Alma Mater Studiorum - Università di Bologna

EndoBayes

Project for the course Fundamentals of Artificial Intelligence and Knowledge Representation (module 3)

Martina Rossini
martina.rossini3@studio.unibo.it

October 2021

Contents

1 Introduction

Endometriosis is a condition where tissue similar to the lining of the uterus is found elsewhere in the body [4]. This causes a chronic inflammatory reaction that may result in the formation of scar tissue within the pelvis and other parts of a woman's anatomy. According to World Health Organization (WHO), about 10% of women of reproductive age are affected by endometriosis, which means currently around 190 million people [7].

However, the general public still has got limited awareness of this condition, which is also, due to its variable and broad symptoms, very much under-diagnosed. In many cases, patients are only able to get an actual diagnosis with a delay that ranges from four to eleven years [3]. Currently, the standard diagnosis is based on visualisation and histological examination of the lesions, which is an invasive procedure that implies laparoscopic surgery.

The aim of this project is to develop a Bayesian Network with specific focus on the signs and symptoms most associated to endometriosis, which should have the double aim of both predicting an eventual diagnosis and helping to spread awareness about this condition. Similar approaches have been widely explored in the scientific literature for other pathologies, like cancer, but rarely for endometriosis [3].

2 Background Information and Dataset

Endometriosis is a complex disease which can greatly decrease an individual's quality of life and whose exact origins/causes are, at present, still not known, though many hypothesis have been formulated [7] - including diet, age and ethnicity. However, as stated in [3], the study of risk factors associated to this condition is the focus of many - often contradictory - studies. For the purpose of this project only two of them were considered, namely another case of endometriosis among family members and an history of invasive pelvic surgeries, both presented in [2] and, for the moment, not contradicted by any other study.

As already said, symptoms associated with endometriosis usually vary quite a lot and some of the most common are: painful periods, pain during

and/or after sexual intercourse, and IBS-like symptoms such as abdominal cramps, nausea, diarrhea, painful bowel movements and bloating [1]. IBS (Irritable Bowel Syndrome) is a common chronic disorder that affects the large intestine and whose symptoms often overlap with those produced by endometriosis - something which makes it harder for doctors trying to discriminate between the two. Moreover, many other ailments have been recognized as endometriosis' comorbidities, which means that patients suffering from endometriosis have an higher probability than other people of having them: according to [5] and [6], some widely recognized comorbidities are IBS, infertility, chronic fatigue, depression / anxiety and ovarian cysts.

The dataset that will be used throughout the project was independently acquired by Shani Cohen, a computer science student at Ariel University (Israel). She collected information regarding the occourrence of 50 common endometriosis symptoms/comorbidities in both healthy people and in individuals affected by the interested condition using a survey. The data was originally intended to be used for a deep learning project aimed at predicting this disease [1] but she agreed to let me access and use it, as open data regarding endometriosis is extremely difficult to find online.

2.1 Data Preprocessing

Shani's dataset contains information about endometriosis' symptoms / comorbidities occurrence in each patient who took part in the survey. Therefore, all the variables only have two values, 0 if the patient doesn't have the specified condition and 1 otherwise. The data pre-processing mainly consists in mergining two variables into a single one and in dropping columns that are not considered relevant for this application. In particular we:

- Merged the columns `Menstrual pain (Dysmenorrhea)` and `Painful cramps during period` into a single column called `Painful Periods`
- Merged the columns `Depression` and `Anxiety` into a single columns called `Depression / Anxiety`
- Deleted all the columns which are not shown in Table 1

[1] You can find her work here

Pain During Intercourse	Abdominal Cramps	Painful Bowel Movements	Nausea	Infertility	Diarrhea	IBS
1	1	1	1	0	1	1
1	1	1	1	0	1	0
0	0	0	1	0	1	0
0	1	0	0	0	0	0

Chronic Fatigue	Ovarian Cysts	IBS-like Symptoms	Bloating	Endometriosis	Painful Periods	Depression/ Anxiety
1	0	1	1	1	1	1
0	1	0	0	1	1	0
1	0	1	1	1	1	1
1	0	0	1	1	0	1

