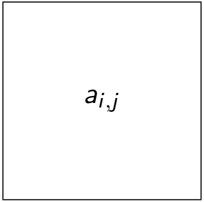
Large Scale Matrix Analysis and Inference

Wouter M. Koolen - **Manfred Warmuth**
Reza Bosagh Zadeh - Gunnar Carlsson - Michael Mahoney

Dec 9, NIPS 2013

Introductory musing — What is a matrix?



1. A vector of n^2 parameters
2. A covariance
3. A generalized probability distribution
4. ...

1. A vector of n^2 parameters

When you regularize with the squared Frobenius norm

$$\min_{\mathbf{W}} \quad ||\mathbf{W}||_F^2 + \sum_n \text{loss}(\mathrm{tr}(\mathbf{W}\mathbf{X}_n))$$

1. A vector of n^2 parameters

When you regularize with the squared Frobenius norm

$$\min_{\mathbf{W}} \quad ||\mathbf{W}||_F^2 + \sum_n \text{loss}(\mathrm{tr}(\mathbf{W}\mathbf{X}_n))$$

Equivalent to

$$\min_{\text{vec}(\mathbf{W})} \quad ||\text{vec}(\mathbf{W})||_2^2 + \sum_n \text{loss}(\text{vec}(\mathbf{W}) \cdot \text{vec}(\mathbf{X}_n))$$

No structure: n^2 independent variables

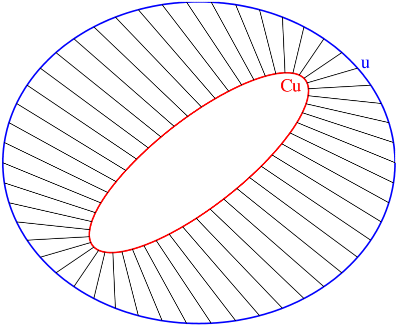
2. A covariance

View the symmetric positive definite matrix $\mathbf{C}$ as a covariance matrix of some random feature vector $\mathbf{c} \in \mathbb{R}^n$, i.e.

$$\mathbf{C} = \mathbb{E}\left((\mathbf{c} - \mathbb{E}(\mathbf{c}))(\mathbf{c} - \mathbb{E}(\mathbf{c}))^\top\right)$$

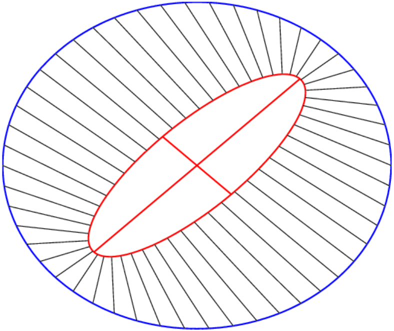
***n* features plus their pairwise interactions**

Symmetric matrices as ellipses



- Ellipse $= \{\mathbf{Cu} : \|\mathbf{u}\|_2 = 1\}$
- Dotted lines connect point $\mathbf{u}$ on unit ball with point $\mathbf{Cu}$ on ellipse

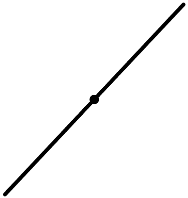
Symmetric matrices as ellipses



- Eigenvectors form axes
- Eigenvalues are lengths

Dyads

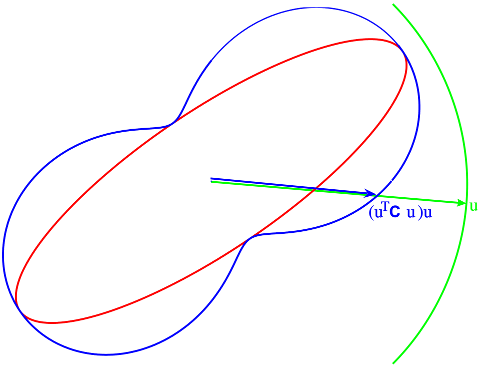
$\mathbf{u}\mathbf{u}^\top$, where $\mathbf{u}$ unit vector



