Proceedings of the 2007 IEEE Symposium on
Foundations of Computational Intelligence (FOCI 2007)

Sparsity Promotion Models for the Choquet Integral

Andres Mendez-Vazquez
Department of Computer and Information
Science and Engineering
University of Florida -P.O. Box 116120
Gainesville, FL 32611-6120 USA
Email: amendez-@cise.ufl.edu

Paul Gader
Department of Computer and Information
Science and Engineering
University of Florida -P.O. Box 116120
Gainesville, FL 32611-6120 USA
Email: pgader@cise.ufl.edu

***Abstract*— In this paper, we present a novel algorithm for learning fuzzy measures for Choquet integration. There are two novel aspects of the algorithm: it seeks to explicitly reduce the number of nonzero parameters in the measure to eliminate noninformative or useless information sources and it uses a Bayesian model for parameter estimation which has not been previously applied to the fuzzy measure learning problem. The method uses a hierarchical model that implements a sparsity promotion algorithm through a Gibbs sampler. This approach builds on the methods proposed by Figueiredo et. al which uses Expectation Maximization (EM) to maximize the Least Absolute Shrinkage and Selection Operator (LASSO) criterion under a distribution that promotes sparsity. Additional constraints are needed to satisfy the requirements of fuzzy measures. Figueiredo's algorithm does not have a mechanism for imposing these constraints. The constraints are imposed by sequentially exploring the lattice tree of the power set and requiring that each fuzzy measure value assigned to a set lies in the domain of a truncated Gaussian determined by the fuzzy measures of supersets of the set under consideration.**

I. Introduction

Methods and motivations for learning fuzzy measures are well-documented in the literature [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11], [12] Previous approaches to learning fuzzy measures have generally relied on heuristic methods or on methods from classical optimization such as gradient descent, quadratic programming, etc. In this paper, we adapt a probabilistic approach to classification and regression model design to the measure learning problem for Choquet integration. This method seeks to simultaneously accomplish the classification or regression task while forcing as many model parameters to zero as possible. The classification or regression task is supported by requiring the difference between actual and desired outputs to come from a Gaussian distribution with zero mean. Parameters are driven to zero by forcing them to have distributions with high probabilities of zero values, such as the Laplacian distribution. These latter distributions are referred to as sparsity promoting distributions. There are multiple motivations for sparsity promotion. One is that sparse algorithms are more resistant to overtraining. Another is that by eliminating parameters (setting them to zero), we may eliminate redundant or useless information sources.

The paper is divided into nine sections. In the first, fuzzy measures and Choquet integration are defined. The second explores the basic model used by Figueiredo et. al [13]. The next five sections develop the idea of using the Gibbs Sampler [14], [15], [16] for learning fuzzy measures used in the Choquet integral to fuse information from different algorithms. In the next section, we describe two experiments that illustrate the capabilities of the new algorithm to handle the learning and sparsity promotion. In the conclusion, we summarize our results and discuss possible future research.

II. Choquet Integral

We first define some of the basic concepts behind the theory of fuzzy measures. These definitions can be found in [17], [5], [1], [18].

Definition 1: Let $X = \{x_1, \ldots, x_n\}$ be any finite set. A discrete **fuzzy measure** on X is a function $\mu : 2^X \to [0, 1]$ with properties

1) $\mu(\emptyset) = 0$ and $\mu(X) = 1$.
2) Given $A, B \in 2^X$, if $A \subset B$ then $\mu(A) \leq \mu(B)$.

For our purposes, the set *X* is considered to contain the names of sources of information (features, algorithms, agents, features, sensors,etc), and for a subset $A \subseteq X$, $\mu(A)$ is considered to be the *worth* of this subset of information.

To fuse evidence supplied by different sources of information from a discrete fuzzy set of *X*, we use the discrete **Choquet integral** [19], [12], [17], [10], [9], [20], [21]

Definition 2: Let f be a function from $X = \{x_1, \ldots, x_n\}$ to $[0, 1]$. Let $\{x_{\sigma(1)}, \ldots, x_{\sigma(n)}\}$ denote a reordering of the set X such that $0 \leq f(x_{\sigma(1)}) \leq \ldots \leq f(x_{\sigma(n)})$, and let $A_{(i)}$ be a collection of subsets defined by $A_{(i)} = \{x_{\sigma(i)}, \ldots, x_{\sigma(n)}\}$. Then, the discrete **Choquet integral** of f with respect to a fuzzy measure μ on X is defined as

$$C_\mu(f) = \sum_{i=1}^{n} \mu(A_{(i)})(f(x_{(i)}) - f(x_{(i-1)})) \tag{1}$$

where we take $f\left(x_{(0)}\right) \equiv 0$, $A_{(n+1)} \equiv \emptyset$ and $x_{(i)} \equiv x_{\sigma(i)}$.

