Apprentissage d'un modèle graphique non orienté hybride parcimonieux par utilisation du gradient proximal stochastique

Romain LABY[1,2], Alexandre GRAMFORT[1], François ROUEFF[1], Cyrille ENDERLI[2], Alain LARROQUE[2]

[1]Laboratoire Traitement et Communication de l'Information
CNRS LTCI, Télécom ParisTech, 46 rue Barrault, F-75634 Paris Cedex 13

[2]Thales Systèmes aéroportés
2 avenue Gay Lussac, CS 80501, 78990 Elancourt, France

romain.laby@telecom-paristech.fr, francois.roueff@telecom-paristech.fr,
alexandre.gramfort@telecom-paristech.fr, cyrille-jean.enderli@fr.thalesgroup.com
alain.larroque@fr.thalesgroup.com

Résumé – Cet article présente une méthode d'apprentissage de modèle graphique hybride non orienté parcimonieux à partir d'un jeu de données, lorsque la structure du modèle n'est pas connue. Le modèle hybride proposé est construit à partir d'un modèle d'Ising pour les variables catégorielles et d'un modèle gaussien pour les variables continues. L'apprentissage est réalisé à l'aide d'une version stochastique de l'algorithme du gradient proximal, dans lequel le gradient de la vraisemblance n'est pas calculable et est estimé par des simulations de type Markov Chain Monte Carlo.

Abstract – This article proposes a method for learning from data undirected sparse hybrid graphical models when the structure is not known. The hybrid model we propose is based on an Ising model for categoriacal variables and a Gaussian model for continuous variables. The learning part is done using a stochastic version of the proximal gradient algorithm. As the gradient of the likelihood is not tractable, we propose to estimate it with a Markov Chain Monte Carlo method.

1 Introduction

Les modèles graphiques sont utilisés pour représenter efficacement des distributions de probabilités et ont des applications dans de nombreux domaines. Plus particulièrement, ils sont une bonne solution au problème de la détection et de la localisation d'anomalies. Si le sujet avait déjà été exploré dans le domaine médical [1], l'application dans le domaine de l'aéronautique pour la prévention des pannes est plus récente [2]. Une caractéristique importante de ce domaine est la présence de variables nombreuses et très diverses, certaines catégorielles (typiquement pour remonter l'état de fonctionnement de systèmes ou composants électronique sous la forme marche/ arrêt/ hors-service) et d'autres quantitatives (mesures physiques telles que des températures ou des tension électrique).

L'apprentissage de modèles non orientés par régularisation ℓ^1 est un sujet qui a été largement étudié ces dernières années, notamment dans le cas des graphes non orientés *par paires* : [3] propose une large revue des méthodes actuelles traitant cette problématique. On s'intéresse ici à l'approche de minimisation de la vraisemblance. Contrairement aux graphes orientés acycliques, les graphes non orientés peuvent facilement être associés à une fonction de vraisemblance concave (voir [4], §20.2.3) qui assure l'absence d'optimum local. En outre, l'absence de contraintes d'acyclicité sur les graphes permet d'utiliser des méthodes de descente de gradient (ou de gradient proximal dans le cas d'une pénalisation de type ℓ^1) pour atteindre l'optimum global. Néanmoins, dans un contexte où le nombre de variables est grand (plus d'une centaine), la vraisemblance et son gradient peuvent être très coûteux à calculer. Pour faire face à cette situation, [5] propose une méthode stochastique basée sur le calcul de la pseudo-vraisemblance, mais cette méthode est connue pour être sous-optimale. Récemment, [6] a proposé une méthode de recherche d'optimal d'une fonction objectif concave régularisée à l'aide d'une version stochastique du gradient proximal. Cette méthode est adaptée à la grande dimension et dans le cas où la fonction objectif ainsi que son gradient seraient coûteux à calculer de manière exacte mais pourraient être approchés par une méthode de Monte Carlo reposant sur une chaîne de Markov (MCMC).

