Towards a Novel Probabilistic Graphical Model of Sequential Data: Fundamental Notions and a Solution to the Problem of Parameter Learning

Edmondo Trentin and Marco Bongini

Dipartimento di Ingegneria dell'Informazione
Università degli Studi di Siena, Siena, Italy
{trentin,bongini}@dii.unisi.it

Abstract. Probabilistic graphical modeling via Hybrid Random Fields (HRFs) was introduced recently, and shown to improve over Bayesian Networks (BNs) and Markov Random Fields (MRFs) in terms of computational efficiency and modeling capabilities (namely, HRFs subsume BNs and MRFs). As in traditional graphical models, HRFs express a joint distribution over a fixed collection of random variables. This paper introduces the major definitions of a proper dynamic extension of regular HRFs (including latent variables), aimed at modeling arbitrary-length sequences of sets of (time-dependent) random variables under Markov assumptions. Suitable maximum pseudo-likelihood algorithms for learning the parameters of the model from data are then developed. The resulting learning machine is expected to fit scenarios whose nature involves discovering the stochastic (in)dependencies amongst the random variables, and the corresponding variations over time.

Keywords: Probabilistic graphical model, Hidden Markov model, Hybrid Random Field, Sequence Classification.

1 Introduction

Probabilistic graphical models [9] have long been one of the hot topics in machine learning. The most popular instances are represented by directed graphical models, or Bayesian Networks (BNs) [10], and by undirected graphical models, namely Markov Random Fields (MRFs) [7]. Albeit intriguing and long-studied, BNs and MRFs present some drawbacks due to their very mathematical nature, and to certain limitations of the corresponding training algorithms [5]. In particular, the class $\mathcal{B}$ of (in)dependence structures that can be modeled via BNs, and the class $\mathcal{M}$ of (in)dependence structures that can be modeled via MRFs are such that $\mathcal{B} \cap \mathcal{M} \neq \varnothing$, $\mathcal{B} \not\subseteq \mathcal{M}$, and $\mathcal{M} \not\subseteq \mathcal{B}$. In other words, there are (in)dependence structures over any given set of random variables which can be modeled via BNs but not via MRFs, and vice-versa. Furthermore (and, possibly even more relevant in the computer science perspective), established learning and inference algorithms for BNs and MRFs present high computational complexity as the number of variables increases. Incidentally, no learning algorithm for

N. Mana, F. Schwenker, and E. Trentin (Eds.): ANNPR 2012, LNAI 7477, pp. 72–81, 2012.

MRFs has qualified as a standard reference so far. These are some of the reasons why real-world applications of BNs and MRFs have been somewhat limited to date. In [3], a new probabilistic graphical model was introduced with the aim of overcoming such drawbacks of traditional paradigms. The model, known as the Hybrid Random Field (HRF), was proven to subsume the modeling capabilities of both BNs and MRFs, meaning that the class $\mathcal{H}$ of (in)dependence structures that can be modeled via HRFs is such that $\mathcal{B} \subset \mathcal{H}$ and $\mathcal{M} \subset \mathcal{H}$ [2]. Moreover, HRFs come with clearly defined, efficient learning algorithms, that are empirically shown to yield models that (i) are at least as good as those obtained via BNs or MRFs, and that (ii) reduce the computational burden of learning (w.r.t. BNs and MRFs) to a dramatic extent [4].

HRFs, as well as traditional probabilistic graphical models (in particular, BNs and MRFs), assume a fixed set of random variables (whose joint probability distribution is actually modeled). Nonetheless, it is a fact that a number of real-world applications involve time-varying phenomena, presenting themselves in the form of (long, and often variable-length) sequences of time-dependent outcomes of a given set of random variables. Examples include a variety of problems in acoustic/speech processing (e.g., speech recognition, speaker identification, word-spotting, emotion recognition, etc.), video processing, bioinformatics (prediction of secondary and tertiary structure from observation of the primary structure of sequences of amino-acids, inference of functional properties from the primary structure, a variety of tasks in genomics and proteomics, etc.), natural language/document processing, etc. For this reason, interest in the development of dynamic extensions (i.e., suitable for sequential data) of the standard graphical models began to flourish in the scientific community. The most popular approach is represented by the dynamic BNs, proposed in [6]. It is noteworthy that standard hidden Markov models (HMMs) [11], which have long been applied to several amongst the aforementioned scenarios, can be shown to be a particular case of dynamic BNs [2]. Dynamic extensions of MRFs can be conceived, as well, although such extensions are not popular for the time being, and no training/inference algorithm has qualified as a standard so far. An alternative approach to the problem is the conditional random field (CRF) [8] which, under slightly different assumptions, found positive application to some of the tasks at hand, especially natural language and document processing.

