The Fuzzy Integral for Missing Data

Muhammad Aminul Islam and Derek T. Anderson
Department of Electrical and Computer Engineering
Mississippi State University
Mississippi State, MS, USA
Email: mi160@msstate.edu,anderson@ece.msstate.edu

Fred Petry, Denson Smith and Paul Elmore
Geospatial Sciences and Technology Branch
Naval Research Laboratory
Stennis Space Center, MS, USA
Email: {Fred.Petry,Denson.Smith,Paul.Elmore}@nrlssc.navy.mil

***Abstract*—Numerous applications in engineering are plagued by *incomplete* data. The subject explored in this article is how to extend the *fuzzy integral* (FI), a parametric nonlinear aggregation function, to missing data. We show there is no universally correct solution. Depending on context, different types of uncertainty are present and assumptions are applicable. Two major approaches exist, use just observed data or model/impute missing data. Three extensions are put forth with respect to just use observed data and a two step process, modeling/imputation and FI extension, is proposed for using missing data. In addition, an algorithm is proposed for learning the FI relative to missing data. The impact of using and not using modeled/imputed data relative to different aggregation operators–selections of underlying fuzzy measure (capacity)–are also discussed. Last, a case study and data-driven learning experiment are provided to demonstrate the behavior and range of the proposed concepts.**

Index terms - fuzzy integral, Choquet integral, missing data

I. Introduction

Incomplete data and information–otherwise referred to hereafter as data unless there is a specific reason to differentiate–is intrinsic to some applications and external to others. For example, in geo-spatial systems, Big Data and the Internet of Things, to name a few, it is common that one or more sensors malfunction or become non-operational and yield no value or noisy data. In a medical context, all data for a diagnosis may not be available, especially expensive and invasive tests. In skeletal age-at-death estimation, remains may not be available due to natural or active intervention (unnatural death) causes. Similarly, companies often rely on incomplete consumer data to develop marketing strategies as consumers may not have enough knowledge–or decline to provide data–for parts of a survey. The point is, we are often faced with fusing and making intelligent robust decisions from incomplete data.

In rare cases missing data can be reacquired. However, often the *cost* of reacquiring outweighs/diminishes the net utility of that data. In [1], Little and Rubin created a taxonomy for missing data. They identified three classes; *missing at random* (MAR), *missing completely at random* (MCAR) and *missing not at random* (MNR). The complexity of missing data is further exacerbated by Donald Rumsfeld's quote–"There are known knowns. These are things we know that we know. There are known unknowns. That is to say, there are things that we now know we don't know. But there are also unknown unknowns. There are things we do not know we don't know. [2]" The point is, missing data is not a trivial problem.

Before we dive into detail, it is important to review assumptions and *resources* that may or may not be available. These factors impact which procedure is selected. In some situations, no external knowledge is available and we are forced to use just what is observed. In other cases, historic data may be present. Other applications have access to other current observations where the present missing data is observed (similar to or the same as historic since data is available). Last, in some cases high-level a priori knowledge is available, e.g., a model (simulation or theoretical).

The bulk of missing data in *machine learning* (ML) literature is focused on topics like probabilistic graphical models (e.g., Bayesian nets), *support vector machines* (SVMs), regression or decision tree classifiers [3]–[9]. In [9], a decision tree classifier was extended by creating multiple test instances, where each took multiple paths along different splits of the missing attributes. A probability per leaf node was estimated based on the frequency of occurrence of training instances along the associated branch and the output was the class with the highest probability. In [10], Saar-Tsechansky and Provost learned multiple models for classification, where each model corresponds to an observed set of inputs encountered during training. Since the estimated model for a given set of observed input is optimal, it is expected to perform better than value based imputation but come at the cost of additional computation and storage. In soft-computing, Zhong et al. constructed information granules around imputed data [11]–which can be interpreted as fuzzification of imputed scalars. Effectiveness was measured in two aspects, cover and specificity. Cover ensures that all the available information is utilized whereas specificity measures the extent of uncertainty, i.e., the range of the interval. In [11], [12], the *fuzzy-c means* (FCM) clustering algorithm was used for imputation. Data is partitioned using just observed data. The grade of membership of a missing sample is determined by substituting the membership of the nearest observed sample.

The *Choquet integral* (ChI), a type of *fuzzy integral* (FI), is a powerful parametric nonlinear aggregation function that has been used in numerous applications like machine learning, multi-sensor fusion and decision theory (e.g., multi-criteria decision making). To the best of our knowledge, no attempt has been made to investigate missing data for the FI. Herein, we focus on two ways to extend the ChI to missing data (illustrated in Figure (1)). First, three normalizations are discussed

978-1-5090-6034-4/17/$31.00

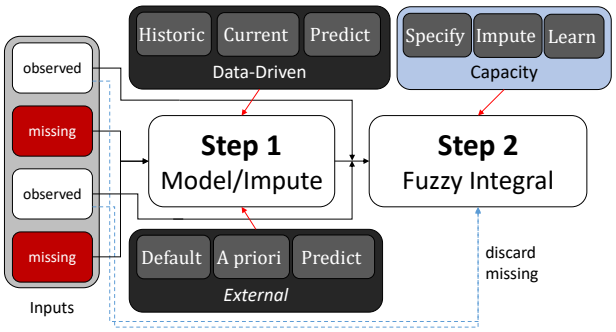


Fig. 1. High-level illustration of main concepts (elaborated on in Section III). The fuzzy integral uses observed data or observed and modeled/imputed data.

relative to using just observed data. Second, a two step process, modeling/imputation and ChI extension, is discussed relative to computing with respect to missing data. In addition, an algorithm is put forth for learning the ChI relative to missing training data. Last, we explore the impact of these choices relative to different aggregation philosophies–selections of underlying *fuzzy measure* (FM).

