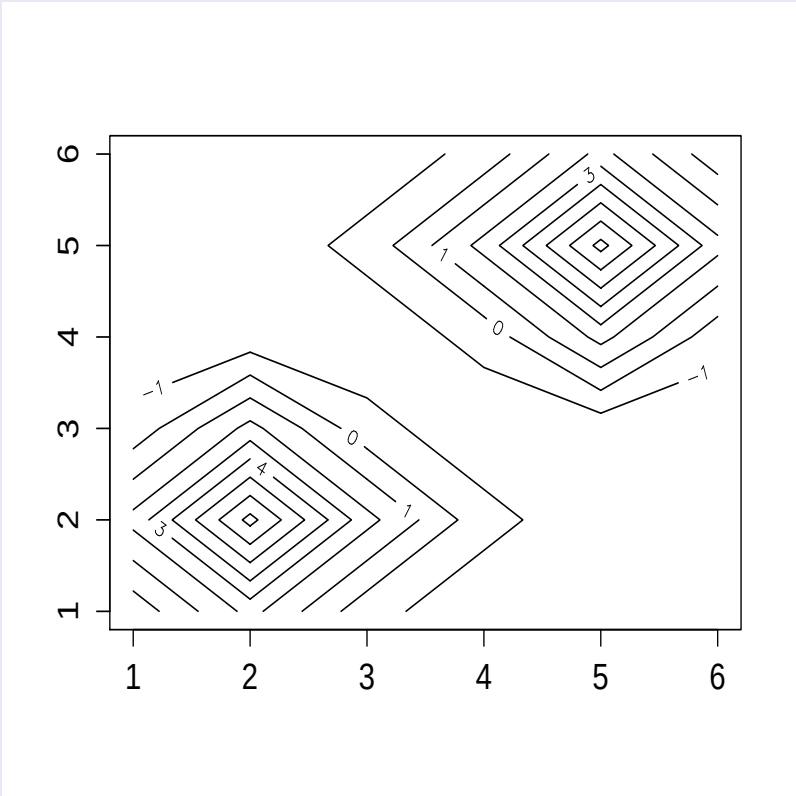
## Principal Component Analysis is an essential tool in high dimensional data analysis

- What is PCA?
- How does PCA uncover patterns in high dimensional data?
- Can we interpret the principal component (eigen) vectors?
- Aligning the principal components with the known patterns in the data.
- Generalizations of principal components.
- Other useful principal components.

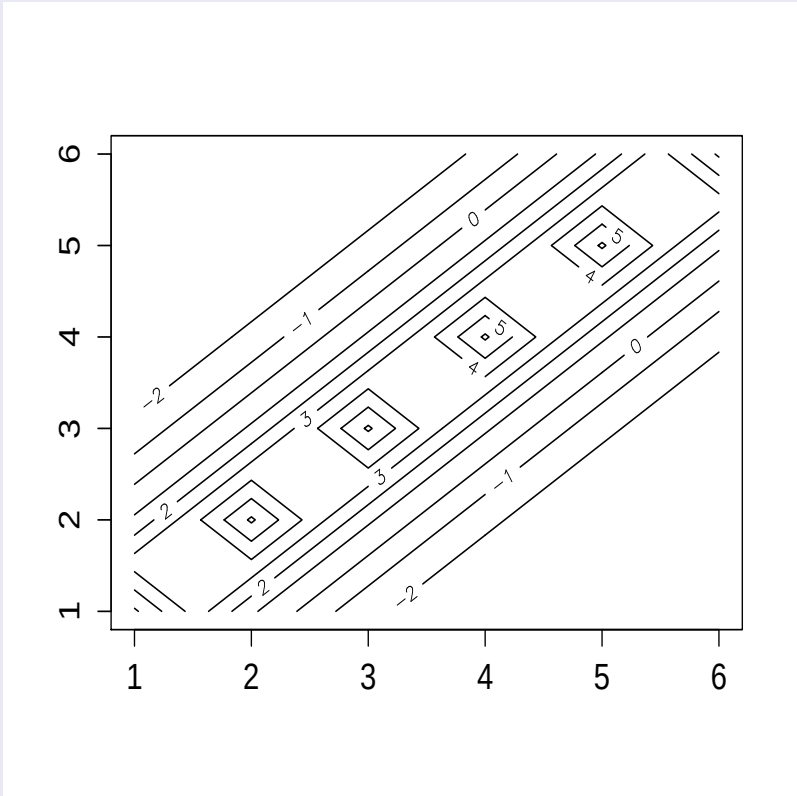
# What is PCA?

## Spatial Patterns

Suppose we have spatial patterns that “generate” our data.




