Sparse pairwise Markov model learning for anomaly detection in heterogeneous data

Romain Laby, Alexandre Gramfort, François Roueff, Cyrille Enderli, Larroque Alain

Sparse pairwise Markov model learning for anomaly detection in heterogeneous data

Romain Laby[1,2], Alexandre Gramfort[1], François Roueff[1], Cyrille Enderli[2], and Alain Larroque[2]

[1]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, Paris, France
[2]Thales Airborne Systems

June 24, 2015

Abstract

An important challenge in the aeronautic industry is to cope with maintenance issues of the products, notably detection and localization of components breakdowns. Modern equipments enjoy better recording and processing capacities, allowing the storage of a large amount of data, on which better maintenance systems are expected to be built. Efficient probabilistic models able to represent the statistic distribution of the collected variables in the "normal state" of the system are needed in order to derive anomaly detection algorithms. Graphical models constitute a rich class of models and are natural candidates to address this task. This article proposes a method for learning undirected hybrid graphical models from heterogeneous data. The data are heterogeneous as they include physical (quantitative) measures as well as a collection of inherently discrete variables for instance describing the state of electronic devices. The model we propose is adapted from the Ising and Gaussian models so that the data don't require to be translated from their original space, allowing the user to easily interpret the dependency graph learned from data. The learning step is carried out by minimizing the negative pseudo-log-likelihood using a proximal gradient algorithm with Lasso and group Lasso penalization for addressing the high dimension of variables. Once the model is learned, we use the penalized negative pseudo-log likelihood as a test statistics for detecting anomalous events.

Introduction

Probabilistic graphical models are used to represent joint distribution over a set of N random variables $X_1, \ldots, X_N$, in an efficient and compact way. When the dimension N is small, the joint distribution can be explicitly represented, whereas for high dimension explicit representations are intractable. Probabilistic graphical models are a framework especially designed for the modeling of complex systems. This approach can be applied in many fields of application, and in particular, it can be used for anomaly detection and localization. Bayesian networks constitute a widespread class of graphical models to achieve this goal, see [1], [2], [3], [4] and the references therein. Namely, given a Bayesian network for the data, the parameters of the conditional distributions are estimated from normal data. Then the computation of the likelihood is easily performed for new records of data to decide weather a record is anomalous or not. Indeed, the lower the likelihood, the higher the probability to have an anomalous record. This method has been successfully applied for network intrusion detection [2] and in the medical fields for disease outbreak detection [3].

The application in the aeronautic field is much more recent: a study about the benefits of Bayesian networks can be found in [5]. The anomaly detection and localization problem is confronted to some industrial constraints: difficulty of data acquisition, high dimensionality of data, computation time constraints, and heterogeneity of the variables. Variables can be categorical, *e.g.* when they correspond to working states of systems or electronic components, and quantitative when they reflect physical measures such as temperatures, pressures or phases.

In the graphical model literature, if the inference in models involving both categorical and quantitative variables – what we refer thereafter as hybrid graphical models (HGM) – has attracted some studies (see [6], [7], [4] and the references therein), few works have addressed the problem of learning the graph structure in the presence of both categorical and quantitative variables. In [8], such hybrid graphs are treated by translating all variables into a common feature space using Mercer kernel. It is interesting to note that all the references above consider models based on Bayesian networks. However, if the Bayesian

network is not given, it can be extremely cumbersome to learn it from data both from a statistical and numerical point of view, especially for a large number of variables.

Here we focus on learning hybrid sparse pairwise undirected graphical models. A sparse network is preferable for mostly two reasons: first it avoids overfitting the model over the data, and secondly it makes probabilistic inference easier. For this purpose, the use of regularization has been widely studied over the last years, especially for Lasso (ℓ_1) and group Lasso (ℓ_1/ℓ_2) regularizations, see [9] for recent reviews and [10] for the use of ℓ_1/ℓ_2 in graph structure estimation. Contrary to directed graphical models, undirected models have the advantage of having a concave likelihood (in the framework of log-linear models, see [7, Section 20.2]), which guarantees the existence of a global optimum of the likelihood. This allows us to use convex optimization algorithms to find the maximum likelihood estimator, that we want to use as reference model for anomaly detection. For that purpose, we use the proximal gradient algorithm. This algorithm is an iterative scheme that is assured to find a global minimum of our objective function, under some hypothesis (see [11], Section 2). Combined with ℓ_1 and ℓ_1/ℓ_2 regularizations, the proximal operator becomes a simple soft-threshold operation. Nevertheless, the complexity of the partition function makes the calculation of the likelihood and its derivatives intractable, especially for hybrid network, since it would require summation over the categorical variables and integration over the quantitative variables, which may not have a closed-form integral. To perform an approximation this quantity, the partition function can be approximated through MCMC simulations. The impact of the stochastic step on the proximal gradient algorithm is studied in [11]. This method is detailed in the context of heterogeneous data in [12]. An alternative method consists in optimizing the pseudo-log-likelihood rather than the log-likelihood (see [13]). The optimization of the ℓ_1-penalized pseudo-log-likelihood is admittedly sub-optimal, but does not require any approximation as it can be expressed in closed-form.