Les travaux réalisés dans le domaine de l'apprentissage de graphes non orientés par régularisation l_1 traitent séparément de données discrètes ou continues, mais très peu de données mélangeant ces deux types de variables. D'autres approches ont été envisagées ; par exemple, [7] propose une solution où les variables utilisées pour l'apprentissage du graphe sont des versions gaussiennes des données initiales, qu'elles soient discrètes ou continues. L'approche utilise des noyaux de Mercer. Ici, on propose en partie 2 un modèle hybride mélangeant des données discrètes (catégorielles) et continues (quantitatives).

On propose en section 3 une méthode d'apprentissage de graphes non orientés, modélisant la distribution jointe de ces variables, en utilisant du gradient proximal stochastique.

2 Présentation du modèle hybride

Parmi les diverses modélisations possibles de réseaux non orientés, la classe des réseaux *par paires* utilise des *potentiels* qui sont des fonctions d'une ou deux variables. Dans de tels réseaux, la densité $p(x_1, \dots, x_N)$ de N variables aléatoires se met sous la forme

$$p(x_1, \dots, x_N) = \frac{1}{Z} \prod_{i=1}^{N} \varphi_i(x_i) \prod_{i,j \in E} \varphi_{ij}(x_i, x_j) \ , \quad (1)$$

où l'ensemble E contient toutes les paires de variables que l'on veut inclure dans le modèle. Le graphe non orienté associé contiendra un arc entre les noeuds i et j si $\varphi_{ij} \neq 1$. Il est aisé de paramétrer les potentiels φ_i et $\varphi_{i,j}$ par des familles de fonctions log-concave en les paramètres. La variable Z est la constante de normalisation nécessaire au calcul de la vraisemblance qui est cependant incalculable si N est grand.

Deux modélisations sont très utilisées pour les réseaux non orientés *par paires* : les modèles d'Ising et les modèles gaussiens. Le cas gaussien est le plus simple puisqu'il revient à dire que p est la densité d'un vecteur gaussien. Dans ce cas le calcul de Z ne pose pas de problème car cela revient à calculer le déterminant d'une matrice $N \times N$. Le modèle d'Ising (IGM, *Ising Graphical Model*) correspond au cas où tous les x_i sont à valeurs binaires, par exemple dans $\{0, 1\}$. Ces modèles ont été généralisés à des ensembles de valeurs plus grands, sous la forme par exemple du modèle de Potts ([8]), mais ce dernier peut être reparamétré sous la forme d'un IGM en utilisant la paramétrisation de [9], §4.3.4. Dans le cas d'un IGM, la densité a la forme

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

où $\Theta = (\theta_{i,j})$ est un paramètre de $\mathbb{R}^{N(N+1)/2}$. Pour des raisons pratiques on considère que Θ est une matrice symétrique $N \times N$. Comme $x_i = x_i^2$ pour $x_i \in \{0, 1\}$, on peut écrire alors

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

Cette forme est essentiellement similaire à une densité gaussienne, mais Θ n'a aucune contrainte puisqu'elle n'est pas associée à une matrice de covariance. Le calcul de la constante de normalisation Z_Θ peut s'avérer très coûteux quand N est grand (somme de 2^N termes).

On propose un nouveau modèle hybride mixant des variables X_i, $i \in \mathcal{C}$, binaires (appelées catégorielles par la suite) et des variables X_i, $i \in \mathcal{Q}$ continues (appelées quantitatives par la suite). Soient $X = (X_\mathcal{C}, X_\mathcal{Q})$ l'ensemble des variables de notre modèle, à valeurs dans $\mathcal{X} = \{0, 1\}^\mathcal{C} \times \mathbb{R}^\mathcal{Q}$. On propose le modèle graphique hybride *par paires* suivant :

