
Article:
Obajemu, O. and Mahfouf, M. orcid.org/0000-0002-7349-5396 (2019) A Dirichlet Process based type-1 and type-2 fuzzy modeling for systematic confidence bands prediction. IEEE Transactions on Fuzzy Systems, 27 (9). pp. 1853-1865. ISSN 1063-6706

https://doi.org/10.1109/TFUZZ.2019.2892347

A Dirichlet Process based Type-1 and Type-2 Fuzzy Modelling for Systematic Confidence Bands Prediction

Olusayo Obajemu and Mahdi Mahfouf

Abstract

This paper presents a new methodology for fuzzy logic systems modelling based on the Dirichlet Process Gaussian Mixture Models (DPGMM). The proposed method simultaneously allows for the systematic elicitation of confidence bands as well as the automatic determination of model complexity. The work is new since existing fuzzy model elicitation techniques use *ad-hoc* methods for confidence band estimations which do not meet the stringent requirements of today's challenging environments where data is sparse, incomplete, and characterised by noise as well as uncertainties. The proposed approach involves an integration of fuzzy and Bayesian topologies and allows for the generation of confidence bands based on both the random and linguistic uncertainties embedded in the data. Additionally, the proposed method provides a 'right-first time approach' to fuzzy modelling as it does not require an iterative model complexity determination. In order to see how the proposed framework performs across a variety of challenging data modelling problems, the proposed approach was tested on a non-linear synthetic dataset as well as two real multi-dimensional datasets generated by the authors from materials science and bladder cancer studies. Results show that the proposed approach consistently provides better generalisation performances than other well-known soft computing modelling frameworks – in some cases, improvements of up to 20% in modelling accuracy were achieved. The proposed method also provides the capability to handle uncertainties via the generation of systematic confidence intervals for informing on model reliability. These results are significant since the generic methodologies developed in the paper should help material scientists as well as clinicians, for example, assess the risks involved in making informed decisions based on model predictions.

Index Terms

fuzzy, uncertainty, non-parametric.

I. INTRODUCTION

Fuzzy models, in contrast to 'black-box' modelling algorithms, can result in models which are more interpretable and easy to maintain because they employ fuzzy sets (*FSs*) and human-like rules. The properties of such fuzzy models consequently allow for easy elicitation of data models from experts. However, for very complex and uncertain systems, such knowledge elicitation from experts may prove 'tricky' or even impossible because of high dimensionality and/or a lack of readily available expert knowledge. When process data are available, a plethora of fuzzy rule generation algorithms have been proposed to allow for automatic fuzzy model extraction from the data [1], [2], [3], [4]. The aforementioned algorithms generally involve some optimisation regime to actively determine the singular 'best' fuzzy model parameters which explain the data whilst completely discarding the other less favourable parameters. These algorithms as well as the modelling assumptions do not provide opportunities or frameworks for even accessing and utilising the other less optimal parameters. These discarded parameters can however provide one with valuable information such as the elicitation of confidence bands in predictions. Though, it is possible to deploy uncertain systems without comprehensively characterising the uncertainties embedded in such systems [5], the ability to quantify the uncertainty in a system remains a valuable tool. For example, knowing the inputs around which a model is uncertain can aid in the design of future experiments [6]. Additionally, the provision of confidence bands for model predictions should help users of the model assess the risks involved when making informed decisions on manufacturing process designs or clinical therapies, for instance, based on how reliable those predictions may be. Only in limited studies have authors attempted to provide algorithms for confidence bands generation, albeit in *ad-hoc* manners. The studies in [7] where the model prediction errors are further analysed to provide confidence bands is a prime example of such an *ad-hoc* approach. Bayesian formalisms can capture the notion of uncertainty in engineering systems more systematically [6]. The authors of this paper are of the belief that an integration of fuzzy topologies and Bayesian statistics can provide a more effective and a systematic approach for determining such confidence bands.

Fuzzy models which allow for automatic rule determination [8] do not require exhaustive, iterative and expensive cluster/rule validations [9], [10], [7] and have become the de facto method for fuzzy systems modelling. Therefore, the motivation for this paper stems from the need to provide a fuzzy modelling framework which not only allows for automatic rule determination

from data (i.e. number of rules and the corresponding membership functions), but also helps to generate systematic prediction confidence bands.

The approach taken in this paper involves viewing the parameters of the conventional fuzzy system as random variables which can be shown later to exploit the advantages of both probabilistic and fuzzy reasoning. The parameters are learnt in a Bayesian manner based on the Dirichlet Process Gaussian Mixture Model (*DPGMM*) [11]. As will later be discussed in the paper, this approach allows for concomitantly determining the number of fuzzy rules as well as providing confidence bands as a result of the intrinsic modelling. The antecedent and consequent parameters are obtained directly from the clusters following a similar procedure discussed in [1].

Attempts at exploiting the strengths of both probabilistic and fuzzy reasoning in a combined manner is not new. In [12], [13], [14], the authors studied the relationship between randomness and fuzziness. In [15], the authors have used what they called probabilistic fuzzy sets (*PFS*), where it is considered that the fuzzy membership functions (*MFs*) are random variables (*RVs*) with a certain probability density function (*PDF*). In [16], the authors used the fuzzy modelling approach to approximate the *PDF* of a random variable. However, how the aforementioned approaches can be used to solve a practical engineering problem remains an open question since many of these studies only analysed the relationship between fuzzy and probabilistic theories without delving into the practical engineering challenges.

Hence, the main contributions of this paper relate to the development of a unifying algorithm which is capable of: 1. systematically determining the complexity of the fuzzy model (via the number of clusters); 2. eliciting a fuzzy model that intrinsically includes systematic confidence intervals for the predictions. This unifying algorithm should be viewed as a right-first-time modelling framework for handling linguistic and random uncertainties.

This paper is organised as follows: Section II introduces the theoretical foundations of fuzzy systems and the Dirichlet Process Gaussian mixture models. Section III provides a visual proof of the limitations of existing fuzzy systems in handling uncertainties and discusses the proposed framework which overcomes these limitations through the use of the Dirichlet process Gaussian mixture modelling approach. Section IV presents the results of testing the proposed framework on synthetic and real datasets with Section V concluding the paper.

II. Fuzzy Systems and Bayesian Reasoning

A. Fuzzy Systems

Perhaps the most important advantage in using fuzzy logic systems (*FLSs*) in modelling is their interpretability. This contrasts with other techniques whose parameters lose meaning especially in very high dimensions. The fuzzy modelling process can be easy to follow as well as being amenable to expert knowledge incorporation because it uses *if-then* fuzzy rules and also the variables are linguistic. To allow for the handling of linguistic uncertainties, type-2 fuzzy sets (*T2 FSs*) theories were formalised by Mendel and Karnik in [9] and [17] whose studies further expanded the already wide applicability of fuzzy systems especially in modelling [18], [19], [20], control [21] and robotics [22].

Despite promising results of the T2 framework as compared to the *T1 FLS* (see [9] for studies showing *T2* methods outperform *T1* ones), its ability to handle random uncertainties remains an open but exciting research topic. By handling, it is meant here being able to quantify the uncertainty embedded in the modelling process in a systematic and unified manner. In this work, it is assumed that the parameters of a *FLS* are random after which a mechanism for eliciting such a *FLS* based on the *DPGMM* is provided. This is a Bayesian scheme which has historically been the *de facto* formalism for handling and quantifying the uncertainties inherent in the data modelling process. Additionally, the proposed approach provides a means for a right-first-time elicitation of such *FLSs* as a consequence of using the *DPGMM* in which complexity (number of rules) is determined by the data. A *T1 FS*, $\boldsymbol{A}$, may be expressed as follows:

$$\boldsymbol{A} = \{x, \mu_{\boldsymbol{A}}(x) | x \in X\} \tag{1}$$

where $\mu_{\boldsymbol{A}}(x)$ is the membership degree of the *T1 FS* of an element x in the universe of discourse X, $0 \leq \mu_{\boldsymbol{A}}(x) \leq 1$.

The *FLS* is a mapping from the input space $\mathcal{X}$ to the output space $\mathcal{Y}$ which can be formulated as follows:

$$\hat{y} = \sum_{j}^{c} \phi_j(\mathbf{x}) \zeta_j \tag{2}$$

where $\hat{y}$ is the output of the *FLS*, $\phi_j(\mathbf{x})$ represents the validity function for the jth rule of c rules for an input $\mathbf{x} \in \mathbb{R}^n$. The nature of ζ_j determines if the *FLS* is of the Mamdani type (in which case ζ_j represents the output/consequent *MF* of the jth rule) or the Takagi-Sugeno-Kang

(*TSK*) type (where $\zeta_j = \mathbf{a}_j^\intercal \mathbf{x} + b_j$). The proposed approach in this paper extends well to both the TSK and Mamdani fuzzy types as will be seen in the succeeding sections[1]. In the *T1 FLS*, $\phi_j(\mathbf{x})$ is simply the normalised product of the *MF* values (firing strength) in each input space. It is worth noting that (2) is also valid for the case of IT2 FLS. However, the validity function $\phi_j(\mathbf{x})$ becomes an interval and the final output is calculated via a type-reduction algorithm such as the Karnik-Mendel algorithms [9].