Pattern 1

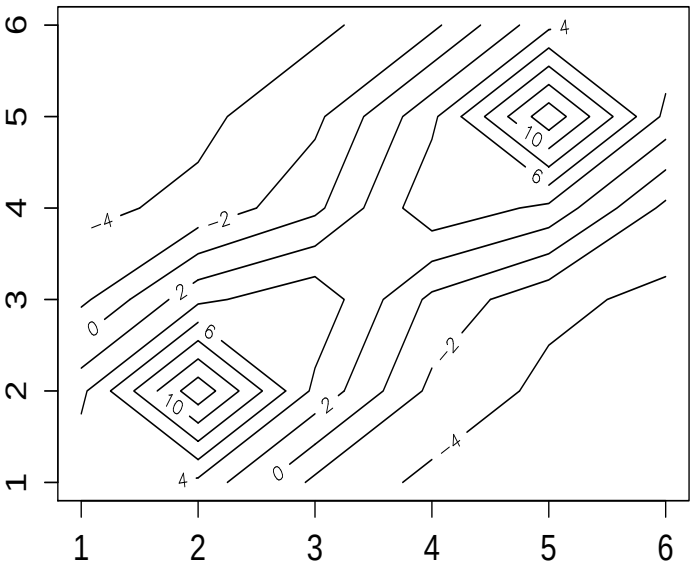



Pattern 2

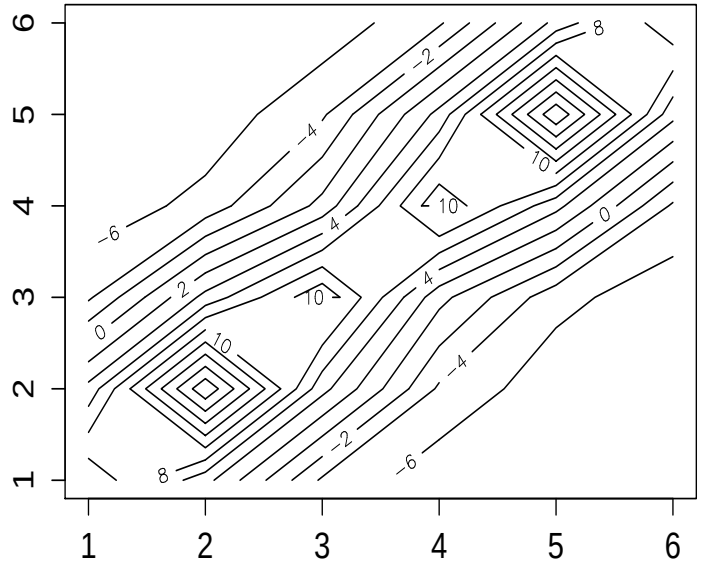
# What is PCA?

## Spatial Patterns

The ideal data are a time evolution of these patterns.




Pattern 1 + Pattern 2

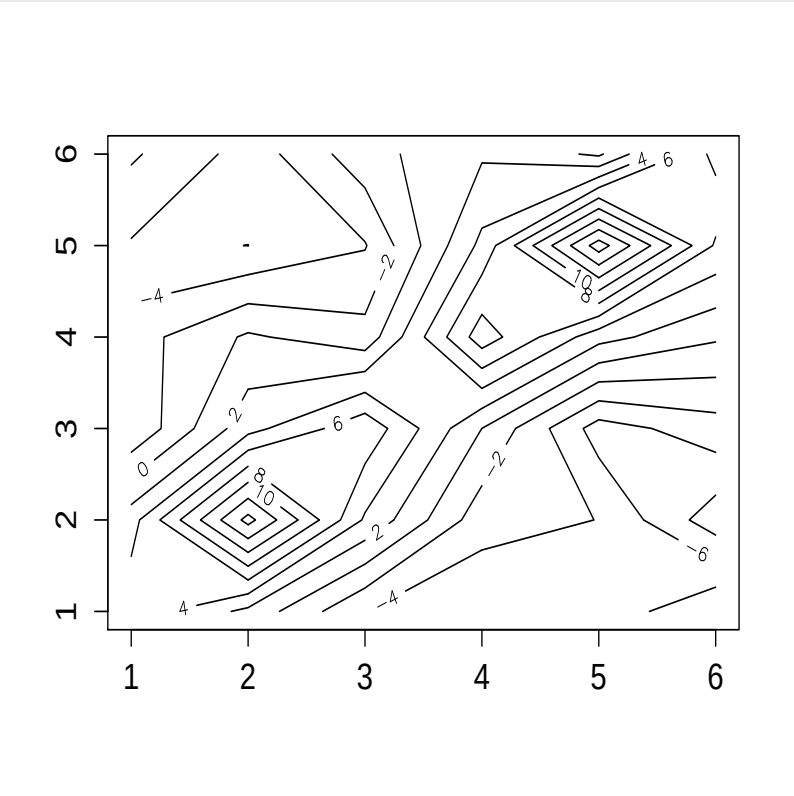



Pattern 1 + 2 × Pattern 2

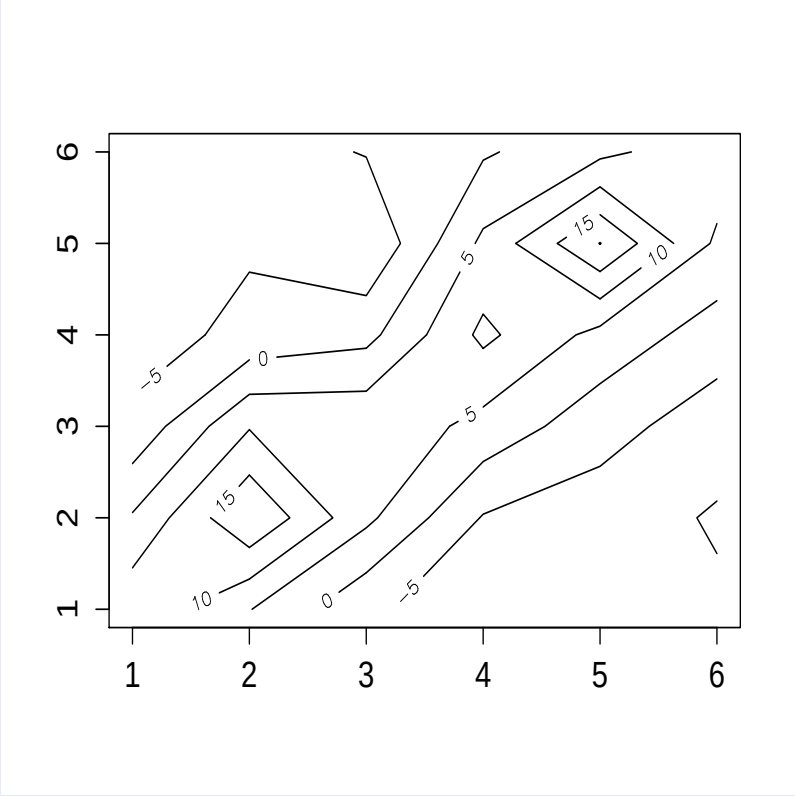
# What is PCA?

## Spatial Patterns

The observed data are a time evolution of these patterns **plus noise**.




Pattern 1 + Pattern 2+noise




Pattern 1 + 2 $\times$ Pattern 2+noise

# What is PCA?

## Spatial Patterns

Principal components analysis is often used as a means of uncovering the underlying component patterns.

# What is PCA?

## Spatial Patterns

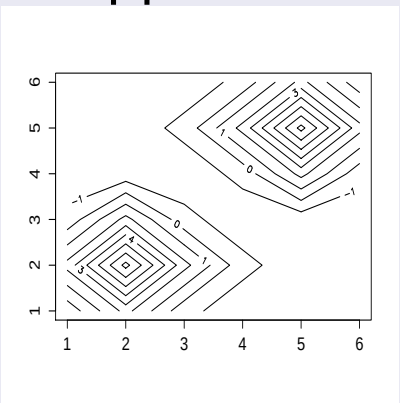
Principal components analysis is often used as a means of uncovering the underlying component patterns.
What happens when the data are the **actual** PCs?

