Graph Learning as a Tensor Factorization Problem

Raphaël Bailly, Guillaume Rabusseau

Graph Learning as a Tensor Factorization Problem

Raphaël Bailly
SAMM, Université Panthéon-Sorbonne, Paris
rbailly@univ-paris1.fr

Guillaume Rabusseau
RLLab, McGill University, Montreal
guillaume.rabusseau@lif.univ-mrs.fr

Abstract

Graphical models (CRFs, Markov Random Fields, Bayesian networks ...) are probabilistic models fairly widely used in machine learning and other areas (e.g. Ising model in statistical physics). For such models, computing a joint probability for a set of random variables relies on a Markov assumption on the dependencies between the variables. Even so, the calculation may be intractable in the general case, forcing one to consider approximation methods (MCMC, loopy belief propagation, etc.). Hence the maximum likelihood estimator is, except in very particular cases, impossible to compute. We propose a very general probabilistic model, for which parameter estimation can be obtained by tensor factorization techniques (similarly to spectral methods or methods of moments), bypassing the calculation of a joint probability thanks to algebraic results.

1 Introduction

This paper addresses the problem of parameter estimation for probability distributions over graphs. We consider the class of probability distributions that can be computed by Graph Weighted Models (GWMs), a computational model on graphs and hypergraphs that has been recently introduced in [3] (see also [8]). This class of distributions is a generalization of the distributions modeled by HMMs on sequences or trees. The observations are provided as graphs, which can be interpreted as graphs of Markov dependencies for a set of underlying latent variables.

We make the strong hypothesis that the joint probability operator that describes on the one hand the dependencies between a latent state and its related observation, and on the other hand between a latent state and its neighbors, is constant over the whole graph. It is worth mentioning that this hypothesis boils down to the Markov property assumed by the HMM model when one considers linear graphs (i.e. strings). For sequences and trees, the linear operators (i.e. *transition matrices and transition tensors*) are usually estimated by maximizing the likelihood (MLE) of the observations. Alternative techniques have been lately developed (spectral methods [4, 6], method of moments [1, 2], etc.) based on direct estimation involving basic algebraic properties of the model with respect to a particular tensor: the *Hankel tensor*.

When it comes to graphs, the maximum likelihood estimation is hardly available, as computing the exact probability of a given configuration is in general intractable. However, methods based on Hankel tensor factorization are still available even though they need to be adapted. We show in this paper that the problem of estimating the transition operators from a set of observations generated by a GWM can be reduced to a 3-way tensor factorization problem, similar to the well-known RESCAL or DEDICOM factorizations [7, 5].

We present two main results related to the problem of learning graph weighted models:

- Under some assumptions, the set of probability distributions modeled by *graph weighted models (GWM)* is dense in the space of probability distributions over graphs, for the $\|\|_1$-norm induced topology.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

- The estimation of the parameters of a GWM can be achieved through a tensor decomposition problem subject to linear constraints on the factors.

We give a formal definition of Graph Weighted Models and present the density result mentioned above in Section 2 and we show how the learning result can be reduced to a tensor factorization problem in Section 3.

Notations. For any integer k we use $[k]$ to denote the set of integers from 1 to k, and we denote by $\mathfrak{S}_k$ the set of permutations of $[k]$. We use lower case bold letters for vectors (e.g. $\mathbf{v} \in \mathbb{R}^{d_1}$), upper case bold letters for matrices (e.g. $\mathbf{M} \in \mathbb{R}^{d_1 \times d_2}$) and bold calligraphic letters for higher order tensors (e.g. $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$). The $d \times d$ identity matrix will be written as $\mathbf{I}_d$. The ith row (resp. column) of a matrix $\mathbf{M}$ will be denoted by $\mathbf{M}_{i,:}$ (resp. $\mathbf{M}_{:,i}$). This notation is extended to slices of a tensor in the straightforward way. We use the $\otimes$ symbol for the Kronecker product and we denote by $(\mathbb{R}^d)^{\otimes k} = \mathbb{R}^{d \times d \times \cdots \times d}$ the space of d-dimensional hypercubic tensors of order k. Given a tensor $\mathcal{T}$, its vectorization will be written as $\mathrm{vec}(\mathcal{T})$.