As already stated, in contrast to *T1 FSs*, *T2 FSs* can help to handle inherent linguistic uncertainties in systems modelling [23]. However, their ability to incorporate randomness in parameter estimation and consequently provide a prediction confidence based on the degree of uncertainty in the modelling process is limited. The aim of this paper is to show that a *FLS* whose parameters are assumed to be random variables can help to systematically handle both random and linguistic uncertainties. The mechanism by which such fuzzy models can be elicited in a systematic manner through the use of the *DPGMM* is then proposed. It is worth noting that, throughout the remainder of the paper, the subscript j relates to a particular cluster/rule (for a total of K or c clusters/rules) while the subscript i relates to a particular data point (for a total of N data points). Inputs are denoted as $\mathbf{x}$ and outputs as y. A combined input/output data point is denoted as $\mathbf{z}$.

B. The Dirichlet Process Gaussian Mixture Model

The Dirichlet Process ($\mathcal{DP}$) is a distribution over distributions [24] which has two parameters α and G_0 (respectively referred to as concentration parameter and the base distribution) such that

$$G \sim \mathcal{DP}(\alpha G_0) \tag{3}$$

where G is a draw from the $\mathcal{DP}$ and the symbol $\sim$ means 'drawn from'. A peculiar property of the $\mathcal{DP}$ is that samples from G are discrete with probability 1 with each placing an infinite collection of points on the G_0 space. The conditional probability of the nth sample from G (θ_n) given the previous $n-1$ samples ($\theta_{1:n-1}$) is given as follows:

$$\theta_n|\theta_{1:(n-1)} \sim \frac{1}{\alpha+n-1}\sum_{i=1}^{n-1}\delta_{\theta_i} + \frac{\alpha}{\alpha+n-1}G_0 \tag{4}$$

[1]From a modelling perspective, the Mamdani fuzzy system is allowed to take a constant in the consequent part. This is similar to the TSK type but with the added restriction that the $\mathbf{a}_j$ parameters are zero [9].

Equation (4) is called the Polya Urn Scheme [25] which shows that draws from G conditioned on previous observations have a clustering property. This clustering property is rather useful for data clustering and will prove useful in the fuzzy model elicitation process. θ_i typically parameterises the distribution of the data points so that the process of data generation is given by the following equations:

$$\begin{aligned} G &\sim \mathcal{DP}(\alpha G_0) \\ \theta_i &\sim G \\ \mathbf{z}_i|\theta_i &\sim h(.|\theta_i) \end{aligned} \tag{5}$$

where $h(.|\theta_i)$ represents the distribution from which data point $\mathbf{z}_i$ is drawn. If one considers the posterior for θ given the data points $\mathbf{z}_{1:N}$, the clustering property would result in observations sharing parameters and belonging to the same cluster. Crucially, however, a new data point can belong to a completely new cluster drawn from G_0 so that the number of clusters is no longer fixed beforehand (as a consequent of 4).

Generally, given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)|i = 1, \cdots, N\}$, $\mathbf{x}_i \in \mathbb{R}^D$ is the input and $y \in \mathbb{R}$ is the output, the probabilistic goal of learning is to find a function $f \ni f(\mathbf{x}_i) \approx y_i \ \forall \ i$. The nature of the function f typifies the learning algorithm. In a FLS, f is determined as shown in (2). In statistical learning, the generative assumption is made whereby for a given input, the observed output is a corrupted version of an unknown deterministic function f given by the following equation:

$$y_i = f(\mathbf{x}_i, \mathbf{w}) + \epsilon \tag{6}$$

where $\mathbf{w}$ represents the parameters of f and ϵ is a random noise variable.

In the Bayesian paradigm, the parameter vector ($\mathbf{w}$) is assumed to be random and then given a prior distribution ($P(\mathbf{w})$). This prior expresses the initial belief/distribution on what the values of these parameters are. The prior is consequently updated in light of observations $\mathcal{D}$ to the posterior $p(\mathbf{w}|\mathcal{D})$ using Bayes' rule as defined by the following equation:

$$P(\mathbf{w}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{\int P(\mathcal{D}|\mathbf{w})P(\mathbf{w})d\mathbf{w}} \tag{7}$$

To elicit predictions for a new input $\mathbf{x}^*$, the output $\mathbf{y}^*$ is determined as follows:

$$P(y^*|\mathcal{D}) = \int P(y^*|\mathbf{w})P(\mathbf{w}|\mathcal{D})d\mathbf{w} \tag{8}$$

$P(y^*|\mathcal{D})$ is called the Posterior Predictive Distribution (*PPD*) and is the essence of the Bayesian formalism which allows for the a systematic determination of confidence bands.

Though the integrals in (7) and (8) are generally intractable, there exist efficient approximate inference techniques such as Variational Inference [27] and Monte Carlo sampling [28].

It is worth noting that the equation defined in (8) is a function of the output y and when one is interested in a representative prediction which best explains this distribution, several types of loss functions $(L(f_{\mathbf{w}}, \mathbf{w}))$ [29] [30] can be defined using the following equation [31]:

$$J = \int L[f_{\mathbf{w}}, \mathbf{w}]p(\mathbf{w}|y)d\mathbf{w} \tag{9}$$

It can easily be shown that the squared loss function corresponds to the posterior mean which is similar to the centroid defuzzification procedure of FLSs.

III. Methodology

This section describes the proposed method for fuzzy model elicitation based on the *DPGMM* and how the proposed method may be more apt at handling uncertainties inherent in a *FLS*. It is worth discussing the limitations of this approach using the following two hypothetical rules:

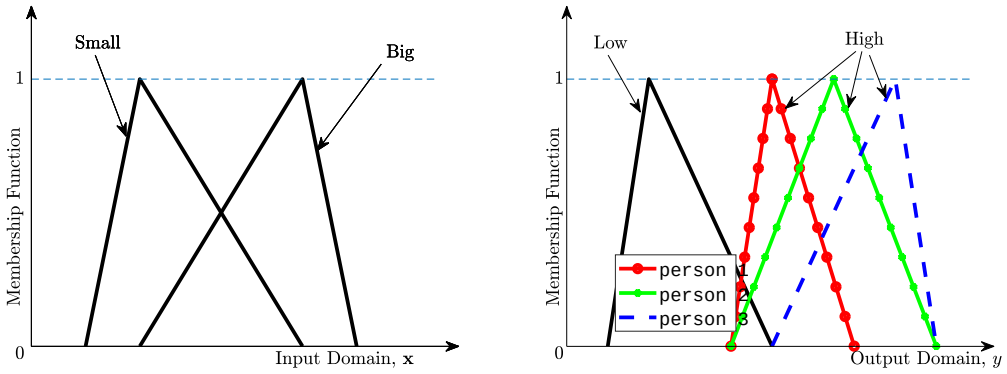
Rule 1 : if the house is *small* then the price is *low*

Rule 2 : if the house is *big* then the price is *high*

Defining the MFs of the rules (for small, big, low and high) can sometimes introduce uncertainties. For example, suppose one assumes that three persons are polled to provide these MFs. It is easily conceivable that different types of MFs will result. To make the analysis simpler, assume that the shape and size of the MFs for small, big and low of the three persons coincide but differ in the MF definition for high as shown in Fig. 1. This type of uncertainty which is called linguistic uncertainty can be handled by defining a T2 FS for the MFs which can be obtained by merging the resulting T1 FS as shown in Fig. 2 [9]. This process transforms the T1 FLS into a T2 FLS.

It should be recalled that after rule aggregation, the resulting output may be found according to the extension principle [32]:

$$C_{\tilde{\boldsymbol{A}}} = \int_{u \in J_{y_1}} \cdots \int_{u \in J_{y_M}} [\eta_{y_1}(u_1) \star \cdots \star \eta_{y_M}(u_M)] \Big/ \frac{\sum_{k=1}^{M} y_k u_k}{\sum_{k=1}^{M} u_k} \tag{10}$$



(a) (b)

Fig. 1: Membership functions of the linguistic values in the input and output domains. There is an uncertainty about the *MF* of the linguistic value *HIGH*.

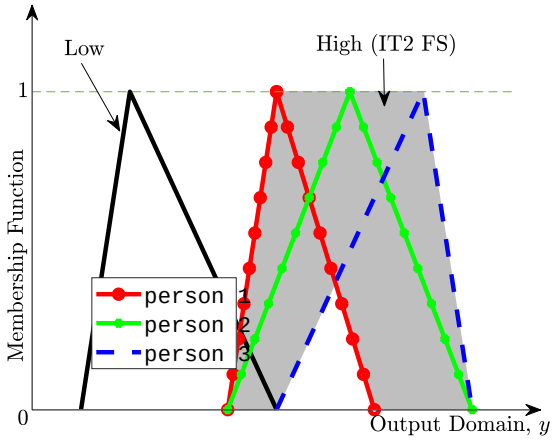
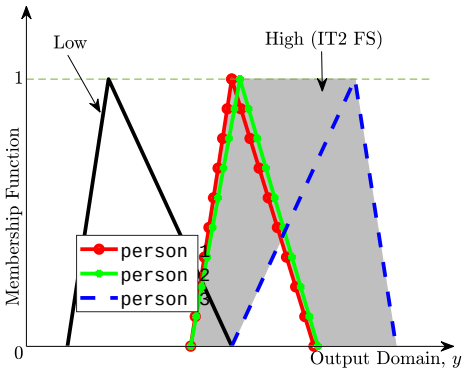


Fig. 2: It is possbible to merge the *MFs* of *HIGH* to make it either a *IT2 FS* or a *GT2 FS*. The *FLS* is now a *T2 FLS*.

