Fault diagnosis in FTTH networks and services

Non-parametric probabilistic graphical models

Amine Echraibi[1,2], Sandrine Vaton[1], Joachim-Flocon Cholet[2], Stéphane Gosselin[2]

[1] *Institut Mines Télécom Atlantique*

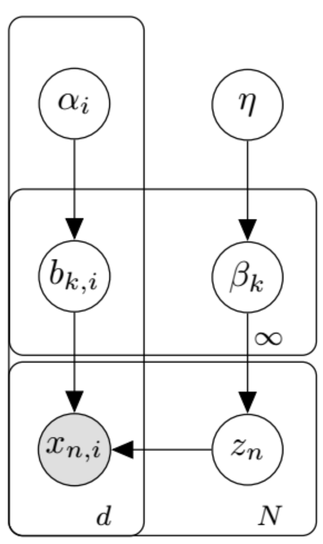
[2] *Orange Labs*

Introduction

- The goal of this research is to identify patterns of faults in telecommunication networks. The increasing size of networks and services makes data exploration by an expert a task that requires high resources (time and expert knowledge).
- This task is a clustering task, the challenge of handling network data is to identify the right model. The data acquired from networks is highly correlated, of different types (categorical variables + continuous). Mixture models seem well suited to treat this problem.
- The number of clusters or patterns to be identified is not known and may increase with time. Therefore, we consider in our research non parametric models such as mixture models with Dirichlet process priors.
- One avenue of interesting future research is to integrate expert knowledge to help the clustering using semi supervised methods.

Model and inference

1. **Notation and data:**
   - We denote random variables as $X_1, X_2, ..., X_d$ where X_i is the ith random variable (Power or alarm, ..).
   - All variables are categorical.
   - We denote x_{ni} the nth instance of variable X_i in the dataset.
   - $\mathcal{D} = (x_{ni})_{n,i}$: The complete dataset.
2. **Plate notation**:



3. **The infinite categorical mixture model:**

$$\beta_k \sim Beta(1, \eta)$$
$$\pi_k = \beta_k \prod_{l=1}^{k-1}(1 - \beta_l) \quad DP$$
$$z_n \sim Cat(.|\pi)$$
$$b_{ki} \sim Dir(.|\alpha_i) \quad prior$$
$$x_{ni}|z_n = k \sim Cat(.|b_{ki}) \quad emission$$

4. **Inference:** inference in the DPCMM is done by computing the posterior distribution:

$$p(\beta, z, b|\mathcal{D}) = \frac{p(\beta, z, b, \mathcal{D})}{p(\mathcal{D})}$$

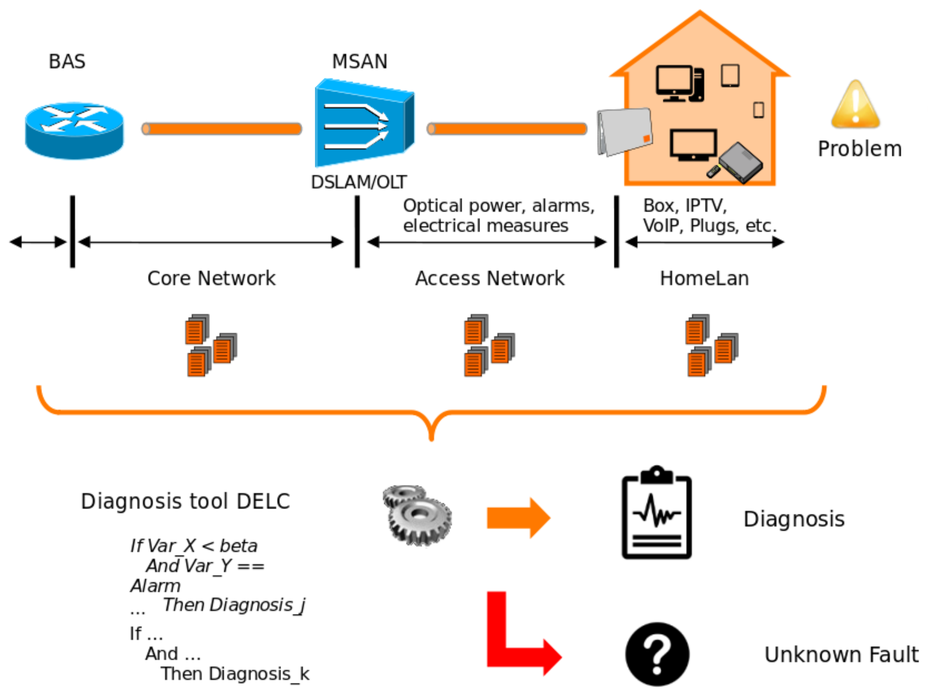
We can drop the constant term $p(\mathcal{D})$ and the objectif is to compute or estimate $p(\beta, z, b, \mathcal{D})$.

