An Introduction to Nonlinear Principal Component Analysis

Adam Monahan

monahana@uvic.ca

School of Earth and Ocean Sciences

University of Victoria

UVic

Overview

- Dimensionality reduction

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA
    - theory

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA
  - theory
  - implementation

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA
  - theory
  - implementation
- Applications of NLPCA

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA
  - theory
  - implementation
- Applications of NLPCA
  - Lorenz attractor

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA
  - theory
  - implementation
- Applications of NLPCA
  - Lorenz attractor
  - NH Tropospheric LFV

UVic

Overview

- Dimensionality reduction
- Principal Component Analysis
- Nonlinear PCA
  - theory
  - implementation
- Applications of NLPCA
  - Lorenz attractor
  - NH Tropospheric LFV
- Conclusions

UVic

Dimensionality Reduction

- Climate datasets made up of time series at individual stations/geographical locations

Dimensionality Reduction

- Climate datasets made up of time series at individual stations/geographical locations
- Typical dataset has $P \sim O(10^3)$ time series

UVic

Dimensionality Reduction

- Climate datasets made up of time series at individual stations/geographical locations
- Typical dataset has $P \sim O(10^3)$ time series
- Organised structure in atmosphere/ocean flows

UVic

Dimensionality Reduction

- Climate datasets made up of time series at individual stations/geographical locations
- Typical dataset has $P \sim O(10^3)$ time series
- Organised structure in atmosphere/ocean flows

$\Rightarrow$ time series at different locations not independent

Dimensionality Reduction

- Climate datasets made up of time series at individual stations/geographical locations
- Typical dataset has $P \sim O(10^3)$ time series
- Organised structure in atmosphere/ocean flows

$\Rightarrow$ time series at different locations not independent

$\Rightarrow$ data does not fill out isotropic cloud of points in $\mathbf{R}^P$, but clusters around lower-dimensional surface (reflecting the “attractor”)

UVic

Dimensionality Reduction

- Climate datasets made up of time series at individual stations/geographical locations
- Typical dataset has $P \sim O(10^3)$ time series
- Organised structure in atmosphere/ocean flows

$\Rightarrow$ time series at different locations not independent

$\Rightarrow$ data does not fill out isotropic cloud of points in $\mathbf{R}^P$, but clusters around lower-dimensional surface (reflecting the "attractor")

- Goal of *dimensionality reduction* in climate diagnostics is to characterise such structures in climate datasets

UVic

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:
- **Theoretical:**

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:
- **Theoretical:**
  - what is the precise definition of "structure"?

UVic

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:
- **Theoretical:**
  - what is the precise definition of “structure”?
  - how to formulate appropriate statistical model?

UVic

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:
- **Theoretical:**
  - what is the precise definition of “structure”?
  - how to formulate appropriate statistical model?
- **Practical:**

UVic

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:
- **Theoretical:**
  - what is the precise definition of “structure”?
  - how to formulate appropriate statistical model?
- **Practical:**
  - many important observational climate datasets quite short, with $O(10) - O(1000)$ statistical degrees of freedom

UVic

Dimensionality Reduction

- Realising this goal has both theoretical and practical difficulties:
- **Theoretical:**
  - what is the precise definition of "structure"?
  - how to formulate appropriate statistical model?
- **Practical:**
  - many important observational climate datasets quite short, with $O(10) - O(1000)$ statistical degrees of freedom
  - what degree of "structure" can be robustly diagnosed with existing data?

UVic

Principal Component Analysis

- A classical approach to dimensionality
  principal component analysis (PCA)

Principal Component Analysis

- A classical approach to dimensionality
  principal component analysis (PCA)
- Look for M-dimensional hyperplane approximation, optimal in least-squares sense

$$\mathbf{X}(t) = \sum_{k=1}^{M} \langle \mathbf{X}(t), \mathbf{e}_k \rangle \, \mathbf{e}_k + \epsilon(t)$$

UVic

Principal Component Analysis

- A classical approach to dimensionality **principal component analysis (PCA)**
- Look for M-dimensional hyperplane approximation, optimal in least-squares sense

$$\mathbf{X}(t) = \sum_{k=1}^{M} \langle \mathbf{X}(t), \mathbf{e}_k \rangle \, \mathbf{e}_k + \epsilon(t)$$

  - minimising $E\left\{||\epsilon^2||\right\}$

UVic

Principal Component Analysis

- A classical approach to dimensionality
  principal component analysis (PCA)
- Look for M-dimensional hyperplane approximation, optimal in least-squares sense

$$\mathbf{X}(t) = \sum_{k=1}^{M} \langle \mathbf{X}(t), \mathbf{e}_k \rangle \, \mathbf{e}_k + \epsilon(t)$$

  - minimising $E\left\{||\epsilon^2||\right\}$
  - inner product often (not always) simple dot product

UVic

Principal Component Analysis

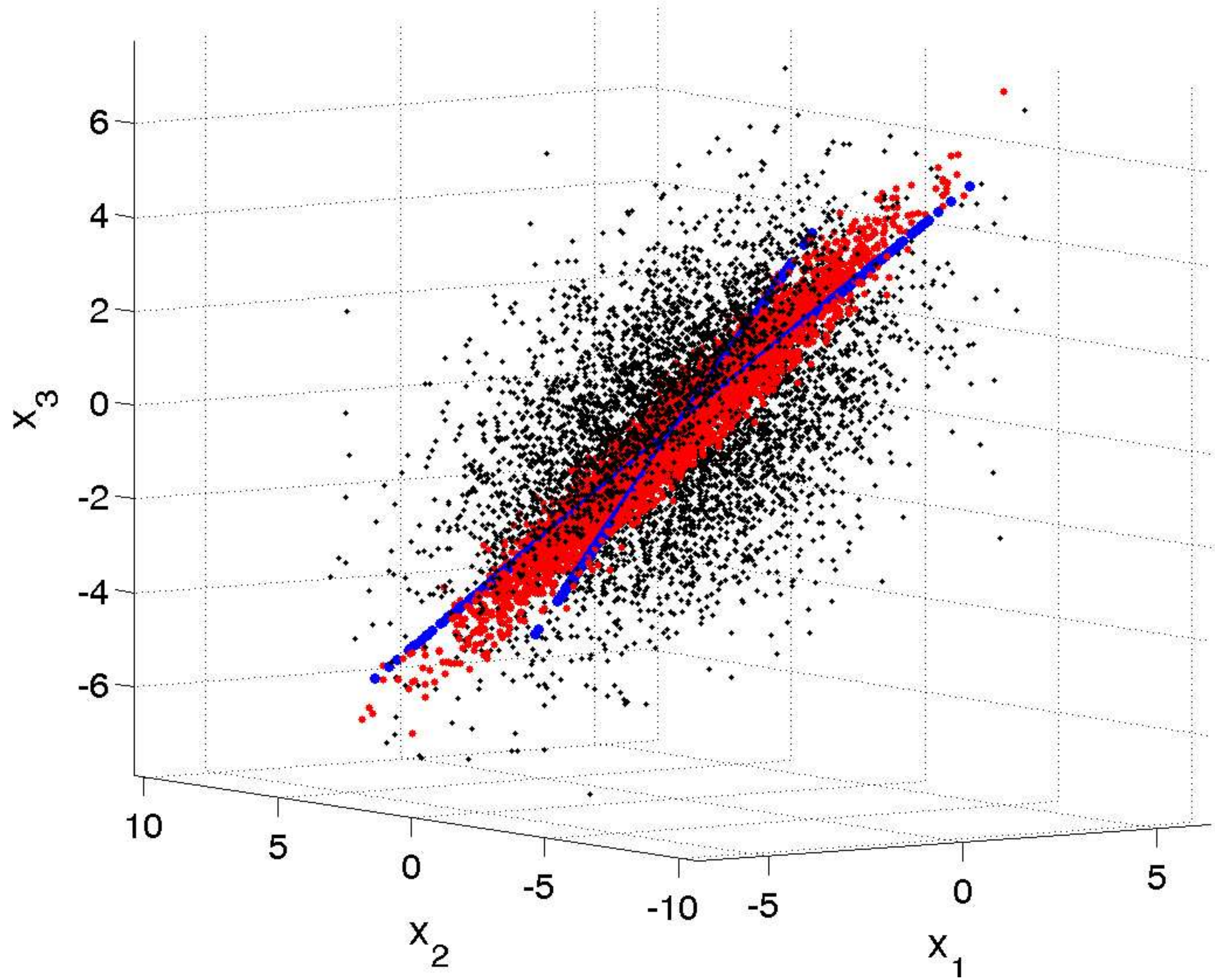
- A classical approach to dimensionality
  principal component analysis (PCA)
- Look for M-dimensional hyperplane approximation, optimal in least-squares sense

$$\mathbf{X}(t) = \sum_{k=1}^{M} \langle \mathbf{X}(t), \mathbf{e}_k \rangle \, \mathbf{e}_k + \epsilon(t)$$

  - minimising $E\left\{||\epsilon^2||\right\}$
  - inner product often (not always) simple dot product
- Vectors $\mathbf{e}_k$ are the **empirical orthogonal functions (EOFs)**

UVic

Principal Component Analysis



Principal Component Analysis

- Operationally, EOFs are found as eigenvectors of covariance matrix (in appropriate norm)

UVic

Principal Component Analysis

- Operationally, EOFs are found as eigenvectors of covariance matrix (in appropriate norm)
- PCA optimally efficient characterisation of Gaussian data