II. ChI for Complete Data

In this section, necessary definitions are provided for the ChI and its data-driven learning relative to complete data. Let $X = \{x_1, \ldots, x_N\}$ be a set of finite elements, e.g., sensors, experts, criteria or attributes in decision making, or algorithms in pattern recognition. A FM is a monotonic set-valued function defined on the power set of X, 2^X, as $\mu : 2^X \to \mathbb{R}^+$ that satisfies two properties; (i) boundary condition, $\mu(\emptyset) = 0$, and (ii) monotonicity, if $A, B \subseteq X$ and $A \subseteq B$, $\mu(A) \leq \mu(B)$. Often an additional constraint is imposed to limit the upper bound, i.e., $\mu(X) = 1$. Herein, without loss of generality we consider this condition for simplicity and convenience, which is useful in contexts like decision-level fusion.

Consider a training set with M observations and labels, $O = \{(\mathbf{o}_j, y_j)\}, j = 1, 2 \ldots, M$, where $\mathbf{o}_j \in \mathbb{R}^N$ is the jth observation, $y_j \in \mathbb{R}$ is the associated label, and $o_j(x_k)$ corresponds to the observed value for jth instance and kth input. Let $\mathbf{u} = [\mu(\{x_1\}), \mu(\{x_2\}), \ldots, \mu(\{x_1, x_2, \ldots, x_l\}), \ldots, \mu(X)]^T$ be the $2^N - 1$ dimensional vector containing all FM variables except $\mu(\emptyset)$. The discrete ChI on $\mathbf{o}_j$ with respect to μ is

$$C_\mu(\mathbf{o}_j) = \sum_{i=1}^{N} [o_j(x_{\pi_j(i)}) - o_j(x_{\pi_j(i-1)})]\mu(S_{\pi_j(i)}), \quad (1)$$

where π_j is a permutation function for observation $\mathbf{o}_j$ on the indices, $\{1, 2, \ldots, N\}$, that satisfies $0 \leq o_j(x_{\pi(1)}) \leq \ldots \leq o_j(x_{\pi(N)})$, where $S_{\pi_j(i)} = \{x_{\pi_j(i)}, x_{\pi_j(i+1)}, \ldots, x_{\pi_j(N)}\}$ and $o_j(x_{\pi_j(0)}) = 0$ [13]. Equation (1) can be written as

$$C_\mu(\mathbf{o}_j) = \mathbf{c}_j^T \mathbf{u}, \quad (2)$$

where $\mathbf{c}_j$ is a column vector containing the $(2^N - 1)$ coefficients for observation $\mathbf{o}_j$. Let k be the index of variable $\mu(B \in 2^X)$ in $\mathbf{u}$. Then the k-th element of $\mathbf{c}_j$ is $c_{jk} = o_j(x_{\pi_j(l)}) - o_j(x_{\pi_j(l-1)})$ if $\exists S_{\pi_j}(l) = B$ for $l \in \{1, \ldots, N\}$, and 0 otherwise. The monotonicity constraints can be written as $\mu(A) \leq \mu(A \cup q), \forall A \subset X$ and $\forall q \in X, q \notin A$. The *sum of squared error* (SSE) between the ChI for all the observations in the training data, O, and corresponding labels is

$$E_1(O, \mathbf{u}) = \sum_{j=1}^{M} (C_\mu(\mathbf{o}_j) - y_j)^2 = \sum_{j=1}^{M} (\mathbf{c}_j^T \mathbf{u} - y_j)^2$$
$$= \sum_{j=1}^{M} (\mathbf{u}^T \mathbf{c}_j \mathbf{c}_j^T \mathbf{u} - 2 y_j \mathbf{c}_j^T \mathbf{u} + y_j^2). \quad (3)$$

Based on this, the least square minimization problem can be expressed as the *quadratic programming* (QP)

$$\text{(OP1)} \ \min_{\mathbf{u}} f_O(\mathbf{u}) = \mathbf{u}^T H \mathbf{u} + \mathbf{d}^T \mathbf{u},$$

$$\mu(A) \leq \mu(A \cup q), \quad \forall A \subset X \text{ and } \forall q \in X, q \notin A,$$
$$\text{(monotonicity conditions)} \quad (4a)$$
$$\mu(\emptyset) = 0, \quad \text{(boundary conditions)} \quad (4b)$$
$$\mu(X) = 1, \quad \text{(normality conditions)} \quad (4c)$$

where $H = \sum_{j=1}^{M} \mathbf{c}_j \mathbf{c}_j^T$ and $\mathbf{d} = -2 \sum_{j=1}^{M} y_j \mathbf{c}_j$. This can be minimized by standard QP solvers. Herein, OP1 is used for the training while Equation (2) is used for prediction.

III. ChI for Missing Data

First, we establish some notation for the following subsections. For instance $\mathbf{o}_j$, inputs (data) are available for the set $X_j^w \subseteq X$ such that $x_i \in X_j^w, i \in I_j^w, I_j^w \subseteq \{1, 2, \ldots, N\}$ and data is not available for $X_j^h = X \backslash X_j^w$. Last, let the cardinality of X_j^w and X_j^h be N_j^w and N_j^h respectively.

Remark 1. We start by exploring if it is acceptable to calculate just the subset of Equation (1) associated with observed data. It turns out that doing so is equivalent to imputing zeros for the missing data. While it does not *mathematically break* the ChI, relative to imputed data, these zeros (or other constants at that) semantically impact aggregation and properties like idempotency (namely our expectations on it). There are better philosophies (following sections) that do not force us to inject potentially misleading/uninformed data and allows us to model and compute with respect to our uncertainty.

A. *Method 1*: ChI for Observed Data Only

In this subsection we restrict the ChI to only using available data and therefore the subset of the FM associated with our observed data. However, since the FM is defined on N, adding or removing inputs can (it is trivial to show) numerically and semantically alter the ChI and result in a measure that violates the properties of a FM, namely the upper boundary condition. We discuss three approaches to overcome this.

