APPROXIMATING COVARIANCE MATRICES USING LOW RANK PERTURBATIONS WITH APPLICATIONS TO ACCENT IDENTIFICATION AND SOCIAL NETWORK CLUSTERING

By

Jonathan Purnell

An Abstract of a Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

The original of the complete thesis is on file
in the Rensselaer Polytechnic Institute Library

Examining Committee:

Malik Magdon-Ismail, Thesis Adviser

Sanmay Das, Member

Mark Embrechts, Member

Mohammed Zaki, Member

Rensselaer Polytechnic Institute
Troy, New York

August 2010
(For Graduation December 2010)

ABSTRACT

In this work, we present a new model, the Low-Rank Gaussian Mixture Model (LRGMM), for modeling data which can be extended to identifying partitions or overlapping clusters. This model is motivated by the effectiveness, yet limited scalability, of the Gaussian Mixture Model (GMM) for the problem of accent identification. The curse of dimensionality that arises in calculating the covariance matrices of the GMM is countered by using low-rank perturbed diagonal matrices. We also demonstrate the LRGMM for finding communities in social networks. We see that the efficiency of the LRGMM allows us to process larger networks than alternative approaches. Altogether, the LRGMM experiments reveal it to be a efficient and highly applicable tool for working with large high-dimensional datasets.