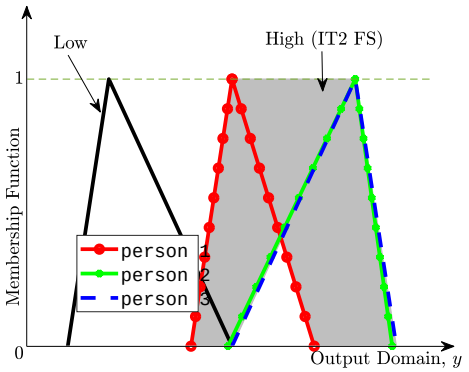
where M is the output domain discretisation level, $\eta_{y_k}(u_k)$ is the secondary grade of u_k at y_k with the t-norm operator defined as $\star$. Equation (10) states that to defuzzify a T2 FS, all the embedded T2 FSs are enumerated and each of these embedded T2 FSs is then defuzzified. The degree of truth of the defuzzified values of the embedded T2 FSs is found by utilising the t-norm operator on the corresponding secondary grades [33].

For an IT2 FS, this process can be simplified by utilising the Karnik-Mendel algorithms which results in finding the left (L) and right (R) end points of a T1 interval. This interval can sometimes be interpreted as corresponding to the linguistic uncertainty embedded in the FLS.

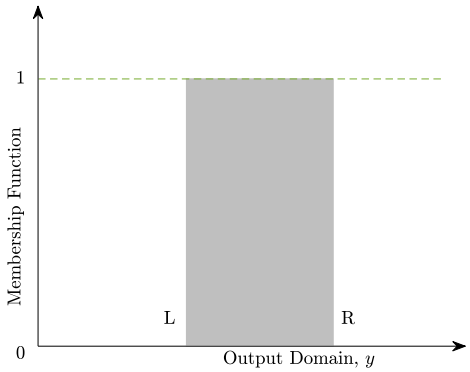
It can be seen in Figs. 3a and 3b that when the values of the UMF and LMF remain unchanged,



(a) Persons 1 and 2 have similar *MF* for the linguistic value *HIGH* (yellow and blue).



(b) Persons 2 and 3 have similar *MF* for the linguistic value *HIGH* (blue and red).

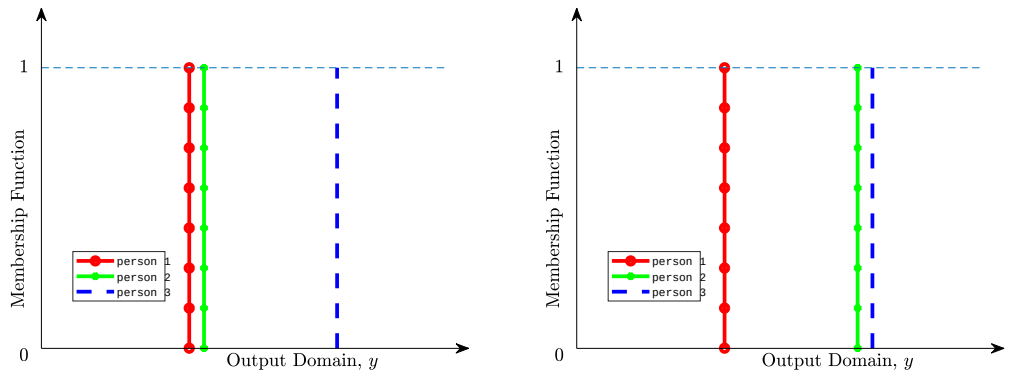


(c) Resulting *T1* interval after type reduction of a *IT2 FS*.

Fig. 3: There is no change in the *LMF* and *UMF* of (a) and (b). Consequently, there is no change in the *L* and *R* switch points (c). Uncertainty not well accounted for.

the L and R switch points remain constant (as shown in Fig. 3c). As a result, the final defuzzified values also remain unchanged even when person 2's MF changes. This means that though person 2 MF is closer to person 1 than person 3, this phenomenon is not accounted for in the final type-reduced set. Ideally, a good measure of the uncertainty in the output should reflect these changes in MF locations. In this paper, it is proposed to leave the MFs as they are without merging which results in the individual defuzzified values as shown in Figs. 4a and 4b.

Interestingly, this can be taken as equivalent to the Bayesian approach and similar to sampling from the posterior weights and formulating the *PPD* (8) for each set of posterior weights. In essence, this approach is akin to having some sort of an hierarchical *FLS* with each corresponding to a particular *FLS* of the different persons polled. This paper presents such a framework for understanding the uncertainties based on assuming that the parameters of a *FLS* are random.



(a) Defuzzified values without merging for (a). (b) Defuzzified values without merging for (b).

Fig. 4: By not merging and leaving the *MFs* as they are, we are able to clearly distinguish the uncertainty in each of the defuzzified values. Consequently, uncertainty is well accounted for.

Clustering is performed on the product space ($\mathcal{X} \times \mathcal{Y}$) based on the *DPGMM* which means that the prior is a Dirichlet Process (*DP*). The antecedent and consequent parameters are directly obtainable from these multidimensional clusters following a similar procedure as described in [1].

A. *Dirichlet Process Gaussian Mixture Modelling-based Fuzzy Modelling*

The finite Gaussian mixture model (*GMM*) (with K components) [34] is given by the following equation:

$$P(\mathbf{z}_i|\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_K) = \sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{z}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{11}$$

where $\boldsymbol{\theta}_j = \{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ defines the suite of parameters for the jth component of the mixture with $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ respectively defined as the mean and covariance matrix of the Normal distribution. $\mathbf{z}_i \in \mathbb{R}^{D+1}$ is a vector representing the ith data point for $i = 1, 2, \cdots, N$ (N being the number of data points in total). If one defines a joint prior G_0 for the component parameters, the generative process of the *GMM* can be described as follows:

$$\begin{aligned} \mathbf{z}_i|c_i, \boldsymbol{\theta} &\sim \mathcal{N}(\boldsymbol{\mu}_{c_i}, \boldsymbol{\Sigma}_{c_i}) \\ c_i|\pi &\sim \text{Discrete}(\pi_1, \cdots, \pi_K) \\ (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &\sim G_0 \\ \boldsymbol{\pi}|\alpha &\sim \text{Dir}(\alpha/K, \cdots, \alpha/K) \end{aligned} \tag{12}$$

where Dir is the Dirichlet distribution of order $K \geq 2$. The *DPGMM* is derived from (12) by allowing the number of mixture components go to infinity (i.e. $K \to \infty$) and integrating out the mixing distribution ($\boldsymbol{\pi}$) so that the component distribution are now drawn from a $G \sim DP(\alpha, G_0)$ (from (4)).

Combining (4) with the likelihood of the data gives rise to the conditional posterior distribution for the parameters of the ith data point as defined by the following equation:

$$\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}|\mathbf{z}_i, \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{-i} \propto \mathcal{L}(\mathbf{z}_i|\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}) \bullet \cdots \left(\frac{1}{\alpha + N - 1} \sum_{j=1, i \neq j}^{N} \delta_{\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}} + \frac{\alpha}{\alpha + N - 1} G_0\right) \quad (13)$$

where $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{-i}$ represents the parameters for all data points except the parameters of the current data point which is to be sampled. The symbol $\bullet$ means product. In (13), it can easily be seen that the expression $(\mathcal{L}(\mathbf{z}_i|\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}) \bullet G_0)$ is the joint distribution (product of the likelihood and the prior) based on observation $\mathbf{z}_i$ alone (similar to the Bayesian equation defined in (7). Thus, this can be calculated by multiplying the marginal likelihood $(\int\int(\mathcal{L}(\mathbf{z}_i|\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \bullet G_0) d\boldsymbol{\mu}_i d\boldsymbol{\Sigma}_i$ by the posterior $(P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_j|\mathbf{z}_i))$. To facilitate the evaluation of this integral, a conjugate prior to the component parameters is defined such that the mean vector $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\Sigma}_i$ are jointly distributed according to the following equation:

$$G_0 \sim P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \text{GIW}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{m}_0, k_0, \rho_0, \mathbf{S}_0) \quad (14)$$

where $\mathbf{m}_0, k_0, \rho_0$ and $\mathbf{S}_0$ are the parameters of the Gaussian-Inverse-Wishart (GIW) distribution. In this research, inference is performed using the Gibbs sampling (a Markov Chain Monte Carlo algorithm) such that the conditional posterior probabilities (13) of the *DPGMM* are updated in turn until convergence. The Gibbs sampling algorithm is similar to that proposed in [11]. The Gibbs sampling algorithm returns the samples from $p(\{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\}, \cdots, \{\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N\}|\mathcal{D})$ with data points sharing parameters belonging to the same cluster. The number of clusters is the number of unique component parameter values in a particular sample (after convergence).

B. Derivation of the Fuzzy Model Parameters

The *DPGMM* defined in Section III-A is used to elicit the the fuzzy model from the product space of the input-output space ($\mathcal{X} \times \mathcal{Y}$) described as follows: let $\mathbf{z_i}$ represent the ith input-output

data point such that

$$\mathbf{z}_i = [\mathbf{x}_i^\intercal \ y_i]^\intercal \tag{15}$$

for $i = 1 \cdots N$, then the data matrix is defined by the following equation:

$$\mathbf{Z}^\intercal = \begin{bmatrix} \mathbf{z}_1^\intercal \\ \mathbf{z}_2^\intercal \\ \vdots \\ \mathbf{z}_N^\intercal \end{bmatrix} \tag{16}$$

A cluster analysis is performed as described in Section III-A so that Gibbs sampling is used to sample from the posterior parameters of the cluster components.