First, we define a "sub-ChI/FM" on instance $\mathbf{o}_j$, $C_{\mu_j^w}(\mathbf{o}_j)$, with respect to all available data (X_j^w),

$$C_{\mu_j^w}(\mathbf{o}_j) = \sum_{i=1}^{N_j^w} [o_j(x_{\pi_j^w(i)}) - o_j(x_{\pi_j^w(i-1)})]\mu_j^w(S_{\pi_j^w(i)}), \quad (5)$$

where π_j^w is the permutation operation for $\mathbf{o}_j$ performed on observed sources. A first idea is to let $\mu_j^w(A) = \mu(A), \forall A \subseteq X_j^w$. However, $\mu(X_j^w)$ may not be equal to 1. Thus, we have a boundary condition difference of

$$\Delta\mu_j^w(X_j^w) = 1 - \mu(X_j^w). \tag{6}$$

Whereas monotonicity is obviously preserved, important properties like boundedness and idempotency are not guaranteed for $C_{\mu_j^w}(\mathbf{o}_j)$. The problem resides in directly taking FM values from μ. In order to realize a *legitimate* (property preserving) $C_{\mu_j^w}(\mathbf{o}_j)$ we need to calculate μ_j^w as a monotonic and boundary condition preserving function of μ. By definition, $\mu(S_{\pi_j^w(i-1)}) \geq \mu(S_{\pi_j^w(i)})$, which gives the condition $\Delta\mu_j^w(S_{\pi_j^w(i-1)}) \geq \Delta\mu_j^w(S_{\pi_j^w(i)})$ on relative change in the FM. This condition along with Equation (6) sets the boundary for $\Delta\mu_j^w(S_{\pi_j^w(i-1)})$ as $0 \leq \Delta\mu_j^w(S_{\pi_j^w(i-1)}) \leq 1 - \mu(X_j^w)$. The question is, how do we determine the $\Delta\mu_j^w$ values?

Upper boundary fix: The first and simplest way to remedy the problem is to fix what is broken; $\mu_j^w(X_j^w) = 1$ and all other variables are $\mu_j^w(A) = \mu(A)$. This method satisfies the boundary conditions and monotonicity. However, an unwanted side effect, when $\Delta\mu_j^w(X_j^w) > 0$, is the introduction of a larger than previously modeled difference between the normalized $\mu_j^w(X_j^w)$ (now inflated) and $\mu_j^w(X_j^w \setminus x_k)$ $(\forall x_k \in X_j^w)$ terms.

Additive normalization: Another idea is to add $\Delta\mu_j^w(X_j^w)$ to each FM variable, except for $\mu_j^w(\emptyset)$. This ensures boundary conditions and monotonicity. This procedure preserves the relative differences between consecutive (increasing cardinality) normalized μ_j^w variables, except for the difference between the $\mu_j^w(\{x_k\})$ $(\forall x_k \in X)$ and $\mu_j^w(\emptyset)$. Furthermore, if $\mu(A) = 0$, $\forall A \subset X \setminus \emptyset$, then $\mu_j^w(A) \geq 0$ when $\Delta\mu_j^w(X_j^w) \geq 0$. That is, a subset that was previously given no weight/importance/utility is no longer zero valued.

Uniform normalization: The last fix considered is to divide each term by $\mu_j^w(X_j^w)$ (when $\mu_j^w(X_j^w) \neq 0$). This ensures boundary conditions, is monotonic and does not have the inflation side effects of upper boundary fix nor additive normalization. This scaling is semantically more justifiable.

Remark 2. (Limitations of the observation-valued ChI) An advantage of using just observed data is we are computing with respect to what we know and the result is bounded between the minimum and maximum observations, i.e., we will not infer values higher or lower that we were informed. However, this advantage is also its shortcoming. If the unobserved data is critical to the task at hand then it will obviously fall short as no attempt was made to model/impute the missing data and our associated uncertainty.

B. *Method 2:* Modeling and Imputing Missing Data

In this subsection, we highlight various approaches for modeling and imputing missing data. This is not a task unique to our paper nor is it a primary contribution of our work. This is actually a strong example of how the field of fuzzy set theory can or already does contribute to missing data problems. In general, we have the following categories.

1) **Discard**; inputs for missing data are ignored.
2) **Default**; a user/system/etc. defined value is used.
3) **Historic data**; prior complete or missing data is used.
4) **Current data**; data from other present instances where the missing data inputs are observed is used.
5) **A priori**; use high-level knowledge external to the data.
6) **Prediction**; use any of the above in combination with the current observed instance data to predict missing data.

Options three to six are research topics. Obviously, discarding and default are trivial. The task of prediction is beyond the scope of this work (see [14] for a recent review). Furthermore, a priori knowledge is task dependent and not explored herein. We touch on historic and/or current data (categories 3 and 4).

Example 1. (Fitting a distribution to historic and/or current data) When historic and/or other current instances are available we can fit a distribution to that data. However, what distribution do we use? There are a number of distributions (e.g., normal, trapezoidal, etc.) and distribution fitting methods (e.g., construction of a type-2 fuzzy set [15], Gaussian mixture model, etc.). In general, the selection of theory depends on the *nature* of the underlying uncertainty. For example, often we follow the central limit theorem and model data via a normal distribution, which is fully determined by its mean and variance. The point is, there are numerous ways (beyond the ChI scope of this work) to derive distributions from historic and/or current observations.

