Optimization on manifolds and data processing

Rodolphe Sepulchre
Department of Electrical Engineering and Computer Science
University of Liège, Belgium

Collaborators: P.A. Absil (U Louvain and U Cambridge)
Robert Mahony (Australian National U)
Michel Journée (U Liège)
Andrew Teschendorff (U Cambridge)

Algorithms on manifolds

Principal manifolds: lines (or surfaces) passing through the middle of the data distribution.

Question: How to define and compute such things when the data are not points in $\mathbb{R}^n$ but points on abstract manifolds?

Motivation: SYMMETRY
In many problems, data represent geometric objects that are invariant under certain transformations.

A three-step approach

- An optimization-based formulation of the computational problem
- Generalization of optimization algorithms on abstract manifolds
- Exploit flexibility and additional structure to build numerically efficient algorithms

Optimization algorithms on matrix manifolds, book in preparation
P.-A. Absil, R. Mahony, R. Sepulchre.

Applications

- Eigenvalue problems
  (Invariant subspace calculation, PCA, SVD, . . .)
- Statistical problems
  (Matrix approximations, ICA, . . .)
- Pose estimation and motion recovery
- . . .

Outline

- Part I: a quick illustration of the three steps
- Part II: ICA and gene expression data analysis

Eigenvalue problems as optimization

Let A a $n \times n$ symmetric matrix.
Find an eigenvalue $\lambda \in \mathbb{R}$ and an eigenvector $y \in \mathbb{R}^n$ such that $Ay = \lambda y$

FACT: Eigenvectors are critical points of the Rayleigh quotient

$$f : \mathbb{R}^n_* \to \mathbb{R} : \quad f(y) = \frac{y^T A y}{y^T y}$$

The global minimum is the leftmost eigenvector.

Manifolds associated to eigenvectors

SYMMETRY: $f(\mu y) = f(y) \quad \forall \mu \in I\!R_*$
$\Rightarrow$ critical points are not isolated in $I\!R^n$.

REMEDY:
Impose a normalization constraint $\| y \| = 1$
$\Rightarrow$ Optimization on the sphere S^{n-1}
or
treat $yI\!R_*$ as one point in the projective space

$$P^{n-1} = \{yI\!R_* : y \in I\!R_*^n\}$$

$\Rightarrow$ Optimization on the projective space P^{n-1}

Generalized eigenvalue problems

- Let A, B $n \times n$ symmetric and B positive definite. Find (λ, y) such that $Ay = \lambda By$
- The cost function is now defined over the full rank $n \times p$ matrices:

$$f(Y) = \mathrm{trace}(\mathrm{Y^T AY (Y^T BY)^{-1}})$$

- Y_* is a global minimizer of f iff Y_* span the leftmost p-dimensional invariant subspace of $B^{-1}A$.

Manifolds for invariant subspaces

SYMMETRY: $f(YM) = f(Y)$ for all full rank $p \times p$ matrix M
$\Rightarrow$ critical points are not isolated in $\mathbb{R}^{n \times p}$.

REMEDY:
Impose a normalization constraint $\| Y^T Y \| = I_p$
$\Rightarrow$ Optimization on the Stiefel manifold $St(p, n)$
or
treat $yGL(p)$ as one point in the Grassmann manifold
$Gr(p, n)$ of p-dimensional subspaces of $\mathbb{R}^n$.

Important matrix manifolds

- $S^{n-1}, St(p,n)$ are examples of *embedded* manifolds in vector spaces.
- $P^{n-1}, Gr(p,n)$ are examples of *quotient* manifolds in vector spaces

The linear structure of the *total* vector space is very helpful for computations!

