HAL open science

Multi-dimensional maximal coherent subsets made easy: illustration on an estimation problem

Loïc Adam, Sébastien Destercke

To cite this version:

Loïc Adam, Sébastien Destercke. Multi-dimensional maximal coherent subsets made easy: illustration on an estimation problem. 10th International Conference on Soft Methods in Probability and Statistics (SMPS 2022), Sep 2022, Valladolid, Spain. pp.1-8, 10.1007/978-3-031-15509-3_1 . hal-03896619

HAL Id: hal-03896619

https://hal.science/hal-03896619v1

Submitted on 13 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Multi-dimensional maximal coherent subsets made easy: illustration on an estimation problem

Loïc Adam and Sébastien Destercke

Abstract Fusing uncertain pieces of information to obtain a synthetic estimation when those are inconsistent is a difficult task. A particularly appealing solution to solve such conflict or inconsistency is to look at maximal coherent subsets of sources (MCS), and to concentrate on those. Yet, enumerating MCS is a difficult combinatorial task in general, making the use of MCS limited in practice. In this paper, we are interested in the case where the pieces of information are multi-dimensional sets or polytopes. While the problem remains difficult for general polytopes, we show that it can be solved more efficiently for hyperrectangles. We then illustrate how such an approach could be used to estimate linear models in the presence of outliers or in the presence of misspecified model.

1 Introduction

This paper deals with the problem of fusing multiple pieces of information (Dubois et al. (2016)). In this problem, handling conflict between contradicting sources of information is one of the most difficult tasks. This is often a mandatory task, including in situations where we want a synthetic estimation from all the sources. Moreover, analyzing the reasons of the contradiction and trying to explain its appearance can be of equal importance, as it can give important insights about the situation.

Dubois et al. (1999) reviewed different methods for aggregating conflict, some requiring additional data like the reliability of the different sources. Yet, having additional data is sometimes difficult or even impossible. A quite appealing way for

Loïc Adam
UMR CNRS 7253 Heudiasyc, Sorbonne Université, Université de Technologie de Compiègne CS 60319 - 60203 Compiègne cedex, France, e-mail: loic.adam@hds.utc.fr

Sébastien Destercke (✉)
UMR CNRS 7253 Heudiasyc, Sorbonne Université, Université de Technologie de Compiègne CS 60319 - 60203 Compiègne cedex, France, e-mail: sebastien.destercke@hds.utc.fr

dealing with contradiction, requiring no additional information, is based on maximal coherent subsets (MCS), which are groups of consistent sources that are as big as possible. MCS have been used in the past both in logic (Manor and Rescher (1970)) and in numerical settings (Destercke et al. (2008)). Here, we illustrate their application to estimation problems.

Detecting and enumerating MCS are NP-hard problems, with intervals being a well-known exception (Dubois et al. (2000)). In this paper, we show that we can extend this exception to hyperrectangles, which can in turn be used as approximations of polytopes.

In Section 2 we further introduce maximal coherent subset and explain the difficulty of enumerating them. We then show in Section 3 that it is easy to list the MCSs of a set of axis-aligned hyperrectangles. Lastly, Section 4 illustrates our approach on linear regression problems.

2 Maximal Coherent Subsets

As mentioned previously, maximal coherent subsets are in theory a nice solution to manage conflict between information sources. Moreover, they can be used with different structures of information, like polytopes. However, we will show that listing the different MCSs is usually a difficult combinatorial problem.

General definition: Let us suppose we have a set $\mathcal{S} = \{S_1, ..., S_N\}$ of sources of information providing a subset $S_i \subseteq \mathcal{X}$ of some space $\mathcal{X}$ of information, for which intersection $\cap$ is well-defined. A maximal coherent subset $c \subseteq \{1, \ldots, N\}$ is a list of source indices such that $\cap_{i \in c} S_i \neq \emptyset$, and for any $j \notin c$, $\cap_{i \in c} S_i \cap S_j = \emptyset$, i.e., the subset c of sources is consistent and is maximal with this property.