# What is PCA?

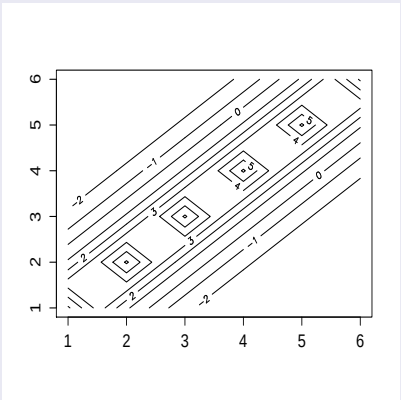
## Spatial Patterns

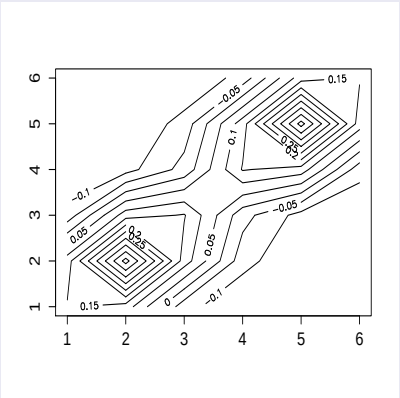
Principal components analysis is often used as a means of uncovering the underlying component patterns.

What happens when the data are the **actual** PCs?

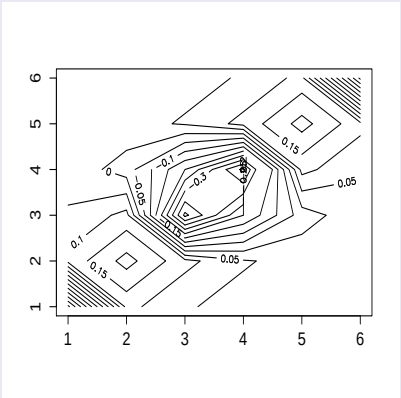


Pattern





PC



# What is PCA?

## Spatial Patterns

What happened when the data are the **actual** PCs?

Principal components must be **uncorrelated**.

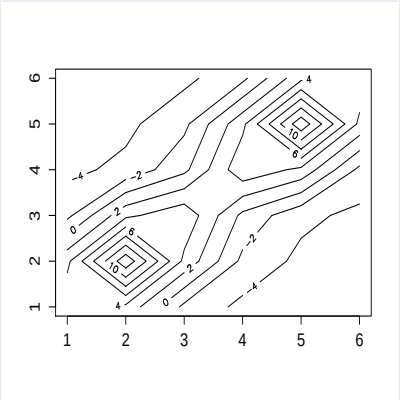
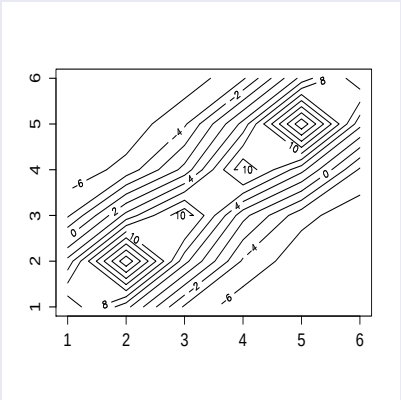
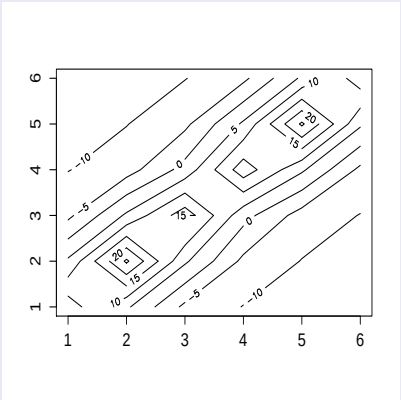
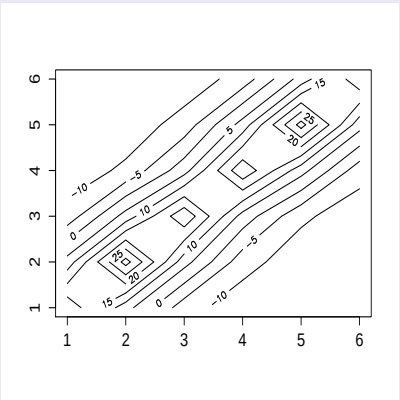
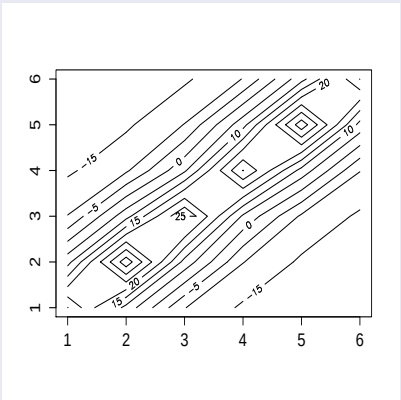
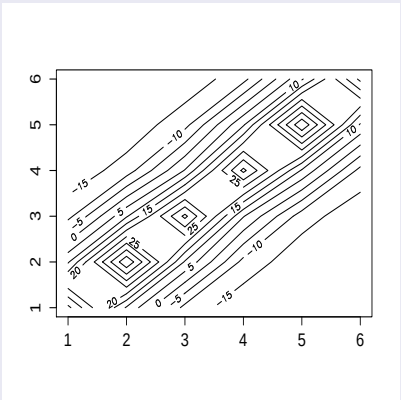
The 2 original patterns had correlation 0.6.

Only 2 PCs can be found (since there are only 2 time points). But they might not be the original patterns.

However, the original patterns are linear combinations of the 2 PCs.

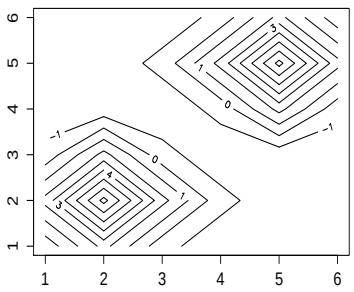
## Spatial Patterns

What happens when the data are the noise-free linear combinations of the 2 PCs "evolving" over time?