In this paper, we present a hybrid model learning algorithm based on the optimization of the pseudo-log-likelihood. The model is derived from the classical Ising model and Gaussian model, thus it can deal with continuous and discrete variables. The Ising model [14] has only binary variables, whereas known Potts [15] model can have discrete non-binary variables. However the Potts model requires all variables to be equally labelled, what is not the case in our application, where variables have different labels with different meanings. We use the 1-of-K encoding scheme to represent every non binary discrete variable taking K values by a K-dimensional binary vector. This transformation requires us to use ℓ_1/ℓ_2 regularization (see [9] and [16, Section 4.3.4]).

This paper is organized as follows: in Section 1 we introduce the hybrid model, and we present some interesting properties for MCMC simulations. In Section 2 we explain the algorithm we use for structure learning. In Section 3, we present the application of our model to the problem of anomaly detection on real industrial data. The application we target is designed in a semi-supervised fashion: we learn a model from normal data set, which is assumed to contain no anomaly. We use that reference model to score new records, and label as anomalies low-scored records. We show an application of our approach on a real aeronautic industrial case of breakdown detection for Active Electronically Scanned Array (AESA) Radars.

1 Presentation of the hybrid model

The concept of graphical models relies on the factorisation of the joint distribution. The density $\mathrm{p}(x_1, \ldots, x_N)$ of N random variables can be factorized over the maximal cliques in an undirected network, according to

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \prod_{C \in \mathcal{I}} \varphi(x_C)$$

where $\mathcal{I}$ is the set of indices of variables involved in the cliques decomposition, $\varphi(x_C)$ is a clique potential, and Z is the normalizing constant (also called partition function). The joint distribution can be specified by using a log-linear model, where the clique potentials are replaced by exponential weighted sum of features, according to

$$p(x_1, \ldots, x_N) = \frac{1}{Z} \exp\left(\sum_{C \in \mathcal{I}} w_C f_C(x_C) \right)$$

where $\{f_C, C \in \mathcal{I}\}$ is the set of features and $\{w_C, C \in \mathcal{I}\}$ the set of associated weights. Several parametrization are available for the choice of the features. The class of pairwise networks has been widely studied (see [9]). In pairwise models, the features are functions of one or two variables. The density p of N random variables takes the form

$$p(x) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} \varphi_i(x_i) + \sum_{i,j \in E} \varphi_{ij}(x_i, x_j) \right), \tag{1}$$

where E is the set of pairs of variables we want to include in the model: the related graph will contain an edge between node i and node j if $\varphi_{ij} \neq 0$. The normalisation constant Z is intractable in high dimension, since its calculation requires the summation/integration over every possible instantiation of $(x_1, \ldots, x_N)$, *e.g.* if each random variable x_i is binary, then this summation has 2^N terms.

There are several choices for the potentials, each leading to several class of models. Among them, we will use the Ising Graphical Model (IGM) and Gaussian Graphical Model (GGM). The Gaussian model has continuous Gaussian random variables, and the density p takes the form of a Gaussian density. In that specific case, the partition function Z is easy to calculate and only requires the calculation of the determinant of a $N \times N$ matrix. The Ising model is one of the earliest studied undirected model for modeling energy of a physical system involving interactions between atoms (see [14]). The Ising model has binary variables, *i.e.* each x_i takes values in $\{-1, 1\}$ or $\{0, 1\}$, depending on the authors. Here we use the state space $\{0, 1\}$.

The Ising model can be generalized for discrete variables, for example with the Potts model [15], but this one can be reparametrized as an IGM using 1-of-K encoding, as explained in [16], 4.3.4 (see Section 2 for more precisions). In the case of an IGM, the density takes the form

$$p_\Theta(x) = \frac{1}{Z_\Theta} \exp\left(\sum_{i=1}^{N} \theta_{ii} x_i + 2 \sum_{i<j}^{N} \theta_{i,j} x_i x_j \right), \tag{2}$$

where $\Theta = (\theta_{i,j})$ is a parameter of $\mathbb{R}^{N(N+1)/2}$. For practical reasons we consider Θ to be a $N \times N$ symmetric matrix. Since $x_i = x_i^2$ for $x_i \in \{0, 1\}$, we can then rewrite

$$p_\Theta(x) = \frac{1}{Z_\Theta} \exp\left(x^T \Theta x\right) . \tag{3}$$

This form is fundamentally similar to a Gaussian density, but Θ has not any constraint because it is not associated to a covariance matrix. The calculation of Z_Θ is costly in high dimension, since it requires the summation over 2^N terms.