The goal of the paper lies in attempting a first, formal definition of a proper dynamic extension of the standard HRF. As seen shortly, the definition involves latent random variables (in addition to the observable quantities) allowing for the modeling of (arbitrary length) sequences of sets of (time-dependent) random variables under Markov assumptions. Consequently, the resulting model is suitable for the recognition of sequential patterns as well, relying on the same classification strategy used in HRFs [2].

Suitable maximum pseudo-likelihood algorithms for learning (e.g., estimating the parameters of the resulting model) from a data sample are proposed, too. The model is expected to inherit the nice theoretical and computational properties of its static counterpart (the regular HRF), in particular as regards the positive

comparison with respect to dynamic BNs and MRFs. In so doing, the practitioner is provided eventually with an alternative, effective tool for facing application tasks whose nature involves discovering the stochastic (in)dependencies amongst random variables and the corresponding variations over time. Although no experiments are presented herein, the companion paper [1] reports on comparative simulation results which corroborate the expectations empirically. The topic of structure learning in the novel graphical model is covered in [1], too.

The remains of the paper stretch out according to the following structure. Section 2 states the formal definition of the proposed model, by referencing to the original definition of HRF, and draws some preliminary conclusions on its modeling capabilities. Section 3 contextualizes properly the problem of learning from data, outlining the fundamental quantities involved in the calculations required in order to develop learning algorithms and the basic recursive scheme used in the following developments. Section 3.1 relies on these notions for coming up with a viable algorithmic solution to the problem of learning the parameters characterizing the different probabilistic distributions which constitute the model. Finally, Section 4 draws some preliminary conclusions.

2 Definitions

A broad-sense definition of HRFs is given in [2]. The reader is referred to the latter for all basic concepts of standard HRFs which are relevant to this paper. The modularity of an HRF is the property of factorizing the overall joint probability of its random variables in terms of a product of local probabilistic quantities defined at the level of the individual BNs embraced by the very definition of HRF. Noteworthily, modularity can be deduced as a property holding for HRFs defined accordingly [2]. A strict-sense, simpler definition can thence be devised, which roughly goes as follows: an HRF is a collection of Bayesian networks which possess the modular property (see, e.g., [3]). In this paper, for simplicity and coherence with the proposed algorithms, the strict-sense definition of HRF is assumed (without loss of generality). That said, we can give a definition of the novel dynamic probabilistic graphical model for sequences in the following terms. The model is referred to as the dynamic hybrid random field (DHRF).

Definition: a dynamic HRF $\mathcal{DH}$ is a tuple $\mathcal{DH} = (\mathbf{X}, S, \pi, \mathcal{F}, \mathbf{a}, \mathcal{H})$ where

1. $\mathbf{X}$ is a set of (observable) random variables $X_1, \ldots, X_n$. Outcomes of the random variables depend on time $t = 1, \ldots, T$, that is we will write $X_i(t)$ whenever we need to make the dependency explicit.
2. S is a set of Q latent radom variables, $S = \{S_1, \ldots, S_Q\}$. It is assumed that sequences of such latent variables are responsible for the generation of sequences of outcomes of the observable variables, and that the variables in S can be thought of as the states of a discrete-time Markov chain (*latent Markov assumption*). We write q_t to denote the state of the Markov chain at time t for $t = 0, \ldots, T$.