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

où $\Omega = (\Theta, \mu, \Delta, \Phi)$ avec $\Theta = (\theta_{ij})_{i,j \in \mathcal{C}}$ matrice symétrique, $\mu = (\mu_i)_{i \in \mathcal{Q}} \in \mathbb{R}^\mathcal{Q}$, $\Delta = (\delta_{uv})_{u,v \in \mathcal{Q}}$ matrice symétrique et $\Phi = (\phi_{iu})_{i,u \in \mathcal{C} \times \mathcal{Q}}$ matrice quelconque. Pour que la fonction p_Ω soit une densité par rapport à la mesure produit composée de la mesure de comptage sur $\{0, 1\}^\mathcal{C}$ et de la mesure de Lebesgue sur $\mathbb{R}^\mathcal{Q}$, il est clair qu'une condition nécessaire et suffisante est que Δ soit une matrice définie positive, ce que nous supposons par la suite. En revanche, aucune condition n'est imposée à Θ, μ et Φ autre que Θ symétrique.

La densité (4) a les propriétés suivantes. Vue comme une fonction de $x_\mathcal{Q}$ uniquement, on a

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

où $\propto$ signifie l'égalité de fonctions à une constante multiplicative près (ici pouvant dépendre de $x_\mathcal{C}$ puisqu'on voit $p_\Omega(x)$ comme une fonction de $x_\mathcal{Q}$ uniquement). On reconnaît dans le membre de droite de la formule précédente une densité gaussienne. On en déduit que, conditionnellement à $X_\mathcal{C}$, $X_\mathcal{Q}$ est un vecteur gaussien de moyenne $\Delta^{-1}\left(\mu + \Phi^T X_\mathcal{C}\right)$ et de covariance Δ^{-1}.

De même, il est aisé de montrer que, conditionnellement à $X_\mathcal{Q}$, $X_\mathcal{C}$ est un modèle IGM. De façon plus surprenante, on peut aussi montrer que la loi non-conditionnelle de $X_\mathcal{C}$ reste un modèle IGM (alors qu'il est clair que la loi non-conditionnelle de $X_\mathcal{Q}$ n'est pas une loi gaussienne mais un mélange de lois gaussiennes). En effet, si on note $p_{\mathcal{C}\Omega}$ la densité de $X_\mathcal{C}$, on obtient

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

On peut interpréter l'intégrale (à une constante multiplicative près) comme une espérance $\mathbb{E}[\exp\left((\mu + \Phi^T x_\mathcal{C})^T U\right)]$ où U est un vecteur gaussien centré de covariance Δ^{-1}. On obtient donc

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

où, à la dernière ligne, on a utilisé $x_i^2 = x_i$. On reconnaît le modèle IGM (3) de paramètre $\Theta + \Phi \Delta^{-1} \Phi^T / 2 + \mathrm{Diag}\left(\Phi \Delta^{-1} \mu\right)$.

3 Apprentissage de structure

On s'intéresse maintenant à l'apprentissage de ces réseaux hybrides, par minimisation de l'opposé de la log-vraisemblance

pénalisée par une régularisation de type ℓ^1. Cette pénalisation est légitime dans un contexte parcimonieux afin de favoriser les paramètres Ω avec peu de connections. On définit notre estimateur par

$$\hat{\Omega} = \underset{\Omega}{\text{Argmin}} \quad -\ell(\Omega) + g(\Omega) \tag{5}$$

où $\ell(\Omega)$ désigne la log vraisemblance et $g(\Omega)$ est une pénalisation sur les arcs des graphes, définie par :

$$g(\Omega) = \lambda_1 \sum_{i<j\in\mathcal{C}} |\theta_{ij}| + \lambda_2 \sum_{u<v\in\mathcal{Q}} |\Delta_{uv}| + \lambda_3 \sum_{i,u\in\mathcal{C}\times\mathcal{Q}} |\Phi_{iu}|$$