UVic

Principal Component Analysis

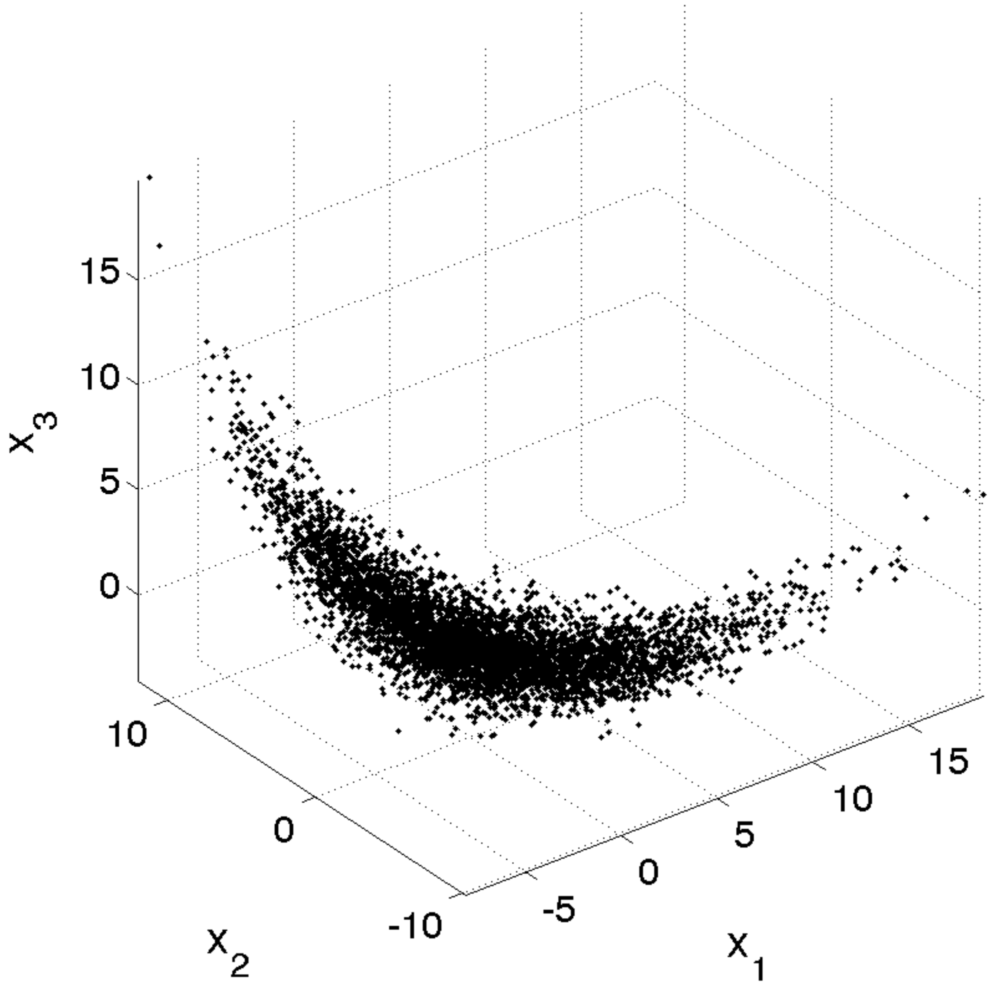
- Operationally, EOFs are found as eigenvectors of covariance matrix (in appropriate norm)
- PCA optimally efficient characterisation of Gaussian data
- More generally: PCA provides optimally parsimonious data compression for any dataset whose distribution lies along orthogonal axes

UVic

Principal Component Analysis

- Operationally, EOFs are found as eigenvectors of covariance matrix (in appropriate norm)
- PCA optimally efficient characterisation of Gaussian data
- More generally: PCA provides optimally parsimonious data compression for any dataset whose distribution lies along orthogonal axes
- But what if the underlying low-dimensional structure is curved rather than straight? (cigars vs. bananas)

Nonlinear Low-Dimensional Structure



Nonlinear PCA

- An approach to diagnosing nonlinear low-dimensional structure is *Nonlinear PCA (NLPCA)*

UVic

Nonlinear PCA

- An approach to diagnosing nonlinear low-dimensional structure is *Nonlinear PCA (NLPCA)*
- Goal: find functions (with $M < P$)

$$\mathbf{s_f} : \mathrm{R}^P \rightarrow \mathrm{R}^M \quad , \quad \mathbf{f} : \mathrm{R}^M \rightarrow \mathrm{R}^P$$

such that

$$\mathbf{X}(t) = (\mathbf{f} \circ \mathbf{s_f})\,(\mathbf{X}(t)) + \epsilon(\mathbf{t})$$

where

UVic

Nonlinear PCA

- An approach to diagnosing nonlinear low-dimensional structure is *Nonlinear PCA (NLPCA)*
- Goal: find functions (with $M < P$)

$$\mathbf{s_f} : \mathrm{R}^P \rightarrow \mathrm{R}^M \quad , \quad \mathbf{f} : \mathrm{R}^M \rightarrow \mathrm{R}^P$$

such that

$$\mathbf{X}(t) = (\mathbf{f} \circ \mathbf{s_f})(\mathbf{X}(t)) + \epsilon(\mathbf{t})$$

where

  - $E\{||\epsilon^2||\}$ is minimised

UVic

Nonlinear PCA

- An approach to diagnosing nonlinear low-dimensional structure is *Nonlinear PCA (NLPCA)*
- Goal: find functions (with $M < P$)

$$\mathbf{s_f} : \mathrm{R}^P \to \mathrm{R}^M \quad , \quad \mathbf{f} : \mathrm{R}^M \to \mathrm{R}^P$$

  such that

$$\mathbf{X}(t) = (\mathbf{f} \circ \mathbf{s_f})\,(\mathbf{X}(t)) + \epsilon(\mathbf{t})$$

  where

  - $E\left\{||\epsilon^2||\right\}$ is minimised
  - $\mathbf{f}(\lambda) \sim$ approximation manifold

UVic

Nonlinear PCA

- An approach to diagnosing nonlinear low-dimensional structure is *Nonlinear PCA (NLPCA)*
- Goal: find functions (with $M < P$)

$$\mathbf{s_f} : \mathrm{R}^P \rightarrow \mathrm{R}^M \quad , \quad \mathbf{f} : \mathrm{R}^M \rightarrow \mathrm{R}^P$$

  such that

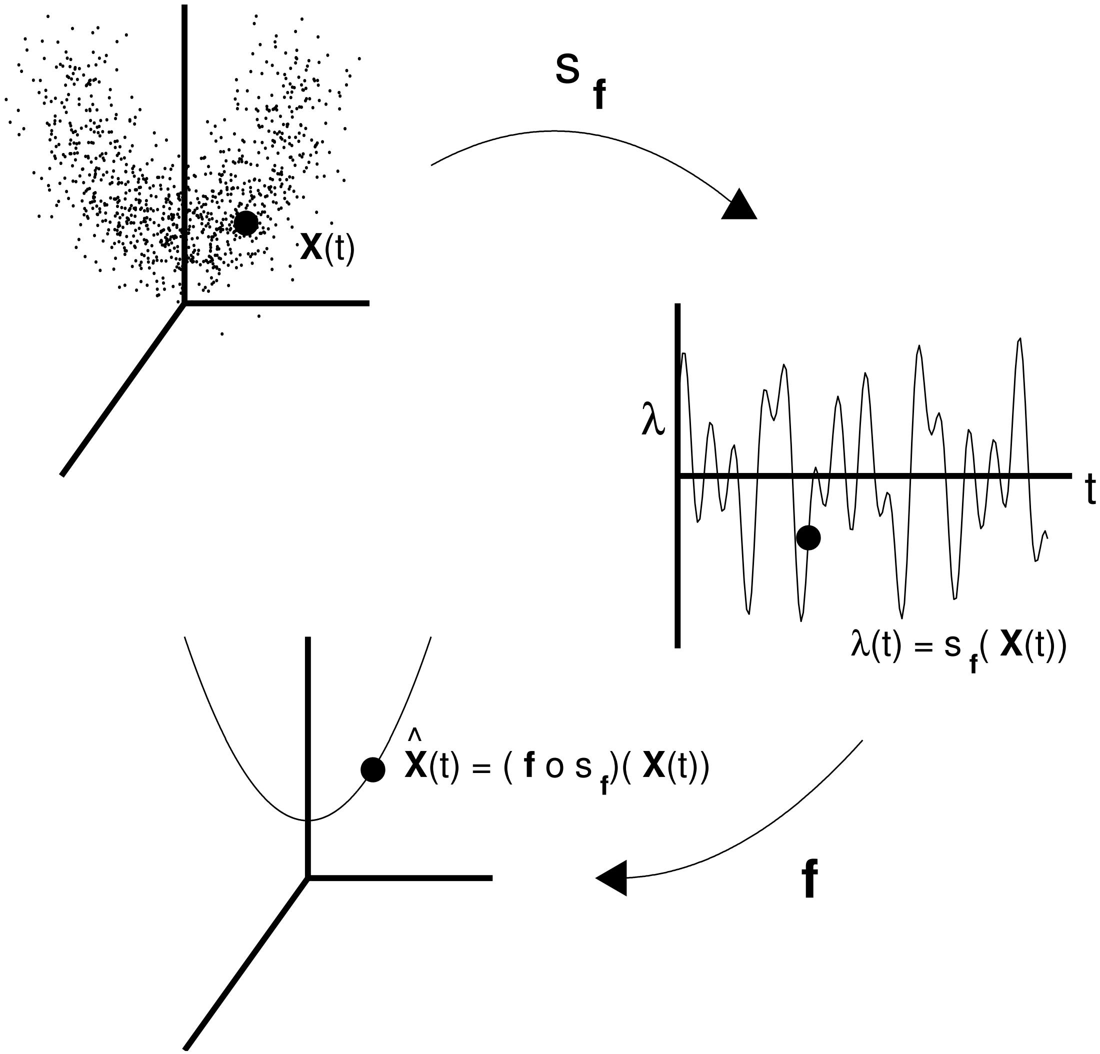
$$\mathbf{X}(t) = (\mathbf{f} \circ \mathbf{s_f})(\mathbf{X}(t)) + \epsilon(\mathbf{t})$$

  where

  - $E\{||\epsilon^2||\}$ is minimised
  - $\mathbf{f}(\lambda) \sim$ approximation manifold
  - $\lambda(t) = \mathbf{s_f}(\mathbf{X}(t)) \sim$ manifold parameterisation (time series)

UVic

Nonlinear PCA



From Monahan, Fyfe, and Pandolfo (2003)

Nonlinear PCA

- As with PCA, “fraction of variance explained” is a measure of quality of approximation

UVic

Nonlinear PCA

- As with PCA, “fraction of variance explained” is a measure of quality of approximation
- PCA is a special case of NLPCA

UVic

Nonlinear PCA

- As with PCA, “fraction of variance explained” is a measure of quality of approximation
- PCA is a special case of NLPCA
- When implemented, NLPCA should reduce to PCA if:

UVic

Nonlinear PCA

- As with PCA, “fraction of variance explained” is a measure of quality of approximation
- PCA is a special case of NLPCA
- When implemented, NLPCA should reduce to PCA if:
  - data is Gaussian

UVic

Nonlinear PCA

- As with PCA, “fraction of variance explained” is a measure of quality of approximation
- PCA is a special case of NLPCA
- When implemented, NLPCA should reduce to PCA if:
  - data is Gaussian
  - not enough data is available to robustly characterise non-Gaussian structure

UVic

NLPCA: Implementation

- Implemented NLPCA using neural networks (convenient, not necessary)

UVic

NLPCA: Implementation

- Implemented NLPCA using neural networks (convenient, not necessary)
- Parameter estimation more difficult than for PCA

UVic

NLPCA: Implementation

- Implemented NLPCA using neural networks (convenient, not necessary)
- Parameter estimation more difficult than for PCA
- PCA model is linear in statistical parameters:

$$\mathbf{Y} = M\mathbf{X}$$

so variational problem has unique analytic solution

UVic

NLPCA: Implementation

- Implemented NLPCA using neural networks (convenient, not necessary)
- Parameter estimation more difficult than for PCA
- PCA model is linear in statistical parameters:

$$\mathbf{Y} = M\mathbf{X}$$

  so variational problem has unique analytic solution
- NLPCA model nonlinear in model parameters, so solution

UVic

NLPCA: Implementation

- Implemented NLPCA using neural networks (convenient, not necessary)
- Parameter estimation more difficult than for PCA
- PCA model is linear in statistical parameters:

$$\mathbf{Y} = M\mathbf{X}$$

  so variational problem has unique analytic solution
- NLPCA model nonlinear in model parameters, so solution
  - may not be unique

UVic

NLPCA: Implementation

- Implemented NLPCA using neural networks (convenient, not necessary)
- Parameter estimation more difficult than for PCA
- PCA model is linear in statistical parameters:

$$\mathbf{Y} = M\mathbf{X}$$

  so variational problem has unique analytic solution
- NLPCA model nonlinear in model parameters, so solution
  - may not be unique
  - must be found through numerical minimisation

UVic

NLPCA: Parameter Estimation

- Two fundamental issues regarding parameter estimation common to *all* statistical models:

NLPCA: Parameter Estimation

- Two fundamental issues regarding parameter estimation common to *all* statistical models:

  Reproducibility

NLPCA: Parameter Estimation

- Two fundamental issues regarding parameter estimation common to *all* statistical models:

  Reproducibility

  - model must be robust to the introduction of new data

UVic

NLPCA: Parameter Estimation

- Two fundamental issues regarding parameter estimation common to *all* statistical models:

  Reproducibility

  - model must be robust to the introduction of new data
  - new observations shouldn't fundamentally change model

NLPCA: Parameter Estimation

- Two fundamental issues regarding parameter estimation common to *all* statistical models:

  Reproducibility
  - model must be robust to the introduction of new data
  - new observations shouldn't fundamentally change model

  Classifiability:

NLPCA: Parameter Estimation

- Two fundamental issues regarding parameter estimation common to *all* statistical models:

  Reproducibility

  - model must be robust to the introduction of new data
  - new observations shouldn't fundamentally change model

  Classifiability:

  - model must be robust to details of optimisation procedure

UVic

NLPCA: Parameter Estimation

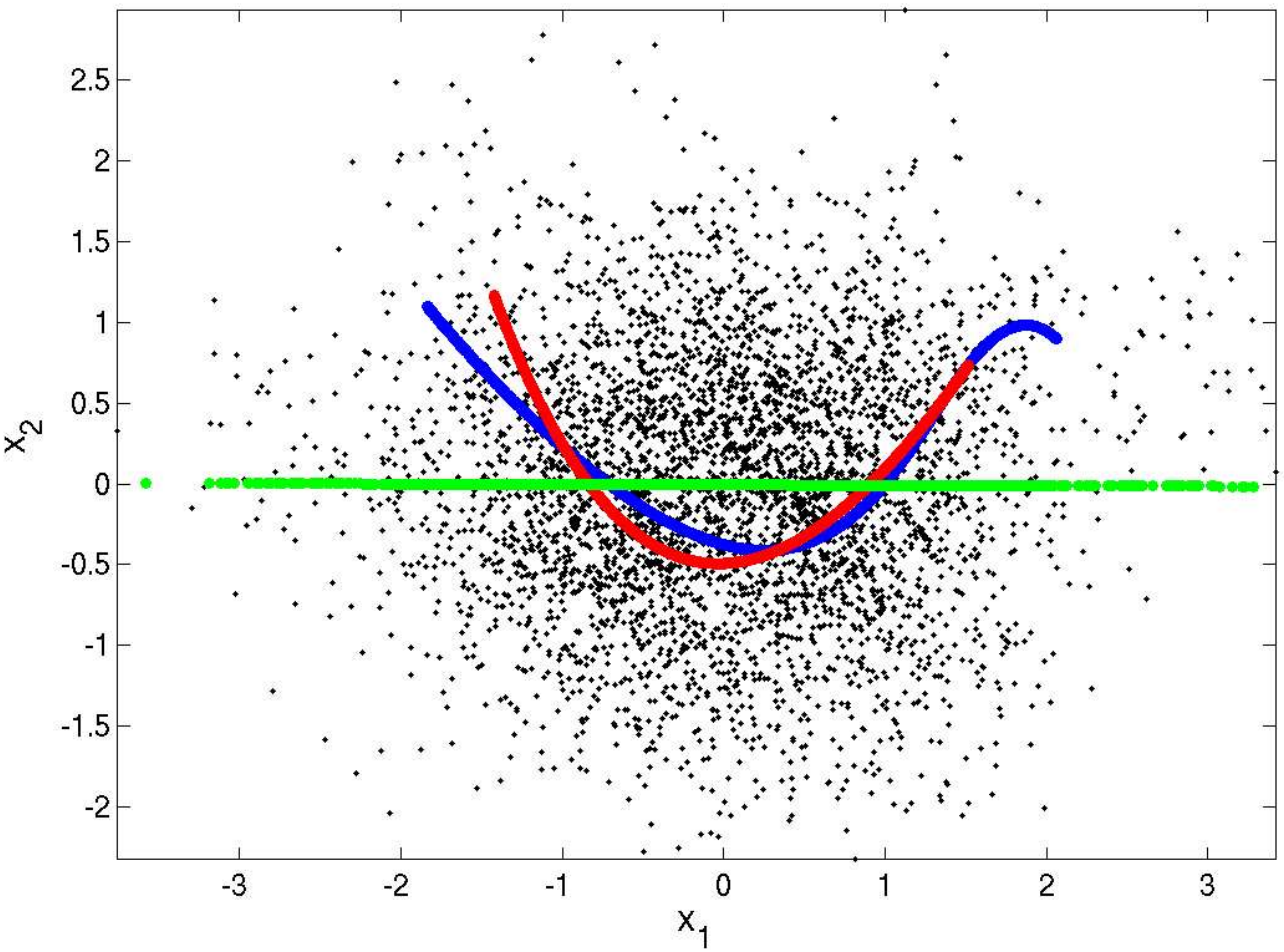
- Two fundamental issues regarding parameter estimation common to *all* statistical models:

  Reproducibility
  - model must be robust to the introduction of new data
  - new observations shouldn't fundamentally change model

  Classifiability:
  - model must be robust to details of optimisation procedure
  - model shouldn't depend on initial parameter values

