Randomization of real-valued matrices for assessing the significance of data mining results

Markus Ojala[1,2], Niko Vuokko[1,2], Aleksi Kallio[3], Niina Haiminen[4], Heikki Mannila[1,2,4]

[1]HIIT [2]TKK, Helsinki University of Technology [3]CSC, The Finnish IT Center for Science [4]University of Helsinki

Abstract

Randomization is an important technique for assessing the significance of data mining results. We study the problem of generating randomized real-valued matrices sharing the row and column means and variances with the original matrix. We describe three alternative algorithms based on local transformations and evaluate their performance on real and generated data. The results imply that the methods are usable in practice for significance testing of data mining results on real-valued matrices.

Basic approach

- Problem:
  - Original $m \times n$ real-valued matrix A
  - Data mining result $S(A) \in \mathbb{R}$
  - For example, clustering error of the matrix
  - How to assess the significance of $S(A)$?
- Solution:
  - Randomization based significance testing
  - Generate randomized matrices $\hat{A}$ which share some statistics with A
  - Compare $S(A)$ against measures $S(\hat{A})$
  - Calculate an empirical p-value
  - Randomization approach: preserve row and column means and variances

Computational task: Given an $m \times n$ real-valued matrix A, generate a 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.

Empirical p-value

- $\hat{\mathcal{A}} = \{\hat{A}_1, \ldots, \hat{A}_k\}$ a set of randomizations of A
- Empirical p-value of the structural measure $\mathcal{S}(A)$, with the hypothesis of $\mathcal{S}(A)$ being small:

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

Why significance testing matters?

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

The matrices A and B share their first two columns x and y having a high correlation, 0.92. In matrix B the values on each row are tightly distributed around the mean of the row, whereas in matrix A the variance of each row is high.

The high correlation between x and y is significant in A but not significant in B when tested using the randomization methods introduced. The corresponding p-values are

$$p_A = 0.001 \qquad p_B = 0.4156.$$

Measuring the error in means and variances

Let A be the original $m \times n$ real-valued matrix and $\hat{A}$ a randomized matrix. Define row sums r_i and square sums R_i of A as

$$r_i = \sum_{j=1}^{n} A_{ij}, \qquad R_i = \sum_{j=1}^{n} A_{ij}^2,$$

and similarly column sums c_j and square sums C_j. Let $\hat{r}_i, \hat{c}_j, \hat{R}_i, \hat{C}_j$ be the corresponding values of $\hat{A}$. The combined error function of row and column means and variances is

$$E(A, \hat{A}) = w_r \sum_{i=1}^{m} \left(|r_i - \hat{r}_i|^2 + w_s|R_i - \hat{R}_i|^2\right) + \sum_{j=1}^{n} \left(|c_j - \hat{c}_j|^2 + w_s|C_j - \hat{C}_j|^2\right).$$

Algorithms

- Three MCMC methods for randomizing matrix while preserving row and columns means and variances approximately

$$\begin{array}{c|ccccccc|}
 & & j_1 & & j_2 & \\
\hline
 & & \vdots & & \vdots & \\
i_1 & \ldots & a & \ldots & b & \ldots \\
 & & \vdots & & \vdots & \\
i_2 & \ldots & b' & \ldots & a' & \ldots \\
 & & \vdots & & \vdots & \\
\hline
\end{array}
\quad \Longrightarrow \quad
\begin{array}{c|ccccccc|}
 & & j_1 & & j_2 & \\
\hline
 & & \vdots & & \vdots & \\
i_1 & \ldots & b' & \ldots & a & \ldots \\
 & & \vdots & & \vdots & \\
i_2 & \ldots & a' & \ldots & b & \ldots \\
 & & \vdots & & \vdots & \\
\hline
\end{array}$$

Local modification: Swap rotation

- SwapMetropolis:
  - Uses Metropolis algorithm to sample from

$$P(\hat{A}) = c \exp\{-wE(A, \hat{A})\}$$

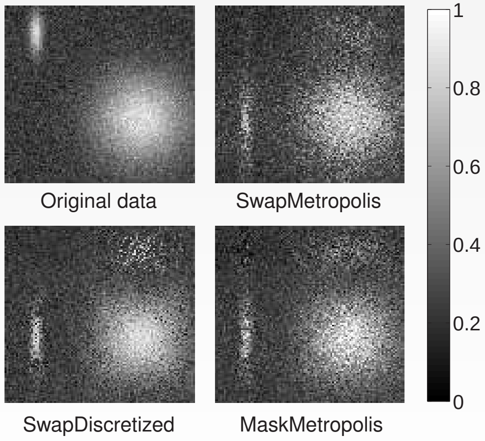
  - Uniform proposal distribution among all possible swap rotations
  - Parameter w is a compromise between efficiency of mixing and the amount of error induced in means and variances
- SwapDiscretized:
  - Discretize values in N classes
  - Require that a and a' are in the same class in a swap as well as b and b'

$$\begin{array}{c|ccccc|}
 & & j_1 & & j_2 & \\
\hline
 & & \vdots & & \vdots & \\
i_1 & \ldots & +\alpha & \ldots & -\alpha & \ldots \\
 & & \vdots & & \vdots & \\
i_2 & \ldots & -\alpha & \ldots & +\alpha & \ldots \\
 & & \vdots & & \vdots & \\
\hline
\end{array}$$

Local modification: Addition mask

- MaskMetropolis:
  - Similar to SwapMetropolis but uses addition mask instead of swap rotation
  - Elements i_1, i_2, j_1, j_1 selected uniformly
  - Addition α selected from $U[-0.1, 0.1]$
  - Restricts the values to the original range

Visual examples of randomizations



Original data is 100×100. The small top left artifact has disappeared in the randomizations.

Significance testing of maximum correlation

Method	Max. correlation	p-value
Random (100×100): random $\mathcal{N}(0,1)$ data		
Original data	0.363	
SwapMetropolis	0.361	0.407
SwapDiscretized	0.361	0.430
MaskMetropolis	0.360	0.406
Gaussian (1000×10): $\mathcal{N}(\mu,1)$, $\mu \sim \mathcal{N}_{10}(0,1)$		
Original data	0.993	
SwapMetropolis	0.992	0.395
SwapDiscretized	0.992	0.398
MaskMetropolis	0.992	0.373
Gene (1375×60): real gene expression data		
Original data	0.995	
SwapMetropolis	0.644	0.001
SwapDiscretized	0.737	0.001
MaskMetropolis	0.657	0.001

- Maximum correlations in Random and Gaussian datasets are insignificant as expected
- Max correlation in dataset Gene is significant
- The methods are usable in significance testing