- One eigenvalue one
- All others zero
- Rank one projection matrix

Directional variance along direction $\mathbf{u}$

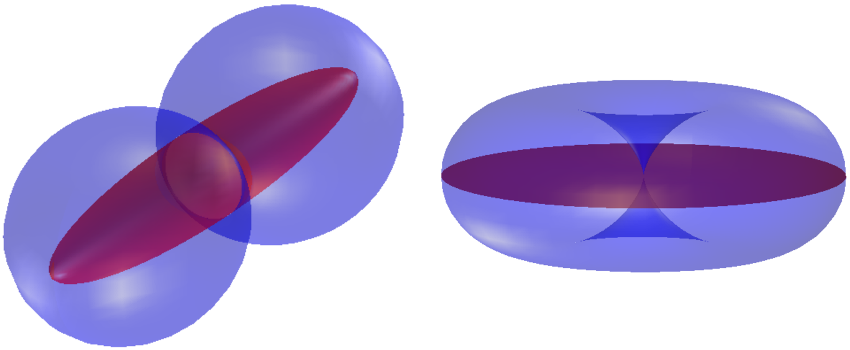
$$\mathbb{V}(\mathbf{c}^\top \mathbf{u}) = \mathbf{u}^\top \mathbf{C} \mathbf{u} = \mathrm{tr}(\mathbf{C}\, \mathbf{u}\mathbf{u}^\top) \;\geq\; 0$$



The outer figure eight is direction $\mathbf{u}$ times the variance $\mathbf{u}^\top \mathbf{C}\, \mathbf{u}$

PCA: find direction of largest variance

3 dimensional variance plots



$\mathrm{tr}(\mathbf{C}\,\mathbf{u}\mathbf{u}^\top)$ is generalized probability when $\mathrm{tr}(\mathbf{C}) = 1$

3. Generalized probability distributions

Probability vector

$$\boldsymbol{\omega} = (.2, .1., .6, .1)^\top = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{e}_i}_{\text{pure events}}$$

Density matrix

$$\mathbf{W} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{w}_i \mathbf{w}_i^\top}_{\text{pure density matrices}}$$

3. Generalized probability distributions

Probability vector

$$\boldsymbol{\omega} = (.2, .1., .6, .1)^\top = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{e}_i}_{\text{pure events}}$$

Density matrix

$$\mathbf{W} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{w}_i \mathbf{w}_i^\top}_{\text{pure density matrices}}$$

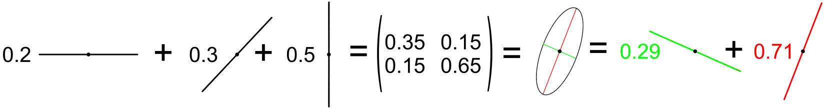
Matrices as generalized distributions

Probability vector
$$\boldsymbol{\omega} = (.2, .1., .6, .1)^\top = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{e}_i}_{\text{pure events}}$$

Density matrix
$$\mathbf{W} = \sum_i \underbrace{\omega_i}_{\text{mixture coefficients}} \underbrace{\mathbf{w}_i \mathbf{w}_i^\top}_{\text{pure density matrices}}$$

Matrices as generalized distributions

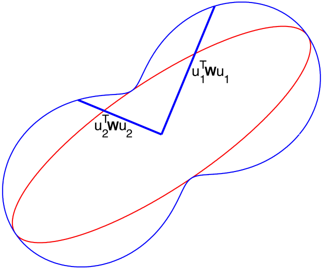
- Many mixtures lead to same density matrix



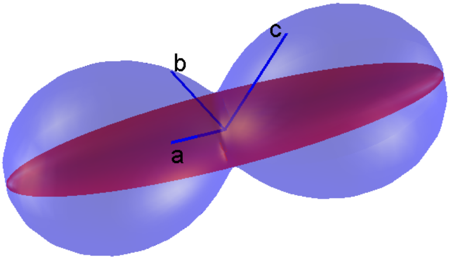
- There always exists a decomposition into n *eigendyads*
- Density matrix: Symmetric positive matrix of trace one

It's like a probability!