Une méthode de recherche d'optimum global de la vraisemblance d'un graphe non orienté est proposée par [6]. L'algorithme est un gradient proximal dans une version stochastique liée à l'estimation de la constante de normalisation Z_Ω qui est d'ordinaire trop complexe à calculer exactement. L'algorithme du gradient proximal peut être vu comme un cas particulier des méthodes de majoration-minimisation (MM), voir [10] section 1.3 et [11]. Sous certaines hypothèses de régularité vérifiées ici (voir [6], H1 à H5), l'algorithme

$$\Omega_{n+1} = \text{Prox}_{\gamma_{n+1}}(\Omega_n + \gamma_{n+1}\nabla\ell(\Omega_n)) \tag{6}$$

converge pour des pas $\{\gamma_n, n \in \mathbb{N}\}$ positifs, où Prox est l'opérateur de gradient proximal de g défini par

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

Avec la pénalisation g ci-dessus, l'opérateur proximal est facilement calculable, puisqu'il est l'opérateur de seuillage doux $s_{\gamma,\lambda}(\Omega)$ défini par

$$s_{\gamma,\lambda}(\Omega) = (s_{\gamma,\lambda_1}(\Theta), s_{\gamma,\lambda_2}(\Delta), s_{\gamma,\lambda_3}(\Phi))$$

et où, hors termes diagonaux pour Θ et Δ,

$$s_{\gamma,\lambda}(\vartheta)_{ij} = \begin{cases} \vartheta_{ij} - \gamma\lambda & \text{si } \vartheta_{ij} \geq \gamma\lambda \ , \\ \vartheta_{ij} + \gamma\lambda & \text{si } \vartheta_{ij} \leq -\gamma\lambda \ , \\ 0 & \text{sinon.} \end{cases}$$

Pour compléter l'algorithme proposé par [6], il s'agit de proposer un calcul approché de $\nabla\ell(\Omega_n)$ à chaque étape n. Nous expliquons la procédure MCMC utilisée pour cela au paragraphe suivant.

4 Calcul approché de la vraisemblance et de son gradient

Dans notre cas, ℓ et $\nabla\ell$ ne sont pas calculables de manière exacte, et on s'oriente vers une version stochastique (voir [6], section 3) dans laquelle $\nabla\ell(\Omega)$ est approchée à l'aide d'une procédure MCMC. Pour des facilités d'écriture, on introduit la statistique F pour écrire la vraisemblance $p_\Omega(X)$:

$$p_\Omega(X) = \frac{1}{Z_\Omega} \exp(\langle \Omega, F \rangle) \ , \tag{7}$$

où $F = (F_1, F_2, F_3, F_4)$ avec :

- F_1 est la matrice indexée sur $\mathcal{C}\times\mathcal{C}$ définie par $F_1 = X_\mathcal{C} X_\mathcal{C}^T$,
- F_2 est le vecteur indexé sur $\mathcal{Q}$ défini par $F_2 = X_\mathcal{Q}$,
- F_3 est la matrice indexée sur $\mathcal{Q}\times\mathcal{Q}$ définie par $F_3 = -\frac{1}{2} X_\mathcal{Q} X_\mathcal{Q}^T$,
- F_4 est la matrice indexée sur $\mathcal{C}\times\mathcal{Q}$ définie par $F_4 = X_\mathcal{C} X_\mathcal{Q}^T$,

et où $\langle \cdot, \cdot \rangle$ est le produit scalaire défini par

$$\langle \Omega, F \rangle = \text{Tr}(\Theta F_1^T) + \mu^T F_2 + \text{Tr}(\Delta F_3^T) + \text{Tr}(\Phi F_4^T). \tag{8}$$

D'après (7), si $\mathcal{D} = \{X^{(k)}\}_{k=1}^{m_\mathcal{D}}$ sont des données d'apprentissage et si on note $F^{(k)}$ la statistique F correspondant à l'observation $X^{(k)}$, l'opposé de la log-vraisemblance s'écrit

$$\ell(\Omega) = \frac{1}{m_\mathcal{D}} \sum_{k=1}^{m_\mathcal{D}} \langle \Omega, F^{(k)} \rangle - \log Z_\Omega \ .$$

