*II PhDay-EIO*
*Santiago de Compostela, 23 de xuño do 2021*

**Nonparametric association tests for discrete data in general spaces, with applications to psychiatric genetics**

Fernando Castro-Prado[1,2], Javier Costas[2], Dominic Edelmann[3], Wenceslao González-Manteiga[1], David R. Penas[1]

[1]Department of Statistics, Mathematical Analysis and Optimisation. Universidade de Santiago de Compostela.
[2]Laboratory of Psychiatric Genetics. Health Research Institute (IDIS), University Hospital of Santiago de Compostela.
[3]Biostatistics Unit. German Cancer Research Centre (DKFZ).

**ABSTRACT**

Testing the nonparametric hypothesis of independence consists in assessing whether a group of two or more variables of interest are associated to one another or not. An overview of relevant problems in biomedicine that boil down to that statistical question will be provided, with special focus on case-control studies of complex human traits.

Understanding epistasis (genetic interaction) may shed some light on the genomic basis of common diseases, including disorders of maximum interest due to their high socioeconomic burden, like schizophrenia. In order to use data from single nucleotide polymorphisms (SNPs), which are discrete by their own nature, a robust statistical tool that detects nonlinear dependence in metric spaces is required (Castro-Prado and González-Manteiga, 2020).

Distance correlation (Székely and Rizzo, 2017) is a recently developed association measure that characterises general statistical independence between random variables, not only the linear one. We (Castro-Prado *et al.*, 2020) propose distance correlation as a novel tool for the detection of epistasis in genome-wide association studies and show some preliminary results using a Galician schizophrenia study (Rodríguez-López *et al.*, 2020).

Then we provide some insight into how this theory is mathematically equivalent to the Hilbert–Schmidt independence criterion (Sejdinovic *et al.*, 2013) that is popular in the machine learning community and how, therefore, all this can be rewritten in the language of the global tests developed by Goeman *et al.* (2006). We conclude with some practical advice regarding how these profound probabilistic results show the advantages of using the aforementioned novel statistical methodology for biomedical scientists.

# References

Castro-Prado, F., Costas, J., González-Manteiga, W. and Penas, D. R. (2020). Searching for genetic interactions in complex disease by using distance correlation. [Preprint.] Available at https://arxiv.org/abs/2012.05285.

Castro-Prado, F. and González-Manteiga, W. (2020). Nonparametric independence tests in metric spaces: What is known and what is not. [Preprint.] Available at https://arxiv.org/abs/2009.14150.

Goeman, J. van de Geer, S. and van Houwelingen, H. (2006). Testing against a high dimensional alternative. *Journal of the Royal Statistical Society (Series B)*, **68,** 477–493.

Rodríguez-López, J., Arrojo, M., Paz, E., Páramo, M. and Costas, J. (2020). Identification of relevant hub genes for early intervention at gene coexpression modules with altered predicted expression in schizophrenia. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **98**.

Sejdinovic, D., Sriperumbudur, B., Gretton, A. and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, **41,** 2263–2291.

Székely, G. J. and Rizzo, M. L. (2017). The energy of data. *Annual Review of Statistics and Its Application* **4,** 447–479.