3. π is a probability distribution of the initial latent variables, i.e. $\pi = \{Pr(S_i \mid t = 0), S_i \in S\}$, where t is the discrete time index. For instance, if the Markov chain over S may equally-likely start with any latent variable, then π is uniform over S. Contrariwise, if a certain S_j can never occur at time $t = 0$, then $\pi(S_j) = 0$, etc.
4. $\mathcal{F} \subseteq S$ is the set of final states, i.e. the latent variables which can legitimately generate sets of outcomes of the observable variables at time T (namely, at the end of sequences).
5. $\mathbf{a}$ is a probability distribution that characterizes the (allowed) transitions between latent variables, that is $\mathbf{a}_{ij} = \{Pr(S_j \text{ at time } t \mid S_i \text{ at time } t - 1), S_i \in S, S_j \in S\}$ where the transition probabilities $\mathbf{a}_{ij}$ are assumed to be independent of time t. Note that the definition is meaningful due to the latent Markov assumption.
6. $\mathcal{H}$ is a set of HRFs over $\mathbf{X}$, $\mathcal{H} = \{\mathcal{H}_1, \ldots, \mathcal{H}_Q\}$, where $\mathcal{H}_q$ is uniquely associated with q-th latent variable S_q such that the joint emission probability $\mathbf{b}(\mathbf{X}) = P(X_1, \ldots, X_n \mid S_q)$ is modeled via HRF $\mathcal{H}_q$ over $\mathbf{X}$, independently of time t, and we assume that the probability distribution of $\mathbf{X}(t)$ is independent of the probability of $\mathbf{X}(t')$ (for all $t' \neq t$) given the latent variable (*emission Markov assumption*). In this definition, bearing in mind the definition of HRF, it turns out that $\mathcal{H}_q$ is a set of Bayesian networks $BN_{q,1}, \ldots, BN_{q,n}$ (with directed acyclic graphs $\mathcal{G}_{q,1}, \ldots, \mathcal{G}_{q,n}$) such that:
   (a) each $BN_{q,i}$ contains X_i plus a subset $\mathcal{R}_q(X_i)$ of $\mathbf{X} \setminus \{X_i\}$, namely the set of relatives of X_i in $BN_{q,i}$;
   (b) for each X_i, $P(X_i|\mathbf{X} \setminus \{X_i, q\}) = P(X_i|\mathcal{MB}_{q,i}(X_i))$, where $\mathcal{MB}_{q,i}(X_i)$ is the set containing the parents, the children, and the parents of the children of X_i in $\mathcal{G}_{q,i}$ (namely, the Markov blanket of X_i in $BN_{q,i}$).
   The Markov assumption holding in HRFs is referred to as the *observable Markov assumption* in the present framework.

Note that the overall DHRF can be thought of as a probabilistic graphical model over the set of random variables $S \cup \mathbf{X}$. Nonetheless, this definition allows for separate sets of BNs (i.e., different HRFs) for each latent variable, meaning that it does not extend regular HRFs to sequences by defining them as sets of dynamic Bayesian networks in a straightforward manner.

This definition is flexible and provides us with useful and efficient algorithmic tools rooted in traditional hidden Markov models (HMM). Before developing an algorithm for parameter learning in DHRFs, it is noteworthy to observe some fundamental properties regarding the modeling capabilities of DHRFs. Let $\mathcal{D}$ be the class of (in)dependence structures that can be represented via DHRFs. First of all, a DHRF with a single latent variable reduces implicitly to a regular HRF. Thence, $\mathcal{H} \subset \mathcal{D}$. Furthermore, since $\mathcal{B} \subset \mathcal{H}$ and $\mathcal{M} \subset \mathcal{H}$ (as we pointed out in Section 1), we have also $\mathcal{B} \subset \mathcal{D}$ and $\mathcal{M} \subset \mathcal{D}$. Given the fact that dynamic BNs (hence, including standard HMMs) are specialized instances of BNs, an immediate consequence of the reasoning is that DHRFs subsume the modeling capability of dynamic BNs (and, of HMMs). Following similar arguments, it is immediately seen that DHRFs subsume also any dynamic extensions (under Markov assumption) of Markov random fields.