III. Hierarchical Sparsity Model

Consider the generalized linear model

$$g(\boldsymbol{x}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \mu_i h_i(\boldsymbol{x}) = H\boldsymbol{\mu},$$
$$y = g(\boldsymbol{x}, \boldsymbol{\mu}) + \epsilon \text{ and } \epsilon \sim N(0, \sigma^2), \tag{2}$$

where y is the desired output for the algorithm, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$ is a vector of model parameters, and $H = (h_1(\boldsymbol{x}), \ldots, h_m(\boldsymbol{x}))$ is a collection of functions of the input vector $\boldsymbol{x}$. It is assumed that the difference between the desired output, and the computed output $g(\boldsymbol{x}, \boldsymbol{\mu})$ is a zero mean Gaussian ϵ. We seek to maximize the likelihood of $\boldsymbol{\mu}$ and σ^2 given the output of the y data, $L(\boldsymbol{\mu}, \sigma^2|y)$. For this, we assume a distribution for each coordinate of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$. In our specific case each μ_j is a measure of an element of the power set 2^X. We would like to eliminate the positions in $\boldsymbol{\mu}$ that do not have much influence on the output. Figueiredo [13] proposed the following hierarchical model to accomplish this

$$\begin{aligned} y_i &\sim N(H_i\boldsymbol{\mu}, \sigma^2)\ \forall i = 1, \ldots, n \\ \mu_j &\sim N(0, \tau_j)\ \forall j = 1, \ldots, m \\ \tau_j &\sim \exp\left(\frac{\gamma}{2}\right)\ \forall j = 1, \ldots, m \\ \gamma &\sim \text{non-informative prior} \\ \sigma^2 &\sim \text{non-informative prior} \end{aligned} \tag{3}$$

where the $y_1, \ldots, y_n$ represent our known data or labels, the $\mu_1, \ldots, \mu_m$ represent the weights to be learned, the $\tau_1, \ldots, \tau_m$ represent the hidden variables that promote sparsity in the weights, γ is the amount of sparsity in the model , and σ^2 is the amount of noise in the model. It can be proved under this model that the μ_j have the following density

$$\mu_j \sim \frac{\sqrt{\gamma}}{2}\exp\left\{-\sqrt{\gamma}|\mu_j|\right\} \text{(Laplacian)}. \tag{4}$$

This density promotes sparsity in the $\boldsymbol{\mu}$ since it has a sharp peak at zero. If the components of $\boldsymbol{\mu}$ are samples from the Laplacian, then they are likely to be small. Assuming $\tau_1, ..., \tau_j$ as hidden data, an EM algorithm can be used to maximize the likelihood

$$\begin{aligned} L(\boldsymbol{\mu}, \sigma^2|y, \tau) = p(y, \tau|\boldsymbol{\mu}, \sigma^2) \propto \\ p(y|\boldsymbol{\mu}, \sigma^2)p(\boldsymbol{\mu}|\tau)p(\sigma^2). \end{aligned} \tag{5}$$

IV. Constraining the Hierarchical Model for Sparsity Promotion

A problem we have in the model (3) is that we cannot constrain elements in $\boldsymbol{\mu}$ that obey complex relations. For example, the fuzzy measure definition has the following relations:

1) $\mu(\emptyset) = 0$ and $\mu(X) = 1$.
2) Given $A, B \in 2^X$, if $A \subset B$ then $\mu(A) \leq \mu(B)$.

Let $K = |X|$. We seek to estimate the $2^K - 2$ parameters $\mu(A), A \subset X$. To simplify our notation, and treat a fuzzy measure as a parameter vector, we order the subsets of 2^X into the succession $\{A_1, A_2, ..., A_{2^K-1}\}$ and write $\mu_j = \mu(A_j)$. Now, we will impose the fuzzy measures relations using the following strategy based on the Gibbs sampler. Given the model for sparsity promotion in the section 3, and taking $j = 1, \ldots, m = |2^X| - 2$, we can try to calculate the joint probability $(\gamma, \sigma^2, \mu_1, \ldots, \mu_m,\ \tau_1, \ldots, \tau_m)$ using the following strategy. Looking at the structure of the power set lattice, for example $X = \{a, b, c\}$ in figure (1), starting at the top (i.e. X) going downward to the singelton elements.