The antecedent parameters for each rule can be obtained from the direct projections of the multi-dimensional clusters to the respective input spaces as discussed in [1] and consequent parameters are derived as given by the following equation:

$$\begin{aligned} \mathbf{a}_j &= {\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{\boldsymbol{j}}}^{-1} \boldsymbol{\Sigma}_{xy}^{j} \\ b_j &= \mu_j^y - \mathbf{a}_j^\intercal \boldsymbol{\mu}_j^x \end{aligned} \tag{17}$$

where $\mathbf{a}_j$ and b_j are the consequent parameters of the jth rule, $\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{\boldsymbol{j}}$ is the marginal covariance matrix of the inputs after integrating out the output, v_j^y is the output mean of the jth cluster and $\boldsymbol{\mu}_j^x$ is the jth cluster mean of the inputs (excluding outputs). It is worth noting that the consequent parameters (defined in (17)) are determined for the Gibbs samples (after convergence) and the output is calculated using (2). In essence, if after convergence, a total of P posterior parameters are sampled, then the antecedent, the consequent parameters as well as the corresponding outputs are derived P times as discussed. The distribution of this output reveals the underlying uncertainty in the predictions from which confidence bands can be derived.

The algorithm for generating the systematic prediction confidence bands for a T1 FLS is given as Algorithm 1.

The computational complexity of this algorithm is $O(NDK^2)$ but the Gibbs sampling algorithm can easily be parallelised as discussed in [35]. It is worth noting that the number of clusters (K) is automatically determined using the training dataset only. The parameters of the fuzzy model are then obtained from these clusters (as described in this section) from which the

Algorithm 1 Algorithm for systematic confidence band estimation based on the *DPGMM*-based fuzzy model.
Inputs: Data vector $\mathbf{z}_i = [\mathbf{x}_i^\intercal \ y_i]^\intercal \in R^{D+1}$ for $i = 1, 2, \cdots, N$, hyperparameters for the priors G_0 $(\mathbf{m}_0, k_0, \rho_0, \mathbf{S}_0)$ and $\boldsymbol{\pi}$ (α). $\mathbf{m}_0 = \mathbf{0}, k_0 = 1, \rho_0 = D + 1, \mathbf{S}_0 = \mathbf{I}_D$ and $\alpha = 1$.
Outputs: Cluster assignments and P posterior samples of $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ for $j = 1, 2, \cdots, K$. K is the number of clusters automatically determined.

The Markov chain state is $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. As a result of the clustering property, some $\boldsymbol{\theta}$ values will belong to the same class.

for $l = 1, 2, \ldots, P$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    Sample $\boldsymbol{\theta}_i$ according to (13).
  end for
  for $j = 1, 2, \ldots, K$ **do**
    Calculate $\mathbf{a}_j$ and b_j according to (17).
  end for
  Calculate $y_i^{(l)}$ $\forall i$ according to (2).
end for

predictions can be made.

Up until now, the proposed algorithm can only elicit a T1 FLS. To facilitate the elicitation of an IT2 FLS, Algorithm 1 is updated by choosing hyperparameters which represent an interval instead of a crisp value. Consequently, the mean hyperparameters ($\mathbf{m}_0$) were considered to be intervals as they directly map into the corresponding fuzzy sets. Therefore, the algorithm for eliciting an ITS FLS is given as Algorithm 2.

It is worth noting that at every sampling iteration, two sets of parameters (determined by the bounds of the selected hyperparameters) are obtained which represent the parameters of the IT2 FLS.

C. Extension of the Algorithms to a Mamdani Fuzzy Model

Algorithms 1 and 2 can easily be extended to derive a fuzzy model of the Mamdani type as described below:

1) Mamdani Type-1: In this study, it is assumed that the consequent MFs are of the Gaussian type with fixed widths. This means that only the centre of each MF is to be determined from the cluster parameters. Consequently, the Mamdani fuzzy model can be taken to be a simplification

Algorithm 2 Algorithm for systematic confidence band estimation based on the *DPGMM*-based interval type-2 fuzzy model.
Inputs: Data vector $\mathbf{z}_i = [\mathbf{x}_i^\intercal \ y_i]^\intercal \in R^{D+1}$ for $i = 1, 2, \cdots, N$, hyperparameters for the priors G_0 $(\mathbf{m}_0, k_0, \rho_0, \mathbf{S}_0)$ and $\boldsymbol{\pi}$ (α). $\mathbf{m}_0 \in [\underline{\mathbf{m}_0} \ \overline{\mathbf{m}}_0], k_0 \in [0.1 \ 2], \rho_0 = D + 1, \mathbf{S}_0 = \mathbf{I}_D$ and $\alpha = 1$.
Outputs: Cluster assignments and P posterior samples of $\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$ for $j = 1, 2, \cdots, K$. K is the number of clusters automatically determined.

The Markov chain state is $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$. As a result of the clustering property, some $\boldsymbol{\theta}$ values will belong to the same class.

for $l = 1, 2, \ldots, P$ **do**
  for $i = 1, 2, \ldots, N$ **do**
    Sample $\boldsymbol{\theta}_i$ according to (13) using the hyperparameters bound. Hence two sets of parameters.
  end for
  for $j = 1, 2, \ldots, K$ **do**
    Calculate $\mathbf{a}_j$ and b_j according to (17).
  end for
  Calculate $y_i^{(l)}$ $\forall i$ according to (2).
end for

of the TSK models where the $\mathbf{a}_j = \mathbf{0}$. One is then only interested in calculating the b_j parameters which can be found as follows:

$$\begin{aligned} \mathbf{a}_j &= \mathbf{0} \\ b_j &= \mu_j^y \end{aligned} \tag{18}$$

2) Mamdani Type-2: In the case of the Mamdani Type-2 fuzzy model, Algorithm 2 is still used for sampling the parameters. As a result of the fact that the hyperparameters are now intervals instead of a fixed single value, after each sampling iteration, one obtains two sets of parameters for each cluster ($\boldsymbol{\theta}_j^*$ and $\boldsymbol{\theta}_j^{**}$). This corresponds to having two sets of parameters which represent intervals. The parameters can then be calculated using the same procedure as in (18).

It is worth re-emphasizing that each component/cluster represents one fuzzy rule. The antecedent MFs are multidimensional Gaussian but which can be projected into each subspace in a manner akin to that described in [1].

IV. Results

In this section, the proposed fuzzy modelling approach is tested on three (3) datasets. The first dataset is a synthetic data derived from a non-linear equation as explained in Section IV-A.

The second dataset is a real dataset from TATA Steel United Kingdom (*U.K.*) from a which a fuzzy model is to be elicited in order to predict the mechanical properties of alloy-steels. The third dataset is also a real dataset which contains disease and patient characteristics of humans diagnosed with bladder cancer. The results of the proposed *DPGMM* (TSK1 - T1 TSK model, TSK2 - T2 TSK model, Mam1 - T1 Mamdani model and Mam2 - T2 Mamdani model) approaches are compared with four algorithms obtained from the literature i.e. Artificial Neural Networks (*ANN*), type-1 fuzzy model (*T1FM*) [19], Mamdani Interval Type-2 Fuzzy Model (*MIT2FM*) [36], Linear Regression (*LR*) and Interval Type-2 Fuzzy Modelling (*IT2FM*) [18]. The performance measure used is the root-mean-square error (*RMSE*). The hyperparameters of the proposed approaches were determined using the 10-fold cross validation technique on the training datasets only.

A. Synthetic Dataset

This section presents the results of applying the proposed approach in modelling a synthetic dataset generated from the non-linear function (shown in Fig. 5a) described by the following equation:

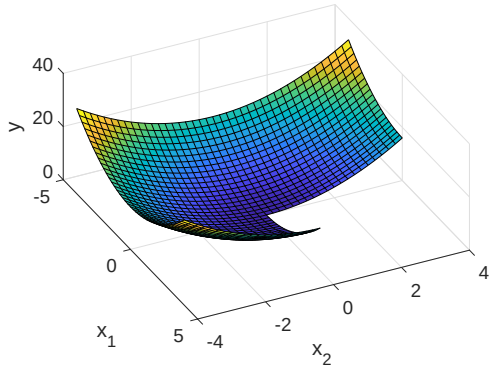
$$y = x_1^2 + x_2^2 \tag{19}$$

There are two inputs (x_1 and x_2) and one output (y). The training dataset is as shown in Fig. 5b. The testing data points (labelled 1-10 in Fig. 5b) were randomly generated from the uniform distribution $\mathcal{U}$ as follows:

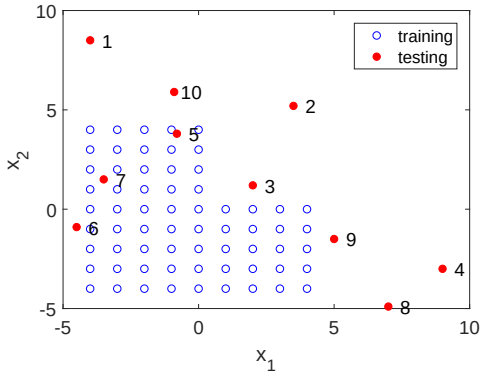
$$(x_1, x_2) \in [-6, 10] \times [-6, 10] \tag{20}$$

It is evident from Fig. 5b, that some testing data points fall on spaces where the training dataset is dense while others fall in the sparse spaces. It is expected that the testing dataset confidence bands should reflect this density differential. Additionally, since the training dataset is involved in eliciting the model, the confidence bands on the training data points are expected to be high.