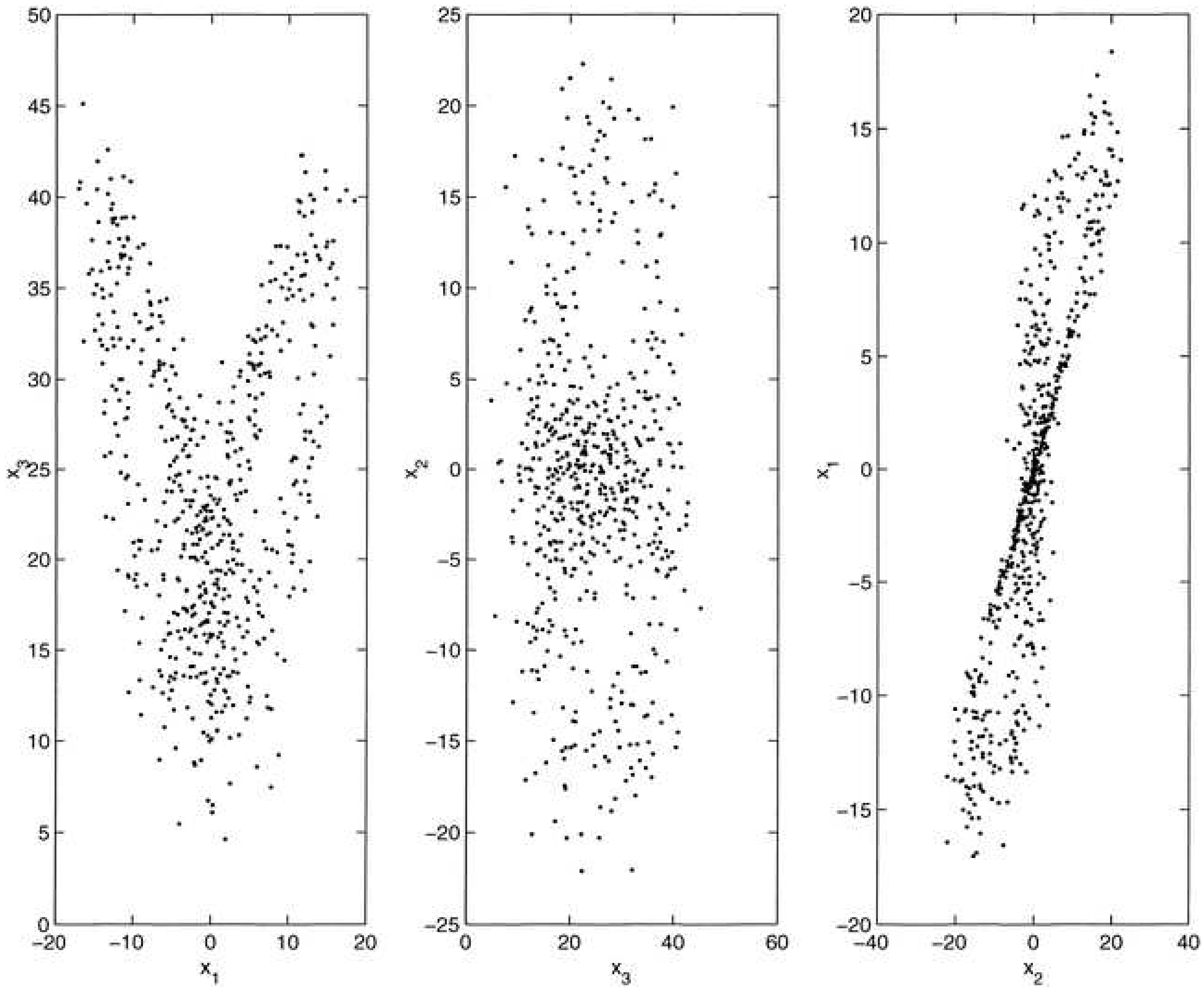
UVic

NLPCA: Synthetic Gaussian Data



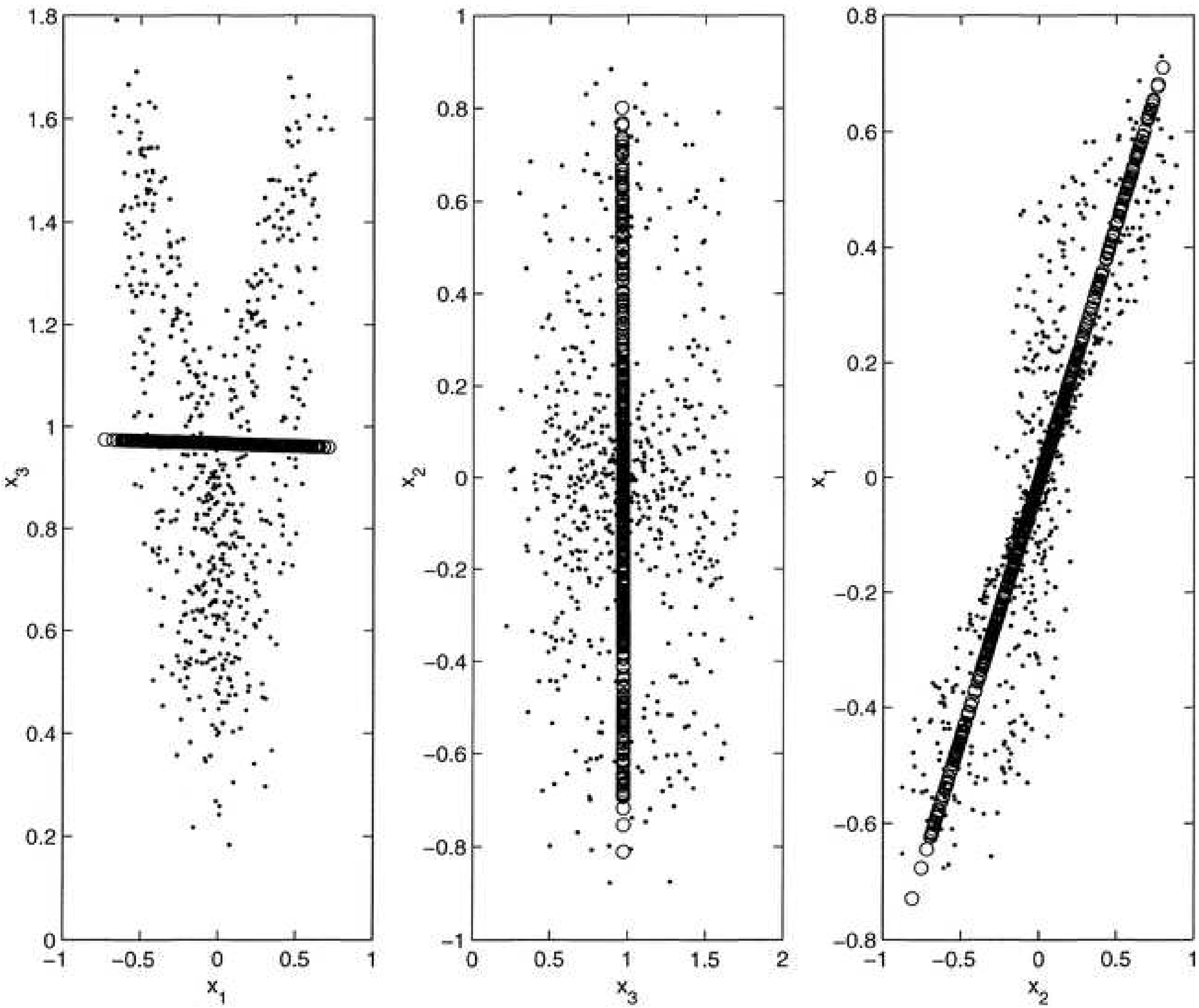
Synthetic Gaussian data

Applications of NLPCA: Lorenz Attractor



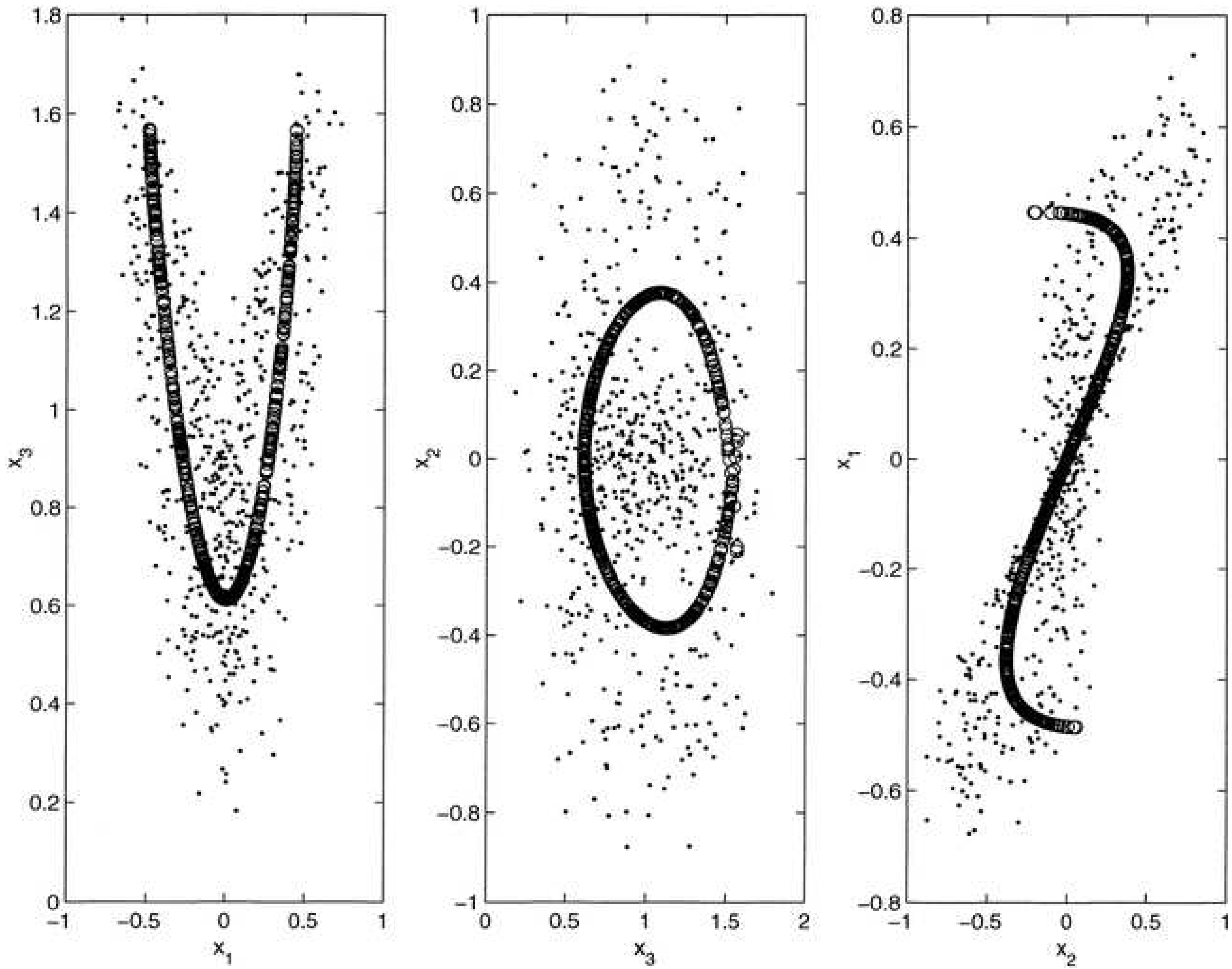
Scatterplots

Applications of NLPCA: Lorenz Attractor



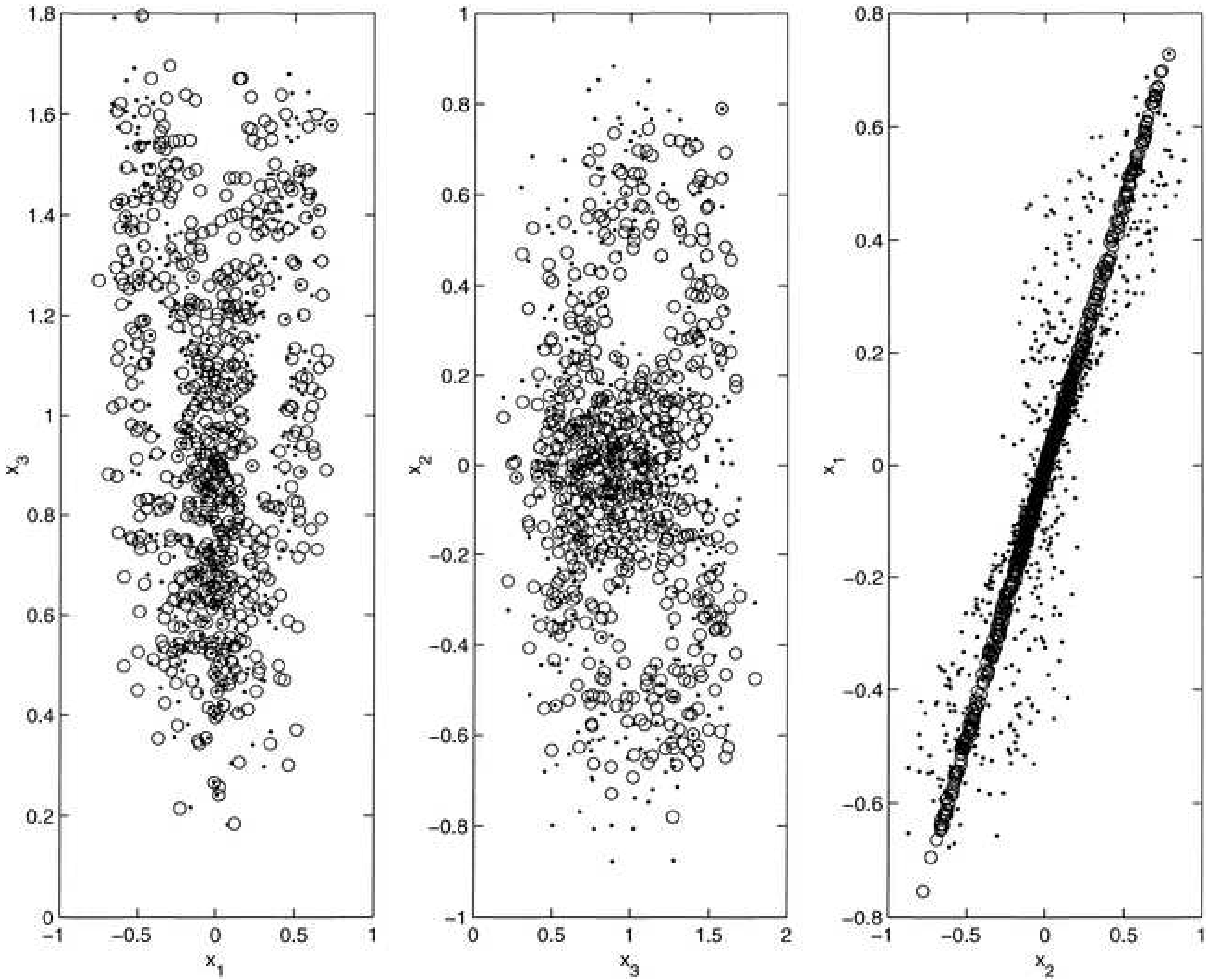