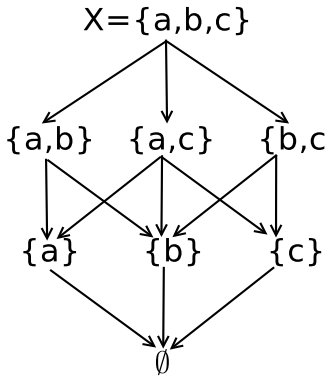

Fig. 1. Example of a lattice. Arrows represent the subset relation.

It is easy to observe that the value of a fuzzy measure on a particular set, for example $\mu(\{a\})$, depends on the previous values for the sets that contain $\mu(\{a\})$, i.e. $\mu(\{a, b\})$, $\mu(\{a, c\})$ and $\mu(\{a, b, c\})$. Because of this, we can constrain the value of each measure μ_j by sampling from a truncated Gaussian on the interval $[0, min\{\mu_{A_s}|A_j \subseteq A_s\}]$. Thus, the hierarchical model from section (3) can be modified in the following way

$$\begin{aligned} y_i &\sim N(H_i\boldsymbol{\mu}, \sigma^2)\ \forall i = 1, \ldots, n \\ \mu_m &= 1 \\ \mu_j &\sim N(0, \tau_j)|_{[0, min\{\mu_{A_s}|A_j \subseteq A_s\}]}\ \forall j = 1, \ldots, m-1 \\ \tau_j &\sim \exp\left(\frac{\gamma}{2}\right)\ \forall j = 1, \ldots, m-1 \\ \gamma &\sim \text{non-informative prior} \\ \sigma^2 &\sim \text{non-informative prior} \end{aligned}$$

In this model the row vectors H_i are composed of zeros and the differences $\left[f(x_{(k)}) - f(x_{(k-1)})\right]$ corresponding to the correct positions of the vector $\boldsymbol{\mu}^T = (\mu_1, \mu_2, ..., \mu_m)^t$. Hence, the Choquet integral can be written as an inner product

$$\begin{aligned} C_\mu(f_i) &= \sum_{k=1}^{n} \mu(A_{(k)})(f_i(x_{(k)}) - f_i(x_{(k-1)})) \\ &= H_i \cdot \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} \end{aligned} \tag{6}$$

V. Gibbs Sampler for the New Model

Knowing that the Gibbs sampler is a variation of the generalized Expectation Maximization algorithm [15], and given that fact that we can constraint our samples to be taken in defined sets. Then, we can use the Gibbs sampler as a way to learn the fuzzy measures.

With the previous idea, we can see that given an initial point $(\gamma_0, \sigma_0^2, \mu_1^0, \ldots, \mu_m^0, \tau_1^0, \ldots, \tau_m^0) = (\gamma_0, \sigma_0^2, \boldsymbol{\mu}^0, \boldsymbol{\tau}^0)$ and assuming the samples $y_1, \ldots, y_n$ identically independent, the

Gibbs sampler looks like:

1) $\mu_j^{t+1} \sim p(\mu_j|\gamma_t, \sigma_t^2, \boldsymbol{\mu}_{-j}^t, \boldsymbol{\tau}^t, y_1, ..., y_n) = \frac{p(\mu_j, y_1, ..., y_n|\gamma_t, \sigma_t^2, \boldsymbol{\mu}_{-j}^t, \boldsymbol{\tau}^t)}{p(y_1, ..., y_n)} \propto p(y_1, ..., y_n|\sigma_t^2, \boldsymbol{\mu}^t) p(\mu_j|\tau_j, \{\mu_s|A_j \subseteq A_s\}) \quad \forall \quad j = 1, \ldots, m-1$ with $\mu_j \sim N(0, \tau_j)|_{[0, min\{\mu_s|A_j \subseteq A_s\}]}$
2) $\tau_j^{t+1} \sim p(\tau_j|\gamma_t, \sigma_t^2, \mu_1^t, \ldots, \mu_m^t, \tau_1^t, \ldots, \tau_m^t, y_1, ..., y_n) = p(t_j|\gamma_t)\ \forall j = 1, \ldots, m-1.$
3) $\gamma_{t+1} \sim$ non-informative prior.
4) $\sigma_{t+1}^2 \sim$ non-informative prior.