2 Graph Weighted Models

Let O be a ranked alphabet where n_o denotes the arity of the symbol o for each $o \in O$. We consider (connected) graphs build on the ranked alphabet O, that is graphs whose vertices are labeled by symbols in O and where *the degree of each vertex coincides with the arity of the symbol it is labeled with*. Formally a graph $G = (V, E, \ell)$ is composed of a set of vertices V, a labeling $\ell : V \to O$ and a set of edges E partitioning the set of ports $P_G = \{(v, i) : v \in V, i \in [n_{\ell(v)}]\}$ into sets of cardinality 2.

2.1 Definitions

Definition 1. *A* Graph Weighted Model *(GWM) with d latent states over an alphabet O is given by a set of tensors $\{\mathcal{A}_o \in (\mathbb{R}^d)^{\otimes n_o}\}_{o \in O}$ (each symbol $o \in O$ is associated with a tensor whose order is the arity of o). It computes a function that maps any graph on O to a real value.*

A GWM maps each graph $G = (V, E, \ell)$ on O to a real number by first taking the tensor product of the tensors associated with all the vertices of G, and then performing contractions according to the edges in G. More formally, the computation of a GWM can be described as follows:

- construct the tensor $\mathcal{M} = \bigotimes_{v \in V} \mathcal{A}_{\ell(v)}$ (each mode of $\mathcal{M}$ corresponds to a port of G),
- obtain a tensor $\mathcal{M}'$ by permuting the modes of $\mathcal{M}$ in such a way that for each edge e of G, the two modes corresponding to the ports connected by e are following each other,
- compute the final value with

$$f(G) = \sum_{i_1=1}^{d} \sum_{i_2=1}^{d} \cdots \sum_{i_{|E|}=1}^{d} \mathcal{M}'_{i_1, i_1, i_2, i_2, \cdots, i_{|E|}, i_{|E|}}$$

One can check that GWMs are a direct generalization of *weighted automata* on strings and trees.

2.2 Polarized Graphs, Covering-Free Families and Density

It can easily be shown that the set of probability distributions that can be computed by GWMs is not dense in the space of probability distributions over graphs:

Example 1. *Let $O = \{a, b\}$ with $n_a = n_b = 1$; it is easy to check that only three connected graphs (up to isomorphisms) can be built on this ranked alphabet: G_1 (resp. G_2) with two vertices labeled by a (resp. b) and G_3 with one vertex labeled by a and the other by b.*

Furthermore a GWM with d states over O is given by two d-dimensional vectors $\mathbf{a} = \mathcal{A}_a$ and $\mathbf{b} = \mathcal{A}_b$. We have $f(G_1) = \mathbf{a}^\top \mathbf{a}$, $f(G_2) = \mathbf{b}^\top \mathbf{b}$ and $f(G_3) = \mathbf{a}^\top \mathbf{b}$, thus $f(G_3)^2 < f(G_1) f(G_2)$ for any function f computed by a GWM.

Even though one cannot hope to obtain density results for GWMs defined over arbitrary family of graphs, it has been shown in [3, 8] that such results can be obtained when one considers *covering free*

families of *polarized* graphs. The notion of polarized graph echoes the traditional notion of directed graph for the definition of graphs we consider here, while covering is a fundamental graph theoretical notion of local isomorphism.

Definition 2. *A family $\mathcal{S}$ of graphs over O is* polarized *if the ports of all symbols $o \in O$ can be partitioned in positive and negative ports in such a way that the edges of any graph in $\mathcal{S}$ only connects ports with opposite signs*[1].

Definition 3. *Let $G = (V, E, \ell)$ and $G' = (V', E'; \ell')$ be two graphs. One says that G is a* covering *of G' if there exists a mapping $\phi : V \mapsto V'$ such that (i) $\ell(v) = \ell'(\phi(v))$ for any $v \in V$ and (ii) for any edge $\{(u, i), (v, j)\} \in E$ of G: $\{(\phi(u), i), (\phi(v), j)\} \in E'$.*

We say that a family $\mathcal{S}$ of graphs is covering free *if for any $G \in \mathcal{S}$ the only coverings of G in $\mathcal{S}$ are graphs that are isomorphic to G.*