Example 1 Let us suppose we have a set of sources $\mathcal{S} = \{S_1, ..., S_4\}$ as shown on Fig. 1. As we can see, $\{1, 2\}$ is a coherent subset, but not a maximal coherent subset, as it is possible to add S_3 and have a non-empty intersection. $\{1, 2, 3\}$ is a maximal coherent subset, as S_4 is contradicting S_3 (empty intersection).

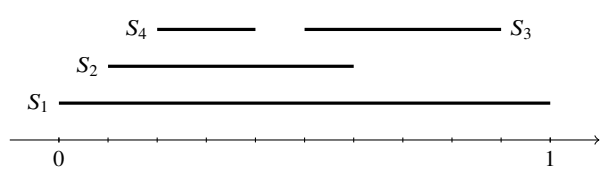

Fig. 1 Visualization of not fully consistent sources $S_i \in [0, ..., 1]$.

MCSs of polytopes: When considering sources of information in the d-dimensional Euclidean space R^d, polytopes are a quite versatile tool to model set-valued information. Such shapes can either be defined through their vertices or

extreme points (V-representation) or through a system of linear constraints defining intersections of half-planes (H-representation).

While finding whether two polytopes given by their V-representation intersect is a NP-hard problem (Tiwary (2008)), the same problem can be solved easily in H-representation through linear programming. Switching between representations is a NP-complete problem [1], with some efficient algorithms for specific H-polytopes (Khachiyan et al. (2009)).

Finding a single MCS in H-representation is thus easy: we add polytopes one by one, and we have a MCS when it is not possible to add another polytope without having an empty intersection. Checking if a set is a MCS is also easy: we check that the intersection of the corresponding H-polytopes is not empty and maximal. However, listing all the MCSs of a set of polytopes $\mathcal{P} = \{P_1, ..., P_N\}$ requires in the worst case to consider all the subsets of $\mathcal{P}$, thus at most 2^N sets for which we need to check if the intersection is not empty. When the number of polytopes is important, it becomes impossible to list all the MCSs. In Section 3, we propose an efficient algorithm to list all the MCSs through an approximation of the polytopes with minimum bounding axis-aligned hyperrectangles.

3 Enumerating the MCSs of axis-aligned hyperrectangles

Enumerating all the MCS of a set of polytopes is a difficult problem in the general case. However, polynomial algorithms (Dubois et al. (2000)) exist in the case of intervals. We will show in this section that such results can also be used in the case where we consider a set $\mathcal{H}$ of axis-aligned hyperrectangles, in order to efficiently determine the set of MCS $C_{\mathcal{H}}$.

We denote by $I^d_{H_i} \in \mathbb{R}$ the projection of H_i onto the dth dimension of the space $\mathbb{R}^D$. We have an important equivalence between the intersection of hyperrectangles and the intersection of their projections:

Proposition 1 *Given a set $\mathcal{H} = \{H_1, ..., H_N\}$ of axis-aligned hyperrectangles in the space $\mathbb{R}^D$, and their projections $I^d_{H_i} \in \mathbb{R}$ onto the different dimensions $d \in \{1, ..., D\}$, we have:*

$$\bigcap_{H_i \in \mathcal{H}} H_i \neq \emptyset \Longleftrightarrow \bigcap_{\mathcal{H}} I^d_{H_i} \neq \emptyset \; \forall d \in \{1, ..., D\}. \tag{1}$$

Proof To see the equivalence, it is sufficient to observe that $\times_{d=1}^{D} \cap_{i=1}^{N} I^d_{H_i} = \bigcap_{H_i \in \mathcal{H}} H_i$. This means in particular that any point $x \in \mathbb{R}^D$ such that its projection $x^d \in \cap_{i=1}^{N} I^d_{H_i}$ for all $d \in \{1, \ldots, D\}$ will also be in $\bigcap_{H_i \in \mathcal{H}} H_i$. Note that this is only true for axis-aligned hyperrectangles. □

The following corollary, which is merely the negation of Proposition 1, will be useful in further demonstrations.

[1] Otherwise the two problems would have the same complexity.

Corollary 1

$$\bigcap_{\mathcal{H}} H_i = \emptyset \Longleftrightarrow \exists d \in \{1, ..., D\} \; s.t. \; \bigcap_{\mathcal{H}} I^d_{H_i} = \emptyset. \tag{2}$$

In the next proof, we show that the MCS of axis-aligned hyperrectangles can be found exactly by combining the MCS of their projections, which we recall can be found in polynomial time.