Example 2. (Interval modeling) Another option is to model missing data as an interval, i.e., $\bar{o}_j(x_k) = [o_j^-(x_k), o_j^+(x_k)]$, $o_j^-(x_k) \leq o_j^+(x_k)$. However, as before, where does this interval come from? For observed data there is no uncertainty, i.e., $o_j^-(x_k) = o_j^+(x_k)$. For missing data, there are a number of possibilities. If uncertainty is modeled as a distribution then we can calculate $\beta\sigma_{x_k}$, i.e., $\bar{o}_j(x_k) = [m_{x_k} - \beta\sigma_{x_k}, m_{x_k} + \beta\sigma_{x_k}]$, where m_{x_k} is the mean and σ_{x_k} is the variance of x_k, and β is an arbitrary constant. Alternatively, if nothing is known about the missing data and no assumptions are applicable, then minimum and maximum values can be used. In the extreme case, we can associate each missing input with $\bar{o}_j(x_k) = [0, 1]$, aka total ignorance.

C. *Method 2:* Using Modeled or Imputed Data

In subsection III-B we touched on different ways to model uncertainty. Note, we do not care about the underlying *meaning* (e.g., probability, possibility, etc.). In this subsection we tackle how to use such interval or set-valued data in the ChI. We start with the $\Re$-valued ChI, followed by the interval-valued ChI then type-1 and type-2 integrand-valued ChIs. Figure (2) shows different options (pathways) between modeling/imputation choices and FI extensions. In general, there is no "winner". Ideally, in the theme of David Marr's Principle of Least Commitment and Principle of Graceful Degradation, we advocate the modeling of computation with respect to *full* (set-valued) data. However, some applications may impose restrictions (eliminate options). Furthermore, dif-

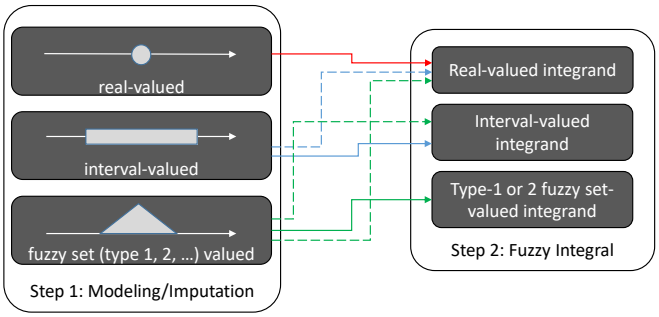


Fig. 2. Illustration of *connections* (pathways) between modeling/imputation choices and FI extensions. Dashed lines indicate type reduction.

ferent applications might require type reduction if lower order uncertainty is desired or scalar-valued outputs are needed.

Scalar-valued ChI: Equation (1) is the simplest and most naive approach considered. The scalar-valued ChI does not take into account uncertainty about missing data. As expected, this action has consequences (elaborated on in Example 3).

Example 3. (Tendency of modeled/imputed data to dictate the ChI) For sake of analysis and without loss of generality, we consider that all observed values lie within $[0, 1]$, i.e., $o_j(x_i) \in [0, 1], \forall i, j$. Let the default imputed value be c, i.e., $o_j(x_i) = c,\ \forall j,\ x_i \in X^h, c \in [0, 1]$. It is trivial to show that this method runs the risk of controlling/determining the ChI. For example, consider a maximum ChI, i.e., $\mu_1(A) = 1, \forall A \subseteq X \setminus \emptyset$ and a minimum ChI, i.e., $\mu_2(A) = 0, \forall A \subset X$. If $c = 1$ then $C_{\mu_1}(o_j) = 1$ regardless of what is observed. Conversely, if $c = 0$ then $C_{\mu_2}(o_j) = 0$. While these are extreme cases, they illustrate the fact that imputed data is important. If the value c is a default constant with little-to-no meaning, then the output of the ChI is questionable, to say the least. For example, if we let $c = 0.5$ to express "ignorance" and all observed values are higher than 0.5 then μ_2 will always yield 0.5 (same argument holds for all observed values less than 0.5 and μ_1). In reality, we have uncertainty and our result does not show that.

Distribution-to-scalar ChI: A logical next step is to type reduce uncertain data for use in Equation (1). For example, if a normal distribution is modeling historic and/or current data then the first moment, the mean, can be used. Semantically, our fusion result now reflects a combination of what we observed and the expected value of missing data. Whereas we improved the modeling of missing data we still sadly *loose* information when we type reduced for the ChI.

Interval-valued ChI: Next, we outline how to use interval-valued uncertainty about missing data in the ChI. Regardless of how the uncertainty is obtained, e.g., type-reduction of a normal distribution into a second moment interval, $\bar{o}_j(x_k) = [m_{x_k} - \beta\sigma_{x_k}, m_{x_k} + \beta\sigma_{x_k}]$, the interval-valued ChI [16],

$$\bar{C}_\mu(\mathbf{o}_j) = [C_\mu(\mathbf{o}_j^-), C_\mu(\mathbf{o}_j^+)], \tag{7}$$

is two $\Re$-valued ChIs, one integral on the interval left endpoints and a different integral on the interval right endpoints.

Remark 3. (Trapezoidal Membership Functions) Consider the trapezoidal membership function, specified by four $\Re$-valued numbers, $a \leq b \leq c \leq d$. It is well-known that a trapezoidal membership function is fully characterized by two level (alpha) cuts, one at 0, associated with $[a, d]$ and one at 1 associated with $[b, c]$. First, we level cut each set. Next, we compute the ChI at each level. Last, we type increase the ChI intervals back into a trapezoidal membership function. Note, there was no information loss. It is trivial to prove this as the ChI is monotonic and the trapezoidal values at each level cut between 0 and 1 are linear equations.

Type-1 and type-2 fuzzy set-valued ChIs: In [15], we put forth two ChI extensions, the gFI and NDFI, for unrestricted (potentially subnormal and non-convex) fuzzy set-valued integrands. In [17], we put forth a ChI extension for type-2 fuzzy sets. There is not sufficient space to review these extensions. The reader can see [15], [17] for mathematical and algorithmic details. The point is, if missing data is modeled by fuzzy sets then there are ChI extensions that can handle this task.

