Objective Bayesian nets for integrating consistent datasets (Extended Abstract)

Jürgen Landes[1] , **Jon Williamson**[2]
[1]Department of Philosophy “Piero Martinetti”, University of Milan
[2]Department of Philosophy and Centre for Reasoning, University of Kent
juergen_landes@yahoo.de, j.williamson@kent.ac.uk

Abstract

This paper addresses a data integration problem: given several mutually consistent datasets each of which measures a subset of the variables of interest, how can one construct a probabilistic model that fits the data and gives reasonable answers to questions which are under-determined by the data? Here we show how to obtain a Bayesian network model which represents the unique probability function that agrees with the probability distributions measured by the datasets and otherwise has maximum entropy. We provide a general algorithm, OBN-cDS, which offers substantial efficiency savings over the standard brute-force approach to determining the maximum entropy probability function. Furthermore, we develop modifications to the general algorithm which enable further efficiency savings but which are only applicable in particular situations. We show that there are circumstances in which one can obtain the model (i) directly from the data; (ii) by solving algebraic problems; and (iii) by solving relatively simple independent optimisation problems.

1 Introduction

It is increasingly common to collect multiple datasets, involving hundreds of variables and thousands of observations, to address a single problem. Different datasets tend to measure different variables, even when the datasets are collected with the same application in mind. For instance, it is common in systems pharmacology—and indeed in systems medicine more generally—to have datasets measuring proteomics, transcriptomics, metabolomics, clinical data, and patient-reported outcomes, and for these datasets to have very few variables in common; see, e.g., (De Pretis, Landes, and Peden 2021; Tricco et al. 2016). How do we integrate all this data?

One approach to data integration is motivated by Objective Bayesian Epistemology (OBE), which holds that a rational agent ought to adopt as a representation of her degrees of belief the probability function with maximum entropy, $P^\dagger$, from all those probability functions that fit her evidence (Jaynes 2003; Williamson 2010; Landes and Williamson 2013; Landes and Williamson 2016). The entropy of a probability function is a measure of the extent to which it equivocates between possible outcomes, and this approach is usually justified on the grounds that $P^\dagger$ is the function that fits the evidence but is maximally non-committal or equivocal in other respects.

In this paper, we apply OBE to the situation in which the agent’s body of evidence consists of a collection of datasets (and nothing else). We take all variables to be discrete and we assume that the datasets have been gathered in such a way that, when a variable occurs in more than one dataset, it is genuinely the same variable, with the same number of values, measured in the same way, in each dataset in which it occurs. Furthermore, we assume that the datasets are large and reliable enough that each dataset distribution provides an accurate estimate of the frequency distribution of the measured variables, and that they are consistent in the sense that these marginal frequency distributions are satisfiable by some joint probability function defined on the set V of all the variables measured by the datasets.

The agent’s belief function $P^\dagger$ is defined on the algebra generated by this larger set of variables. OBE holds that $P^\dagger$ should agree with each marginal distribution of measured frequencies, and should otherwise have maximum entropy.

In general, finding the function in a convex set of probability functions which has maximum entropy is a computationally hard optimisation problem (Paris 1994, Chapter 10). Indeed, this has been viewed as a criticism of the maximum entropy approach (Pearl 1988, p. 463). In this paper, we employ Bayesian networks to reduce the dimension of the problem in realistic cases, and thereby reduce its complexity. A Bayesian net representation of the probability function $P^\dagger$ which is motivated by OBE is called an *Objective Bayesian Net* or OBN (Williamson 2005). In this paper we develop a general algorithm, OBN-cDS, which generates an OBN and does so efficiently in realistic cases, and we show that this algorithm is preferable to a brute-force approach to entropy maximisation. Furthermore, we explore particular situations in which one can generate an OBN even faster than is possible even by OBN-cDS.

This adds to the state-of-the-art in the following ways: (a) it develops the OBE approach to data integration, (b) it shows how Bayesian net algorithms can be used to determine a maximum entropy probability function more efficiently, (c) it explores algebraic means to solve the maximum entropy optimisation problem and (d) it provides philosophical underpinnings for a particular kind of ‘statistical matching’ technique.