A three-step approach

- An optimization-based formulation of the computational problem
- Generalization of optimization algorithms on abstract manifolds
- Exploit flexibility and additional structure to build numerically efficient algorithms

How different is an algorithm in a vector space and on a manifold?
Illustration: line-search algorithm

Line search in a vector space

$$x_{k+1} = x_k + t_k \eta_k$$

The vector η_k is a search direction
The scalar t_k dictates the step length
$\approx$ discretized version of the continuous-time descent gradient flow

$$\dot{x} = -\mathrm{grad} f(x)$$

Line search on a manifold

Let M an abstract Riemannian manifold.

$$x_{k+1} = \mathsf{Exp}_{x_k}(t_k \xi) = \gamma(t_k : x_k, \xi)$$

Start at x_k; choose a direction ξ in the tangent space $T_x M$; follow for t_k units the geodesic passing at x_k and tangent to ξ.
(Luenberger, 73; Gabay, 82).
Conceptually elegant and useful; numerically unpractical.

Optimization on manifolds

- Newton method (Smith 93, Mahony 94)
- Conjugated gradients (Edelman 96)
- Trust region method (Absil et al. 04)
- . . .

Translation of corresponding algorithms in vector spaces + *convergence* theory.

A three-step approach

- An optimization-based formulation of the computational problem
- Generalization of optimization algorithms on abstract manifolds
- Exploit flexibility and additional structure to build numerically efficient algorithms

Does this approach lead to competitive numerical algorithms?
Illustration: line-search algorithm

Retractions

$$x_{k+1} = R_{x_k}(t_k \xi)$$

The convergence theory of line search methods still holds if the exponential mapping is replaced by ANY mapping $R : TM \to M$ satisfying $R_x(0_x) = x$ and $DR_x(0_x) = idT_xM$.

Examples of retractions

Use the linear structure of the total space:

On S^{n-1}: $R_x(\xi) = \frac{x+\xi}{\|x+\xi\|}$

On $Gr(p,n)$: $R_{\text{span}Y}(\xi) = \text{span}(Y + \overline{\xi}_Y)$ with $\overline{\xi}_Y$ the horizontal lift of ξ

Good retractions may turn the algorithm into a numerically efficient procedure.

State of the art

Brute force trust-region algorithms applied to the Rayleigh quotient cost on $Gr(p, n)$ (Absil et al, 04) compete with the best available numerical algorithms for large-scale problems.

Some benefits of the approach

- A solid framework for convergence analysis;
- A geometric interpretation of existing heuristics;
- Sometimes, new and competitive algorithms

More in
Optimization algorithms on matrix manifolds,
Princeton University Press, 2007.
P.-A. Absil, R. Mahony, R. Sepulchre.

Extracting Independent Components of Gene Expression Data

Michel Journée, Rodolphe Sepulchre, Pierre-Antoine Absil

Department of Electrical Engineering and Computer Science
University of Liège, Belgium

Workshop on Principal Manifolds, Leicester, August 2006

Independent Component Analysis

- Blind source separation based on the statistical independence of the sources.
- It assumes a linear, instantaneous and noisy mixture of sources,

$$\mathbf{x} = H\mathbf{s} + \mathbf{v}, \quad H \in \mathbb{R}^{n \times p}.$$

➠ Given the observations $\mathbf{x}$, identify the mixing matrix $\mathbf{H}$ and the independent sources $\mathbf{s}$.

Outline

- ICA algorithms are optimization algorithms on manifolds.
- The application of ICA to gene expression data raises central issues. (Cost function, manifold, optimization algorithm?)

The basic ICA algorithm

1. Let assume a linear demixing model: $\mathbf{z} = W^T \mathbf{x}, \quad W \in \mathbb{R}^{n \times p}$.

2. Measure the statistical independence of the estimated sources z_i ($\Rightarrow$ contrast).

3. Select the W^* that maximizes that measure.

➠ Two main features: the contrast and the optimization algorithm.

The contrast

- Definition:

  A function $\gamma(\cdot) : W \in \mathcal{M} \rightarrow \gamma(W) \in \mathbb{R}$ that measures the statistical independence of the z_i.

- Different types of contrast:

  - Based on the mutual information (MI is zero at the independence and otherwise always positive).
  - Diagonalization of the rth-order cumulant tensor (usually r=4).
  - Joint approximate diagonalization of a set of matrices (SOBI, JADE, etc.).
  - The constrained covariance: $\sup_{f,g} \operatorname{cov}(f(z_1), g(z_2))$.
  - ...