1D PCA approximation (60%)

Applications of NLPCA: Lorenz Attractor



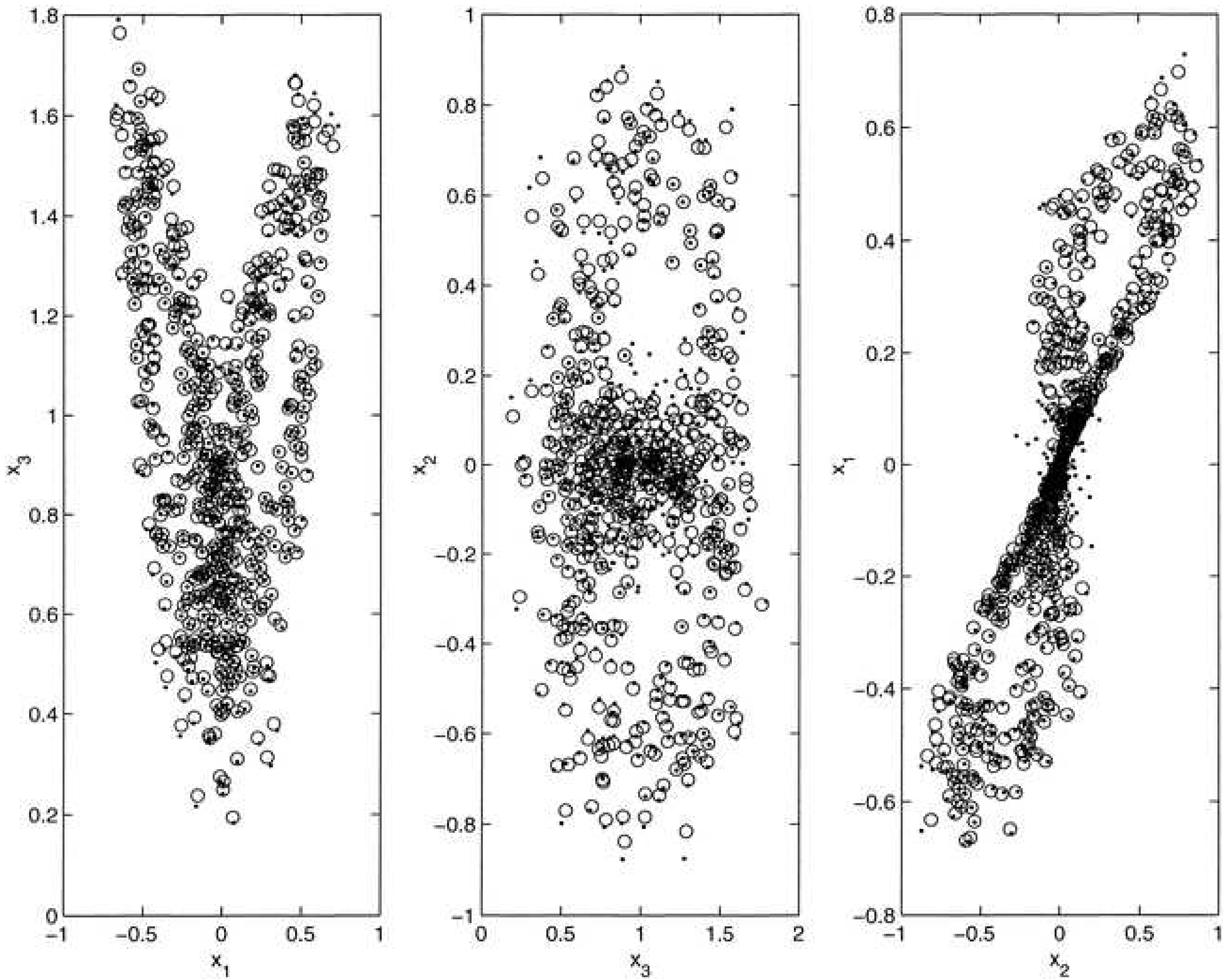
1D NLPCA approximation (76%)

Applications of NLPCA: Lorenz Attractor



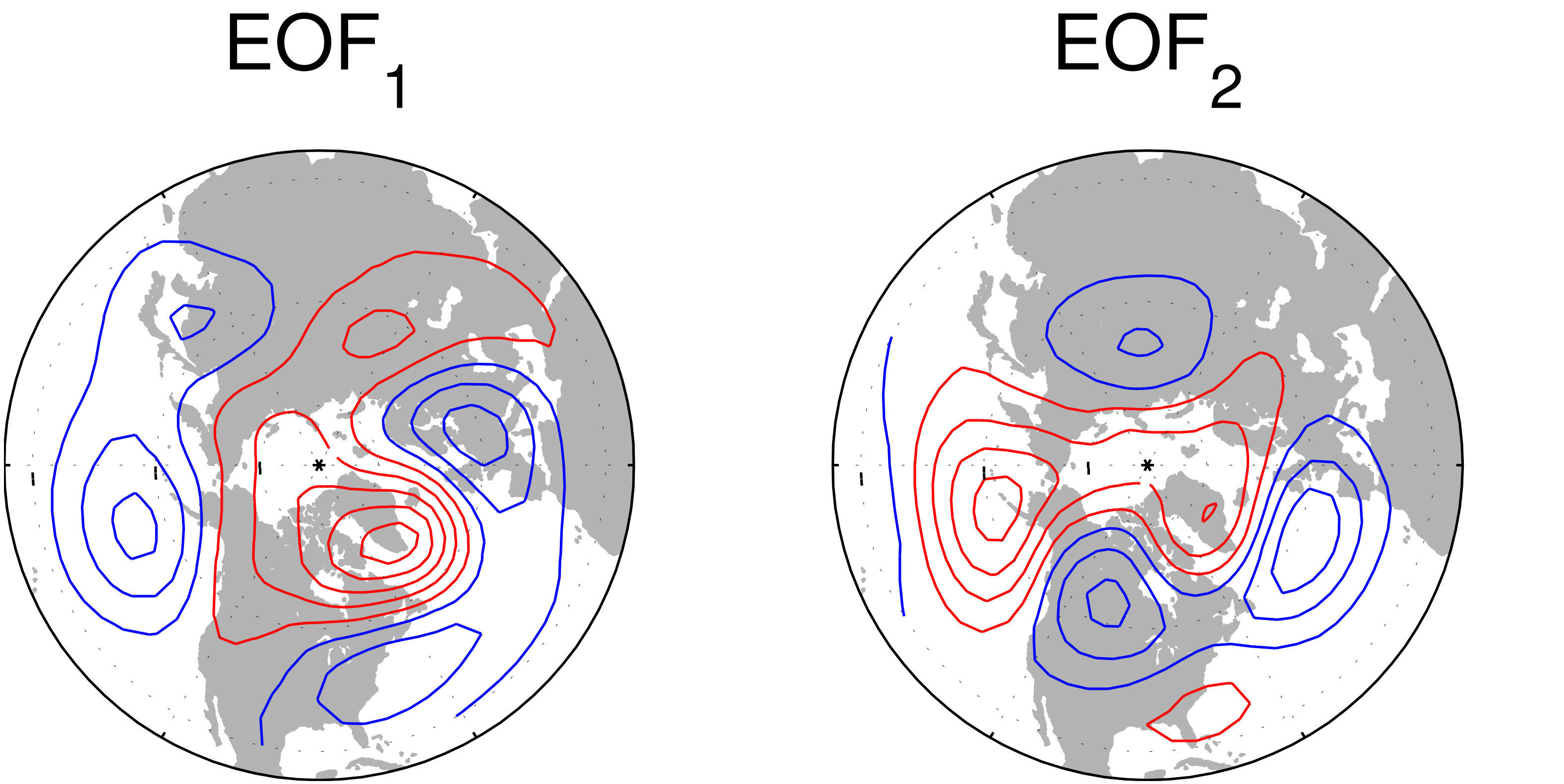
2D PCA approximation (94%)

Applications of NLPCA: Lorenz Attractor



2D NLPCA approximation (97%)