2 Main Findings

Algorithm 1 Pseudo Code of OBN-cDS

```
Input: Consistent datasets $DS_1, \dots, DS_h$
Output: Objective Bayesian Net with DAG $\mathcal{H}$ and conditional probabilities as determined in Step 5.
1: For all $i$ learn a Markov network structure $\mathcal{G}_i$ from $DS_i$ representing independences of $P_i^*$.
2: Set overarching undirected graph $\mathcal{G}$ as the union of the $\mathcal{G}_i$.
3: Compute a minimal triangulation $\mathcal{G}^T$ of $\mathcal{G}$.
4: Orientate $\mathcal{G}^T$ to give DAG $\mathcal{H}$.
5: For each vertex in $\mathcal{H}$, determine its probability distribution conditional on its parents:
   a) For all vertices for which there exists a dataset which measures this vertex and all its parents.
   b) For all other vertices determine conditional probabilities by solving the optimisation problem
```

Proposition 1 (Correctness of OBN-cDS)**.** *If each dataset distribution P_i^* satisfies the conditional probabilistic independence relationships represented by $\mathcal{G}_i$, for $i = 1, \dots, h$, then OBN-cDS outputs an OBN that represents the probability function $P^\dagger$, from all those that agree with all measured marginal probability distributions P_k^*, that has maximal entropy.*

Proposition 2 (Computational Complexity of OBN-cDS)**.** *As long as the maximal degree of a vertex is bounded by a polynomial, the complexity of learning the initial Markov net structures is polynomial (Step 1). Steps 2–4 can be computed in P-Time in terms of* input graphs*. Every application of Step 5a runs in linear time, if arities of variables and the number of parents is bounded by some constant. Step 5b requires the solution of an optimisation problem, which has fewer variables than the brute-force maximum entropy optimisation problem. For these reasons, the computational complexity of OBN-cDS compares favourably with that of the brute-force method.*

See Figure 1 for a realistic problem instance. Run times of our Matlab implementation are reported in the published paper (Landes and Williamson 2022).

If there are only two datasets, $h = 2$, then we do not need to optimise:

Proposition 3 (2 Datasets)**.** *In the case of two consistent datasets, OBN-2cDS finds an OBN without underdetermined variables, where OBN-2cDS is a slight modification of OBN-cDS.*

Proposition 4 (Independent Optimisation Problems)**.** *If all under-determined variables of $\mathcal{H}$ are leaves, then the problem of computing an OBN can be reduced to independently computing the conditional probabilities at these leaves.*

3 Future Work

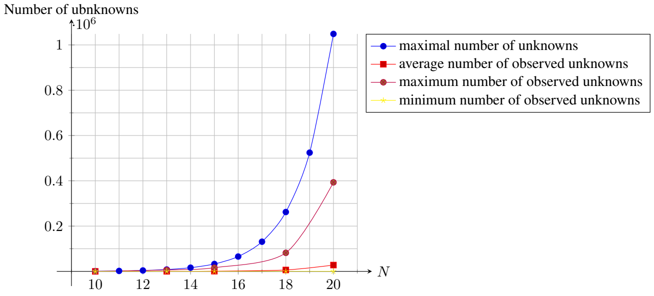

Figure 1: The number of unknowns needing to be determined by optimisation during Step 5b of OBN-cDS for 3 datasets. N is the total number of variables. For $N = 10, 13, 15, 18, 20$ we created 3 datasets 200 times. Plots are provided for the average observed number of unknowns in red, the maximum in purple and the minimum in yellow.

A number avenues for further research stand out to us: (i) to identify further efficiency savings to OBN-cDS; (ii) to extend the methodology to include inconsistent datasets; and (iii) to apply the Matlab implementation to real-world datasets.

References

De Pretis, F.; Landes, J.; and Peden, W. J. 2021. Artificial Intelligence Methods For a Bayesian Epistemology-Powered Evidence Evaluation. *Journal of Evaluation in Clinical Practice* 27(3):504–512.

Jaynes, E. T. 2003. *Probability theory: the logic of science*. Cambridge: Cambridge University Press.

Landes, J., and Williamson, J. 2013. Objective Bayesianism and the maximum entropy principle. *Entropy* 15(9):3528–3591.

Landes, J., and Williamson, J. 2016. Objective Bayesian nets from consistent datasets. In Giffin, A., and Knuth, K. H., eds., *Proceedings of MaxEnt*, volume 1757, 020007–1 – 020007–8. AIP.

Landes, J., and Williamson, J. 2022. Objective Bayesian Nets for Integrating Consistent Datasets. *Journal of Artificial Intelligence Research* 74:393–458.

Paris, J. B. 1994. *The Uncertain Reasoner's Companion*. Cambridge: Cambridge University Press.

Pearl, J. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo CA: Morgan Kaufmann.

Tricco, A. C.; Antony, J.; Soobiah, C.; Kastner, M.; MacDonald, H.; Cogo, E.; Lillie, E.; Tran, J.; and Straus, S. E. 2016. Knowledge synthesis methods for integrating qualitative and quantitative data: a scoping review reveals poor operationalization of the methodological steps. *Journal of Clinical Epidemiology* 73:29–35.

Williamson, J. 2005. Objective Bayesian nets. In Artemov, S.; Barringer, H.; d'Avila Garcez, A. S.; Lamb, L. C.; and Woods, J., eds., *We Will Show Them! Essays in Honour of Dov Gabbay*, volume 2. London: College Publications. 713–730.

Williamson, J. 2010. *In defence of objective Bayesianism*. Oxford: Oxford University Press.