The optimization algorithm

- Optimization on a matrix manifold: $W^* = \underset{W \in \mathcal{M}}{\operatorname{argmax}}\, \gamma(W)$.

- Which manifold $\mathcal{M}$?

  Inherent symmetries of ICA:

  - Continuous symmetry: $W \sim W\Lambda$, with Λ an invertible diagonal matrix.
  - Discrete symmetry: $W \sim WP$, with P a permutation matrix.

Choice of a manifold

- Optimization on the orthogonal group:

$$\mathcal{O}_p = \{Y \in \mathbb{R}^{p \times p} : Y^T Y = I_p\}.$$

  - Jacobi algorithms (JADE, SOBI, RADICAL), KernelICA.

- Optimization on the orthogonal Stiefel manifold:

$$\mathrm{St}(n, p) = \{Y \in \mathbb{R}^{n \times p} : Y^T Y = I_p\}.$$

  - FastICA (one-unit algorithm used in a deflation scheme).

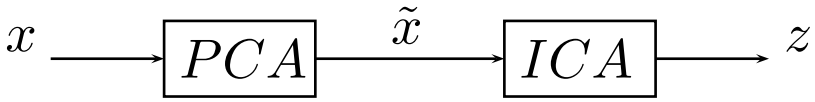
- Optimization on the oblique manifold [P.-A. Absil and K.A. Gallivan, 2006]:

$$\mathcal{OB}(n, p) = \{Y \in \mathbb{R}^{n \times p} : \mathrm{diag}(Y^T Y) = I_p\}.$$

  - Trust region optimization.

Prewhitening in ICA

- ICA is usually used in conjunction with PCA.



- Motivations for prewhitening:
  - Good-conditioning of the ICA problem.
  - Reduction of the dimensions of the ICA problem.
  - Restriction of the ICA optimization to the orthogonal Stiefel manifold (prewhitening-based algorithms).

Discussion about prewhitening

- The prewhitening step is biased in the presence of noise and outliers.

  Optimization on orthogonal manifolds is not able to compensate for these errors.

  Optimization on non-orthogonal manifolds is more accurate.

- Optimization algorithms on orthogonal manifolds are usually better conditioned.

  Optimization on non-orthogonal manifolds might be less robust.

- The compromise between performance and robustness is rarely discussed in the literature, especially for high-dimensional problems.

Outline

- ICA algorithms are optimization algorithms on manifolds.
- The application of ICA to gene expression data raises central issues. (Cost function, manifold, optimization algorithm?)

What are gene expression data?

- Gene expression denotes the relevance of a specific gene on the biological functions to be fulfilled in the cell.
- DNA microarrays are intensively used in biochemistry and biomedicine to estimate the gene expression levels.
- They provide a huge amount of data (typically, $\sim$10.000 genes and $\sim$100 experiments).

➠ Dimensionality reduction methods are needed for the analysis of these data.