Now we present the hybrid model mixing binary variables $X_\mathcal{C} = \{X_i, i \in \mathcal{C}\}$ (called categorical thereafter), and continuous variables $X_\mathcal{Q} = \{X_u, u \in \mathcal{Q}\}$ (called quantitative thereafter). Let $X = (X_\mathcal{C}, X_\mathcal{Q})$ be the variables of our model with values in $\{0, 1\}^{|\mathcal{C}|} \times \mathbb{R}^{|\mathcal{Q}|}$. We study the pairwise hybrid model

$$p_\Omega(x) = \frac{1}{Z_\Omega} \exp\Bigg(x_\mathcal{C}^T \Theta x_\mathcal{C} + \mu^T x_\mathcal{Q} - \frac{1}{2} x_\mathcal{Q}^T \Delta x_\mathcal{Q} + x_\mathcal{C}^T \Phi x_\mathcal{Q} \Bigg), \tag{4}$$

where $\Omega = (\Theta, \mu, \Delta, \Phi)$ with $\Theta = (\theta_{ij})_{i,j \in \mathcal{C}}$ is a symmetric matrix, $\mu = (\mu_i)_{i \in \mathcal{Q}} \in \mathbb{R}^\mathcal{Q}$, $\Delta = (\delta_{uv})_{u,v \in \mathcal{Q}}$ is a symmetric matrix and $\Phi = (\phi_{iu})_{i,u \in \mathcal{C} \times \mathcal{Q}}$ is a general matrix. In order that p_Ω would be a valid density with respect to the product measure made up by the counting measure over $\{0, 1\}^\mathcal{C}$ and the Lebesgue measure over $\mathbb{R}^\mathcal{Q}$, one only requires Δ to be positive-definite, hypothesis we will make thereafter. On the other hand, no condition is imposed to Θ, μ and Φ, other than Θ symmetric.

The density (4) has interesting properties. Seen as a function of $x_\mathcal{Q}$ only, we get

$$p_\Omega(x) \propto \exp\left(\left(\mu^T + x_\mathcal{C}^T \Phi\right) x_\mathcal{Q} - \frac{1}{2} x_\mathcal{Q}^T \Delta x_\mathcal{Q} \right),$$

where $\propto$ means equality between functions up to a constant multiplier (that here depends on $x_\mathcal{C}$). We recognize a Gaussian density, we thus conclude that given $X_\mathcal{C}$, $X_\mathcal{Q}$ is a Gaussian vector with mean $\Delta^{-1}\left(\mu + \Phi^T X_\mathcal{C}\right)$ and covariance matrix Δ^{-1}.

Likewise it is easy to prove that, given $X_\mathcal{Q}$, $X_\mathcal{C}$ is an Ising model. More surprisingly, we can show that the non-conditional law of $X_\mathcal{C}$ is still an Ising model (whereas it is clearly not the case for $X_\mathcal{Q}$, where the non-conditional law of $X_\mathcal{Q}$ isn't Gaussian but is a mixture of Gaussian). Indeed, if $p_{\mathcal{C}\Omega}$ is the density of $X_\mathcal{C}$, we get

$$p_{\mathcal{C}\Omega}(x_\mathcal{C}) \propto \exp\left(x_\mathcal{C}^T \Theta x_\mathcal{C}\right) \int_{\mathbb{R}^{|\mathcal{Q}|}} \exp\left((\mu + \Phi^T x_\mathcal{C})^T x_\mathcal{Q} - \frac{1}{2} x_\mathcal{Q}^T \Delta x_\mathcal{Q} \right) dx_\mathcal{Q} .$$

We can interpret the integral term (up to a constant multiplier) as an the expectation $\mathbb{E}[\exp\left((\mu+\Phi^T x_{\mathcal{C}})^T U\right)]$ where U is a Gaussian vector with zero mean and covariance matrix Δ^{-1}. Thus we get

$$\begin{aligned} p_{\mathcal{C}\Omega}(x_{\mathcal{C}}) &\propto \exp\left(x_{\mathcal{C}}^T\Theta x_{\mathcal{C}} + \frac{1}{2}(\mu+\Phi^T x_{\mathcal{C}})^T\Delta^{-1}(\mu+\Phi^T x_{\mathcal{C}})\right) \\ &\propto \exp\left(x_{\mathcal{C}}^T(\Theta+\Phi\Delta^{-1}\Phi^T/2)x_{\mathcal{C}} + \mu^T\Delta^{-1}\Phi^T x_{\mathcal{C}}\right) \\ &\propto \exp\left(x_{\mathcal{C}}^T\ \left(\Theta+\Phi\Delta^{-1}\Phi^T/2 + \mathrm{Diag}\left(\Phi\Delta^{-1}\mu\right)\right)x_{\mathcal{C}}\right)\ , \end{aligned}$$

where we used in the last line that $x_i^2 = x_i$. Here $\mathrm{Diag}\,(U)$ denotes the diagonal matrix with diagonal entries given by the vector U. We recognize the Ising model (3) with parameter $\Theta + \Phi\Delta^{-1}\Phi^T/2 + \mathrm{Diag}\left(\Phi\Delta^{-1}\mu\right)$.