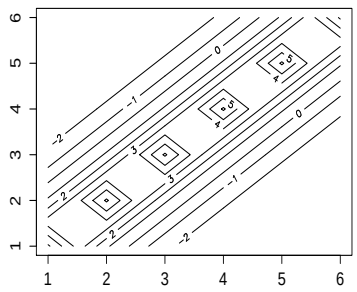







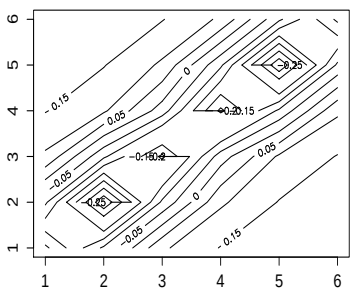
## Spatial Patterns

PCs from the noise-free linearly evolving patterns

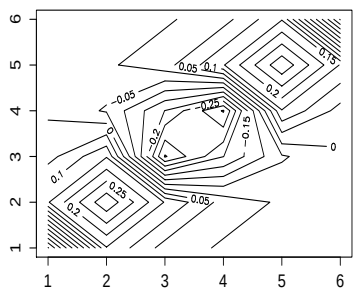


Pattern





PC



# What is PCA?

## Computing the Principal Components

In the context of climate data:

- Some variable (e.g. sea level pressure) has been interpolated to a spatial grid.
- “Observations" have been averaged to a discrete set of times at each grid point.
- At each grid point, the time average of the variable is subtracted off.
- A data matrix $X$ is formed by stringing out the 2-D (or even 3-D) spatial data into the columns.
- Each time point is a row.

# What is PCA?

## Computing the Principal Components

In the context of climate data:

- The area-weighted (empirical) covariance matrix of the spatial data is computed.
- $V = X^\top AX$ where $A$ is a diagonal matrix of areas.
- To avoid notation problems, we are going to replace $X$ by $X\sqrt{(A)}$, so we can write $V = X^\top X$.
- The ordered eigenvectors of the covariance matrix are the principal components.
- The eigenvalues are the variance of the projection of the area-weighted data on the principal components.

# What is PCA?

## Computing the Principal Components

In the context of climate data:

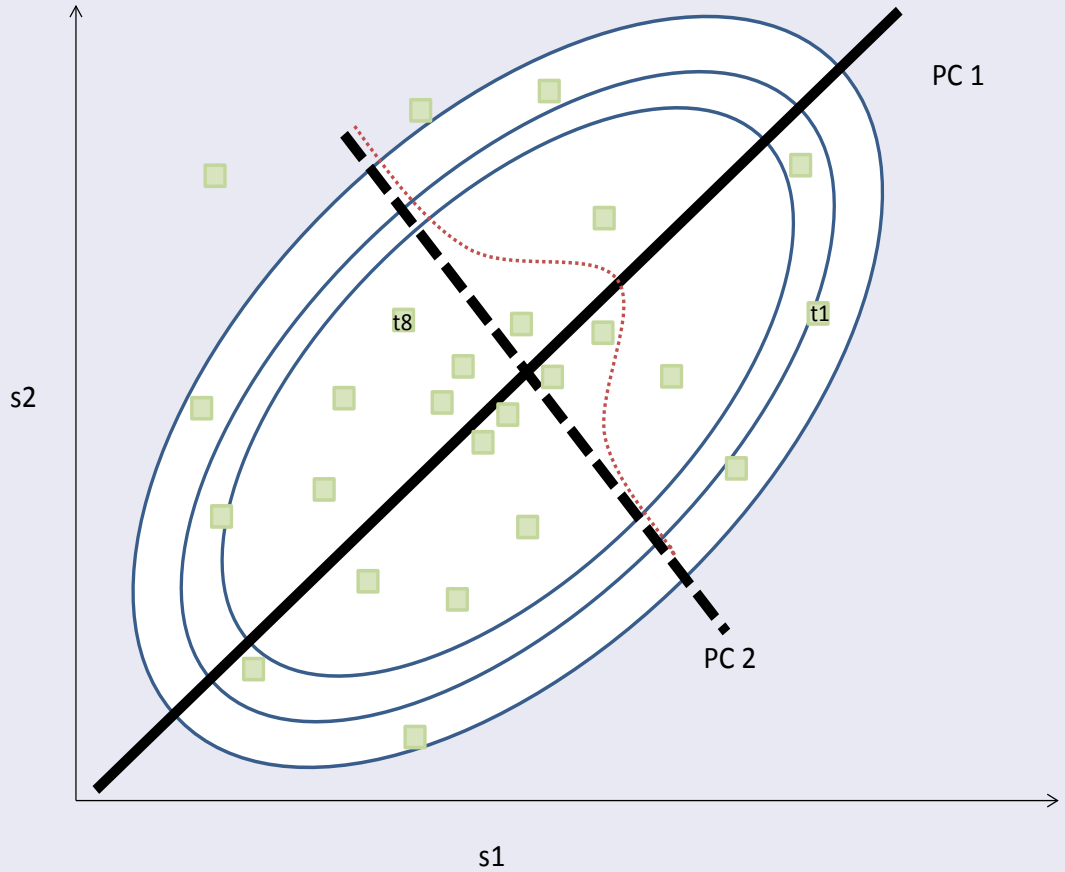
- The area-weighted (empirical) covariance matrix of the spatial data is computed.
- $V = X^\top AX$ where $A$ is a diagonal matrix of areas.
- To avoid notation problems, we are going to replace $X$ by $X\sqrt{(A)}$, so we can write $V = X^\top X$.
- The ordered eigenvectors of the covariance matrix are the principal components.
- The eigenvalues are the variance of the projection of the area-weighted data on the principal components.

Notice that the spatial information is not used - only variability over time.

# What is PCA?

## Elliptical Distributions

For elliptical distributions, the PCs are the major axes of the density ellipsoid.



# Finding Patterns in Data

## The Latent Factor Model

- Suppose we have $r$ “spatial patterns" the latent factors.
- These patterns are the oscillations or modes of variation.
- Each pattern can be summarized by a number at each location.
- This leads to an $r \times p$ matrix $Q$ with $r < p$ and full rank $r$.