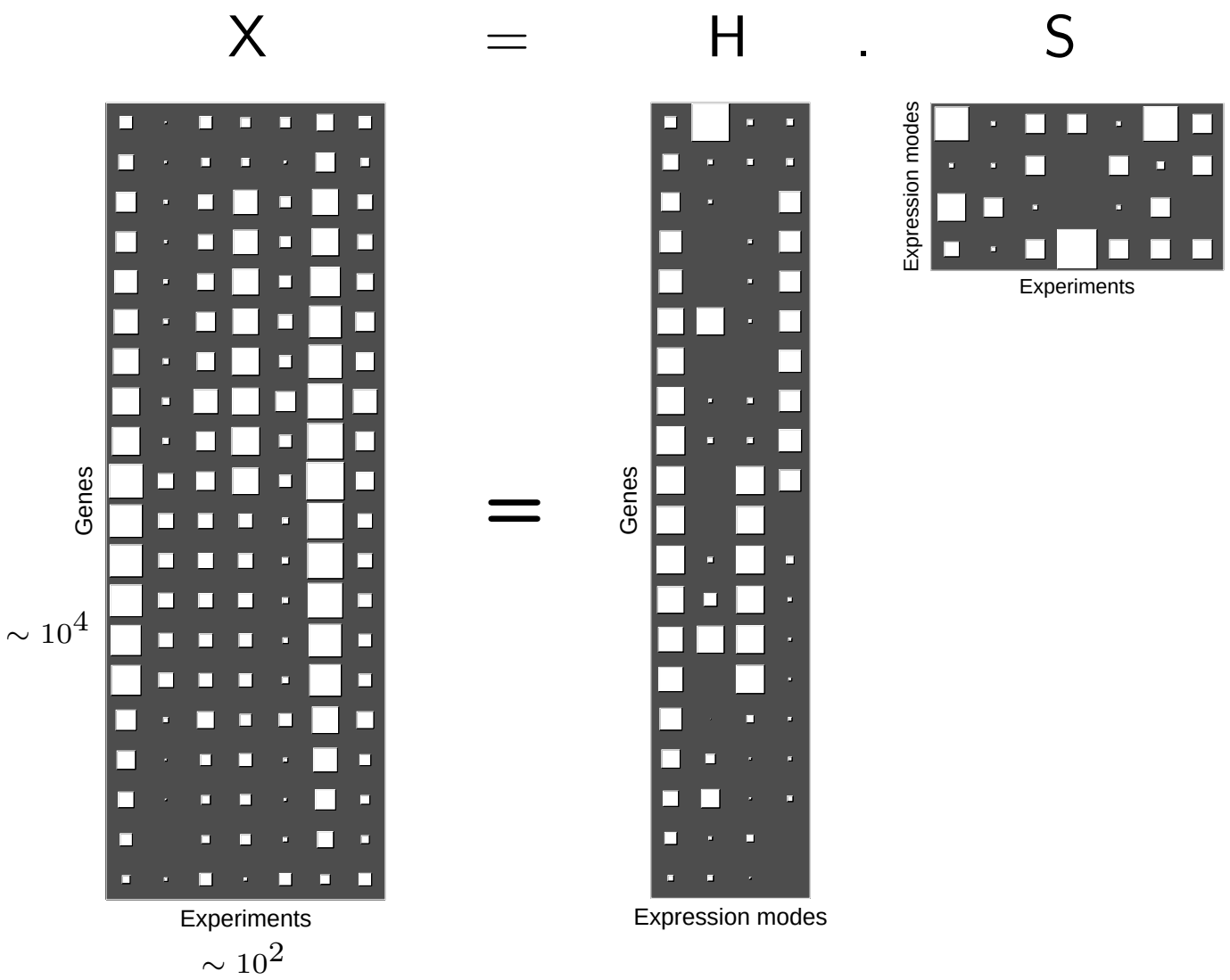
Dimensionality reduction by ICA: Motivation

- Each biological function relies on a subset of genes (expression mode).
- Gene expression levels result from several biological processes that take place **independently**.
- Gene expression is assumed to be a **linear** function of the expression modes.

➠ Independence and linearity are the basic requisites for ICA[1].

[1] First application of ICA to microarrays:
W. Liebermeister, *Linear modes of gene expression determined by independent component analysis*, Bioinformatics **18** (2002), 51–60.

ICA for the analysis of gene expression data



Preliminary results

- Application of standard ICA algorithms to breast cancer databases[2].

- Performance:
  ICA seems to outperform PCA in relating expression modes to biological pathways (i.e., groups of genes that participate together when a certain biological function is required).

[2] In collaboration with A.E. Teschendorff, Department of Oncology, University of Cambridge

Challenges

Standard ICA algorithms are not well adapted for gene expression data.
(i.e., few experiments, many observations, lot of outliers and noise.)

- New algorithmic developments are needed, i.e, cost functions, manifolds and optimization algorithms specially dedicated to this kind of data sets.

Conclusion

- ICA performs dimensionality reduction by assuming that the observations arise from several independent sources.
- ICA algorithms are optimization-based algorithms on manifolds.
- ICA seems promising for the analysis of microarrays but raises central robustness and performance issues.