Overcoming computational inability to predict clinical outcome from high-dimensional biomedical data using Bayesian methods

A. S. S. Shalabi[1] **and A. C. C. Coolen**[1]
[1] Institute for Mathematical and Molecular Biomedicine, King's College London, London, U.K.

Clinical outcome prediction from high-dimensional data is problematic in the common setting where there is only a relatively small number of samples. The imbalance causes data overfitting, and outcome prediction becomes computationally expensive or even impossible. We propose a Bayesian outcome prediction method that can be applied to data of arbitrary dimension d, from an arbitrary number J of outcome classes, and that reduces overfitting without any approximations at parameter level. This is achieved by avoiding numerical integration or approximation, and solving the Bayesian integrals *analytically*. We thereby reduce the dimension of numerical integrals from Jd dimensions to $2J$, for any d. For large d, this is reduced further to $2J-1$, and we obtain a simple outcome prediction formula without integrals in leading order for very large d. We compare our method to the *mclustDA* method (Fraley and Raftery 2002), using simulated and real data sets. Our method performs as well as or better than *mclustDA* in low dimensions d. In large dimensions d, *mclustDA* breaks down due to computational limitations, while our method provides a feasible and computationally efficient alternative.

Keywords: discriminant analysis; Bayesian prediction; overfitting; curse of dimensionality; Bayesian integrals in high dimensions; multi-class prediction

References:

Fraley, C. Raftery A.E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of American Statistical Association* 2002; **97**(458): 611-631.