Bayesian Networks and Gaussian Mixture Models in Multi-Dimensional Data Analysis with Application to Religion-Conflict Data

by

Hui Liu

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved April 2012 by the
Graduate Supervisory Committee:

Thomas Taylor, Co-Chair
Douglas Cochran, Co-Chair
Junshan Zhang

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

This thesis examines the application of statistical signal processing approaches to data arising from surveys intended to measure phychological and sociological phenomena underpinning human social dynamics. The use of signal processing methods for analysis of signals arising from measurement of social, biological, and other non-traditional phenomena has been an important and growing area of signal processing research over the past decade. Here, we explore the application of statistical modeling and signal processing concepts to data obtained from the Global Group Relations Project, specifically to understand and quantify the effects and interactions of social psychological factors related to intergroup conflicts.

We use Bayesian networks to specify prospective models of conditional dependence. Bayesian networks are determined between social psychological factors and conflict variables, and modeled by directed acyclic graphs, while the significant interactions are modeled as conditional probabilities. Since the data are sparse and multi-dimensional, we regress Gaussian mixture models (GMMs) against the data to estimate the conditional probabilities of interest. The parameters of GMMs are estimated using the expectation-maximization (EM) algorithm. However, the EM algorithm may suffer from over-fitting problem due to the high dimensionality and limited observations entailed in this data set. Therefore, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are used for GMM order estimation.

To assist intuitive understanding of the interactions of social variables and the intergroup conflicts, we introduce a color-based visualization scheme. In this scheme, the intensities of colors are proportional to the conditional probabilities observed.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

LIST OF SYMBOLS

$\boldsymbol{x}_i$	The ith data observation	12
n	Variable dimensionality	12
N	Data size	12
M	Number of mixture components	12
w_i	Weight of the ith mixture component	12
μ_i	Expectation of the ith mixture component	12
Σ_i	Variance or covariance matrix of the ith mixture component	12
$\boldsymbol{X}$	Integration of data samples	13
$\boldsymbol{Y}$	Hidden state (the mixture component information in GMM)	13
$\boldsymbol{\theta}$	Parameter for true density function	13
$\hat{\boldsymbol{\theta}}$	Estimated parameter for the density function	14
ϕ_i	Parameter of the ith mixture component	15
$\hat{\boldsymbol{\theta}}_o$	Estimated parameter based on partial observations	20
K	Number of free parameters	22

Chapter 1

INTRODUCTION

Modern signal processing comprises a corpus of statistical, analytical, and algorithmic techniques that have proven effective across a wide variety of applications. Traditional uses of signal processing, and the ones in context of which much of the current subject were developed, involve signals transduced from physical phenomena that are described by classical models, such as those for electromagnetism, fluid dynamics, and Newtonian mechanics. Many standard assumptions and models in signal processing are explicitly or implicitly associated with these phenomena, and in some cases signal processing has been instrumental in the process of understanding and modeling the behavior of systems governed by physical laws.

Over the past decade, the use of signal processing with non-traditional signals has been of increasing interest within the research community. Such signals include measurements associated with biological, sociological, and psychological phenomena. It is widely understood that standard assumptions and models, and methods predicated upon or optimized for these, will generally not apply in such domains. Nevertheless, much recent research is based on the premise that the underlying mathematical principles of signal processing are often compatible with adaptation or generalization to such non-traditional settings.

This thesis explores the utility of statistical and model fitting ideas familiar in signal processing to data collected with the intention of understanding intergroup conflict in human social systems. The foundations of intergroup conflict have long been of significant interest in the social sciences, and the practical relevance of grasping these foundations has been acutely advanced by the rise of technologies enabling the rapid spread of information across social groups that may be globally distributed. A wide literature has explored a number of different hypotheses that have been advanced for the primary causal factors of intergroup conflict. Previous research [1, 2, 3, 4, 5, 6, 7] has explored social factors such as primordial affiliation, traditions, ancient hatred, value incompatibility, cultural difference between groups, competition over resources, economics and power, collective fears of the other group, etc. In spite of religious riots, murders and pogroms, religion has been largely discounted as the "true" motivating cause of intergroup conflicts and has been viewed as a cloak for other motivations.

By contrast, the research discussed sought to explore the extent to which religion may drive or influence intergroup conflict, in terms of an aggregate religious variable which has been labeled "religious infusion". Our analysis is founded on Global Group Relations Project surveys [8, 9] that elicit information about socio-political and religious variables in conflict and non-conflict situations. We seek to discover if and to what extent some combinations of these variables can serve as predictors of conflict at some level. In this discussion the word "conflict" is interpreted to be much broader than violence or shedding blood. It includes the five main intergroup conflict variables: prejudice, interpersonal discrimination, symbolic aggression, individual violence and collective violence [8, 9]. These are thought of roughly in this order on a scale representing increasing severity of conflict.

In fact, the survey measured a large number (169) of socio-political variables, but with limited observations: 731 with missing data, and 310 with complete data for the analysis. The social scientists working on this project focussed on just a few of the these variables and aggregated these down to just three main predictor variables: competition over resources and power, incompatibilities between groups' values and their religious infusion.

In this thesis, statistical modeling is used to quantify the relationships between the predictors and the conflict variables. Solutions to religious-conflict problems based on structural equation modeling (SEM) have been given in the previous work of the Global Group Relations Project [8]. This model is founded on the assumption of a normal distribution over, and a linear relationship between, all variables. As a result, upon the regression of dependent (also referred as endogenous) variables on the independent (also referred as exogenous) variables, the resulting linear model is unable to explain nonlinear relationships in the data. Furthermore, the small sample size hampers the ability of the SEM regression to deal effectively with more than one conflict variable and two predictor variables at the same time; it is thus impossible to take all three predictors into consideration at the same time. In this thesis, we present alternative, more modern, methods for statistical analysis of the religious conflict problem that are innovative within the context of statistical techniques for political analysis. The new approach improves the performance significantly by carrying fewer assumptions and by optimizing probability representations, which enables the observation of non-linearity in the relationship and interactions of higher dimensionality (3-D and 4-D analysis). In addition, it provides a means to visualize the high dimensional data in an RGB color map, and thereby facilitate understanding of relationships. The introduction of Bayesian networks provides an extremely simple and straightforward model of the impact of the social predictors on the conflict outcomes. To set the parameters of these Bayesian networks we implement Gaussian mixture models for different combinations of predictors and conflict variables in which parameters are estimated and optimized to maximize the likelihood function and the number of free parameters is optimized in the sense of information theory, as instantiated in the Bayesian Information Criterion [10] and/or the Akaike Information Criterion [11].

The form of the GMM provides a means by which the results of this statistical modeling may be visualized — and this is an innovative improvement for the visualization improves our ability to interpret what the statistical models say about the data — in which conditional probabilities have been turned into colors. In the visualizations, bright red paints the cases where a high likelihood of severe conflict exists, while bright blue paints the cases where there is almost no conflict. These results are discussed in the penultimate section of this thesis.

Besides analysis of the socio-political data, the statistical techniques described above, especially GMM, are suitable for implementation in engineering applications for traditional signal processing as well. Details of such applications are discussed in some recent papers on speaker identification [12], object detection [13], face recognition [14], medical image processing [15], etc.

Chapter 2

DISCUSSION OF TECHNIQUES

2.1 Bayesian networks

Introduction to Bayesian networks

When even a moderate number of simple variables are measured, e.g., 35 binary variables, the space of possible outcomes is exponentially large, in this case $2^{35} \approx 34$ billion. Inferential statements regarding probabilities of unmeasured outcome variables marginalized on hidden variables involve summation over these combinatorial outcomes, and hence are computationally intractable. Efficient bookkeeping methods utilizing known conditional independencies can help to manage the complexity of uncertainty and dependence of variables. In 1985, Pearl initiated the use of graphical models to efficiently tabulate statistical relationships between variables [16, 17]. Motivated by Bayes' rule which expresses the relationship between opposite conditional dependencies $P(A|B)$ and $P(B|A)$ in terms of marginal probabilities $P(A)$ and $P(B)$, this technique uses graphical representation of guide iterated applications of Bayes' rule for extended inference across multiple variables and multiple statistical relationships. This structure, termed "Bayesian network", has also been referred to as "influence network" for the objective of illustrating the influence among variables. Bayesian networks use directed acyclic graphical (DAG) models to present the knowledge of uncertainty and conditional dependence [18]. Statistical dependencies can be encoded in the structure of Bayesian networks; these are often obtained from domain experts' knowledge [19].

Bayesian networks provide a straightforward mathematical language to express relations between variables in a clear form [20]. Applications of Bayesian networks have been useful tools in engineering, including the areas of speech recognition [21, 22], image processing [23, 24], wireless communications [25, 26], biomedical engineering [27, 28] and others [29, 30, 31]. The networks, derived from uncertainty and causality, provide systematic and localized solutions for the probabilistic information structuring while supported by inference algorithms [32].

Definitions

Since Bayesian networks are based on DAG models, we first give some basic terms from graph theory to prepare for the discussion of Bayesian networks.

Definition 1 (Graph, Directed Graph)**.** *A finite **graph** $\mathcal{G} = (V, E)$ consists of a finite set of **nodes** V and an **edge** set E, where each edge indicates a unique connection between two nodes so that the elements of $E \subseteq V \times V$ consists of two-element subsets of V. Specifically, if e connects distinct nodes α, β, then $e = \{\alpha, \beta\}$. By contrast, for a **directed graph** $\mathcal{G} = (V, E)$ the **edge** set E consists of unique directed edges, each **from** some vertex α **to** some other vertex β; i.e., a directed edge is an ordered pair (α, β). Following common notation, an **undirected edge** connecting α and β is denoted as $\langle \alpha, \beta \rangle$.*

Definition 2 (Path, Directed Path)**.** *Let $\mathcal{G} = (V, E)$ denote a graph. A **path** of length m from a node α to a node β is a sequence of distinct nodes $(\tau_0, \ldots, \tau_m)$ such that $\tau_0 = \alpha$ and $\tau_m = \beta$ such that $(\tau_{i-1}, \tau_i) \in E$ for each $i = 1, \ldots, m$. The **path** is a **directed path** if all edges (τ_{i-1}, τ_i) for $i = 1, \ldots, m$ in the path are directed edges.*

Definition 3 (Directed Acyclic Graph)**.** *A graph $\mathcal{G} = (V, E)$ is a **directed acyclic graph** if each edge is directed and, for any node $\alpha \in V$, there does not exist any set of distinct nodes $\tau_1, \ldots, \tau_m$ such that $\alpha \neq \tau_i$ for all $i = 1, \ldots, m$ and $(\alpha, \tau_1, \ldots, \tau_m, \alpha)$ forms a directed path.*

Definition 4 (Parent, Child). *In a directed graph $\mathcal{G} = (V, E)$, an ordered pair of nodes $(\alpha, \beta) \in E$, β is referred to as a child of α and α as a parent of β.*

With the basic terminology of graph theory above, we now give the definition of a Baysian network.