Thus, we only need to devise a method for the posterior sampling of the μ_j's, and select the appropriate non-informative priors for γ and σ^2. We need to point out that the selection of the non-informative prior for the random variables γ and σ^2 can be extremely difficult. For example, it is known that some of the non-informative priors, like the Jeffrey prior $p(\gamma) = \frac{1}{\gamma}$, are not proper probabilities. Thus, this part of our model is still an open problem.

VI. Sampling from a Posterior Distribution of μ_j

Given $(Y_1, ..., Y_n)$ a sample from multivariate normal variable with distribution $N(H\boldsymbol{\mu}, \sigma^2 I)$, where $\boldsymbol{\mu}^T = (\mu_1, \mu_2, ..., \mu_m)^T$ represent the mapping of the lattice $2^X - \emptyset$ into a vector, and H represent the sorting of the values $\left[f(x_{(k)}) - f(x_{(k+1)})\right]$. Thus, we can see from the model in section 3 for the Gibbs sampler that

$$\mu_j \sim p(y_1, ..., y_n|H\boldsymbol{\mu}, \sigma^2 I) p(\mu_j|\tau_j, \{\mu_s|A_j \subseteq A_s\}) \quad (7)$$

Consider for a moment that $p(\mu_j|\tau_j, \{\mu_{A_s}|A_j \subseteq A_s\})$ is not truncated that is

$$p(\mu_j|\tau_j, \{\mu_{A_s}|A_j \subseteq A_s\}) = p(\mu_j|\tau_j). \quad (8)$$

Then, $p(y_1, ..., y_n|H\boldsymbol{\mu}, \sigma^2 I) p(\mu_j|\tau_j)$ is a univariate Gaussian distribution with mean θ and variance σ^2. Thus, we only need to devise a closed form expression for this Gaussian and truncate it. Now, we write the complete expression of $p(\mu_j|\theta, \sigma^2)$ as

$$p(\mu_j|\theta, \sigma^2) = p(y_1, ..., y_n|H\boldsymbol{\mu}, \sigma^2 I) p(\mu_j|\tau_j) \propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - H_i\boldsymbol{\mu})^2\right\} * \exp\left\{-\frac{1}{2\tau_j}\mu_j^2\right\}, \quad (9)$$

where H_i is the i row in the design matrix $H = (H_1, \ldots, H_n)^T$ and μ_j is the j position in the vector $\boldsymbol{\mu}$. Thus, we can then rearrange the terms in the exponential part of $p(\mu_j|\theta, \sigma^2)$ to be

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}((y_i - H_i^{-j}\boldsymbol{\mu}_{-j}) - H_{ij}\mu_j)^2 - \frac{1}{2\tau_j}\mu_j^2\right\} \quad (10)$$

where $H_i^{-j} = (H_i^1, ..., H_i^{j-1}, H_i^{j+1}, ..., H_i^m)$ and $\boldsymbol{\mu}_{-j} = (\mu_1, ..., \mu_{j-1}, \mu_{j+1}, ..., \mu_m)^T$. Equation (10) can be rearranged to be

$$\exp\left\{-\frac{\sum_{i=1}^{n}((H_{ij})^2\mu_j^2)}{2\sigma^2} + \frac{\sum_{i=1}^{n}2(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij})\mu_j}{2\sigma^2} - \frac{\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})^2}{2\sigma^2} - \frac{\mu_j^2}{2\tau_j}\right\} \quad (11)$$

$$(12)$$

The exponential term can be rewritten as

$$-\frac{\sum_{i=1}^{n}(\tau_j(H_{ij})^2\mu_j^2)}{2\sigma^2\tau_j} + \frac{2\sum_{i=1}^{n}\tau_j\left[\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right]\mu_j}{2\sigma^2\tau_j} - \frac{\sum_{i=1}^{n}\tau_j(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})^2 + \sigma^2\mu_j^2}{2\sigma^2\tau_j}. \quad (13)$$

Thus, we can have simplify this to be equal to

$$-\frac{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right) + \sigma^2)\mu_j^2}{2\sigma^2\tau_j} + \frac{2\tau_j\left[\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right]\mu_j}{2\sigma^2\tau_j} - \frac{\sum_{i=1}^{n}\tau_j(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})^2}{2\sigma^2\tau_j}. \quad (14)$$