Proposition 2 *Given a set $\mathcal{H} = \{H_1, ..., H_N\}$ of axis-aligned hyperrectangles in the space $\mathbb{R}^D$, its set of MCSs $C_{\mathcal{H}}$ and the sets of MCSs C_d on their projection on the d-dimension, we have:*

$$C_{\mathcal{H}} = \left\{ \cap_{d=1}^{D} c_d \mid c_d \in C_d \; \forall d \in \{1, ..., D\}, \cap_{d=1}^{D} c_d \neq \emptyset \right\}. \tag{3}$$

Proof We proceed by showing a double inclusion for a given MCS $c_h \in C_{\mathcal{H}}$:

- Let us first show that there exists $c_i \in \left\{ \cap_{d=1}^{D} c_d \mid c_d \in C_d \right\}$ s.t. $c_h \subseteq c_i$,i.e. c_i is an outer approximation of c_h. Because $\cap_{i \in c_h} H_i \neq \emptyset$ (it being a MCS), Proposition 1 tells us that for any dimension d, $\cap_{i \in c_h} I^d_{H_i} \neq \emptyset$, meaning that there will be a MCS $c_d \in C_d$ such that $c_h \subseteq c_d$. Since this is true for all $d \in \{1, \ldots, D\}$, this means that $c_h \subseteq \cap_{d=1}^{D} c_d$ for some collection of c_d, showing the inclusion.
- To show the other inclusion, we consider a set $c_d, d \in \{1, \ldots, D\}$ of MCSs on dimension d which outer-approximate $c_h \subseteq c_d$, that we know exists from the first part of the proof. We will then demonstrate that $j \notin c_h$ implies $j \notin \cap_{d=1}^{D} c_d$, therefore $\cap_{d=1}^{D} c_d \subseteq c_h$. To see this, simply consider the set of hyperrectangles $H_k, k \in c_h \cup \{j\}$, then by Corollary 1, there will be a dimension d such that $\cap_{k \in c_h \cup \{j\}} I^d_{H_k} = \emptyset$, yet $\cap_{k \in c_h} I^d_{H_k} \neq \emptyset$ (c_h being a MCS). This shows that $j \notin c_h$ implies $j \notin \cap_{d=1}^{D} c_d$.

Proposition 2 provides us with an easy approach to get MCS: we start by projecting the hyperrectangles onto the different dimensions, in order to obtain intervals. Then we enumerate the MCS on each dimension, which is polynomial. Lastly we determine $C_{\mathcal{H}}$ as the set of common sources among all the enumerations, i.e., the conjunctive combination (i.e., intersection) of the different sources such that for each element c_h of $C_{\mathcal{H}}$. By Proposition 2 and Equation (3), this gives us exactly the set of MCS.

4 Illustration on linear estimation

In this section, we illustrate our method on small imprecise linear regression problems with only two dimensions. Note that the purpose of this section is purely illustrative, so as to show the potential usefulness of our result when performing estimation from a logical, set-theoretic standpoint.

Given a data set $\{y_i, x_i\}_{i=1}^{N}$ with a single input and a single output, a linear model assumes that the relationship between the response variable y_i and the input variable

x_i is linear:

$$y_i = \beta_0 + \beta_1 x_i \tag{4}$$

Usually, the y_i are considered to be observed with a (normal) noise ϵ, and a statistical regression is performed. In our case, we will adopt a more logical, version space point of view (Mitchell (1982)): we assume that data is set-valued, and will consider the linear models consistent with it.