Table 1: Example of the final dataset

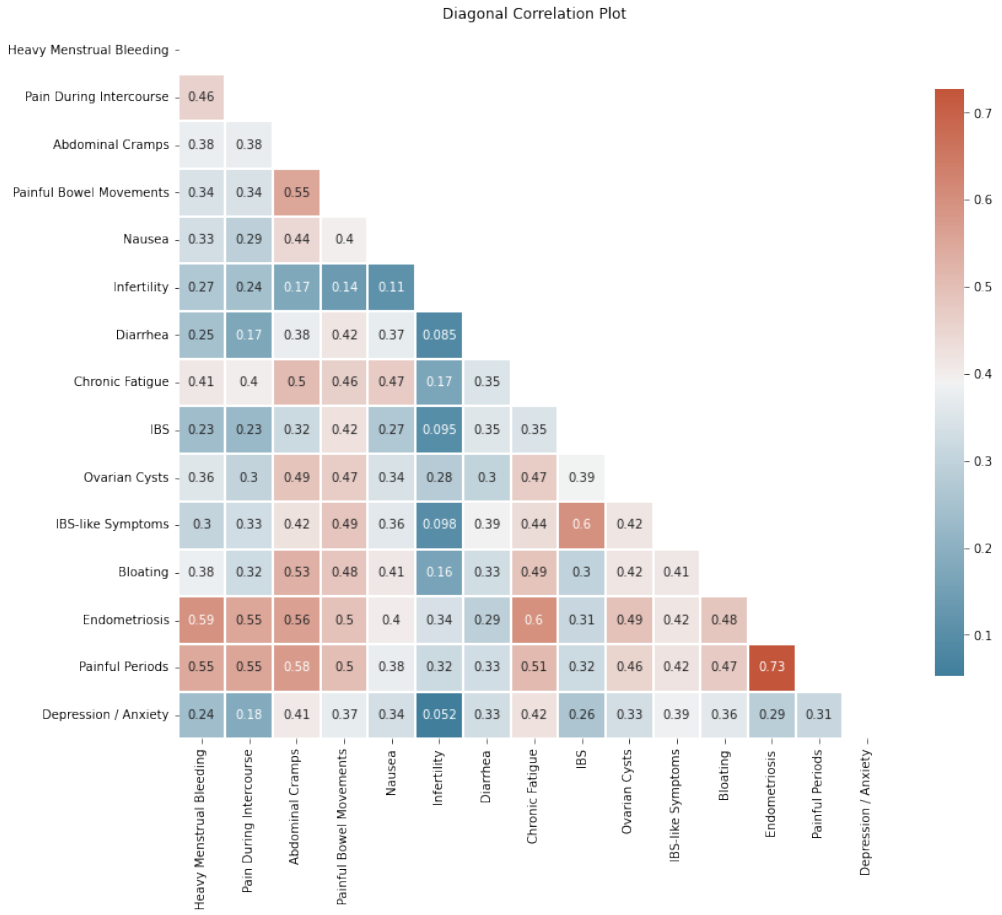

Figure 1: Correlation matrix between the variables

We then also prepared some visualizations related to the dataset, in particular Figure 1 shows the diagonal correlation matrix between the 15 columns we decided to keep.

3 Proposed Bayesian Network

In addition to the variables representing symptoms and comorbidities shown in Table 1, we also decided to include in our network the two risk factors introduced in Section 2, namely `Family History` and `Invasive Pelvic Procedure`. Note that they were not included in Shani's data and there was no way to derive their values from the other known variables, thus we decided to treat them as latent variables. However, thanks to previous literature [3, 2], we knew their probability tables beforehand and we could use them when estimating the network parameters.

3.1 The Structure

Similarly to what was proposed in [3], we used medical idioms in order to guide the structure of our bayesian model. Figure 2 shows how we can use medical reasoning patterns and the corresponding idioms in order to bridge the gap between medical knowledge and decision making, simplifying the construction of our network.

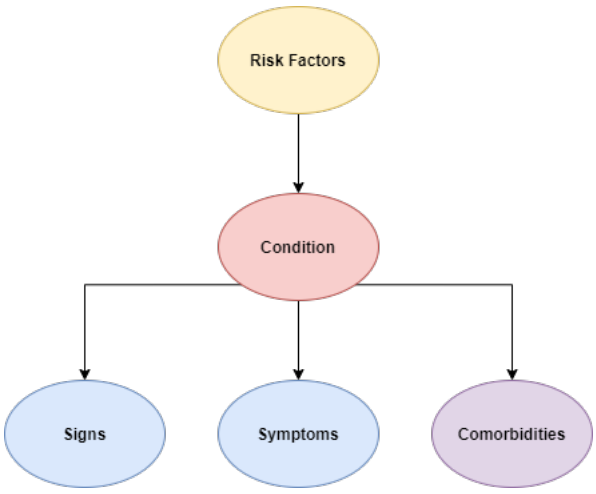

Figure 2: How medical idioms can help in building a PGM

The initial network we proposed using the help of the aforementioned medical reasoning patterns can be found in Figure 3.
However, we soon noticed the presence of some V-structures between `IBS-like Symptoms`, `IBS` and each of their five common children (`Nausea`, `Painful Bowel Movements`, `Diarrhea`, `Bloating` and `Abdominal Cramps`): this implies that we are considering `IBS` and `IBS-like Symptoms` to be mutually independent, at least until we have evidence concerning one of their children, - an assumption which doesn't really reflect our reality.

