

## SUPPLEMENTAL FIGURES

Nguyen, D.V. and Rocke, D.M. (2002), "Multi-class cancer classification via partial least squares with gene expression profiles," *Bioinformatics*.

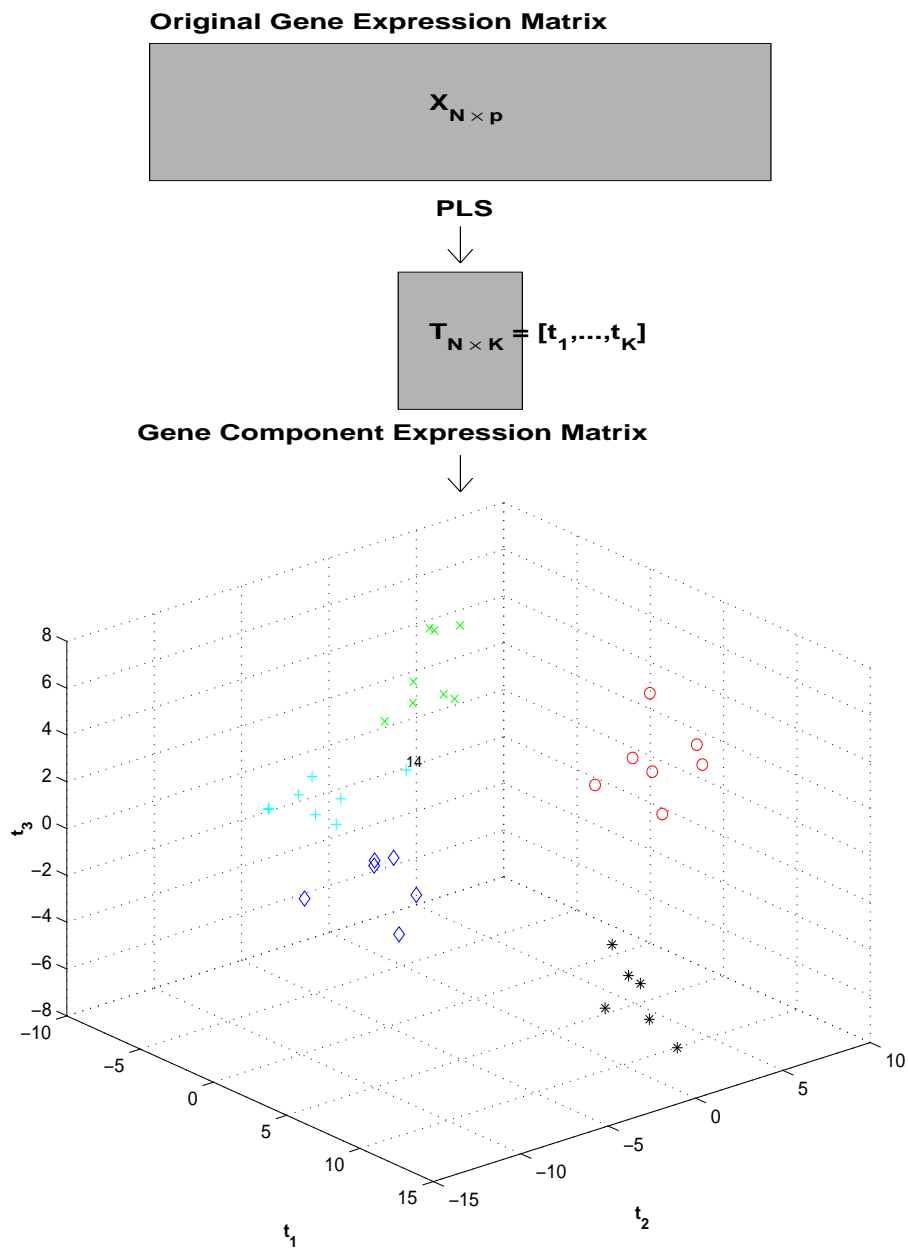


Figure 1: **Illustration of dimension reduction for NCI60 data.** For the NCI60 data, the “original” gene expression data set used here is  $\mathbf{X}_{35 \times 167}$  and  $K = 3$  PLS gene components are constructed giving  $\mathbf{T}_{35 \times 3} = [\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3]$ . The 3-dimensional PLS gene components plot illustrate the separability of the cancer classes: leukemia=\*, colon=○, melanoma=+, renal=×, and CNS=◇.