3 Parameter Learning

Let $\mathcal{DH}$ be a DHRF and let $O = O_1, O_2, ..., O_T$ be a training sequence of outcomes of the observable variables, i.e. $O_t = (x_1, \ldots, x_n)$ for $t = 1, \ldots, T$. Training algorithms can be devised by exploiting a recursive scheme which is similar (to some extent) to the popular forward-backward procedures for HMMs [11].

The training criterion function is the (pseudo-)likelihood $P^*(O|\mathcal{DH})$, as in regular probabilistic graphical models (see [2] for a justification of why the pseudo-likelihood is used instead of the bare likelihood criterion). We define the (pseudo) forward terms

$$\alpha_t(i) = P(O_1, ..., O_t, q_t = S_i|\mathcal{DH}) \tag{1}$$

which can be recursively computed as

$$\alpha_t(i) = b_{i,t} \sum_j a_{ji} \alpha_{t-1}(j) \tag{2}$$

where $b_{i,t}$ denotes the emission probability of observation O_t given i-th latent variable. We say that the *forward step* of the following algorithms is the recursive computation of the α's according to Equation 2. Note that $P(O|\mathcal{DH}) = \sum_{i=1}^{Q} \alpha_T(i)$.

Also, the (pseudo) backward terms are defined as:

$$\beta_t(j) = P(O_{t+1}, ..., O_T|q_t = S_j, \mathcal{DH}) \tag{3}$$

which are recursively computed as

– initialization:

$$\beta_T(i) = \begin{cases} 1 \text{ if } S_i \in \mathcal{F} \\ 0 \text{ otherwise.} \end{cases} \tag{4}$$

– recursion:

$$\beta_t(j) = \sum_{i=1}^{Q} a_{ji} b_i(O_{t+1}) \beta_{t+1}(i) \text{ for } t = T-1, T-2, ..., 1 \text{ and } 1 \leq j \leq Q \tag{5}$$

We refer to this recursive computation as the *backward step* of the following algorithms. We then define the quantity

$$\gamma_t(i) = P(q_t = S_i|O, \mathcal{DH}) \tag{6}$$

and we can write

$$\gamma_t(i) = \frac{P(q_t = S_i, O|\mathcal{DH})}{P(O|\mathcal{DH})} = \frac{P(O_1, ..., O_t, q_t = S_i, O_{t+1}, ..., O_T|\mathcal{DH})}{P(O|\mathcal{DH})} \tag{7}$$

$$= \frac{P(O_1, ..., O_t, q_t = S_i|\mathcal{DH}) P(O_{t+1}, ..., O_T|O_1, ..., O_t, q_t = S_i, \mathcal{DH})}{P(O|\mathcal{DH})} \tag{8}$$

$$= \frac{P(O_1, ..., O_t, q_t = S_i|\mathcal{DH})P(O_{t+1}, ..., O_T|q_t = S_i, \mathcal{DH})}{P(O|\mathcal{DH})} \tag{9}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{Q} P(q_t = S_j, O|\mathcal{DH})} \tag{10}$$

$$= \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{Q} \alpha_t(j)\beta_t(j)} \tag{11}$$

where the emission Markov assumption was used to rewrite eq. 8 in the form of eq. 9. Put into words, $\gamma_t(i)$ can thus be calculated from the α's and β's during the backward step. Finally, we define:

$$\xi_t(i, j) = P(q_t = S_i, q_{t+1} = S_j \mid O, \mathcal{DH}) \tag{12}$$

i.e.

$$\xi_t(i, j) = \frac{P(q_t = S_i, q_{t+1} = S_j, O \mid \mathcal{DH})}{P(O|\mathcal{DH})} \tag{13}$$

$$= \frac{P(q_t = S_i, q_{t+1} = S_j, O|\mathcal{DH})}{\sum_{i=1}^{Q}\sum_{j=1}^{Q} P(q_t = S_i, q_{t+1} = S_j, O|\mathcal{DH})} \tag{14}$$

$$= \frac{\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)}{\sum_{i=1}^{Q}\sum_{j=1}^{Q} \alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j)} \tag{15}$$

which is computed during the backward step for $t = T, T-1, ..., 1$, where $1 \leq i, j \leq Q$.