As a consequence of this observation, we slightly modified the structure of the bayesian model and obtained the architecture shown in Figure 4.

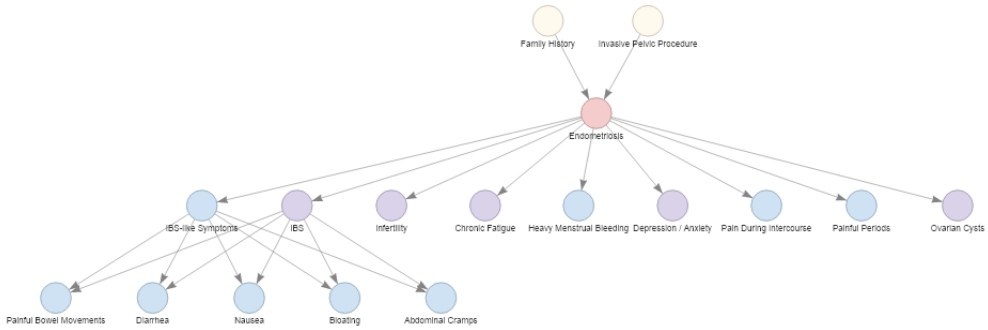

Figure 3: First proposed model

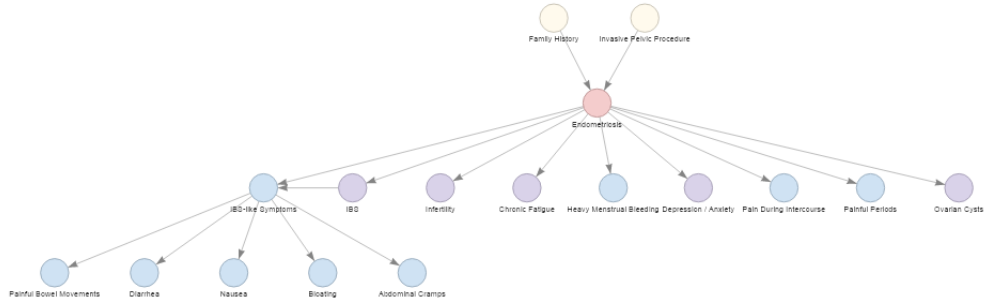

Figure 4: Final proposed model

3.2 Parameter Learning

When learning the model's parameters, we had to decide how to deal with the latent variables: traditionally, their CPT can be estimated from the available data using the Expectation Maximization algorithm, however, in our case, we were able to get some estimates from the literature. This meant that some of our network parameters had to be learned from data, while the others could be manually inserted: we explored three different approaches in order to correctly combine these two techniques, which are explained in detail in the following Sections.

3.2.1 Using fake data

The first thing we tried was to generate fictitious columns in our dataset for the missing variables, using $P(Family\ History \mid Endometriosis)$ and $P(Invasive\ Pelvic\ Procedure \mid Endometriosis)$ which we got from the literature, thus creating a sort of augmented dataset.

We then estimated the model's CPT from this new dataset using a Bayesian Estimator. Note that we choose not to use a Maximum Likelihood Estimator (MLE), as it's known to sometimes overfit in case of small datasets: indeed, if we provide too few observations, the frequencies computed by MLE are not representative of the probabilities on the whole population.
The Bayesian Estimator instead makes use of a prior distribution in order not to rely only on the input data, and is thus intrinsically more robust than MLE. Among the priors available in pgmpy we choose BDeu, which generates N uniform samples for each variable and uses these samples to compute the pseudo-counts. Note that this first method requires our network not to have any latent nodes, as the Bayesian Estimator only works on models with all observed variables, thus we had to define a variation of our PGM.

3.2.2 Using Expectation Maximization

As a second approach, we simply tried to use Expectation Maximization (EM) to estimate the latent CPTs from the available data. However, as already pointed out, our two risk factors are not easy to estimate just from their associated condition and its symptoms, thus this approach often produced CPTs which were extremely different from those found in the literature. Moreover, since the process is randomized, these probability values

varied quite a lot when using a different random seed.

