

**University of Leeds**  
**School of Mathematics**  
**Suggested Project titles for MATH 3752/3**  
Supervisor: Dr. Arief Gusnanto

### 1. Ridge regression for highly correlated data

The standard least-squares regression (LS) is useful for modelling if the explanatory variables  $X$  of size  $n \times p$  are approximately independent and its parameters can only be estimated if the number of variables  $p$  is greater than  $n$ . However, this 'ideal' situation may not exist in certain experiments. In chemometrics, calibration of near-infrared instruments produces data with thousands of variables, but from limited number of samples. This is exacerbated with the fact that the variables are highly correlated with correlation coefficient ranges from 0.96 to 1. With these characteristics, LS either fails due to  $n < p$ , or unstable due to high variance of the estimates. Ridge regression (RR) is one of several methods that can be employed to deal with the problem. RR works in such situation due to 'regularisation' of parameter estimation compared to that of LS. This project will explore the use of RR with applications to chemometrics data.

#### References

- Miller, J., and Miller, J. (2005). *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall
- Brereton, R. (2007). *Applied chemometrics for scientists*, John Wiley & Sons
- Hoerl, A., and Kennard, R. (1970). Ridge Regression: Biased Estimation for Non orthogonal Problems, *Technometrics*, **12**: 55–67

### 2. Partial Least Squares Regression

In standard least-squares linear regression (LS), we try to explain a relationship between a vector of response  $y$  with explanatory variables in matrix  $X$  of size  $n \times p$ . LS would be estimable if  $n > p$ . However, experiments in chemometrics and microarray produces data with a characteristic of  $n \ll p$ . In dealing with the problem of  $n \ll p$ , partial least squares regression (PLS) constructs new explanatory variables, often called factors or components, that have maximal covariance with  $y$ . We then regress  $y$  on the components. Interestingly, if  $y$  is binary, it can be shown that PLS can be used for discrimination. This project will explore the use of PLS with applications in chemometrics and microarray data.

#### References

- Miller, J., and Miller, J. (2005). *Statistics and Chemometrics for Analytical Chemistry*, Prentice Hall
- Brereton, R. (2007). *Applied chemometrics for scientists*, John Wiley & Sons

Geladi, P., and Kowalski, B. (1986). Partial least-squares regression: a tutorial, *Analytica Chimica Acta*, **185**: 1–17

Barker, M., and Rayens, W. (2003). Partial Least Squares for discrimination, *Journal of Chemometrics*, **17**(3):166–173

### 3. Multiple testing and false discovery rate in microarray data analysis

Microarray technology enables us to obtain expressions of thousands of genes simultaneously, although usually from a limited number of samples. The main objective of the experiments is usually to identify genes that are differentially expressed between two conditions or groups, such as treatment versus control. So, a test is carried out for each genes with the null hypothesis that the gene's expressions are the same between the two groups, i.e. they are 'equally expressed'. However, with thousands of tests done simultaneously, getting false positives is inevitable. To deal with this, a multiple testing adjustment has been proposed, by controlling the probability of family-wise type-I error, or family-wise error rate (FWER). However, using FWER control may be too restrictive for microarray data, where we generate biological hypotheses rather than testing them. Therefore, controlling for false discovery rate (FDR) has been seen as a reasonable method, where we control for the proportion of false positive among significant results. This project will explore the different false positives control methods with applications to microarray data.

#### References

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., Ploner, A. (2005). False discovery rate, sensitivity, and sample size for microarray studies, *Bioinformatics*, **21**, 3017–3024

Dudoit, S., Shaffer, J.P., and Boldrick J.C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, **18**(1): 71–103

Gusnanto, A., Calza, S., Pawitan, Y. (2007). Identification of differentially expressed genes and false discovery rate in microarray studies. *Current Opinion in Lipidology*, **18**(2):187-193 (available from the first author)