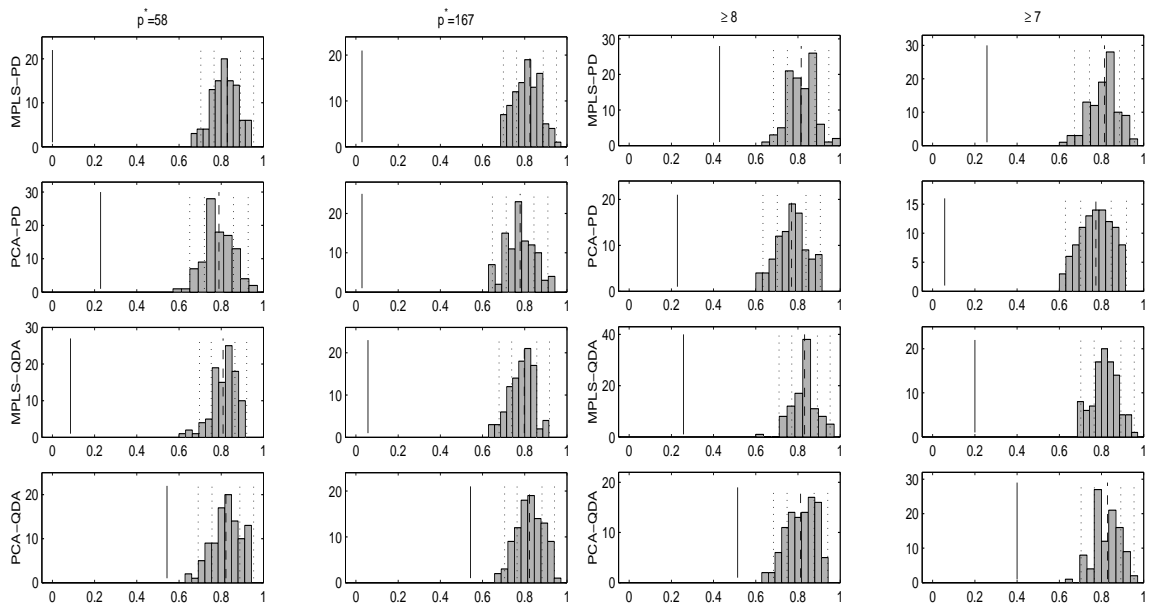


Figure 2: **Classification of NCI60 data under randomization-A1 & A2.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets using algorithm A1 (columns 1, 2) and A2 (columns 3, 4). The observed gene expression profiles were randomly assigned cancer labels (leukemia, colon, melanoma, renal, or CNS). Class sizes (the  $n_i$ s) were the same as the original data set. The corresponding observed error rates given in Table 2.A1 and Table 2.A2 are indicated by the solid lines in the histograms.