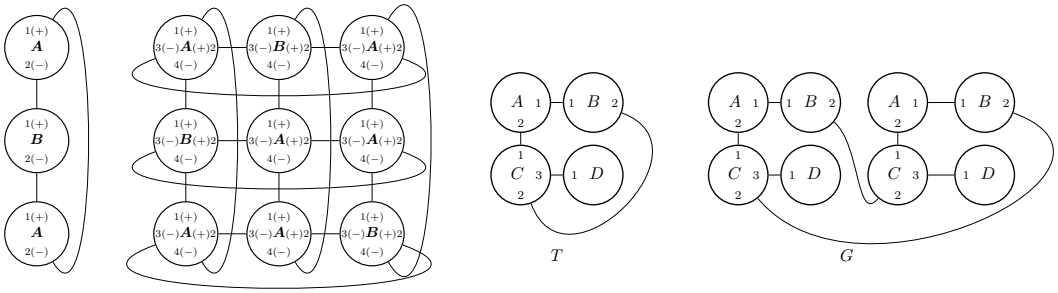

Figure 1: (left) Two polarized graphs. (right) The graph G is a covering of the graph T.

3 Learning Graph Weighted Models

For sake of clarity, we consider that all symbols $o \in O$ have the same arity r. Let $\mathcal{S}$ be a family of graphs on the ranked alphabet O and let $A = \{\mathcal{A}_o\}_{o \in O}$ be a GWM with d latent states computing the function $f^* : \mathcal{S} \to \mathbb{R}$ that we wish to infer. Observe that each tensor $\mathcal{A}_o$ is in $(\mathbb{R}^d)^{\otimes r}$.

We introduce the notion of *contexts* which are graphs with holes. A 2-context C is a graph from which an edge with its two vertices has been removed; we denote the set of all 2-contexts on O by $\mathcal{C}$. Given two symbols $o, o' \in O$ and two permutations $\sigma, \sigma' \in \mathfrak{S}_r$ of the port numbers $[r]$, we denote by $C[o_\sigma, o'_{\sigma'}]$ the graph obtained by pluging the hole in C with two vertices labeled respectively by o and o', these two vertices being connected by an edge between their $\sigma^{-1}(1)$ and $\sigma'^{-1}(1)$ ports (i.e. the first ports after permutation by σ and σ' are connected to each other while the other ports are connected to the context C).

With any 2-context C we associate the tensor $\mathcal{T}_C \in (\mathbb{R}^d)^{\otimes 2r-2}$ obtained by performing the computation of the GWM A on C[2] and we denote by $\mathbf{T}_C \in \mathbb{R}^{d^{r-1} \times d^{r-1}}$ the corresponding matricization (where modes corresponding to the first removed vertex are mapped to rows and the others to columns). For any permutation $\sigma \in \mathfrak{S}_r$, let $\mathbf{\Pi}_\sigma$ be the corresponding permutation matrix, that is $\mathbf{\Pi}_\sigma \mathrm{vec}(\mathcal{A}_o)$ is the vectorization of the tensor $\mathcal{A}_o$ after permutation of its modes with σ. Then, one can show that

$$f^*(C[o_\sigma, o'_{\sigma'}]) = \mathrm{vec}(\mathcal{A}_o)^\top \mathbf{\Pi}_\sigma^\top (\mathbf{I}_d \otimes \mathbf{T}_C) \mathbf{\Pi}_{\sigma'} \mathrm{vec}(\mathcal{A}_{o'})$$

for any 2-context C, symbols $o, o' \in O$ and permutations $\sigma, \sigma' \in \mathfrak{S}_r$. Observe that the identity matrix $\mathbf{I}_d$ in the product above corresponds to the contraction operation performed by the GWM A for the edge between the two vertices plugged in C.

Furthermore, let $\mathfrak{O} \in \mathbb{R}^{O \times O \times \mathcal{C} \times \mathfrak{S}_r \times \mathfrak{S}_r}$ be the 5th order tensor defined by

$$\mathfrak{O}_{o,o',C,\sigma,\sigma'} = f^*(C[o_\sigma, o'_{\sigma'}])$$

[1] Observe that in the example above (used to show that GWMs are not dense when defined over arbitrary families of graphs) the graphs G_1 and G_2 cannot be polarized.

