Mixture of Factor Analyzers (MoFA) Models for the Design and Analysis of SAR Automatic Target Recognition (ATR) Algorithms

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Tarek Abdel-Rahman, B.S., M.S.

Graduate Program in Electrical and Computer Engineering Department

The Ohio State University

2017

Dissertation Committee:

Emre Ertin, Advisor
Randolph Moses
Bradley Clymer
Thomas Santner
Abstract

We study the problem of target classification from Synthetic Aperture Radar (SAR) imagery. Target classification using SAR imagery is a challenging problem due to large variations of target signature as the target aspect angle changes. Previous work on modeling wide angle SAR imagery has shown that point features, extracted from scattering center locations, result in a high dimensional feature vector that lies on a low dimensional manifold. We propose to use rich probabilistic models for these target manifolds to analyze classification performance as a function of Signal-to-noise ratio (SNR) and Bandwidth. We employ Mixture of Factor Analyzers (MoFA) models to approximate the target manifold locally, and use error bounds for the estimation and analysis of classification error performance. We compare our performance predictions with the empirical performance of practical classifiers using simulated wideband SAR signatures of civilian vehicles.

We then extend this work to design optimal maximally discriminative projections (MDP) for the manifold structured data. An optimization algorithm is proposed that maximizes the Kullback Leibler (KL)-divergence between two mixture models through optimizing the
closed-form "Variational Approximation" of the KL-divergence between the MoFA models. We then propose to generalize our MDP dimensionality reduction technique to multi-class using non-linear constrained optimization through minimax quasi-Newton methods. The proposed MDP algorithm is compared to existing dimensionality reduction techniques using simulated Civilian Vehicles datadome dataset and real-world MSTAR data.
This is dedicated to my lovely, supporting parents...
Acknowledgments

I would like to thank and express gratitude to my advisor Prof. Emre Ertin, for his continuous and generous support, patience and guidance throughout my PhD. His immense knowledge, intelligence and guidance helped me throughout my journey. I could not have imagined having a better advisor and mentor for my PhD study.

I would also like to thank Prof. Lee Potter and Prof. Randolph Moses for their fruitful comments and suggestions that contributed significantly to my research and dissertation document; and Prof. Bradley Clymer for his outstanding teaching and support during my PhD.

I am very thankful to many friends and colleagues in the field that supported me with suggestions and encouragement during my journey. Particularly, I would like to thank Nithin Sugavanam, Karim Khalil, Shehab Yomn and Siddharth Baskar.

Finally, I would like to extend my deepest gratitude and love to my parents and sister for their love and support.
Vita

July 8, 1988 ................................................................. Born - Giza, Egypt

2010 ................................................................. B.S. Electronics and Communications Engineering, Cairo University

2013 ................................................................. M.S. Electrical and Computer Engineering, The Ohio State University

2013-present ....................................................... Graduate Research Associate, The Ohio State University.

Publications

Research Publications

T. Abdelrahman “Multi-resolution tracking using dyadic trees”. OhioLINK Electronic Theses and Dissertations Center, The Ohio State University, Columbus, Ohio, Jul 2013.


Fields of Study

Major Field: Electrical and Computer Engineering Department
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>ii</td>
</tr>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>x</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xi</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Background and Problem Statement</td>
<td>1</td>
</tr>
<tr>
<td>1.1.1 Synthetic Aperture Radar Imaging</td>
<td>1</td>
</tr>
<tr>
<td>1.1.2 Automated Target Recognition</td>
<td>6</td>
</tr>
<tr>
<td>1.2 Statement of Contributions and Dissertation Outline</td>
<td>10</td>
</tr>
<tr>
<td>2. Modeling SAR target signatures using Mixture of Factor Analyzers</td>
<td>14</td>
</tr>
<tr>
<td>2.1 Point feature extraction from SAR imagery</td>
<td>14</td>
</tr>
<tr>
<td>2.2 Mixture of Factor Analyzers (MoFA) models</td>
<td>17</td>
</tr>
<tr>
<td>2.3 Out-of-library Rejection Performance for MoFA Models</td>
<td>21</td>
</tr>
<tr>
<td>3. Classification Performance using Theoretical Error Bounds</td>
<td>30</td>
</tr>
<tr>
<td>3.1 Theoretical Performance Bounds of the Bayes Classification Error</td>
<td>31</td>
</tr>
<tr>
<td>3.2 High SNR Approximation of the Error Bounds</td>
<td>34</td>
</tr>
<tr>
<td>3.3 Simulation Results</td>
<td>37</td>
</tr>
</tbody>
</table>
Bibliography
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1 Table comparing processing time of Multi-Class MDP in terms of mixture components per class</td>
<td>70</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>The SAR image shows the reflectivity of the Ohio Stadium, (The Shoe), and surrounding campus, to radio waves at 10 GHz. All rights reserved to General Dynamics Advanced Information Systems.</td>
</tr>
<tr>
<td>1.2</td>
<td>Ka-band image of C-130s on flight line at Kirtland AFB (Courtesy of Sandia National Laboratories, Airborne ISR).</td>
</tr>
<tr>
<td>1.3</td>
<td>Flow Chart of a Feature-Based SAR ATR System</td>
</tr>
<tr>
<td>2.1</td>
<td>Block Diagram showing the Feature Extraction Algorithm</td>
</tr>
<tr>
<td>2.2</td>
<td>Feature Extraction of Civilian Vehicle Data Domes</td>
</tr>
<tr>
<td>2.3</td>
<td>Illustrative diagram of modeling SAR feature images using MoFA Models</td>
</tr>
<tr>
<td>2.4</td>
<td>MSTAR SAR Imagery</td>
</tr>
<tr>
<td>2.5</td>
<td>MSTAR Out-of-library Rejection Performance of the MoFA models based on Maximum Conditional Likelihood Classification compared to K-NN Classification, 16 Mixture Components per class are considered</td>
</tr>
<tr>
<td>2.6</td>
<td>Plotting Receiver Operating Characteristic (ROC) curves of the MSTAR MoFA Model based on Maximum Conditional Likelihood Classification compared to K-NN Classification, 16 Mixture Components per class are considered</td>
</tr>
<tr>
<td>2.7</td>
<td>Determining the Optimal Number of Mixture Components per Class based on the $P_D = 0.9$ Probability of Detection standard operating point</td>
</tr>
<tr>
<td>2.8</td>
<td>Determining the Optimal Number of the MoFA Intrinsic Dimension based on the $P_D = 0.9$ Probability of Detection standard operating point</td>
</tr>
</tbody>
</table>
3.1 Upper bound on misclassification error vs SNR for 1 GHz Bandwidth using different number of mixture components. Empirical performance of the kNN and LDA classifiers are superimposed. .......................... 38

3.2 Upper bound on misclassification error vs SNR for a bandwidth of 0.5, 1 and 2 GHz. ................................................................. 39

3.3 Upper bound on misclassification error vs number of principal components per mixture for fixed SNR and Bandwidth. ......................... 40

3.4 Dividing Feature Image Data of Civilian Vehicle Data Domes into $M$ Mixtures equispaced in Azimuth ................................. 42

4.1 Comparing the Variation Approximation and Union Bound of the KL divergence of the mixture models of Camry vs Jeep classes .......... 54

4.2 Comparing the Variation Approximation of the KL divergence of the mixture models of Camry class vs the rest of the dataset using the MDP Optimal Projector compared to various dimensionality reduction techniques, 4 mixture components per class .......................... 57

4.3 Comparing Misclassification Error using SVM Classification of Camry Class using the MDP vs PCA projector using different proportionally-sized training samples ........................................ 58

4.4 Comparing Probability of Detection using SVM Classification of Camry Class using the MDP vs PCA projector using different proportionally-sized training samples ........................................ 59

5.1 Multi-Class MDP classification performance compared to other dimensionality reduction techniques of MSTAR data using separate depressions in training and testing ........................................ 69

5.2 Zoomed view of Multi-Class MDP classification performance with various different number of mixture components per class cases at NuMax’s RIP range .................................................. 70

5.3 Worst-Case Divergence Comparison of Multi-Class MDP versus various Dimensionality Reduction techniques .......................... 71
Chapter 1: Introduction

1.1 Background and Problem Statement

1.1.1 Synthetic Aperture Radar Imaging

Synthetic aperture radar (SAR) based automatic target recognition (ATR) serves as a powerful tool in intelligence, surveillance, and reconnaissance (ISR) applications. SAR is an attractive choice over other sensor modalities for long range surveillance applications due to its robustness to atmospheric and illumination conditions. SAR, first introduced by Wiley in [1], provides an imaging dimension to the radar sensing previously reserved for detection and tracking using one dimensional echo returns. Civilian applications of SAR include natural disaster assessment, forestry, and oceanography. For SAR imaging, a radar sensor collects one dimensional backscatter returns over a flight trajectory and coherently combines these one dimensional returns to produce an estimate of the location and scattering coefficients of the reflectors in the scene. Two or three dimensional estimates of the reflector locations can be produced based on the geometry of the flight trajectory. The SAR image in figure 1.1 shows the reflectivity of the Ohio Stadium, commonly known as "The Shoe", to radio waves at 10 GHz. Figure 1.2 is another example of a SAR image collected
A classical problem in any imaging modality is to achieve fine resolution for obtaining images that could be used for making inferences of the scene interrogated by the sensor. For radar, high range resolution is proportional to the bandwidth of the sensor, which can be extended using pulse compression techniques [28], that modulates an RF carrier to sweep over a large bandwidth. However, achieving comparable cross-range resolution using a
stationary radar array is much more challenging, as improving cross-resolution requires either a higher frequency of operation or use an impractical aperture comprising of multiple transmitters and/or receivers. SAR overcomes this problem by constructing a large virtual phased array. SAR “synthesizes” a large aperture by operating a radar platform comprising of a collocated transmitter and receiver by sequentially moving this radar platform through successive positions of a larger array [43].

Figure 1.2: Ka-band image of C-130s on flight line at Kirtland AFB (Courtesy of Sandia National Laboratories, Airborne ISR).
SAR operates in different modes mainly depending on the relative geometry of the radar platform relative to the scene center. A *sidelooking strip-map SAR* is locked into the side-looking position, and the array face is perpendicular to the flight path. Although the strip-map covers a larger surveillance region, the cross-range resolution is determined by the beam width of the illuminating antenna. This limitation in cross-range resolution can be compensated by keeping the scene under interrogation within the beam width of the physical antenna element and rotating the radar platform around the region of interest [43]. This mode of operation is termed as *spotlight* SAR.

In our work, we focus on the *spotlight* mode of operation for the radar sensor, which was first introduced by John Walker [3]. The *spotlight* mode implies that the antenna focuses on the scene center during the flight path of the radar platform. Based on the extent of the aperture synthesized in the azimuth angle domain, we have two classes of measurement schemes namely- normal mode and wide-angle mode [34]. If the azimuth angle moved by the platform covers the entire 360 degrees in azimuth, then the sensor operates in the wide-angle mode also named as circular SAR. Our work focuses on the wide-angle mode of operation since it produces a diverse set of measurements or views of the object, which is ideal for the task of target recognition [10]. Wide-angle signature of a target is encoded in complex-baseband images obtained at different center azimuth angles. The platform typically transmits multiple pulses, and the received back-scattered energy from the scene also termed as phase history measurements. For a narrow synthetic aperture the
phase history measurements can be directly linked to samples of the spatial Fourier transform of the scene. Therefore, the problem of estimating scene reflectivity can be posed as a non-uniformly sampled inverse Fourier problem. Many inversion algorithms have been proposed in the literature based on various interpolation/Fourier inversion techniques or time domain inversion methods termed as back-projection algorithms that directly invert the associated Fourier integral equations [50] using adjoint operators. The relation between the phase history measurements and spatial Fourier transform of the scene does not hold for wide-angle aperture collections, as reflectors in the scene do not persist for the wide angle and only provide backscatter energy for part of the aperture. For the case of wide-angle measurements, the measurements can be divided into sub-apertures, where scene reflectivity remains constant. The images are formed over these sub-apertures coherently processes phase history measurements restricted to a specified small azimuth extent. This method of partitioning yields a sequence of images of the scene. In a surveillance application where the scene includes targets of interest, these sequence images which provide a rich description of the target as a function of spatial coordinates and the pose angle of the target with respect to the radar platform. We note that target wide-angle signatures consists of high resolution images indexed by pose angle result in high dimensional data structure of size \(10^6\) dimensions. Since the underlying target phenomena has much fewer degrees of freedom, the wide-angle target signature lives on a low dimensional manifold of much smaller intrinsic dimension. In particular, Ertin [13] explored manifold learning methods on SAR feature vectors that are invariant to phase variations introduced by nuisance parameters and
showed that point features extracted from SAR imagery are well approximated by a low
dimensional nonlinear manifold.

Our work in this dissertation will first focus on statistical modeling of these high-
dimensional wide-angle target signatures exploiting the low dimensional manifold struc-
ture [2]. Next, we use these statistical models for the analysis of SAR ATR performance
under various operating conditions and for synthesizing novel dimensionality reduction
techniques that outperform classical and state-of-the art techniques in classification tasks.