We have finally that

$$-\frac{\mu_j 2 - \frac{2\tau_j\left[\sum_{i=1}^{n}((y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right]}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right)+\sigma^2)}\mu_j + \frac{\sum_{i=1}^{n}\tau_j(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})^2}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right)+\sigma^2)}}{\frac{2\sigma^2\tau_j}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right)+\sigma^2)}}. \quad (15)$$

We have the following after completing squares

$$p(\mu_j|\theta, \sigma^2) \propto \exp\left\{-\frac{\left[\mu_j - \frac{\tau_j\left[\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right]}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right)+\sigma^2)}\right]^2}{\frac{2\sigma^2\tau_j}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right)+\sigma^2)}}.\right\} \quad (16)$$

Then, we have that

$$\mu_j \sim N\left(\frac{\tau_j\left[\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right]}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right) + \sigma^2)}, \frac{\sigma^2\tau_j}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right) + \sigma^2)}\right) \quad \forall j = 1, \ldots, m. \quad (17)$$

Using this distribution, we can then sample from the truncated normal in the interval $[0, min\{\mu_s|A_j \subseteq A_s\}]$. Finally, the

truncated distribution looks like

$$\mu_j \sim N\left(\tau_j \frac{\left[\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right]}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right) + \sigma^2)}, \frac{\sigma^2\tau_j}{(\tau_j\left(\sum_{i=1}^{n}(H_{ij})^2\right) + \sigma^2)}\right)|_{[0,min\{\mu_s|A_j\subseteq A_s\}]}\forall j = 1,\ldots,m. \quad (18)$$

VII. Problems when sampling from a Gaussian with small variance.

An immediate problem that this method has is that variance for the normal sampling distribution tends to be small. This produces inaccuracies in the sampling of the measures. A way to solve this problem is based in the following observation. Imagine that we are sampling from a Gaussian distributions, it is clear that we have a higher probability to sample from the places near to mean than away from it. If the standard deviation decrease, the probability to sample things near to the mean is higher. Now, if we are trying to sample from an interval that is far away from the mean with respect to the standard deviation, the curve of the Gaussian function tends to be flat. Thus, we can assume that we are sampling from a uniform distribution in the interval of interest. Although, this looks like an inefficient solution, our experiments indicate that this simple approach works. We need to remark that this is not a theoretically correct solution. The correct solution is based in the distribution $\mu_j \sim K * \exp\left[-\frac{(\mu-\theta)2}{2\sigma^2}\right]$, where *K* is the inverse of the integral of the exponential term in an interval. For this reason, we are looking for more efficient ways to circumvent this problem.

VIII. Improving the Sparsity Promoting Model

Something that is clear from sections (IV) - (V) is that we are trying to promote Sparsity using a Gaussian distributions. This is clear if we look at the first equation in the Gibbs sampler,

$$\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - H_i\boldsymbol{\mu})^2\right\} * \exp\left\{-\frac{1}{2\tau_j}\mu_j^2\right\}. \quad (19)$$

It is clear that the first exponential represents the minimization term for the classification and the second exponential represents the sparsity promoting term. In the moment we understand this, we realize that given $\mu_j \in [0,1]$, we shall we using a better sparsity promoting distribution. When looking at Figueiredo [13], we realized that we should be using directly the Laplacian, or in our case the exponential distribution, to promote more sparsity in our model. Thus, the new model looks like

$$\begin{aligned} y_i &\sim N(H_i\boldsymbol{\mu}, \sigma^2)\ \forall i = 1,\ldots,n \\ \mu_m &= 1 \\ \mu_j &\sim \exp\left(\frac{\gamma}{2}\right)|_{[0,min\{\mu_{A_s}|A_j\subseteq A_s\}]}\ \forall j = 1,\ldots,m-1 \\ \gamma &\sim \text{non-informative prior} \\ \sigma^2 &\sim \text{non-informative prior} \end{aligned}$$

Thus, the new model eliminate the intermidiate τ and uses directly a exponential distribution for the measures to be learned. Now, the Gibbs sampler looks like

1) $\mu_j^{t+1} \sim p(\mu_j|\gamma_t, \sigma_t^2, \boldsymbol{\mu}_{-j}^t, y_1, ..., y_n) = \frac{p(\mu_j, y_1,\ldots,y_n|\gamma_t,\sigma_t^2,\boldsymbol{\mu}_{-j}^t)}{p(y_1,\ldots,y_n)} \propto p(y_1, ..., y_n|\sigma_t^2, \boldsymbol{\mu}^t)p(\mu_j|\gamma_t, \{\mu_s|A_j \subseteq A_s\}) \quad \forall j = 1,\ldots,m$ with $\mu_j \sim \exp(\frac{\gamma}{2})|_{[0,min\{\mu_s|A_j\subseteq A_s\}]}$
2) $\gamma_{t+1} \sim$ non-informative prior.
3) $\sigma_{t+1}^2 \sim$ non-informative prior.