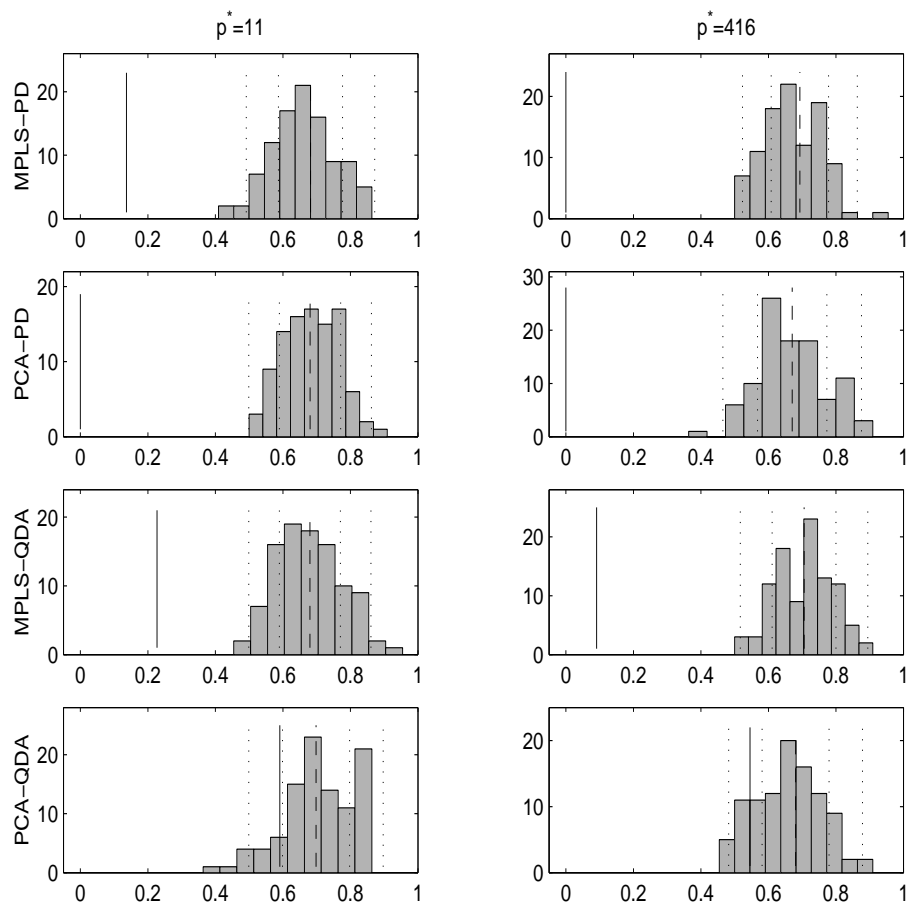


Figure 3: **Classification of breast cancer data under randomization-A1.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets (A1), where the observed gene expression profiles were randomly assigned cancer labels ( $BRCA1$ ,  $BRCA2$ , or Sporadic). Class sizes (the  $n_i$ s) were the same as the original data set. The dashed line locates the mean and the dotted lines indicate one and two standard deviations above/below the mean. The corresponding observed error rates given in Table 1.A1 is indicated by the solid lines in the histograms.

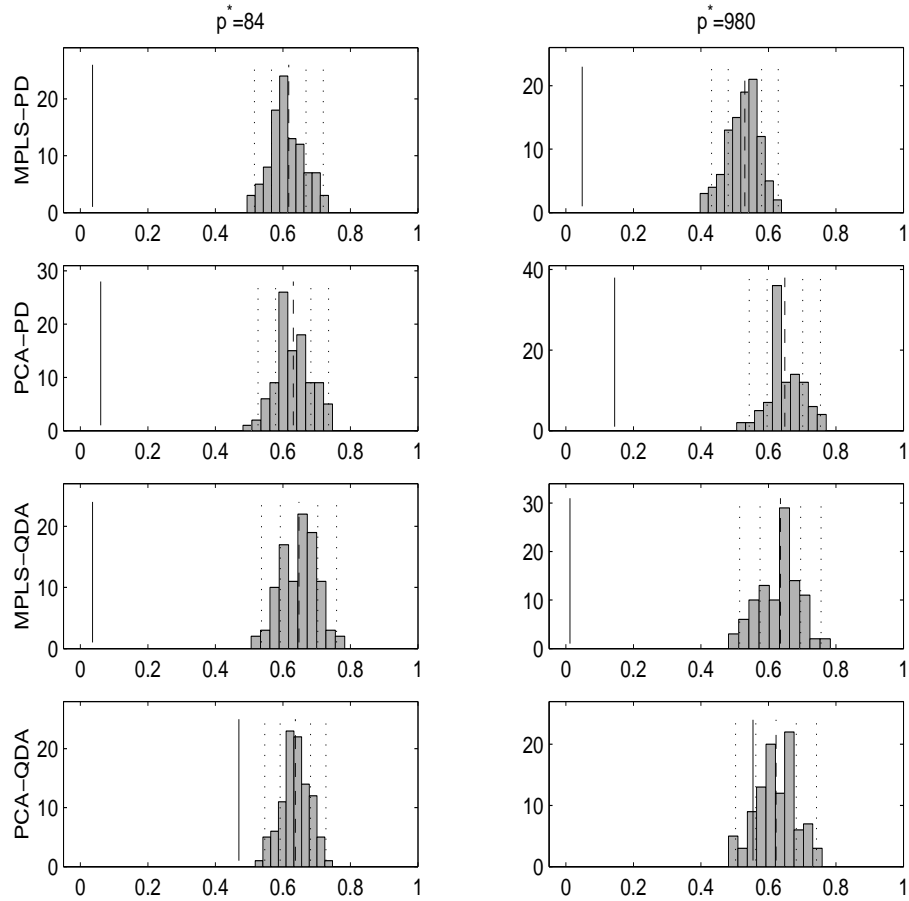


Figure 4: **Classification of lymphoma data under randomization-A1.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets (A1), where the observed gene expression profiles were randomly assigned cancer labels (DLBCL, BCLL, or FL). Class sizes (the  $n_{iS}$ ) were the same as the original data set. The corresponding observed error rates given in Table 3.A1 is indicated by the solid lines in the histograms.