Remark 4. (Amount of missing data) The previous sections are focused on modeling, imputation and ChI computation. In Example 3 it was observed that the answer can be drastically skewed by modeled or imputed data. The question investigated here is, what is the impact of different amounts of missing data? It turns out that this is difficult-to-impossible to solve as the answer depends on the missing data, but more importantly the selection of aggregation operator (underlying capacity). Meaning, if an *extreme* operator such as maximum or minimum (t-conorm (union) and t-norm (intersection) respectively) is selected then one poorly modeled or imputed data can have a catastrophic impact, regardless of how much data is missing. On the other hand, a robust operator like the expected value (e.g., mean, median, etc.) can be expected to perform better– of course degrading with respect to amount of missing data. This holds regardless of modeling/imputation and subsequent ChI computation method. For example, Example 3 naturally extends to the interval-valued ChI (two $\Re$-valued ChIs) and fuzzy set-valued ChI (which is decomposable to interval ChIs and then $\Re$-valued ChIs). The point of this remark is not mathematical characterization but observation that we need to select our aggregation operator with care for missing data.

IV. Missing Data-Driven ChI Learning

In this section, a data-driven method is put forth for observed data only ChI learning. In [18], we put forth a data-driven QP and regularization (for minimal model complexity) approach to ChI learning. However, like all other ChI learning algorithms to date it was assumed that no data is missing. Here, we outline a way to learn the ChI with respect to a set of training data with missing inputs.

Since OP1 is expressed in terms of the full lexicographically encoded μ, one possibility is to represent μ_j^w in terms of μ to facilitate optimization of a common set of variables. First, we represent each Δ as a factor of $1 - \mu(X_j^w)$,

$$\Delta\mu_j^w(S_{\pi_j^w(i)}) = \alpha(S_{\pi_j^w(i)})(1 - \mu(X_j^w)),$$

where $\alpha(S_{\pi_j^w(1)}) = 1$ and $0 \leq \alpha(S_{\pi_j^w(i)}) \leq 1,\ i = 2, \ldots, N_j^w$. The parameter $\alpha(S_{\pi_j^w(i)})$ is user defined and it can be selected per layer or node in the FM[1]. Note, we already discussed three applicable strategies; e.g., upper boundary fix, $\alpha(S_{\pi_j^w(1)}) = 1$ and $\alpha(S_{\pi_j^w(i)}) = 0$ for $i = 2, \ldots, N_j^w$. The ChI with respect to μ_j^w, in terms of μ, is

$$
\begin{aligned}
C_{\mu_j^w}(\mathbf{o}_j) &= \sum_{i=1}^{N_j^w} [o_j(x_{\pi_j^w(i)}) - o_j(x_{\pi_j^w(i-1)})][\mu(S_{\pi_j^w(i)}) \\
&+ \alpha(S_{\pi_j^w(i)})(1 - \mu(X_j^w))] \\
&= \sum_{i=2}^{N_j^w} [o_j(x_{\pi_j^w(i)}) - o_j(x_{\pi_j^w(i-1)})]\mu(S_{\pi_j^w(i)}) + \\
&\sum_{i=1}^{N_j^w} [o_j(x_{\pi_j^w(i)}) - o_j(x_{\pi_j^w(i-1)})]\alpha(S_{\pi_j^w(i)}) - \sum_{i=2}^{N_j^w} [o_j(x_{\pi_j^w(i)}) \\
&- \mu(X_j^w) o_j(x_{\pi_j^w(i-1)})]\alpha(S_{\pi_j^w(i)}),
\end{aligned}
$$

which can be expressed in matrix form as

$$C_{\mu_j^w}(\mathbf{o}_j) = \mathbf{c}_j^{wT}\mathbf{u} + b, \tag{8}$$

where the coefficient c_{jk}^w for the kth variable, $\mu(B)$, is $o_j(x_{\pi_j^w(l)}) - o_j(x_{\pi_j^w(l-1)})$ if $\exists S_{\pi_j^w}(l) = B, l \in \{2, \ldots, N_j^w\}$, $-\sum_{i=2}^{N_j^w}[o_j(x_{\pi_j^w(i)}) - o_j(x_{\pi_j^w(i-1)})]\alpha(S_{\pi_j^w(i)})$, if $S_{\pi_j^w}(1) = B$, 0 otherwise and $b = \sum_{i=1}^{N_j^w}[o_j(x_{\pi_j^w(i)}) - o_j(x_{\pi_j^w(i-1)})]\alpha(S_{\pi_j^w(i)})$. The SSE for the data is therefore

$$E_2(O, \mathbf{u}) = \sum_{j=1}^{M} (\mathbf{c}_j^{wT}\mathbf{u} + b - y_j)^2 = \sum_{j=1}^{M} (\mathbf{c}_j^{wT}\mathbf{u} - y_j^w)^2,$$

where $y_j^w = y_j - b$. This SSE equation has the same form as Equation (3) and therefore OP1 can be used to learn the FM – $\mathbf{c}_j$ and y_j need to be replaced by $\mathbf{c}_j^w$ and y_j^w respectively. Equation (8) can be used to compute the ChI of data with respect to μ and scaling coefficients $\alpha(S_{\pi_j^w(i)})$.

V. Case Study and Synthetic Experiment

The previous sections are focused on concepts and methods for fusing missing data with the FI. This section explores these ideas via a case study and synthetic experiment. Energy is not expended on proving that the FI is a useful tool for any one application, e.g., sensor data fusion, computer vision, multi-criteria decision making, etc., this has already been well-established by the field.