In this new model, the μ_j will be sampled from the distribution

$$\mu_j \sim N\left(\frac{\left(\sum_{i=1}^{n}(y_i - H_i^{-j}\boldsymbol{\mu}_{-j})H_{ij}\right) + \sigma^2\gamma}{\sum_{i=1}^{n}(H_{ij})^2}, \frac{\sigma^2}{\sum_{i=1}^{n}(H_{ij})^2}\right)|_{[0,min\{\mu_s|A_j\subseteq A_s\}]}\forall j = 1,\ldots,m. \quad (20)$$

Other possible improvement for this new model could be assuming not a single γ as a rate of sparsity for each measure but multiple γ's for each measure to be learned. In addition we would like to be able to have a feedback from the sparsity ratio to this γ's to increase the level of sparsity if it is necessary. In order to to this, we need to include an extra distribution for the γ's which cannot be a non-informative prior.

IX. Interpreting the results from sparsity promotion

A tool used to measure the importance of each input in the Choquet integration is the **Shapley index**. The Shapley index [22] for input x_i with respect to the measure μ is given by

$$\phi_{x_i}(\mu) = \sum_{x_i\notin S\subseteq X} \frac{|S|!(|X| - |S| - 1)!}{|X|!}(\mu(S\bigcup x_i) - \mu(S)). \quad (21)$$

A comparison of the definition of the Shapley index with the definition of the Choquet integral shows how the Shapley index measures the importance of each element in the set *X* by essentially computing the average value of the coefficients involving a particular element. More specifically, any weight in the Choquet integral involving a particular element x_i will have the form $\mu(S\bigcup x_i) - \mu(S)$ where $x_i \notin S$. If all these values are zero, then x_i has no influence on the output at all. If all these values are large, then x_i has a significant influence on the output.

X. Comparison with quadratic programming

A classical way to learn a complete measure is the optimization method proposed by Grabisch et. al [23]. In this method a quadratic objective function under constraints is defined to obtain the optimal measures. The number of constraints is a exponential function of the dimensionality of the input and they are stored in a matrix. Therefore, solving the quadratic

programming quickly becomes intractable. Although the Gibbs sampler has a similar limitation than the quadratic program, the exponential nature of the input, the Gibbs sampler does not depend on a sparse matrix of constraints, but on a clever way of imposing the constraints which decreases the complexity in calculating the solution.

It has been seen empirically, using the same hardware for both methods, that the Gibbs sampler is faster than the quadratic programming, but a complete complexity analysis will be performed.

XI. Experiments

In this section, we present results on two synthetic problems to illustrate the sparsity promotion capability of the system. Each of them is a two class synthetic problem with 1000 samples in each class with four features in each sample. In addition, each Gibbs sampler chain has a length of 10000 iterations.

A. Case I

In this case the only separable features are in the odd positions which contain samples for Class 1 from the distribution $N(0.2, 0.1)$, and Class 2 from the distribution $N(0.8, 0.1)$. The second features and the fourth feature contain uniform noise. For features bigger than one and less than zero, we clipped the values. If we plot the samples in their three first features (figure 2), we can see the separation between the classes.

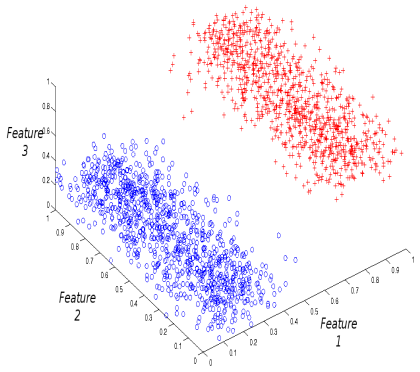

Fig. 2. Plot of samples for class 1 'o' and class 2 '+' for the first three features in Case I. Note that feature 2 has not value for classification.