# Finding Patterns in Data

## The Latent Factor Model

- Suppose we have $r$ “spatial patterns" the latent factors.
- These patterns are the oscillations or modes of variation.
- Each pattern can be summarized by a number at each location.
- This leads to an $r \times p$ matrix $Q$ with $r < p$ and full rank $r$.
- At any given time $t$ there is a weighting vector $c_t$ and we observe $\mu + c_t^\top Q + \epsilon_t$ at the locations, where $\mu_s$ is the field mean at location $s$ and $\epsilon_t$ is a noise component uncorrelated with $c_t$.
- PCA treats the $c_t$'s as independent random vectors with unit variance.
- The $c_t$'s for the $n$ observation times are then accumulated into an $n \times r$ matrix $C$ and the $\epsilon_t$'s are accumulated into an $n \times p$ noise matrix.
- Then the centered data are $X = CQ + \epsilon$.

## The Latent Factor Model

Under the assumption

$$X = CQ + \epsilon$$

we have

$$(n-1)Var(X) = E(X^\top X) = Q^\top Q + \sigma_\epsilon^2 I$$

PCA can "recover" $Q$ in the sense that an spatial pattern that is a linear combination of the $r$ latent factors can also be expressed as a linear combination of the first $r$ principal components.

This is because the $r$ eigenvectors of $Q$ associated with the non-zero eigenvalues are also eigenvectors of Var(X).

And the eigenvalues of $Q^\top Q + \sigma_\epsilon^2 I$ are the eigenvalues of $Q^\top Q + \sigma_\epsilon^2$.

## Empirical PCA

The result on the previous slide used the “expected value" of the sample variance which we do not get to see.

We have only the observed value $X^\top AX$.

It turns out we do not need to appeal to large sample theory to use empirical PCA.

This is because of the properties of the singular value decomposition (SVD).

## Singular Value Decomposition

Let $M$ be an $n \times p$ matrix.
Then $M = udv^\top$ where

- $u$ is $n \times n$ orthonormal (left singular vectors).
- $d$ is $n \times p$ upper diagonal with $d_1 \geq d_2 \cdots d_p \geq 0$ (singular values).
- $v$ is $p \times p$ orthonormal (right singular vectors).
- If the singular values are unique then $u$ and $v$ are well-defined.

## Singular Value Decomposition

Let $M$ be an $n \times p$ matrix.
Then $M = udv^\top$ where

- $u$ is $n \times n$ orthonormal (left singular vectors).
- $d$ is $n \times p$ upper diagonal with $d_1 \geq d_2 \cdots d_p \geq 0$ (singular values).
- $v$ is $p \times p$ orthonormal (right singular vectors).
- If the singular values are unique then $u$ and $v$ are well-defined.

## SVD and empirical PCA

Let $X$ be the centered data matrix with SVD $X = udv^\top$.
$X^\top X = vd^2v^\top$ so:

## Singular Value Decomposition

Let $M$ be an $n \times p$ matrix.
Then $M = udv^\top$ where

- $u$ is $n \times n$ orthonormal (left singular vectors).
- $d$ is $n \times p$ upper diagonal with $d_1 \geq d_2 \cdots d_p \geq 0$ (singular values).
- $v$ is $p \times p$ orthonormal (right singular vectors).
- If the singular values are unique then $u$ and $v$ are well-defined.

## SVD and empirical PCA

Let $X$ be the centered data matrix with SVD $X = udv^\top$.
$X^\top X = vd^2v^\top$ so:

- The principal components of $X$ are the right singular vectors of $X$.

## Singular Value Decomposition

Let $M$ be an $n \times p$ matrix.
Then $M = udv^\top$ where

- $u$ is $n \times n$ orthonormal (left singular vectors).
- $d$ is $n \times p$ upper diagonal with $d_1 \geq d_2 \cdots d_p \geq 0$ (singular values).
- $v$ is $p \times p$ orthonormal (right singular vectors).
- If the singular values are unique then $u$ and $v$ are well-defined.

## SVD and empirical PCA

Let $X$ be the centered data matrix with SVD $X = udv^\top$.
$X^\top X = vd^2v^\top$ so:

- The principal components of $X$ are the right singular vectors of $X$.
- The variances associated with the principal components are the squares of the singular values of $X$.

# Finding Patterns in Data

Principal components analysis is effective at finding patterns in data because of two strongly related properties of the right singular vectors of $X$.

Let $B$ $n \times k$ and $W$ $p \times k$ both have rank $k$.

For a matrix $M$ let $M[+k]$ be the matrix of the first $k$ columns of $M$.

## Matrix nearness and SVD

A pair minimizing $\| X - BW^\top \|$ is $W = v[+k]$ and $B = Xv[+k]$.

# Finding Patterns in Data

Principal components analysis is effective at finding patterns in data because of two strongly related properties of the right singular vectors of $X$.

Let $B$ $n \times k$ and $W$ $p \times k$ both have rank $k$.

For a matrix $M$ let $M[+k]$ be the matrix of the first $k$ columns of $M$.

## Matrix nearness and SVD

A pair minimizing $\| X - BW^\top \|$ is $W = v[+k]$ and $B = Xv[+k]$.

## Latent variable regression

A pair minimizing $\| X - XBW^\top \|$ is $B = W = v[+k]$.

# Finding Patterns in Data

Principal components analysis is effective at finding patterns in data because of two strongly related properties of the right singular vectors of $X$.

Let $B$ $n \times k$ and $W$ $p \times k$ both have rank $k$.

For a matrix $M$ let $M[+k]$ be the matrix of the first $k$ columns of $M$.

## Matrix nearness and SVD

A pair minimizing $\| X - BW^\top \|$ is $W = v[+k]$ and $B = Xv[+k]$.

## Latent variable regression

A pair minimizing $\| X - XBW^\top \|$ is $B = W = v[+k]$.

Furthermore

$$\| X - Xv[+k]v[+k]^\top \|^2 = \sum_{i=k+1}^{p} \delta_i^2$$

# Finding Patterns in Data

## PCs, matrix nearness and prediction

- Matrix nearness and SVD: The PC's give the best low rank approximation to the centered data.
- Latent variable regression: The PC's are the best predictors of X (for given rank $k$).

## BUT ...

The matrix nearness and best predictor results are not unique. Replace $W = v[+k]$ by $WG$ for any invertible $G$.

## Latent variable regression

- Denoising
  - The last $p - r$ principal components are essentially *noise*.
  - So $Xv[+r]v[+r]^\top$ is a *denoised* version of $X$.