These properties yield an algorithm to sample from (4), which provides a numerical approximation of the partition function Z_Ω and could lead to a stochastic optimization algorithm to minimize the penalized negative log-likelihood, but this quantity is complex to calculate and would be approximated with MCMC simulations.

2 Structure Learning

We show now how to learn hybrid sparse networks by minimizing a likelihood function penalized by ℓ_1 regularization and ℓ_1/ℓ_2 regularization. This penalization is reasonable, since it allows to learn networks with few connections between nodes. Group Lasso is useful for penalizing when additional structures in the data are known *a priori*. Here, since we use binary variables, each categorical variable is transformed in a set of binary variables using 1-of-K encoding scheme, as proposed in [16, Section 4.3.4] and [9]. The principle is the following: for $i \in \mathcal{C}$, if x_i takes values in $1,\dots,m_i$, we use instead the binary vector $t^{(i)} \in \{0,1\}^{m_i}$, with $t_{k_0}^{(i)} = 1$ if $x_i = k_0$, and $t_k^{(i)} = 0$ elsewhere for $k \neq k_0$. This transformation will only be done for categorical variables and thus will only impact Θ and Φ, whose dimensions will be consequently increased. Thereafter in this paper, when we use the notation X and $X_{\mathcal{C}}$, we will suppose that the discrete data were already transformed following this scheme.

Since we force a structure over the data, we need to penalize variables by groups. The penalization g we use involves ℓ_1 and ℓ_1/ℓ_2 penalty, plus a compact constraint on Δ. That constraint is compulsory to ensure that Δ remains inside a compact set included in the cone of positive-definite matrices. It follows that our learning criterion is L-Lipschitz, what is a required hypothesis for proximal gradient (see [11], H1).

For any $0 < \rho < 1$, denote by $\mathcal{K}_\rho$ the compact subset of positive definite symmetric matrices defined by

$$\begin{aligned} \mathcal{K}_\rho = \{\, &\Delta_0^{1/2}(I+\epsilon)\Delta_0^{1/2}\ :\ \epsilon \text{ is symmetric with} \\ &-\rho < \lambda_{\min}(\epsilon) < \rho\} \end{aligned}$$

where Δ_0 is the empiric precision, I is the identity matrix, $\lambda_{\min}$ denotes the minimal eigenvalue and $\lambda_{\max}$ denotes the maximal one. Observe that $\mathcal{K}_\rho$ can be seen as the ball of symmetric matrices endowed with the Euclidean operator norm, centered at Δ_0 and with radius ρ. Here ρ is arbitrary chosen to ensure the convergence of the numerical optimization. In practice, one needs to check that the obtained optimizer is in the interior of the compact set.

Thus the penalization we use is

$$\begin{aligned} g(\Omega) = \lambda_\theta \sum_{g\neq g'\in G_\Theta} \|\theta_{gg'}\|_2 + \mathbb{1}_{\{\mathcal{K}_\rho\}}(\Delta) \\ + \lambda_\Delta \sum_{u<v\in\mathcal{Q}} |\Delta_{uv}| + \lambda_\Phi \sum_{g\in G_\Theta, u\in\mathcal{Q}} \|\Phi_{gu}\|_2\ , \end{aligned} \quad (5)$$

where $\theta_{gg'} = (\theta_{ii'})_{i\in g, i'\in g'}$ and $\phi_{gu} = (\phi_{iu})_{i\in g}$ with $G_\Theta = \{g_1,\dots,g_{|\mathcal{C}|}\}$ and, for all $i \in \mathcal{C}$, g_i is the set of indexes of binary variables created after applying 1-of-K scheme over non binary discrete variable x_i.

The general problem we want to solve is finding the estimator

$$\hat{\Omega} = \underset{\Omega}{\mathrm{Argmin}} \quad -\ell(\Omega) + g(\Omega)\ , \quad (6)$$

where ℓ is a likelihood function, and g the penalization (5) we describe above. Usually one uses the log-likelihood, which is a concave function (see [7], corollary 20.1).

Here, we rather focus on the minimization problem with the negative pseudo-log-likelihood $-p\ell(\Omega)$. This approach is admittedly sub-optimal, however it does not require any approximation and all involved quantities can explicitly be calculated. Indeed, we define the pseudo-log-likelihood, for a sample X, by

$$p\ell(\Omega \mid X) = \log p_\Omega(X_\mathcal{Q} \mid X_\mathcal{C}) + \sum_{i \in \mathcal{C}} \log p_\Omega(X_i \mid X_{-i}) \,, \tag{7}$$

where X_{-i} represents all the variables except X_i. Note that this pseudo-likelihood is defined over a parametrization space such that Θ and Δ are, respectively, symmetric and symmetric positive-definite. Remark also that $p\ell$ is also concave like the log-likelihood. Since, given $X_\mathcal{C}$, $X_\mathcal{Q}$ admits a conditional normal density with mean $\Delta^{-1}(\mu + \Phi^T X_\mathcal{C})$ and covariance matrix Δ^{-1}, we have