Definition 5 (Bayesian Network). *A **Baysian network** is a pair (G, P), where $G = (V, D)$ is a directed acyclic graph (DAG) consisting of a set of nodes $V = \{\alpha_1, \ldots, \alpha_n\}$ and directed edge set D between variable nodes such that each node α_v has a set of parents $\pi_v = (\alpha_{v_1}, \ldots, \alpha_{v_m})$, and there is an assigned potential $P(\alpha_v|\pi_v)$. The joint probability is $P(\alpha_1, \ldots, \alpha_n) = \prod_{v=1}^{n} P(\alpha_v|\pi_v)$.*

Although the notion of Bayesian network does not impose any *a priori* constraint on the form of the distributions $P(\alpha_v|\pi_v)$, in this thesis we assume each variable $\alpha \in V$ in a Bayesian network has a finite number of mutually exclusive states. A given joint distribution may have more than one Bayesian network representation. For example, it is always the case that $P(\alpha_1, \ldots, \alpha_n) = \prod_i P(\alpha_i|\alpha_{i+1}, \ldots, \alpha_n)$, and any permutation of the variables $\alpha_{j_1}, \ldots, \alpha_{j_n}$ has a corresponding representation of $P(\alpha_1, \ldots, \alpha_n) = \prod_i P(\alpha_{j_i}|\alpha_{j_{i+1}}, \ldots, \alpha_{j_n})$.

Examples of Bayesian networks

In the case of three random variables A, B and C, the above factor model gives the joint probability distribution

$$P(A, B, C) = P(C|A, B)P(B|A)P(A). \tag{2.1}$$

and can be associated to the directed acyclic graph in Figure 2.1.

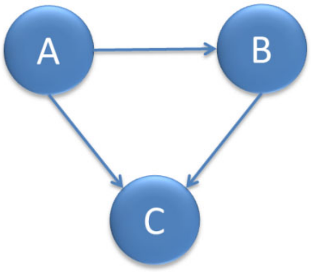


Figure 2.1: A Bayesian network example

The network model in Figure 2.1 clearly describes the variables in a probabilistic sense. Another issue that should be taken into consideration is conditional independence [33]. Three typical cases regarding independence that are significant components of more complex Bayesian networks are discussed as follows.

Case 1 Diverging connections

In the above example, if

$$P(C|A,B) = P(C|A), \tag{2.2}$$

then C is independent of B conditioned on A; i.e.,

$$P(B,C|A) = P(C|A,B)P(B|A) = P(C|A)P(B|A). \tag{2.3}$$

Although B and C are not independent, the equation indicates that B and C are independent of each other conditioned on A.

Therefore, conditioning on A introduces independence to the variables B and C and the network can be simplified by removing the directed edge from B to C, as shown in Figure 2.2. In this case, we say that A, B and C are subject to diverging connections [34].

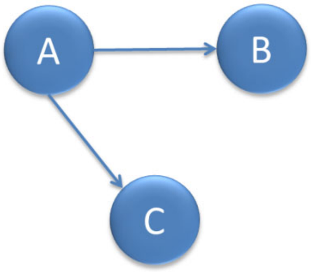


Figure 2.2: Diverging connections

Case 2 Converging connections

Another network can be derived from the basic example in Figure 2.1 by removing the directed edge from A to B. Random variables A and B are independent when no observations have been made; i.e.,

$$P(A, B) = P(A)P(B). \tag{2.4}$$

However, conditioning on C, introduces dependency between A and B according Bayes' rule:

$$P(A, B|C) = \frac{P(A, B, C)}{P(C)} = \frac{P(A|B, C)P(B|C)P(C)}{P(C)} = P(A|B, C)P(B|C). \tag{2.5}$$

This conditional dependence is represented using the Bayesian network shown in Figure 2.3. In this case, we say that A, B and C are subject to converging connections [34].

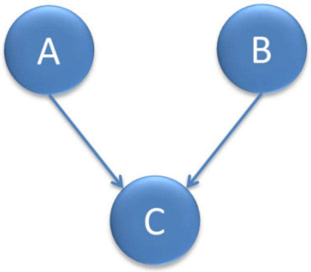


Figure 2.3: Converging connections

Case 3 Serial connections

The third network removes the directed edge from A to C based on the basic network in Figure 2.1. Then the only connection from A to C is through the influence on B. Therefore, when B is known, the connection has been "blocked" and random variables A and C are thus independent of each other conditional on B.

$$P(A, C|B) = \frac{P(A, B, C)}{P(B)} = \frac{P(A)P(B|A)P(C|B)}{P(B)} = P(A|B)P(C|B). \quad (2.6)$$

This conditional dependence is represented using the Bayesian network shown in Figure 2.4. In this case, we say that A, B and C are subject to serial connections [34].

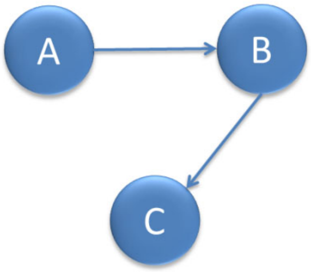


Figure 2.4: Serial connections

In this thesis all Bayesian networks considered will be of converging type because our subject matter experts wish to evaluate the hypothesis that, and the extent to which, conflict variables are predicted by social variables.

2.2 GAUSSIAN MIXTURE MODELS

Introduction

The Gaussian mixture model (GMM) efficiently models the distribution of data observations as a weighted sum of parameterized Gaussian distributions. As such it provides a computationally feasible non-Gaussian generalization of the linear-Gaussian model that is standard of classical statistics. Various attempts have been made exploring GMM both practically and theoretically after their initial introduction by Pearson to classify two subspecies of crabs from the Bay of Naples in 1894 [35].

While regressing the data samples in a descriptive Gaussian mixture model, we consider two significant concerns in this thesis. An obvious issue in model fitting is estimating the parameters given observations. While the idea of using maximum-likelihood (ML) estimation started in the 1930s [36, 37, 38], the advent of EM algorithm in 1977 [37] has proven to be effective and popular method for the ML fitting of a Gaussian mixture model.

The GMM yields estimates of the parameters and weighting of mixture components for a fixed finite number of components. However, the choice of the number of mixture model components in the GMM is another issue. Too large a number of components can lead to over-fitting, which in turn may result in extra computational complexity and the loss of universality. Techniques for choosing the number of components that consider a penalized form of likelihood, such as the Akaike information criterion [11], the Bayesian information criterion [10] and many other criteria have provided methods to address the problem as described above.

By using a sufficient number of Gaussian distributions and adjusting the weights, means and variances, any continuous density can be approximated to arbitrary accuracy by a Gaussian mixture [33]. This also leads to an extremely flexible method for clustering, especially for the data having asymmetrical distributions. GMMs have been widely used in applications including astronomy, biology, genetics, medicine, psychiatry, economics, engineering and marketing [39]. In this thesis, GMM is applied to the social psychological problem of quantifying the relationship between social factors and intergroup conflicts.

Definition of GMM

As previously mentioned, a GMM models the distribution of data by a weighted sum of parameterized Gaussian distributions. On the n-dimensional Euclidean space $\mathbb{R}^n$, we suppose that independent n-dimensional data observations $\boldsymbol{x}_i$ have been drawn as $\{\boldsymbol{x}_i : i = 1, \ldots, N\}$ where N denotes the size of the data sample. The probability density of GMM for M mixture components built for the above variable is written as

$$\begin{aligned} p(\boldsymbol{x}) &= \sum_{i=1}^{M} w_i \mathcal{N}(\boldsymbol{x}; \mu_i, \Sigma_i) \\ &= \sum_{i=1}^{M} w_i \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_i|^{\frac{1}{2}}} \exp\left(-\frac{(\boldsymbol{x} - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (\boldsymbol{x} - \mu_i)}{2} \right). \end{aligned} \tag{2.7}$$

In the above equation, M reflects the number of mixture components, while μ_i, and Σ_i are the mean and covariance matrix for the ith mixture component. The weights accordingly are represented by w_i and satisfying

$$0 \leq w_i \leq 1 \qquad (i = 1, \ldots, M) \tag{2.8}$$

and

$$\sum_{i=1}^{M} w_i = 1. \tag{2.9}$$

Given data observations, the mixture model can be obtained by the maximum-likelihood approach using expectation-maximization algorithm.

EM algorithm

General derivation of EM algorithm

Assume $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is the list of all observations, which are each statistically independent of the others and drawn from the distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$. The joint density function of all the observations is thus

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}),$$

where, in the case of interest to us, each $p(\boldsymbol{x}|\boldsymbol{\theta})$ is a Gaussian mixture with parameters

$$\boldsymbol{\theta} = (w_1, \ldots, w_M, \mu_1, \ldots, \mu_M, \Sigma_1 \ldots \Sigma_M).$$

For the sake of computation, we take the logarithm of the density function to form the log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{X})$ which is referred as likelihood function of the parameters given the data; i.e.,

$$\ell(\boldsymbol{\theta}|\boldsymbol{X}) \triangleq \log p(\boldsymbol{X}|\boldsymbol{\theta}) = \log \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}).$$

In a parameter estimation problem, the objective is to find a value of $\boldsymbol{\theta}$ that maximizes likelihood function as

$$\boldsymbol{\theta}^{opt} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\boldsymbol{X}). \tag{2.10}$$

Now we assume the observations $\boldsymbol{X}$ are generated from some Gaussian mixture distribution by a process of 1) a random (and unobservable, or hidden) draw $\boldsymbol{Y} = i \in \{1, 2, \cdots, M\}$ according to the probability distribution $w_1, w_2, \cdots w_M$, followed by 2) a draw from the multivariate normal distribution $\mathcal{N}(\mu_i, \Sigma_i)$ with parameters $\boldsymbol{\theta}$. Then we can define the set of observations and hidden states as $(\boldsymbol{X}, \boldsymbol{Y})$, therefore the joint density function is

$$p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta}) = p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta}), \tag{2.11}$$

while the function we seek to optimize is the conditional expectation (marginal) $\sum\limits_{i} p(\boldsymbol{X}, \boldsymbol{Y} = i|\boldsymbol{\theta})$. Accordingly, we define a joint log-likelihood $\ell(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) = \log p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})$. Before looking at the optimization problem, we introduce Jensen's inequality. For a concave function defined on an interval $\mathbb{I}$, and coefficients λ_i such that $\sum\limits_{i=1}^{m} \lambda_i = 1$ and $\lambda_1, \ldots, \lambda_m \geq 0$, if $x_1, \ldots, x_m \in \mathbb{I}$,

$$f(\sum_{i=1}^{m} \lambda_i x_i) \geq \sum_{i=1}^{m} f(\lambda_i x_i). \tag{2.12}$$