3.2.3 Using EM with priors

The final approach we used is based on a modified implementation of the Expectation Maximization algorithm provided by pgmpy, which can be found in `parameter_estimators.py`. We now take into consideration eventual CPTs already manually added to the network, instead of just replacing them with random ones at the start of the computation. Basically, we are making it possible for the EM algorithm to accept manual CPT priors and start from those, instead of starting from random ones every time.

This final method produces realistic results which are also quite stable no matter the random seed used in the process; moreover, it does not require any modification to the original pgmpy network. Thus, this is the preferred way to estimate the network parameters.

4 Network Analysis

After defining the network and learning its parameters, we'll now analyze its architecture in terms of Markov blankets and active trails. Regarding eventual independencies, we can immediately see that the final model presents a V-structure between `Family History`, `Invasive Pelvic Procedure` and `Endometriosis`, which implies that the two risk factors are independent one from the other - something which we found quite reasonable in our application. Finally, note that the aforementioned independency is no longer guaranteed by the V-structure when we have `Endometriosis` in our evidence.

4.1 Markov Blankets

The Markov blanket for a given node consists of the set of its parents, children and children's parents. Once we know the blanket of a given node A, it becomes independent from every other variable in the network. Below, in Figure 5, we show the Markov blaket produced for the most important node in our network - the one representing the condition of interest. We also though it interesting to show the blankets for `IBS` and `IBS-like Sympoms`, seeing as the interaction between these two variables were the reason why we

choose to change our initial network structure; they can be found respectively in Figures 6 and 7. Note that in all the plots of this Section the node for which we are showing the blanket is coloured in light orange, the variables in its blanket are in red and all the other nodes are in gray.

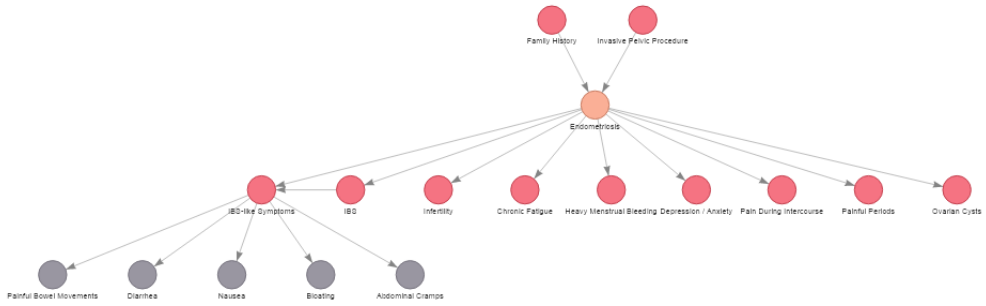

Figure 5: Markov Blanket for `Endometriosis`.

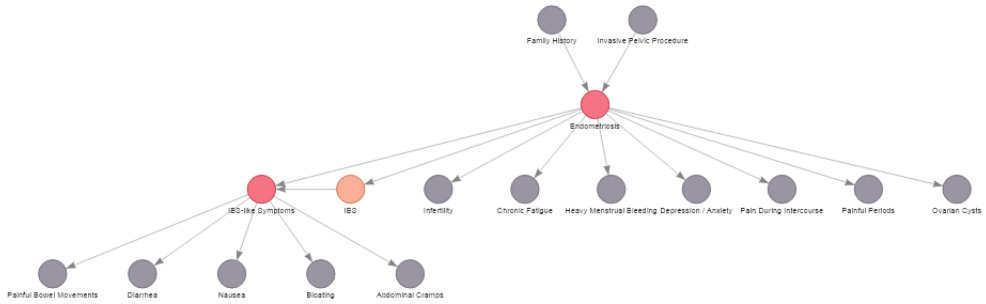

Figure 6: Markov Blanket for `IBS`.

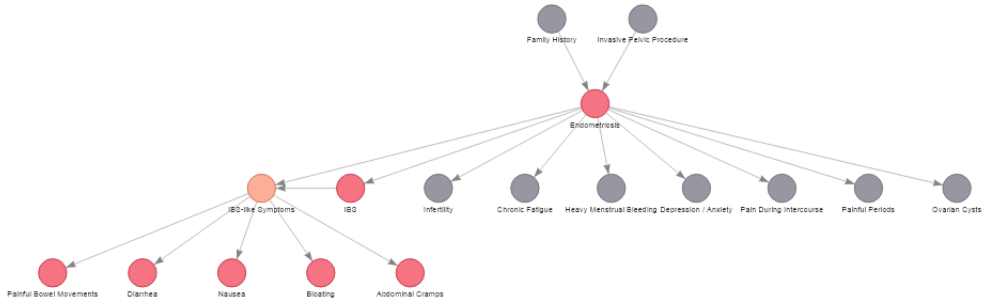

Figure 7: Markov Blanket for `IBS-like Symtoms`.