$$\begin{aligned}\log p_\Omega(X_\mathcal{Q} \mid X_\mathcal{C}) = &-\frac{1}{2} X_\mathcal{Q}^T \Delta X_\mathcal{Q} + (\mu + \Phi^T X_\mathcal{C})^T X_\mathcal{Q} \\ &- \frac{1}{2}(\mu + \Phi^T X_\mathcal{C})^T \Delta^{-1} (\mu + \Phi^T X_\mathcal{C}) \\ &+ \log[(2\pi)^{-\frac{|\mathcal{Q}|}{2}} |\Delta|^{\frac{1}{2}}] \,.\end{aligned}$$

That formula involves only linear terms in $X_\mathcal{Q}$ and a term independent of $X_\mathcal{Q}$, so $p_\Omega(X_\mathcal{Q} \mid X_\mathcal{C})$ is a special case of log-linear model and thus it has a concave log-likelihood. Differentiating in Δ, Φ and μ yields the gradients

$$\begin{aligned}\partial_\Delta \log p_\Omega(X_\mathcal{Q} \mid X_\mathcal{C}) = &-\frac{1}{2}\left[X_\mathcal{Q} X_\mathcal{Q}^T - \Delta^{-1}\right. \\ &\left. + \Delta^{-1}(\mu + \Phi^T X_\mathcal{C})(\mu + \Phi^T X_\mathcal{C})^T \Delta^{-1}\right] \\ \partial_\Phi \log p_\Omega(X_\mathcal{Q} \mid X_\mathcal{C}) = &\, X_\mathcal{C} X_\mathcal{Q}^T - X_\mathcal{C}(\mu + \Phi^T X_\mathcal{C})^T \Delta^{-1} \\ \partial_\mu \log p_\Omega(X_\mathcal{Q} \mid X_\mathcal{C}) = &\, X_\mathcal{Q} - \Delta^{-1}\mu - \Delta^{-1}\Phi^T X_\mathcal{C} \,.\end{aligned}$$

Concerning the Ising part of $p\ell$, for $i_0 \in \mathcal{C}$, we calculate the conditional probability of X_{i_0} given X_{-i_0}. We get the logistic-like formula

$$P_\Omega(X_{i_0} = 1 \mid X_{-i_0}) = \frac{\mathrm{e}^{q_\Omega(X, i_0)}}{1 + \mathrm{e}^{q_\Omega(X, i_0)}} \,,$$

with

$$q_\Omega(X, i_0) = \theta_{i_0 i_0} + \Theta_{i_0, -i_0} X_{-i_0} + X_{-i_0}^T \Theta_{-i_0, i_0} + \Phi_{i_0, \mathcal{Q}} X_\mathcal{Q} \,,$$

where Θ_{-i_0, i_0} represents $(\theta_{i_0, j})_{j \neq i_0}$, the i_0th line of Θ without the i_0th element, and $\Phi_{i_0, \mathcal{Q}}$ represents the i_0th line of Φ. Note that X_{-i_0} here denotes the vector $X_\mathcal{C}$ with the i_0 entry removed, whereas previously it denoted the whole X with the i_0 entry removed. Thus we get

$$\begin{aligned}\log p_\Omega(X_{i_0} \mid X_{-i_0}) &= X_{i_0} \log P_\Omega(X_{i_0} = 1 \mid X_{-i_0}) \\ &\quad + (1 - X_{i_0}) \log(1 - P_\Omega(X_{i_0} = 1 \mid X_{-i_0})) \\ &= X_{i_0} q_\Omega(X, i_0) - \log(1 + \exp q_\Omega(X, i_0)) \,,\end{aligned}$$

This formula also involves only linear terms in X_{i_0} and a term independent of X_{i_0}, so $p_\Omega(X_{i_0} \mid X_{-i_0})$ is a special case of log-linear model and thus it has a concave log-likelihood. As a sum of concave terms, the pseudo-log-likelihood (7) is actually concave. Observe that, for all $i, j \in \mathcal{C}$,

$$\partial_{\Theta_{i,j}} q_\Omega(X, i_0) = \begin{cases} \mathbb{1}_{\{i = i_0\}} & \text{if } i = j, \\ \mathbb{1}_{\{i = i_0\}} X_j + \mathbb{1}_{\{j = i_0\}} X_i & \text{if } i \neq j. \end{cases}$$