Assume the estimate of parameters is $\hat{\boldsymbol{\theta}}$. Since the logarithm is a concave function, and $\sum\limits_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) = 1$ we can derive the following inequality:

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{X}) - \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{X}) &= \log\left(\sum_{\boldsymbol{Y}} p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta})p(\boldsymbol{Y}|\boldsymbol{\theta})\right) - \log p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) \\
&= \log\left(\sum_{\boldsymbol{Y}} p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta})p(\boldsymbol{Y}|\boldsymbol{\theta})\frac{p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}})}{p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}})}\right) - \log p(\boldsymbol{X}|\hat{\boldsymbol{\theta}}) \\
&\geq \sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \log\left(\frac{p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta})p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}})}\right) \\
&- \sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \log p(\boldsymbol{X}|\hat{\hat{\boldsymbol{\theta}}}) \\
&= \sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \log\left(\frac{p(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{\theta})p(\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}})p(\boldsymbol{X}|\hat{\boldsymbol{\theta}})}\right) \\
&= \sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \log\left(\frac{p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{Y}|\hat{\boldsymbol{\theta}})}\right).
\end{aligned}$$

Thus the increment of the log-likelihood can be written in the form

$$\ell(\boldsymbol{\theta}|\boldsymbol{X}) - \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{X}) = \sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}) \log\left(\frac{p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{Y}|\hat{\boldsymbol{\theta}})}\right) \triangleq D(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$$

where the equal sign holds when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Thus by maximizing the $D(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\theta}$, it is also guaranteed that $\ell(\boldsymbol{\theta}|\boldsymbol{X})$ is not smaller than $\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{X})$.

Therefore we propose an iterative method of maximizing $\ell(\boldsymbol{\theta}|\boldsymbol{X})$ where at each step $D(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}})$ is maximized with respect to $\boldsymbol{\theta}$:

$$\begin{aligned}
\hat{\boldsymbol{\theta}}^{(m+1)} &= \arg\max_{\boldsymbol{\theta}} D(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(m)}) \\
&= \arg\max_{\boldsymbol{\theta}} \left[\ell(\hat{\boldsymbol{\theta}}^{(m)}|\boldsymbol{X}) + \sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{\theta}}^{(m)}) \log\left(\frac{p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})}{p(\boldsymbol{X},\boldsymbol{Y}|\hat{\boldsymbol{\theta}}^{(m)})}\right)\right] \\
&= \arg\max_{\boldsymbol{\theta}} \left[\sum_{\boldsymbol{Y}} p(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{\theta}}^{(m)}) \log p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})\right] \\
&= \arg\max_{\boldsymbol{\theta}} \left[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{\theta}}^{(m)}}[\log p(\boldsymbol{X},\boldsymbol{Y}|\boldsymbol{\theta})]\right] \\
&= \arg\max_{\boldsymbol{\theta}} \left[\mathbb{E}_{\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{\theta}}^{(m)}}[\ell(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y})]\right].
\end{aligned} \tag{2.13}$$

The above expression suggests two main steps, expectation step (E-step) and maximization step (M-step), for parameter optimization. In the E-step, the expectation of $\ell(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{Y})$ with respect to $(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{\theta}}^{(m)})$ can be computed using the previous step estimation and the knowledge of the model. Then the expectation is maximized over $\boldsymbol{\theta}$ which is defined as the M-step. Therefore this algorithm is referred to as the EM algorithm [37, 40].

EM algorithm in GMM

Now we consider an incomplete data density function in the form of a finite mixture model as

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_{j=1}^{M} w_j p_j(\boldsymbol{x}|\phi_j) \tag{2.14}$$

in which the $\boldsymbol{\theta}$ is composed of the weights w_j and parameters ϕ_j when the index of the mixture component $j = 1, \ldots, M$. The weights are subject to constraints given in equations (2.8) and (2.9). The log likelihood expression for the mixture density model is

$$\ell(\boldsymbol{\theta}|\boldsymbol{X}) = \log \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{M} w_j p_j(\boldsymbol{x}_i|\phi_j) \right). \tag{2.15}$$

Maximization turns out to be difficult considering the logarithm of a summation.

To simplify this problem, the EM algorithm introduces the hidden states $\boldsymbol{Y}$ which is defined corresponding to the mixture components that the sample data belongs to in the mixture model. So $y_i \in \{1, 2, \ldots, M\}$ and

$$p(y_i|\boldsymbol{\theta}) = \frac{w_{y_i}}{\sum\limits_{y_i=1}^{M} w_{y_i}} = w_{y_i} \tag{2.16}$$

demonstrating that the weight w_j can also be interpreted as the probability that a particular sample belongs to jth mixture component. Therefore, the optimization problem is significantly simplified as shown in equation (2.13).

We first examine the term $\ell(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y})$.

$$\begin{aligned}
\ell(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}) &= \log(p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})) \\
&= \log \prod_{i=1}^{N} p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log(p(\boldsymbol{x}_i, y_i|\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} \log(p(y_i|\boldsymbol{\theta}) p(\boldsymbol{x}_i|y_i, \boldsymbol{\theta})) \\
&= \sum_{i=1}^{N} \log(w_{y_i} p_{y_i}(\boldsymbol{x}_i|\phi_{y_i})).
\end{aligned} \tag{2.17}$$

According to the objective of the EM algorithm in equation (2.13), we then optimize the expectation of this term with respect to $(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}^{(m)})$

$$p(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\theta}^{(m)}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{x}_i, \boldsymbol{\theta}^{(m)}). \tag{2.18}$$

Since $p_j(\boldsymbol{x}_i|\phi_j)$ can be easily obtained given the statistical model. According to Bayes' rule and equations (2.14) and (2.16), we have

$$p(y_i|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)}) = \frac{p(y_i, \boldsymbol{x}_i|\hat{\boldsymbol{\theta}}^{(m)})}{p(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}^{(m)})} = \frac{p(\boldsymbol{x}_i|y_i, \hat{\boldsymbol{\theta}}^{(m)})p(y_i|\hat{\boldsymbol{\theta}}^{(m)})}{p(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}}^{(m)})} = \frac{\hat{w}_{y_i}^{(m)} p_{y_i}(\boldsymbol{x}_i|\hat{\phi}_{y_i}^{(m)})}{\sum\limits_{j=1}^{M} \hat{w}_j^{(m)} p_j(\boldsymbol{x}_i|\hat{\phi}_j^{(m)}))}. \tag{2.19}$$

In this case, equation (2.13) is formed as

$$\begin{aligned}
&\mathbb{E}[\log p(\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{\theta})|\boldsymbol{X}, \hat{\boldsymbol{\theta}}^{(m)}] \\
&= \sum_{\boldsymbol{Y}} \log(\ell(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{Y}))p(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{\boldsymbol{Y}} \sum_{i=1}^{N} \log(w_{y_i} p_{y_i}(\boldsymbol{x}_i|\phi_{y_i})) \prod_{k=1}^{N} p(y_k|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \log w_{y_i} p_{y_i}(\boldsymbol{x}_i|\phi_{y_i})) \prod_{k=1}^{N} p(y_k|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \sum_{i=1}^{N} \sum_{l=1}^{M} \delta_{l,y_i} \log w_l p_l(\boldsymbol{x}_i|\phi_l)) \prod_{k=1}^{N} p(y_k|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} \log(w_l p_l(\boldsymbol{x}_i|\phi_l)) \sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{l,y_i} \prod_{k=1}^{N} p(y_k|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} \log(w_l p_l(\boldsymbol{x}_i|\phi_l)) \\
&\qquad \left(\sum_{y_1=1}^{M} \sum_{y_2=1}^{M} \cdots \sum_{y_N=1}^{M} \delta_{l,y_i} \prod_{k=1, k\neq i}^{N} p(y_k|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^{(m)}) \right) p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} \log(w_l p_l(\boldsymbol{x}_i|\phi_l)) \prod_{k=1, k\neq i}^{N} \left(\sum_{y_k=1}^{M} p(y_k|\boldsymbol{x}_k, \hat{\boldsymbol{\theta}}^{(m)}) \right) p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} \log(w_l p_l(\boldsymbol{x}_i|\phi_l)) p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)}) \\
&= \sum_{l=1}^{M} \sum_{i=1}^{N} \log(w_l) p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)}) + \sum_{l=1}^{M} \sum_{i=1}^{N} \log(p_l(\boldsymbol{x}_i|\theta_l)) p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)})
\end{aligned} \tag{2.20}$$

in which the expectation is multiplied by $\sum\limits_{l=1}^{M} \delta_{l,y_i} = 1$ to simplify the expression and l indexes the mixture components.