# Finding Patterns in Data

## Latent variable regression

- Missing value imputation
- Pattern searching

# Finding Patterns in Data

## Latent variable regression

- Missing value imputation
  - Regression of $X$ on the first $k$ principal components provides the best possible linear predictor of $X$ on any $k$-dimensional basis.
  - So imputing missing values by their predicted values provides the best linear imputation.
- Pattern searching

# Finding Patterns in Data

## Latent variable regression

- Missing value imputation
  - Regression of $X$ on the first $k$ principal components provides the best possible linear predictor of $X$ on any $k$-dimensional basis.
  - So imputing missing values by their predicted values provides the best linear imputation.
- Pattern searching
  - $v_1 \cdots v_k$ are a basis of the predictor space ordered by $R^2$.

# Finding Patterns in Data

## Latent variable regression

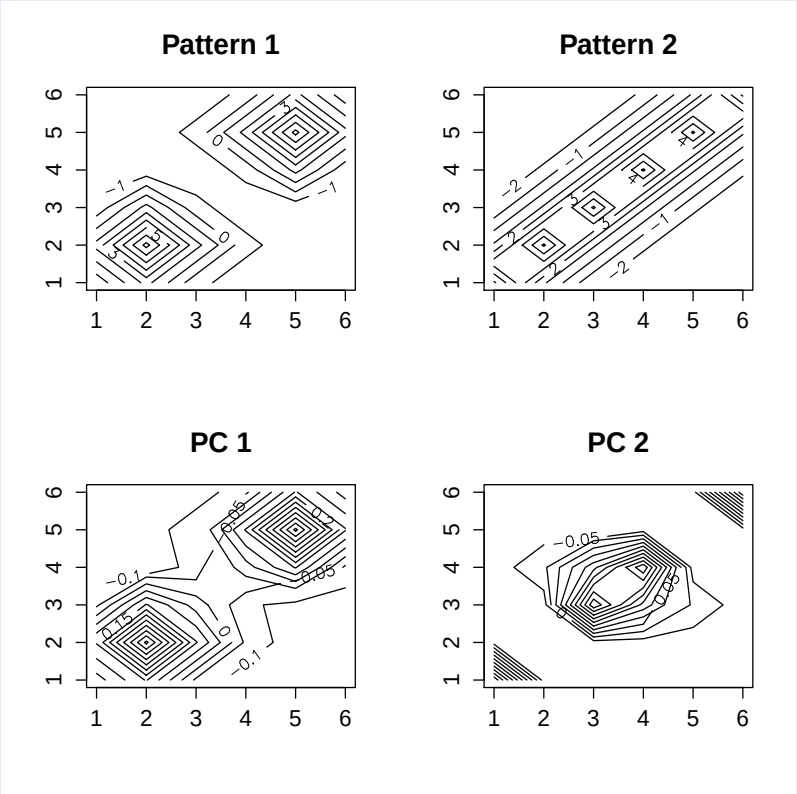
- Missing value imputation
  - Regression of $X$ on the first $k$ principal components provides the best possible linear predictor of $X$ on any $k$-dimensional basis.
  - So imputing missing values by their predicted values provides the best linear imputation.
- Pattern searching
  - $v_1 \cdots v_k$ are a basis of the predictor space ordered by $R^2$.
  - So, $v_1$ could be thought of as the dominant signal in the data.
  - $v_i$ is the dominant signal orthogonal to the signals with higher SNR.

# Interpreting PCA

## What are the PCs?

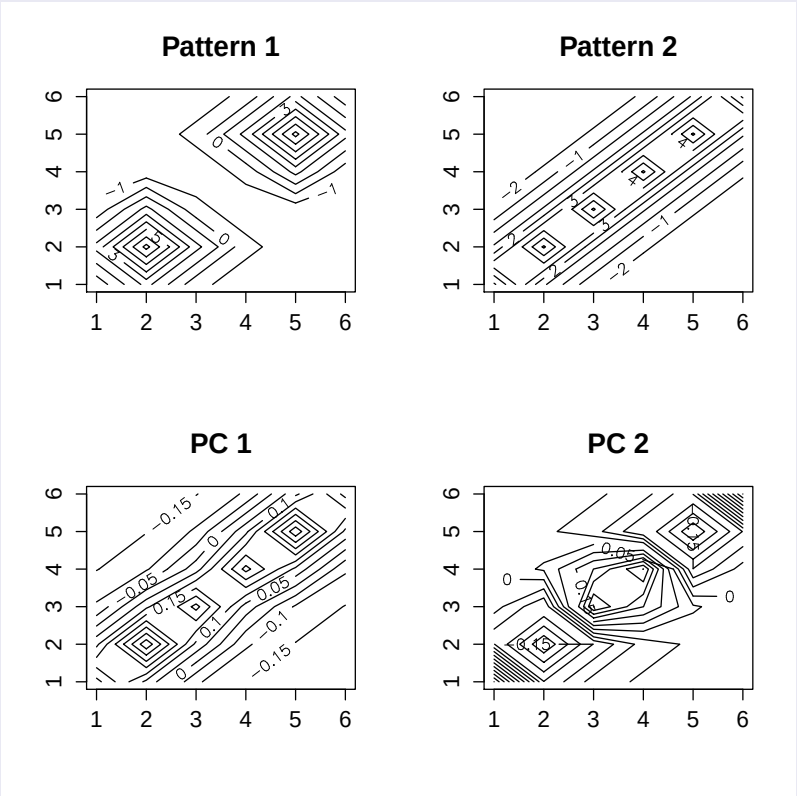
Problems arise when you try to interpret the PCs - especially the higher order PCs.

For example, suppose there are 4 times with patterns 1,1,1,2 and no noise.



## What are the PCs?

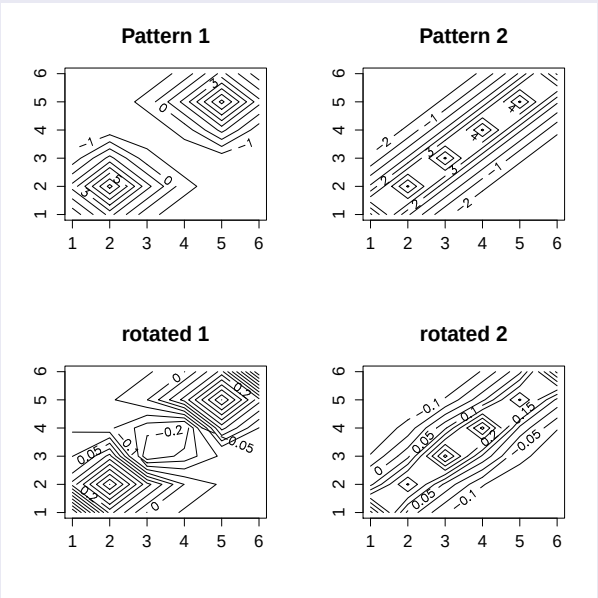
Suppose instead there are 4 times with patterns 1,2,2,2 and no noise.