It is seen that the following properties hold true:

1. $\gamma_t(i) = \sum_{j=1}^{Q} \xi_t(i, j)$
2. $\sum_{t=1}^{T} \gamma_t(i)$ represents the expected number of instances of the latent variable S_i during the generation of the observed sequence O
3. $\sum_{t=1}^{T-1} \gamma_t(i)$ is the expected number of transitions starting from variable S_i (to any other variable)
4. $\sum_{t=1}^{T-1} \xi_t(i, j)$ is the expected number of transition that occur from S_i to S_j

Thus, an algorithm for re-estimating the probabilistic quantities involved in the definition of the DHRF $\mathcal{DH}$ and yielding a new (more accurate) DHRF $\widehat{\mathcal{DH}}$ involves the following formulas:

$\widehat{\pi}_i = \gamma_1(i)$
$\widehat{a}'_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$ (from properties 3 and 4)

which yield the re-estimation of initial and transition probabilities. The calculations accomplished so far underly the structure learning algorithm presented in the companion paper [1], as well. As regards re-estimation of the HRFs within the DHRF, an ad-hoc algorithm for parameter learning is covered in the next section.

3.1 Learning the Parameters of the HRFs within the DHRF

In this Section we propose a solution to the problem of learning the conditional probability tables (CPTs) [2] of the Bayesian networks modeling the local conditional distributions for each HRF $\mathcal{H}_1, \ldots, \mathcal{H}_Q$ associated with the latent variables. That is, for each latent variable $q = 1, \ldots, Q$ and for each observable variable X_i, the task is to learn the parameters of the conditional distribution $P(X_i|q, mb_{q,i}(X_i))$, for each state $mb_{q,i}$ of the variables in $\mathcal{MB}_{q,i}$, where $\mathcal{MB}_{q,i}$ is the Markov blanket [2] (in $\mathcal{H}_q$) for i-th observable variable and q-th latent variable. This requires that the structure of $\mathcal{H}_q$ has been previously fixed, i.e. that the directed acyclic graph (DAG) $\mathcal{G}_{q,i}$ associated in $\mathcal{H}_q$ with each variable X_i has been specified (the issue of learning an adaptive structure which is not fixed and pre-defined is covered in the companion paper [1]). In order to learn the parameters of each $BN_{q,i}$ from the training sequence $O = O_1, O_2, ..., O_T$, we use the technique described below. In order to denote the parents of node X_i in $\mathcal{H}_q$ we will use the notation $\mathcal{PA}_q(X_i)$, rather than using $\mathcal{PA}_{q,i}(X_i)$, since indexing the DAG only makes sense in the specific context of HRFs.

Let us assume that each observation O_j, $j = 1, \ldots, T$, is an n-dimensional vector $(x_{1_j}, \ldots, x_{n_j})$ of discrete values of the variables $X_1, \ldots, X_n$. The simplest case in parameter learning is the case of a variable X_i having no parents in the DAG within $\mathcal{H}_q$. In this case, we only need to estimate the absolute distribution $P(X_i|q)$. For each value x_{i_k} of X_i, our estimate reduces to the expected number of observations of x_{i_k} along the observed sequence O while being in the latent variable q, normalized by the expected number of presences in q, i.e.:

$$\widehat{P}(X_i = x_{i_k}|q) = \frac{\sum_{t, O_{t,k} = x_{i_k}} \gamma_t(q)}{\sum_t \gamma_t(q)} \tag{16}$$

where the sums over t are extended to $t = 1, \ldots, T$, $O_{t,k}$ denotes the k-th observed variable at time t, and the notation $\widehat{P}(X = x|q)$ refers to the new (improved) estimate of $P(X = x|q)$.