A. Case Study from Forensic Anthropology

In this subsection, we discuss fusion for skeletal-age-at-death estimation in forensic anthropology. In [16], [19]–[22], the task is to determine the age that an individual died, via natural or by active intervention (unnatural death), based on recovered skeletal remains and established aging methods. Different bones are usually present–making it a missing data problem. Sadly, this was not recognized, at a minimum appreciated, until now. For each bone and aging method, we obtain an input fuzzy set defined on the "age domain" (e.g., $[0, 120]$) where the membership degree is the support in age-at-death. For example, we might have evidence of age-at-death based on the Pubic Symphysis, Auricular Surface and Ectocranial Valt Suture Closure (three inputs from three different bones and aging methods, e.g., Todd 1920, Lovejoy et al. 1985b and Meindl and Lovejoy 1985). Input one might tell us $[25, 26]$, input two might say $[25, 29]$ while input three says $[24, 60]$. Along with each input we have a *bone quality*. We assign a score of $\{1.0, 0.8, 0.6, 0.4, 0.2, 0.1, 0.0\}$ based on the weathering of the bone (corresponding to Stages 0 to 6 for a modified version of Behrensmeyer weathering stages provided in Standards). The result is typically a subnormal-valued convex fuzzy set. In addition, we use known accuracies (correlation coefficients) for each aging method. These values are the FM singleton variables (aka densities). As we did not have access to the full FM, the Sugeno λ FM [23] (a well-known imputation method) was used to assign the remaining FM variables relative to the densities for observed data.

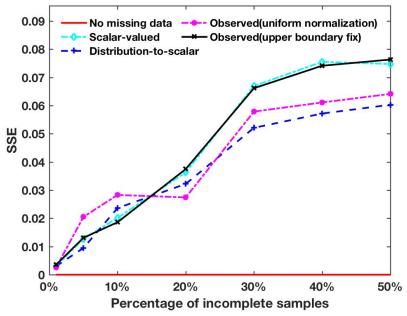
This forensic application is discussed herein because it's an interesting example of the observation only FI. In our current article, one idea is to sample the relevant subset of the FM and normalize it. However, in our forensic application we only have access to the densities–which could be seen as a missing "data" challenge for μ. As the two and above tuples are unknown, an imputation algorithm (the Sugeno λ formula) is used to assign FM variable values. If we had privilege to the full FM then we would follow the procedure outlined in this article. In [24] Nguyen et al. showed that the Sugeno λ FM is, from a purely mathematical standpoint, equivalent to a probability (i.e., additive) measure. There is a computational advantage to using the Sugeno λ FM versus reducing it into its equivalent probability measure. Specifically, they showed that we can perceive the Sugeno λ FM as a re-scaling of the underlying probability measure. The point is, our prior forensic work addressed missing data for both h (observation only FI) and μ (an additive measure-based imputation formula that is monotonic and normality preserving).

Last, in many cases a remain (e.g., skull) is missing. Previously, we discarded such inputs. Based on the current article, we plan to next explore the use of interval and fuzzy set-valued modeling and imputation forensic algorithms for missing data. Whereas complete ignorance, e.g., $[0, 120]$, might prove to be too extreme, if recovered remains allow us to narrow the range of other missing data then we would like to explore the benefits of computing with respect to uncertainty. At the moment, our current approach is the observation only FI, which is subject to the limitation discussed in Remark 2.

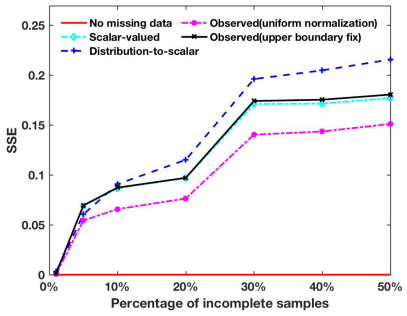
B. Synthetic Example: Data-Driven ChI Learning

In this subsection, we conduct a synthetic experiment for data-driven ChI learning. A controlled experiments is preferred to "real data" because we can better investigate benefits and drawbacks related to missing data versus limiting ourself to the

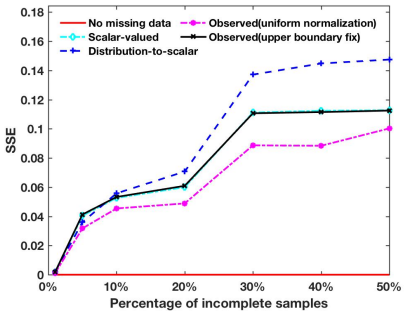
[1]Alternatively, the scaling function α can be learned in conjunction with the FM, possibly via alternating optimization (subject of future work).



(a) SSE of learned FM and ground truth FM

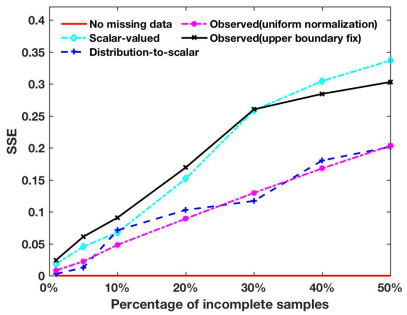


(b) SSE for training data

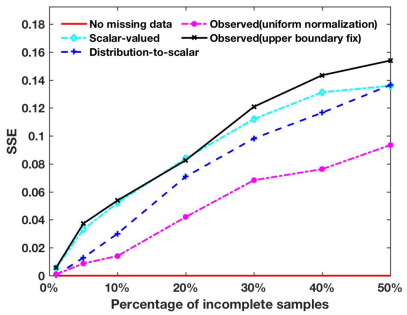


(c) SSE for test data

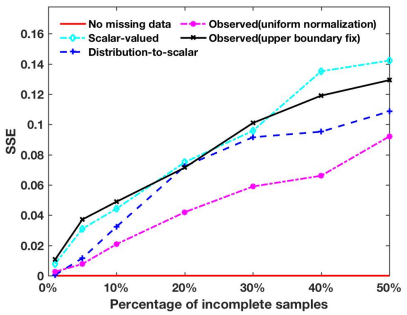
Fig. 3. Results for soft-max aggregation operator.