One thousand (1000) posterior FLS parameters are sampled according to the Gibbs sampling procedure as explained in Section III-A for both the Mamdani and TSK approaches. Fig. 6 shows the confidence bands of the testing dataset which have been obtained by finding the difference between the minimum and maximum predictions as shown in Fig. 7. The bands on



(a) Plot of the function $y = x_1^2 + x_2^2$ with some sections cut out.



(b) Plot of the scatter of the function $y = x_1^2 + x_2^2$ with labels.

Fig. 5: Synthetic dataset.

the training dataset are shown in Fig. 8 which, as expected, show a high confidence in training output predictions.

Table I shows the modelling results of the proposed approach as compared with different modelling methods. These results show that the proposed framework compares favourably with these approaches with the added advantage of being able to provide confidence bands in predictions. In particular, it can be observed that the proposed framework produces better performance than the other algorithms by consistently providing better generalisation capabilities. The Mamdani Type-2 DPGMM approach provides the best result on the testing dataset with an RMSE of 8.72. The Mamdani DPGMM approaches provide the best generalisation performances. This may be because of the fact that in some cases (e.g. for data points 1 and 2), the models are attempting to extrapolate beyond the data domain. Since the Mamdani approaches are restrictive versions of the TSK approaches, they tend to provide a better generalisation performance than the TSK ones.

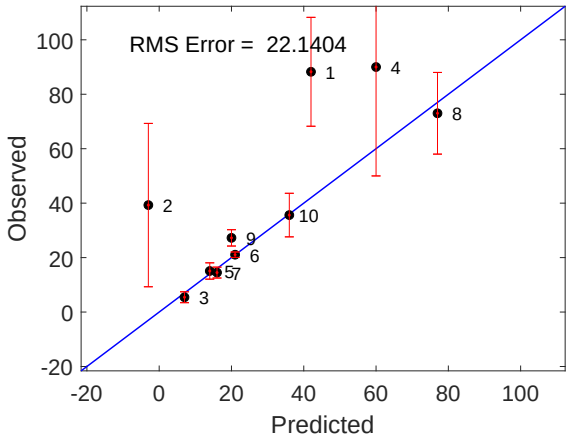


Fig. 6: Prediction results and confidence on the synthetic dataset. Data points in the sparse area produce less reliable predictions as can be seen from the high confidence bands.

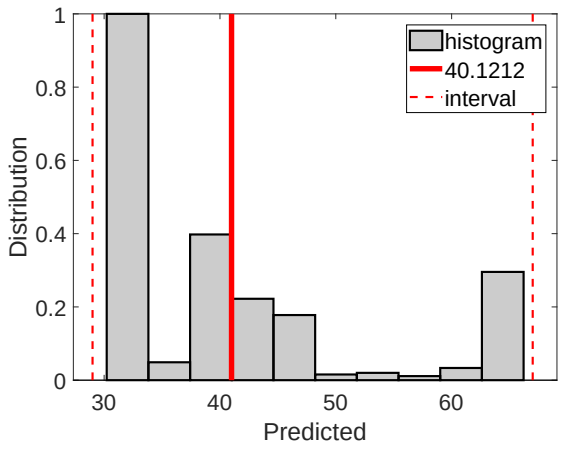


Fig. 7: Histogram of samples from selected testing data points using the Gibbs sampling procedures in Algorithms 1 and 2. The prediction interval is determined from the difference between the maximum and minimum of the samples. A single prediction is made by finding the mean of the samples.

The ANN model provides the worst generalisation performance which indicates that although it is good at interpolation, it does not generalise well to data points which have not been seen during model training.

B. Ultimate Tensile Strength (UTS) of Steel Model

The proposed approach was also tested on a real-life industrial problem which relates to the prediction of the Ultimate Tensile Strength (UTS) of alloy-steels prior to production. This dataset

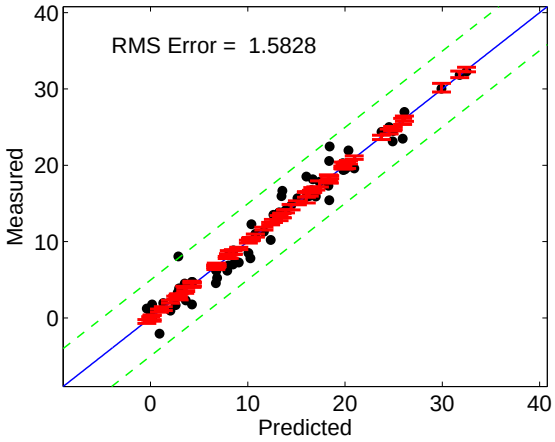


Fig. 8: Training dataset performances on the synthetic dataset. The prediction confidence is high (confidence bands small) for all data points in the training dataset.

TABLE I: Synthetic dataset prediction results. All numbers represent the average RSME for a total of 10 runs. The numbers in the brackets are the standard deviations .

Method	**Training**	**Testing**
DPGMM (TSK1)	1.58 ($\pm$0.04)	22.14 ($\pm$1.25)
DPGMM (TSK2)	1.52 ($\pm$0.05)	27.73 ($\pm$1.73)
DPGMM (Mam1)	1.68 ($\pm$0.041)	9.01 ($\pm$2.09)
(8 Rules)	1.65 ($\pm$0.03)	11.73 ($\pm$2.22)
(12 Rules)	1.61 ($\pm$0.03)	10.24 ($\pm$2.04)
DPGMM (Mam2)	1.74 ($\pm$0.027)	8.72 ($\pm$1.94)
(8 Rules)	1.60 ($\pm$0.04)	10.09 ($\pm$1.29)
(12 Rules)	1.58 ($\pm$0.034)	10.63 ($\pm$1.21)
T1FM	1.77 ($\pm$0.063)	30.87 ($\pm$4.21)
MIT2FM	1.69 ($\pm$0.066)	32.07 ($\pm$3.23)
(8 Rules)	1.63 ($\pm$0.071)	33.31 ($\pm$4.22)
(12 Rules)	1.62 ($\pm$0.064)	36.76 ($\pm$4.39)
ANN	1.24 ($\pm$0.018)	49.02 ($\pm$20.32)
LR	8.40 ($\pm$0.013)	29.64 ($\pm$1.14)
IT2FM	1.59 ($\pm$0.07)	29.40 ($\pm$4.56)

was obtained from TATA Steel, U.K. and contains two sets of data.

The first set included 3760 data points with 15 inputs representing chemistry, geometry of the specimens and various heat treatment temperatures. The second set was obtained at a later date and includes a further twelve (12) data points with the same set of inputs. These two sets of data are independent time-wise but result from the same manufacture setting as well as the same production cycle. A decision was made not to combine these datasets in order to validate the proposed approach on a completely remote dataset. Fig. 9 shows the density distribution of a selected input variable (% of Silicon) against the output (which is the UTS).

Fig. 10 includes the equivalent top-view plot of the distribution of the data which highlights locations of four (4) selected data points taken from the independent dataset. It can be seen that the data distribution shows varying degrees of density across the input space. The training procedure consists of dividing the first set of data into two (70% for training and 30% for testing) with the testing dataset used to test the generalization performance of the elicited model. The hyperparameters were determined via the 10-fold cross validation technique on the training dataset only.

A visual inspection of the second independent dataset revealed some uncertainties. For example, though data points 4 and 5 contain exactly the same values of inputs, they exhibit different output values. This is due to the random uncertainty embedded in the data which may have resulted from insufficient number of input variables or measurement precision. This uncertainty may even have resulted from errors at the computerised data entry stage. Hence, the proposed framework can help one overcome this by providing a degree of confidence in the predictions.

The performance of the model on the training, testing and independent datasets are shown in Figs. 11-12 and Table II. It is worth emphasizing that the uniqueness of the model lies in its ability to correctly determine the confidence bands as shown in Fig. 12. For example, as compared to data points 3 and 4, data point 12 includes more uncertainty (lower confidence). As can be seen, data points 12, 2 and 1, which are on the edges of the distribution (Figs. 9 and 10), tend to have lower confidence bands than data points closer to the middle of the plot. This behaviour is reasonable because when one observes the data distribution of Figs. 12 and 13, it can be seen that the dataset is denser in the middle area. Hence, the model may not be reliable in areas where the data distribution is sparse.

In Table II, it can be observed that the DPGMM-fuzzy algorithm for eliciting an IT2 FLS provides comparable or better performance when compared with the other selected modelling

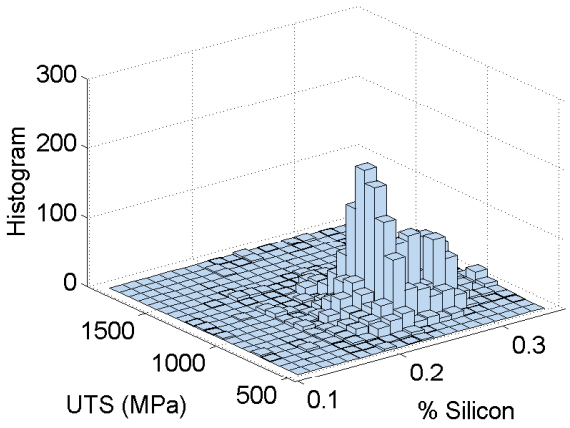


Fig. 9: Silicon (%) against UTS (MPa) distribution.