A more general case occurs in learning the conditional distribution of a node X_i in HRF $\mathcal{H}_q$ having parents $\mathcal{PA}_q(X_i)$. In this case, we need to estimate a distribution $P(X_i|pa_q(X_i))$ for each possible state $pa_q(X_i)$ of $\mathcal{PA}_q(X_i)$. For each value x_{i_k} of X_i, we will estimate these conditional probabilities from the training sequence O as the expected number of occurrences of outcome x_{i_k} for observable variable X_i jointly with $\mathcal{PA}_q(X_i) = pa_q(X_i)$ while being in latent variable q, normalized by the expected number of occurrences of the outcome $pa_q(X_i)$ of variables $\mathcal{PA}_q(X_i)$ while in q :

$$\widehat{P}(X_i = x_{i_k}|pa_q(X_i)) = \frac{\sum_{t, O_{t,k} = x_{i_k}, \mathcal{PA}_q(X_i) = pa_q(X_i)} \gamma_t(q)}{\sum_{t, \mathcal{PA}_q(X_i) = pa_q(X_i)} \gamma_t(q)} \tag{17}$$

The strategy proposed in Equations 16–17 has to be accomplished over the whole training set, i.e. not limited to an individual training sequences O (expectations need to be estimated over all training sequences), as in regular HMMs. Nonetheless, the technique suffers from the following problem. If a particular

value x_{i_k} of X_i is never observed in the training set, or if it is never observed together with a particular configuration $pa_q(X_i)$ of $\mathcal{PA}_q(X_i)$, then our estimate of $P(X_i = x_{i_k}|q)$ (or of $P(X_i = x_{i_k}|q, pa_q(X_i))$) will be zero. This result is not acceptable in all cases where any event is possible, i.e. where every state of the network can be observed in principle. A solution for this difficulty appeals to the notion of an *equivalent state-occurrence expectation*, which we denote by $N^{(q)}$. It is the expected number of occurrences of latent variable S_q over a theoretical data sample which we assume to have been observed before the actual dataset. In other words, it is the expected number of occurrences of S_q over a *prior* sample. Within this prior sample, we assume to have observed any particular value x_{i_k} of X_i while in state q for a number of times equal to $p^{(q)} \cdot N^{(q)}$, where p stands for the prior probability that X_i has value x_{i_k} while in latent variable q.

Going back to Equations 16–17, we revise them as follows:

$$\widehat{P}(X_i = x_{i_k}|q) = \frac{\sum_{t,O_{t,k}=x_{i_k}} \gamma_t(q) + p_{i_k}^{(q)} N^{(q)}}{\sum_t \gamma_t(q) + N^{(q)}} \tag{18}$$

$$\widehat{P}(X_i = x_{i_k}|pa_q(X_i)) = \frac{\sum_{t,O_{t,k}=x_{i_k},\mathcal{PA}_q(X_i)=pa_q(X_i)} \gamma_t(q) + p_{i_k}^{(q)} N_{pa_i}^{(q)}}{\sum_{t,\mathcal{PA}_q(X_i)=pa_q(X_i)} \gamma_t(q) + N_{pa_i}^{(q)}} \tag{19}$$

where $N_{pa_i}^{(q)}$ is a parameter related to $N^{(q)}$ in a way we will explain shortly. An important question concerning Equations 18–19 is what values we choose for the parameters $p_{i_k}^{(q)}$, $N^{(q)}$, and $N_{pa_i}^{(q)}$. In typical applications we assign uniform prior probabilities to the different values of each variable. Therefore, our choice for $p_{i_k}^{(q)}$ will be the following:

$$p_{i_k}^{(q)} = \frac{1}{|\mathcal{D}_i|} \tag{20}$$

where $\mathcal{D}_i$ is the domain of variable X_i. The value we assign to $N^{(q)}$ is instead:

$$N^{(q)} = \max_{1 \le i \le n} |\mathcal{D}_i| \tag{21}$$

An intuitive justification of Equation 21 appeals to two different aims. On the one hand, we want to keep the equivalent state-occurrence expectation as small as possible, so as to prevent prior probabilities from biasing learning too heavily. On the other hand, we want the equivalent state-occurrence expectation to be large enough to contain at least one occurrence for all values of each variable. Therefore, the choice made in Equation 21 seems to be an optimal trade-off. Given $N^{(q)}$, we define $N_{pa_i}^{(q)}$ as follows:

$$N_{pa_i}^{(q)} = \frac{N^{(q)}}{|\mathcal{D}_{\mathcal{PA}_{q,i}}|} \tag{22}$$

where $\mathcal{D}_{\mathcal{PA}_{q,i}}$ is the set of all possible states of $\mathcal{PA}_q(X_i)$. As a result of Equation 22, each value x_{i_k} of a non-root node X_i is expected to be observed $\frac{N^{(q)}}{|\mathcal{D}_i|}$ times