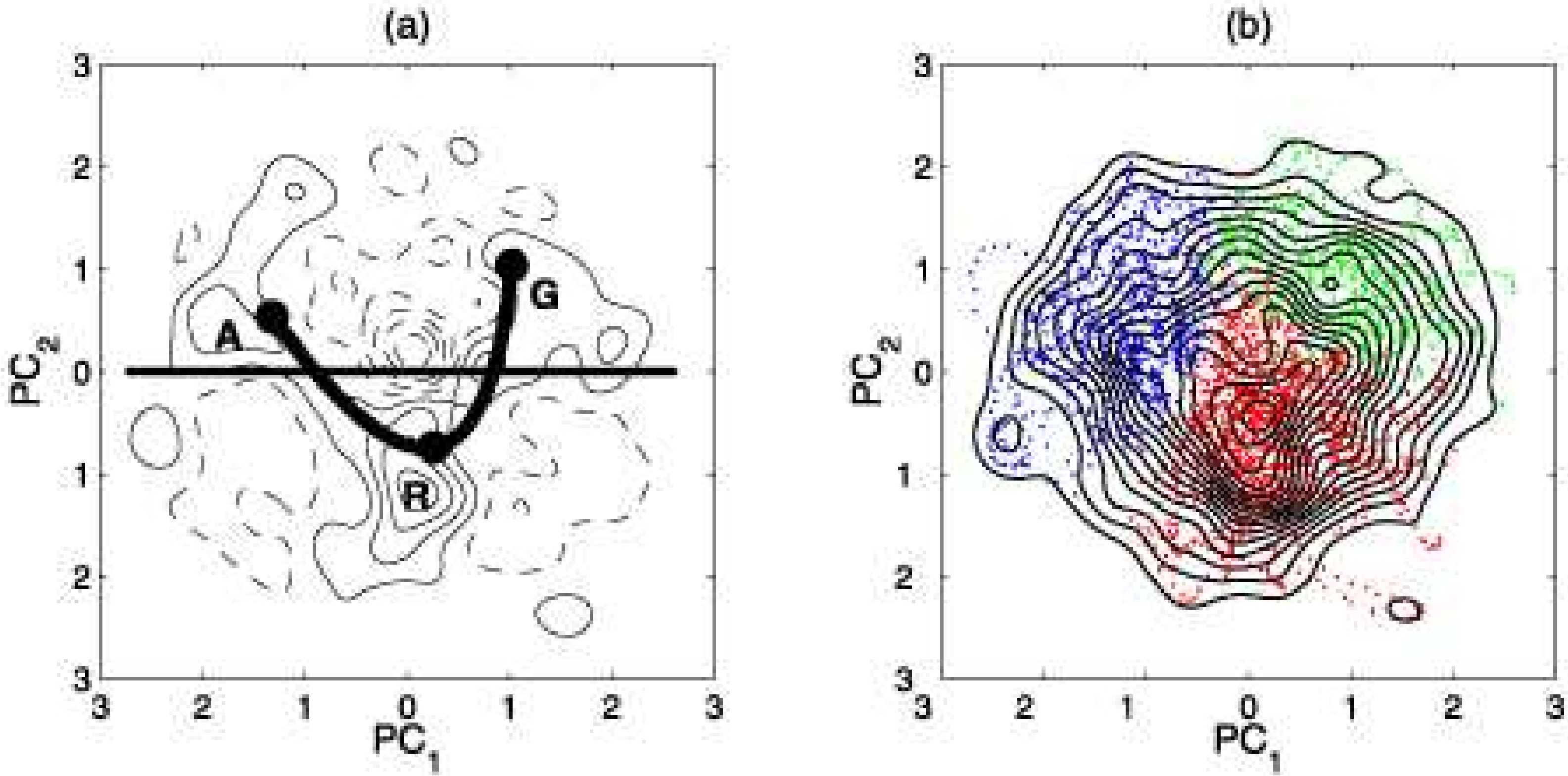
Applications of NLPCA: NH Tropospheric LFV



10-day lowpass-filtered 500 hPa geopotential height EOFs

UVic

Applications of NLPCA: NH Tropospheric LFV

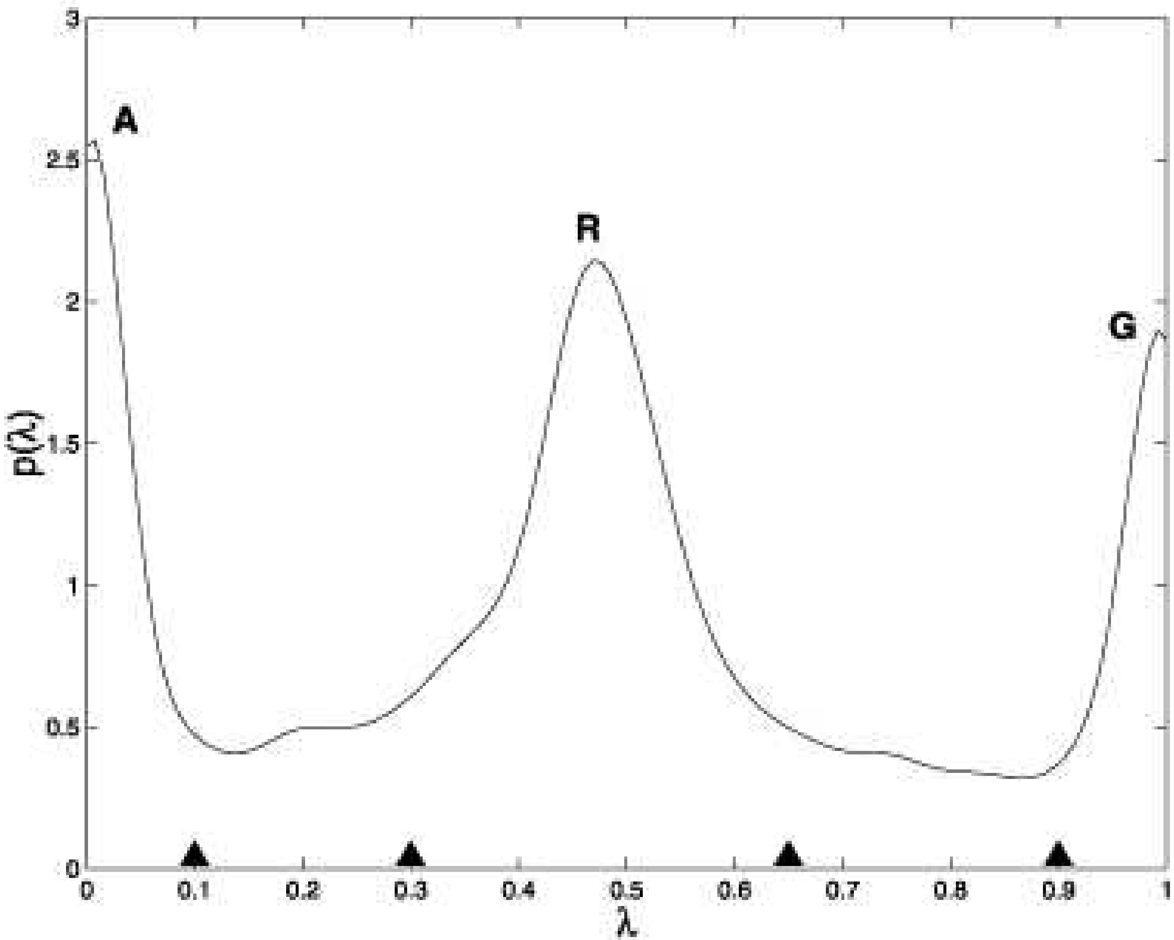


1D NLPCA Approximation: spatial structure

(PCA: 14.8%; NLPCA 18.4%)