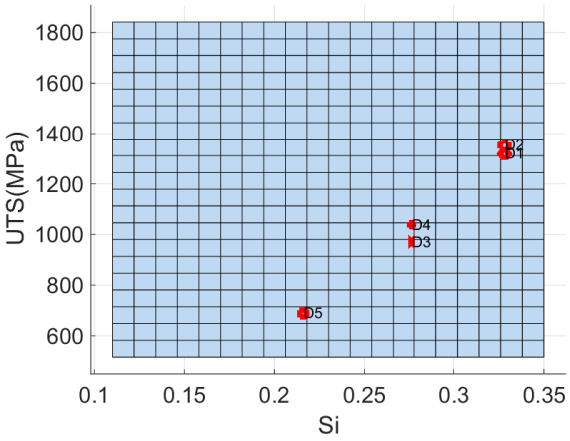
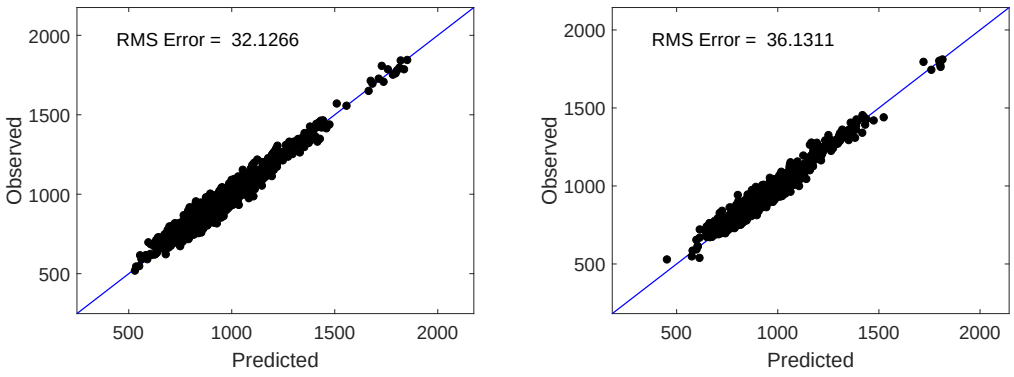


Fig. 10: Top view of the distribution of amount of Silicon (%) against UTS (MPa).

frameworks. The DPGMM-fuzzy framework for eliciting a T1 FLS also provides a better performance than an ANN as well as its type-1 counterpart. In the case of the Mamdani systems, it can be seen that as the number of fuzzy rules increased (the number of fuzzy rules can be controlled through the α parameter), the proposed approach tended to give a consistent performance. However, the non-Bayesian Mamdani approach sees a considerably increase in the generalisation error (44.50 for 12 rules). This is to be expected for, as the model complexity increases, the Bayesian approaches provide implicit regularisation which allows for maintaining excellent generalisation capabilities. On average, the proposed approaches produce consistently better performances by more than 10%.

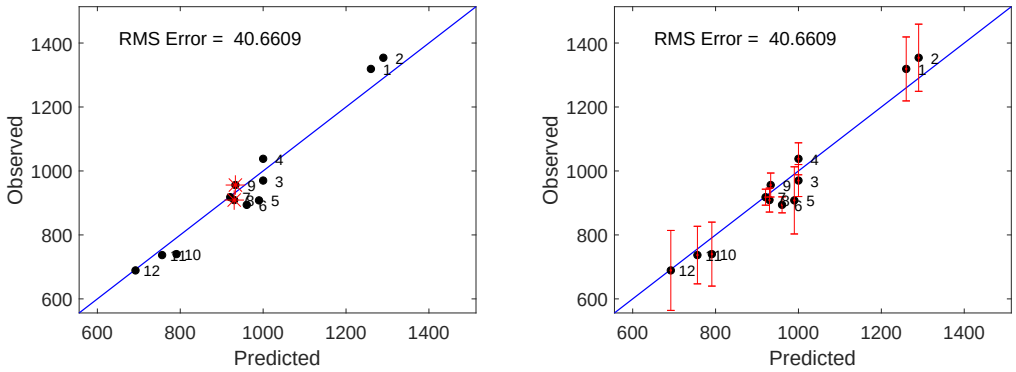


(a) Training Data Results. (b) Testing Data Results.

Fig. 11: Results on the prediction of *UTS* of steel (DPGMM-TSK1).

TABLE II: *UTS* of steel prediction results. All numbers represent the average RSME for a total of 10 runs while the numbers in bracket specifies the corresponding standard deviation.

Method	**Training**	**Testing**	**Validation**
DPGMM (TSK1)	32.12 (1.0)	36.13 (1.3)	40.66 (2.1)
DPGMM (TSK2)	33.16 (1.3)	34.34 (1.3)	35.24 (2.3)
DPGMM (Mam1)	36.26 (1.3)	38.90 (1.1)	39.88 (1.1)
(8 Rules)	35.14 (1.3)	37.16 (1.1)	37.83 (1.3)
(12 Rules)	35.08 (1.2)	38.23 (1.3)	36.09 (1.0)
DPGMM (Mam2)	35.73 (1.1)	37.06 (1.4)	39.15 (1.5)
(8 Rules)	35.04 (1.2)	38.43 (1.3)	36.22 (1.4)
(12 Rules)	35.31 (1.2)	37.66 (1.2)	35.77 (1.1)
T1FM	38.14 (1.1)	39.54 (1.3)	37.37 (1.0)
MIT2FM	36.33 (2.1)	40.52 (2.1)	34.77 (2.5)
(8 Rules)	35.39 (1.3)	38.43 (1.2)	43.48 (1.8)
(12 Rules)	35.18 (1.4)	39.34 (1.2)	44.50 (1.4)
ANN	45.78 (3.1)	46.02 (3.7)	46.23 (4.0)
LR	47.98 (0.8)	48.07 (0.9)	49.21 (1.0)
IT2FM	33.40 (1.9)	35.90 (2.1)	34.17 (2.3)



(a) UTS independent dataset results highlighting data points 8 and 9.

(b) Predictions with confidence bands.

Fig. 12: UTS independent dataset results.

C. Bladder Cancer Data Model

The last dataset on which the proposed modelling framework was tested is a real clinical dataset. The dataset was obtained from a study of patients diagnosed with bladder cancer at the Royal Hallamshire Hospital in Sheffield, U.K. The hospital collected this database over 16-year period between 1 January, 1994 and 31 December, 2009 [19]. A total of 613 patients with complete follow-up times were included in this study. The input variables contain patient characteristics (such as age, sex), disease properties (such as stage, grade, tumour pathology) as well as treatment type (cystectomy and radiotherapy). Medical datasets have historically been the most challenging of datasets to model because of noise and high dimensionality [37]. Existing fuzzy models mainly predict a singular value without providing a measure of reliability on these predictions. When predictions are found to be inaccurate, the model is deemed unreliable and is often discarded. The proposed approach is particularly useful in such a situation as the model is able to provide its own measure of reliability which can aid in the decision making process. This measure of reliability can also help in the design of future experiments such as providing the areas to concentrate on in the data gathering process. This section investigates how the proposed approach generalises to a complex bladder cancer data modelling problem. The aim is to predict the time of death of patients diagnosed with bladder cancer. The data was divided into two parts: 70% for training and 30% for testing. 10-fold cross validation was performed on the training dataset in selecting the hyperparameters.

As can be seen in Table III, there is an improvement of approximately 48% and 53% in

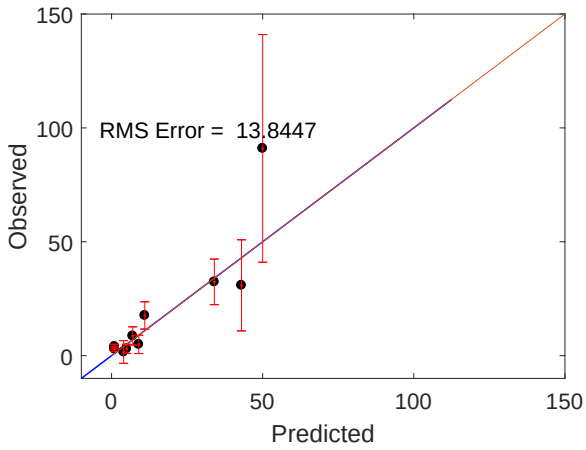


Fig. 13: Predictions and confidence bands of 10 randomly selected data points from the bladder cancer dataset. The model provides a confidence band for each model prediction which will inform clinicians on its reliability.

training and testing results when the proposed framework is compared with linear regression which buttresses the good modelling strength of the proposed framework. In particular, the DPGMM TSK2 approach provides the best generalisation capability with an RMSE of 12.11. As the number of rules increased, the proposed approach (Mam1 and Mam2), when compared with the non-Bayesian Mamdani model (MIT2FM), maintains its good generalisation performances. On average, the proposed framework outperforms the other algorithms with which results were compared by up to 20% decrease in the testing data RMSE.

Compared with the other selected non-linear modelling methods (Table III), the proposed framework is able to provide consistent improvements in modelling performance over all these methods by an average of 11% with the type-2 based framework providing the best results for an equivalent number of fuzzy rules. Fig. 13 shows the prediction results (DPGMM TSK1) as well as the confidence bands of 10 randomly selected data points from the bladder cancer dataset. It can be seen that the accuracy of the model is further improved as the observed time of death of the patients decreases. However, the model also provides a measure of reliability which indicates that when the patient time of death is high, the model may not be as reliable as for when the patient time of death is low.

