# XERp a novel approach to variable selection for classification

**Mari van Reenen**
North-West University, South Africa

I intend to discuss recently published work on two statistical methods for variable selection and classification namely, ERp and XERp, as well as future prospects.

**Problem Statement:** Variable selection and classification can become complex due to the large number and sources of missing values often present in data, specifically data generated in GC-MS based metabolomics experiments. Missing values are set to zero under certain conditions resulting in a mixture distributions which are difficult for most statistical approaches to account for.

**Methodology:** ERp is a variable selection and classification method based on minimized classification error rates, from a control and experimental group. ERp tests the null hypothesis that there is no difference between the distributions of the two groups. Significant variables are can discriminate between the two groups and provide insight into the biological mechanisms driving group differences. XERp is an extension of ERp that takes zero-inflated data into account. XERp addresses two sources of zero-valued observations: (i) zeros reflecting the complete absence of a metabolite from a sample (true zeros); and (ii) zeros reflecting a measurement below the detection limit.

**Findings:** XERp performs well with regard to bias and power. XERp was also applied to a GC-MS dataset from a metabolomics study on *tuberculosis meningitis* in infants and children and generated a list of discriminatory variables which informed the biological interpretation of the data. XERp also attained satisfactory classification accuracy for previously unseen cases in a leave-one-out cross-validation context.

**Conclusion & Significance:** XERp is able to identify variables that discriminate between two groups by simultaneously extracting information from the difference in the proportion of zeros and shifts in the distributions of the non-zero observations. XERp uses simple rules to classify new subjects and a weight pair to adjust for unequal sample sizes or sensitivity and specificity requirements.

## Recent Publications

- Van Reenen, M., Westerhuis, J. A., Reinecke, C. J., & Venter, J. H. (2017). Metabolomics variable selection and classification in the presence of observations below the detection limit using an extension of ERp. BMC bioinformatics, 18(1), 83.
- Irwin, C., van Reenen, M., Mason, S., Mienie, L. J., Westerhuis, J. A., & Reinecke, C. J. (2016). Contribution towards a Metabolite Profile of the Detoxification of Benzoic Acid through Glycine Conjugation: An Intervention Study. PloS one, 11(12).
- Van Reenen, M., Reinecke, C. J., Westerhuis, J. A., & Venter, J. H. (2016). Variable selection for binary classification using error rate p-values applied to metabolomics data. BMC bioinformatics, 17(1), 33.
- Moutloatse, G. P., Bunders, M. J., van Reenen, M., Mason, S., Kuijpers, T. W., Engelke, U. F., & Reinecke, C. J. (2016). Metabolic risks at birth of neonates exposed in utero to HIV-antiretroviral therapy relative to unexposed neonates: an NMR metabolomics study of cord blood. Metabolomics, 12(11), 175.
- Mason, S., van Furth, A. M. T., Solomons, R., Wevers, R. A., van Reenen, M., & Reinecke, C. J. (2016). A putative urinary biosignature for diagnosis and follow-up of *tuberculous meningitis* in children: outcome of a metabolomics study disclosing host–pathogen responses. Metabolomics, 12(7), 1-16

## Biography

Mari van Reenen is head of Bioinformatics at the Centre for Human Metabolomics, North-West University (Potchefstroom Campus), South Africa as well as a PhD candidate at the Department of Statistics, Faculty of Natural Sciences North-West University (Potchefstroom Campus), South Africa. The new statistical methods presented here form part of her PhD study aimed at developing new statistical approaches that can account for the nuances of metabolomics data. Her continued work and research aims to bridge the gap between the assumptions statistical tools often require and the reality of experimental work, from experimental design to data analysis.

van.reenen.mari@gmail.com