and, for all $i \in \mathcal{C}$ and $v \in \mathcal{Q}$,

$$\partial_{\Phi_{i,v}} q_\Omega(X, i_0) = \mathbb{1}_{\{i = i_0\}} X_v \,.$$

It follows that

$$\begin{aligned}\partial_\Theta \sum_{i_0 \in \mathcal{C}} \log p_\Omega(X_{i_0} \mid X_{-i_0}) = &-\mathrm{Diag}\left(E_\Omega(X, \mathcal{C}) \circ (2X_\mathcal{C} - 1)\right) \\ &+ 2X_\mathcal{C} X_\mathcal{C}^T - (E_\Omega(X, \mathcal{C}) X_\mathcal{C}^T + X_\mathcal{C} E_\Omega(X, \mathcal{C})^T) \,,\end{aligned}$$

where here Diag (A) denotes the vector with entries given by the diagonal of the square matrix A, $A \circ B$ denotes the Hadamard product of two matrices A and B, and $E_\Omega(X, \mathcal{C})$ is the vector defined by

$$E_\Omega(X, i) = \frac{\mathrm{e}^{q_\Omega(X,i)}}{1 + \mathrm{e}^{q_\Omega(X,i)}}, \quad i \in \mathcal{C} .$$

Similarly, we get that

$$\partial_\Phi \sum_{i_0 \in \mathcal{C}} \log P_\Omega(X_{i_0} \mid X_{-i_0}) = X_{\mathcal{C}} X_{\mathcal{Q}}^T - E_\Omega(X, \mathcal{C}) X_{\mathcal{Q}}^T .$$

Note also that another way to write $q_\Omega(X, \mathcal{C}) = (q_\Omega(X, i))_{i \in \mathcal{C}}$ is to set

$$q_\Omega(X, \mathcal{C}) = (\Theta + \Theta^T) X_{\mathcal{C}} + \mathrm{Diag}\,(\Theta) \circ (1 - 2X_{\mathcal{C}}) + \Phi X_{\mathcal{Q}} .$$

Those equations carry out an algorithm for structure learning, using proximal gradient. This algorithm is particularly relevant for convex optimization when the regularization is not differentiable, as it is the case here. The algorithm is proposed in [11]. If Ω_0 denotes the starting estimates, and $\{\gamma_n, n \in \mathbb{N}\}$ a sequence of positive step sizes, then given Ω_n, we compute

$$\Omega_{n+1} = \mathsf{Prox}_{\gamma_{n+1}}(\theta_n + \gamma_{n+1} \nabla p\ell(\Omega_n)) , \tag{8}$$

where Prox_γ is the proximal operator defined by

$$\mathsf{Prox}_\gamma(\theta) = \underset{\vartheta}{\mathrm{Argmin}} \left\{ \frac{1}{2\gamma} \|\vartheta - \theta\|^2 + g(\vartheta) \right\} .$$

The penalization g that we are using, defined in (5), is the sum of ℓ_1 and ℓ_1/ℓ_2 regularization over different variables, plus a compact constraint. Note that there is no overlapping groups here. With such ℓ_1 and ℓ_1/ℓ_2 regularizations, the proximal operator can be reformulated as a component-wise soft-threshold $\sigma_{\lambda,\gamma}(\Omega)$ defined by

$$\sigma_{\lambda,\gamma}(\Omega) = (\tilde{s}_{\lambda,\gamma}(\Theta), s_{\lambda,\gamma}(\mu), s_{\lambda,\gamma}(\Delta), \tilde{s}_{\lambda,\gamma}(\Phi)) ,$$

where, for a general matrix A, $s_{\lambda,\gamma}(A)$ is the matrix defined by blocks, for any $g, g' \in G_\Theta$, by

$$s_{\lambda,\gamma}(A)_{ij} = \begin{cases} A_{ij} - \lambda\gamma & \text{if } A_{ij} > \lambda\gamma \\ A_{ij} + \lambda\gamma & \text{if } A_{ij} < \lambda\gamma \\ 0 & \text{elsewhere,} \end{cases}$$

and $\tilde{s}_{\lambda,\gamma}(A)$ is the matrix defined by blocks, for $g, g' \in G_\Theta$, by

$$\tilde{s}_{\lambda,\gamma}(A)_{gg'} = \begin{cases} A_{gg'} - \lambda\gamma \frac{A_{gg'}}{\|A_{gg'}\|_2} & \text{if } \|A_{gg'}\|_2 > \lambda\gamma , \\ 0 & \text{elsewhere.} \end{cases}$$

With the compact constraint $\mathbb{1}_{\{\mathcal{K}_\rho\}}$, the proximal operator is the orthogonal projection on $\mathcal{K}_\rho$, which is the map $\Pi_{\mathcal{K}_\rho}$ such that $\|\Delta - \Pi_{\mathcal{K}_\rho}(\Delta)\| = \min_{\Delta'} \|\Delta - \Delta'\|$.

Since our penalization is composed of two regularization terms, we solve (6) using the generalized forward-backward splitting algorithm [17]. Note that applying the compact constraint does not insure that Δ_n will remain in K_ρ for every iteration, it only tends to bring Δ_n back inside K_ρ. To guarantee that Δ_n remains definite positive during the learning process, one must control the gradient step γ_n.

