Master's Thesis:

Randomization of Real-valued Matrices for Assessing the Significance of Data Mining Results

Markus Ojala

March 12, 2008

Acknowledgments

- Supervisor: Academy Professor Heikki Mannila
- Instructor: Doctor Kai Puolamäki
- Collaborators: Niko Vuokko, Aleksi Kallio and Niina Haiminen

Background

- *Data mining* is the process of analyzing large amounts of data to find out relevant information
- Many data mining methods are suitable for analyzing real-valued matrices
- Real-valued matrices arise naturally in various applications areas such as in bioinformatics
- The significance testing of data mining results on real-valued matrices is studied

General approach

- Randomization-based significance testing
- Original data $A \in \mathbb{R}^{m \times n}$
- Structural measure $\mathcal{S}(A) =$ data mining result
- $\hat{\mathcal{A}} = \{\hat{A}_1, \ldots, \hat{A}_k\}$ a set of independent randomizations of A sharing some statistics with A
- *Empirical p-value* of the result $\mathcal{S}(A)$:

$$p = \frac{\left|\left\{\hat{A} \in \hat{\mathcal{A}} \,|\, \mathcal{S}(\hat{A}) \leq \mathcal{S}(A)\right\}\right| + 1}{k + 1}$$

Randomization task

- Approach: preserve the mean values and variances of the rows and columns of a matrix in randomization
- Data mining result is interesting if it is not explained by the row and column means and variances of the matrix

Problem

Given an $m \times n$ real-valued matrix A, generate a random matrix $\hat{A}$ chosen independently and uniformly from the set of $m \times n$ real-valued matrices having approximately the same row and column means and variances as A

Example

x	y			
.46	.36	.21	.68	.45
.44	.29	.64	.21	.04
.74	.87	.32	.84	.03
.04	.06	.96	.63	.31
.75	.66	.73	.13	.01
.85	.81	.41	.21	.38
.80	.98	.74	.61	.68
.70	.72	.27	.63	.09
.30	.37	.44	.37	.04
.57	.41	.93	.58	.61

Matrix A

x	y			
.46	.36	.56	.51	.53
.44	.29	.49	.52	.38
.74	.87	.90	.79	.80
.04	.06	.03	.11	.05
.75	.66	.68	.75	.71
.85	.81	.83	.81	.90
.80	.98	.88	.90	.81
.70	.72	.67	.79	.63
.30	.37	.37	.35	.43
.57	.41	.46	.44	.41

Matrix B

- Data mining: correlation between columns x and y ($= 0.92$)
- Significance testing (1000 samples): $p_A = 0.001$, $p_B = 0.4156$

Methods

- Three MCMC methods developed: *SwapDiscretized*, *SwapMetropolis* and *MaskMetropolis*
- Based on local transformations
- Start from the original state, randomize until convergence
- Using the approach introduced by Besag *et al.* to overcome the problem of dependent samples in significance testing

Error measure

- Row and column sums and square sums of A:

$$r_i = \sum_{j=1}^{n} A_{ij}, \quad c_j = \sum_{i=1}^{m} A_{ij}, \quad R_i = \sum_{j=1}^{n} A_{ij}^2, \quad C_j = \sum_{i=1}^{m} A_{ij}^2$$

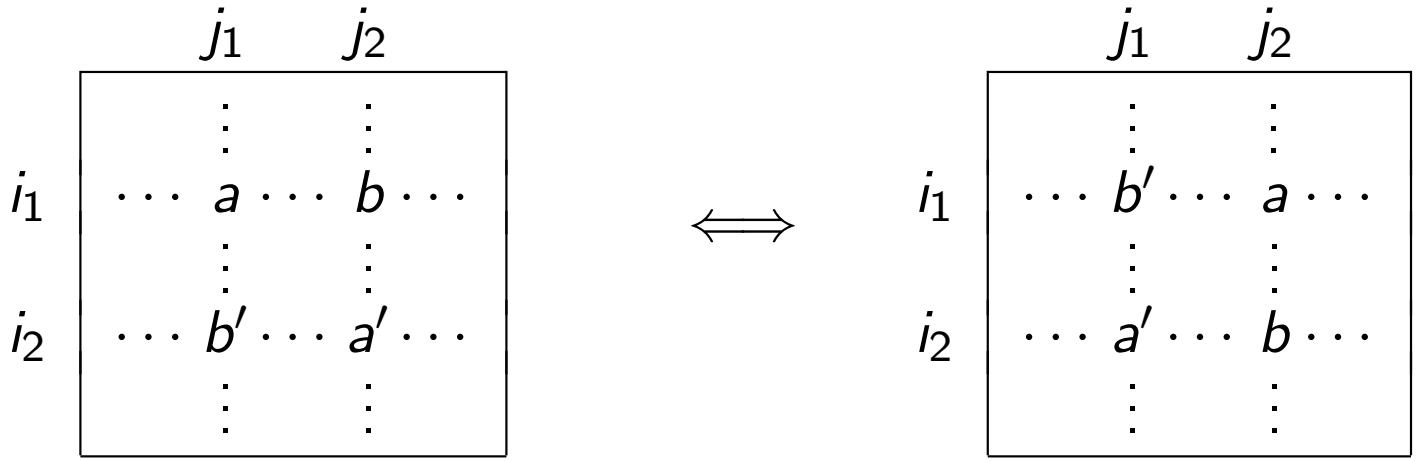
- Similarly $\hat{r}_i$, $\hat{c}_j$, $\hat{R}_i$, $\hat{C}_j$ for randomized matrix $\hat{A}$
- Row and column sum and square sum errors:

$$E(r_i) = |r_i - \hat{r}_i|, \qquad E(c_j) = |c_j - \hat{c}_j|$$
$$E(R_i) = |R_i - \hat{R}_i|, \qquad E(C_j) = |C_j - \hat{C}_j|.$$

- Combined error function:

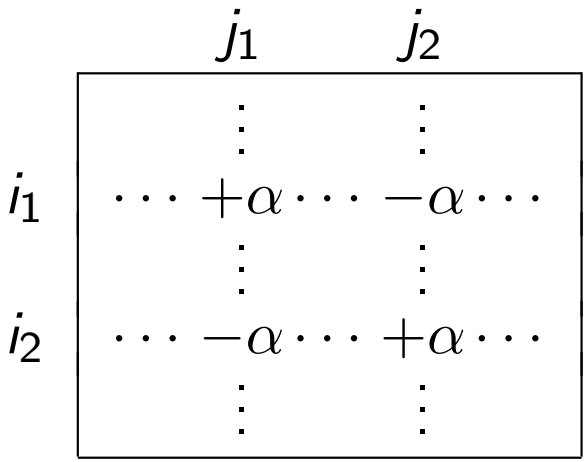
$$E(A, \hat{A}) = w_r \sum_{i=1}^{m} \left(E(r_i)^2 + w_s E(R_i)^2 \right) + \sum_{j=1}^{n} \left(E(c_j)^2 + w_s E(C_j)^2 \right).$$

Swap rotation methods



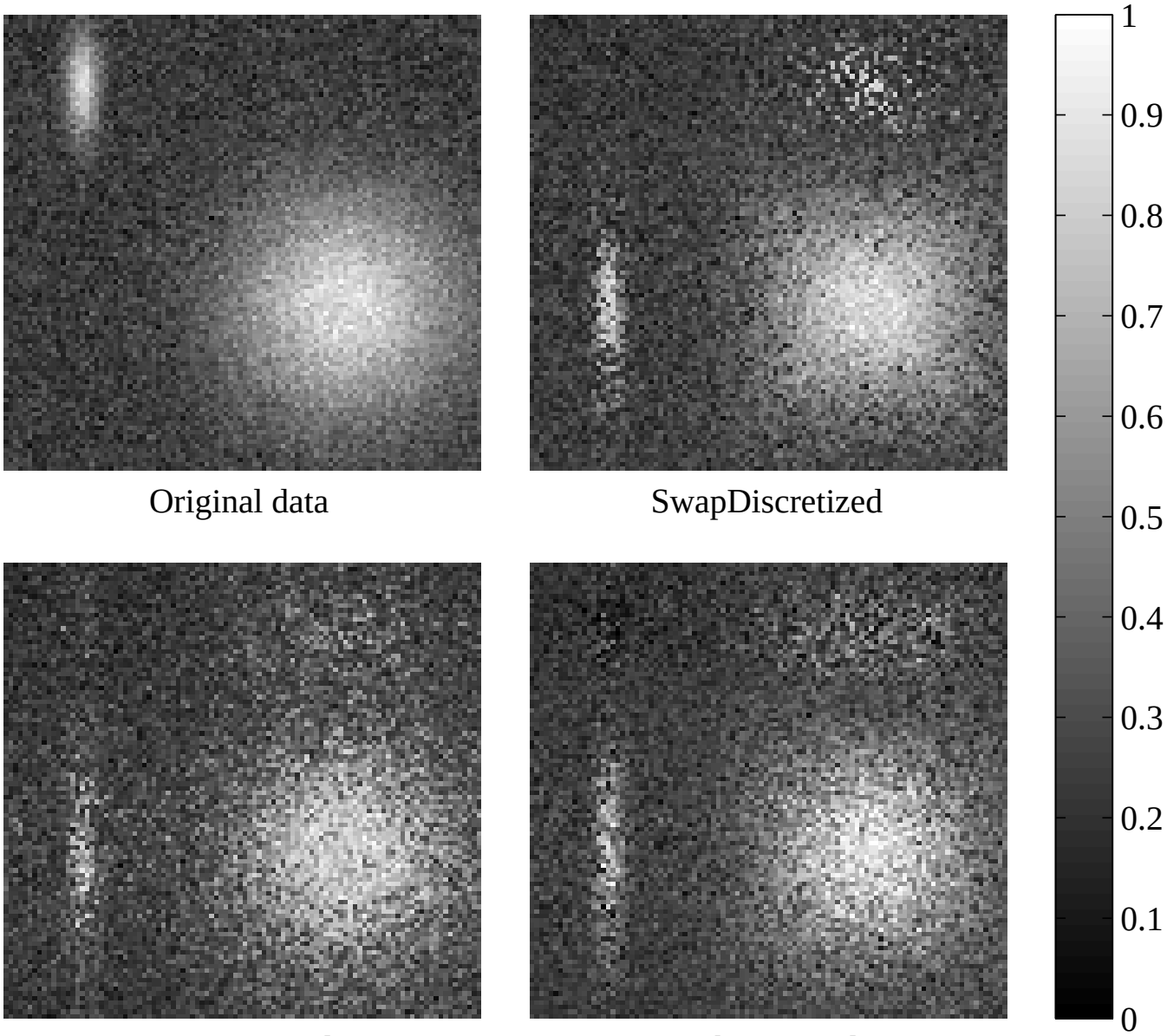
- SwapDiscretized:
  - Discretize the values in N classes
  - Require that a and a' in the same class as well as b and b'
- SwapMetropolis:
  - Direct implementation of Metropolis algorithm
  - $\pi(\hat{A}) = c \exp(-wE(A, \hat{A}))$