1.1.2 Automated Target Recognition

Target identification from SAR imagery is a challenging task due to the high variabil-
ity of target signatures as a function of aspect. Conventional methods in SAR ATR were
based on measuring the consistency of several target types and the pose compared to a
given test image. Mean-square error (MSE) between the given test image and all in-library
hypothesized target reference images at the hypothesized pose is typically used as a metric
for measuring the accuracy. Templates are constructed from the image data by applying
transformations, which are used as references for comparing it with the test data. These
reference images are commonly known to as “templates” and therefore arises the name,
“template-based methods” or “image-based methods”. MIT Lincoln Laboratory utilized
this approach in the classification stage of several ATR systems [36], [37]. Template-based
methods were also utilized by the DARPA/AFRL Moving and Stationary Target Acqui-
sition and Recognition (MSTAR) program, which has provided a unique breakthrough to
assess the progress in SAR ATR algorithm development [44]. Even though, template-based methods are easier to apply, they suffer from significant problems.

One of the main drawbacks of template-based methods is that the test images are compared to reference templates that are processed to eliminate background, clutter and radar-shadow regions. The MSE metric, in specific, performs poorly with reference templates that are constructed from model-based predictions and, furthermore, radar shadow information is not utilized, which leads to a degradation in performance due to loss of essential information. For the target in white additive noise model, template-based methods provide a simple detector structure. In many cases, the training data is limiting the applicability of constructing representative templates. These template-based classifiers perform poorly in the presence of model mismatch due to occlusion and discrete clutter.

"Feature-based methods” are more accurate and provide more robust performance for the task of target recognition. Instead of comparing the different hypothesis directly on the image space, features are first extracted from raw images and target recognition is performed in the feature domain. Irving [26] considered feature extraction through Quantized Grayscale Matching (QSM). Unlike, the traditional approach of template matching using a mean-square error (MSE) metric, Irving quantizes the pixel values, and subsequently fits statistical models and then utilizes them to perform a maximum likelihood test over all possible target hypothesis.

Even though emerging feature based ATR methods have shown to provide immunity to discrete clutter and occlusions [38] [10] [27], the performance prediction of feature-based
methods are complicated by the lack of statistical models of feature variability as target features are masked, or non-target features are included in the feature vector under test. Modeling and learning these variations from the training data is further complicated due to the high dimensionality of feature vectors. To further elaborate, figure 1.3 provides a high-level flow chart of a feature-based SAR ATR system.

Initially, phase history data gathered by the synthetic-aperture radar is extracted using SAR Image formation algorithms to produce complex SAR images. These images are normalized to achieve dynamic range consistency for all the images. For example, if the dynamic range of an image is between 50 to 180 and the desired range is between 0 and 255, we subtract 50 from each pixel image transforming the range from 0 to 130. Next, we multiply each pixel by $\frac{255}{130}$, making the range 0 to 255, where we can refer to as dynamic range expansion. Next, windowing is performed to reduce the effect of undesired side lobe and thresholding may be performed to reduce the effect of background noise.

After pre-processing, we input our data to the Feature Extraction algorithm to extract the informative features that describe the variations in the data. We will explain this in more detail in the next chapter. The feature vectors, obtained using the feature extraction algorithm, are highly correlated, which deteriorates the efficiency of learning the classifier. Therefore, we reduce the dimensionality of these feature vectors, while maintaining a suitable performance statistic such as divergence measures or distances between the feature vectors of the different classes (Chapter 4). This reduction in dimensionality can substantially reduce the computational complexity of classifiers and can lead to improvement in
the performance determined by the generalization error. However, many of the divergence measures utilized in dimensionality reduction, require parametric models and statistical fitting, therefore, statistical modeling of SAR feature images is performed (Chapter 2) and exploited to optimize a projection matrix that’s goal is to preserve minimizing the misclassification probability of error through class-discrimination.

After applying dimensionality reduction to the feature based models, we train the classifier in the feature space. Typically, the classifiers have certain parameters that control the complexity. A general rule in statistical learning theory is to penalize the complexity of the classifier to avoid over-fitting to the training samples and improve the generalization performance of the proposed method. A classical method to avoid over-fitting is called as cross-validation. Cross-validation with $k$-folds [19], is an example of a cross-validation technique performed, to extract the optimized parameters of the learned classifier. After the classification model is produced, it will be utilized in the prediction stage to classify class labels. The testing images, similar to training, will be feature extracted and then its dimensionality will be reduced, using the projection matrix produced from the dimensionality reduction algorithm during training. After we have our final form of low-dimensional feature-based test images, prediction and classification will be performed on them, based on the computed scores using the statistical model of each class.
1.2 Statement of Contributions and Dissertation Outline

In this dissertation, we first derive statistical models for wide-angle target signatures exploiting the low dimensional manifold structure of the target response in high dimensional measurements in image domain. Specifically, we construct Mixture of Factor Analyzer (MoFA) models of low-dimensional target manifolds to produce a generative probabilistic model, with explicit latent variables encoding pose information. These generative probabilistic target models can be used for both analysis of classification performance of ATR systems and synthesis of new building blocks for ATR systems.

Second, we apply these MoFA models to an important problem common to all ATR systems. The problem, "out-of-library signature rejection", of discriminating between targets used in the modeling phase from targets not present in the training data, is crucial for improving robustness of the system by rejecting signatures not well represented in the training data set. We use a simple threshold based rule on rejecting measurements whose statistical distance to the nearest target model is large. We characterize the performance of the proposed technique empirically using the MSTAR database [35].

Third, we use MoFA models of wide angle target signatures in SAR ATR classification performance prediction under various operating conditions, such as bandwidth and signal-to-noise ratio. We employ multi-class Bayes error bounds to analyze the classification performance between target classes as a function of SNR. We compare our theoretical
predictions with empirical results from practical algorithms. We also provide an asymptotic theoretical analysis for high SNR case.

In the remaining of the dissertation, we explore manifold learning techniques for constructing novel dimensionality reduction algorithms. Dimensionality reduction attempts to map the original data space to a small number of dimensions without losing information about the underlying structure. Linear dimensionality reduction techniques are the main basis for analyzing high dimensional data because of their simplicity in terms of geometrical interpretations and lower relatively computational requirements [9]. Examples of linear dimensionality techniques include Principal Component Analysis (PCA) [47] [12], Multidimensional Scaling (MDS) [8] [53] [6], Linear Discriminant Analysis (LDA) [14] [41], Locality Preserving Projections (LPP) [21] [20], Independent Component Analysis (ICA) [25], Distance Metric Learning (DML) [30] and probabilistic PCA [52] [45] [49]. However, linear methods of dimensionality reduction often perform poorly if the underlying data space is a nonlinear manifold. Manifold learning techniques [21] [46] [48] [51] [5] have provided methods of solving the high dimensionality problem while retaining information of the latent variables of interest. Features extracted from SAR image chips are typically high dimensional vectors with redundant and correlated features.

Training SAR ATR algorithms directly on these high dimensional signatures lead to poor generalization performance when the number of training image chips is limited. Dimensionality reduction techniques attempt to map the data to small number of dimensions where variations of the feature vector due to latent variables of interest (e.g. class and pose)
are preserved. Direct dimensionality reduction techniques to complex valued imagery, of SAR imagery, is problematic due to nuisance parameters (layover, range migration, etc.) that affect the absolute phase of each pixel.

We pursue an alternative strategy of extracting generative mixture models from training data and optimizing parameters of a linear projection matrix to maximize a statistical measure of discriminability between two MoFA models. Next, we extend our algorithm termed as Maximally Discriminative Projections to the multi-class class case by optimizing a minimax formulation of multi-class discriminability using nonlinear constrained minimax quasi-Newton methods.

The dissertation is structured as follows. In chapter 2, we review the details of the feature extraction algorithm used in the analysis of SAR imagery and present our modeling strategy based on MoFA models. We also discuss mixture model fitting procedure for estimating models from training data and study out-of-library rejection performance. Next, in chapter 3, we review multi-class Bayes error bounds and study and analyze classification performance using the Bhattacharyya error bound and derive a high-SNR approximation of these error bounds. In chapter 4, we link mutual information to misclassification error probability through divergence bounds where we propose our novel dimensionality reduction algorithm: Maximally Discriminative Projection. Next, we extend our MDP algorithm to multi-class in chapter 5 using non-linear constrained quasi-Newton methods. Finally, we conclude and summarize in chapter 6 and propose future work.
Figure 1.3: Flow Chart of a Feature-Based SAR ATR System
Chapter 2: Modeling SAR target signatures using Mixture of Factor Analyzers

In the previous chapter, we described the challenges of implementing a SAR based ATR system and highlighted the advantages of feature based methods over template based methods having high immunity against clutter, occlusions, and changes in radar-target geometry. We described the lack of statistical models that fit these feature-based methods and the curse of high dimensionality that complicates learning classifiers from training data. In this chapter, we will present our feature-based method and model the extracted SAR imagery features using low-ranked statistical model termed as Mixture of Factor Analyzers (MoFA). Finally, we test the performance of MoFA model’s for the task of detecting in-library and rejecting out-of-library testing images.

2.1 Point feature extraction from SAR imagery

Feature-based methods for SAR ATR attempt to provide stable signatures for targets by extracting target features that are invariant to nuisance parameters such as absolute range, phase errors, speckle and layover effects. Attributed scattering center models [38] use physics-based predictions of canonical scattering structures (trihedral, dihedral, dipole, and
curved surfaces) to decompose the SAR image into primitive scattering mechanisms. Both the relative location and type of these scattering centers provide information about target type and pose. We focus on a single scattering mechanism, point (corner) reflector and model the information embedded in the spatial distribution of point reflectors of the target image. We consider a point process model with a non-uniform density, which encodes the spatial distribution of reflectors that are part of the target. Our feature extraction algorithm serves as an example of the proposed modeling strategy and can be extended to more sophisticated feature extraction algorithms that provide additional attributes such as polarization, beamwidth and surface curvature. The spatial density of point reflectors is computed from a given SAR image using the processing steps outlined in Figure 2.1.

First, back projection algorithm with azimuth and frequency windowing is used to create SAR imagery with square pixels from phase history data. Next, magnitude images are thresholded to remove the background due to the noise floor. The images are then smoothed using a Gaussian Filter to avoid multiple local maxima. Next, watershedding segmentation followed by local maxima computation is used to extract the local peaks. After identifying
Figure 2.2: Feature Extraction of Civilian Vehicle Data Domes
the position of the peaks, point features are centered around their spatial mean. Finally, a Parzen density estimate is constructed by convolving the extracted features with a Gaussian Kernel. Figure 2.2 shows examples of the outputs, in log scale, generated at each block of the feature extraction algorithm in log scale. For each azimuth angle, we generate multiple shifted copies of the point density to model centering errors common in the test data. We consider shifts on a grid extending over a support of two times the range and cross range resolution.

In this section, we presented our feature extraction algorithm. In the next section, we will use the feature vectors extracted to fit statistical models that will approximate the target manifold locally and will be used in the analysis of classification error performance.

2.2 Mixture of Factor Analyzers (MoFA) models

Modeling and learning feature variability in SAR is a difficult task [13] as the target features can be masked or non-target features can be included in the feature vector under test. The high dimensionality of these feature vectors complicates the problem even further, as many of these features are highly correlated and redundant. Mapping the high dimensional features to a low dimensional descriptor of only a few latent parameters could be an important step towards addressing the curse of dimensionality and improving the performance of the learned classifier. MoFA models by Ghahramani and Hinton [17] combine clustering with local dimensionality reduction and can be an effective tool in modeling nonlinear manifolds. The dataset is divided into clusters, and each cluster is modeled as an
affine linear hyperplane with uncorrelated factors modeling random variations within each cluster. Here we use the MoFA model where the data is clustered around uniformly spaced azimuth centers. Typically, MoFA clusters are identified using an unsupervised learning algorithm, here we assume the data is labeled in azimuth, and therefore we consider clusters equispaced in azimuth rather than learning the cluster locations and membership from the data.

Figure 2.3 illustrates the algorithm of utilizing MoFA models to describe feature SAR imagery of two different classes plotted above each other. Each point represents a high dimensional feature image, after applying feature extraction (section 2.1), to its corresponding complex-valued SAR original image. As can be observed from the figure, feature images are clustered locally with respect to the factor variants in the data. In other words, data points with similar features, which are highly correlated, are grouped together. Within each cluster, local dimensionality reduction is performed. The clusters, in the figure, can be shown as the linear planes such that the mixture of these planes approximates the class’ non-linear manifold. This approach can be thought as low-ranked Gaussian Mixture Models (GMM). Later on, in future chapters, we aim to exploit these generative parametric models in designing linear embeddings to maximize the divergence between different mixture models of different classes.

We model the $D \times 1$ feature vector $x$ for the $k$’th target class using MoFA model with $M$ low rank mixture components per class. The class feature vector $x$ is assumed to be
Figure 2.3: Illustrative diagram of modeling SAR feature images using MoFA Models

sampled from the $M$ mixtures as follows:

\begin{equation}
    p_k(x) = \sum_{m=1}^{M} \int \pi_m p_k(x|\theta, \omega_m) p_\Theta(\theta|\omega_m) d\theta
\end{equation}

where $\pi_m$ is the prior probability for each mixture components $\omega_m$ is the value of the discrete random variable indexing mixture $m$, and $p \times 1$ random vector $\theta$ correspond to the factors modeling variations within each mixture component modeled as independent Gaussian random variables with unit variance:

\begin{equation}
    p_\Theta(\theta|\omega_m) = \mathcal{N}(0, I_p)
\end{equation}
For equal prior probabilities, the class feature vector $x$ is assumed to be sampled from the $M$ mixtures as follows:

$$p_k(x) = \sum_{m=1}^{M} \int \frac{1}{M} p_k(x|\theta, \omega_m)p_{\theta}(\theta|\omega_m) d\theta$$  \hspace{1cm} (2.3)$$

The conditional distribution of the feature vector $p_k(x|\theta, \omega_m)$ when sampled from the $m$’th mixture of class $k$ approximates an affine subspace model with Gaussian disturbances:

$$p_k(x|\theta, \omega_m) = \mathcal{N}(\mu_{km} + V_{km}\theta, \Sigma_{m_k} + \sigma^2_n I)$$  \hspace{1cm} (2.4)$$

where $V_{km}$ is the $D \times p$ factor loading matrix which represents the affine subspace approximating the cluster of mixture component $m$ of class $k$, $\mu_{km}$ is the mean vector for mixture component $m$ and $\Sigma_{m_k}$ is the covariance of the deviations from the subspace model. Finally, we consider white additive Gaussian noise on the feature vector components with variance $\sigma^2_n$.