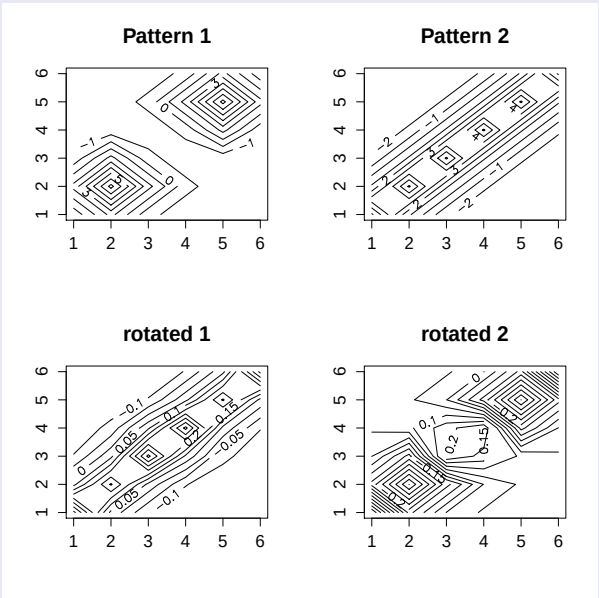
# Interpreting PCA

## What are the PCs?

One idea is to use varimax rotations to “align" the PCs with the data.




Pattern 1,1,1,2




Pattern 1,2,2,2

# Interpreting PCA

## What are the PCs?

These days statisticians use “sparse" PCA to force some of the loadings to be zero.
This is done through a penalized latent variable regression.
This gave

- PC1 = pattern 1, PC2 = pattern 2 for 1,1,1,2
- PC1 = pattern 2, PC2 = pattern 1 for 1,2,2,2
- which is exactly correct.

# Latent variable regression

Latent variable regression provides a framework for extensions of PCA.

## Extended PCA via regression

- Minimum Distance SVD
- Penalized PCA

# Extending PCA via regression

## Minimum Distance

- Robust PCA:
- Minimum distance SVD
  - Bregman SVD (generalized PCA)
  - Maximum Likelihood SVD

# Extending PCA via regression

## Minimum Distance

- Robust PCA:
  - To reduce sensitivity to outliers, use robust regression instead of least squares regression in the latent variable regression.
- Minimum distance SVD
  - Bregman SVD (generalized PCA)
  - Maximum Likelihood SVD

# Extending PCA via regression

## Minimum Distance

- Robust PCA:
  - To reduce sensitivity to outliers, use robust regression instead of least squares regression in the latent variable regression.
  - There is another form of robust PCA using the eigen-decomposition of a robust variance estimator.
- Minimum distance SVD
  - Bregman SVD (generalized PCA)
  - Maximum Likelihood SVD

# Extending PCA via regression

## Minimum Distance

- Robust PCA:
  - To reduce sensitivity to outliers, use robust regression instead of least squares regression in the latent variable regression.
  - There is another form of robust PCA using the eigen-decomposition of a robust variance estimator.
- Minimum distance SVD
  - Replace $L_2$ distance with another metric or divergence in the SVD matrix nearness problem.
  - Bregman SVD (generalized PCA)
  - Maximum Likelihood SVD

# Extending PCA via regression

## Minimum Distance

- Robust PCA:
  - To reduce sensitivity to outliers, use robust regression instead of least squares regression in the latent variable regression.
  - There is another form of robust PCA using the eigen-decomposition of a robust variance estimator.
- Minimum distance SVD
  - Replace $L_2$ distance with another metric or divergence in the SVD matrix nearness problem.
  - Bregman SVD (generalized PCA)
  - In linear exponential families, do uncentered SVD using Bregman divergence (generalized linear model, MLE)
  - Maximum Likelihood SVD

# Latent variable regression

## Extended PCA via regression

- Penalized PCA

# Latent variable regression

## Extended PCA via regression

- Penalized PCA
  - e.g. Sparse PCA (Zou, Hastie and Tibshirani, 2006)
    - Let $x_i$ be the $i^{th}$ row of the data matrix $x$.
    - Let $A$ be an orthonormal $n \times k$ matrix.
    - Let $W$ be an $n \times k$ matrix with rank $k$ and columns $W_i$.
  - To extract $k$ penalized "principal components" minimize:

$$\sum_{i=1}^{n} \| x_i - x_i A W^\top \|^2 + \sum_{i=1}^{k} P_\lambda(W_i)$$

Related methods penalize $W$ in the matrix nearness formulation.

# Latent variable regression

## Extended PCA via regression

- Penalized PCA
  - e.g. Sparse PCA (Zou, Hastie and Tibshirani, 2006)
    - Let $x_i$ be the $i^{th}$ row of the data matrix $x$.
    - Let $A$ be an orthonormal $n \times k$ matrix.
    - Let $W$ be an $n \times k$ matrix with rank $k$ and columns $W_i$.
  - To extract $k$ penalized "principal components" minimize:

$$\sum_{i=1}^{n} \| x_i - x_i A W^\top \|^2 + \sum_{i=1}^{k} P_\lambda(W_i)$$

  - This could also be put in the context of Bayesian models, for which the penalty is the prior log-likelihood.

Related methods penalize $W$ in the matrix nearness formulation.

# Latent variable regression

## Extended PCA via regression

- Kernel PCA

# Latent variable regression

## Extended PCA via regression

- Kernel PCA
  - A set of nonlinear basis functions $\phi_i : \Re^p \rightarrow \Re$ is selected $i = 1 \cdots p > P$.
  - The Euclidean covariance matrix is replaced by the kernel covariance matrix $\hat{\Sigma} = \sum_{j=1}^{n} \phi(x_j)\phi(x_j)^\top$ where $x_j$ is the $j^{th}$ row of the data matrix.
  - Kernel PCA finds a basis for approximating nonlinear functions of the data.