As parts of $\boldsymbol{\theta}$, the weights w and parameters ϕ are uncorrelated and therefore can be considered separately. We now consider the first term with respect to estimating the weight of a certain mixture component w_l. Given the constraints in equation (2.9), we can obtain the maximum using a Lagrange multiplier

$$\frac{\partial}{\partial w_l}\left[\sum_{l=1}^{M}\sum_{i=1}^{N}\log(w_l)p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})+\lambda\left(\sum_l w_l-1\right)\right]=0,$$

$$\sum_{i=1}^{N}\frac{1}{w_l^{(m+1)}}p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})+\lambda=0,$$

And by computation we obtain that $\lambda=-N$, so

$$w_l^{(m+1)}=\frac{1}{N}\sum_{i=1}^{N}p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)}).$$

Now we have successfully estimated the weights. The parameter ϕ_l is comprised of the mean μ_l and variance Σ_l in a Gaussian mixture model shown in equation (2.7). Taking the logarithm and substituting into equation (2.20), we get

$$\begin{aligned}&\sum_{l=1}^{M}\sum_{i=1}^{N}\log(p_l(\boldsymbol{x}_i|\mu_l,\Sigma_l))p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})\\&=\sum_{l=1}^{M}\sum_{i=1}^{N}\left(-\frac{1}{2}\log(|\Sigma_l|)-1/2(\boldsymbol{x}_i-\mu)^T\Sigma_l^{-1}(\boldsymbol{x}_i-\mu_l)\right)p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)}).\end{aligned}\tag{2.21}$$

Taking the partial derivative with respect to μ_l and setting it equal to zero yields

$$\hat{\mu}_l^{(m+1)}=\frac{\sum_{i=1}^{N}\boldsymbol{x}_i p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})}{\sum_{i=1}^{N}p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})}.\tag{2.22}$$

Similarly, differentiating with respect to Σ_l gives

$$\hat{\Sigma}_l^{(m+1)}=\frac{\sum_{i=1}^{N}(\boldsymbol{x}_i-\hat{\mu}_l^{(m+1)})(\boldsymbol{x}_i-\hat{\mu}_l^{(m+1)})^T p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})}{\sum_{i=1}^{N}p(l|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}^{(m)})}.\tag{2.23}$$

To sum up, in a Gaussian mixture model, the EM algorithm provides estimates of the parameters w, μ and Σ iteratively based on the previous step estimates

$$w_l^{(m+1)} = \frac{1}{N}\sum_{i=1}^{N} p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)}),$$

$$\hat{\mu}_l^{(m+1)} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)})}{\sum_{i=1}^{N} p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)})},$$

$$\hat{\Sigma}_l^{(m+1)} = \frac{\sum_{i=1}^{N} (\boldsymbol{x}_i - \hat{\mu}_l^{(m+1)})(\boldsymbol{x}_i - \hat{\mu}_l^{(m+1)})^T p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)})}{\sum_{i=1}^{N} p(l|\boldsymbol{x}_i, \hat{\boldsymbol{\theta}}^{(m)})}.$$

Order of a GMM

The EM algorithm is able to achieve an arbitrarily accurate approximation in the sense that the mean-square error approaches zero, for GMMs with arbitrarily large numbers of components. However, zero error is achieved in the limit of one $\mathcal{N}(\mu_i = \boldsymbol{x}_i, \Sigma_i = 0)$ multivariate normal per data point $\boldsymbol{x}_i$, which is surely over-fitting. Therefore, it becomes important to determine the number of mixture components, which is also referred as the order of the mixture model. There exist a variety of solutions for order optimization for mixture models, including graphical tools [41, 42, 43], information theoretic critera [10, 11], kernel techniques [44, 45], moment-based methods [46, 47, 48], and some other non-parametric estimation techniques [49, 50, 51].

In this thesis we concentrate on selecting the order of Gaussian mixture models using information theoretic criteria, including the Akaike information criterion and Schwarz' Bayesian information criterion.

Akaike information criterion

Before discussing the AIC, we first introduce basic ideas of information theory. Optimal model choice, in our case GMM order, can be approached in terms of the Kullback-Leibler (K-L) information which measures the information loss of an optimal model estimate from reality [52]; i.e., the K-L distance from the estimated model to the true distribution [53]. Here we define a good model or a good estimate to be close to the true distribution in the sense of having a small K-L value. Let $f(\boldsymbol{x}|\boldsymbol{\theta})$ denote the "true" model, in which $\boldsymbol{x}$ represents the random variable and $\boldsymbol{\theta}$ is the true parameter value, and let $g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ denote a model estimate based on the whole data set and estimated parameter $\hat{\boldsymbol{\theta}}$. The K-L information of $f(\boldsymbol{x}|\boldsymbol{\theta})$ with respect to $g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ is defined as

$$\begin{aligned} I\{f(\boldsymbol{x}|\boldsymbol{\theta}); g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})\} &\triangleq \int f(\boldsymbol{x}|\boldsymbol{\theta}) \log \frac{f(\boldsymbol{x}|\boldsymbol{\theta})}{g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})}\, d\boldsymbol{x} \\ &= \int f(\boldsymbol{x}|\boldsymbol{\theta}) \log f(\boldsymbol{x}|\boldsymbol{\theta})\, d\boldsymbol{x} - \int f(\boldsymbol{x}|\boldsymbol{\theta}) \log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})\, d\boldsymbol{x} \\ &= \mathrm{const} - \mathbb{E}_{\boldsymbol{x}} \left[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \right] \end{aligned} \tag{2.24}$$

which measures the K-L divergence between $f(\boldsymbol{x}|\boldsymbol{\theta})$ and $g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$. The objective of model selection is to minimize K-L information. Since the first term is a function of the truth which is a constant, the minimization is equivalent to maximizing the second term on the right side. To further explore the model selection problem, it is better to remove the uncertainty of parameter estimation. So we add another expectation with respect to the $\hat{\boldsymbol{\theta}}$. The problem now becomes minimizing $\mathbb{E}_{\hat{\boldsymbol{\theta}}}\mathbb{E}_{\boldsymbol{x}} \left[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \right]$.

The logarithm $\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})$ can be expanded to second order in a Taylor series around an estimate by partial observations $\hat{\boldsymbol{\theta}}_o$ as follows:

$$\begin{aligned} \log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \approx& \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o)) + \left[\frac{\partial \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial \hat{\boldsymbol{\theta}}} \right]^{\mathrm{T}} [\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o] \\ &+ \frac{1}{2} [\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o]^{\mathrm{T}} \left[\frac{\partial^2 \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial^2 \hat{\boldsymbol{\theta}}} \right] [\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o] \end{aligned} \tag{2.25}$$

So the expectation is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] \approx & \mathbb{E}_{\boldsymbol{x}}\left[\log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))\right] + \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial \hat{\boldsymbol{\theta}}}\right]^{\mathrm{T}}[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o] \\ & + \frac{1}{2}[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o]^{\mathrm{T}}\left[\mathbb{E}_{\boldsymbol{x}}\frac{\partial^2 \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial^2 \hat{\boldsymbol{\theta}}}\right][\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o]\end{aligned} \tag{2.26}$$

In the first-order term, we observe that $\mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial \hat{\boldsymbol{\theta}}}\right]$ can be derived from differentiation of the K-L information $I\{g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}); g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o)\}$. It is known that the minimum of the K-L information occurs at the best estimate value, where $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_o$. Thus we know that

$$\begin{aligned}\left[\frac{\partial I\{g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}); g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o)\}}{\partial \hat{\boldsymbol{\theta}}}\right]_{\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}_o} &= \left[\frac{\partial \int g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) \log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o)\, d\boldsymbol{x}}{\partial \hat{\boldsymbol{\theta}}}\right]_{\hat{\boldsymbol{\theta}}=\hat{\boldsymbol{\theta}}_o} \\ &= \mathbb{E}_{\boldsymbol{x}}\left[\frac{\partial log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial \hat{\boldsymbol{\theta}}}\right] = 0\end{aligned} \tag{2.27}$$

And in the second-order term, we assume

$$\frac{\partial^2 \log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))}{\partial^2 \hat{\boldsymbol{\theta}}} \triangleq I(\hat{\boldsymbol{\theta}}_o) \tag{2.28}$$

Now the expectation to be maximized can be written as

$$\mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] \approx \mathbb{E}_{\boldsymbol{x}}\left[\log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))\right] + \frac{1}{2}\left[[I(\hat{\boldsymbol{\theta}}_o)][\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o][\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o]^{\mathrm{T}}\right] \tag{2.29}$$

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] \approx \mathbb{E}_{\boldsymbol{x}}\left[\log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))\right] + \frac{1}{2}tr\left[[I(\hat{\boldsymbol{\theta}}_o)]\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o][\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o]^{\mathrm{T}}\right] \tag{2.30}$$

If we see $\hat{\boldsymbol{\theta}}$ as a random variable with mean $\hat{\boldsymbol{\theta}}_o$, then the term $\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o][\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o]^{\mathrm{T}}$ is the covariance matrix Σ.

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] \approx \mathbb{E}_{\boldsymbol{x}}\left[\log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))\right] - \frac{1}{2}\operatorname{tr}\left[[I(\hat{\boldsymbol{\theta}}_o)]\Sigma\right] \tag{2.31}$$

Similarly, $\mathbb{E}_{\boldsymbol{x}}\left[\log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))\right]$ is also expanded to second order in a Taylor series and we get

$$\mathbb{E}_{\boldsymbol{x}}\left[\log(g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_o))\right] \approx \mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] - \frac{1}{2}\operatorname{tr}\left[[I(\hat{\boldsymbol{\theta}}_o)]\Sigma\right] \tag{2.32}$$

Then by substituting equation (2.32) into equation (2.31)

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}\mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] \approx \mathbb{E}_{\boldsymbol{x}}[\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}})] - \mathrm{tr}\left[[I(\hat{\boldsymbol{\theta}}_o)]\Sigma\right] \tag{2.33}$$

Conventionally, the information criterion is in the form of minimizing

$$-2\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + 2\,\mathrm{tr}[[I(\hat{\boldsymbol{\theta}}_o)]\Sigma] \tag{2.34}$$

By assuming $\mathrm{tr}[[I(\hat{\boldsymbol{\theta}}_o)]\Sigma] = K$ where K is the total number of free parameters in the mixture model, we can use AIC for model selection by minimizing

$$-2\log g(\boldsymbol{x}|\hat{\boldsymbol{\theta}}) + 2K \tag{2.35}$$

Bayesian information criterion

Another approach for the model selection is derived from the Bayesian framework. The derivation of BIC holds both the model set and the data-generating model fixed as sample size goes to infinity. It is also clear that if the model contains the true model, then BIC selection converges with probability one. A critical quantity to be approximated is the marginal probability of the data:

$$\int \left[\prod_{i=1}^{n} g(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})\right] \pi(\hat{\boldsymbol{\theta}})\, d\hat{\boldsymbol{\theta}} \tag{2.36}$$

which can be rewritten in the form of likelihood

$$\int [\ell(\hat{\boldsymbol{\theta}}|\boldsymbol{x}, g)]\pi(\hat{\boldsymbol{\theta}})\, d\hat{\boldsymbol{\theta}} \tag{2.37}$$

where $\boldsymbol{x}$ represents the data. Under general regularity conditions, as sample size increases, the log likelihood function can be approximated using a second-order Taylor series as

$$\log \ell(\hat{\boldsymbol{\theta}}|\boldsymbol{x}, g) = \log \ell(\hat{\boldsymbol{\theta}}_o|\boldsymbol{x}, g) - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o)^{\mathrm{T}} I(\hat{\boldsymbol{\theta}}_o)(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o) \tag{2.38}$$

Therefore the marginal probability of the data is

$$\ell(\hat{\boldsymbol{\theta}}_o|\boldsymbol{x}, g)\int \exp\left[-\frac{1}{2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o)^{\mathrm{T}} I(\hat{\boldsymbol{\theta}}_o)(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o)\right] \tag{2.39}$$