Addition mask method



- MaskMetropolis:
  - Another direct implementation of Metropolis algorithm
  - $\pi(\hat{A}) = c \exp(-wE(A, \hat{A}))$
  - Addition α selected uniformly from $[-s, s]$
  - Restricts the values into the original range

Visual examples of randomizations



Properties of the methods

- Convergence of the methods were analyzed empirically:
  - By monitoring the Frobenius distance between A and $\hat{A}$
  - Around $100mn$ steps needed to randomize an $m \times n$ matrix
  - E.g., few seconds needed to randomize an 1000×100 matrix
- Randomization differed significantly from the original matrix and from each other
- Error in the row and column means and variances were few parts per thousand

Significance testing

- Performed significance testing of three data mining task
  - K-means clustering error with 10 clusters
  - Maximum correlation between rows
  - The fraction of variance explained by the first five principal components
- Four generated datasets and one real dataset:

Dataset	Rows	Columns	Mean	Std
Random	100	100	0.473	0.132
Cluster	1117	100	0.509	0.081
Gaussian	1000	10	0.529	0.142
Component	1000	50	0.278	0.116
Gene	1375	60	0.578	0.110

Significance testing: K-means

Dataset	Method	Measure	p-value
Random	Original data	147.02	
	SwapDiscretized	146.74 (0.52)	0.702
	SwapMetropolis	146.71 (0.55)	0.713
	MaskMetropolis	147.35 (0.54)	0.261
Cluster	Original data	457.33	
	SwapDiscretized	659.47 (0.77)	0.001
	SwapMetropolis	661.95 (0.63)	0.001
	MaskMetropolis	656.31 (0.93)	0.001
Gene	Original data	525.53	
	SwapDiscretized	592.38 (1.24)	0.001
	SwapMetropolis	610.70 (0.99)	0.001
	MaskMetropolis	592.29 (1.24)	0.001

Significance testing: Maximum correlation

Dataset	Method	Measure	p-value
Random	Original data	0.363	
	SwapDiscretized	0.361 (0.028)	0.430
	SwapMetropolis	0.361 (0.029)	0.407
	MaskMetropolis	0.360 (0.029)	0.406
Gaussian	Original data	0.993	
	SwapDiscretized	0.992 (0.002)	0.398
	SwapMetropolis	0.992 (0.002)	0.395
	MaskMetropolis	0.992 (0.002)	0.373
Gene	Original data	0.995	
	SwapDiscretized	0.737 (0.046)	0.001
	SwapMetropolis	0.644 (0.026)	0.001
	MaskMetropolis	0.657 (0.024)	0.001

Significance testing: PCA

Dataset	Method	Measure	p-value
Random	Original data	0.173	
	SwapDiscretized	0.174 (0.003)	0.625
	SwapMetropolis	0.173 (0.003)	0.486
	MaskMetropolis	0.174 (0.003)	0.607
Component	Original data	0.941	
	SwapDiscretized	0.765 (0.001)	0.001
	SwapMetropolis	0.736 (0.001)	0.001
	MaskMetropolis	0.769 (0.000)	0.001
Gene	Original data	0.605	
	SwapDiscretized	0.454 (0.001)	0.001
	SwapMetropolis	0.433 (0.001)	0.001
	MaskMetropolis	0.456 (0.001)	0.001

Conclusions

- Considered significance testing of data mining results on real-valued matrices
- Approach used: randomize matrix while preserving row and column means and variances
- Introduced three methods to solve the problem
- Analyzed the methods both theoretically and empirically
- The results imply that the methods are usable in assessing the significance of data mining results