In summary, one can see that the proposed approach consistently provides better generalisation performances than the other well known algorithms with which results were compared. The Mamdani DPGMM approaches provided the better performances on the synthetic dataset while

TABLE III: Comparative average RMSE results between the proposed framework and other well known algorithms for the bladder cancer dataset. The numbers in the brackets specify the standard deviation for a total of 10 runs.

Method	**Training**	**Testing**
DPGMM (TSK1)	12.04 (0.7)	12.16 (0.7)
DPGMM (TSK2)	11.73 (0.7)	12.11 (0.8)
DPGMM (Mam1)	15.64 (0.8)	16.09 (0.9)
(8 Rules)	14.04 (0.8)	18.26 (0.9)
(12 Rules)	13.17 (0.9)	17.47 (0.9)
DPGMM (Mam2)	14.02 (0.7)	16.71 (0.9)
(8 Rules)	13.19 (0.6)	16.44 (0.8)
(12 Rules)	12.96 (0.8)	16.32 (1.0)
T1FM	16.22 (1.0)	19.43 (1.1)
MIT2FM	14.04 (1.2)	18.26 (1.3)
(8 Rules)	13.52 (1.2)	19.49 (1.3)
(12 Rules)	13.30 (1.2)	19.90 (1.4)
ANN	11.19 (3.7)	26.77 (6.4)
LR	23.16 (0.5)	25.84 (0.6)
IT2FM	13.04 (0.8)	13.96 (0.9)

the TSK DPGMM approaches performed better on the two real datasets. This can be attributed to the fact the TSK approaches are better suited for highly non-linear and high dimensional modelling problems. The TSK approaches include more parameters which provide additional degrees of freedom. The Mamdani-based approaches, however, are better suited for low dimensional modelling problems where extrapolation beyond the data domain may be required.

V. Conclusion

This paper has presented a new approach for fuzzy systems elicitation from data. In contrast to existing methodologies, this new approach has the capability of providing systematic confidence intervals in predictions which should reflect uncertainty handling. Additionally, the proposed method should also be viewed as a right-first-time approach to fuzzy modelling since complexity/number of fuzzy rules can be determined by the data under investigation through the use, for the first time, in the context of fuzzy logic, of the Dirichlet Process Gaussian Mixture

model. The parameters of the fuzzy system are directly obtainable from the models using an earlier approach developed in the literature. When tested on real, noisy and highly non-linear data modelling problems, the results show that the proposed method leads to improved accuracy and, more importantly, to the systematic handling and quantification of uncertainties when compared to other modelling frameworks of similar configuration. The proposed method is generic and has been applied to successfully elicit type-1 as well as interval type-2 fuzzy logic systems for both the Mamdani and the TSK types.

APPENDIX

A. *Model Predictions after derivation of Antecedent and Consequent Parameters*

Given a ***FLS*** with D inputs ($\mathbf{x} \in \mathbb{R}^D$), $x_1 \in X_1, x_2 \in X_2, \cdots, x_D \in X_D$ and one output $y \in Y$, with a rule-base consisting of c rules, the jth rule of a typical IF-THEN fuzzy rule base may be expressed in the form:

$$R^j: \text{IF} \quad x_1 \text{ is } \mathbf{A}_1^j \text{ and } x_2 \text{ is } \mathbf{A}_2^j \ \cdots \text{ and } x_D \text{ is } \mathbf{A}_D^j, \\ \text{THEN } y^j \ is \ \mathbf{B}^j \quad (21)$$

After rule-aggregation, the defuzzified output is as given in (2). After clustering the product space as described in Algorithm 1, then one obtains samples of the parameters of the components. The jth component represents the jth rule so that ith output predicted by the elicited ***FLS*** whose antecedent and consequent parameters have been determined is given as follows:

$$\begin{aligned}
\hat{y}_i = f(\mathbf{x}_i) &= \mathbf{E}[y|\mathbf{x} = \mathbf{x}_i] = \int yP(y|\mathbf{x})dy \\
&= \int y\frac{P(\mathbf{x}, y)}{P(\mathbf{x})}dy = \int y\frac{P(\mathbf{z})}{P(\mathbf{x})}dy \\
&= \int y \sum_{j=1}^{c} \underbrace{\frac{P(\mathbf{z}_j|c_j)P(c_j)}{P(\mathbf{x})}}_{\text{from (5)}} dy \\
&= \sum_{j=1}^{c} \frac{P(\mathbf{x}|c_j)p(c_j)}{P(\mathbf{x})} \underbrace{\int yP(y|\mathbf{x}, c_j)dy}_{[\mathbf{x}^\intercal \ 1]\boldsymbol{\gamma}_j}
\end{aligned} \quad (22)$$

where the symbol $\mathbf{E}[\chi]$ represents the expectation of a random variable χ. The expression $[\mathbf{x}^\intercal\, 1]\boldsymbol{\gamma}_j$ is the output *MF* of the jth rule of a *TSK* model where $\boldsymbol{\gamma}_j$ is a vector containing the consequent parameters of the jth rule (defined as $\mathbf{a}_j$ and b_j) in (2). The equations in (22) and (2) are essentially the same with $\frac{P(\mathbf{x}|c_j)p(c_j)}{P(\mathbf{x})}$ representing the validity function ($\phi_j(\mathbf{x})$). It has been shown [1] that this validity function is equal to

$$\phi_j(\mathbf{x}) = \frac{\frac{P(c_j)}{2\pi^{D/2}\sqrt{|\boldsymbol{\Sigma}^{\boldsymbol{j}}_{\boldsymbol{xx}}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^\intercal(\boldsymbol{\Sigma}^{\boldsymbol{j}}_{\boldsymbol{xx}})^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right)}{\sum_i^c \frac{P(c_j)}{2\pi^{D/2}\sqrt{|\boldsymbol{\Sigma}^{\boldsymbol{j}}_{\boldsymbol{xx}}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^\intercal(\boldsymbol{\Sigma}^{\boldsymbol{j}}_{\boldsymbol{xx}})^{-1}(\mathbf{x}-\boldsymbol{\mu}_j)\right)} \tag{23}$$

B. Equations for Sampling the Fuzzy Model Parameters

The following series of equations detail the computational flow from deciding on the hyper-parameters of the prior parameters to performing the Gibbs sampling procedures in Algorithms 1 and 2 and finally to deriving the fuzzy logic parameters from the posterior samples.

The generative process is as shown in (12). There are two classes of prior hyper-parameters. One for the Dirichlet Process (which is the α) and the hyperparameters for the component parameters (which is the Gaussian Inverse Distribution described in (14)). The meaning and influence of the hyper-parameters for these two categories are listed as follows:

1) α is an hyper-parameter for the mixing component. This can be used to control the number of components/rules;
2) $\mathbf{m}_0$ is the mean of the mean of the clusters.
3) k_0 determines the scatter around the mean.
4) ρ_0 determines the scatter around the expectation around the covariance matrix.
5) $\mathbf{S}_0$ is the expectation of the covariance matrix.

Once the hyper-parameters have been determined (usually through cross-validation), Gibbs sampling is performed in turn as follows:

1) initialise the parameters ($\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}$) and the number of clusters (K) and assign each data point to one of the clusters. The number of clusters may change at each iteration since a data point is allowed to form a new cluster (when parameter sampling returns a new parameter set different from existing parameters).
2) For each data point ($\mathbf{z}_i$), sample the corresponding parameters ($\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$) according to the following:

- choose the parameters of the current data point to be equivalent to that of an existing cluster j proportional to the following value:

$$\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \propto \sum_{j=1, i\neq j}^{N} \delta(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)(\mathcal{L}(\mathbf{z}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{24}$$

or

- choose a new parameter with probability defined as proportional to:

$$\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\} \propto \alpha P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{z}_i) \bullet \cdots \int\int (\mathcal{L}(\mathbf{z}_i|\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\} \bullet G_0) d\boldsymbol{\mu}_j d\boldsymbol{\Sigma}_j \tag{25}$$

It should be recalled that the prior is a GIW distribution which can be defined as follows:

$$\begin{aligned} G_0 &\sim p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \\ &= \text{GIW}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{m}_0, k_0, \rho_0, \mathbf{S}_0) \\ &= \mathcal{N}\left(\boldsymbol{\mu}_i|\mathbf{m}_0, \frac{1}{k_0}\boldsymbol{\Sigma}_i\right) \mathcal{W}^{-1}\left(\boldsymbol{\Sigma}_i|\rho_0, \mathbf{S}_0\right) \\ &= \frac{1}{Q(D, k_0, \rho_0, \mathbf{S}_0)} |\boldsymbol{\Sigma}_i|^{-\frac{\rho_0+D+2}{2}} \cdots \\ &\exp(-\frac{\rho_0}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^\intercal \boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0) - \cdots \\ &\frac{1}{2}\text{Tr}(\boldsymbol{\Sigma}_i^{-1}\mathbf{S}_0)) \end{aligned} \tag{26}$$

where $\mathcal{N}(\boldsymbol{\mu}_i|\mathbf{m}_0, \frac{1}{k_0}\boldsymbol{\Sigma}_i)$ and $\mathcal{W}^{-1}(\boldsymbol{\Sigma}_i|\rho_0, \mathbf{S}_0)$ are the Normal distribution and Inverse Wishart distribution respectively. $Q(D, k_0, \rho_0, \mathbf{S}_0)$ is the normalising constant defined as follows:

$$Q(D, k_0, \rho_0, \mathbf{S}_0) = 2^{\frac{(\rho_0+1)D}{2}} \pi^{\frac{D(D+1)}{4}} |\mathbf{S}_0)|^{-\frac{\rho_0}{2}} \prod_{i=1}^{D} \Gamma\left(\frac{\rho_0 + 1 - i}{2}\right) \tag{27}$$