L'identité de Fisher $\mathbb{E}_\Omega[\nabla\ell(\Omega)] = 0$ donne donc

$$\nabla \log Z_\Omega = \mathbb{E}_\Omega[F] = \mathbb{E}_\Omega[\mathbb{E}_\Omega[F \mid X_\mathcal{C}]] \tag{9}$$

Ce résultat suggère une méthode pour estimer $\nabla \log Z_\Omega$:

(a) La loi conditionnelle de $X_\mathcal{Q}$ sachant $X_\mathcal{C}$ étant connue (voir § 2), on calcule $\mathbb{E}_\Omega[F \mid X_\mathcal{C}]$ explicitement comme une statistique dépendant de $X_\mathcal{C}$. On obtient plus précisément

$$\begin{aligned} &\mathbb{E}_\Omega(F_1|X_\mathcal{C}) = F_1 \\ &\mathbb{E}_\Omega(F_2|X_\mathcal{C}) = \mathbb{E}(X_\mathcal{Q} \mid X_\mathcal{C}) = \Delta^{-1}(\mu + \Phi^T X_\mathcal{C}) \\ &\mathbb{E}_\Omega(F_3|X_\mathcal{C}) = -\frac{1}{2}\Delta^{-1} - \frac{1}{2}\mathbb{E}_\Omega(F_2|X_\mathcal{C})\mathbb{E}_\Omega(F_2|X_\mathcal{C})^T \\ &\mathbb{E}_\Omega(F_4|X_\mathcal{C}) = X_\mathcal{C}\mathbb{E}_\Omega(F_2|X_\mathcal{C})^T \end{aligned}$$

(b) La loi non-conditionnelle de $X_\mathcal{C}$ suit un modèle d'Ising (voir § 2), que l'on peut simuler comme la loi stationnaire d'une chaîne de Markov (voir [12], qui propose un algorithme ayant de bonnes propriétés de mixage).

La succession de ces étapes fournit donc un algorithme MCMC pour le calcul de $\nabla \log Z_\Omega$ donné par (9).

L'ensemble des étapes de l'algorithme du calcul de $\hat{\Omega}$ défini par (5) est donc finalement décrit par une initialisation Ω_0 arbitraire, puis en itérant : pour l'étape $n \geq 0$, étant donné Ω_n,

1. simuler une chaîne de Markov de loi stationnaire $p_\Omega(x_\mathcal{C})$,
2. calculer une approximation MCMC du gradient $\nabla \log Z_\Omega$ donné par (9) en suivant les étapes (a) et (b),
3. calculer l'approximation H_{n+1} de $\nabla\ell(\Omega_n)$,
4. calculer $\Omega_{n+1} = \text{Prox}_{\gamma_{n+1}}(\Omega_n + \gamma_{n+1} H_{n+1})$.

L'algorithme fait apparaître à chaque itération une complexité temporelle $\mathcal{O}(N^3 + m_n|\mathcal{C}|^2)$, où m_n est la longueur de la chaîne de Markov générée à l'itération n, $|\mathcal{C}|$ le nombre de variables catégorielles et N le nombre total de variables.

5 Illustrations et Exemples

Pour illustrer le modèle proposé et la méthode d'estimation décrite, nous montrons quelques simulations en faible dimension (3 catégorielles, 2 quantitatives). La figure 1 montre des simulations de $X_\mathcal{Q}$ dans le cas $\Phi = 0$ (les variables quantitatives

sont indépendantes des variables catégorielles et suivent donc une loi gaussienne), et dans le cas $\Phi \neq 0$ (les variables quantitatives sont dépendantes des variables catégorielles et suivent donc une loi qui est un mélange de gaussiennes). La partie gauche de la figure 2 montre la minimisation de l'opposé de la log-vraisemblance pénalisée (comme expliqué en début de section § 3) dans la configuration $\Phi \neq 0$. On présente maintenant des résultats en plus grande dimension. On considère 100 variables : 40 variables catégorielles et 60 variables quantitatives. On génère 500 échantillons à partir d'un modèle parcimonieux où en moyenne 5% des coefficients de Φ et des coefficients sous la diagonale (diagonale non incluse) de Θ et Δ sont non-nuls (en conservant le caractère défini positif de Δ). Pour cette simulation, nous avons pris γ_n constant égal à 10^{-3}, et $\lambda_1 = \lambda_2 = \lambda_3 = 0.1$, obtenus après différents essais.