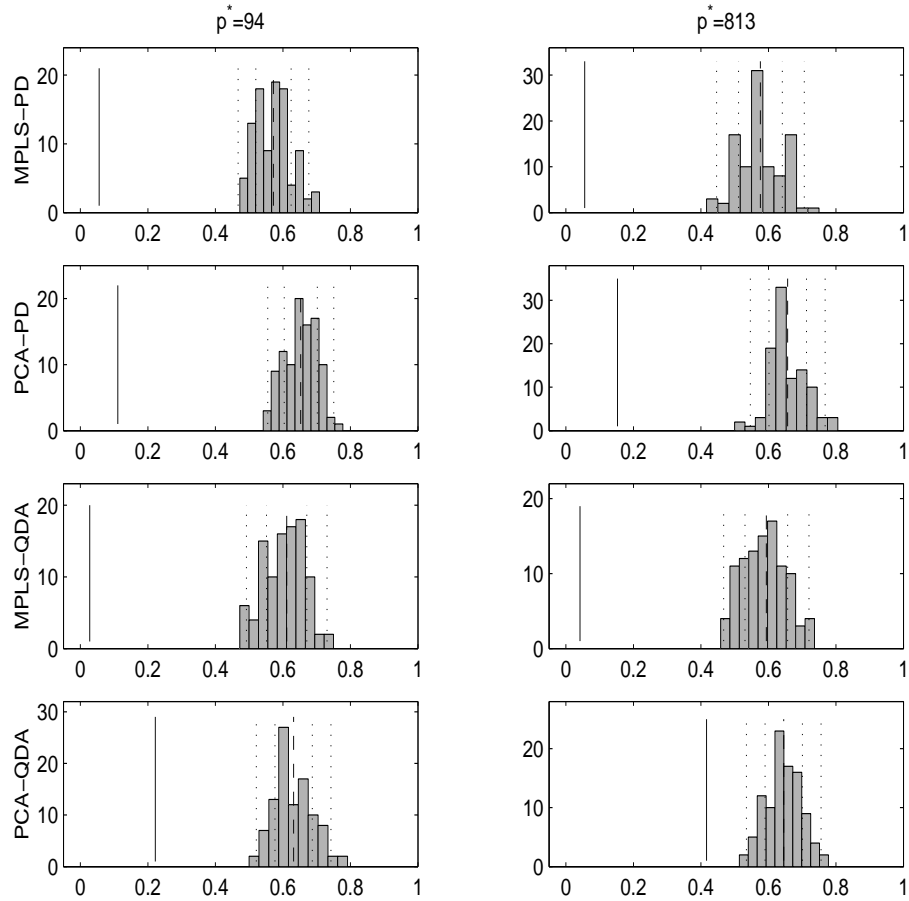


Figure 5: **Classification of leukemia data under randomization-A1.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets (A1), where the observed gene expression profiles were randomly assigned cancer labels (AML, B-ALL, or T-ALL). Class sizes (the  $n_i$ s) were the same as the original data set. The corresponding observed error rates given in Table 4.A1 is indicated by the solid lines in the histograms.

