Ordinal discriminative dimensionality reduction of functional profiles

Giulia Patanè[1], Federica Nicolussi[1], Alexander Krauth[3], Bianca Maria Colosimo[2], Luca Dede'[1], and Alessandra Menafoglio[1]

[1] MOX, Department of Mathematics, Politecnico di Milano, Milan MI, Italy
[2] Department of Mechanics, Politecnico di Milano, Milan MI, Italy
[3] University of Tübingen, Tübingen TÜ, Germany
giulia.patane@polimi.it
federica.nicolussi@polimi.it
alexander.krauth@uni-tuebingen.de
bianca.colosimo@polimi.it
luca.dede@polimi.it
alessandra.menafoglio@polimi.it

Abstract. Optical biosensors, utilizing biological components like DNA and antibodies, are effective analytical tools. However, analyzing sensor signals, especially in large datasets such as reflectometric imaging sensors, is computationally demaning both in terms of time and of memory usage. In this communication, we employ data from a reflectometric imaging sensor to track the antibody-antigen reaction progression using video images of the biosensor surface. Analyzing temporal changes in light intensity provides insights into reaction progression; however, the need for an automatic detector of biological process reaction requires dimensionality reduction of the data from the sensor, which arrive in the form of functional profiles from video signals. We illustrate a workflow which includes cleaning light disturbances, condensing video data, and reducing the dimensionality of the obtained functional data. Departing from traditional methods like Functional Principal Component Analysis (FPCA), we discuss functional-ordinal Canonical Correlation Analysis to optimize correlation between the high-dimensional predictor and the outcome. The conclusion highlights that, the combination of preprocessing and CCA for effective discrimination among different reagent concentration levels, enables one to project the video signal into a 2-dimensional space. This innovative approach enhances our ability to detect virus vitality in biomanufacturing processes.

Keywords: Functional Data Analysis, Ordinal Data, Canonical Correlation Analysis, Biosensors

1 Introduction

Optical biosensors are a powerful analytical tool for diagnostic purposes. As a sensing element biological components such as DNA, antibodies, enzymes, cell

receptors, and even whole cells can be used [14–16]. In the case of large data sets – for example in reflectometric imaging sensors – the analysis of the sensor signals can be very time-consuming and requires huge amounts of memory space while the output of these analyses often is a single parameter, for example, the concentration of the analyte. In the present study, a reflectometric imaging sensor was used to monitor the progression of the reaction between an antibody and an antigen through video images of the biosensor surface. Here, the statistical challenge consists in analysing the temporal profiles of the light intensity (i.e., functional observations) in Regions of Interest (ROIs), in response to the binding of the antibody to the antigen spot of pre-specified concentrations onto the sensor surface. Given the necessity for simultaneous and real-time identification of reactions across numerous ROIs, an automatic, efficient and fast reaction detector becomes imperative. This underscores the need for dimensionality reduction in sensor data, typically presented as video signals. In this communication, we introduce the dimensionality reduction method proposed in [12], with a specific emphasis on discerning the presence of a reaction in a designated region and on predicting the reagent concentration level. For this purpose, we illustrate a pipeline composed of three steps. Since our focus is on video signals, the initial phase involves cleaning of the light disturbances. The subsequent phase is dedicated to extracting a functional profile from the video frames corresponding to the ROI. The final phase centers on reducing dimensionality of the functional data. Regarding the last step, traditional literature suggests employing Principal Component Analysis (PCA) for dimensionality reduction in predictive models, where the first k components are utilized as predictors. While PCA seeks components maximising the variance, we here follow [12] and conduct the dimensionality reduction by optimizing the correlation between the high-dimensional predictor and the outcome. In this sense, we shall rely on a variant of Canonical Correlation Analysis named functional-ordinal CCA, to select components that effectively discriminate among units with different concentration levels.

2 Dataset and preprocessing

The dataset consists of a video signal, recording one image per second, captured through a reflectometric sensor. Each video frame represents the light intensity of the reflected light. Specifically, 1035 spots, each containing a certain concentration of the antigen, are arranged in a regular grid on the sensor surface, identifying 1035 ROIs. In practice, some spots may not exist or contain minimal reagent concentration. Note that visually detecting the presence of a spot is straightforward, as it appears darker than the background in the reflectometric image. In this section we report the preprocessing of the video signal, from the light cleaning to the construction of the final dataset, where the statistical unit is the ROI and the covariates are (i) a univariate functional datum that represents the light intensity temporal evolution and (ii) an ordinal variable that represents the level of concentration of reagent.

