УДК 519.6

Групповой анализ данных на основе блочного канонического разложения тензоров.

П.В. Харюк [1,2]

[1]Московский государственный университет имени М.В. Ломоносова, факультет ВМК
[2]Институт вычислительной математики РАН

Достаточно большое число исследований связано с анализом данных, объединённых по некоторому признаку. Часто в условиях подобных задач требуется выделить некоторую общую информацию. Например, в задачах когнитивных, медицинских, иных исследований головного мозга у разных людей собирают информацию об электрической активности мозга с помощью многоканальной ЭЭГ записи, о динамике изменения кровотока с помощью функциональной МРТ и т.д. с тем, чтобы обнаружить групповые эффекты или зависимости [1]. Кроме того, особый интерес вызывает возможность выявления индивидуальных особенностей в дополнение к групповым эффектам.

В общем виде задачу группового анализа данных можно определить как извлечение общей информации из набора различных данных, объединённых в группу. Остановимся на случае, когда данные внутри одной группы представляют собой однородные (имеющие одинаковые размеры) массивы (тензоры). Предполагается, что для каждого массива справедливо выражение в виде суммы общей информации для всего набора, индивидуальной информации и ошибки приближения для k-го массива.

В данной работе предлагается метод извлечения групповой информации на основе блочного канонического разложения тензоров [2]. Такая вариация канонического разложения имеет блочную структуру: в нём задаётся число блоков и блочное разбиение фактор-матриц (будем обозначать их как A) по столбцам. Для части мод при этом подматрицы соответствующих фактор-матриц представляют собой повторение отдельного вектор-столбца, которые в совокупности, в свою очередь, образуют отдельные матрицы с числом столбцов, равным числу блоков (обозначим такие матрицы как C).

Имея набор из N образцов (d-1)-мерных однородных данных, мы можем ввести новую, групповую ось, которую расположим на последнем месте (с номером d). Представление в виде суммы общей и индивидуальной частей даёт основания для введения следующих ограничений на вид разложения: пусть $R = N+1$ и матрица C имеет блочную структуру вида «единичная матрица порядка N, вектор-столбец неотрицательных чисел с суммой равной 1».

В подобных предположениях каноническое разложение будет представлено (N+1)-м блоком, из которых первые N соотнесены с индивидуальной частью, последний — с общей (групповой). Дальнейшие предположения накладываются на блоки и формирующие их подматрицы различных мод.

Одна из размерностей может представлять особый интерес (например, временной ряд / пиксель / воксель), в связи с чем измерения по такой размерности часто моделируют как реализации некоторых случайных величин (источников), смешанных друг с другом. Процесс смешивания при этом часто полагается линейным, но подобное смешивание может происходить также по разным модам одновременно, что превращает такой процесс уже в мультилинейный; например, в [3] матрица смешивания полагается представимой в виде произведения Хатри-Рао двух фактор-матриц, которые вместе с матрицей источников составляют каноническое (трилинейное) разложение исходных данных. В данной работе предлагается применять условия к моде, представляющей интерес (обозначим её номер за k), именно через подматрицы соответствующей фактор-матрицы. Например, условие ортогональности первых N подматриц последней подматрице фактор-матрицы A интересующей моды

Подобное условие также было использовано в варианте матричного группового анализа, предложенного в работе [4].

Предложенный в данной работе вариант разложения также можно отнести к тензорной вариации метода анализа связанных компонент (linked multi-way component analysis, LMWCA [5]).

Работа выполнена при поддержке гранта РФФИ, проект № 16-31-00494-мол_а.

Литература

1. Vince D. Calhoun, Jingyu Liu, Tulay Adali. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. // Neuroimage, 2009, V. 45, N. 1, P. 163-172.
2. Laurent Sorber, Marc Van Barel, Lieven De Lathauwer. Optimization-based algorithms for tensor decompositions: Canonical polyadic decomposition, decomposition in rank-(Lr, Lr, 1) terms, and a new generalization. // SIAM Journal on Optimization, 2013, V. 23, N. 2, P. 695-720.
3. Christian F. Beckmann, Stephen M. Smith. Tensorial extensions of independent component analysis for multisubject fMRI analysis. // Neuroimage, 2005, V. 25, N. 1, P. 294-311.
4. Guoxu Zhou, Andrzej Cichocki, Yu Zhang, Danilo Mandic. Group component analysis for multiblock data: Common and individual feature extraction. // IEEE Transactions on Neural Networks and Learning Systems, 2015, P. 1-14.
5. Andrzej Cichocki, Danilo Mandic, Lieven De Lathauwer, Guoxu Zhou, Qibin Zhao, Cesar Caiafa, Anh Huy Phan. Tensor decompositions for signal processing applications: From two-way to multiway component analysis. // IEEE Signal Processing Magazine, 2015, V. 32, N. 2, P. 145-163