5. **Variational Inference:**

$$q^* = \arg\min_q \mathbb{KL}[q||p]$$

$$\log q_j^*(\theta_j) = const + \mathbb{E}_{\theta \setminus \{\theta_j\} \sim q^*}[\log p(\theta, \mathcal{D})]$$

Environment and data

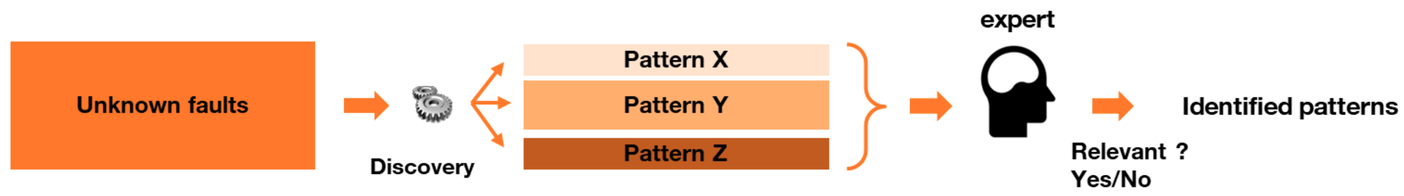


- 90k tests/day
- 3-4k variables (continuous + discrete)
- rules designed by experts
- hardware changes.
- new services.

How to identify patterns of faults from data ?

Advantages in contrast to current approaches

- Automatic identification of patterns of faults:



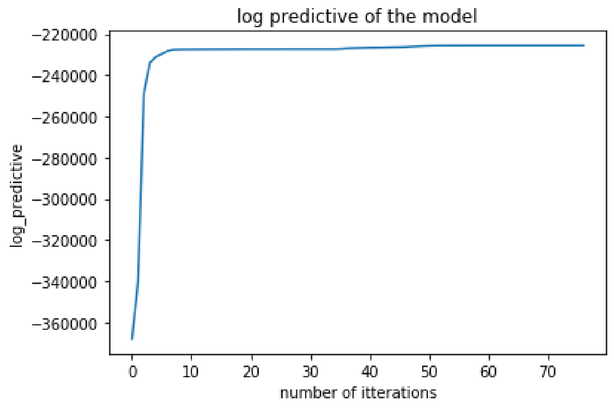
- Expert intervention : validation (no exploration of raw data).
- Correct patterns $\Longrightarrow$ new or updated expert rules.
- Proactive system: automatic detection of faults.
- Better customer experience.

Preliminary results

- **Clustering results:**

- **Convergence:**

$$\mathcal{L} = \mathbb{E}_{\theta \sim q^*}[\log p(\theta, \mathcal{D})]$$



Conclusion

This approach will permit experts to automatically identify patterns of network faults in a broad range of applications. The expert role will be to validate the results without raw data exploration.

Acknowledgements

This work is funded by Orange Labs, in the context of a PhD thesis at IMT Atlantique. I would like to thank my thesis adviser and my tutors for valuable contributions to this work.