[2] In this case we obtain a tensor rather than a real value since not all the modes of the tensor $\mathfrak{M}'$ are contracted during the computation.

and let $\mathbf{A} \in \mathbb{R}^{O \times d^r}$ be the matrix whose rows are the vectorizations of the tensors $\mathcal{A}_o$, i.e. $\mathbf{A}_{o,:} = \mathrm{vec}(\mathcal{A}_o)$ for each $o \in O$. Then one can check that for any 2-context C and permutations σ, σ', the matrix $\mathfrak{O}_{:,:,C,\sigma,\sigma'}$ can be factorized as

$$\mathfrak{O}_{:,:,C,\sigma,\sigma'} = \mathbf{A}(\mathbf{\Pi}_\sigma^\top(\mathbf{I}_d \otimes \mathbf{T}_C)\mathbf{\Pi}_{\sigma'})\mathbf{A}^\top.$$

As shown in the following theorem, it turns out that finding a matrix $\hat{\mathbf{A}} \in \mathbb{R}^{O \times d^r}$ satisfying these factorizations of the tensor $\mathfrak{O}$ is sufficient to recover a GWM computing the target function f^*. Similarly to the classical spectral method for weighted automata that relies on the redundancy of the information stored in the Hankel matrix, our method relies on the fact that the tensor $\mathfrak{O}$ is a highly redundant representation of the function f^*.

Theorem 1. *Let $\mathfrak{O} \in \mathbb{R}^{O \times O \times \mathcal{C} \times \mathfrak{S}_r \times \mathfrak{S}_r}$ be the 5th order tensor defined above. Suppose that for any 2-context $C \in \mathcal{C}$ and permutations $\sigma, \sigma' \in \mathfrak{S}_r$ we have the factorization*

$$\mathfrak{O}_{:,:,C,\sigma,\sigma'} = \hat{\mathbf{A}}(\mathbf{\Pi}_\sigma^\top(\mathbf{I}_d \otimes \mathbf{T}_C)\mathbf{\Pi}_{\sigma'})\hat{\mathbf{A}}^\top \tag{1}$$

for some matrix $\hat{\mathbf{A}} \in \mathbb{R}^{O \times d^r}$ and matrices $\mathbf{T}_C \in \mathbb{R}^{d^{r-1} \times d^{r-1}}$ for $C \in \mathcal{C}$.

Then, under the condition that $|O| \geq d^r$*, the GWM $\hat{A} = \{\hat{\mathcal{A}}_o\}_{o \in O}$ defined by the relation $\mathrm{vec}(\hat{\mathcal{A}}_o) = (\hat{\mathbf{A}}_{o,:})^\top$ for all $o \in O$ computes a mapping $\hat{f}$ which is equal, up to a normalization factor Z, to the target function f^* (i.e $\exists Z : \forall G, \hat{f}(G) = Z f^*(G)$).*

(sketch of proof). We will say that a matrix $\mathbf{M}$ can be factorized (to the left) by a matrix $\mathbf{X}$ if there exists a matrix $\mathbf{Y}$ such that $\mathbf{M} = \mathbf{X}\mathbf{Y}$. Let $G = (V, E)$ be a graph structure with N vertices $v_1, \cdots, v_n$, without any labeling of the vertices. We consider the tensor $\mathcal{H}_G \in (\mathbb{R}^O)^{\otimes N}$ that gathers the values of f^* for all possible labelings of G, i.e. $(\mathcal{H}_G)_{o_1,\cdots,o_N} = f^*(G[o_1, \cdots, o_N])$ where $G[o_1, \cdots, o_N]$ is the graph obtained using the labeling $v_i \mapsto o_i$. Since any matricization $(\mathbf{H}_G)_{(e)}$ of $\mathcal{H}_G$ along two modes connected by an edge e in G is a part of the tensor $\mathfrak{O}$, one can check that $(\mathbf{H}_G)_e$ can be factorized to the left by $(\hat{\mathbf{A}} \otimes \hat{\mathbf{A}})(\mathbf{I}_{d^{r-1}} \otimes \mathrm{vec}(\mathbf{I}_{d^2}) \otimes \mathbf{I}_{d^{r-1}})$ (which follows from Eq. (1)), where we assumed for sake of simplicity that the edge e connects the last port of the first vertex to the first port of the second one (hence the vector $\mathrm{vec}(\mathbf{I}_{d^2})$ performs the contraction corresponding to the edge e). Similarly, one can check that this implies that any matricization $(\mathbf{H}_G)_{(e,e')}$ of $\mathcal{H}_G$ along three modes connected by two edges e, e' in G, can be factorized both by $(\hat{\mathbf{A}} \otimes \hat{\mathbf{A}} \otimes \hat{\mathbf{A}})(\mathbf{I}_{d^{r-1}} \otimes \mathrm{vec}(\mathbf{I}_{d^2}) \otimes \mathbf{I}_{d^{r-1}} \otimes \mathbf{I}_{d^r})$ and $(\hat{\mathbf{A}} \otimes \hat{\mathbf{A}} \otimes \hat{\mathbf{A}})(\mathbf{I}_{d^r} \otimes \mathbf{I}_{d^{r-1}} \otimes \mathrm{vec}(\mathbf{I}_{d^2}) \otimes \mathbf{I}_{d^{r-1}})$ (where we assumed again that the last port of a vertex is connected to the first port of the following one). It can then be shown (under the condition $|O| \geq d^r$) that $(\mathbf{H}_G)_{(e,e')}$ can be factorized by $(\hat{\mathbf{A}} \otimes \hat{\mathbf{A}} \otimes \hat{\mathbf{A}})(\mathbf{I}_{d^{r-1}} \otimes \mathrm{vec}(\mathbf{I}_{d^2}) \otimes \mathbf{I}_{d^{r-2}} \otimes \mathrm{vec}(\mathbf{I}_{d^2}) \otimes \mathbf{I}_{d^{r-1}})$. By induction, the vector $\mathrm{vec}(\mathcal{H}_G)$ can be factorized by $(\hat{\mathbf{A}}^{\otimes N})\mathbf{v}_E$ where $\mathbf{v}_E$ is a vector that represents all the contraction operations performed by a GWM on the graph structure G; that is $(\mathcal{H}_G)_{o_1,\ldots,o_N} = f^*(G[o_1, \ldots, o_N])$ is equal to $(\hat{\mathbf{A}}_{o_1,:} \otimes \hat{\mathbf{A}}_{o_2,:} \otimes \cdots \otimes \hat{\mathbf{A}}_{o_N,:})\mathbf{v}_E = \hat{f}(G[o_1, \ldots, o_N])$ for any $o_1, \cdots, o_N \in O$ up to a normalization factor. □