On the other hand,

$$\int (2\pi)^{-K/2} |I(\hat{\boldsymbol{\theta}}_o)|^{1/2} \exp\left[-\frac{1}{2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o)^{\mathrm{T}} I(\hat{\boldsymbol{\theta}}_o)(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o)\right] d\hat{\boldsymbol{\theta}} = 1 \tag{2.40}$$

$$H(\hat{\boldsymbol{\theta}}_o) = nH_1(\hat{\boldsymbol{\theta}}_o) \tag{2.41}$$

$H_1(\hat{\boldsymbol{\theta}}_o)$ is independent of sample size and will converge to $H_1(\hat{\boldsymbol{\theta}})$. Then,

$$\begin{aligned} &\log \int \left[\prod_{i=1}^{n} g(\boldsymbol{x}_i|\hat{\boldsymbol{\theta}})\right] \pi(\hat{\boldsymbol{\theta}})\, d\hat{\boldsymbol{\theta}} \\ =&\log(\ell(\hat{\boldsymbol{\theta}}_o|\boldsymbol{x}, g)) - \frac{1}{2}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o)^{\mathrm{T}} I(\hat{\boldsymbol{\theta}}_o)(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_o) \\ =&\log(\ell(\hat{\boldsymbol{\theta}}_o|\boldsymbol{x}, g)) + \frac{K}{2}\log(n) - \frac{K}{2}\log(2\pi) - \log(|I(\hat{\boldsymbol{\theta}}_o)|) \end{aligned} \tag{2.42}$$

In previous literature, the last two terms with higher orders are usually dropped and the BIC value is defined as

$$-2\log(\ell(\hat{\boldsymbol{\theta}}_o|\boldsymbol{x}, g)) + K\log(n) \tag{2.43}$$

Minimizing the BIC value is another useful method for model selection.

Chapter 3

APPLICATION AND RESULTS

3.1 Bayesian networks for religion-conflict data

To provide a straightforward representation of interactions of the predictors and conflict variables from the religion-conflict data, Bayesian networks are implemented encoding the "relationship" as conditional probabilities. To facilitate comparison with SEM models [8], the model is built to represent the interactions of each pair of the three social conditions interacting with one conflict variable at a time as shown in figure in which the "conflict" block/node indicates one of the five conflict variables (e.g., one of the models examines the effect of value incompatibility and resource-power differential over prejudice). In the networks, the states of the nodes can be evaluated according to survey investigated by social psychology experts in the Global Group Relations Project [9].

According to the domain experts' knowledge, we assume that the social conditions are independent of each other in the absence of a conflict condition, which can be modeled as converging connections.

On the other hand, however, it is not the case that all conflict variables are independent of each other conditional on any of the social conditions. Diverging and serial connections are inappropriate for the religion-conflict data analysis. Therefore, in Figure 3.1, conflict variables are considered separately, i.e., only one conflict variable is considered at a time.

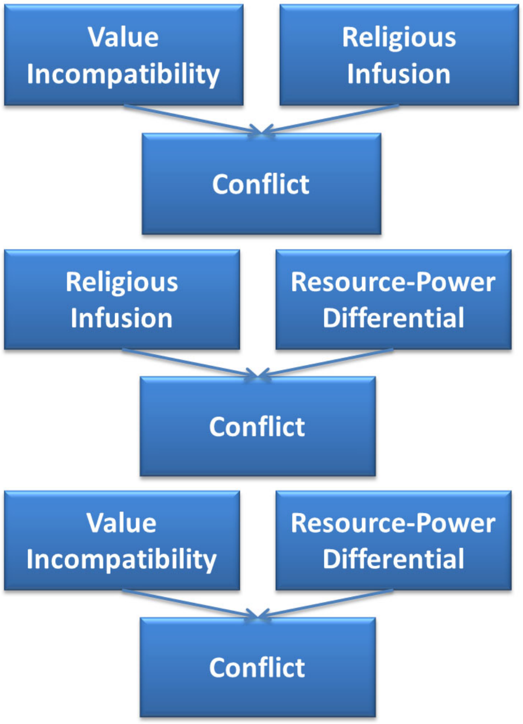


Figure 3.1: Bayesian networks of two social conditions and one conflict variable

If we use P_1, P_2, P_3 to represent the predictors and $C_i (i = 1, 2, 3, 4, 5)$ denotes one conflict variable, the Bayesian network above suggests that the essence of the problem lies in the probabilities $P(C_i|P_1, P_2)$, $P(C_i|P_1, P_3)$, and $P(C_i|P_2, P_3)$ $(i = 1, 2, 3, 4, 5)$. Moving beyond what was feasible with SEM modeling, we consider the impact of all three predictors simultaneously on each conflict variable in the network shown in Figure 3.2.

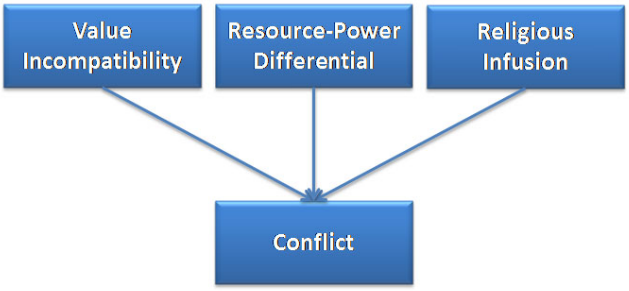


Figure 3.2: Bayesian networks of all social conditions and one conflict variable

Similarly, the 4-D model is built to explore probabilities $P(C_i|P_1, P_2, P_3)(i = 1, 2, 3, 4, 5)$.

According to the definition of conditional probability $P(A|B) = \frac{P(A,B)}{P(B)}$, a simple method to compute the conditional probabilities is obtaining the joint probability. Therefore, Gaussian mixture model is used to estimate the joint probability distribution.

3.2 Gaussian mixture model

Methods analysis and validation

In among the main ideas about optimizing the number of mixture components for mixture models, we build mixture model for the purpose of understanding interactions of social psychology concerns. For the multidimensional Gaussian mixture, the limited number of available observations (310) is not sufficient for a non-parametric method of estimating the order. And as suggested in literature [39], information criteria based on a penalized form of the likelihood are adequate for the problem of estimating unknown distributional shapes and density.

Before implementing the methods to the observations, the performance of AIC, BIC and the basic method maximizing the log-likelihood have been compared by testing using the EM algorithm to decompose the mixture models on artificial data. Gaussian mixtures are generated by pre-set mean, variance and number of random samples selected from each mixture component. Various scenarios have been designed for the testing. Issues have been taken into consideration, such as number of observations, number of mixture components, sparsity of the means for the clusters and the overlapping issue.

For the sake of visualizing the comparison, we first view the cases designed as one-dimensional Gaussian mixture models with two mixture components. Case 1 presents the scenario that two evenly weighted mixtures have mean values very close to each other and exactly the same variance. Case 2 builds one of the mixture components with relatively large variance. Case 3 weights one mixture component much more significantly than the other. The results are shown in Table 3.1 below, and the distributions are plotted in Figure 3.3.

Table 3.1: Testing scenarios

	Weights	Means	Variances	AIC	BIC	Log-likelihood
Case 1	$0.5, 0.5$	$1, 1.2$	$2, 2$	1	1	1
Case 2	$0.5, 0.5$	$0, 2$	$3, 10$	2	2	9
Case 3	$0.9, 0.1$	$0, 5$	$3, 3$	2	2	15

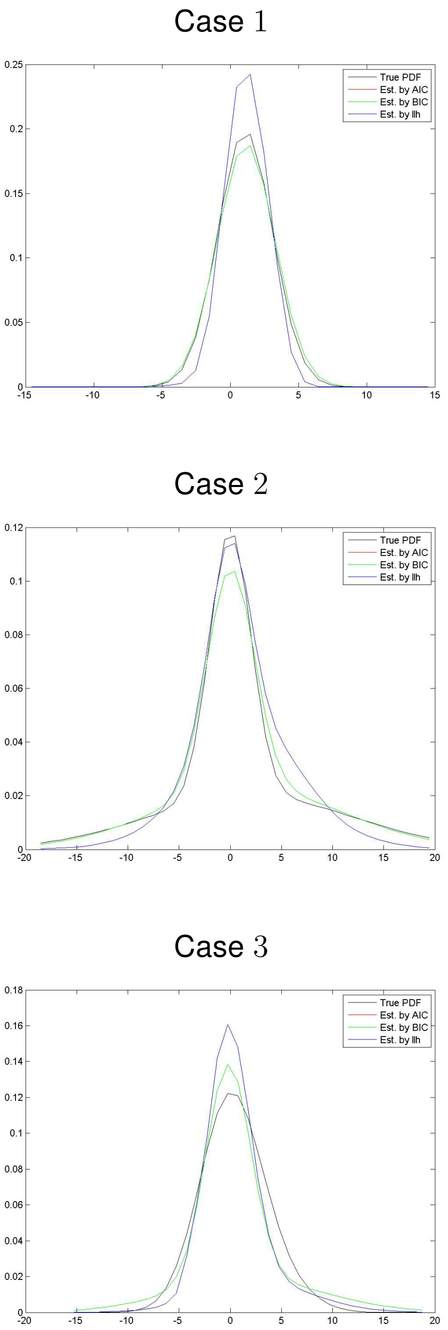


Figure 3.3: True distribution and estimated distributions

Although none of these methods is able to distinguish very close distributions, it is apparent that the over-fitting problem in the EM algorithm has been significantly avoided here by implementing AIC and BIC.

Table 3.2: Example of an artificial data set

3-D Data	4 Mixture Components			
	1st component	2nd component	3rd component	4th component
Weight	0.17	0.34	0.23	0.26
Mean	$\begin{bmatrix} 8.54 \\ 7.37 \\ 6.02 \end{bmatrix}$	$\begin{bmatrix} 0.86 \\ 0.47 \\ 8.78 \end{bmatrix}$	$\begin{bmatrix} 9.19 \\ 9.34 \\ 2.43 \end{bmatrix}$	$\begin{bmatrix} 4.26 \\ 1.57 \\ 8.27 \end{bmatrix}$
Covariance	$\begin{bmatrix} 1.42 & -0.61 & -0.10 \\ -0.61 & 1.37 & 0.03 \\ -0.10 & 0.03 & 0.55 \end{bmatrix}$	$\begin{bmatrix} 4.60 & -2.64 & 2.44 \\ -2.64 & 6.02 & 2.48 \\ 2.44 & 2.48 & 6.71 \end{bmatrix}$	$\begin{bmatrix} 6.72 & 0.34 & 1.12 \\ 0.34 & 8.46 & 1.28 \\ 1.12 & 1.28 & 5.61 \end{bmatrix}$	$\begin{bmatrix} 0.90 & -0.02 & 0.06 \\ -0.02 & 2.19 & -1.45 \\ 0.06 & -1.45 & 3.37 \end{bmatrix}$

Now we move forward using high-dimensional data to approximate the real-world data. According to the religion-conflict problem we try designing data similar to the real cases of 3-D and 4-D using expectations within the survey range and random covariance matrices.

