**CH 5440 Multivariate Data Analysis in Process Monitoring and Diagnosis**

Assignment 5

Consider the problem of developing a correlation between pressure and saturated temperature (boiling point). For pure components, the Antoine equation given below generally fits the data well



For n-hexane, the values of the constants are A = 14.0568, B = 2825.42, and C = 230.44 where Psat is given in kPa and T in deg C. Using this correlation a data set consisting of 100 samples have been generated in the temperature range 10 - 70 deg C. Gaussian measurements errors to both the true temperature and saturated pressures with standard deviations of 0.18 deg C and 2 kPa, respectively, have been added to generate the measurements (available in *vpdata.mat*)

1. Apply Kernel PCA to obtain a nonlinear correlation between saturated pressure and temperature using Gaussian Kernels. Use the first 70 samples for developing the model and the remaining 30 samples for cross validation. Choose the optimal hyper-parameters (Kernel width as well as the number of PC's in feature space to be chosen) using PRESS on the cross validation samples. It is better to first shift the temperature measurements using the mean and scale using the maximum range or the standard deviation of the temperature before applying the nonlinear transformation. Note that you should shift and scale the temperature as you do for the training set before using the KPCR model for predicting the saturated pressures for the test data set.
2. Test the accuracy of predicting the saturated pressures using the above KPCA model for a temperature of 55 deg C and temperature of 100 deg C and compare with the values obtained using the Antoine equation. What are your observations from the predicted saturated pressures? Provide reasons for the same.

*Answers*

*(a) Kernel width of 5 and number of PCs 5 give the minimum PRESS of around 10.5*

*(b) For T = 55 deg C, estimated psat = 64.05 kPa and true psat = 63.96 kPa. For T = 100 deg C, estimated psat = 2569 kPa and true psat = 246.24 kPa. KPCR models (like most nonlinear regression models) is good for interpolation but not good at extrapolation.*