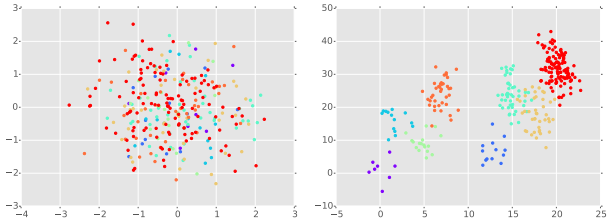

FIGURE 1 – Réalisations i.i.d. de X_Q en dimension 2. Les valeurs de $X_C \in \{0,1\}^3$ sont représentées par 2^3 couleurs différentes. A gauche $\Phi = 0$, à droite $\Phi \neq 0$.

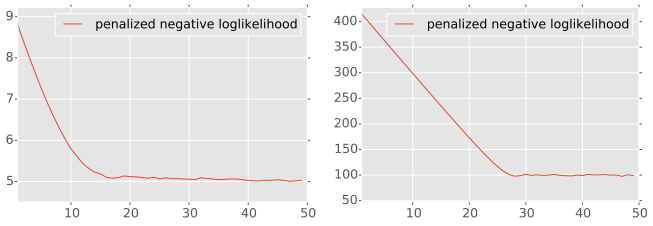

FIGURE 2 – Évolution du critère à minimiser en fonction du nombre d'itérations de l'algorithme en faible dimension (à gauche) et en plus grande dimension (à droite).

La figure 2 montre à droite la minimisation de l'opposé de la log-vraisemblance pénalisée, et la figure 3 montre l'évolution des taux de faux et vrais positifs.

L'apprentissage du modèle hybride 4 trouve une application dans le contexte industriel de détection et de localisation de pannes sur radars aéroportés, voir [13], où le modèle est appris en minimisant la pseudo-log-vraisemblance, qui peut être calculée analytiquement.

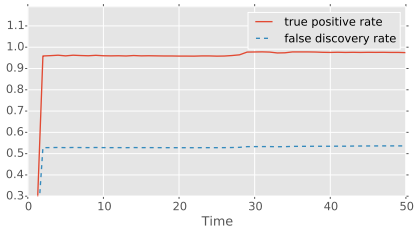

FIGURE 3 – Evolution des taux de connections réelles (en rouge) et taux de connections erronées (en bleu). Augmenter le poids des pénalités permet de diminuer le taux de fausses détections, mais diminue aussi le taux de vraies détections. Le ratio fausses détection - vraies détection ne peut être amélioré qu'en augmentant le nombre d'observations.

Références

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

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

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

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

[5] H. Höfling and R. Tibshirani, "Estimation of sparse binary pairwise markov networks using pseudo-likelihoods," *The Journal of Machine Learning Research*, vol. 10, pp. 883–906, 2009.

[6] Y. F. Atchade, G. Fort, and E. Moulines, "On stochastic proximal gradient algorithms," *arXiv preprint arXiv :1402.2365*, 2014.

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

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

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

[10] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.

[11] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in optimization*, vol. 1, no. 3, pp. 123–231, 2013.

[12] A. Barbu and S.-C. Zhu, "Generalizing swendsen-wang to sampling arbitrary posterior probabilities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 8, pp. 1239–1253, 2005.

[13] R. Laby, A. Gramfort, F. Roueff, C. Enderli, and L. Alain, "Sparse pairwise Markov model learning for anomaly detection in heterogeneous data." June 2015.