In Table 3.2 we present an example set of three-dimensional artificial data with four mixture components and parameters as stated in the table.

Similar data sets are created of 3-D and 4-D data with mixture components every third number from 1 to 16 (i.e., $M = 1, 4, 7, 10, 13, 16$). And five groups using different combinations of random parameters for a certain number of components have been tested by AIC, BIC, and maximizing the likelihood. The numbers of components obtained for the artificial data are shown in Table 3.3.

Table 3.3: Estimated numbers of mixture components for artificial data sets

Dimensionality	Real number	AIC	BIC	Log-likelihood
3-dimensional data	1	$29, 29, 30, 20, 26$	$1, 1, 1, 1, 1$	$30, 30, 30, 30, 30$
	4	$30, 29, 28, 29, 27$	$4, 3, 2, 4, 4$	$30, 30, 30, 29, 30$
	7	$30, 27, 29, 30, 23$	$2, 6, 4, 4, 4$	$30, 30, 30, 30, 30$
	10	$28, 28, 30, 30, 28$	$3, 5, 6, 5, 6$	$28, 29, 30, 29, 30$
	13	$30, 28, 29, 29, 29$	$5, 3, 2, 5, 5$	$30, 28, 30, 29, 30$
	16	$29, 25, 30, 30, 29$	$3, 6, 4, 2, 3$	$30, 30, 30, 30, 30$
4-dimensional data	1	$30, 28, 30, 30, 28$	$1, 1, 1, 1, 1$	$30, 30, 30, 30, 30$
	4	$28, 27, 30, 29, 30$	$4, 3, 3, 3, 3$	$30, 30, 30, 30, 30$
	7	$29, 26, 29, 25, 30$	$5, 3, 6, 5, 4$	$29, 30, 29, 30, 30$
	10	$29, 26, 28, 29, 28$	$6, 6, 3, 4, 6$	$29, 30, 30, 29, 30$
	13	$28, 26, 27, 30, 29$	$4, 4, 4, 3, 4$	$30, 29, 30, 30, 29$
	16	$30, 30, 30, 30, 26$	$5, 4, 4, 5, 4$	$30, 30, 30, 30, 29$

The results reveal that BIC acts with better robustness and brings about higher accuracy especially when dealing with smaller number of mixtures. The result of BIC for the example given in Table 3.2 turns out to be accurate, as shown in Table 3.4. Here we present Table 3.5 comparing the original model and the recovered model from the data. The comparison is made over each mixture component. Error of weight is interpreted as the absolute value of the differences between corresponding Gaussian mixture component of the artificial model w_o and the estimation w_e,

$$\Delta_w = |w_e - w_o|. \tag{3.1}$$

Error of the estimated mean vectors μ_e from the truth μ_o is measured by Euclidean distance on $\mathbb{R}^n$

$$d(\mu_e, \mu_o) = \|\mu_e - \mu_o\|. \tag{3.2}$$

The covariance matrices are compared by listing their eigenvalues.

However, when it comes to complicated mixture models, such as cases with large number mixture components, BIC underestimates the number of components with the limited number of data samples.

Table 3.4: Results for the artificial data set example estimated by BIC

3-D Data	4 Mixture Components			
	1st component	2nd component	3rd component	4th component
Weight	0.17	0.34	0.23	0.26
Mean	$\begin{bmatrix} 8.57 \\ 7.27 \\ 5.96 \end{bmatrix}$	$\begin{bmatrix} 0.94 \\ 0.38 \\ 8.88 \end{bmatrix}$	$\begin{bmatrix} 9.40 \\ 9.44 \\ 2.49 \end{bmatrix}$	$\begin{bmatrix} 4.44 \\ 1.48 \\ 8.41 \end{bmatrix}$
Covariance	$\begin{bmatrix} 1.52 & -0.68 & -0.19 \\ -0.68 & 1.34 & 0.01 \\ -0.19 & 0.01 & 0.61 \end{bmatrix}$	$\begin{bmatrix} 4.09 & -2.28 & 2.44 \\ -2.28 & 5.31 & 2.20 \\ 2.44 & 2.20 & 6.52 \end{bmatrix}$	$\begin{bmatrix} 7.35 & 0.82 & 1.18 \\ 0.82 & 7.01 & 1.61 \\ 1.18 & 1.61 & 6.66 \end{bmatrix}$	$\begin{bmatrix} 0.92 & -0.06 & 0.03 \\ -0.06 & 2.01 & -1.18 \\ 0.03 & -1.18 & 3.17 \end{bmatrix}$

Table 3.5: Model estimation compared with truth

3-D Data		4 Mixture Components			
		1st component	2nd component	3rd component	4th component
Weight Error		0	0	0	0
Mean Error		0.11	0.15	0.24	0.24
Eigenvalues of Covariance Matrices	True Model	$2.01, 0.80, 0.54$	$8.88, 7.84, 0.62$	$9.18, 6.93, 4.68$	$4.35, 1.21, 0.90$
	Estimation	$2.12, 0.83, 0.52$	$8.35, 7.02, 0.54$	$9.42, 6.46, 5.15$	$3.90, 1.29, 0.91$

Solving the problem of density estimation of multi-dimensional religion-conflict data, since survey has provided very limited number of observations (310 after removing the missing data), avoiding over-fitting is the most significant issue. Therefore, BIC is selected for the analysis of religion-conflict data set.

Gaussian mixture model for religion-conflict data

After exploring the techniques, we now build a GMM for the religion-conflict data extracted from the survey and represented by score numbers. Three social predicting factors and five intergroup conflict variables are selected by social scientists to be the main focus. Two of the social factors (religious infusion and value incompatibility) are evaluated in the scale of 1 through 9, and the resource-power differential is measured by number between -6 and 6 where a positive value means the group has relatively greater resources and power than the other group and a negative value means a greater scarcity. All conflict variables come with the value somewhere between 1 and 9. In the statistical model estimation, we make an assumption that all data sets are continuous and unbounded.

According to the Bayesian networks created for the religion-conflict data, the probabilities of interests are probabilities of conflict variables conditional on the predictors. There are three social factors acting as predictors, and five conflict variables, thus we have $\binom{2}{3}\binom{1}{5} = 15$ sets of three-dimensional data and $\binom{3}{3}\binom{1}{5} = 5$ sets of four-dimensional data.

Given the observations, mixture model has been built for each set of data implementing the EM algorithm using BIC to determine the number of mixtures. More specifically, to achieve good estimation, convergence is defined by a threshold of 10^{-6} and the maximum number of iterations is set to be 400. Three hundred different random initializations have been attempted for the minimum BIC value.

Here we only present the result of mixture components for the religion-conflict data shown in Table 3.6 and later the probabilities will be shown via a straightforward color visualization technique.

Table 3.6: Number of mixture components for religion-conflict data

Conflict Type	Predictor Variables			
	Value and Resource-power	Religious Infusion and Value	Religious Infusion and Resource-power	All Three predictors
Prejudice	11	7	5	5
Interpersonal Discrimination	12	4	4	4
Symbolic Aggression	12	6	6	7
Individual Violence	6	7	10	8
Collective Violence	8	6	8	7

After successfully estimating the joint probabilities of the predictors and the variables using Gaussian mixture models, the next step is to achieve the goal shown in the Bayesian networks by obtaining the conditional probabilities. Again to simplify the expression, we use P_1, P_2, P_3 to represent the predictors and $C_i(i = 1, 2, 3, 4, 5)$ to be the conflict variable. Here we take two predictor cases for example of further derivation, noting that the same method can be implemented in three predictor cases. The joint probability distribution is estimated by the mixture model and can be represented using conditional probability densities

$$\begin{aligned} p(C_i, P_j, P_k) &= \sum_{l=1}^{M} w_l p_l(C_i, P_j, P_k) \\ &= \sum_{l=1}^{M} w_l p_l(C_i|P_j, P_k) p_l(P_j, P_k) \end{aligned} \tag{3.3}$$

where for each mixture we have

$$p_l(C_i, P_j, P_k) = p_l(C_i|P_j, P_k) p_l(P_j, P_k) \tag{3.4}$$

If we view the probability $p_l(P_j, P_k)$ as the marginal probability for the predictors, since we know the joint probability, this marginal can be derived by integrating over C_i

$$\begin{aligned} p(P_j, P_k) &= \int\limits_{C_j} \sum_{l=1}^{M} w_l p_l(C_i, P_j, P_k)\, dC_j \\ &= \sum_{l=1}^{M} w_l \int\limits_{C_j} p_l(C_i, P_j, P_k) \\ &= \sum_{l=1}^{M} w_l p_l(P_j, P_k). \end{aligned} \tag{3.5}$$

Therefore, the conditional probability density function is

$$\begin{aligned} p(C_j|P_j, P_k) &= \frac{p(C_i, P_j, P_k)}{p(P_j, P_k)} \\ &= \frac{\sum\limits_{l=1}^{M} w_l p_l(C_i|P_j, P_k) p_l(P_j, P_k)}{\sum\limits_{r=1}^{M} w_r p_r(P_j, P_k)} \\ &= \sum_{l=1}^{M} \frac{w_l p_l(P_j, P_k)}{[\sum\limits_{r=1}^{M} w_r p_r(P_j, P_k)]} p_l(C_i|P_j, P_k) \end{aligned} \tag{3.6}$$

which is also a mixture model with weights $\frac{w_l p_l(P_j,P_k)}{[\sum\limits_{r=1}^{M} w_r p_r(P_j,P_k)]}$ for $l = 1, \ldots, M$.

