УДК 519.254

Бурнаев Е.В. [1]

[1] Московский физико-технический институт (ГУ)

Об алгоритме подсчета главных компонент в равномерной метрике

Снижение размерности данных представляет собой такое преобразование многомерных данных, которое снижает размерность данных до внутренней размерности или близкой к ней (под внутренней размерностью данных понимается такое количество параметров, которое достаточно для того, чтобы объяснить наблюдаемые свойства данных [1]).

Снижение размерности широко применяется во многих областях науки и техники, поскольку оно позволяет существенно увеличить эффективность различных алгоритмов обработки данных за счет ослабления нежелательных эффектов, вызванных “проклятием размерности” [2,3].

Метод главных компонент (ГК), по всей видимости, в силу своей простоты, аналитических свойств и наличия эффективных алгоритмов подсчета является наиболее распространенным методом снижения размерности [4,5].

Существует достаточно большое количество модификаций ГК, учитывающих те или иные практические нужды. Например, в [6,7] рассмотрены подходы к подсчету ГК для данных, принимающих дискретные значения. В [8,9] предложены варианты метода ГК, с помощью которых можно обрабатывать неполные данные. В [10,11] разработаны алгоритмы подсчета ГК, которые позволяют получать разреженную матрицу нагрузок.

Известно, что ГК минимизируют сумму квадратов евклидовых расстояний между исходными векторами выборки и их восстановленными аналогами [4]. Классические алгоритмы подсчета ГК подвержены влиянию выбросов, поэтому в [8] были предложены робастные варианты подсчета ГК. Также в [8] за счет введения весов предлагалось: учитывать “надежность” того или иного наблюдения из выборки, используемой для оценки ГК, а также влиять на точность восстановления определенных координат по сжатым векторам.

Однако до сих пор ни в одной из работ не был разработан алгоритм подсчета таких ГК, которые бы минимизировали среднюю (по выборке) взвешенную максимальную (по координате) ошибку восстановления векторов выборки по сжатым данным. Использование ГК, подсчитанных на основе такого алгоритма, позволило бы учесть требования, что точность восстановления векторов выборки по разным координатам должна быть разной, а максимальная ошибка восстановления не должна превышать заданное значение. Эти требования, например, важны при сжатии описания аэродинамических профилей крыла самолета.

В работе предлагается один из возможных алгоритмов для подсчета ГК, обладающих указанными свойствами. Приводятся результаты моделирования на искусственных и реальных данных, иллюстрирующие эффективность предложенного алгоритма.

СПИСОК ЛИТЕРАТУРЫ

1. *Fukunaga K.* Introduction to Statistical Pattern. – San Diego: Academic, 1990.
2. *Bellman R.* Adaptive Control Processes: A Guided Tour. – Princeton: Princeton Univ. Press, 1961.
3. *Scott D.W.* Multivariate Density Estimation. Theory, Practice, and Visualization. – New York: Wiley, 1992.
4. *Jackson J.E.* A User's Guide to Principal Components. – New York: Wiley, 1991.
5. *Jolliffe I.T.* Principal Component Analysis. Springer Series in Statistics. – Berlin: Springer-Verlag, 1986.
6. *Collins M., Dasgupta S., Schapire R.E.* A Generalization of Principal Components Analysis to the Exponential Family // Advances in Neural Information Processing Systems 14 (NIPS). – Cambridge: MIT Press, 2002.
7. *Sajama B., Orlitsky A.* Semi-parametric Exponential Family PCA. // Advances in Neural Information Processing Systems 17 (NIPS). – Vancouver, 2002.
8. *Skocaj D., Leonardis A., Bischof H.* Weighted and Robust Learning of Subspace Representations // Pattern Recognition. – 2007. – V. 40, N. 5. – P. 1556-1569.
9. *Roweis S.* EM Algorithms for PCA and SPCA // Proc. 1997 Conf. on Advances in Neural Information Processing Systems 10. – 1998. – P. 626-632.
10. *Zou H., Hastie T., Tibshirani R.* Sparse Principal Component Analysis // J. Computational and Graphical Statistics. – 2006. – V.15, N. 2. – P. 265-286.
11. *Chipman H.A., Gu H.* Interpretable Dimension Reduction // J. Appl. Statistics. – 2005. V. 32, N. 9. – P. 969-987.