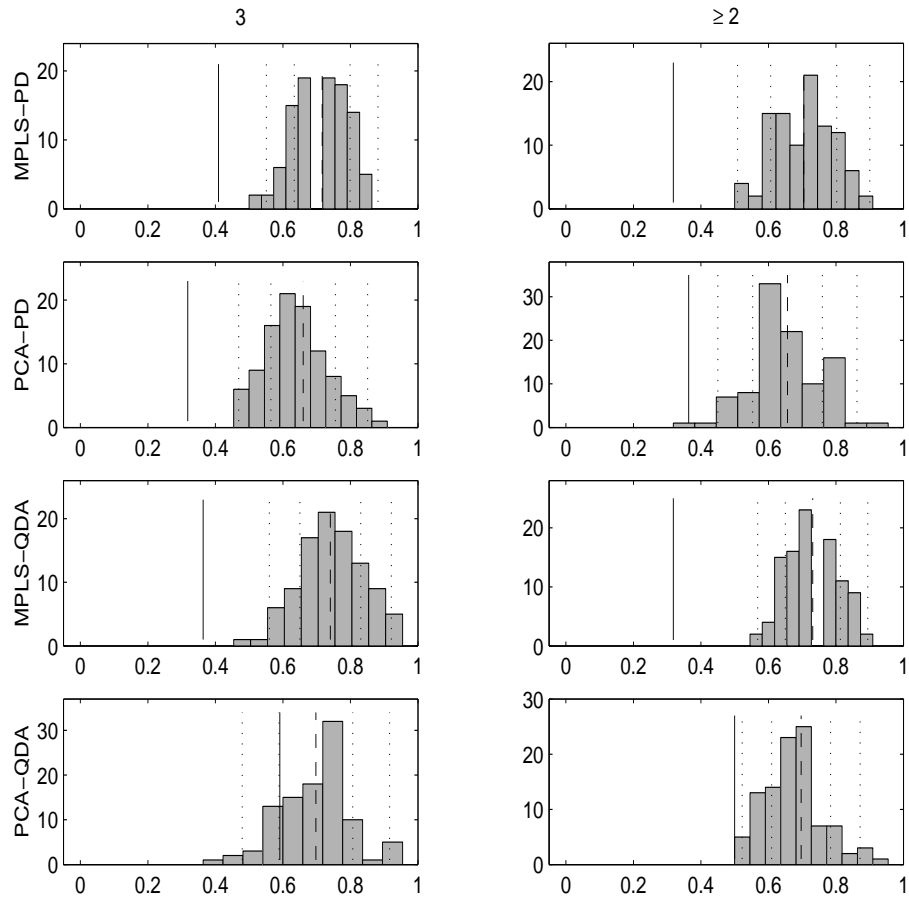


Figure 6: **Classification of breast cancer data under randomization-A2.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets (A2), where the observed gene expression profiles were randomly assigned cancer labels ( $BRCA1$ ,  $BRCA2$ , or Sporadic). Class sizes (the  $n_i$ s) were the same as the original data set. The corresponding observed error rates given in Table 1.A2 is indicated by the solid lines in the histograms.