Based on equation (3.4), we now discuss the case of a single mixture component of the joint probability. This component is subject to the Gaussian distribution

$$\begin{aligned} p_l(C_i, P_j, P_k) = \sum_{l=1}^{M} w_l \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_l(C_i, P_j, P_k)|^{\frac{1}{2}}} \\ \exp\left(- \frac{(\boldsymbol{x} - \mu_l(C_i, P_j, P_k))^{\mathrm{T}} \Sigma_l(C_i, P_j, P_k)^{-1} (\boldsymbol{x} - \mu_l(C_i, P_j, P_k))}{2} \right) \end{aligned} \tag{3.7}$$

in which the parameters are

$$\Sigma_l(C_i, P_j, P_k) = \begin{bmatrix} \mathrm{var}_l(C_i) & \mathrm{cov}_l(C_i, P_j) & \mathrm{cov}_l(C_i, P_k) \\ \mathrm{cov}_l(P_j, C_i) & \mathrm{var}_l(P_j) & \mathrm{cov}_l(P_j, P_k) \\ \mathrm{cov}_l(P_k, C_i) & \mathrm{cov}_l(P_k, P_j) & \mathrm{var}_l(P_k) \end{bmatrix} \tag{3.8}$$

$$\mu_l(C_i, P_j, P_k) = \begin{bmatrix} \mu_l(C_i) \\ \mu_l(P_j) \\ \mu_l(P_k) \end{bmatrix} \tag{3.9}$$

from which we can observe that $p_l(C_i|P_j, P_k)$ is Gaussian with

$$\mu_l(C_i|P_j, P_k) = \mu_l(C_i) + \begin{bmatrix} \mathrm{cov}_l(C_i, P_j) & \mathrm{cov}_l(C_i, P_k) \end{bmatrix}$$

$$\begin{bmatrix} \mathrm{var}_l(P_j) & \mathrm{cov}_l(P_j, P_k) \\ \mathrm{cov}_l(P_k, P_j) & \mathrm{var}_l(P_k) \end{bmatrix}^{-1} \left\{ \begin{bmatrix} P_j \\ P_k \end{bmatrix} - \begin{bmatrix} \mu(P_j) \\ \mu(P_k) \end{bmatrix} \right\},$$

and

$$\Sigma_l(C_i|P_j, P_k) = \underset{l}{\mathrm{var}}(C_i) - \begin{bmatrix} \mathrm{cov}_l(C_i, P_j) & \mathrm{cov}_l(C_i, P_k) \end{bmatrix}$$

$$\begin{bmatrix} \mathrm{var}_l(P_j) & \mathrm{cov}_l(P_j, P_k) \\ \mathrm{cov}_l(P_k, P_j) & \mathrm{var}_l(P_k) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{cov}(P_j, C_i) \\ \mathrm{cov}(P_k, C_i) \end{bmatrix}.$$

Therefore, $p(C_j|P_j, P_k)$ has been proven to be another Gaussian mixture model with known parameters. Similarly, we have $p(C_j|P_1, P_2, P_3)$ to be a Gaussian mixture model as well. The next step is presenting a clear visualization for the 3-D and 4-D Gaussian mixture distributions.

3.3 RGB color visualization

Visualization of probability densities enables analysts to directly view trends and shapes of the data distributions. It usually gives impetus to significant analysis and novel findings. Curves, surfaces and meshes have been served as helpful tools for illustrating probabilities. However, depending on the particular nature of this intergroup religion-conflict interaction research, three or four variables should be taken into consideration together plus the probabilities. Dimensionality now becomes the main issue for two-dimensional visualization. Here we propose a color mapping method inspired by the idea of heat map which is able to use two-dimensional Figure to display three-dimensional data [54]. Two-dimensional images are organized as x-by-y pixels/cells with respect to the integer values within the scales of the two predictors. The color encodes level of conflict while the intensity of color encodes the conditional probability.

In the visualization of the mixture model results for religion-conflict data, we use RGB color intensities to represent the probability of each conflict conditional in every cell (which means given values of the social predictors). The color intensities of red, green and blue corresponding to the probabilities range from 0 to 1, where 0 represents the lowest intensity and 1 represents the highest intensity. According to the survey design [9] and psychology experts' knowledge, we focus on the analysis of particular values of intergroup conflicts. Levels are set for conflict variables and integrates the conditional probabilities over each level setting. Probability of low conflict value $P(C_i \leqslant 2|P_j, P_k)$ or $P(C_i \leqslant 2|P_1, P_2, P_3)$ is encoded by the intensity of blue, and probability of high conflict value $P(C_i > 6|P_j, P_k)$ or $P(C_i > 6|P_1, P_2, P_3)$ encoded by the intensity of red. Probability of medium conflict value $P(2 < C_i \leqslant 6|P_j, P_k)$ or $P(2 < C_i \leqslant 6|P_1, P_2, P_3)$ is simply represented by black $(R = 0, G = 0, B = 0)$ to avoid confusion or negative effect over the analysis of the severe cases of conflicts.

Through the novel visualization method, explorations have been made for all five conflicts. The displays in Figure 3.4 show the prediction of prejudice by two social factors at a time.

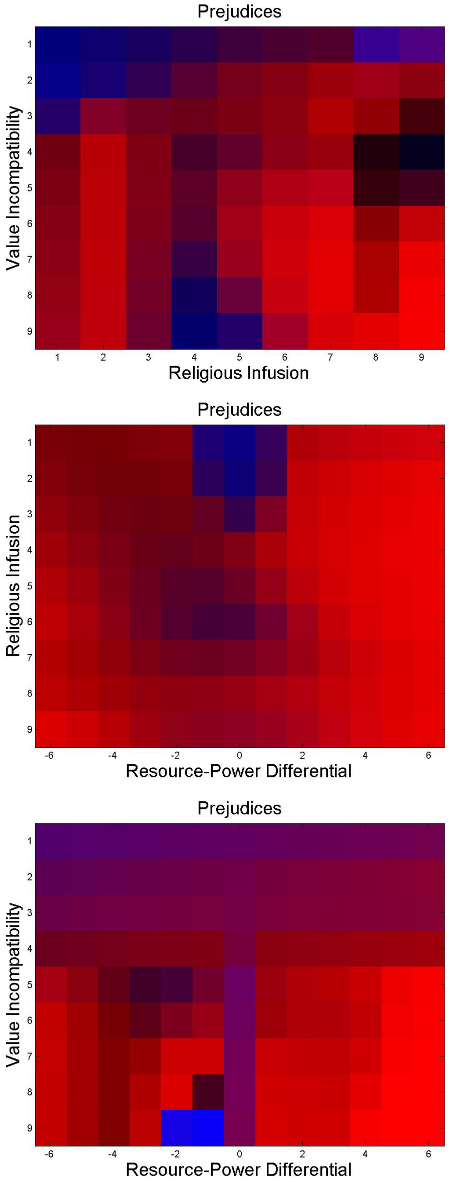


Figure 3.4: Visualization for two-predictor analysis of prejudice

From the figures above we can observe that the increment of the probability of high prejudice and the decrement of low prejudice follow the increasing value incompatibility, religious infusion, and/or the severe case of resource-power differential, both negative and positive. And the interaction with religious infusion enables the value incompatibility to have a stronger effect over prejudice. Similar patterns and trends as shown above are discovered in the analysis of interpersonal discrimination as well.

However, some different patterns are discovered while exploring other conflicts including aggression, assaults and collective violence. The following figures illustrate the interactions between social predictors and the collective violence.

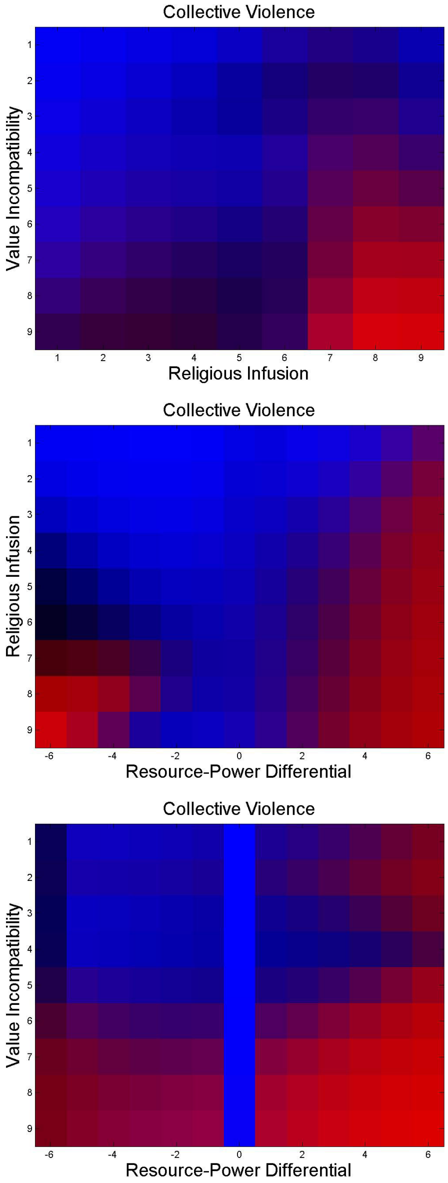
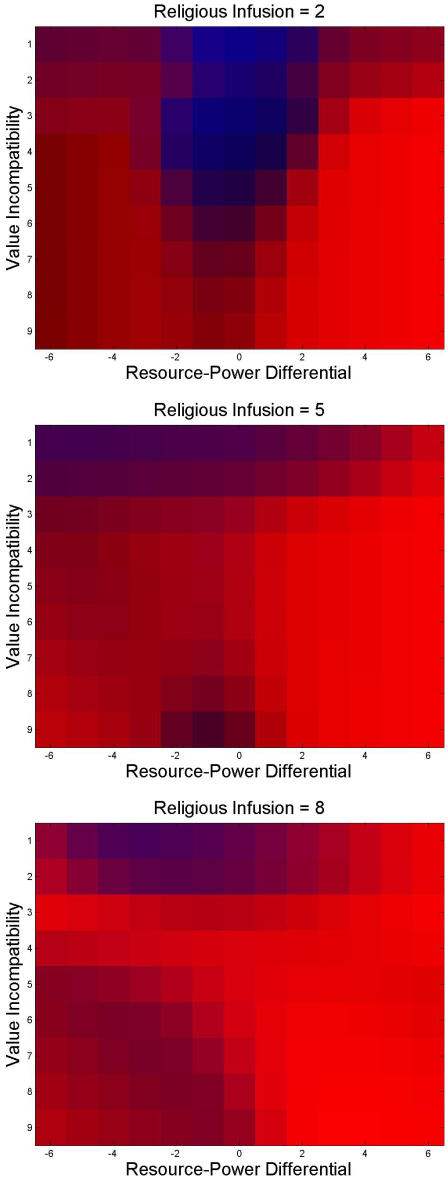


Figure 3.5: Visualization for two-predictor analysis of collective violence

We still can see as the predictors' value increases, the conflict of collective violence has a higher probability to be strong. But in this case, the color maps reveal that in the groups where religious infusion has a low value, and the resource-power differential is negative, people avoid having any strong collective violence. And these patterns can also be detected in the analysis of symbolic aggression and individual violence.

