

M532 Mathematical Modeling of Large Data Sets

Problem Set Two

Due Thursday, March 13, 2008

1 Theory

1. Problem 1.5
2. Problem 2.4
3. Problem 2.17
4. Problem 2.18
5. Problem 2.25 (can use matlab)
6. Problem 3.23
7. Problem 3.26
8. Verify equation (3) in the paper by Eastment and Krzanowski.

2 Computing

Compute the (spanning) dimension of the data set 'peaks100049.mat' using the cross-validatory PCA algorithm. Include a table of your results which shows the value of i) $\text{Press}(m)$, ii) W statistic and iii) the eigenvalues λ_m of the covariance matrix of the full data matrix for $m = 1, \dots, 10$.

Propose and implement at least two methods for accelerating this basic algorithm and compare.

Extra Credit: Repeat this dimension calculation on another data set of particular interest.

For additional details on this algorithm see the paper *Cross-Validatory Choice of the Number of Components from a Principal Component Analysis*, by Eastment and Krzanowski, *Technometrics*, Vol. 24, No. 1, 73-77.