For each ROI, the nominal concentration level is associated with an ordinal label. In order to detect magnitude outliers among the ROIs, we eliminate exogenous sources of light from the reflectometric images, such as external lights and irregularities in the surface coating. In our case study, it is realistic to assume that the exogenous sources of light are constant over time. We model the logarithm of light intensity at time 0 as a Generalized Additive Model (GAM) [6], using as predictor the pixels's spatial coordinates (x, y), with a nonlinear contribution, and the categorical information of whether or not the pixel belongs to a ROI. By subtracting the model-based expected light disturbance on the reflectometric images, we effectively minimize variations in light that do not stem from the presence of the antigen spot. Net of the light cleaning, we can easily check if a pixel belongs to random spots or determine if a randomly sampled pixel is located within or outside a ROI.

Light intensity is a variable ranging from 0 to $2^{16} - 1$, making it compositional in its nature. To conduct standard statistical analyses (such as dimensionality reduction, models or clustering) it is essential to work with absolute information rather than relative, as is the case with compositional data [8]. Several transformations for compositional data are suggested in the literature to ensure meaningful and interpretable results; amongst the most popular methodologies, we mention that based on the log-ratio approach (firstly proposed by Aitchison, see, e.g., [4]) which is also used in [12]. Following these works, the univariate functional datum representing the light intensity of the i-th ROI is built as follow. We subtract from the raw signal the GAM model of light disturb, obtaining a cleaned time-varying light intensity, $f_j(t)$ for each pixel j in the image D. For each ROI i, we compute the 3%-trimmed mean of the (cleaned) signals $f_j(t)$ corresponding to $N = 100$ randomly extracted pixels within the ROI, obtaining $\bar{f}_i^{ROI}(t)$. Note that this guarantees robustness for the estimated mean function within the ROI, which is then used as representative of the signal for the ROI. We also sample $N = 100$ pixels in the neighborhood of each ROI i and we compute the 3%-trimmed mean $\bar{f}_i^{BG}(t)$, in order to have a robust estimation of the mean in the background area adjacent to the ROI i. In order to take into account for the compositional nature of the signals, we compute the log-ratio:

$$\bar{x}_i(t) = log(\frac{\bar{f}_i^{ROI}(t)\bar{f}_i^{BG}(0)}{\bar{f}_i^{ROI}(0)\bar{f}_i^{BG}(t)})$$

We smooth the log-ratio through smoothing splines regression [9] with 20 cubic splines, equally distributed knots, and penalty parameter equal to 10^5 selected via GCV, in order to obtain a meaningful functional profile. In conclusion, every ROI i is associated with a functional datum $x_i(t)$ and a level L_i of concentration of reagent. In Figure 1 one can observe the functional boxplot of the log-ratio by level of concentration.

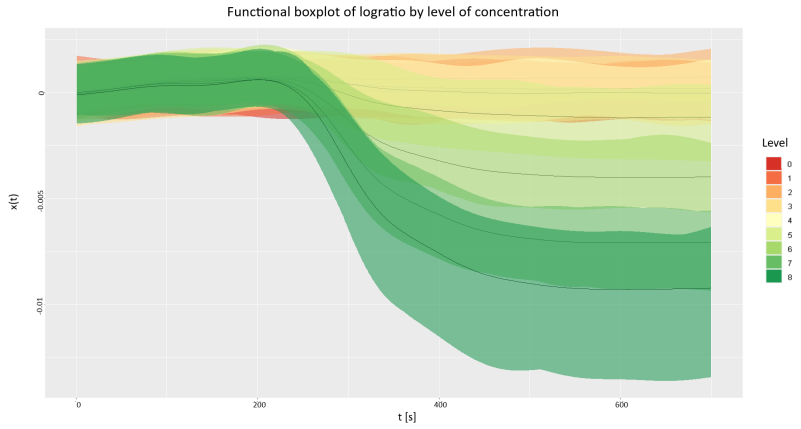

Fig. 1. Functional boxplot of a sample of smoothed log-ratio, stratified by concentration level.