Total variance along orthogonal set of directions is 1

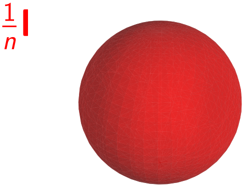


$$\mathbf{u}_1^\top \mathbf{W} \mathbf{u}_1 + \mathbf{u}_2^\top \mathbf{W} \mathbf{u}_2 = 1$$



$$a + b + c = 1$$

Uniform density?



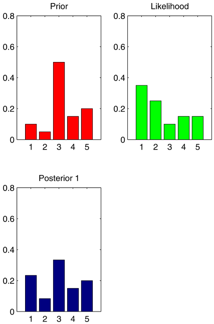
- All dyads have generalized probability $\frac{1}{n}$

$$\mathrm{tr}(\frac{1}{n}\mathbf{I}\ \mathbf{u}\mathbf{u}^\top) = \frac{1}{n}\,\mathrm{tr}(\mathbf{u}\mathbf{u}^\top) = \frac{1}{n}$$

- Generalized probabilities of n orthogonal dyads sum to 1

Conventional Bayes Rule

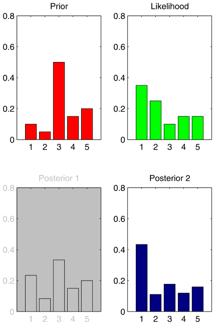
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



- 4 updates with the same data likelihood
- Update maintains uncertainty information about maximum likelihood
- **Soft max**

Conventional Bayes Rule

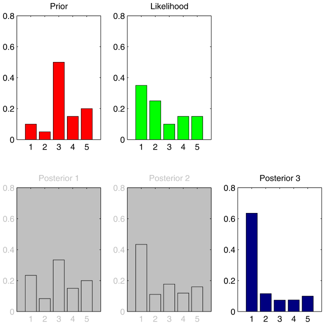
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



- 4 updates with the same data likelihood
- Update maintains uncertainty information about maximum likelihood
- **Soft max**

Conventional Bayes Rule

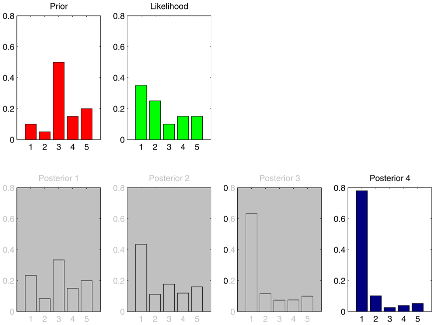
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



- 4 updates with the same data likelihood
- Update maintains uncertainty information about maximum likelihood
- **Soft max**

Conventional Bayes Rule

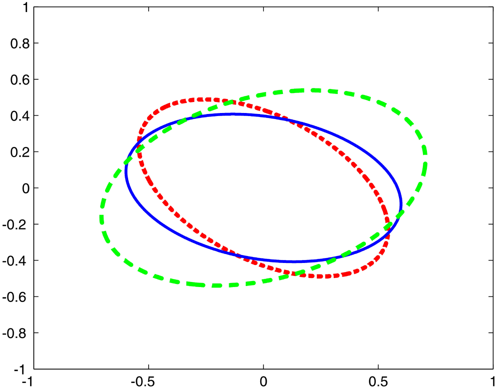
$$P(M_i|y) = \frac{P(M_i)P(y|M_i)}{P(y)}$$



- 4 updates with the same data likelihood
- Update maintains uncertainty information about maximum likelihood
- **Soft max**

