# 論 文 の 内 容 の 要 旨

論文題目　Statistical Analysis for Pattern Extraction from Relational Data Matrix

（関係データ行列のパターン抽出に関する統計解析）

氏　名　渡邊　千紘

Relational data matrices are everywhere around us, including user-movie rating data and document-word occurrence data. Pattern extraction from such relational data matrices has been extensively studied in the literature to capture and visualize a latent global structure of a given matrix. Specifically, in this paper, we focus on the two well-known problems of relational data analysis: *biclustering* and *matrix reordering*. In the biclustering problem, we assume that a data matrix contains some homogeneous submatrices, say *biclusters*, and estimate the locations of such biclusters. The matrix reordering problem deals with more general structural patterns than biclusters, and its purpose is to find the optimal row/column permutations for a given matrix with which some latent pattern appears (biclusters are special cases here).

An open problem in the biclustering and matrix reordering is that we need to accept various kinds of assumptions in its procedure, such as the number of biclusters or features to be used for row/column orderings. In practice, however, we do not always know the validity of such assumptions in advance. Therefore, we need some evaluation method for the reliability of the model, features, and estimation results of these problems.

In this dissertation, we propose three new approaches for solving this problem in both biclustering and matrix reordering: evaluations of the number of biclusters, the estimated bicluster structure, and the row and column features used for matrix reordering. The first two methods are based on the statistical hypothesis tests, whereas in the last one we maximize the “goodness” of the row/column features in terms of the reconstruction error of the original matrix by a new neural network model.

First, we develop a statistical test on the number of biclusters in a given data matrix $A$. For a given hypothetical number of biclusters $(K_0, H_0)$, we test whether matrix $A$ consists of $K_0 \times H_0$ biclusters or more. The proposed test statistic is based on the largest singular value of the

standardized data matrix, and its asymptotic distribution based on the null hypothesis is derived by using random matrix theory. We also give a theoretical guarantee for the proposed test statistic under the alternative hypothesis. Based on these results, we propose an asymptotically valid sequential testing on the number of biclusters.

Second, we construct a test on the estimated bicluster structure that have been selected based on a given data matrix $A$ and a specific loss function. Such data-driven selection of a hypothetical bicluster structure is a natural choice when we have no knowledge about the latent structure of a matrix $A$ in advance. However, to construct a statistically valid test (i.e., the Type I error is controlled by a given significance rate), we need to take the *selective bias* into account. If we derive the $p$-value of the test statistic based on an invalid assumption that the hypothetical bicluster structure is independent of the data matrix, the test is biased towards optimistic. To avoid this difficulty, we develop a statistical test based on the framework of *selective inference*, where we derive the null distribution of the test statistic under the condition that the hypothetical bicluster structure is selected based on the data matrix.

Finally, we propose a new neural network model for matrix reordering, which automatically extracts row and column features from a given matrix, which are later used for determining the row/column orderings. To evaluate the goodness of the extracted features, we assume a generative model of a data matrix with an autoencoder-like architecture, and evaluate the features based on the reconstruction error of the original matrix. By using the trained neural network model, we can not only determine the row/column orderings of a given matrix $A$ but also visualize a global structural pattern in the matrix $A$ as the output of the model.

Our results provide a clue for examining the validity of the assumptions used in the relational data analysis, and they can be used as first-step analysis tools for acquiring knowledge about the latent structure of a given data matrix.