4.2 Active Traits

We know that two variables X and Z in a model are independent if there's no active trail between them. We say that a path $X_1 \to X_2 \to \cdots X_n$ is active given as evidence the variables in E if, for each consecutive triplet in the path, we have:

1. The triplet is of the form $X_{i-1} \to X_i \to X_{i+1}$ and X_i is not observed ($X_i \notin E$). Note that this must also be true when all the arrows point in the other direction and is called *head-to-tail* active trail.

2. The triplet is of the form $X_{i-1} \leftarrow X_i \to X_{i+1}$ and X_i is not observed ($X_i \notin E$). This is called *tail-to-tail* active trail.

3. The triplet is of the form $X_{i-1} \to X_i \leftarrow X_{i+1}$ and X_i or one of its descendents is observed. This is called *head-to-head* active trail.

Thus, active trails in bayesian network can be of three main types (head-to-tail, tail-to-tail and head-to-head): we'll now graphically depict them and show, for each of these structures, an example in our network.

4.2.1 Head-to-Tail

An example of this type of structure is shown in Figure 8. In our network, when we have no evidence, we can detect two active trails going from `Endometriosis` to `Nausea` (and vice versa, following an evidential reasoning pattern instead of a causal one). If we add `IBS` to our evidence however, one of the two trails gets broken: still, the information can flow

from `Endometriosis` to `Nausea` using the path `Endometriosis` → `IBS-like Symptoms` → `Nausea` as shown in Figure 9.

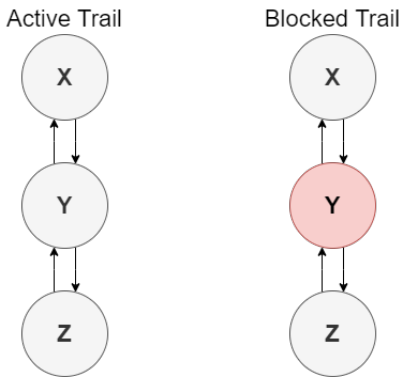

Figure 8: Head-to-Tail active path

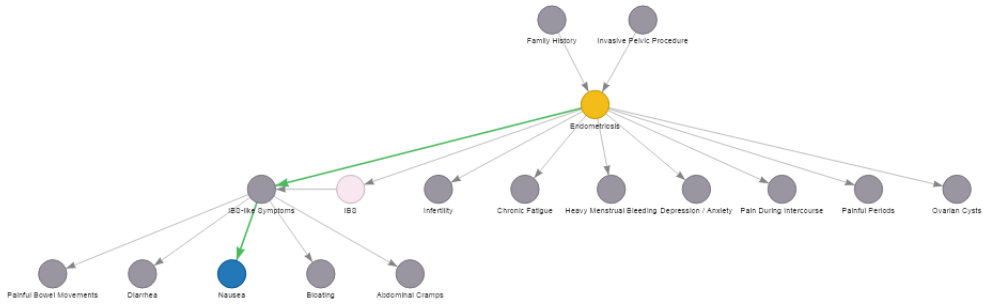

Figure 9: Active Trail not broken by `IBS` in the evidence.

To break both the active paths, effectively making `Endometriosis` and `Nausea` independent one from the other, we need to know the value of `IBS-like Symptoms`. Indeed, now both trails are broken, as we can see in Figures 10 and 11.

4.2.2 Tail-to-Tail

An example of this type of structure can be found in Figure 12. In our network, a tail-to-tail path is active between `Chronic Fatigue` and `Ovarian Cysts` without any evidence, as can be seen in Figure 13. Once we know

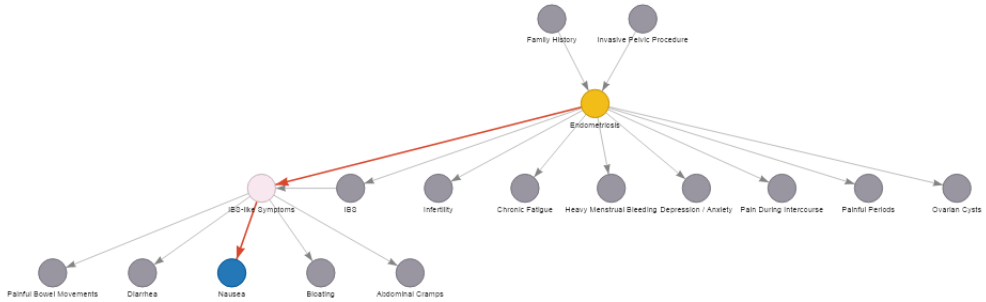

Figure 10: First active trail broken by `IBS-like Symptoms`.