The factor loading matrix $V_{km}$ and $\mu_{km}$ can be computed from the data by computing cluster covariance and means. The covariance matrices are obtained using Singular Value Decomposition (SVD) after subtracting $\mu_{km}$ from each training vector in mixture $m$ of class $k$ and using $p$ principal singular vectors scaled with the square root of the corresponding singular value. Modeling error covariance $\Sigma_{m_k}$ can be computed similarly by averaging the covariance matrix of the unmodeled $D - p$ dimensions of the $M$ mixtures.
2.3 Out-of-library Rejection Performance for MoFA Models

ATR systems are trained using a set of limited number of in-library (mission) targets. However, in actual use, the system is bound to encounter images from targets not present in the training data. Forcing an ATR decision, results in poor performance when out-of-library targets are present in test phase. Many ATR systems employ a two stage classification rule [44], where a hard decision classifier is preceded with a pre-screener, to reduce the number of classification errors caused by out-of-library confusers.

In this section, we will test the accuracy, of our statistical MoFA model, to reject out-of-library class test samples and detect in-library test samples. We achieve this by computing the maximum conditional likelihood and plotting Receiver Operating Characteristic (ROC) curves as we vary the MoFA intrinsic dimension, $p$, and the number of mixture components per class, $M$.

To test our model’s out-of-library rejection performance, we will be using a set of figures of merit that are used as metrics for the evaluation of SAR ATR:

- Probability of Detection ($P_d$): number of in-library targets detected / number of targets tested.

- Conditional Probability of Correct Class ($P_{ID}$): number of targets correctly classified / number of targets detected (conditioned on detection by pre-screener)
- Probability of False Alarm ($P_{FA}$) which represents a false positive decision: probability that an out-of-library confuser will be detected as in-library/ number of out-of-library confusers

Our decision rule will be based on Maximum Conditional Likelihood decision. For a given test sample $y \in \mathbb{R}^D$, the classical Bayes Optimal Classifier can be shown as

$$k^* = \arg\max_k p_k(y)$$  \hspace{1cm} (2.5)

We consider the simple threshold decision rule

$$\max_k p_k(y) \begin{cases} \begin{align*} H_1 & \quad \text{H}_1 \setminus \text{H}_0 \gamma \end{align*} \end{cases}$$  \hspace{1cm} (2.6)

where, $H_1$, represents the in-library hypothesis, where the test sample is detected by the pre-screener, given that its likelihood exceeds the pre-screener’s threshold, $\gamma$. While, $H_0$, represents the out-of-library hypothesis, where the test sample’s likelihood did not exceed the pre-screener’s threshold.

We consider a challenging real-world dataset. The Moving and Stationary Target Acquisition and Recognition (MSTAR) public database, [35], contains baseline X-band SAR imagery with, 1ft. x 1 ft. resolution, of 10 target types. These 10 target types include military vehicles such as BMP2 (tank), BTR70, BTR60 (armored personnel carriers), T72 (tank), T62 (tank), D7 (a bulldozer), 2S1 (cannon), ZSU (cannon), BRDM (truck) and
ZIL131 (truck). "Clutter" will be used as the confuser class, or in other words, the out-of-library class during testing out-of-library rejection performance. Images were taken over a full range of azimuth \((0^\circ - 360^\circ)\) at two different depression angles, 17 and 15, for training and testing, respectively. A pre-processing step is performed to convert the images to log scale, Next, we saturate the images to get rid of shadows, as shadows contain low saturated values. Next, we linearly stretch the log scale and finally, we window our images to obtain a centered \(32 \times 32\) feature image.

Figure 2.4 shows the ten different military vehicles \([55]\). The first two rows (a) and (b) are visible images for BMP2, BTR70, T72, BTR60, 2S1, BRDM2, D7, T62, ZIL131, and ZSU23/4, while rows (c) and (d) show their corresponding SAR images at 45 degree azimuth.

Figure 2.5 plots the MSTAR Out-of-library Rejection Performance of the MoFA models as we vary the pre-screener’s threshold. We consider 16 mixture components per class in this example. As can be seen, as the threshold increases and less test samples will be detected by the pre-screener, \(P_d\) and \(P_{FA}\) decreases, while \(P_{ID}\) increases, which represents the correctly classified test samples detected by the pre-screener.

Plotting the ROC curves in figure 2.6, we test our MoFA models’ in-library detection and out-of-library rejection performance using Maximum Likelihood (ML) classification and compare it to k-NN classification. Cross-validation on the training dataset is used to determine the optimal number of neighbors. We can observe from the figure that utilizing MoFA models in ML classification provides a better conditional probability of correct class
\( P_iD \) with less probability of false alarm \( P_{FA} \). A \( P_d \) of 0.9 is typically used and recommended as a standard operating point. We can see that at \( P_d \) of 0.9, \( P_iD \) is approximately 0.96 with a \( P_{FA} \) of 0.5 compared to \( P_iD \) of approximately 0.86 with a \( P_{FA} \) of 0.62 using K-NN classification. We included the case of marking the D7 bulldozer trained class, as a confuser. Even though \( P_{FA} \), is considerably lower for the training vehicle, \( P_iD \) almost coincides with the normal practical case of using an untrained class as a confuser.

To study how the MoFA model’s rejection performance varies with the number of mixture components per class \( M \), figure 2.7 plots \( P_iD \) and \( P_{FA} \) vs \( M \) based on the Probability of Detection standard operating point \( (P_d = 0.9) \). A good choice of \( M \) should be a trade off between the highest \( P_iD \) and the lowest \( P_{FA} \). As can be seen from the figure, starting \( M = 16 \), \( P_{FA} \) starts to converge while it provides the highest \( P_iD \) (0.966) which will be considered the most suitable number of mixture components per class for this experiment.

Similarly, figure 2.8 studies the optimal MoFA Intrinsic dimension, \( p \). We plot \( P_iD \) and \( P_{FA} \) vs \( p \) and as can be observed, \( p = 10 \) provides the most suitable trade off between the highest \( P_iD \) and lowest \( P_{FA} \).

In this chapter, we presented modeling SAR image features with MoFA. We tested its out-of-library rejection performance and its ability to classify in-library testing samples. Next, we will show the benefit of exploiting these parametric, low ranked models in the analysis of classification performance using classification error bounds.
Figure 2.4: MSTAR SAR Imagery
Figure 2.5: MSTAR Out-of-library Rejection Performance of the MoFA models based on Maximum Conditional Likelihood Classification compared to K-NN Classification, 16 Mixture Components per class are considered
Figure 2.6: Plotting Receiver Operating Characteristic (ROC) curves of the MSTAR MoFA Model based on Maximum Conditional Likelihood Classification compared to K-NN Classification, 16 Mixture Components per class are considered.
Figure 2.7: Determining the Optimal Number of Mixture Components per Class based on the $P_D = 0.9$ Probability of Detection standard operating point
Figure 2.8: Determining the Optimal Number of the MoFA Intrinsic Dimension based on the $P_D = 0.9$ Probability of Detection standard operating point
Chapter 3: Classification Performance using Theoretical Error Bounds

It is costly to develop, deploy and validate large scale ATR systems. Therefore, system developers require tools for predicting performance of ATR systems under various operating conditions such as Signal-to-Noise ratio (SNR), Signal-to-Clutter ratio (SCR), relative geometry of the target with respect to sensor platform, bandwidth and resolution. Performance prediction, typically, employs simulation studies followed by a baseline ATR system implementation and scoring; a process that is tedious and likely to be suboptimal as the ATR algorithms are not tuned for varying operating conditions. In this chapter, we propose an alternative strategy, based on statistical modeling of signatures. We derive theoretical upper-bounds on performance based on MoFA modeling of target manifolds and study how the error performance varies with different operating conditions, such as Bandwidth and SNR.

First, we review multi-class Bayes classification parametric error bounds based on pairwise classification errors. Next, we utilize these bounds to derive a probability of error union bound for the 2-class $M$-mixture case for the proposed MoFA models. After that,
we derive high-SNR approximations on the error bounds following the work of [42], and finally, analyze the classification performance using simulated CV data domes.

3.1 Theoretical Performance Bounds of the Bayes Classification Error

Bayesian Decision theory is a well-known basis for the design of parametric classifiers where the performance of these classifiers is evaluated using the Bayes risk or error. The error obtained for Bayes optimal classifier serves as a benchmark for evaluating the performance of alternative algorithms. Unfortunately, the exact closed-form expression for the average error probability of Bayesian classifiers is quite difficult to obtain. This complexity is mainly due to the presence of integration of multivariate functions over convoluted boundaries; thus, approximations and bounds of the average probability of error have been computed.

Garber [16] and Tuner [54] have obtained the upper and lower bounds for the Bayes error of an $N_c$-class problem, which can be recursively calculated as

$$P_e^{N_c} \leq \min_{\alpha \in \{0, 1\}} \left( \frac{1}{N_c - 2\alpha} \sum_{i=1}^{N_c} (1 - P(c_i))P_{e,i}^{N_c-1} + \frac{1 - \alpha}{N_c - 2\alpha} \right)$$

and

$$P_e^{N_c} \geq \frac{N_c - 1}{N_c(N_c - 2)} \sum_{i=1}^{N_c} (1 - P(c_i))P_{e,i}^{N_c-1}$$

(3.1)

where $P_e^{N_c}$ is the Bayes error for an $N_c$-class problem, $P_{e,i}^{N_c-1}$ is the Bayes error of the $(N_c - 1)$-class subproblem, where the $i$th class has been removed, and $\alpha$ is an optimization
parameter. Therefore, the Bayes error for an $N_c$ class problem can be computed starting from the $\binom{N_c}{2}$ pairwise errors.

For a 2-class problem, the Bayes error can be bounded using Mahalanobis and Bhattacharyya distances. The Mahalanobis distance is defined by:

$$P_e \leq \frac{2P(c_1)P(c_2)}{1 + P(c_1)P(c_2)\Delta} \quad (3.2)$$

where $\Delta = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$, $\Sigma = P(c_1)\Sigma_1 + P(c_2)\Sigma_2$, $\mu_1$ and $\mu_2$ are the means of class 1 and class 2, respectively, and $\Sigma_1$ and $\Sigma_2$ are their corresponding covariance matrices.

Assuming Gaussian class densities, the Bhattacharyya bound reduces to:

$$P_e^{UB} = \sqrt{P(c_1)P(c_2)e^{-K_{i,j}}}, \quad (3.3)$$

where

$$K_{i,j} = \frac{1}{8} \left\{ (\mu_i - \mu_j)^T \left[ \frac{\Sigma_i + \Sigma_j}{2} \right]^{-1} (\mu_i - \mu_j) \right\}$$

$$+ \frac{1}{2} \log \left( \frac{\det \left( \frac{\Sigma_i + \Sigma_j}{2} \right)}{\sqrt{\det(\Sigma_i)\det(\Sigma_j)}} \right). \quad (3.4)$$

In the following, we will use Bhattacharyya bound the error between a pair of mixture components sampled from the two classes. These pairwise error probabilities can be combined to bound the two-class classification error by using the union bound.

$$P_e^{UB} = \sum_{i=1}^{M} \sum_{j=1}^{M} \sqrt{P(c_1)\frac{1}{M}P(c_2)\frac{1}{M}e^{-K_{1i,2j}}}; \quad (3.5)$$
For equal class priors, \( P(c_1) = P(c_2) = \frac{1}{2} \), the misclassification error between two classes, of \( M \) factor analyzers each, can be written as

\[
P_e^{UB} = \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{1}{2M} e^{-K_{1i,2j}}, \quad (3.6)
\]

Using the model for each mixture component from Chapter 2, and plugging it to the Bhattacharyya Upper Bound in Equation 3.4, the Bhattacharyya distance between mixture \( i \) of class 1 and mixture \( j \) of class 2 can be given as

\[
K_{1i,2j} = \frac{1}{8} \left\{ (\mu_{1i} - \mu_{2j})^T \left[ \frac{(V_{1i}V_{1i}^T + V_{2j}V_{2j}^T) + \Sigma_m + \Sigma_m + \Sigma_m + 2\sigma_n^2\mathbf{I}}{2} \right]^{-1} (\mu_{1i} - \mu_{2j}) \right\}
\]

\[
+ \frac{1}{2} \log \left( \frac{\text{det} \left( \frac{(V_{1i}V_{1i}^T + V_{2j}V_{2j}^T) + \Sigma_m + \Sigma_m + 2\sigma_n^2\mathbf{I}}{2} \right)}{\sqrt{\text{det} (V_{1i}V_{1i}^T + \Sigma_m + \sigma_n^2\mathbf{I}) \text{det} (V_{2j}V_{2j}^T + \Sigma_m + \sigma_n^2\mathbf{I})}} \right) \quad (3.7)
\]