Bayes Rule for density matrices

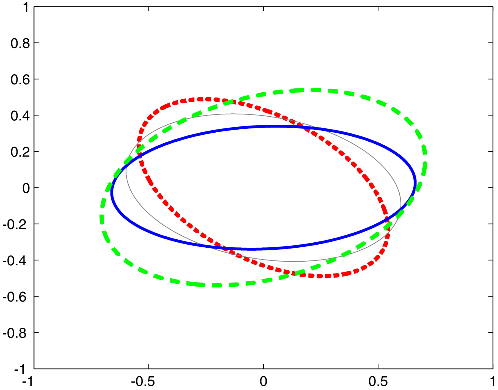
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\,(\text{above matrix})}$$



- 1 update with data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- **Soft max eigenvalue calculation**

Bayes Rule for density matrices

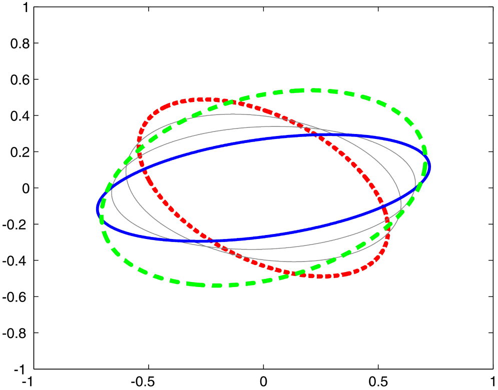
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\mathbf{exp}\left(\mathbf{log}\,\mathbf{D}(\mathbb{M}) + \mathbf{log}\,\mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\,(\text{above matrix})}$$



- 2 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- **Soft max eigenvalue calculation**

Bayes Rule for density matrices

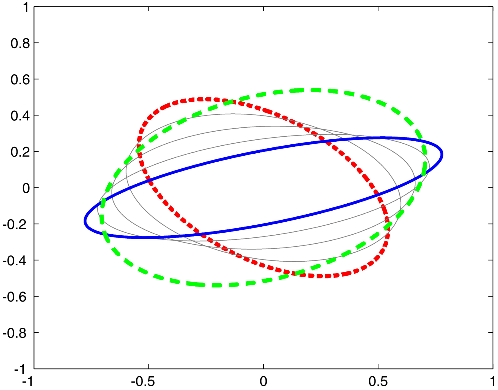
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\,(\text{above matrix})}$$



- 3 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- **Soft max eigenvalue calculation**

Bayes Rule for density matrices

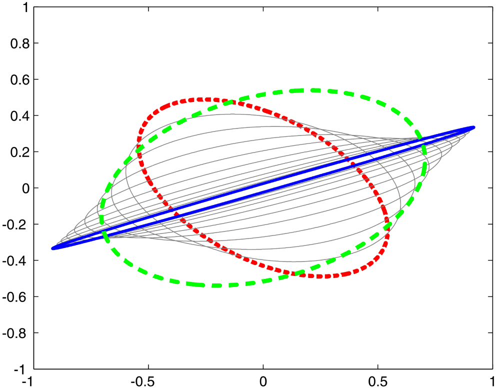
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\,(\text{above matrix})}$$



- 4 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- **Soft max eigenvalue calculation**

Bayes Rule for density matrices

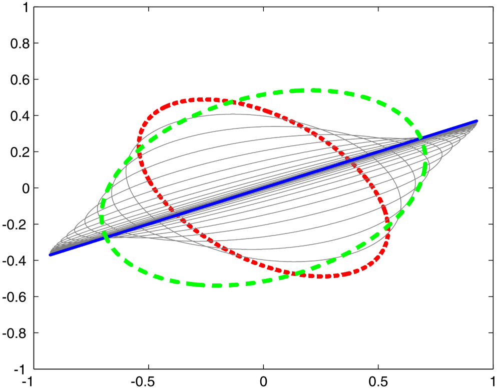
$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\left(\text{above matrix}\right)}$$



- 10 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- **Soft max eigenvalue calculation**