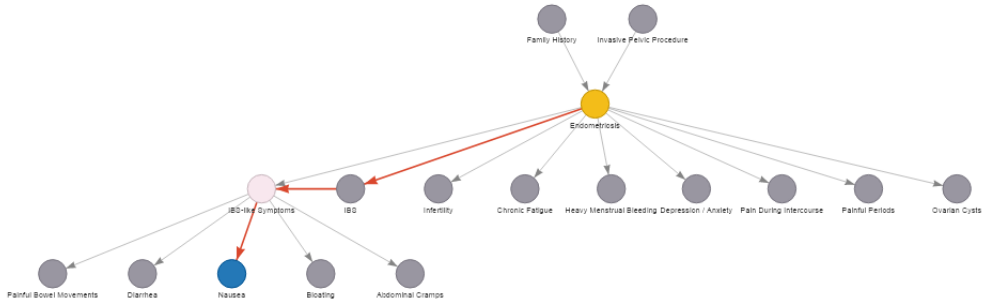

Figure 11: Second active trail broken by `IBS-like Symptoms`.

that the patient suffers from `Endometriosis` however, the path gets broken, as depicted in Figure 14.

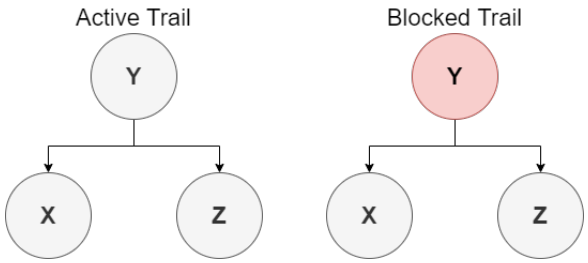

Figure 12: Tail-to-Tail active trail.

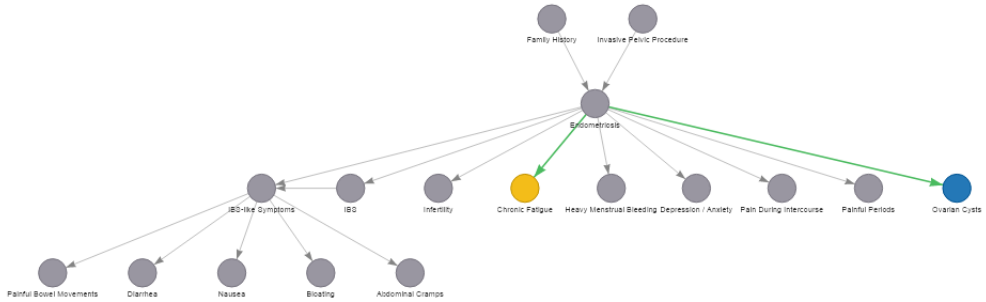

Figure 13: Active trail between `Chronic Fatigue` and `Ovarian Cysts`.

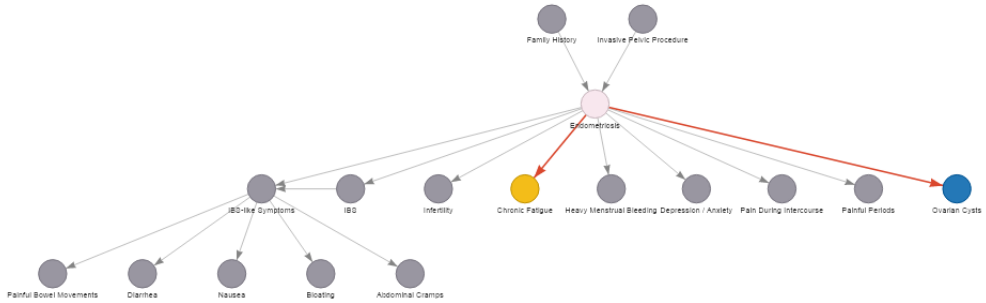

Figure 14: Broken trail between `Chronic Fatigue` and `Ovarian Cysts`.

4.2.3 Head-to-Head

Finally, an example of this V-structure can be seen in Figure 15 and, in our network, between the two risk factors and the main condition of interest. Figures 16 and 17 show how this trail is initially blocked but becomes active once we know that the patient suffers from `Endometriosis`.