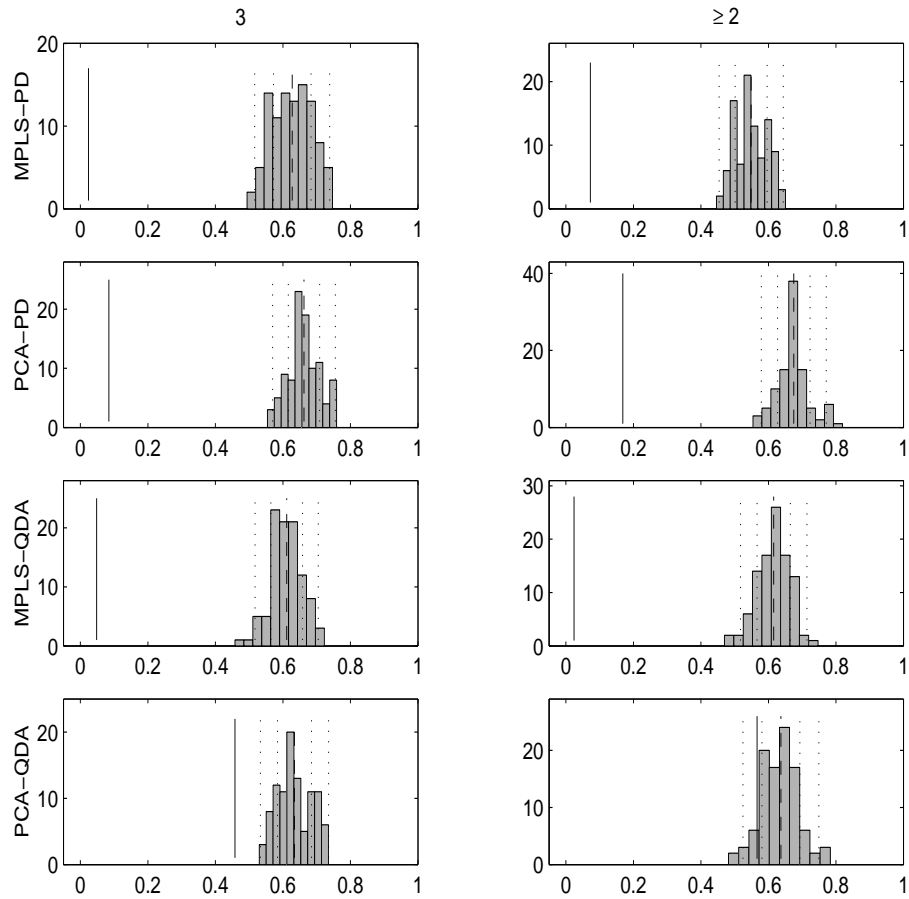


Figure 7: **Classification of lymphoma data under randomization-A2.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets (A1), where the observed gene expression profiles were randomly assigned cancer labels (DLBCL, BCLL, or FL). Class sizes (the  $n_{iS}$ ) were the same as the original data set. The corresponding observed error rates given in Table 3.A2 is indicated by the solid lines in the histograms.

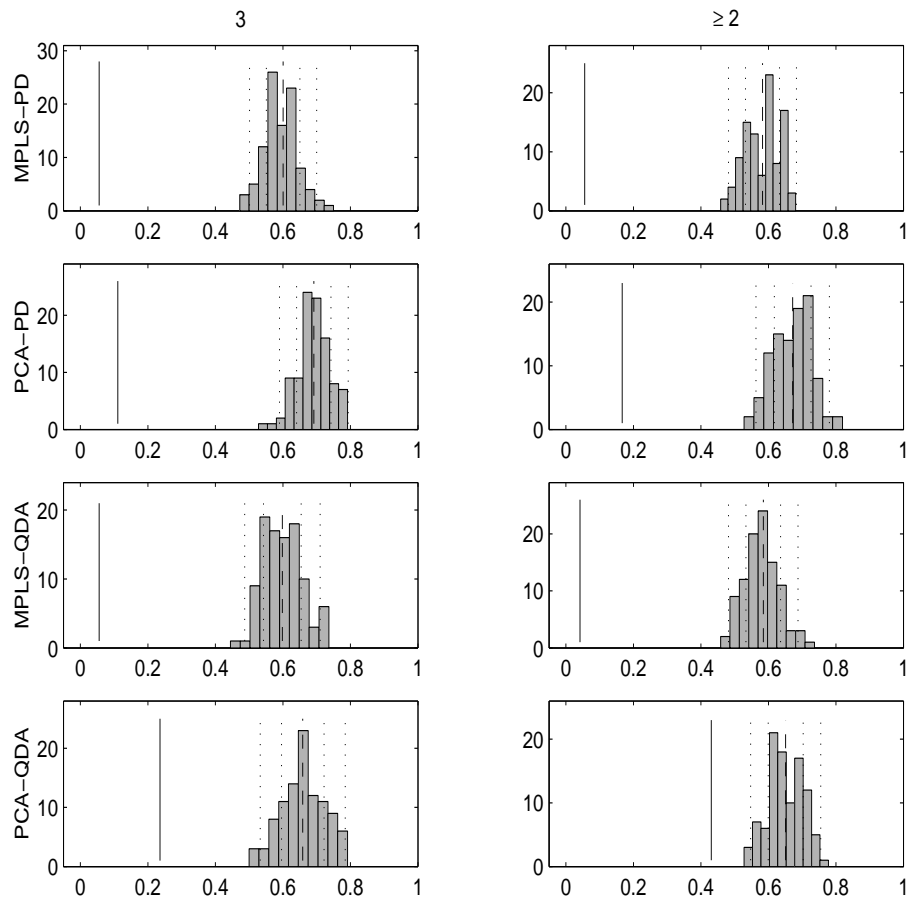


Figure 8: **Classification of leukemia data under randomization-A2.** Each histogram is of  $B = 100$  classification error rates from  $B$  randomized data sets (A1), where the observed gene expression profiles were randomly assigned cancer labels (AML, B-ALL, or T-ALL). Class sizes (the  $n_i$ s) were the same as the original data set. The corresponding observed error rates given in Table 4.A2 is indicated by the solid lines in the histograms.