(a) SSE of learned FM and ground truth FM

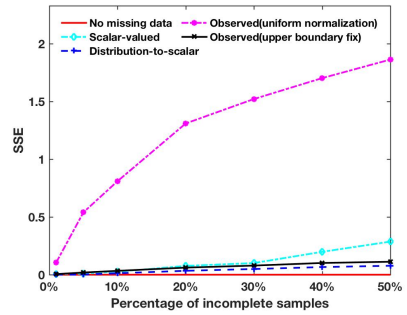


(b) SSE for training data

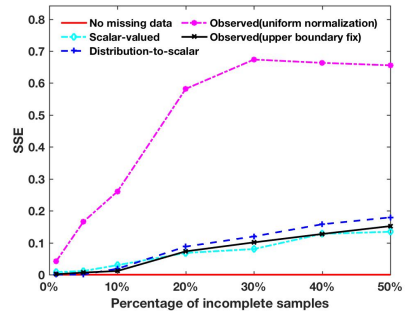


(c) SSE for test data

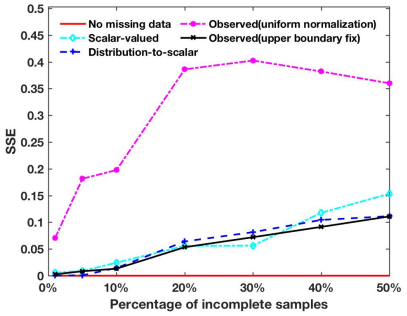
Fig. 4. Results for mean aggregation operator.



(a) SSE of learned FM and ground truth FM



(b) SSE for training data



(c) SSE for test data

Fig. 5. Results for soft-min aggregation operator.

"scope" of available data or accuracy results for a process that we do not truly know the underlying answer for. We consider three *ordered weighted averages* (OWAs) [25], which are just *linear combinations of order statistics* (LCOS) when the inputs and weights are $\Re$-valued, which the ChI can generate. The three OWAs are soft-max (t-conorm, union like operator), mean (expected value operator) and soft-min (t-norm, intersection like operator). These three operators are helpful because they cover the "spectrum" of commonly encountered aggregation philosophies (optimistic, expected value and pessimistic). We let $N = 5$, $M = 150$ and the OWA weights are $(0.71, 0.155, 0.077, 0.039, 0.0019)$ for soft-max, $\frac{1}{5}$ for mean and $(0.0019, 0.039, 0.077, 0.155, 0.71)$ for soft-max,. The data was (pseudo)randomly generated from a truncated normal distribution in $[0, 1]$ with means $[0.50\ 0.17\ 0.33\ 0.67\ 0.83]$ and standard deviation 0.45. The performance of the different proposed methods are evaluated for different percentages of missing data, $[1\%\ 5\%\ 10\%\ 20\%\ 30\%\ 40\%\ 50\%]$. One input was removed from each sample marked as missing. Three-fold cross-validation is used. The average SSE for the FM to the ground truth FM and also the SSE on the training and test data are used as performance metrics.

A value of 0.5 is selected for scalar-valued imputation–as it debatably best represents ignorance, 0 being no support and 1 being full support. For distribution-to-scalar, we modeled data using a normal distribution and it is type reduced to the mean. Last, both uniform normalization and upper boundary fix are explored for the observation value only FI. Figures (3), (4), and (5) are experiment results. Before we delve into specifics, we highlight that different aggregation philosophies (i.e., selections of FM) result in different trends and no global pattern is present (as expected). This is further elaborated on at the end of the subsection.

Figure (3) are the results for soft-max (optimistic aggregation). Overall, the observed-value method (the "safe approach") with uniform normalization leads in each performance criteria. While distribution-to-scalar does well with respect to modeling the true underlying FM, its training and prediction SEE is the worst. Observed-value with upper boundary fix has relatively poor performance for soft-max as imputation depends on the minimum of the observed values. On the other hand, uniform scaling puts more weight on the higher observed values, yielding lower error for soft-max.

Results for the mean aggregation operator are reported in Figure (4). Overall, observed data only with uniform normalization is best again. However, distribution-to-scalar is now second best (versus worst). This is expected as we are using the means of the missing data distributions. Note, observed-value with uniform normalization considers only the observed inputs. As such, its scaling of the FM terms is equivalent to imputing missing values proportional to the FM of the observed sources. Therefore, the imputed value in mean FM is proportional to the sum (or mean) of the observed values.

Figure (5) are the results for soft-min (pessimistic operator). Whereas observed data with uniform normalization was best before, it is by far the worst here. The reason is, in soft-min the top node of the sub-lattice is relatively small and therefore the FM values are scaled by a larger factor than in soft-max–making the normalized sub-FM closer to a FM of all ones (which is the maximum operator for the ChI). Distribution-to-scalar and observed data with upper boundary fix provide better results, where the imputation of the latter depends on the minimum of the observed values, which partially helps to preserve the aggregation behavior.

Overall, as stated in the start of this subsection, the most interesting takeaway is the performance variation across aggregation *philosophies* (minimum, average and maximum). There is no global winner across percentages of missing data and aggregations, which is expected based on the theory and remarks in previous sections. This is interpreted as follows. Missing data is not a simple problem. Care should be given with respect to studying all that is possible for for a task at hand. The next step is to explore the different options, compute with respect to just observed data or model/impute, and it all should be done relative to underlying aggregation operator (selection of FM).