It can be seen that the algorithm is able to separate the classes because of the confusion matrix in table (I). The fuzzy measures are in the table(II) and the Shapley values are in table (III). Note that the Shapley indices of the informative features (features 1 and 3) are approximately 20 times larger than those of the non-informative features (features 2 and 4).

CM	class 1	class
class1	1000	0
class2	0	1000

TABLE I
Confusion Matrix for Case I

Measures	Mean	std
$\mu(x_1)$	0.4506	0.033273
$\mu(x_2)$	0.01005	0.0078438
$\mu(x_3)$	0.46807	0.038589
$\mu(x_4)$	0.0074683	0.0068834
$\mu(x_3, x_4)$	0.52648	0.04968
$\mu(x_2, x_4)$	0.035829	0.016663
$\mu(x_2, x_3)$	0.50671	0.055892
$\mu(x_1, x_4)$	0.49988	0.042394
$\mu(x_1, x_3)$	0.96511	0.020748
$\mu(x_1, x_2)$	0.49782	0.04617
$\mu(x_1, x_2, x_3)$	0.99138	0.011033
$\mu(x_1, x_2, x_4)$	0.53374	0.050566
$\mu(x_1, x_3, x_4)$	0.98527	0.014186
$\mu(x_2, x_3, x_4)$	0.5743	0.055924
$\mu(X)$	1	0

TABLE II
Measures Case I with mean and std of the Markov chains

B. Case II

In this case the the only separable feature is the first one which contain samples for Class 1 from $N(0.2, 0.1)$, and Class 2 from $N(0.8, 0.1)$. The rest of features contain uniform noise. The same clipping strategy is used. The confusion matrix is shown in table (IV). The measures are in the table (V) and the Shapley values ae in the table (VI).

XII. Conclusion

A novel method for learning fuzzy measures for Choquet integration has been presented. The method uses a maximum likelihood approach to learn mappings from inputs to outputs coupled with a sparsity promoting term that reduces the influence of uninformative features. In our future work, we are planning to explore how the sparsity in the fuzzy measures affects the Shapley index. In this regard, we would like to

Feature	Shapley Value
1	0.46229
2	0.024708
3	0.48755
4	0.025449

TABLE III
Shapley values for the features in Case I

CM	class 1	class
class1	998	2
class2	0	1000

TABLE IV
CONFUSION MATRIX FOR CASE II

Measures	Mean	std
$\mu(x_1)$	0.4846	0.04257
$\mu(x_2)$	0.019192	0.010254
$\mu(x_3)$	0.022071	0.010317
$\mu(x_4)$	0.025016	0.011738
$\mu(x_3, x_4)$	0.049664	0.016633
$\mu(x_2, x_4)$	0.051202	0.021779
$\mu(x_2, x_3)$	0.03418	0.016434
$\mu(x_1, x_4)$	0.60286	0.069511
$\mu(x_1, x_3)$	0.98603	0.035203
$\mu(x_1, x_2)$	0.60599	0.064725
$\mu(x_1, x_2, x_3)$	0.99137	0.013091
$\mu(x_1, x_2, x_4)$	0.6642	0.086878
$\mu(x_1, x_3, x_4)$	0.99648	0.020278
$\mu(x_2, x_3, x_4)$	0.1134	0.014942
$\mu(X)$	1	0

TABLE V
MEASURES CASE II WITH THE MEAN AND STD OF THE MARKOV CHAINS

Feature	Shapley Value
1	0.72993
2	0.029855
3	0.20466
4	0.035554

TABLE VI
SHAPLEY VALUES FOR THE FEATURES IN CASE I

prove a direct relation between the Sparsity and the variance of the Shapley index. In addition we are developing a minimum error classification error using the logistic distribution and the Gibbs sampler

ACKNOWLEDGMENT

Research was sponsored by the U. S. Army Research Office and U. S. Army Research Laboratory and was accomplished under Cooperative Agreement Number DAAD19-02-2-0012. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, Army Research Laboratory, or the U. S. Government. The U. S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] M. Sugeno, "Theory of fuzzy integrals and its applications," Doctoral Thesis, Tokyo Institute of Technology, Tokyo, Japan, 1974.

[2] H. Tahani and J. Keller, "Information fusion in computer vision using the fuzzy integral," *IEEE Transactions on Systems, Man and Cybernetic*, vol. 20, no. 3, pp. 733–741, 1990, reprinted as an appendix to G. Klir and Z. Wang, fuzzy Measure Theory, Plenum Press, 1992.