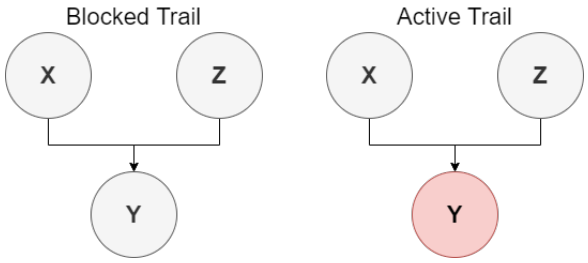

Figure 15: Head-to-Head active trail.

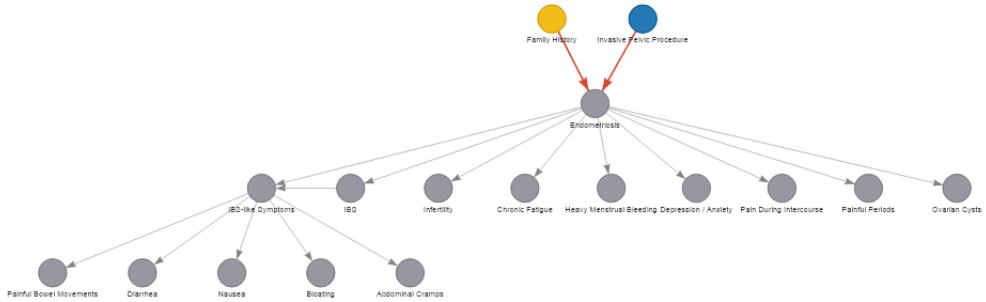

Figure 16: Broken trail between the two risk factors.

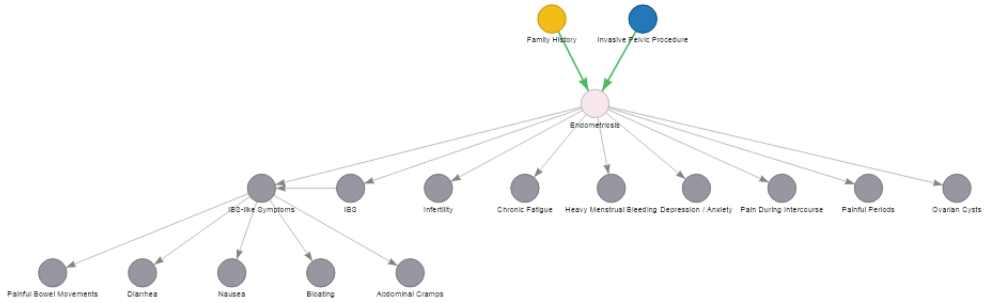

Figure 17: Active trail between the two risk factors

5 Inference

We'll now show, through two simple case studies, how this network can actually be used to help in diagnosis. In particular, the first case study will be solved using exact inference, namely the Variable Elimination algorithm included in the pgmpy library, while the second case study will be solved through Gibbs Sampling, a Markov chain Monte Carlo (MCMC) algorithm for approximate inference.

5.1 Exact

Our case study for exact inference will be the following: we need the probability that a patient has endometriosis, knowing they feel pain during intercourse, are bloated and have a family history of endometriosis, but they do not have neither IBS nor painful periods.
In the pgmpy's Variable Elimination method, the order in which variables will be eliminated depends on a heuristic function, which assigns a cost to the removal of every node apart from query and evidence nodes. The library provides 4 different heuristics:

1. *MinFill*: the cost of removing the node is equal to the number of edges that need to be added to the network after its elimination.

2. *MinNeighbors*: removing a node has a cost equal to the number of neighbors it has in the current architecture.

3. *MinWeight*: to every node is assigned a weight given by its domain cardinality; then, the cost of removing a node is equal to the product of the weights of its neighbors

4. *WeightedMinFill*: to every edge is assigned a weight given by the product of the domain cardinality of its vertices; then, the cost of removing node A is given by the sum of the weights of the edges we'll need to add to the network after its elimination.

We tried to solve our simple case study using all of the aforementioned heuristics, in order to compare them in terms of their performances. By repeating each trial 20 times and then averaging the results, we were able to determine that, for this simple network and simple case study, the Variable Elimination process takes about the same amount of time, no matter the

heuristic. Still, note that using *MinWeight* and *WeightedMinFill* we get a slightly bigger average than with the others, as shown in Table 2.

Finally, they obviously all produce the same probabilities for `Endometriosis` but, in our case, they also always result in the same elimination order, which is `[Invasive Pelvic Procedure, IBS-like Symptoms]`.

