Technical Summary

Recent advances in theory and practice, have introduced a wide variety of tools from machine
learning that can be applied to data intensive chemical engineering problems. This thesis covers
applications of statistical learning spanning a range of relative importance of data versus existing
detailed theory. In each application, the quantity and quality of data available from experimental
systems are used in conjunction with an understanding of the theoretical physical laws governing
system behavior to the extent they are available.

A detailed generative parametric model for optical spectra of multicomponent mixtures is intro-
duced. The application of interest is the quantification of uncertainty associated with estimating
the relative abundance of mixtures of carbon nanotubes in solution. This work describes a detailed
analysis of sources of uncertainty in estimation of relative abundance of chemical species in solution
from optical spectroscopy. In particular, the quantification of uncertainty in mixtures with para-
metric uncertainty in pure component spectra is addressed. Markov Chain Monte Carlo methods
are utilized to quantify uncertainty in these situations and the inaccuracy and potential for error in
simpler methods is demonstrated. Strategies to improve estimation accuracy and reduce uncertainty
in practical experimental situations are developed including when multiple measurements are avail-
able and with sequential data. The utilization of computational Bayesian inference in chemometric
problems shows great promise in a wide variety of practical experimental applications.

A related deconvolution problem is addressed in which a detailed physical model is not available,
but the objective of analysis is to map from a measured vector valued signal to a sum of an unknown
number of discrete contributions. The data analyzed in this application is electrical signals generated
from a free surface electro-spinning apparatus. In this information poor system, MAP estimation is
used to reduce the variance in estimates of the physical parameters of interest. The formulation of
the estimation problem in a probabilistic context allows for the introduction of prior knowledge to
compensate for a high dimensional ill-conditioned inverse problem. The estimates from this work are
used to develop a productivity model expanding on previous work and showing how the uncertainty
from estimation impacts system understanding.

A new machine learning based method for monitoring for anomalous behavior in production oil
wells is reported. The method entails a transformation of the available time series of measurements
into a high-dimensional feature space representation. This transformation yields results which can
be treated as static independent measurements. A new method for feature selection in one-class
classification problems is developed based on approximate knowledge of the state of the system.
An extension of features space transformation methods on time series data is introduced to handle
multivariate data in large computationally burdensome domains by using sparse feature extraction
methods.

Thesis Supervisor: Richard D. Braatz
Title: Edwin R. Gilliland Professor