within the prior (i.e., equivalent) sample, where $\frac{N^{(q)}}{|\mathcal{D}_i|} \geq 1$. To realize why $\frac{N^{(q)}}{|\mathcal{D}_i|} \geq 1$, we have to keep in mind that x_{i_k} is expected to be observed $p_{i_k}^{(q)} N_{pa_i}^{(q)}$ times for each possible state $pa_q(X_i)$ of $\mathcal{PA}_q(X_i)$. In other words, we have that $\frac{N^{(q)}}{|\mathcal{D}_i|} = p_{i_k}^{(q)} N_{pa_{q,i}}^{(q)} \cdot |\mathcal{D}_{\mathcal{PA}_{q,i}}|$.

4 Conclusions

As we said, probabilistic graphical models are a flexible, intriguing branch of machine learning. Traditional paradigms, being defined over fixed sets of random variables, are suitable for modeling joint distributions within "static" scenarios. A number of applications of the utmost interest, on the other hand, involve sequential data. This requires the capability of modeling time-varying stochastic (in)dependencies amongst random variables. The paper introduced a novel probabilistic graphical model for the modeling of sequences of random variables. Basically, its definition involves an underlying HMM structure combined with state-specific HRFs. This formulation is quite general under the Markov assumption, and subsumes dynamic Bayesian networks (including the traditional HMM itself) and any dynamic extensions of Markov random fields (provided that the time-dependencies satisfy Markov assumptions). An algorithm for learning the parameters of DHRFs was given, as well. The companion paper [1] focuses on the development of an algorithm for learning the structure of a DHRF, and presents empirical evidence (in the form of computer simulations) which corroborate the expectation that originally motivated the development of the present framework, that is: (i) having a dynamic graphical model which subsumes the capabilities of modeling (in)dependence structures offered by dynamic BNs and MRFs, whilst (ii) reducing their computational burden to a dramatic extent.

Acknowledgments. The authors would like to thank Antonino Freno for his invaluable support and the many stimulating discussions.

References

1. Bongini, M., Trentin, E.: Towards a Novel Probabilistic Graphical Model of Sequential Data: A Solution to the Problem of Structure Learning and an Empirical Evaluation. In: Mana, N., Schwenker, F., Trentin, E. (eds.) ANNPR 2012. LNCS (LNAI), vol. 7477, pp. 82–92. Springer, Heidelberg (2012)
2. Freno, A., Trentin, E.: Hybrid Random Fields: A Scalable Approach to Structure and Parameter Learning in Probabilistic Graphical Models. Springer (2011)
3. Freno, A., Trentin, E., Gori, M.: A Hybrid Random Field Model for Scalable Statistical Learning. Neural Networks 22, 603–613 (2009)
4. Freno, A., Trentin, E., Gori, M.: Scalable Pseudo-Likelihood Estimation in Hybrid Random Fields. In: Elder, J.F., Fogelman-Souli, F., Flach, P., Zaki, M. (eds.) Proceedings of the 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2009), pp. 319–327. ACM (2009)

5. Freno, A., Trentin, E., Gori, M.: Scalable Statistical Learning: A Modular Bayesian/Markov Network Approach. In: Proceedings of the International Joint Conference on Neural Networks (IJCNN 2009), pp. 890–897. IEEE (2009)
6. Ghahramani, Z.: Learning Dynamic Bayesian Networks. In: Giles, C.L., Gori, M. (eds.) IIASS-EMFCSC-School 1997. LNCS (LNAI), vol. 1387, pp. 168–197. Springer, Heidelberg (1998)
7. Kindermann, R., Laurie Snell, J.: Markov Random Fields and Their Applications. American Mathematical Society, Providence (1980)
8. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proc. 18th International Conf. on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2001)
9. Lauritzen, S.L.: Graphical Models. Oxford University Press (1996)
10. Pearl, J.: Bayesian networks: A model of self-activated memory for evidential reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society, pp. 329–334. University of California, Irvine (1985)
11. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)