Reboredo and Calderbank [42] showed Bhattacharyya bound can be simplified for low rank models as:

\[
K_{1i,2j} = T_1 + T_2 \quad (3.8)
\]

such that

\[
T_1 = \frac{1}{8} \left\{ (\mu_{1i} - \mu_{2j})^T \left[ \frac{(V_{1i}V_{1i}^T + V_{2j}V_{2j}^T) + \Sigma_m + \Sigma_m + 2\sigma_n^2\mathbf{I}}{2} \right]^{-1} (\mu_{1i} - \mu_{2j}) \right\} \quad (3.9)
\]
and

$$T_2 = \frac{1}{2} \log \left( \frac{\text{det} \left( \frac{(V_1 V_1^T + V_2 V_2^T) + \Sigma_{m_1} + \Sigma_{m_2} + 2\sigma^2 I}{2} \right)}{\sqrt{\text{det} \left( V_1 V_1^T + \Sigma_{m_1} + \sigma^2 I \right) \text{det} \left( V_2 V_2^T + \Sigma_{m_2} + \sigma^2 I \right)}} \right)$$

(3.10)

Let $M_{1i,2j} = \{(\mu_{1i} - \mu_{2j}) (\mu_{1i} - \mu_{2j})^T\}$

We can express $T_1$ as

$$T_1 = \frac{1}{4} tr \left( M_{1i,2j} ((V_1 V_1^T + V_2 V_2^T) + 2\Sigma_{m_k})^{-1} \right)$$

(3.11)

$$= \frac{1}{4} tr \left( M_{1i,2j} U_{1i,2j} (\Lambda_{1i,2j} + 2\sigma^2 I)^{-1} U_{1i,2j}^T \right)$$

(3.12)

$$= \frac{1}{4} \sum_{l=1}^{r_{1i,2j}} \frac{1}{\Lambda_{1i,2jl} + 2\sigma^2} u_{1i,2jl}^T M_{1i,2j} u_{1i,2jl} + \frac{1}{8\sigma^2} \sum_{l=r_{1i,2j}+1}^{D} u_{1i,2jl}^T M_{1i,2j} u_{1i,2jl}$$

(3.13)

where $r_{1i,2j}$ is the rank of $V_1 V_1^T + V_2 V_2^T + 2\Sigma_{m_k}$. Simplifying the Bhattacharyya bound for low-rank models is very beneficial for simulational purposes as it helps to avoid ill-conditioned matrices and it reduces the computational complexity during the bound computation.

### 3.2 High SNR Approximation of the Error Bounds

In this section, we will derive high SNR approximation of the Bhattacharyya error bound following the same method as in [42]. We are interested in studying the asymptotic behavior, of the upper bound of the misclassification probability of error, for the 2-class problem with $M$-mixture components per class.

We will first consider the special case of one mixture component per class ($M = 1$) and equal means $\mu_1 = \mu_2$. 

\[34\]
Therefore, $P_{e}^{UB} = \sqrt{P(c_1)P(c_2)} \exp(-K_{1,2})$.

From equations 3.7 and 3.8, $T_1 = 0$ and thus, $K_{1,2} = T_2$.

Therefore, $K_{1,2}$ can be written as with model errors $\Sigma_{m_1}$ and $\Sigma_{m_2}$ included

$$K_{1,2} = T_2 = \frac{1}{2} \log \left( \frac{\det \left( \frac{(V_1V_1^T + V_2V_2^T) + \Sigma_{m_1} + \Sigma_{m_2} + 2\sigma_n^2 I}{2} \right)}{\sqrt{\det (V_1V_1^T + \Sigma_{m_1} + \sigma_n^2 I) \det (V_2V_2^T + \Sigma_{m_2} + \sigma_n^2 I)}} \right)$$

(3.14)

Suppose if the modelling errors are not taken into account, we get

$$T_2 = \frac{1}{2} \log \left( \frac{\det \left( \frac{(V_1V_1^T + V_2V_2^T) + 2\sigma_n^2 I}{2} \right)}{\sqrt{\det (V_1V_1^T + \sigma_n^2 I) \det (V_2V_2^T + \sigma_n^2 I)}} \right)$$

(3.15)

$$= \frac{1}{2} \log \left[ 2^{-r_{12}} \sigma^2 \frac{r_1 + r_2 - r_{12}}{2} \frac{\prod_{i=1}^{r_{12}} (\lambda_{1i} + \sigma_n^2)}{\prod_{i=1}^{r_1} \sqrt{\lambda_{1i} + \sigma_n^2} \prod_{i=1}^{r_2} \sqrt{\lambda_{1i} + \sigma_n^2}} \right]$$

(3.16)

where

$$r_{12} = \text{rank}(V_1V_1^T + V_2V_2^T)$$

(3.17)

$$r_1 = \text{rank}(V_1V_1^T)$$

(3.18)

$$r_2 = \text{rank}(V_2V_2^T).$$

(3.19)

The possible cases are given below

1. if the two classes completely coincide or lie on the same subspace then

$$\frac{r_1 + r_2}{2} = r_{12}.$$ The probability of misclassification error from 3.3 becomes

$$T_2 = \frac{1}{2} \log \left[ 2^{-r_{12}} \frac{\prod_{i=1}^{r_{12}} (\lambda_{1i} + \sigma_n^2)}{\prod_{i=1}^{r_1} \sqrt{\lambda_{1i} + \sigma_n^2} \prod_{i=1}^{r_2} \sqrt{\lambda_{1i} + \sigma_n^2}} \right]$$

(3.20)

$$P_{e}^{UB} = \sqrt{P(c_1)P(c_2)} \left[ 2^{-r_{12}} \frac{\prod_{i=1}^{r_{12}} (\lambda_{1i} + \sigma_n^2)}{\prod_{i=1}^{r_1} \sqrt{\lambda_{1i} + \sigma_n^2} \prod_{i=1}^{r_2} \sqrt{\lambda_{1i} + \sigma_n^2}} \right]^{-0.5}$$

(3.21)
2. if the union of subspaces of the two classes is greater than each of the individual subspaces then \( r_1, r_2 < r_{12} \leq D \), then we have

\[
\lim_{\sigma^2 \to 0} T_2 \to \infty
\]  \hspace{1cm} (3.22)

\[
\lim_{\sigma^2 \to 0} P^{UB}_e \to 0.
\]  \hspace{1cm} (3.23)

In order to include the modelling error we modify the definitions as given below

\[
r_{12} = \text{rank}(V_1 V_1^T + \Sigma_{m1} + V_2 V_2^T + \Sigma_{m2})
\]  \hspace{1cm} (3.24)

\[
r_1 = \text{rank}(V_1 V_1^T + \Sigma_{m1})
\]  \hspace{1cm} (3.25)

\[
r_2 = \text{rank}(V_2 V_2^T + \Sigma_{m2}).
\]  \hspace{1cm} (3.26)

This leads to the same results established above.

1. if the two classes completely coincide or lie on the same subspace then

\[
\frac{r_1 + r_2}{2} = r_{12}, \text{ then we have } P^{UB}_e > 0.
\]

2. if the union of subspaces of the two classes is greater than each of the individual subspaces then \( r_1, r_2 < r_{12} \leq D \), then we have

\[
\lim_{\sigma^2 \to 0} P^{UB}_e \to 0
\]  \hspace{1cm} (3.27)

Now, let us consider the case of \( M \) mixture components per class with the modelling error included. Similarly, the possible cases are:
1. if the two classes completely coincide or lie on the same subspace then \( \frac{r_1 + r_2}{2} = r_{12} \), then we have

\[
\lim_{\sigma^2 \to 0} P_{\text{UB}} = \lim_{\sigma^2 \to 0} \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{1}{M} P_{\text{UB}} \quad (3.28)
\]

The upper bound misclassification probability will exhibit an error floor at low noise levels and may fail to distinguish between both classes perfectly.

2. if the union of subspaces of the two classes is greater than each of the individual subspaces then \( r_1, r_2 < r_{12} \leq D \), then we have

\[
\lim_{\sigma^2 \to 0} P_{\text{UB}} = \lim_{\sigma^2 \to 0} \sum_{i=1}^{M} \sum_{j=1}^{M} \frac{1}{M} P_{\text{UB}} \quad (3.29)
\]

The subspaces of the parametric model of class 1 and class 2 will not overlap entirely, and classification between both classes will be performed perfectly.

### 3.3 Simulation Results

In this section, we will present empirical results using simulated target signatures. We use simulated X-band scattering data set for civilian vehicles [11], which contains fully polarimetric far-field monostatic returns for 360 degrees azimuth and elevation angles from 30 to 60 degrees. Here we use a single elevation cut at 30 degrees at VV polarization with 360 images that are generated for each azimuth degree for the two class problem of Jeep93.
Figure 3.1: Upper bound on misclassification error vs SNR for 1 GHz Bandwidth using different number of mixture components. Empirical performance of the kNN and LDA classifiers are superimposed.

and Camry. Each image was generated over a 10 degree azimuth window using azimuth and frequency windowing with $501 \times 501$ pixels with square pixels at 2.4 cm spacing.

Figure 3.1 compares the upper bound on misclassification error for a different number of mixture components for 1 GHz Bandwidth. As described in chapter 2, We divide 360 images into mixtures that are equispaced in azimuth, as can be shown in figure 3.4, and extract point feature density using the feature extraction algorithm explained in Chapter 2, Section 2.1. We approximate the feature space using the MoFA model and calculate the probability of misclassification for these two classes. At high SNR, as the number of
mixtures increase, the probability of error decreases and reaches an error floor representing fixed modeling error. We observe that deviations from the locally linear model and the corresponding error floor decrease as the number of mixture components increase.

Also, we compare our bound empirically with the performance of two algorithms: Linear Discriminant Analysis (LDA) and k-Nearest Neighbors (k-NN). As the number of mixtures is increased, the error bound approaches to the k-NN classifier error probability. For estimation of probability of error of the k-NN and LDA classifiers, 70% of the data was used in training and 30% was used in testing and cross-validation was used in choosing the number of nearest neighbors. For the case of single mixture model (M=1) the error
Figure 3.3: Upper bound on misclassification error vs number of principal components per mixture for fixed SNR and Bandwidth.

Figure 3.2 shows how the misclassification error bound behaves as the bandwidth is varied from 0.5 GHz to 2 GHz for $M = 18$ and 36 mixture components. As expected as the bandwidth is increased, classification performance improves. However, at high bandwidth ($> 2$ GHz), the performance starts to saturate and suggesting limited returns from further resources spent on increasing the bandwidth.

Figure 3.3 plots the number of principal components per mixture vs the probability of error upper bound for 1 GHz bandwidth and an SNR level of 20 dB. The results show,
for this example, that beyond 5 principal components results in negligible changes in performance and therefore provides a good tradeoff between computational complexity and model accuracy.

In this Chapter, we utilized the proposed MoFA models to present error bounds to analyze classification performance theoretically as a function of bandwidth and SNR. We compared our performance predictions with the empirical performance of practical classifiers using simulated Wideband SAR signatures of civilian vehicles. In the next chapter, we will extend our work and exploit our MoFA models to derive a novel dimensionality reduction method that will improve the classification performance through class discrimination utilizing divergence metrics.
Figure 3.4: Dividing Feature Image Data of Civilian Vehicle Data Domes into $M$ Mixtures equispaced in Azimuth
Chapter 4: Designing Maximally Discriminative Projections (MDP)

ATR systems employ classifiers trained using labeled SAR images from the different target classes capturing pose and other variations in the target signatures. While classifiers can be designed to work in the high dimensional feature space directly without exploiting manifold structure of the target signatures, the number of required training vectors grows exponentially with the number of dimensions to learn separating boundaries between the classes. Dimensionality reduction techniques reduce the storage requirements of classification algorithm by removing the redundancy between feature vectors to improve the efficiency of the learning and reduces the size of the required training dataset. In this chapter, we utilize the MoFA models for the design of a novel linear embedding method that maximizes the discrimination between mixture models in the projected space.

In the first section of this chapter, we introduce dimensionality reduction and give a review of some of the well-known dimensionality reduction techniques that will be used as a reference to compare the proposed method. In the second section, we will show the
connection between our optimized metric, the KL divergence, and the probability of mis-
classification error. Next, we will utilize the MoFA models for the design of a novel dimen-
sionality reduction technique that maximizes the discrimination between mixture models
in the projected space. Finally, we present supporting simulation results for our proposed
algorithm.