3 Application to anomaly detection

Graphical models are particularly well fitted to the anomaly detection task. A classic method (see [1]) for that purpose is to learn a model from data and then to estimate the likelihood on new records; records with low likelihood can be labelled as anomalies. [2] and [3] have successfully applied that approach in the network security and medical fields, respectively for intrusion and diseases detection.

The potential of Bayesian networks for breakdown detection in the aeronautic field was shown in [5]. However, this study assumed that every variable were categorical. Also the learning of Bayesian networks by optimization of a criterion (BIC, MDL ...) implies local search and heuristics (see [7] chapter 18). There is therefore no warranty that the learned model is actually an optimum of the learning criterion.

Our algorithms were designed for breakdown detection in complex electronic systems, especially Active Electronically Scanned Array Radars (AESA Radars). The complexity of the systems growing alongside the capacity of storage, new maintenance systems had to be developed. The amount of variables can be very high (around 100 000), but for this application we used a restrained subset of 310 variables. Our dataset was built over 140 categorical variables and 20 quantitative variables. After the categorical data have been transformed with 1-of-K scheme – as explained in 2 – we had a final data set of 290 categorical variables and 20 quantitative variables. Figure 1 shows the minimisation of the penalized negative pseudo-log-likelihood.

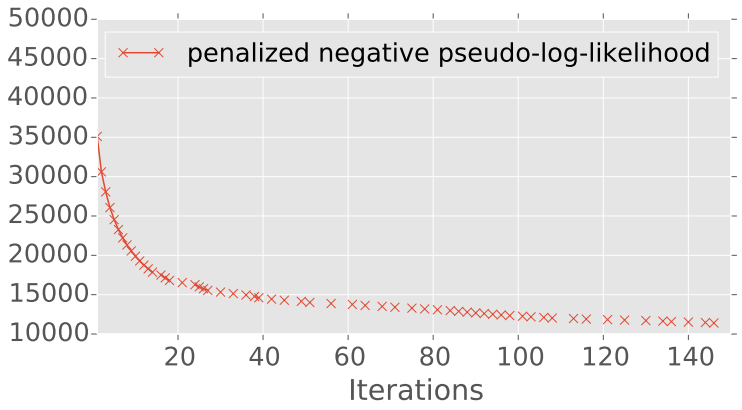

Figure 1: Minimization of the penalized negative pseudo-log-likelihood when learning graph structure. The $+\infty$ values of score (caused by the compact constraint in 5) are not represented.

The anomaly detection is then performed by analysing the pseudo-log-likelihood of new records. Figure 3. shows histograms of pseudo-log-likelihood, for new records without anomalies (at the top) and records with anomalies (at the bottom). Figure 2 shows the pseudo-log-likelihood of each sample of normal records, followed by anomalous records. One can observe that the pseudo-log-likelihood allows to score and spot efficiently anomalous records.

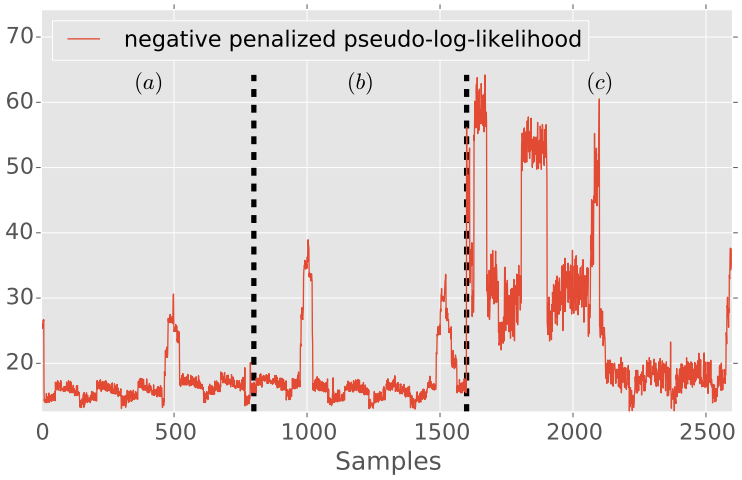

Figure 2: Negative penalized pseudo-log-likelihood of: (a) 800 training samples, (b) 800 new records without anomalies and (c) 1000 new records with anomalies in the first 500 samples.