MinFill	**MinNeighbors**	**MinWeight**	**WeightedMinFill**
0.031145	0.031902	0.033308	0.032409

Table 2: Timing information for the different heuristic functions.

5.2 Approximate

For approximate inference, instead, our case study is the following: we need to know how likely it is for a patient to also have IBS when they have endometriosis and suffer from nausea, bloating and ovarian cysts.
As we already mentioned, the required probability will be approximated using Gibbs Sampling; in particular, we'll try many different sample sizes and - since the sampling process has a random component - we'll also repeat the experiment 10 times in order to average the results. The estimated mean probability for every sample size along with the 95% confidence interval for `Endometriosis` equal to 1 (i.e.: the patient suffers from it) is shown in Figure 18.

6 Conclusions and Future Work

Overall, the results we got were reasonable but the model absolutely needs to take into account many more factors (like age, ethnicity, socio-economical status, etc..) before if can really be useful in a diagnosis. Moreover, as already discussed, we only used two widely accepted risk factors and some of the most common symptoms and comorbidities but others could be added in order to improve the model's precision. Not only that, but the network could be extended to consider also the impact of endometriosis on an individual's quality of life, much like it was done in [3].

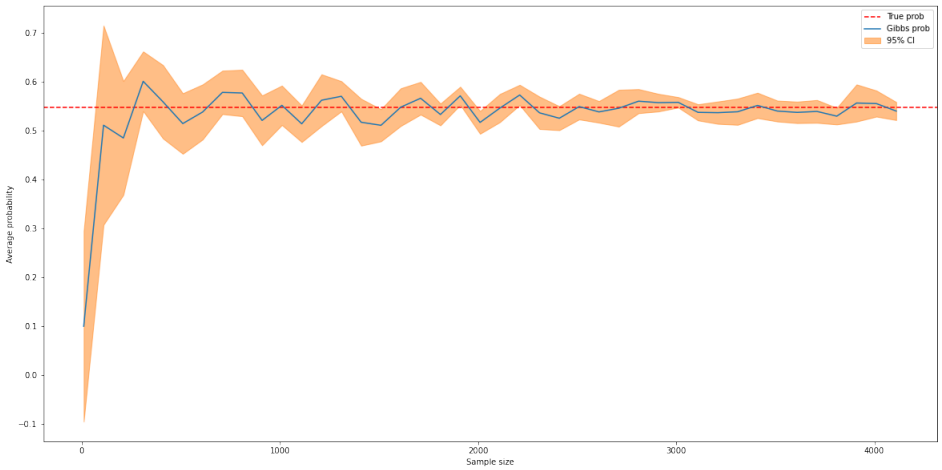

Figure 18: Gibbs Sampling output

References

[1] *About Endometriosis and IBS: Symptoms, Diagnosis, Treatment & More*. en. June 2019. URL: https://www.healthline.com/health/womens-health/endometriosis-and-ibs (visited on 10/21/2021).

[2] Mahnaz Ashrafi et al. "Evaluation of Risk Factors Associated with Endometriosis in Infertile Women". In: *International journal of fertility & sterility* 10 (Apr. 2016), pp. 11–21. DOI: 10.22074/ijfs.2016.4763.

[3] Rachel Collins and Norman Fenton. "Bayesian network modelling for early diagnosis and prediction of Endometriosis". In: (). DOI: 10.1101/2020.11.04.20225946. URL: https://doi.org/10.1101/2020.11.04.20225946.

[4] Stephen Kennedy et al. "ESHRE guideline for the diagnosis and treatment of endometriosis". In: *Human Reproduction* 20.10 (June 2005), pp. 2698–2704. ISSN: 0268-1161. DOI: 10.1093/humrep/dei135. eprint: https://academic.oup.com/humrep/article-pdf/20/10/2698/1651419/dei135.pdf. URL: https://doi.org/10.1093/humrep/dei135.

[5] Eric S. Surrey et al. "Risk of Developing Comorbidities Among Women with Endometriosis: A Retrospective Matched Cohort Study". eng. In:

Journal of Women's Health (2002) 27.9 (Sept. 2018), pp. 1114–1123. ISSN: 1931-843X. DOI: 10.1089/jwh.2017.6432.

[6] *Women with Endometriosis Have Higher Rates of Some Diseases*. en. URL: https://www.nichd.nih.gov/newsroom/releases/endometriosis (visited on 10/21/2021).

[7] World Health Organisation. *Endometriosis*. https://www.who.int/news-room/fact-sheets/detail/endometriosis, Last accessed on 2021-07-20. 2021.