OUTLIER EXPLANATION BASED ON SHAPLEY VALUES FOR VECTOR- AND MATRIX-VALUED OBSERVATIONS

Peter Filzmoser [1] and Marcus Mayrhofer[1]

[1] Institute of Statistics and Mathematical Methods in Economics, TU Wien, Austria, (e-mail: Peter.Filzmoser@tuwien.ac.at, Marcus.Mayrhofer@tuwien.ac.at)

ABSTRACT: Shapley values are a practical tool from Explainable AI used to interpret model outcomes on the observation level. Their usefulness has also been demonstrated in the context of multivariate outlier detection, where the contributions of single variables to the overall outlyingness are evaluated. This allows for an alternative view to cellwise outlyingness, where the interest is in identifying deviating cells of a data matrix. The concept of outlier explanation based on Shapley values can be extended to outlyingness for matrix-valued observations, which is an interesting new topic in robustness by itself.

KEYWORDS: Anomaly explanation, Shapley value, Mahalanobis distance.

1 Shapley Values for Vector-valued Observations

Shapley values have been introduced in cooperative game theory, where they evaluate the collective payoff of a coalition of players (Shapley, 1953). In the context of multivariate data, each observation is analyzed separately. A player would be an individual variable, and one can be interested in a subset of variables' effect on an outcome. For example, for a black-box method in classification, we might want to know why an observation has been assigned to a particular class. Shapley values allow evaluating how the variables contributed to the classifier's decision (Lundberg & Lee, 2017).

Also, in the context of multivariate outlier detection, it is of interest why an observation has been declared outlying. A traditional tool for multivariate outlier detection is the Mahalanobis distance (Mahalanobis, 1936). To reliably identify outliers, it is essential to robustly estimate mean and covariance (Rousseeuw & Zomeren, 1990), and one option is to use the Minimum Covariance Determinant (MCD) estimator (Rousseeuw & Driessen, 1999). Shapley values can be adapted to the setting of squared Mahalanobis distances: One can obtain a decomposition of this distance measure into an outlyingness score for each variable, which can be interpreted as the average marginal contribution to the outlyingness of an observation (Mayrhofer & Filzmoser, 2023). The sum of all these contributions is identical to the squared Mahalanobis distance of the observation. Another interesting feature is that the computational complexity of determining the Shapley values reduces to a very simple problem in the context of Mahalanobis distances, and thus the computations are very time-efficient, also in higher dimensions.

While the Shapley values inform about the contribution of the variables to the outlyingness of an observation, they do not inform about the values these cells would have if the observation would not be contaminated. This, however, is the goal of cellwise outlyingness methods (Rousseeuw & Bossche, 2018). A modification in the calculations of Shapley values also allows getting this information by which amount a cell needs to be modified to make the observation non-outlying (Mayrhofer & Filzmoser, 2023). As an outcome, one can obtain diagnostics regarding cellwise outlyingness.

2 Shapley Values for Matrix-valued Observations

Another important class of data structures are matrix-valued observations. Thus, the information is represented in the rows and columns of a matrix, and a prominent example are image data. Often, matrix-valued observations are vectorized; for example, the pixel information of an image can be arranged in a long vector, which then forms one row of a "traditional" data matrix. This leads to very high-dimensional data in which the neighborhood relationship of the pixels is lost.

The concept of matrix-valued data is not new at all, and a prominent distribution in this context is the matrix normal distribution (Dawid, 1981). There are different proposals in the literature on how to estimate the parameters of this distribution (Dutilleul, 1999). It is also possible to define a Mahalanobis distance, and the concept of the MCD estimator can be modified to obtain robust estimators. Finally, Shapley values can be used, and their contributions again sum up to an observation's squared Mahalanobis distance. In the context of image data, for example, one can identify outlying images and explain which pixels contribute to this outlyingness. A more detailed background, as well as illustrative examples, will be provided in the presentation.

References

DAWID, A PHILIP. 1981. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, **68**(1), 265–274.

DUTILLEUL, PIERRE. 1999. The mle algorithm for the matrix normal distribution. *Journal of Statistical Computation and Simulation*, **64**(2), 105–123.

LUNDBERG, SCOTT M, & LEE, SU-IN. 2017. A Unified Approach to Interpreting Model Predictions. *Pages 4765–4774 of:* GUYON, I., LUXBURG, U. V., BENGIO, S., WALLACH, H., FERGUS, R., VISHWANATHAN, S., & GARNETT, R. (eds), *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc.

MAHALANOBIS, PRASANTA CHANDRA. 1936. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, **2**, 49–55.

MARONNA, RICARDO A, MARTIN, R DOUGLAS, YOHAI, VICTOR J, & SALIBIÁN-BARRERA, MATÍAS. 2019. *Robust statistics: theory and methods (with R)*. John Wiley & Sons.

MAYRHOFER, MARCUS, & FILZMOSER, PETER. 2023. Multivariate outlier explanations using Shapley values and Mahalanobis distances. *Econometrics and Statistics*. To appear.

MOLNAR, CHRISTOPH. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.

RAYMAEKERS, JAKOB, & ROUSSEEUW, PETER J. 2022. The Cellwise Minimum Covariance Determinant Estimator. *arXiv preprint arXiv:2207.13493*.

RIBEIRO, MARCO, SINGH, SAMEER, & GUESTRIN, CARLOS. 2016 (02). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

ROUSSEEUW, PETER. 1985. Multivariate Estimation With High Breakdown Point. *Mathematical Statistics and Applications Vol. B*, 01, 283–297.

ROUSSEEUW, PETER, & ZOMEREN, BERT. 1990. Unmasking Multivariate Outliers and Leverage Points. *Journal of The American Statistical Association - J AMER STATIST ASSN*, **85**(06), 633–639.

ROUSSEEUW, PETER J., & BOSSCHE, WANNES VAN DEN. 2018. Detecting Deviating Data Cells. *Technometrics*, **60**(2), 135–145.

ROUSSEEUW, PETER J., & DRIESSEN, KATRIEN VAN. 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41**(3), 212–223.

SHAPLEY, LLOYD S. 1953. A value for n-person games. *Contributions to the Theory of Games*, **2**(28), 307–317.

ZHANG, YICHI, SHEN, WEINING, & KONG, DEHAN. 2022. Covariance estimation for matrix-valued data. *Journal of the American Statistical Association*, 1–12.