

INCREASING CLASSIFIER ACCURACY BY EXPLOITING BACKGROUND VARIABLE INFORMATION USING LPLS

S.Sæbø^{†1}, T.Almøy¹, A.H.Aastveit¹, H.Martens^{1,2,3}

¹*Norwegian University of Life Sciences, Ås, NORWAY*; ²*Matforsk, Ås, NORWAY*; ³*CIGENE, Center for Integrative Genetics, Ås, NORWAY*

[†] E-mail: *solve.sabo@umb.no*

A challenge facing scientist within areas such as medicine, bioinformatics and biology is to combine multiple sources of information in some sort of integrated data analysis. LPLS regression (LPLSR) is a new approach for integrated analysis of multiple data sources, and is an extension of regular Partial Least Squares Regression (PLSR). Whereas PLSR typically is used to study the relationship between two data tables (Y and X), LPLSR is suited for studying the patterns of covariation between three data matrices (Y, X and Z) arranged in a corner shaped system. This facilitates the potential of exploiting extra information (summarized in Z) about the variables in X (or Y). Classification of health status from gene expression data, exploiting background knowledge on genes (e.g. pathway information, gene ontology), will serve as an example for illustration. By examples involving both simulated and real data we show how relevant background information may reduce the noise-level in the data and increase classifier performance.