As in the original spectral algorithm for weighted automata, one can consider a finite number of contexts, as long as they correspond to *sufficient statistics*, i.e. there are enough statistics to identify the distribution. Moreover, the linearity of the problem allows one to sum over the contexts, hence to consider generalized contexts (e.g. specifying only immediate neighborhoods). In practice, the condition $|O| \geq d^r$ can be relaxed provided a correct regularization. Experiments also show that in the case of a noisy observation tensor (e.g. empirical distribution deduced from an i.i.d. sample), the error is linearly dependent on the noise magnitude, which is what is expected.

For the more general case where not all symbols have the same arity, the problem can be solved by considering simultaneous tensor factorizations.

4 Conclusion

In this paper, we provide a way to estimate the parameters of a probability distribution defined over graphs. We show that the estimation problem can be reduced to a particular tensor factorization problem. The next steps include finding efficient methods to perform this factorization and explore different types of regularization to enhance the stability of the model.

References

[1] Anima Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models (A survey for ALT). In *Algorithmic Learning Theory - 26th International Conference, ALT 2015, Banff, AB, Canada, October 4-6, 2015, Proceedings*, pages 19–38, 2015.

[2] Anima Anandkumar, Prateek Jain, Yang Shi, and U. N. Niranjan. Tensor vs. matrix methods: Robust tensor decomposition under block sparse perturbations. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics, AISTATS 2016, Cadiz, Spain, May 9-11, 2016*, pages 268–276, 2016.

[3] Raphaël Bailly, François Denis, and Guillaume Rabusseau. Recognizable series on hypergraphs. *CoRR*, abs/1404.7533, 2014.

[4] Raphaël Bailly, François Denis, and Liva Ralaivola. Grammatical inference as a principal component analysis problem. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, pages 33–40, 2009.

[5] R. A. Harshman. Models for analysis of asymmetrical relationships among n objects or stimuli. *First Joint Meeting of the Psychometric Society and the Society for Mathematical Psychology, McMaster University, Hamilton, Ontario, August, 1978.*

[6] Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. In *COLT 2009 - The 22nd Conference on Learning Theory, Montreal, Quebec, Canada, June 18-21, 2009*, 2009.

[7] Maximilian Nickel. *Tensor factorization for relational learning*. PhD thesis, Ludwig Maximilians University Munich, 2013.

[8] Guillaume Rabusseau. *A Tensor Perspective on Weighted Automata, Low-Rank Regression and Algebraic Mixtures*. PhD thesis, Aix-Marseille Université, 2016.