[3] K. Xu, Z. Wang, P.-A. Heng, and K.-S. Leung, "Classification by nonlinear integral projections," *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 2, pp. 187–2001, 2003.

[4] A. Hocaoglu, P. Gader, and J.-H. Chiang, "Comments on Choquet fuzzy integral-based hierarchical network for decision analysis," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 6, pp. 48–53, December 1999.

[5] M. Grabisch and M. Sugeno, "Multi-attribute classification using fuzzy integral," in *IEEE international conference on Fuzzy Systems*, March 1992, pp. 47–54.

[6] M. Grabisch, T. Murofushi, and M. Sugeno, *Fuzzy Measures and Integrals. Theory and Applications*, ser. Studies in Fuzziness and Soft Computing. Physica Verlag, Heidelberg, 2000.

[7] J.-L. Marichal, "Aggregation of interacting criteria by means of the discrete Choquet integral," in *Aggregation Operators: New Trends and Applications*, ser. Studies in Fuzziness and Soft Computing, T. Calvo, G. Mayor, and R. Mesiar, Eds. Physica Verlag, Heidelberg, 2002, vol. 97, pp. 224–244.

[8] M. Grabisch, "A new algorithm for identifying fuzzy measures and its application to pattern recognition," in *Fourth IEEE International Conference on Fuzzy Systems*, Yokohama, Japan, March 1995, pp. 145–150.

[9] ——, "Fuzzy integral for classification and feature extraction," in *Fuzzy Measures and Integrals. Theory and Applications*, M. Grabisch, T. Murofushi, and M. Sugeno, Eds. Physica Verlag, 2000, pp. 348–374.

[10] ——, "Modelling data by the Choquet integral," in *Information Fusion in Data Mining*, V. Torra, Ed. Physica Verlag, Heidelberg, 2003, pp. 135–148.

[11] P. D. Gader, R. Grandhi, W.-H. Lee, and J. N. Wilson, "Integration of ordered weighted averaging operators with feed-forward neural networks for feature subset selection and pattern classification," submitted to IEEE Trans. Fuzzy Systems.

[12] P. D. Gader, B. Nelson, A. Hocaoglu, S. Auephanwiriyakul, and M. Khabou, "Neural versus heuristic development of choquet fuzzy integral fusion algorithms for land mine detection," in *Neuro-fuzzy Pattern Recognition*, H. Bunke and A. Kandel, Eds. World Scientific Publ. Co., 2000, pp. 205–226.

[13] M. A. T. Figueiredo, "Adaptive sparseness for supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1150–1159, 2003.

[14] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 6, no. 6, pp. 721–741, Nov. 1984.

[15] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[16] B. Walsh, "Markov chain monte carlo and gibbs sampling," 2004. [Online]. Available: http://nitro.biosci.arizona.edu/courses/EEB581-2004/handouts/Gibbs.pdf

[17] P. Gader, L. Wen-Hsiung, and A. Mendez-Vazquez, "Continuous Choquet integrals with respect to random sets with applications to landmine detection," in *IEEE International Conference on Fuzzy Systems*, IEEE. IEEE, July 2004, pp. 523–528 vol 1.

[18] Z. Wang and G. J. Klir, *Fuzzy Measure Theory*. Norwell, MA, USA: Kluwer Academic Publishers, 1993.

[19] G. Choquet, "Theory of capacities," Annales de l'Institut Fourier, 1955, vol 5, pp 131-295.

[20] M. Grabisch, H. Nguyen, and E. Walker, *Fundamentals of Uncertainty Calculi, with Applications to Fuzzy Inference*. Kluwer Academic Publishers, Dordrecht, 1995.

[21] J.-H. Chiang, "Choquet fuzzy integral-based hierarchical network for decision analysis," *IEEE Transactions on Fuzzy Systems*, vol. 7, no. 1, pp. 63–71, February 1999.

[22] L. S. Shapley., "A value for n-person games," in *Contributions to the Theory of Games Volume II*, ser. Annals of Mathematical Studies, H. Kuhn and A. Tucker, Eds. Princeton University Press, 1953, vol. 28, pp. 307–317.

[23] M. Grabisch and J. Nicolas, "Classification by fuzzy integral: Performance and tests," *Fuzzy Sets and Systems*, vol. 65, no. 2-3, pp. 255–271, 1994.