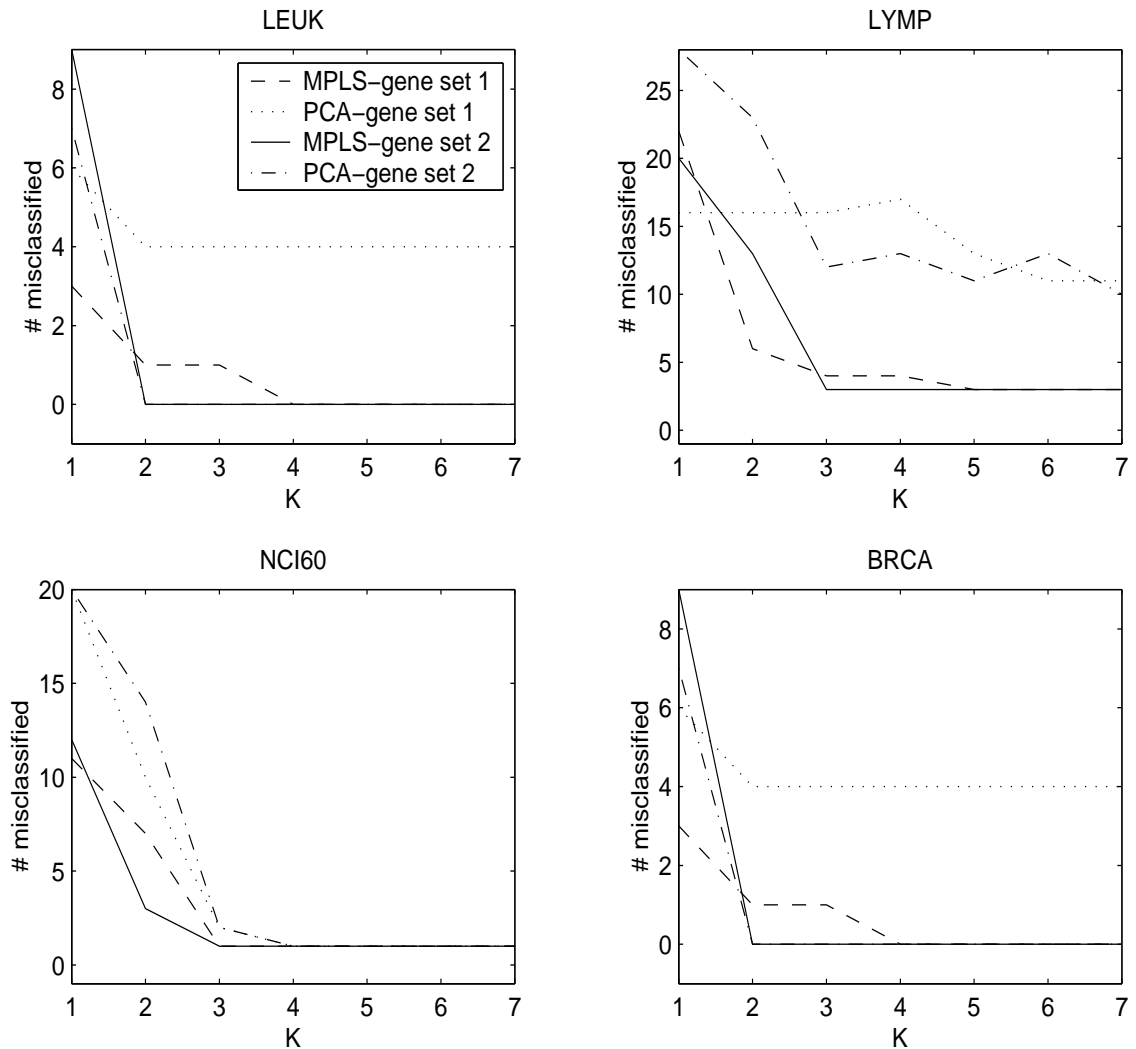


Figure 9: **Classification results for various dimension  $K$ .** Given on each plot is # of misclassification as the dimension,  $K$ , varies from 1 to 7. The plots suggests the choice of  $K = 3$  is reasonable for PLS dimension reduction. Results are given for algorithm A1 with classifier DLDA. Gene sets 1 and 2 refers to the set of 94 and 813 genes for the leukemia data, for example. Refer to Tables 1-3 for details on the gene sets 1 and 2 for the breast cancer, NCI60, and lymphoma data sets respectively.