4.1 Background: Manifold Learning and Dimensionality Reduction
Techniques

Dimensionality reduction plays an important role in reducing the computational com-
plexity of the algorithms used for solving some common problems in statistical learning
such as classification or regression. The main motivation for reducing the dimensionality
comes from the fact that the sources that generate the data have a small set of degrees of
freedom. This condition, in turn, governs the intrinsic dimensionality of the data, which is
usually assumed to arise from a structured manifold such as a linear subspace or a union
of subspaces. The key idea is to reduce the rank of the projection operator while imposing
constraints on some performance metric such as energy, pairwise distances, inter-class and
intra-class distances, task-driven metrics such as the probability of error or information-
theoretic approaches, which are surrogates to the task-specific metrics. In this chapter, we
take a look at some common approaches that will be used to compare our proposed method.
4.1.1 Principal Component Analysis (PCA)

The principal component analysis is a simple method that statistically treats the data arising from a Gaussian distribution. Since the first two moments completely describe the Gaussian distribution, this method utilizes the second order moment or the covariance matrix. The goal of PCA is to find projection matrices that preserve the variance subject to a fixed rank constraint. The problem can be formally stated as follows

\[
\min_{A} \text{Tr}(A^T X^T X A), \text{ subject to } A^T A = I,
\]

(4.1)

where \( A \in \mathbb{R}^{d \times D} \) is the projection matrix such that \( d < D \), and \( X \in \mathbb{R}^{D \times N_s} \) is the data matrix with \( N_s \) samples. The term \( XX^T \) is used as the maximum likelihood estimate for the covariance matrix in the objective function. The solution of this problem can be computed as the solution of the eigen-decomposition of the covariance matrix and rank reduction by choosing the eigen vectors corresponding to the \( d \) largest eigen values given by

\[
XX^T = V \Sigma V^T
\]

\[
A = V_{1:d}.
\]

(4.2)

4.1.2 Linear Discriminant Analysis (LDA)

The LDA method is appropriate for the task of classification since it uses the class label information. LDA assumes that the distribution of the data conditioned on the class label
follows Gaussian distribution such that the covariance matrix is shared by all the classes also termed as the homoscedastic assumption. The dimensionality of the projector is always equal to \( N_c - 1 \), where \( N_c \) is the number of classes. Similar to PCA, the optimization problem can be stated as follows

\[
\min_A \operatorname{Tr} \left( \left( A^T S_B A \right) \left( A^T S_C A \right)^{-1} \right), \text{ subject to } A^T A = I, \tag{4.3}
\]

where \( A \in \mathbb{R}^{N_c - 1 \times D} \) is the projection matrix, \( S_B \) is the between class covariance matrix computed as the covariance matrix of the means of all the classes, and \( S_C \) is the average covariance matrix of all the covariance matrices computed by conditioning on the class label. The solution to this problem is termed as the generalized Eigen-value problem given by

\[
S_B A = \lambda S_C A. \tag{4.4}
\]

### 4.1.3 Locality Preserving Projections (LPP)

The previous methods aimed at reducing the dimensionality of data arising from a subspace, while this method assumes that the data is sampled from a non-linear manifold embedded in a higher dimensional space. Since the manifold is assumed to be smooth locally, this method preserves the distances of samples that lie in a specified neighborhood.
The optimization problem [21] is given by

$$\min_A \sum_{i,j} \|Ax_i - Ax_j\|_2^2 W_{i,j},$$

(4.5)

$$A = \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_d^T \end{bmatrix},$$

where $A \in \mathbb{R}^{d \times D}$ is the projection matrix, and $W_{i,j}$ is an appropriately chosen weight function that gives more importance to samples that are within a close neighborhood. Equivalently, this problem can be written using graph-theoretic notation given by forming a graph where each node in the graph is a training example and $W$ is the weight matrix that denotes the weight of the connection between each node. The alternative formulation is given by

$$\min_{a_i} (a_i^T XLX^T a_i), \text{ subject to } a_i^T XDX^T a_i = 1,$$

(4.6)

where $D \in \mathbb{R}^{N_s \times N_s}$ is the diagonal degree matrix given by diagonal elements $D_{i,i} = \sum_j W_{i,j}$, and $L \in \mathbb{R}^{N_s \times N_s}$ is the Laplacian matrix given by $L = D - W$. The final solution is obtained by solving the generalized Eigen-value problem given by

$$XLX^T a_i = \lambda XDX^T a_i.$$

(4.7)

We can easily obtain the PCA, and LDA solutions as special cases in this formulation.

### 4.1.4 Random projections

We also compare projection matrices with entries sampled from Gaussian distribution of 0 mean and variance $\sigma = \frac{1}{\sqrt{d}}$. Such projection matrices have been shown to satisfy the Johnson-Lindenstrauss (JL) Lemma [29] with high probability.
Prior stating the JL Lemma we need first to define the Restricted Isometry Property (RIP) [32]:

**Definition 4.1.1.** Consider a dataset $X = \{x_1, x_2, ..., x_{N_S}\}$, the projection matrix $A \subset \mathbb{R}^{d \times D}$ satisfies the restricted isometry property on $X$ if there exists a positive constant $\delta > 0$ such that, for every $x, x'$ in $X$, the following relations hold:

$$
(1 - \delta) \|x - x'\|_2^2 \leq \|Ax - Ax'\|_2^2 \leq (1 + \delta) \|x - x'\|_2^2 
$$

where $\delta$ is the isometry constant, $0 < \delta < 1$

Now, we can define the JL Lemma:

**Lemma 4.1.1.** Consider a dataset $X = \{x_1, x_2, ..., x_{N_S}\} \subset \mathbb{R}^D$. Let $d \geq O(\delta^{-2} \log N_S)$. Construct a matrix $A \in \mathbb{R}^{d \times D}$ by drawing each element of $A$ independently from a Gaussian distribution with mean $0$ and variance $\frac{1}{d}$. Then, with high probability, the linear operator $A : \mathbb{R}^D \rightarrow \mathbb{R}^d$ satisfies the RIP on $X$.

Random Projections thereby preserve the pairwise distances of samples obtained from a manifold of sufficiently small intrinsic dimension and bounded curvature. These random matrices are also agnostic to the structure of the manifold as long as the technical conditions are satisfied.

### 4.1.5 Nuclear norm minimization with max-norm constraints (NuMax)

Similar to LPP, NuMax [22] searches for the projection matrix that preserves the distance of sample points that are sampled from a manifold. LPP assumes that the manifold is
locally linear and preserves the pairwise distances in that local regime whereas NuMax preserves pairwise distances for all \(\binom{N_s}{2}\) points, where \(N_s\) is the number of samples. The pairwise differences are computed and denoted as the secant set \(S = \left\{ \frac{x_i - x_j}{\|x_i - x_j\|_2^2} : i \neq j \right\}\). Assuming that the manifold has a sufficiently small intrinsic dimensionality and well-defined and bounded curvature dictated by the condition-number, JL Lemma holds true, which guarantees that there exists a subspace denoted by the projection matrix \(A\) of appropriate rank such that \(|\|As\|_2^2 - 1| \leq \delta\). This guarantees that all the pairwise distances are preserved. The optimization problem formulated using the Gram-matrix formed using the projection matrix denoted by \(P = AA^T\), is given by

\[
\min_P \text{rank}(P), \text{ subject to } \|A(P) - 1_S\|_\infty \leq \delta,
\]  

(4.8)

where \(A(P)\) computes the quantity \(s^TPs, \forall s \in S\) and returns a vector of dimensionality equal to that of the cardinality of set \(S\), and \(\delta\) denotes the user-specified isometry constant.

The rank is a non-convex function and the optimization problem is typically intractable. The problem is further relaxed by using the Nuclear norm as the surrogate for the rank.

The Nuclear norm is defined as

\[
\|P\|_* = \sum_{i=1}^{d} \sigma(P)_i,
\]  

(4.9)

where \(\sigma(P)_i\) is the \(i^{th}\) singular value of the matrix \(P\). The relaxed problem is given by

\[
\min_P \|P\|_*, \text{ subject to } \|A(P) - 1_S\|_\infty \leq \delta.
\]  

(4.10)

The projection matrix \(A\) is obtained using the eigen-decomposition of the matrix \(P\). This method does not utilize the class label information for learning the projection matrix.
With a simple modification, this method has been extended to the multi-class setting. The secant set defined earlier is re-formulated as two separate sets denoted by inter-class $S_B$ and intra-class $S_C$ secant set formed using data points from different classes and the same class, respectively. The optimization problem is given by

$$\min_{P} \|P\|_s, \text{ subject to }$$

$$\|A_B(P) - 1_{S_B}\|_\infty \geq \delta,$$

$$\|A_C(P) - 1_{S_C}\|_\infty \leq \delta,$$

$$P \succeq 0,$$  \hspace{2cm} (4.12)

where $A_B(P)$, and $A_C(P)$ computes the quantity $s^T Ps$, $\forall s \in S_B, \forall s \in S_C$ and returns vectors of dimensionality equal to that of the cardinality of set $S_B$ and $S_C$, respectively.

### 4.2 Bounds on Probability of Misclassification using Information Theoretic Inequalities

While dimensionality reduction techniques can decrease the computational complexity significantly and improve learning of classification algorithms, a typical concern is preserving the classification performance in the projected space. We hypothesize that by finding projection operators that maximize the divergence or distance between the statistical models conditioned on the class label, the classification performance will be maintained in the projected space. To this end, we consider the Kullback–Leibler (KL) divergence, which is related to the probability of error in the case of binary classification. We model the
class conditional probability distribution using the MoFA model and utilize the KL divergence between the classes for the binary case and the worst case KL divergence for the multi-class problem. We approximate the KL divergence using variational methods to get tractable objective function and obtain projection operators that optimize the approximation. We empirically verify the effectiveness of this approach in both simulated as well as real datasets.

Lin [31] derived bounds linking mutual information to the probability of error through the Jensen-Shannon (JS) divergence. The upper bound for the Bayes probability of error can be shown as

\[ P_e(p_1, p_2) \leq \frac{1}{2} (H(\pi_1, \pi_2) - \frac{JS(\pi(p_1||p_2)}{\log(2)}) \]  

where \( H(\pi_1, \pi_2)) = -\pi \log_2(\pi_1) - \pi \log_2(\pi_2) \), and \( JS(p_1||p_2) \) is the JS divergence measure. \( JS(p_1||p_2) > 0 \) and is equal to zero when \( p_1 = p_2 \).

The lower bound for the Bayes probability of error can be shown to be

\[ P_e \geq \frac{1}{4} (H(\pi_1, \pi_2) - \frac{JS(\pi(p_1||p_2)}{\log(2)})^2 \]  

where

\[ H(\pi_1, \pi_2)) = -\pi \log \pi_1 - \pi \log \pi_2 \]  

Derivations of the upper and lower Bayes probability of error bounds are given in appendix C.
The probability of error upper bound and lower bound in equations 4.13 and 4.14 explain the inverse relationship between the probability of error and the divergence between the parametric statistical models. For example, as the JS-divergence increases, the upper bound for the probability of error decreases and becomes a tighter bound. Thus, we can utilize this inverse relationship to minimize the probability of error by maximizing the divergence between our proposed low-rank mixture models.

The JS divergence measure is a symmetrized and smoothed version of the Kullback Leibler (KL) divergence [33], and can be defined as

\[
JS(p_1, p_2) = \frac{1}{2} \int p_1(x) \log \frac{2p_1(x)}{p_1(x) + p_2(x)} dx + \frac{1}{2} \int p_2(x) \log \frac{2p_2(x)}{p_1(x) + p_2(x)} dx \quad (4.16)
\]

The KL divergence can be used as a measure to discriminate between two different distributions and is also known as the relative entropy [24]. It is defined as

\[
D(p_1, p_2) = \int p_1(x) \log \frac{p_1(x)}{p_2(x)} dx \quad (4.17)
\]

Therefore, it can be written in terms of the KL divergence \( D(p_1||p_2) \) as

\[
JS(p_1||p_2) = \frac{1}{2} \left( D(p_1||\frac{p_1 + p_2}{2}) + D(p_2||\frac{p_1 + p_2}{2}) \right) \quad (4.18)
\]

The closed formed expression of the KL divergence between two Gaussian distributed components can be given as
\[
D(p_1 \Vert p_2) = \frac{1}{2} \left[ \log \left| \frac{\Sigma_2}{\Sigma_1} \right| + Tr[\Sigma_2^{-1} \Sigma_1] - D(p_1 \Vert p_2) \right] + (\mu_1 - \mu_2)^T \Sigma_2^{-1} (\mu_1 - \mu_2)
\] (4.19)

For Mixture Models, as shown by Hershey and Olsen [24], we adopt to use the Variational Approximation which provides the most accurate closed-form approximation of the KL divergence.

Figure 4.1 shows a simulation example comparing the Variational Approximation with the Union bound of the KL divergence of all possible mixture components between Camry and Jeep classes. The figure shows the tightness of the Variational Approximation is to Monte Carlo Sampling (50,000 samples) when compared to the sub-optimal results of the union bound.

The expression of the Variational Approximation between two mixture models can be given as

\[
D_{\text{variational}}(p_1 \Vert p_2) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(p_{1a} \Vert p_{1a'})}}{\sum_b \pi_b e^{-D(p_{1a} \Vert p_{2b})}}
\] (4.21)

where \(\pi_a\), \(\pi_b\) are, respectively, the prior probabilities of the low-ranked Gaussian distributed mixture component, \(a\), of Class 1 and the mixture component, \(b\), of Class 2. The derivation of the Variational Approximation can be found in appendix B.

Next, we will exploit the Variational Approximation of the KL divergence and the inverse relationship, between the probability of error and the divergence measures, to design
maximally discriminative projectors in the projected space which surrogates minimizing the probability of error.

### 4.3 Maximally Discriminative Projection (MDP) Algorithm

We consider the problem of obtaining the optimized projector matrix $A_{MDP} \in \mathbb{R}^{d \times D}$ that maximizes the Variational Approximation of the KL divergence between two low-ranked Gaussian mixture models.