When we consider the interaction of three predictors and the conflict variables, the results turn out as depicted in Figure 3.6.

In the analysis considering all the three predictors at the same time, the visualizations are of considerable value in interpreting the data. However, the higher the dimension the more strongly the curse of dimensionality enters as data sparsity, which may result in a estimation with much higher error rate. Therefore, we use this model as a suggestive material of the analysis. When larger data sets are obtained from further surveys, these methods should be more helpful.



Religious Infusion = 2

Religious Infusion = 5

Religious Infusion = 8

Figure 3.6: Visualization for three-predictor analysis of prejudice

Chapter 4

CONCLUSIONS AND FUTURE WORK

In this thesis, we studied signal processing methods of Bayesian networks, Gaussian mixture models, and information criteria for model selection. We also developed a novel visualization method for the multi-dimensional distributions. Artificial data has created for testing the methods, and Gaussian mixture models achieved using AIC and BIC were compared to true distributions. The testing results suggest that AIC works well with relatively small data samples, and low dimensionality, but tends to give large estimated numbers of mixture components (larger than 20 in the testing scenario). But the Bayesian method works well in dealing with large data samples and relatively small numbers of mixture components (most of which are smaller than 10 in the testing scenario).

We used combinations of these methods to solve a non-traditional signal problem in which the social psychology survey data are treated as the signal. To facilitate the analysis, statistical models were effectively built for the multi-dimensional and sparse data samples using BIC for the Gaussian mixture model selection to avoid over-fitting problem. According to the Bayesian networks built for the religion-conflict data, the interactions were modeled as conditional probabilities from the Gaussian mixture models.

A RGB color visualization technique were proposed to display the probabilities of interest by intensities of red and blue. The color representations enabled intuitive and direct observations about the effect of religion, interacting with value incompatibility and resource-power differential, in predicting different levels of intergroup conflicts.

In this first attempt to implement Bayesian techniques and statistical models into the religion-conflict problem analysis, it was assumed that the survey data may be sufficiently well approximated by a model involving a continuous and unbounded distribution. It was also assumed that the mixture model components are Gaussian and the mixture components were estimated based on Bayesian information criterion. Also, by integrating the probability densities over three ranges, we may have lost some valuable information. Therefore, future work is expected including development and improvement of the mixture model structure selection and model parameter estimation. And as complexity of the model increases, the visualization method needs to be improved using, for example, HSV color space instead of RGB color maps.

REFERENCES

[1] D. L. Horowitz, “Structure and strategy in ethnic conflict: A few steps toward synthesis,” *Annual World Bank Conference on Development Economics*, vol. 1, pp. 345–370, 1998.

[2] D. A. Lake and D. Rothchild, “Containing fear: The origins and management of ethnic conflict,” *International Security*, vol. 21, pp. 41–75, 1996.

[3] L. E. Cederman, N. B. Weidmann, and K. S. Gleditsch, “Horizontal inequalities and ethnonationalist civil war: A global comparison,” *The American Political Science Review*, vol. 105, pp. 478–495, 2011.

[4] P. Collier and A. Hoeffler, “Greed and grievance in civil war,” *Oxford Economic Papers*, vol. 56, pp. 563–595, 2004.

[5] C. Kimball, *When Religion Becomes Evil*, Harper San Francisco, 2002.

[6] M. Rokeach, *Beliefs, Attitudes and Values: A Theory of Organization and Change*, Jossey-Bass, 1972.

[7] R. Inglehart and P. Norris, *Sacred and Secular: Religion and Politics Worldwide*, Cambridge University Press, 2004.

[8] C. Warner, S. Neuberg, S. Mistler, E. Hill, and A. Berlin, “Religious infusion and intergroup conflict: Results from the global group relations project,” in *APSA 2011 Annual Meeting*, 2011.

[9] S. L. Neuberg, C. M. Warner, S. A. Mistler, A. Berlin, E. D. Hill, J. D. Johnson, P. Mahanti, H. Liu, G. Filip-Crawford, R. E. Millsap, T. J. Taylor, G. Thomas, M. Winkelman, B. J. Broome, and J. Schober, “Religious infusion predicts enhanced resource- and values-linked intergroup conflict,” unpublished.

[10] G. Schwarz, “Estimating the dimension of a model,” *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] H. Akaike, “A new look at the statistical model identification,” *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[12] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using Gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[13] I. Bilik, J. Tabrikian, and A. Cohen, “GMM-based target classification for ground surveillance Doppler radar,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 42, no. 1, pp. 267–278, 2006.

[14] J. Y. Kim, D. Y. Ko, and S. Y. Na, “Implementation and enhancement of GMM face recognition systems using flatness measure,” in *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication*, 2004, pp. 247–251.

[15] H. Greenspan, A. Ruf, and J. Goldberger, “Constrained Gaussian mixture model framework for automatic segmentation of MR brain images,” *IEEE Transactions on Medical Imaging*, vol. 25, no. 9, pp. 1233–1245, 2006.

[16] J. Pearl, “Bayesian networks: A model of self-activated memory for evidential reasoning,” Tech. Rep., UCLA, Computer Science Department, 1985.

[17] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, 1988.

[18] R. E. Neapolitan, *Learning Bayesian Networks*, Pearson Prentice Hall, 2004.

[19] D. Heckerman and D. M. Chickering, “Learning Bayesian networks: The combination of knowledge and statistical data,” *Machine Learning*, vol. 20, pp. 197–243, 1995.

[20] K. Timo and M. N. John, *Bayesian Networks: An Introduction*, Wiley, 2009.

[21] G. Zweig, “Bayesian network structures and inference techniques for automatic speech recognition,” *Computer Speech and Language*, vol. 17, pp. 173–193, 2003.

[22] L. Xie and H. W. Yang, “Dynamic Bayesian network inversion for robust speech recognition,” *IEICE Transactions on Information and Systems*, vol. E90D, pp. 1117–1120, 2007.

[23] L. Zhang and Q. Ji, “A Bayesian network model for automatic and interactive image segmentation,” *IEEE Transactions on Image Processing*, vol. 20, pp. 2582–2593, 2011.

[24] K. Jayech and M. M. Ali, “Clustering and Bayesian network for image of faces classification,” *International Journal of Advanced Computer Sciences and Applications*, 2011.

[25] G. Quer, H. Meenakshisundaram, B. R. Tamma, B. S. Manoj, R. Rao, and M. Zorzi, “Using Bayesian networks for cognitive control of multi-hop wireless networks,” in *Proceedings of the Military Communications Conference, MILCOM 2010*, 2010.

[26] F. Markowetz and R. Spang, “Inferring cellular networks: A review,” *BMC Bioinformatics*, vol. 8, pp. S5, 2007.

[27] J. Lin and P. Huang, “Exploiting missing clinical data in Bayesian network modeling for predicting medical problems,” *Journal of Biomedical Informatics*, vol. 41, pp. 1–14, 2008.

[28] Y. Yang, “A consistency contribution based Bayesian network model for medical diagnosis,” *Journal of Biomedical Science and Engineering*, vol. 3, pp. 488–495, 2010.

[29] H. Langseth and L. Portinale, “Bayesian networks in reliability,” *Reliability Engineering and System Safety*, vol. 92, pp. 92–108, 2007.

[30] D. Woof, M. Goldstein, and F. Coolen, “Bayesian graphical models for software testing,” *IEEE Transactions on Software Engineering*, vol. 28, pp. 510–525, 2002.

[31] V. Gowadia, C. Farkas, and M. Valtorta, “PAID: A probabilistic agent-based intrusion detection system,” *Computers and Security*, vol. 24, pp. 529–545, 2005.

[32] A. Darwiche, “Bayesian networks,” *Communications of the ACM*, vol. 53, pp. 80–90, 2010.

[33] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[34] F. V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[35] K. Pearson, “Contributions to the mathematical theory of evolution,” *Philosophical Transactions of the Royal Society of London*, vol. 185, 1894.

[36] H. Jeffreys, “An alternative to the rejection of observations,” *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 137, no. 831, pp. 78–87, 1932.

[37] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[38] R. W. Butler, “Predictive likelihood inference with applications,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 48, pp. 1–38, 1986.

[39] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2000.

[40] S. Borman, “The expectation maximization algorithm: A short tutorial,” Unpublished paper, available online.

[41] J. P. Harding, “The use of probability paper for the graphical analysis of polymodal frequency distributions,” *Journal of the Marine Biological Association of the United Kingdom*, vol. 28, pp. 141, 1949.

[42] R. M. Cassie, “Some uses of probability paper in the analysis of size frequency distributions,” *Australian Journal of Marine and Freshwater Research*, vol. 5, pp. 513–522, 1954.

[43] E. B. Fowlkes, “Some methods for studying the mixture of two normal (lognormal) distributions,” *Journal of the American Statistical Association*, vol. 74, pp. 561–575, 1979.

[44] B. W. Silverman, “Using kernel density estimates to investigate multimodality,” *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 43, pp. 97–99, 1981.

[45] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*, vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, 1986.

[46] J. J. Heckman, R. Robb, and J. R. Walker, “Testing the mixture of exponentials hypothesis and estimating the mixing distribution by the methods of moments,” *Journal of the American Statistical Association*, vol. 85, pp. 582–589, 1990.

[47] W. D. Furman and B. G. Lindsay, "Testing for the number of components in a mixture of normal distributions using moment estimators," *Computational Statistics and Data Analysis*, vol. 17, pp. 473–492, 1994.

[48] D. Dacunha-Castelle and E. Gassiat, "The estimation of the order of a mixture model," *Bernoulli*, vol. 3, pp. 279–299, 1997.

[49] M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, vol. 90, pp. 577–588, 1995.

[50] C. P. Robert, *Mixtures of distributions: Inference and estimation*, chapter 24, pp. 441–464, Chapman and Hall, 1996.

[51] K. Roeder and L. Wasserman, "Practical density estimation using mixtures of normals," *Journal of the American Statistical Association*, vol. 92, pp. 894–902, 1997.

[52] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[53] D. R. Anderson and K. P. Burnham, *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*, Springer, 2002.

[54] M. T. Freedman and T. Osicka, "Heat maps: An aid for data analysis and understanding of ROC CAD experiments," *Academic Radiology*, vol. 15, pp. 249–259, 2008.