INTERNATIONAL CENTRE
OF EXCELLENCE FOR
EDUCATION IN
MATHEMATICS

Dimension Reduction via Ordinary Least Squares Regression
Marcus Tonelly, Department of Statistical Science, La Trobe University

The multiple index model for dimension reduction is of the form

$$y = f(\beta_1^{\mathrm{T}} x, \ldots, \beta_K^{\mathrm{T}} x, \varepsilon)$$

where *x* is a p dimensional predictor vector, $\beta_1, \ldots, \beta_K$ are unknown directions, the error term ε is independent of x, and f is the unknown link function. When $K < p$, the *p*-dimensional *x* can be replaced with the *K* dimensional $\beta_1 x, \ldots, \beta_K x$ without loss of information.

The single index model is a subclass of the multiple index model that restricts *K* to one. In this setting it is the purpose of dimension reduction methods to estimate $b = c\beta_1$ for some $c \in \Re$ (denoted $\hat{b}$) and plot the observed y_i's versus the $\hat{b}^T x_i$'s in order to determine the structural relationship between *y* and *x*.

Brillinger (1977, 1983) showed that for a normally distributed *x*, Ordinary Least Squares (OLS) regression could be used to estimate *b*. This result remained relatively hidden until it was reinvestigated by Duan & Li (1989) who expressed their surprise at its existence. Duan & Li extended the result to include less stringent conditions on *x* and other regression methods.

The purpose of my AMSI summer project was to study and provide a detailed proof, similar to that illustrated in Prendergast (2005) with respect to Sliced Inverse Regression, that OLS is an applicable method under some mild distributional conditions for *x* by looking at the structure of Cov(*x*,*y*). Varying simulated models were also considered to emphasize the usefulness of OLS applied to the single index model.

I have enjoyed the AMSI scholarship program as it has given me a real taste of what is the life of research. I will be continuing studies in this area for my honours thesis and will also be considering the robustness of single index model regressions.

References:

Brillinger, D. R. 1977. The Identification of a Particular Nonlinear Time Series System. *Biometrika*, **64**(3), 509-515.

Brillinger, D. R. 1983. A Generalized Linear Model with "Gaussian" Regression Variables. *Pages* 97-114 *of:* Bickel, P.J., Doksum, K.A., & Hodges Jr. J.L. (eds). *A Festschrift for Erik L Lehmann.* Wadsworth International Group Belmont. California.

Li, K,-C., Duan N. 1989. Regression Analysis Under Link Violation*. The Annuals of Statistics*. **17**(3), 1009-1052.

Prendergast, L.A. 2005. Influence Functions for Sliced Inverse Regression. *Scand. J. Statist.***32**, 385-404.

Marcus received an ICE-EM Vacation Scholarship in December 2005.
See www.ice-em.org.au/students.html#scholarships05