$A_{MDP}$ can be obtained by solving the following optimization problem:
$$A_{MDP} = \arg\max_A D_{\text{variational}}(p_1 || p_2, A) \quad (4.22)$$

We employ Gradient Ascent Optimization to maximize the Variational Approximation of the KL divergence. The Gradient of the Variational Approximation of the KL divergence can be derived to the following closed form expression

$$\frac{\partial D_{\text{variational}}(p_1 || p_2)}{\partial A} = \sum_a \pi_a \frac{\Sigma_b D'(p_{1a} || p_{2b}) e^{-D(p_{1a} || p_{2b})}}{\Sigma_b e^{-D(p_{1a} || p_{2b})}} - \frac{\Sigma_{a'} D'(p_{1a} || p_{1a'}) e^{-D(p_{1a} || p_{1a'})}}{\Sigma_{a'} e^{-D(p_{1a} || p_{1a'})}} \quad (4.23)$$

where, \(D(p_1 || p_2)\), is the KL divergence between two low-ranked Gaussian distributed components and can be shown as

$$D(p_1 || p_2) = \frac{1}{2} \left[ \log \left| A \Sigma_2 A^T \right| \right] + \text{Tr} \left[ (A \Sigma_2 A^T)^{-1} A \Sigma_1 A^T \right] - d \quad (4.24)$$

$$+ (\mu_1 - \mu_2)^T A^T (A \Sigma_2 A^T)^{-1} A (\mu_1 - \mu_2)$$

where \(d\) is the projected dimension and \(D'(p_1 || p_2)\) is the gradient of \(D(p_1 || p_2)\), and can be shown as

$$D'(p_1 || p_2) = (\Sigma_2 A^T (A \Sigma_2 A^T)^{-1})^T - (\Sigma_1 A^T (A \Sigma_1 A^T)^{-1})^T$$

$$+ \left( \Sigma_1 A^T (A \Sigma_1 A^T)^{-1} - \Sigma_2 A^T (A \Sigma_2 A^T)^{-1} A \Sigma_1 A^T (A \Sigma_2 A^T)^{-1} \right)^T$$

$$+ (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T A^T (A \Sigma_2 A^T)^{-1}$$

$$- \Sigma_2 A^T (A \Sigma_2 A^T)^{-1} A (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T A^T (A \Sigma_2 A^T)^{-1}$$

55
Let

\[ B_1 = (A\Sigma_1 A^T)^{-1}A \]
\[ B_2 = (A\Sigma_2 A^T)^{-1}A \]
\[ \bar{\mu} = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \]

\( D'(p_1||p_2) \) can be simplified to

\[ D'(p_1||p_2) = -B_1 \Sigma_1 + B_2(\Sigma_1 + \Sigma_2 + \bar{\mu} - (\Sigma_1 + \bar{\mu})A^T B_2 \Sigma_2) \quad (4.25) \]

The Maximally Discriminative Projection matrix \( A \) could be obtained recursively using Gradient Ascent.

### 4.4 Simulation Results

In this section, we will study the performance of our proposed technique compared to traditional dimensionality techniques. We build on the results from Section 3.3 by considering all 10 classes. We then model it using MoFA as explained in chapter 2. After that, we apply our proposed MDP algorithm to extract the optimal projection matrix that maximizes the Variational Approximation of the KL divergence between a desired class (Class 1) and the 9 remaining classes (Class 2).

In figure 4.2, we measure the Variational Approximation of the KL divergence between Camry versus the rest of the dataset (9 classes) for different number of measurements. 4 mixture components are considered per class. We compare our optimal projection
Figure 4.2: Comparing the Variation Approximation of the KL divergence of the mixture models of Camry class vs the rest of the dataset using the MDP Optimal Projector compared to various dimensionality reduction techniques, 4 mixture components per class algorithm versus various dimensionality reduction techniques such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Locally Preserved Projections (LPP) and random projections. Details about the various dimensionality reduction techniques can be found in Section 4.1.

We can observe the superiority of our technique verses projecting using the other conventional dimensionality reduction techniques. The above blue dashed line is the optimal KL divergence for the unprojected case, which serves as an upper bound where the solid blue line will eventually converge as we increase the number of dimensions.
Figure 4.3: Comparing Misclassification Error using SVM Classification of Camry Class using the MDP vs PCA projector using different proportionally-sized training samples

In figure 4.3, we then show classification performance using support vector machines, it can be shown that the optimal MDP design provides lower classification error compared to PCA. Also, there is not a big difference in classification error for our optimal design when we decrease the training samples to 25% vs 50%.

Figure 4.4 compares the detection rate of Camry using PCA and using the optimal MDP Projector. It can be seen that the optimal design provides better detection rate. Also, even for smaller training set (25% of the aspects), MDP still performs well and provides a much higher detection rate than PCA.
Figure 4.4: Comparing Probability of Detection using SVM Classification of Camry Class using the MDP vs PCA projector using different proportionally-sized training samples

In this Chapter, we first introduced the inverse relation between the misclassification error and the KL divergence, and elaborated choosing the KL divergence as an optimization criteria for our proposed dimensionality reduction technique. We showed that increasing the KL divergence surrogates decreasing the misclassification error, and thus improving the classification performance, through bounds that link mutual information to the probability of error. Next, we introduced the Variational Approximation of the KL divergence as a metric to obtain the KL divergence between 2 mixture models and as an optimization criteria to maximize a statistical measure of discriminability between between 2-class mixture models. After that, we presented our MDP dimensionality reduction technique through
optimizing the Variational Approximation of the KL divergence to obtain an optimal max-
imally discriminative projection matrix. We compared our technique to other conventional
linear embedding techniques in terms of divergence metrics and classification performance
and showed the superiority of our technique.
Chapter 5: Extension of MDP to Multi-class using Non-linear Constrained Optimization based on minimax quasi-Newton methods

In this Chapter, we derive an optimal projector for multi-class discrimination that can be used prior to multi-class classification algorithms like Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN).

Previously, in section 4.3, for a 2-class problem we maximized the Variational Approximation of the KL-divergence, $D_{\text{variational}}$, by calculating its gradient and solving the objective function for an optimal MDP projection matrix using gradient ascent. Now in order to generalize it for the multi-class case, we are interested in solving the projection matrix, $A_{MDP}$, that maximizes the worst-case KL divergence between all combinations of mixture model, class $i$, and its corresponding complement, which we will denote as class $i^c$, where $i^c \in \{i_1, i_2, ..., i_{NC}\}$, $i^c \notin i$ and $NC$ is the number of classes which we will denote as $D_{\text{variational}}(p_i||p_{i^c}, A)$

The maxmin multi-class MDP optimization problem can be expressed as

$$A_{MDP} = \arg\max_A \min_{(i, i^c)} D_{\text{variational}}(p_i||p_{i^c}, A)$$  \hspace{1cm} (5.1)
where, $D_{\text{variational}}(p_i \| p_{i^c}, A)$, between two mixture models, $i$ and $i^c$ can be shown as

$$D_{\text{variational}}(p_i \| p_{i^c}, A) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(p_{a} \| p_{a'}^c, A)}}{\sum_b w_b e^{-D(p_{a} \| p_{a^c}, A)}}$$  \hspace{1cm} (5.2)

where, $D(p \| q, A)$, is the KL divergence between two low-ranked Gaussian distributed single components and can be shown as

$$D(p_i \| p_j) = \frac{1}{2} \left[ \log \left| A_i \Sigma_j A_i^T \right| + \text{Tr} \left[ (A_i \Sigma_j A_i^T)^{-1} A_i \Sigma_i A_i^T \right] - d + (\mu_i - \mu_j)^T A_i^{-1} (A_i \Sigma_j A_i^T)^{-1} A_i (\mu_i - \mu_j) \right]$$

The maxmin optimization problem can be transformed to the minimax problem as follows:

$$\max_A \min_{(i, i^c)} D_{\text{variational}}(p_i \| p_{i^c}, A) = -\min_A \max_{(i, i^c)} -D_{\text{variational}}(p_i \| p_{i^c}, A)$$  \hspace{1cm} (5.3)

For simplicity, we will rewrite, $-D_{\text{variational}}(p_i \| p_{i^c}, A)$, as, $f_i(A)$, such that our new minimax equation is:

$$-\min_A \max_{(i)} f_i(A)$$  \hspace{1cm} (5.4)

The main problem arising in solving Equation 5.4, is that the function $\phi(A) = \max_{(i)} f_i(A)$, is not continuous in $A$. Following the notation in [7], we will therefore relax the objective functions, $f_i(A)$, with subject to a parameter, $\gamma$, such that $f_i(A) \leq \gamma$.

Now we can define our optimization problem as
minimize \[A, \gamma\] \[\gamma\] \tag{5.5}

subject to \[f_i(A) \leq \gamma, \ i = 1, \ldots, N_c.\]

Note that by transforming our unconstrained problem to a constrained one, not only it solves the problem of \(f_i(A)\) not being continuous, the optimizer can make informed decisions regarding directions of search and step length.

In the next section, we will address our method to solve the above optimization problem.

5.1 Multi-Class MDP using Brayton’s algorithm for minimax problems

In this section, we will employ work of Han [18] and Powell [39] for nonlinearly constrained optimization. A review is given in Appendix A.

In order to solve our optimization problem in Equation 5.5, for completeness, we will state and employ Brayton’s [7] algorithm for minimax problems as described from the following theorem

**Theorem 5.1.1.** Let \(H_a\) be positive definite denoted by \(H_a > 0\) which is an approximation of the Hessian. Details about the approximation and updating it will be given in a later section.

The step \(\Delta a \neq 0\) obtained from solving the nonlinear constrained optimization problem:

\[
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \Delta a^T H_a \Delta a + \Delta \gamma \\
\text{subject to} & \quad f_i(a) + \left(\frac{\partial f_i}{\partial a}\right)^T \Delta a \leq \gamma + \Delta \gamma
\end{align*}
\] \tag{5.6}

is in the direction of decent for the function \(\phi(a) = \max_{(i)} f_i(a)\).
where, \( a \in \mathbb{R}^{dD} \) is the vectorization of \( A \in \mathbb{R}^{d \times D} \)

**Proof.** Let us first show that \( \Delta \gamma \) must be negative.

Since \( \Delta a \neq 0 \) and \( H_a > 0 \), then from 5.6, we can show that

\[
\frac{1}{2} \Delta a^T H_a \Delta a > 0
\]

and hence \( \Delta \gamma \) is negative.

Now expanding \( f_i(a) \) about \( a \) for any function \( f_i(a) = \Delta \gamma = \max_{(j)} f_j(a) \),
we obtain

\[
f_i(a + \epsilon \Delta a) = f_i(a) + \epsilon \left( \frac{\partial f_i}{\partial a} \right)^T \Delta a + \mathcal{O}(\epsilon^2) \leq \gamma + \epsilon \Delta \gamma + \mathcal{O}(\epsilon^2) \quad (5.8)
\]

and since \( \Delta \gamma < 0 \),
then for any small \( \epsilon \) we have

\[
f_i(a + \epsilon \Delta a) \leq \gamma = \max_{(i)} f_i(a) \quad (5.9)
\]

Furthermore, for any \( f_i(a) < \gamma \), there will always exist an \( \epsilon \) such that

\[
f_i(a + \epsilon \Delta a) \leq \gamma = \max_{(i)} f_i(a) \quad (5.10)
\]

Therefore, the step \( \Delta a \) is in the direction of descent for the function \( \phi(a) = \max_{(i)} f_i(a) \).
The solution is used to form a new iterate until the optimal $a^*$ is found:

$$a_{k+1} = a_k + \beta \Delta a_k.$$  \hfill (5.11)

Proceeding finding the optimal $a^*$, we transform it back to its original matrix form $A^*$, where we have our final maximally discriminative projection matrix $A_{MDP} = A^*$

### 5.1.1 Approximating the Hessian Matrix

A positive definite quasi-Newton approximation of the Hessian of the Lagrangian function, $H_a$ is calculated at each major iteration using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method [15] as follows

$$H_{k+1} = H_k = \frac{q_k q_k^T}{q_k s_k} - \frac{H_k s_k s_k^T H_k^T}{s_k^T H_k s_k}.$$  \hfill (5.12)

where

$$s_k = a_{k+1} - a_k$$  \hfill (5.13)

and

$$q_k = \left( \nabla f\left(a_{k+1}\right) + \sum_{i}^{m} \lambda_i \nabla g_i\left(a_{k+1}\right) \right) - \left( \nabla f\left(a_k\right) + \sum_{i}^{m} \lambda_i \nabla g_i\left(a_k\right) \right).$$  \hfill (5.14)

Theorem 5.1.1 is based on $H_a$ being positive definite. $H_a$ is initialized to be positive definite. Powell [40] recommends keeping the Hessian positive definite even though it
not be the case at the solution point. As can be seen from Equation 5.12, the Hessian approximation will always be positive given two conditions. The first condition is for the Hessian approximation to be initialized as positive definite \((H_0)\). The second condition is for \(q_k^T s_k\) to be always positive. Thus, using a modification forcing it to be positive definite will ensure the validity of Theorem 5.1.1.