4 Conclusion

In this paper, we proposed a method to learn an hybrid Markov model from heterogeneous data, composed by categorical (non necessarily binary) and quantitative variables. The learnt graphical model is sparse, because of the use of an ℓ_1 and ℓ_1/ℓ_2 penalty. It reveals interactions between variables, that can be easily interpreted and checked by the user. To our knowledge, this approach is new in the context of hybrid

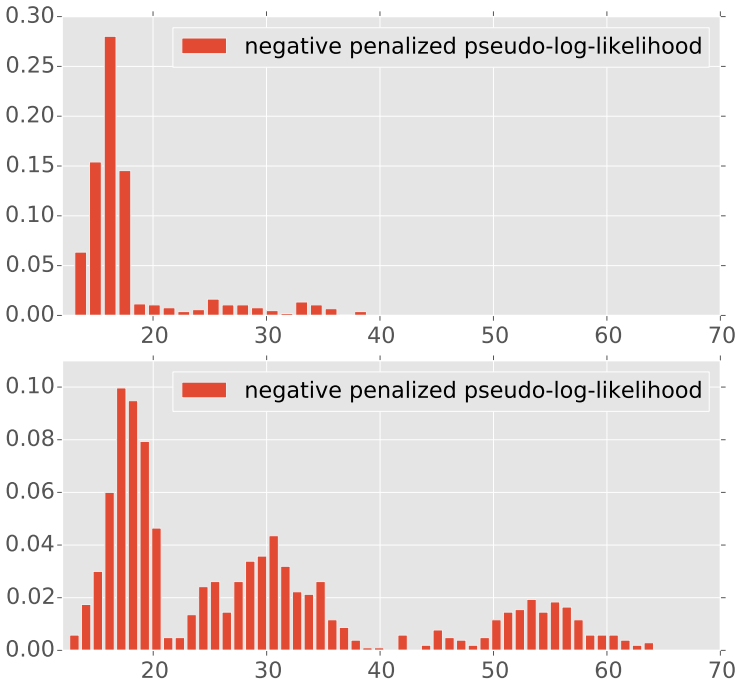

Figure 3: Histogram of negative penalized pseudo-log-likelihood of new AESA radar normal records (at the top) and records with anomalies (at the bottom). The anomalous records are producing new lobes of hight negative pseudo-log-likelihood.

undirected graphical models without translating the data into new feature spaces. We successfully applied our learning method to the problem of anomaly detection in the specific case of breakdown detection in AESA radars. In the field of aeronautic industry, and more precisely of the AESA radars, the learned graph provides a solution for practical maintenance issues. It brings a way to find interactions between hardware and/or software components of complex systems. When learned from normal data, the model can be considered as a reference of a good-working system, and subsequent comparisons with new records are obtained in the form of a pseudo-likelihood statistic, which can be computed online.

References

[1] L. Rashidi, S. Hashemi, and A. Hamzeh, "Anomaly detection in categorical datasets using bayesian networks," in *Artificial Intelligence and Computational Intelligence*, pp. 610–619, Springer, 2011.

[2] N. Ye and M. Xu, "Probabilistic networks with undirected links for anomaly detection," *IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop*, pp. 175–179, 2000.

[3] W.-K. Wong, A. Moore, G. Cooper, and M. Wagner, "Bayesian network anomaly pattern detection for disease outbreaks," in *ICML*, pp. 808–815, 2003.

[4] U. Lerner, R. Parr, D. Koller, G. Biswas, *et al.*, "Bayesian fault detection and diagnosis in dynamic systems," in *AAAI/IAAI*, pp. 531–537, 2000.

[5] S. Kemkemian, A. Larroque, and C. Enderli, "The industrial challenges of airborne aesa radars," 2013.

[6] T. Heskes and O. Zoeter, "Generalized belief propagation for approximate inference in hybrid bayesian networks," in *Artificial Intelligence and Statistics*, 2003.

[7] N. Friedman and D. Koller, *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

[8] F. R. Bach and M. I. Jordan, "Learning graphical models with mercer kernels," in *Advances in Neural Information Processing Systems*, pp. 1009–1016, 2002.

[9] M. Schmidt, *Graphical model structure learning with l1-regularization.* PhD thesis, University Of Bristish Columbia (Vancouver), 2010.

[10] G. Varoquaux, A. Gramfort, J.-B. Poline, and B. Thirion, "Brain covariance selection: better individual functional connectivity models using population prior.," in *NIPS* (J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, eds.), pp. 2334–2342, Curran Associates, Inc., 2010.

[11] Y. F. Atchade, G. Fort, and E. Moulines, "On stochastic proximal gradient algorithms," *arXiv preprint arXiv:1402.2365v2*, 2015.

[12] R. Laby, A. Gramfort, F. Roueff, C. Enderli, and A. Larroque, "Apprentissage d'un modèle graphique non orienté hybride parcimonieux par utilisation du gradient proximal stochastique." Submitted.

[13] J. Besag, "Statistical analysis of non-lattice data," *The Statistician*, vol. 24, 1975.

[14] E. Ising, "Beitrag zur theorie des ferromagnetismus," *Zeitschrift fur Physik*, vol. 31, pp. 253–258, 1925.

[15] R. Potts, "Some generalized order-disorder transformations," *Proc. Cambridge Philosophie Soc*, 1953.

[16] C. Bishop, *Pattern Recognition and Machine Learning.* 2006.

[17] H. Raguet, J. Fadili, and G. Peyré, "A generalized forward-backward splitting," *SIAM Journal on Imaging Sciences*, vol. 6, no. 3, pp. 1199–1226, 2013.