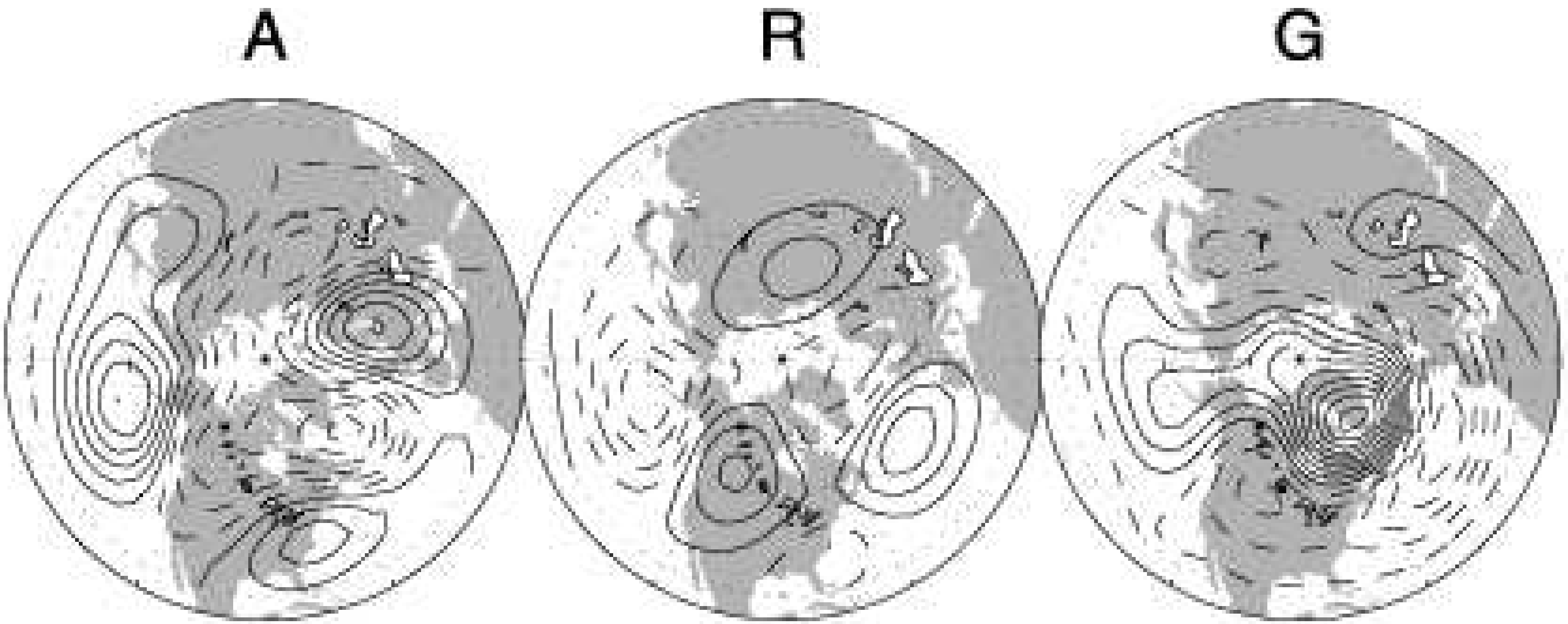
UVic

Applications of NLPCA: NH Tropospheric LFV



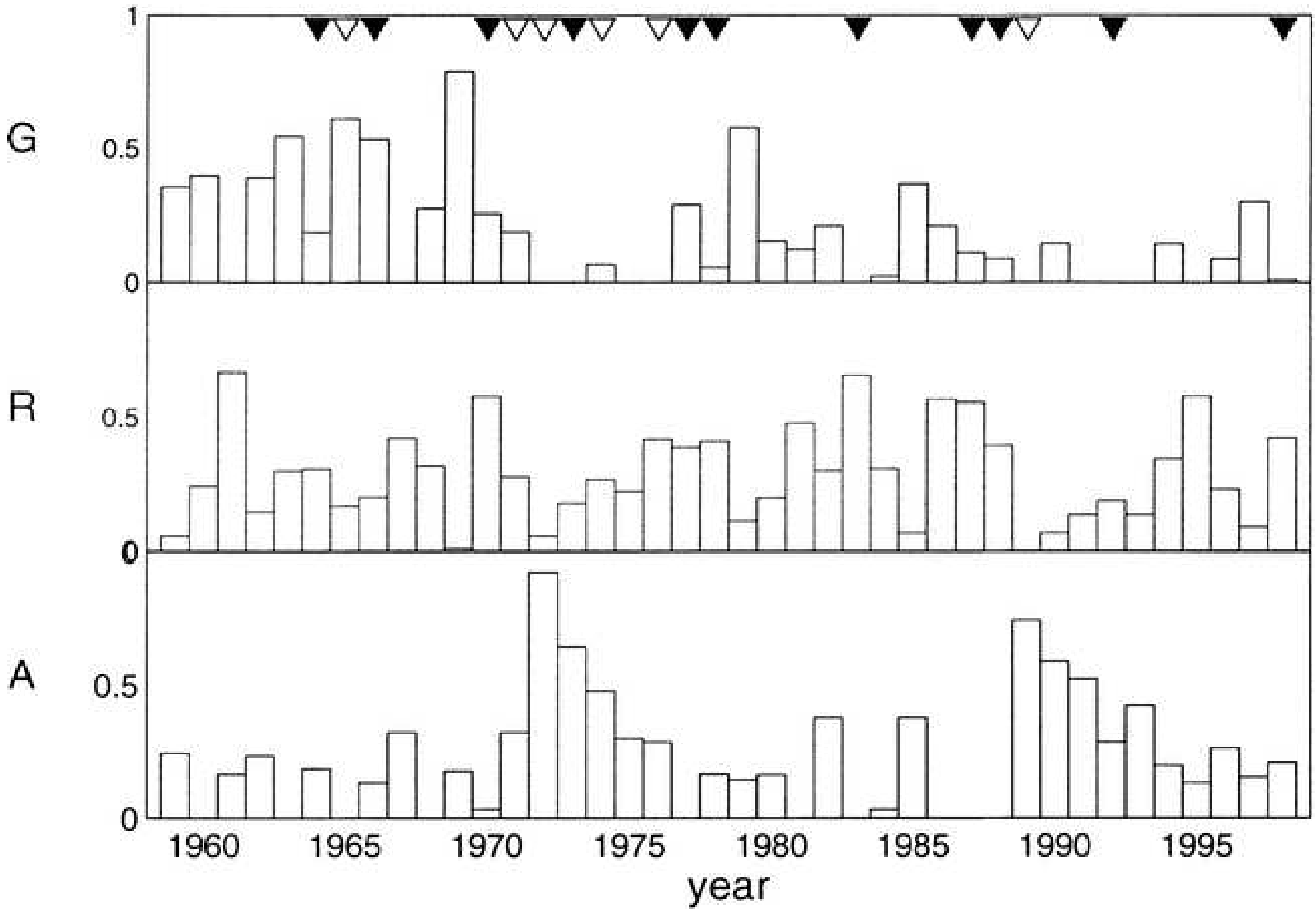
1D NLPCA Approximation: pdf of time series

Applications of NLPCA: NH Tropospheric LFV



1D NLPCA Approximation: regime maps

Applications of NLPCA: NH Tropospheric LFV



1D NLPCA Approximation: interannual variability

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

UVic

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

$\Rightarrow$ insufficiently careful analysis leads to spurious results (e.g. Christiansen, 2005)

UVic

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

$\Rightarrow$ insufficiently careful analysis leads to spurious results (e.g. Christiansen, 2005)

- Theoretical underpinning of NLPCA is weak

UVic

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

$\Rightarrow$ insufficiently careful analysis leads to spurious results (e.g. Christiansen, 2005)

- Theoretical underpinning of NLPCA is weak

$\Rightarrow$ no “rigorous” theory of sampling variability

UVic

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

$\Rightarrow$ insufficiently careful analysis leads to spurious results (e.g. Christiansen, 2005)

- Theoretical underpinning of NLPCA is weak

$\Rightarrow$ no “rigorous” theory of sampling variability

- Information theory may provide new tools with

UVic

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

$\Rightarrow$ insufficiently careful analysis leads to spurious results (e.g. Christiansen, 2005)

- Theoretical underpinning of NLPCA is weak

$\Rightarrow$ no “rigorous” theory of sampling variability

- Information theory may provide new tools with
  - better sampling properties

UVic

NLPCA: Limitations and Drawbacks

- Parameter estimation in NLPCA (as in any nonlinear statistical model) must be done very carefully to ensure robust approximation

$\Rightarrow$ analysis time-consuming, data hungry

$\Rightarrow$ insufficiently careful analysis leads to spurious results (e.g. Christiansen, 2005)

- Theoretical underpinning of NLPCA is weak

$\Rightarrow$ no "rigorous" theory of sampling variability

- Information theory may provide new tools with
  - better sampling properties
  - better theoretical basis

UVic

Conclusions

- Traditional PCA optimal for dimensionality reduction only if data distribution falls along orthogonal axes

UVic

Conclusions

- Traditional PCA optimal for dimensionality reduction only if data distribution falls along orthogonal axes
- Can define nonlinear generalisation, NLPCA, which can robustly characterise nonlinear low-dimensional structure in datasets

UVic

Conclusions

- Traditional PCA optimal for dimensionality reduction only if data distribution falls along orthogonal axes
- Can define nonlinear generalisation, NLPCA, which can robustly characterise nonlinear low-dimensional structure in datasets
- NLPCA approximations can provide a fundamentally different characterisation of data than PCA approximations

UVic

Conclusions

- Traditional PCA optimal for dimensionality reduction only if data distribution falls along orthogonal axes
- Can define nonlinear generalisation, NLPCA, which can robustly characterise nonlinear low-dimensional structure in datasets
- NLPCA approximations can provide a fundamentally different characterisation of data than PCA approximations
- Implementation of NLPCA difficult and lacking in underlying theory; represents a first attempt at a big (and challenging) problem

UVic

Acknowledgements

- William Hsieh (UBC)
- Lionel Pandolfo (UBC)
- John Fyfe (CCCma)
- Qiaobin Teng (CCCma)
- Benyang Tang (JPL)

UVic

Parameter Estimation in NLPCA

- An ensemble approach was taken

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:

UVic

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:
  - data was randomly split into *training* and *validation* sets (taking autocorrelation into account)

UVic

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:
  - data was randomly split into *training* and *validation* sets (taking autocorrelation into account)
  - a random initial parameter set was selected

UVic

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:
  - data was randomly split into *training* and *validation* sets (taking autocorrelation into account)
  - a random initial parameter set was selected
- For each ensemble member, iterative minimisation procedure carried out until either:

UVic

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:
  - data was randomly split into *training* and *validation* sets (taking autocorrelation into account)
  - a random initial parameter set was selected
- For each ensemble member, iterative minimisation procedure carried out until either:
  - error over training data stopped changing

UVic

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:
  - data was randomly split into *training* and *validation* sets (taking autocorrelation into account)
  - a random initial parameter set was selected
- For each ensemble member, iterative minimisation procedure carried out until either:
  - error over training data stopped changing
  - error over validation data started increasing

Parameter Estimation in NLPCA

- An ensemble approach was taken
- For a large number N (~ 50) of trials:
  - data was randomly split into *training* and *validation* sets (taking autocorrelation into account)
  - a random initial parameter set was selected
- For each ensemble member, iterative minimisation procedure carried out until either:
  - error over training data stopped changing
  - error over validation data started increasing
- Method **does not** look for global error minimum

UVic

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

- Candidate models compared

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

- Candidate models compared
  - if they share same shape and orientation

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

- Candidate models compared
  - if they share same shape and orientation

$\Rightarrow$ approximation is robust

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

- Candidate models compared
  - if they share same shape and orientation

$\Rightarrow$ approximation is robust

  - if they differ in shape and orientation

UVic

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

- Candidate models compared
  - if they share same shape and orientation

$\Rightarrow$ approximation is robust
  - if they differ in shape and orientation

$\Rightarrow$ approximation is not robust

UVic

Parameter Estimation in NLPCA

- Ensemble member becomes **candidate model** if

$$\langle ||\epsilon||^2 \rangle_{validation} \leq \langle ||\epsilon||^2 \rangle_{training}$$

- Candidate models compared
  - if they share same shape and orientation

$\Rightarrow$ approximation is robust

  - if they differ in shape and orientation

$\Rightarrow$ approximation is not robust

- If approximation not robust, model simplified & procedure repeated until robust model found