It can easily be shown that as a result of the conjugacy property, the posterior distribution ($P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{z}_i)$) given a data point $\mathbf{z}_i$ is also a GIW distribution which is given as follows:

$$P(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{z}_i) = \text{GIW}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i|\mathbf{m}_i, k_i, \rho_i, \mathbf{S}_i) \tag{28}$$

where each of the parameters is defined as follows:

$$\begin{aligned}\mathbf{m}_i &= \frac{k_0\mathbf{m}_0 + \mathbf{z}_i}{k_0+1}\\ k_i &= k_0 + 1\\ \rho_i &= \rho_0 + 1\\ \mathbf{S}_i &= \mathbf{S}_0 + \frac{k_0}{k_0+1}(\mathbf{z}_i - \mathbf{m}_0)(\mathbf{z}_i - \mathbf{m}_0)^\intercal\end{aligned} \tag{29}$$

which can be included into (25). Additionally the marginal distribution (the integral in (25)) can be calculated as follows:

$$\begin{aligned}\int\int (\mathcal{L}(\mathbf{z}_i|\{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\} \bullet G_0) d\boldsymbol{\mu}_j d\boldsymbol{\Sigma}_j = \\ \left(\frac{k_0}{\pi(1+k_0)}\right)^{\frac{D}{2}} \bullet \frac{\prod_{i=1}^{D}\Gamma\left(\frac{\rho_0+2-i}{2}\right)}{\prod_{i=1}^{D}\Gamma\left(\frac{\rho_0+1-i}{2}\right)} \bullet \cdots \\ \frac{|\mathbf{S}_0 + \frac{k_0}{1+k_0}(\mathbf{z}_i - \mathbf{m}_0)(\mathbf{z}_i - \mathbf{m}_0)^\intercal|^{-\frac{\rho_0+1}{2}}}{|\mathbf{S}_0|^{-\frac{\rho_0}{2}}}\end{aligned} \tag{30}$$

3) What one obtains at each iteration t is a parameter chain which corresponds to the parameters of each data point. The data points which share the same parameter values belong to the same cluster.
4) The last step involves obtaining the antecedent and consequent parameters for each unique clusters from the unique parameters.

REFERENCES

[1] J. Abonyi, R. Babuska, F. Szeifert, Modified gath-geva fuzzy clustering for identification of takagi-sugeno fuzzy models, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 32 (5) (2002) 612–621.

[2] T. A. Johansen, R. Babuska, Multiobjective identification of takagi-sugeno fuzzy models, IEEE Transactions on Fuzzy Systems 11 (6) (2003) 847–860.

[3] N. Kasabov, Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 31 (6) (2001) 902–918.

[4] J. R. Castro, O. Castillo, P. Melin, A. Rodríguez-Díaz, A hybrid learning algorithm for a class of interval type-2 fuzzy neural networks, Information Sciences 179 (13) (2009) 2175–2193.

[5] G. Lai, Z. Liu, Y. Zhang, C. P. Chen, Adaptive fuzzy tracking control of nonlinear systems with asymmetric actuator backlash based on a new smooth inverse, IEEE Transactions on Cybernetics 46 (6) (2016) 1250–1262.

[6] Y. Gal, Uncertainty in deep learning, Ph.D. thesis, PhD thesis, University of Cambridge (2016).

[7] G. Zhang, M. Mahfouf, M. Abdulkareem, S.-A. Gaffour, Y.-Y. Yang, O. Obajemu, J. Yates, S. A. Soberanis, C. Pinna, Hybrid-modelling of compact tension energy in high strength pipeline steel using a gaussian mixture model based error compensation, Applied Soft Computing 48 (2016) 1–12.

[8] S. Oh, W. Pedrycz, Identification of fuzzy systems by means of an auto-tuning algorithm and its application to nonlinear systems, Fuzzy sets and Systems 115 (2) (2000) 205–230.

[9] J. M. Mendel, Uncertain rule-based fuzzy logic systems: Introduction and new directions, Prentice-Hall, Upper-Saddle River, NJ.

[10] M. F. Zarandi, B. Rezaee, I. Turksen, E. Neshat, A type-2 fuzzy rule-based expert system model for stock price analysis, Expert Systems with Applications 36 (1) (2009) 139–154.

[11] R. M. Neal, Markov chain sampling methods for Dirichlet process mixture models, Journal of computational and graphical statistics 9 (2) (2000) 249–265.

[12] A. Meghdadi, M.-R. Akbarzadeh-T, Probabilistic fuzzy logic and probabilistic fuzzy systems, in: Fuzzy Systems, 2001. The 10th IEEE International Conference on, Vol. 3, 2001, pp. 1127–1130. doi:10.1109/FUZZ.2001.1008853.

[13] G. Gonzlez-Rodrguez, A. Colubi, W. Trutschnig, Simulation of fuzzy random variables, Information Sciences 179 (5) (2009) 642 – 653, special Section - Quantum Structures: Theory and Applications Workshop Quantum Structures. doi:http://dx.doi.org/10.1016/j.ins.2008.10.018.

[14] P. Liang, F. Song, What does a probabilistic interpretation of fuzzy set mean?, IEEE Transactions on Fuzzy Systems, 4 (2) (1996) 200–205. doi:10.1109/91.493913.

[15] G. Zhang, H.-X. Li, A probabilistic fuzzy learning system for pattern classification, in: Systems Man and Cybernetics (SMC), 2010 IEEE International Conference on, 2010, pp. 2336–2341.

[16] O. Osoba, S. Mitaim, B. Kosko, Bayesian inference with adaptive fuzzy priors and likelihoods, IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 41 (5) (2011) 1183–1197.

[17] J. M. M. Karnik N. N., Q. Liang, Type-2 fuzzy logic systems, IEEE Transactions on Fuzzy Systems 7 (1999) 643–658.

[18] O. Obajemu, M. Mahfouf, A new interval type-2 fuzzy clustering algorithm for interval type-2 fuzzy modelling with application to heat treatment of steel, 19th IFAC World Congress, 2014.

[19] O. Obajemu, M. Mahfouf, J. W. Catto, A new fuzzy modelling framework for integrated risk prognosis and therapy of bladder cancer patients, IEEE Transactions on Fuzzy Systems, 2017.

[20] A. Ramos-Soto, A. Bugarín, S. Barro, On the role of linguistic descriptions of data in the building of natural language generation systems, Fuzzy Sets and Systems 285 (2016) 31–51.

[21] H. Wang, Z. Wang, Y.-J. Liu, S. Tong, Fuzzy tracking adaptive control of discrete-time switched nonlinear systems, Fuzzy Sets and Systems 316 (2017 35–48.

[22] H. Hagras, A hierarchical type-2 fuzzy logic control architecture for autonomous mobile robots, IEEE Transactions on Fuzzy Systems, 12 (4) (2004) 524–539.

[23] J. M. Mendel, On KM algorithms for solving type-2 fuzzy set problems, IEEE Transactions Fuzzy Systems 21 (3) (2013) 426–446.

[24] J. Sethuraman, A constructive definition of dirichlet priors, Statistica sinica (1994) 639–650.

[25] D. Blackwell, J. B. MacQueen, Ferguson distributions via pólya urn schemes, The annals of statistics (1973) 353–355.

[26] C. E. Rasmussen, C. K. I. Williams, Gaussian processes for machine learning, MIT Press.

[27] D. M. Blei, M. I. Jordan, et al., Variational inference for Dirichlet process mixtures, Bayesian analysis 1 (1) (2006) 121–143.

[28] C. Andrieu, N. De Freitas, A. Doucet, M. I. Jordan, An introduction to MCMC for machine learning, Machine learning 50 (1-2) (2003) 5–43.

[29] A. Zellner, Bayesian econometrics, Econometrica: Journal of the Econometric Society (1985) 253–269.

[30] P. Congdon, Bayesian statistical modelling, Vol. 704, John Wiley & Sons, 2007.

[31] C. P. Robert, Bayesian computational methods, Springer, 2012.

[32] N. N. Karnik, J. M. Mendel, Centroid of a type-2 fuzzy set, Information Sciences 132 (1-4) (2001) 195–220.

[33] Q. Liang, J. M. Mendel, Interval type-2 fuzzy logic systems, IEEE Transactions on Fuzzy Systems 8 (2000) 583–587.

[34] D. Görür, C. E. Rasmussen, Dirichlet process gaussian mixture models: Choice of the base distribution, Journal of Computer Science and Technology 25 (4) (2010) 653–664.

[35] J. Gonzalez, Y. Low, A. Gretton, C. Guestrin, Parallel gibbs sampling: From colored fields to thin junction trees., Vol. 15, 2011, pp. 324–332.

[36] S. Wang, M. Mahfouf, A new computationally efficient Mamdani interval type-2 fuzzy modelling framework, Proc. IEEE FUZZ Conference, Brisbane Australia, 2012.

[37] K. J. Cios, G. W. Moore, Uniqueness of medical data mining, Artificial intelligence in medicine 26 (1) (2002) 1–24.