The main goal of this modification is to minimize the distortion of the elements of \(q_k^T s_k\) that lead to a positive definite contribution. Thus, we will divide our modification into two parts. Initially, we will repeatedly divide the most negative element of \(q_k^T s_k\) into half until it is greater than or equal to a small negative tolerance parameter. If this procedure fails to make \(q_k^T s_k\) to be positive, we will follow the second part of the modification by adding a vector \(v\) to \(q_k\) such that

\[
q_k = q_k + wv
\]

where \(w\) is a constant scalar and for each element \(i\) of vector \(v\), we can denote \(v_i\) as

\[
v_i \begin{cases} 
\nabla g_i(a_{k+1}) g_i(a_{k+1}) - \nabla g_i(a_k) g_i(a_k) & \text{if } (q_k)_i w < 0 \text{ and } (q_k)_i (s_k)_i < 0 \\
0 & \text{otherwise}
\end{cases}
\]

In this way, \(w\) will increase systematically until \(q_k^T s_k > 0\).

### 5.2 Simulation Results

We consider a challenging real-world dataset to illustrate the performance of the multi-class dimensionality reduction technique derived in the previous sections. The Moving and
Stationary Target Acquisition and Recognition (MSTAR) public database, [35], contains baseline X-band SAR imagery with, 1 ft. x 1 ft. resolution, of 10 target types such as BMP2 (an armored personnel carrier), BTR70 (an armored personnel carrier), and T72 (a tank), D7 (a bulldozer), and ZIL131 (a truck). Images were taken over a full range of azimuth ($0^\circ - 360^\circ$) at two different depression angles, 15 and 17. A pre-processing step is performed where we convert the images to log scale, Next, we saturate the images to get rid of shadows, as shadows contain low saturated values. After that, we linearly stretch the log scale and finally we window our images to centered $32 \times 32$ features.

To further test the limits of our dimensionality reduction technique, we will use two different depressions during training and testing. A good classification system should be able to recognize the difference of a $2^\circ$ depression system without having trained on such images. Also, a good dimensionality reduction technique should preserve the information so that the classification performance almost stays the same. In other words, a good projection matrix applied to the data matrix, should not increase the misclassification error when applied to a classifier, compared to applying the unprojected data matrix to the classifier directly.

In figure 5.1, we compare the performance of our multi-class MDP approach compared to NuMax, PCA and Random Projections. We use a k-NN-classifier for classification performance. To determine the optimal number of neighbors, cross-validation was used. For NuMax, the Projected Dimension, $d$, can be varied by varying the isometry constant, $\delta$, between, $0 - 1$, through an inverse proportional relationship as shown from equation 4.12.
This is because as we decrease the isometry constant, the more preserved pairwise distances should be included, the higher the rank of the projection matrix becomes. The projected dimensions’ range, that correspond to varying the isometry constant, will be denoted as NuMax’s Restricted Isometry Parameter (RIP) range. We perform our experiment for two separate gathered datasets at 17° and 15° for training and testing, respectively. The number of mixture components considered for multi-class MDP here is 4. As observed from the figure, the classification performance using Multi-class MDP converges approximately to only 1% error. Also, it can be observed from the figure that Multi-class MDP performs better than the remaining dimensionality reduction techniques, including NuMax, this is because MDP is optimized to minimize the misclassification error through maximizing the KL-divergence as we have shown in chapter 3.

Figure 5.2 provides a closer zoomed version of figure 5.1 at NuMax’s RIP range (0 < \( \delta \) < 1). It also includes various different number of mixture components per class cases for multi-class MDP in order to see how varying the number of mixture components will affect the classification performance. As expected, as we increase the number of mixture components, the performance improves and the misclassification error decreases. For 2 mixture components, the performance is not optimal and can be comparable to NuMax and PCA. However, as we increase it to 4 and 8 mixture components, multi-class MDP outperforms NuMax, PCA and random projections. As observed, the misclassification performance of the 8 mixture components and the 4 mixture components cases are nearly equal. However, the processing time for the 8 mixture components case is significantly
greater than the 4 mixture components case as can be shown from table 5.1. Thus, we suggest choosing 4 mixture components per class, for this case, which provides a good trade off between processing time and classification performance.

5.2.1 Worst-Case Divergence Performance

We are interested in presenting an analysis of our Multi-Class MDP performance, in terms of the worst case divergence, between all combinations of pairs of class $i$ and the remaining classes $i^C$. The divergence metric, that will be used, is the Variational Approximation of the KL Divergence.
Figure 5.2: Zoomed view of Multi-Class MDP classification performance with various different number of mixture components per class cases at NuMax’s RIP range

<table>
<thead>
<tr>
<th>Projected Dimension</th>
<th>4 Mixture Components</th>
<th>8 Mixture Components</th>
<th>Processing Time Ratio Increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1.6599</td>
<td>12.8183</td>
<td>672.23%</td>
</tr>
<tr>
<td>16</td>
<td>2.1268</td>
<td>301.5308</td>
<td>14077.67%</td>
</tr>
<tr>
<td>36</td>
<td>47.54138</td>
<td>125.1179</td>
<td>163.17%</td>
</tr>
<tr>
<td>48</td>
<td>90.4931</td>
<td>165.8853</td>
<td>83.31%</td>
</tr>
<tr>
<td>58</td>
<td>99.0825</td>
<td>366.1769</td>
<td>269.56%</td>
</tr>
</tbody>
</table>

Table 5.1: Table comparing processing time of Multi-Class MDP in terms of mixture components per class
We will define the Worst-Case Divergence (WCD) metric as

\[
\min_{(i, i^C)} D_{\text{variational}}(p_i, ||p_{i^C}, A)
\]  \hspace{1cm} (5.17)

In other words, the WCD metric measures the robustness of a given dimensionality reduction technique in terms of how well can all combinations of a given class, \(i\), be discriminated from their corresponding complement class, \(i^C\).
Figure 5.3 compares the WCD performance of Multi-Class MDP to other dimensionality reduction techniques. As can be observed, Multi-Class MDP outperforms the rest and provides the highest worst-case discrimination performance.
Chapter 6: Conclusion

6.1 Summary

In this dissertation, we considered the problem of ATR-SAR system analysis and design using generative probabilistic models of SAR target signatures. We proposed the use of low-ranked parametric mixture models to analyze classification performance theoretically as a function of bandwidth and SNR, through Bayes performance bounds and compared our performance predictions with the empirical performance of practical classifiers using simulated wideband SAR signatures of civilian vehicles. Next, we tested the performance of our MoFA models in reject out-of-library class test samples and detect closed-library test samples by computing the maximum conditional likelihood and plotting Receiver Operating Characteristic (ROC) curves varying the MoFA intrinsic dimension and the number of mixture components per class. Then, we reviewed information theoretic inequalities that link probability of error to mutual information through divergence bounds and proposed a novel maximally discriminative projector based on maximizing the Variational Approximation of the KL divergence. We compared our method to conventional dimensionality reduction techniques and showed the superiority of our algorithm. Finally, we extended
our MDP method to multi-class using non-linear constrained optimization and maxim
quasi-Newton methods. We showed that our dimensionality reduction method outperforms
other methods suggested in the literature when used in combination with SVM and KNN-
classifiers. Experiments were performed on both simulated data through the Civilian Vehi-
 cle data domes and real-world challenging data through the well-known MSTAR dataset.

6.2 Future Work

6.2.1 Optimal Projected Dimension for a given Worst-Case Divergence

In this dissertation, our proposed MDP algorithm was employed to maximize the dis-
 crimination between classes and find the optimal projection matrix for a specified number
of projected dimensions.

In future work, we are interested in looking to the problem from a different angle,
which involves finding the optimized number of projected dimensions given a worst-case
divergence threshold. This approach is inspired by the NuMax optimization framework
[22].

The new optimization framework can be shown utilizing the Gram-matrix formed using
the projection matrix denoted by \( P = AA^T \). The optimized number of Projected Dimen-
sions can be computed by solving the following optimization problem:

\[
\min_{P} \text{rank}(P), \quad \text{subject to } D_{\text{variational}}(p_i \parallel p_i^v, P(A)) \geq D_{\lambda}, \forall i = 1, 2, \ldots N_C
\]
where the Projected Dimension, $d = \text{rank}(P)$. The projection matrix $A$ is obtained using the Eigen-decomposition of the matrix $P$.

The problem of the non-convexity of the rank function can be solved by further relaxing the problem by using the *Nuclear norm* as the surrogate for the rank.

### 6.2.2 The JS divergence as an alternative MDP optimization criteria

In both our 2-class and multi-class MDP problems, we utilized the Variational Approximation of the KL divergence as an optimization metric to discriminate between different mixture models of classes. The KL divergence, which is an asymmetric metric, was used to measure how different, the probability distribution is of a certain class’ mixture model, is compared to the remaining classes.

In future work, we are interested in investigating a different approach, where we will utilize the JS divergence, the symmetric version of the KL divergence and that is defined in chapter 4, as the new optimization criteria. It can be written as a function of the KL divergence as shown in equation 4.18.

In this work, we focus on finding the worst-case divergence measure that maximizes the discrimination between all the possible combinations of a given class $i$ from the remaining classes, denoted by the complement class $i^C$, where we analyze the problem from the class $i$’s perspective. This approach leads to an asymmetric measure. Instead, we will investigate methods to obtain the worst-case measure that maximizes the discrimination between all possible combinations of given classes $i$ and $j$, where $i \neq j$. Please note, that in this case,
we are considering the problem from both classes’ point of view and thus, a symmetric version of divergence measure could be used, such as the JS divergence.

It might be interesting to study and analyze the difference in classification performance compared to using the KL divergence metric. Also, processing time will be a factor in comparison. Notice beforehand, our algorithm using the KL divergence metric considered optimizing between $N_C$ combinations of class $i$ and its complement class $i^C$. Using the JS divergence metric will lead to $\binom{N_C}{2}$ combinations.

The expression of JS Variational Approximation between two mixture models, using equation 4.18, can be given as

$$JS_{\text{variational}}(p_1||p_2) = \frac{1}{2}D_{\text{variational}}(p_1||\frac{p_1 + p_2}{2}) + \frac{1}{2}D_{\text{variational}}(p_2||\frac{p_1 + p_2}{2})$$ (6.2)

For simplicity, let $p_3 = \frac{p_1 + p_2}{2}$.

Thus, substituting equation 4.21 in equation 6.2, the JS Variational Approximation between two mixture models can be given as

$$JS_{\text{variational}}(p_1||p_3) = \frac{1}{2} \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(p_1||p_1a')}}{\sum_c \pi_c e^{-D(p_1||p_3c)}}$$ (6.3)

$$+ \frac{1}{2} \sum_b \pi_b \log \frac{\sum_{b'} \pi_{b'} e^{-D(p_2||p_2b')}}{\sum_c \pi_c e^{-D(p_2||p_{3c})}}$$ (6.4)

The maxmin multi-class MDP optimization problem, using the JS divergence Variational Approximation metric, can be expressed as
\[
\max_{A} \min_{(i,j)} JS_{\text{variational}}(p_i||p_j, A), i \neq j
\] (6.5)

### 6.2.3 Multi-Class MDP based on Class-Specific Feature (CSF) Classifier

For a given multidimensional sample of data, \(x \in \mathbb{R}^D\), the classical Bayes classifier that minimizes the probability of error for equi-probable prior probabilities can be shown as:

\[
i^* = \arg \max_i p(x|H_i)
\] (6.6)

Assume lower dimensional projected feature sets \(z_i = T_i(x)\) exists and we are interested in tailoring the Bayes classifier to function in the lower dimensional space. This will be very computationally efficient, especially for large dimensional data. Of course, the more the transformation \(T_i\) maintains the discrimination between the different hypothesis/classes, the better the performance of the classifier will be in the lower dimension.

We are interested in the following decision rule:

\[
i^* = \arg \max_i p(z_i|H_i)
\] (6.7)

Unfortunately, the classifier modeled is invalid because comparisons of densities on different feature spaces are unfeasible.
From the PDF projection theorem in [4], Baggenstoss showed that, provided we know the PDF under some reference hypothesis $H_0$ at both the input and output of transformation $T_i(x)$, we can approximate $p_x(x|H_1)$ as:

$$
\hat{p}_x(x|H_1) = \frac{p_x(x|H_0(z))}{p_z(z|H_0(z))} \hat{p}_z(z|H_1), z = T(x) \quad (6.8)
$$

The theorem provides a means of creating PDF approximations on the high-dimensional input data space without dimensionality penalty using low-dimensional feature PDFs and provides a way to optimize the approximation by controlling both the reference hypothesis $H_0$ as well as the features themselves. The beauty of the theorem is that the resulting function remains a PDF whether or not the features are sufficient statistics. Since sufficiency means optimality of the classifier, approximate sufficiency means PDF approximation and approximate optimality. The Class-Specific Features (CSF) classifier is an application of the projected PDF theorem as a Bayesian classifier. This can be shown by plugging 6.8 in 6.6, the Decision Rule is

$$
i^* = \arg\max_i \frac{p(x|H_{0,i})}{p(z_i|H_{0,i})} p(z_i|H_i) P(H_i), z_i = T_i(x) \quad (6.9)
$$

Note that an important decision is choosing the reference hypothesis, $H_{0,i}$. The implicit assumptions in, 6.7, is that that feature $z_i = T_i(x)$ is a sufficient (or an approximately sufficient) statistic for the binary test between class $H_i$ and $H_{0,i}$. Approximate sufficiency means that
\[
\frac{p(x|H_i)}{p(x|H_{0,i})} \approx \frac{p(z_i|H_i)}{p(z_i|H_{0,i})}
\] (6.10)

The feature set of interest in our problem is the lower dimensional projected feature sets, \(z_i = T_i(x)\), where \(T_i(x) = A_i x\) is a many to one linear transformation and \(A_i\) is the MDP Projection Matrix that maximizes the Variational Approximation of the KL divergence between class \(H_i\) and class \(H^C_i\) which will serve as our reference hypothesis, \(H_{0,i}\), where class \(H^C_i\) are classes \(1, 2, \ldots, N_C \neq i\).