UVic

Parameter Estimation in NLPCA

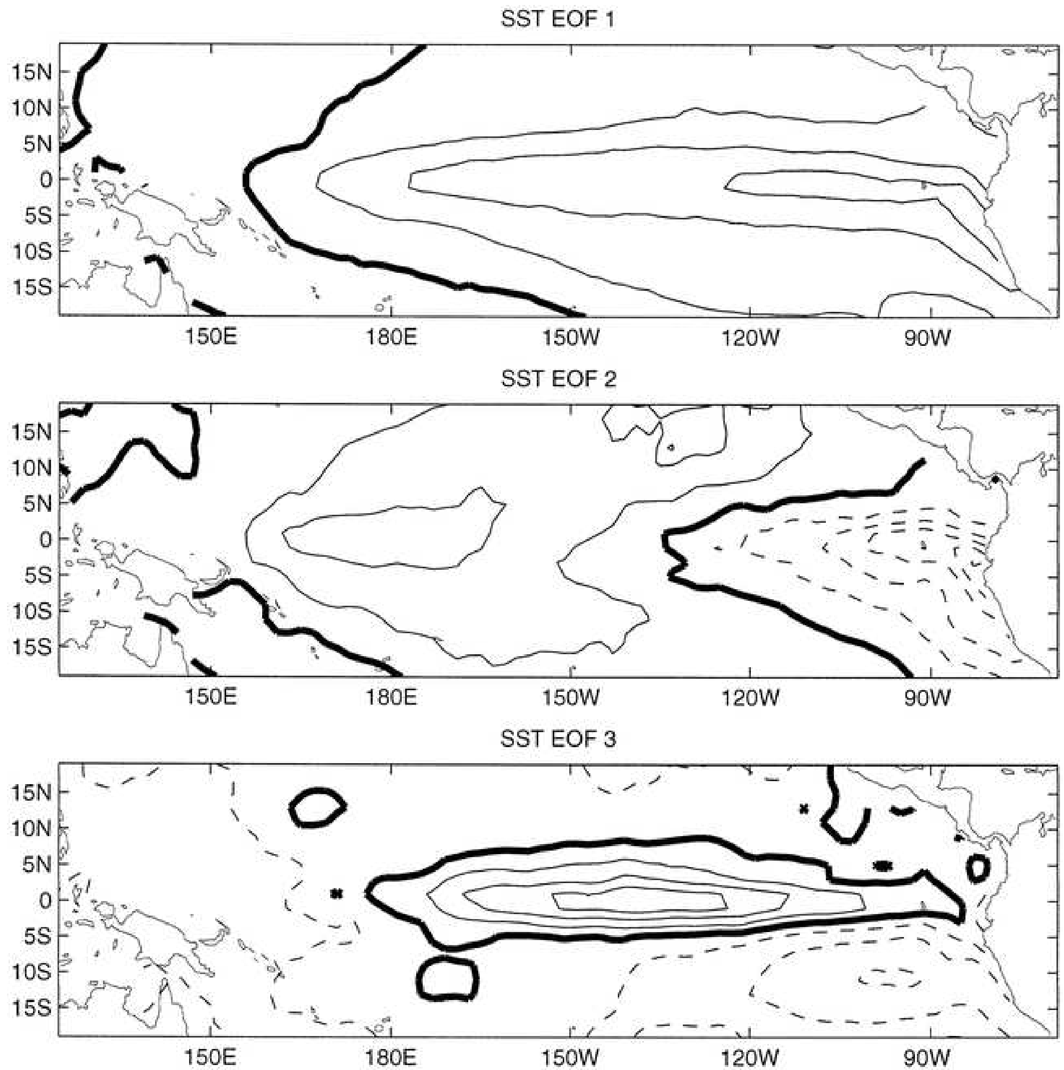
- Procedure will ultimately yield PCA solution if no robust non-Gaussian structure present

Parameter Estimation in NLPCA

- Procedure will ultimately yield PCA solution if no robust non-Gaussian structure present
- Such a careful procedure necessary to avoid finding spurious non-Gaussian structure

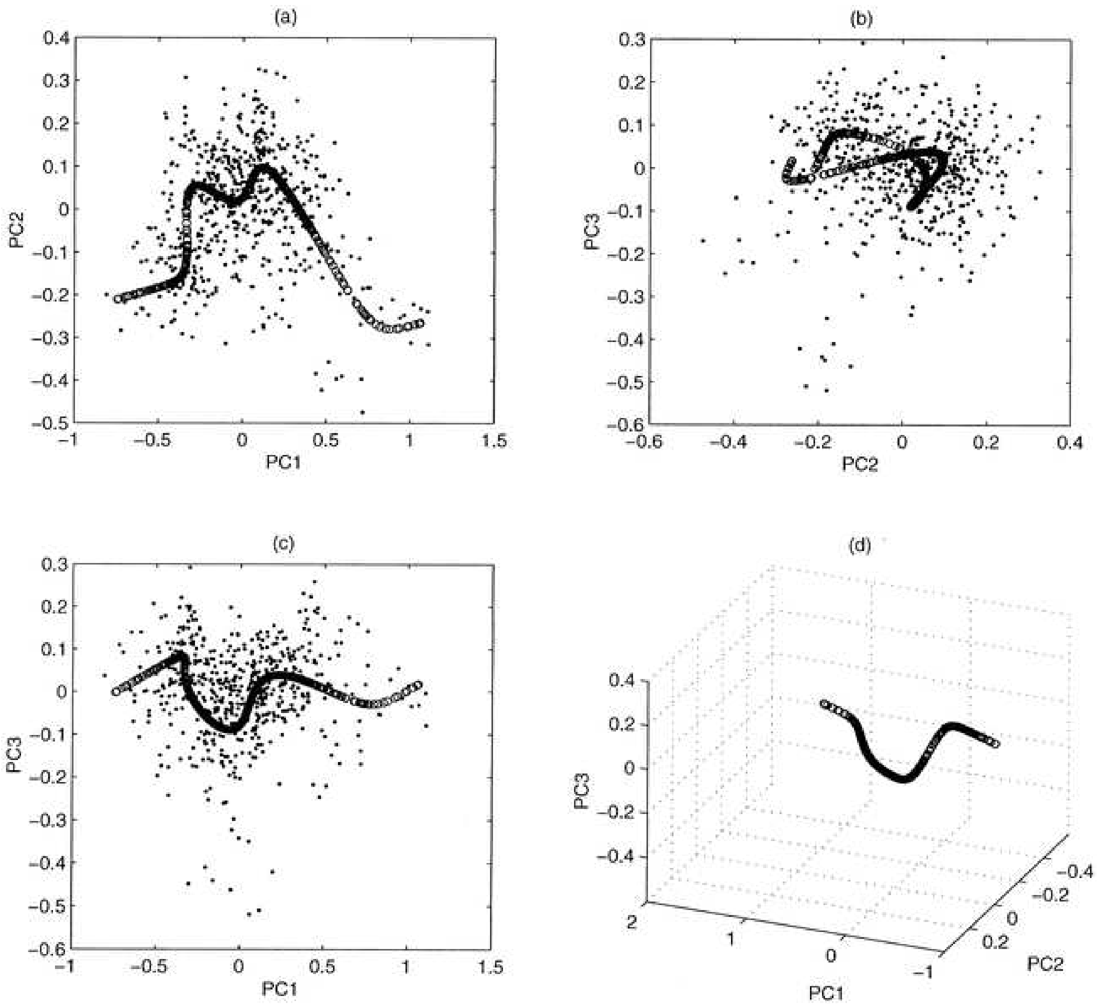
UVic

Applications of NLPCA: Tropical Pacific SST



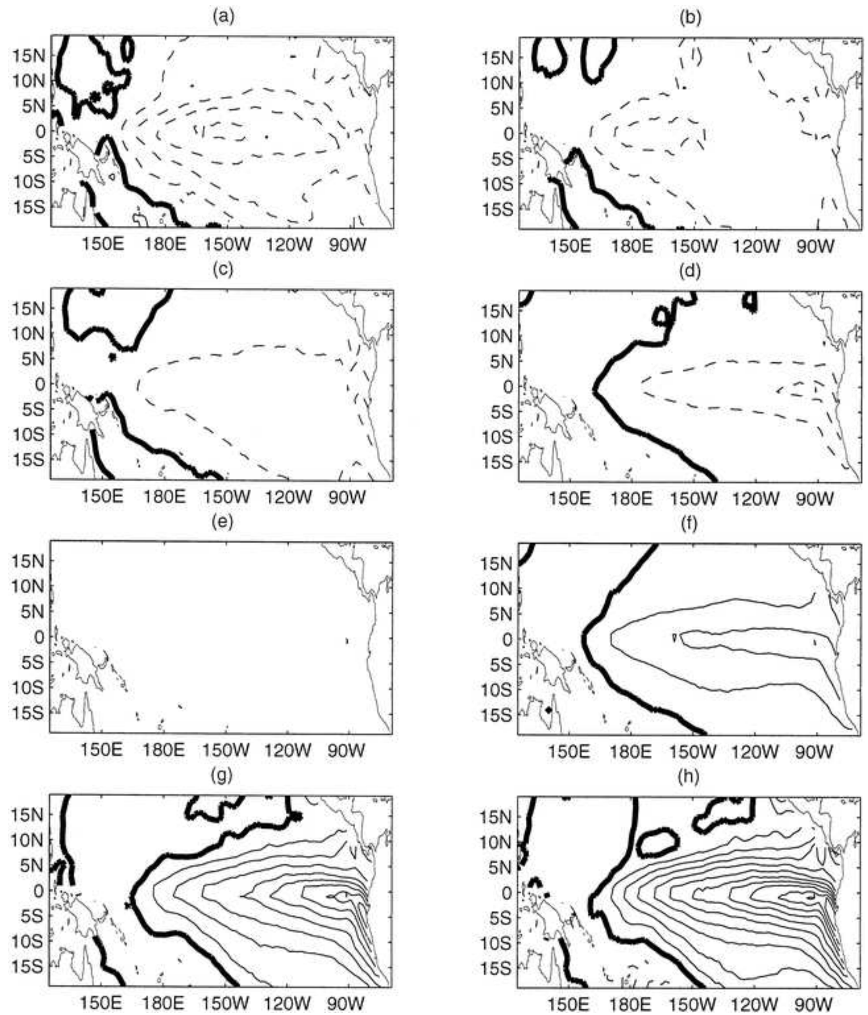
EOF Patterns

Applications of NLPCA: Tropical Pacific SST



1D NLPCA Approximation: spatial structure

Applications of NLPCA: Tropical Pacific SST



1D NLPCA Approximation: spatial structure