4.1 Estimating possible linear models with MCS and rectangles

In our setting, we assume that we observe imprecise data points $R_i = ([\underline{y_i}, \overline{y_i}], [\underline{x_i}, \overline{x_i}])$. Given the Cartesian equation of a line $\mathrm{L} = \{(x, y) \mid ax + by = c\}$ and two imprecise points R_i and R_j, finding all the lines that intersect both rectangles is formalized as:

$$\mathcal{L}_{ij} = \left\{(a, b) \mid (L \cap R_i \neq \emptyset) \wedge (L \cap R_j \neq \emptyset)\right\} \tag{5}$$

n points can be intersected by a single line if and only if all the corresponding $\mathcal{L}_{ij}$ ($\binom{n}{2}$) in total) have a non-empty intersection, i.e., their indices belong to a single coherent subset. It is maximal if no other points can be intersected by the same line. Listing all the MCSs is hard, as mentioned in Subsection 2, but our results tell us that if we approximate the different $\mathcal{L}_{ij}$ with minimum bounding axis-aligned rectangles, minimal outer approximations of the $\mathcal{L}_{ij}$ with axis-aligned rectangles, then enumerating their MCSs is easily done as shown in Section 3.

Determining the minimum axis-aligned bounding rectangle $H_{ij} = [\underline{a}_i, \overline{a}_i] \times [\underline{b}_i, \overline{b}_i]$ of $\mathcal{L}_{ij}$, i.e., the minimal volume that is fully enclosing $\mathcal{L}_{ij}$, is equivalent to finding the minimum and maximum values of the parameters a and b. It can be done quite easily, as maximizing (respectively minimizing) a is equivalent to minimizing (respectively maximizing) b.

4.2 Application

We first start with the case where the model is indeed linear, but where one data point is an outlier (box 2), as pictured on Fig. 2. As we can see, the statistical regression model, due to the outlier, does not capture the true model. In contrast, Fig. 3 shows the 5 different MCS (and their intersection) obtained using axis-aligned rectangle approximations. We can see that the biggest MCS is c_3, and it includes the true parameters, while the other MCSs are smaller and point out possible model outliers (i.e., observation 2 is in all the remaining MCSs, but not in the biggest one). Note that such an approach is quite different from standard imprecise regressions, that still uses least-square inspired approaches (Ferraro et al. (2010)).

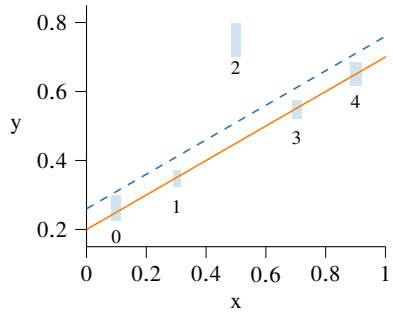

Fig. 2 Example of an imprecise linear regression problem. The dotted line corresponds to the linear regression including the outlier. The continuous line corresponds to the ground truth.

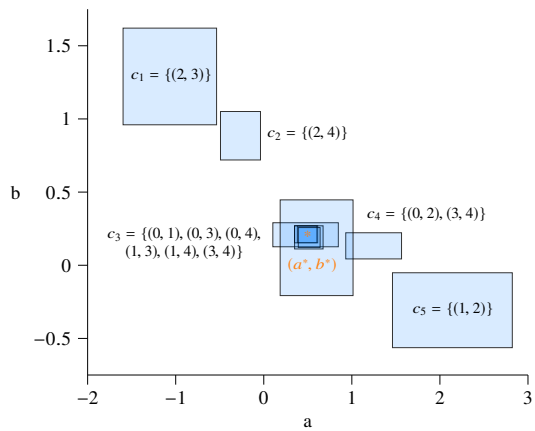

Fig. 3 Approximation with rectangles of the different $\mathcal{L}_{ij}$ from the example shown on Fig. 2

The second illustration considers the case where the observations are not faulty (the model goes through all observations), but where the model assumption is wrong, as we have a piecewise linear regression as shown on Fig. 4, but not a linear one. We have two partitions over x: $[0, 0.5]$ and $[0.5, 1]$. The two ground truths are very different, and a single statistical linear regression (in dotted blue) over the whole domain of x fits poorly to the data.

Fig. 5 shows what happens when we consider a MCS approach in this case. Instead of having one large MCS and quite smaller ones, we have now two quite large MCSs c_1 and c_2, along with other isolated ones. Those two distinct MCSs indicates that the error is likely in the model, that may only be locally valid. It also shows that it is not clear whether observation 3 belongs to one line or the other, being in both MCS. This ambiguity also cause c_1 to not include the true parameter (a_2^*, b_2^*).