VI. Conclusion and Future Work

Herein, the question of how to extend the FI to problems involving missing data was investigated. We identified two paths, use just observed data or fuse modeled/imputed data with observed data. Three strategies were put forth to make the ChI work with respect to observed data and a complete FM. Furthermore, different modeling and FI extensions were discussed for the inclusion and fusion with respect to missing data. It was shown that there is no "winning method". Instead, the different choices are dictated by the properties of an application (context). It was shown that whereas the observation-valued ChI is a safe route, it runs the risk of not including uncertainty about our missing data. However, while modeled/imputed data leads to a more informed result, if we are not careful about what aggregation operator we use then missing data often dictates, in a potentially destructive fashion, the result. Last, a data-driven algorithm was presented for historic and/or other current observed data.

This article is just a first step towards missing data and the FI. In future work, we will expand our scope to include both missing integrand (h) and missing FM (μ), the latter only slightly touched on in our case study. We will also attempt to simultaneously solve for the FM in conjunction with how to normalize it (versus specify the normalization). We will also explore pathways involving "up sampling" (type increasing). We also want to provide a field guide for practitioners so a user can connect their applications properties to the best set of modeling, imputation and FI extension choices. Next, we put forth an extension for data driven learning. In future work we will study interval and fuzzy set valued label driven learning, or learning of such data in light of scalar-valued inputs. Last, we observed that selection of aggregation operator has a big impact, to say the least. In future work a goal will be to connect different missing data choices to the most appropriate aggregation operator (i.e., underlying FM).

Acknowledgment

Dr. Petry and Dr. Elmore would like to thank the Naval Research Laboratory's Base Program, Program Element No. 0602435N for sponsoring this research.

References

[1] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[2] D. Rumsfeld, "Defense department briefing," Feburary 2002.

[3] M. Ramoni and P. Sebastiani, "Robust learning with missing data," *Machine Learning*, vol. 45, no. 2, pp. 147–170, 2001.

[4] ——, "Learning bayesian networks from incomplete databases," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1997, pp. 401–408.

[5] K. Mohan, G. Van den Brock, A. Choi, and J. Pearl, "An efficient method for bayesian network parameter learning from incomplete data," in *Causal Modeling and Machine Learning Workshop*, vol. 951, 2014, p. 2014.

[6] A. J. Smola, S. Vishwanathan, and T. Hofmann, "Kernel methods for missing variables." in *AISTATS*. Citeseer, 2005.

[7] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an em approach," *Advances in neural information processing systems*, pp. 120–120, 1994.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[9] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[10] M. Saar-Tsechansky and F. Provost, "Handling missing values when applying classification models," *Journal of machine learning research*, vol. 8, no. Jul, pp. 1623–1657, 2007.

[11] C. Zhong, W. Pedrycz, D. Wang, L. Li, and Z. Li, "Granular data imputation: A framework of granular computing," *Applied Soft Computing*, vol. 46, pp. 307–316, 2016.

[12] I. B. Aydilek and A. Arslan, "A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm," *Information Sciences*, vol. 233, pp. 25–35, 2013.

[13] P. Benvenuti, R. Mesiar, and D. Vivona, "Monotone set functions-based integrals," *Handbook of measure theory*, vol. 2, pp. 1329–1379, 2002.

[14] P. J. García-Laencina, J. L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: a review," *Neural Computing and Applications*, vol. 19, no. 2, pp. 263–282, 2010. [Online]. Available: http://dx.doi.org/10.1007/s00521-009-0295-6

[15] C. Wagner, S. Miller, J. M. Garibaldi, D. T. Anderson, and T. C. Havens, "From interval-valued data to general type-2 fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 2, pp. 248–269, April 2015.

[16] D. T. Anderson, T. C. Havens, C. Wagner, J. M. Keller, M. F. Anderson, and D. J. Wescott, "Extension of the fuzzy integral for general fuzzy set-valued information," *IEEE Transactions on Fuzzy Systems*, vol. 22, no. 6, pp. 1625–1639, Dec 2014.

[17] T. C. Havens, D. T. Anderson, and J. M. Keller, "A fuzzy choquet integral with an interval type-2 fuzzy number-valued integrand," in *International Conference on Fuzzy Systems*, July 2010, pp. 1–8.

[18] D. T. Anderson, S. R. Price, and T. C. Havens, "Regularization-based learning of the choquet integral," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, July 2014, pp. 2519–2526.

[19] M. F. Anderson, D. T. Anderson, and D. J. Wescott, "Estimation of adult skeletal age-at-death using the sugeno fuzzy integral," *American Journal of Physical Anthropology*, vol. 142, no. 1, pp. 30–41, 2010. [Online]. Available: http://dx.doi.org/10.1002/ajpa.21190

[20] D. T. Anderson, T. C. Havens, C. Wagner, J. M. Keller, M. F. Anderson, and D. J. Wescott, "Sugeno fuzzy integral generalizations for sub-normal fuzzy set-valued inputs," in *2012 IEEE International Conference on Fuzzy Systems*, June 2012, pp. 1–8.

[21] D. T. Anderson, J. M. Keller, M. Anderson, and D. J. Wescott, "Linguistic description of adult skeletal age-at-death estimations from fuzzy integral acquired fuzzy sets," in *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, June 2011, pp. 2274–2281.

[22] D. T. Anderson, P. Elmore, F. Petry, and T. C. Havens, "Fuzzy choquet integration of homogeneous possibility and probability distributions," *Information Sciences*, vol. 363, pp. 24 – 39, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0020025516302961

[23] M. Sugeno, "Fuzzy measure and fuzzy integral," *Fuzzy Measure and Fuzzy Integral*, vol. 8, no. 2, pp. 94–102, 1972.

[24] H. Nguyen, V. Kreinovich, J. Lorkowski, and S. Abu, "Why sugeno -measures," *Technical Report: UTEP-CS-15-17*, 2015.

[25] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Transactions on systems, Man, and Cybernetics*, vol. 18, no. 1, pp. 183–190, 1988.