3 Dimesionality reduction

We aim at reducing the dimensionality of the functional data while maximizing the ability to predict the ordinal level in the reduced space. Following [12], we rely on Canonical Correlation Analysis [7] which finds the directions that maximize the correlation between two datasets. In [3] the concept of CCA was extended to functional datasets [1], where the smoothness of the functional directions was imposed by introducing appropriate penalizations in the objective functional. As a further challenge, we here aim to use the concept of canonical correlation in a context in which one wants to find the functional canonical directions that enables one to discriminate an ordinal variable (the concentration), from the reduced functional signal (the time-varying preprocessed light intensity). By ordinal variable we here refer to discrete but not categorical variable for which labels have a well defined order [5]. For an introduction to a useful approach for their statistical analysis we refer to [13]; we here limit to recall that this case needs a suitable encoding of the ordinal variable, in order to take account the order and the possible non-equidistance of consecutive levels. In [12] we propose an encoding criterion for the ordinal variable, suitable to use CCA in the functional-ordinal framework, and we provide an interpretation of the canonical directions. On this basis, [12] develops a novel methodology, named functional-ordinal CCA, which allows one to identify the (smoothed) directions of maximum association between a functional and an ordinal dataset, robustly to the presence of magnitude and shape outliers [10] [11]. In this communication, we shall discuss the functional-ordinal CCA methodology and its application for biosensors. In particular, the analysis of the available data suggests that a functional-ordinal CCA dimensionality reduction of dimension 2 allows to effectively discriminate

among different levels of concentrations, unlike classical dimensionality reduction methods such as PCA.

4 Conclusions and further analyses

In this communication, we discuss the combination of the preprocessing showed in Section 2 with dimensionality reduction through the functional-ordinal CCA, in which we encode the ordinal datum in a suitable way that allows the levels to be non-equidistant. This strategy allows one to project a high dimensional object such as the video of light intensity into a lower dimensional space, maximizing the ability to discriminate among different levels of the ordinal variable. In [12] we validate the functional-ordinal CCA methodology, through k-fold Cross Validation. Moreover, we show through extensive simulations that functional-ordinal CCA out-performs the FPCA, in terms of classification accuracy, even in situations for which the dataset shows a higher variability than that explained by the ordinal variable only, as it is typically the case for the class of target datasets in the biosensor applications considered in this work.

Acknowledgements. This research has received funding by the European Commission under the "HORIZON-CL4-2021-DIGITALEMERGING-01 project BioProS - Biointelligent Production Sensor to Measure Viral Activity" (grant agreement no. 101070120), 2022-2026". The present research is part of the activities of "Dipartimento di Eccellenza 2023–2027", MUR, Italy, Dipartimento di Matematica, Politecnico di Milano.

References

1. Leurgans S. E., Moyeed R. A. and Silverman B. W. (1993) Canonical correlation analysis when the data are curves. *Journal of the Royal Statistical Society*, 55, 725–740.
2. Silverman B. W. (1996) Smoothed functional principal components analysis by choice of norm. *Annals of Statistics*, 24, 1–24.
3. Ramsay J. O., Silverman B. W. (2005) Functional Data Analysis, *Springer*.
4. Aitchison J. (2011) The Statistical Analysis of Compositional Data, *Monographs on statistics and applied probability*, *Springer*.
5. Agresti A. (2013) Categorical Data Analysis (3 ed.). *Hoboken, New Jersey: John Wiley Sons*.
6. Hastie T., Tibshirani R. (1986) Generalized Additive Models. Statist. Sci. 1 (3) 297 - 310.
7. Härdle W., Simar L. (2007) Canonical Correlation Analysis, *Applied Multivariate Statistical Analysis*, *Springer*.
8. Greenacre M. (2021) Compositional Data Analysis, *Annual Review of Statistics and Its Application*.
9. Reinsch C.H. (1967) Smoothing by spline functions, Numer. Math. 10, 177–183.
10. Lopez-Pintado S. and Romo J. (2012) A half-region depth for functional data, *Computational Statistics and Data Analysis*, 55, 1679-1695.

11. Arribas-Gil A. and Romo J. (2014) Shape outlier detection and visualization for functional data: the outliergram, *Biostatistics*, 15(4), 603-619.
12. Patanè G., Nicolussi F., Krauth A., Colosimo B. M., Dede' L., Menafoglio A. (2024) Functional-Ordinal Canonical Correlation Analysis With Application to Data from Bioprinting Sensors, *MOX Report*.
13. Hastie T., Buja A. and Tibshirani R. (1995) Penalized discriminant analysis. *Annals of Statistics*, 23, 73–102.
14. Fechner P., Gauglitz G., Proll G. (2022) Through the looking-glass - Recent developments in reflectometry open new possibilities for biosensor applications, *TrAC Trends in Analytical Chemistry*, 156, 2022, 116708
15. Gauglitz G. (2010) Direct optical detection in bioanalysis: an update, *Anal Bioanal Chem*, 398, 2363–2372
16. Hutterer J., Proll G., Fechner P. et al. (2022) Parallelized label-free monitoring of cell adhesion on extracellular matrix proteins measured by single colour reflectometry, *Anal Bioanal Chem*, 414, 575–585