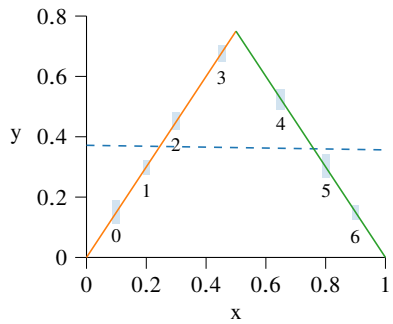

Fig. 4 Example of an imprecise piecewise linear regression problem. The dotted line corresponds to the full linear regression. The continuous lines correspond to the piecewise ground truths.

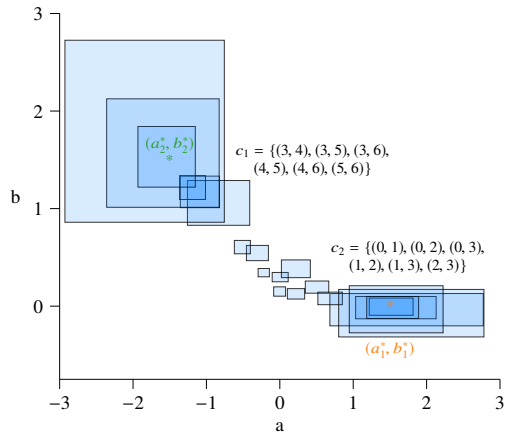

Fig. 5 Approximation with rectangles of the different $\mathcal{L}_{ij}$ from the example shown on Fig. 4

5 Conclusion

MCS is a theoretically very interesting notion for handling conflicting observations or pieces of information, but hard to enumerate in practice. We showed that such an enumeration was easier with hyperrectangles, as we can use known polynomial algorithms on their interval-valued projection to perform it. We illustrated their possible use on estimation problems, showing that their different behaviors could provide useful information (erroneous observations vs erroneous models).

In the future, we plan to perform some comparisons of such estimation approaches with Bayesian approaches in case of model misspecification, similarly to previous works intending to solve inverse problems (Shinde et al. (2021)). Another case where our results may be useful is in the repair of inconsistent preference information (Adam and Destercke (2021)).

Acknowledgements

This work was done within the PreServe Project, funded by grant ANR-18-CE23-0008 of the national research agency (ANR).

References

Adam L, Destercke S (2021) Possibilistic preference elicitation by minimax regret. In: Uncertainty in Artificial Intelligence, PMLR, pp 718–727

Destercke S, Dubois D, Chojnacki E (2008) Possibilistic information fusion using maximal coherent subsets. IEEE Transactions on Fuzzy Systems 17(1):79–92

Dubois D, Prade H, Yager R (1999) Merging fuzzy information. In: Fuzzy sets in approximate reasoning and information systems, Springer, pp 335–401

Dubois D, Fargier H, Prade H (2000) Multiple-sources informations fusion-a pratical inconsistency-tolerant approach. In: 8th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU 2000), Consejo Superior de Investigaciones Científicas

Dubois D, Liu W, Ma J, Prade H (2016) The basic principles of uncertain information fusion. an organised review of merging rules in different representation frameworks. Information Fusion 32:12–39

Ferraro MB, Coppi R, Rodríguez GG, Colubi A (2010) A linear regression model for imprecise response. International Journal of Approximate Reasoning 51(7):759–770

Khachiyan L, Boros E, Borys K, Gurvich V, Elbassioni K (2009) Generating all vertices of a polyhedron is hard. In: Twentieth Anniversary Volume:, Springer, pp 1–17

Manor R, Rescher N (1970) On inference from inconsistent premises. Theory and Decision 1:179–219

Mitchell TM (1982) Generalization as search. Artificial intelligence 18(2):203–226

Shinde K, Feissel P, Destercke S (2021) Dealing with inconsistent measurements in inverse problems: set-based approach. International Journal for Uncertainty Quantification 11(3)

Tiwary HR (2008) On the hardness of computing intersection, union and minkowski sum of polytopes. Discrete & Computational Geometry 40(3):469–479