Bayes Rule for density matrices

$$\mathbf{D}(\mathbb{M}|\mathbf{y}) = \frac{\exp\left(\log \mathbf{D}(\mathbb{M}) + \log \mathbf{D}(\mathbf{y}|\mathbb{M})\right)}{\mathrm{tr}\,(\text{above matrix})}$$



- 20 updates with same data likelyhood matrix $\mathbf{D}(\mathbf{y}|\mathbb{M})$
- Update maintains uncertainty information about maximum eigenvalue
- **Soft max eigenvalue calculation**

Bayes' rules

	vector	matrix
Bayes rule	$P(M_i \mid y) = \frac{P(M_i) \cdot P(y \mid M_i)}{\sum_j P(M_j) \cdot P(y \mid M_j)}$	$\mathbf{D}(\mathbb{M} \mid \mathbf{y}) = \frac{\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y} \mid \mathbb{M})}{\mathrm{tr}(\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y} \mid \mathbb{M})}$ $\mathbf{A} \odot \mathbf{B} := \mathbf{exp}(\mathbf{log}\, \mathbf{A} + \mathbf{log}\, \mathbf{B})$

Bayes' rules

	vector	matrix
Bayes rule	$P(M_i\|y) = \frac{P(M_i) \cdot P(y\|M_i)}{\sum_j P(M_j) \cdot P(y\|M_j)}$	$\mathbf{D}(\mathbb{M}\|\mathbf{y}) = \frac{\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}\|\mathbb{M})}{\mathrm{tr}(\mathbf{D}(\mathbb{M}) \odot \mathbf{D}(\mathbf{y}\|\mathbb{M}))}$
		$\mathbf{A} \odot \mathbf{B} := \mathbf{exp}(\mathbf{log}\, \mathbf{A} + \mathbf{log}\, \mathbf{B})$
Regularizer	Entropy	Quantum Entropy

Vector case as special case of matrix case

- Vectors as diagonal matrices
- All matrices same eigensystem
- Fancy $\odot$ becomes $\cdot$

- Often the hardest problem
  ie bounds for the vector case "lift" to the matrix case

Vector case as special case of matrix case

- Vectors as diagonal matrices
- All matrices same eigensystem
- Fancy $\odot$ becomes $\cdot$

- Often the hardest problem
  ie bounds for the vector case "lift" to the matrix case
- This phenomenon has been dubbed the "free matrix lunch"

Size of matrix = size of vector = n

PCA setup

Data vectors $\mathbf{C} = \sum_n \mathbf{x}_n \mathbf{x}_n^\top$

$$\underbrace{\max_{\text{unit } \mathbf{u}} \quad \mathbf{u}^\top \mathbf{C}\, \mathbf{u}}_{\text{not convex in } \mathbf{u}} \quad = \quad \max_{\text{dyad } \mathbf{u}\mathbf{u}^\top} \quad \underbrace{\mathrm{tr}(\mathbf{C}\mathbf{u}\mathbf{u}^\top)}_{\text{linear in } \mathbf{u}\mathbf{u}^\top}$$

Corresponding vector problem $\qquad \max_{\mathbf{e}_i} \quad \underbrace{\mathbf{c}^\top \mathbf{e}_i}_{\text{linear in } \mathbf{e}_i}$

Vector problem is matrix problem when everything happens in the same eigensystem

Uncertainty over unit: probability vector
Uncertainty over dyads: density matrix
Uncertainty over k-sets of units: capped probability vector
Uncertainty over rank k projection matrices: capped density matrix

For PCA

- Solve the vector problem first
- Do all bounds
- Lift to matrix case: essentially replace $\cdot$ by $\odot$
- Regret bounds stay the same
- Free Matrix Lunch

Questions

- When can you “lift” vector case to matrix case?
- When is there a free matrix lunch?
- Lifting matrices to tensors?
- Efficient algorithms for large matrices?
  - Approximations of $\odot$
  - Avoid eigenvalue decomposition by sampling
  - ...