Future work involves applying the PDF theorem by re-writing equation 6.11 to be suitable for our problem. The CSF classifier for MDP feature sets can be shown as:

\[
i^* = \arg\max_i \frac{p(x|H^C_i)}{p(z_i|H^C_i)} p(z_i|H_i) p(H_i), z_i = (A_{MDP})_i x
\] (6.11)

where \((A_{MDP})_i\) can be denoted by

\[
(A_{MDP})_i = \arg\max_A D_{\text{variational}}(p_i||p_{i^*}, A)
\] (6.12)

The above decision rule can be used to generalize classification, using MDP feature sets, from the 2-class problem to the multi-class problem without the urge to obtain a multi-class, universal MDP projection matrix and going through the complexity of minimax problems.

For future work, it will be interesting analyzing the above MDP multi-class extension and comparing it to our proposed method in chapter 5 in terms of classification performance and processing time.
Appendix A: Background on Constrained Nonlinear Optimization and Sequential Quadratic Programming (SQP)

Early work on constrained optimization, focused on transforming the constrained optimization problem into an unconstrained problem with respect to a penalty function, which by solving a sequence of parametrized unconstrained optimizations converged to the constrained problem. These methods however are inefficient and have been replaced by methods based on the solution of the Kuhn-Tucker (KKT) equations, which are necessary conditions for constrained optimization and serve as a basis for many nonlinear programming algorithms. If however, the objective function, $f(x)$ and the constraints, $g(x)$ are convex functions, then the KKT equations are both necessary and sufficient conditions for a global minimizer solution.

We are interested in the work of Biggs, Han [18] and Powell [39] for nonlinearly constrained optimization. These methods are based on accumulating second-order information regarding the KKT equations using quasi-Newton methods. These methods closely mimics Newton’s method for constrained optimization just as it is done with quasi-Newton methods for unconstrained optimization.
Initially, we will study the constrained optimization problem with equal constraints only and later on, we will solve it for inequality constraints. This review is based on following the notation of [7].

The optimization problem for equal constraints can be denoted by

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad g(x) = 0
\end{align*}$$

(A.1)

where $g(x) = (g_1(x), g_2(x), \ldots, g_m(x))^T$ are the equality constraints. Based on the KKT equations, there exist Lagrange multipliers $\lambda_i$, such that

$$h(x, \lambda) = \frac{\partial f}{\partial x} + \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x} = 0, \quad i = 1, \ldots, m$$

(A.2)

and the Lagrangian function can be shown as

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

(A.3)

Lagrange multipliers balance deviations in magnitude of the objective function and the constraint gradients. They also ensure that only active constraints at the solution point are considered so that the canceling between the gradients of the objective function and the active constraints happen. Inactive constraints should not be included in equation A.2 and there corresponding Lagrange multipliers equal to 0.

Notice that the equation A.2 has $n + m$ equations in $n + m$ unknowns which represent the dimension of $x$ and $\lambda$, respectively.
This standard Newton method that solves this equation can be given by

$$\begin{bmatrix} h(x, \lambda) \\ g(x) \end{bmatrix} + \begin{bmatrix} f_{xx} + \sum \lambda_i g_{ixx} \\ g_x \\ 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = 0 \tag{A.4}$$

where $f_{xx}$ and $g_{ixx}$ are the second derivative matrices of $f$ and $g_i$, respectively. Each iteration, we will solve for $\Delta x$ and $\Delta \lambda$ and update $x$ and $\lambda$ till they converge through the following update equations:

$$x^* = x + \Delta x$$

$$\lambda^* = \lambda + \Delta \lambda \tag{A.5}$$

Quasi-Newton’s approach operates on the formulation of a QP sub-problem by approximating the Lagrangian function, A.3 by a quadratic function as follows

$$L(x + \Delta x, \lambda) = \frac{1}{2} \Delta x^T H \Delta x + h^T (x, \lambda) \Delta x + L(x, \lambda) \tag{A.6}$$

We can approximate A.4 by

$$\begin{bmatrix} h(x, \lambda) \\ g(x) \end{bmatrix} + \begin{bmatrix} H \\ g_x \\ 0 \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta \lambda \end{bmatrix} = 0 \tag{A.7}$$

These KKT equations are similar as if we were to solve the quadratic program:

$$\begin{array}{ll}
\text{minimize} & \frac{1}{2} \Delta x^T H \Delta x + f_x^T \Delta x + f(x) \\
\text{subject to} & g_i(x) + g_{ix}^T \Delta x = 0, \ i = 1, \ldots, m
\end{array} \tag{A.8}$$
where $H$, is a positive definite approximation of the Hessian matrix of the Lagrangian function. $H$ can be updated by various of quasi-Newton methods where the most popular one is BFGS method, which is used by Powell.

Similarly for the inequality constraints case, the optimization problem can be given as

$$
\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0 \quad i = 1, \ldots, m
\end{align*}
$$

where $g(x) = (g_1(x), g_2(x), \ldots, g_m(x))^T$

and the quadratic approximation to the Lagrangian function corresponding to equation A.8 can be given as

$$
\begin{align*}
\text{minimize} & \quad \frac{1}{2} \Delta x^T H \Delta x + f^T x \Delta x + f(x) \\
\text{subject to} & \quad g_i(x) + \left( \frac{\partial g_i}{\partial x} \right)^T \Delta x \leq 0, \quad i = 1, \ldots, N_c
\end{align*}
$$
Appendix B: Derivation of the Variational Approximation of the KL divergence

To derive the Variational Approximation, as shown by Hershey and Olsen [24], we first introduce another form of the KL-divergence by relating it to the likelihood, where

$$ D(p_1||p_2) = L_{p_1}(p_1) - L_{p_1}(p_2) \quad \text{(B.1)} $$

$L_{p_1}(p_2)$, the likelihood, can be denoted by

$$ L_{p_1}(p_2) = E_{p_1(x)} \log p_2(x). \quad \text{(B.2)} $$

We introduce the variational parameters $\phi_{b|a} > 0$ such that $\sum_b \phi_{b|a} = 1$.

The variational lower bound to the likelihood can be derived by exploiting Jensen’s inequality as follows
\[ L_{p_1}(p_2) = E_{p_1(x)} \log p_2(x) \]  
\[ = E_{p_1(x)} \log \sum_b w_b p_{2_b}(x) \]  
\[ = E_{p_1(x)} \log \sum_b \phi_{b|a} w_b p_{2_b}(x) \]  
\[ \geq E_{p_1(x)} \sum_b \phi_{b|a} \log \frac{w_b p_{2_b}(x)}{\phi_{b|a}} \]  
\[ = L_{p_1}(p_2, \phi). \]  
(B.3)  
(B.4)  
(B.5)  
(B.6)  
(B.7)

where we define \( L_{p_1}(p_2, \phi) \) as the lower bound of \( L_{p_1}(p_2) \).

We maximize \( L_{p_1}(p_2, \phi) \) to tighten the bound. The optimal variational parameter that maximizes \( L_{p_1}(p_2, \phi) \) is

\[ \hat{\phi}_{b|a} = \frac{w_b e^{-D(p_{1,a} || p_{2,b})}}{\sum_{b'} \pi_{b'|a} e^{-D(p_{1,a} || p_{2,b'})}}. \]  
(B.8)

Similarly,

\[ L_{p_1}(p_1) = E_{p_1(x)} \log p_1(x) \]  
\[ = E_{p_1(x)} \log \sum_{a'} \pi_{a'p_{1,a'}}(x) \]  
\[ = E_{p_1(x)} \log \sum_{a'} \psi_{a'|a} \frac{\pi_{a'p_{1,a'}}(x)}{\psi_{a'|a}} \]  
\[ \geq E_{p_1(x)} \sum_{a'} \psi_{a'|a} \log \frac{\pi_{a'p_{1,a'}}(x)}{\psi_{a'|a}} \]  
\[ = L_{p_1}(p_1, \psi). \]  
(B.9)  
(B.10)  
(B.11)  
(B.12)  
(B.13)
where we define $L_{p_1}(p_1, \psi)$ as the lower bound of $L_{p_1}(p_1)$.

We maximize $L_{p_1}(p_1, \psi)$ to tighten the bound. The optimal variational parameter that maximizes $L_{p_1}(p_1, \psi)$ is

$$
\hat{\psi}_{a'|a} = \frac{\pi_{a'} e^{-D(p_{1,a}||p_{1,a'})}}{\sum_{\hat{a}} \pi_{\hat{a}} e^{-D(p_{1,a}||p_{1,\hat{a}})}}.
$$  \hfill (B.14)

By defining $D_{\text{variational}}(p_1||p_{p_2}) = L_{p_1}(p_1, \hat{\psi}) - L_{p_1}(p_2, \hat{\phi})$ and substituting equations B.8 and B.14, the result simplifies to the closed form Variational Approximation

$$
D_{\text{variational}}(p_1||p_2) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(p_{1,a}||p_{1,a'})}}{\sum_b w_b e^{-D(p_{1,a}||p_{2,b})}}
$$  \hfill (B.15)
Appendix C: Bounds on Probability of Misclassification using Information Theoretic Inequalities

As derived in [31], Lin linked mutual information to probability of misclassification error through using divergence bounds.

**Theorem C.0.1** (Theoretical Upper Bound).

\[
P_e(p_1, p_2) \leq \frac{1}{2}(H(\pi_1, \pi_2) - JS_\pi(p_1||p_2) \log(2)) \tag{C.1}
\]

where \( H(\pi_1, \pi_2) = -\pi \log_2(\pi_1) - \pi \log_2(\pi_2) \)

**Proof.** From [23]

\[
P_e(1, 2) \leq \frac{1}{2} * H(C|X) \tag{C.2}
\]

where \( H(C|X) \) is the conditional entropy

\[
H(C|X) = \sum_{x \in X} p(x)H(C|x) \tag{C.3}
\]

\[
= - \sum_{x \in X} p(x) \sum_{c \in C} p(c|x) \log p(c|x) \tag{C.4}
\]
Also,

\[ H(C|X) = H(C) + H(X|C) - H(X) \]  \hspace{1cm} (C.5)

For the two class problem

\[ H(C) = H(p(c_1) + p(c_2)) \]  \hspace{1cm} (C.6)

and

\[ H(X|C) = p(c_1)H(X|c_1) + p(c_2)H(X|c_2) = \pi_1 H(p_1) + \pi_2 H(p_2) \]  \hspace{1cm} (C.7)

Also since

\[ p(x) = \pi_1 p_1(x) + \pi_2 p_2(x) \]  \hspace{1cm} (C.8)

Therefore,

\[ H(X) = H(\pi_1 p_1 + \pi_2 p_2) \]  \hspace{1cm} (C.9)

Combining (C.9), (C.7), (C.6) into (C.5), from inequality (C.2), we obtain

\[ P_e \leq \frac{1}{2} \left( H(\pi_1, \pi_2) + \pi_1 H(p_1) + \pi_2 H(p_2) - H(\pi_1 p_1 + \pi_2 p_2) \right) \]

\[ = \frac{1}{2} \left( H(\pi_1, \pi_2) - \frac{JS_{\pi}(p_1||p_2)}{\log(2)} \right) \]  \hspace{1cm} (C.10)

\[ \square \]
where $JS(p_1||p_2)$ is the Jensen-Shannon (JS) divergence measure which was introduced by Lin [31]. $JS(p_1||p_2) > 0$ and is equal to zero when $p_1 = p_2$
**Theorem C.0.2** (Theoretical Lower Bound).

\[ P_e \geq \frac{1}{4}(H(\pi_1, \pi_2) - \frac{JS_\pi(p_1||p_2)}{\log(2)})^2 \]  
\hspace{1cm} (C.11)

where

\[ H(\pi_1, \pi_2) = -\pi \log \pi_1 - \pi \log \pi_2 \]  
\hspace{1cm} (C.12)

**Proof.** From (C.5) and the Cauchy inequality,

\[ H^2(C|X) \leq (\sum_{x \in X} p(x))(\sum_{x \in X} p(x)H^2(C|x)) \]
\hspace{1cm} (C.13)

\[ = -\sum_{x \in X} p(x)H^2(C|x) \]

Also, for any \(0 \leq t \leq 1 - t\)

\[ \frac{1}{2}H(t, 1 - t) \leq \sqrt{t(1 - t))} \]  
\hspace{1cm} (C.14)

Therefore inequality (C.13) can be written as

\[ H^2(C|X) \leq 4 \sum_{x \in X} p(x)(p(c_1|x)p(c_2|x)) \]  
\hspace{1cm} (C.15)

\[ \leq 4 \sum_{x \in X} p(x) \min(p(c_1|x)p(c_2|x)) \]  
\hspace{1cm} (C.16)

\[ = 4 \sum_{x \in X} \min(\pi_1p_1(x), \pi_2p_2(x)) = 4P_e(p_1, p_2) \]  
\hspace{1cm} (C.17)

From (C.5) - (C.9),
$$P_e \geq \frac{1}{4}(H(\pi_1, \pi_2) - \frac{JS_{\pi}(p_1||p_2)}{\log(2)})^2$$

(C.18)
Bibliography