# Other Ideas

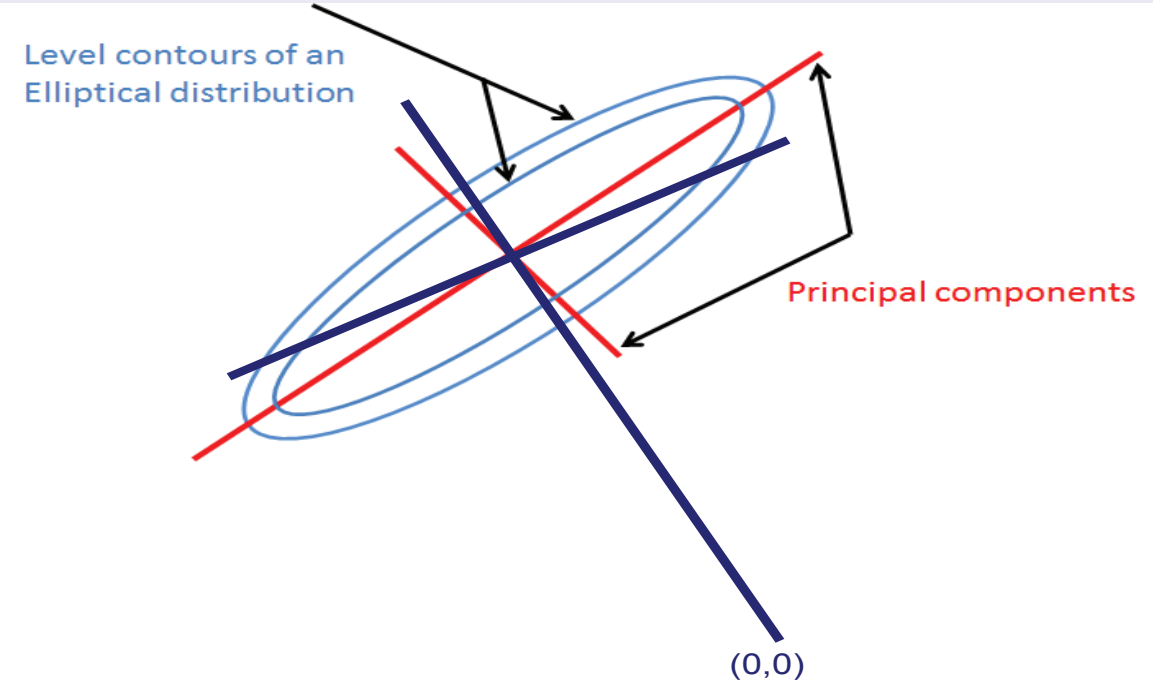
## Centering

- Since the right singular vectors are the important quantities, should we center?
- It makes little difference to the fitted values, but affects interpretation.
- Discussed in detail in Cadima and Jolliffe(2009).

# Other Ideas

## Centering

- Since the right singular vectors are the important quantities, should we center?
- It makes little difference to the fitted values, but affects interpretation.
- Discussed in detail in Cadima and Jolliffe(2009).



# Other Ideas

## Other Matrix Factorizations

- If the data are non-negative, non-negative matrix factorizations (NMF) may be more interpretable.
- Independent Component Analysis (ICA)
- Factor Analysis

# Other Ideas

## Other Matrix Factorizations

- If the data are non-negative, non-negative matrix factorizations (NMF) may be more interpretable.
- This looks like (generalized) SVD with non-negativity constraints but has some surprising properties (non-nested).
- Independent Component Analysis (ICA)
- Factor Analysis

# Other Ideas

## Other Matrix Factorizations

- If the data are non-negative, non-negative matrix factorizations (NMF) may be more interpretable.
- This looks like (generalized) SVD with non-negativity constraints but has some surprising properties (non-nested).
- NMF tends to be sparse.
- Can be used for clustering (similar to k-means).
- Independent Component Analysis (ICA)
- Factor Analysis

# Conclusions

## PCA is excellent for finding a low dimensional basis for climate patterns.

# Conclusions

**PCA is excellent for finding a low dimensional basis for climate patterns.**

- Latent variable regression and matrix nearness (i.e. SVD) explain many of the good properties of PCA.

# Conclusions

**PCA is excellent for finding a low dimensional basis for climate patterns.**

- Latent variable regression and matrix nearness (i.e. SVD) explain many of the good properties of PCA.
- Latent variable regression and matrix nearness unify methods for extending PCA by using penalized versions or different norms.

# Conclusions

**PCA is excellent for finding a low dimensional basis for climate patterns.**

- Latent variable regression and matrix nearness (i.e. SVD) explain many of the good properties of PCA.
- Latent variable regression and matrix nearness unify methods for extending PCA by using penalized versions or different norms.
- Other matrix decompositions may be at least as useful - e.g. various NMF methods.

# Conclusions

**PCA is excellent for finding a low dimensional basis for climate patterns.**

- Latent variable regression and matrix nearness (i.e. SVD) explain many of the good properties of PCA.
- Latent variable regression and matrix nearness unify methods for extending PCA by using penalized versions or different norms.
- Other matrix decompositions may be at least as useful - e.g. various NMF methods.
- Interpretation of the PCs as climate patterns or modes is not scientifically justified.

# Conclusions

**PCA is excellent for finding a low dimensional basis for climate patterns.**

- Latent variable regression and matrix nearness (i.e. SVD) explain many of the good properties of PCA.
- Latent variable regression and matrix nearness unify methods for extending PCA by using penalized versions or different norms.
- Other matrix decompositions may be at least as useful - e.g. various NMF methods.
- Interpretation of the PCs as climate patterns or modes is not scientifically justified.
- If there are “known" patterns, regress these out and use PCA on the residuals!

# Thanks to


## NSF

For funding part of the work through DMS 1007801.

## NIH

For funding part of the work through NIH UL1RR033184.

## SAMSI (NSF)

For funding part of the work and providing an environment in which the work could be pursued and discussed.

## Participants of SAMSI Working Groups: Inference, Streaming and Sketching, Datamining and Clustering

For stimulating discussions of the role of PCA in various settings, robust PCA and for discussions of earlier versions of the work.

## Wei Luo (Baruch College), Garvesh Raskutti (U. Wisconsin), Frank Shen (Penn State)

**Penn State University**