Predictive Random Fields: Latent Variable Models Fit by Multiway Conditional Probability with Applications to Document Analysis

Andrew McCallum, Xuerui Wang and Chris Pal
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003 USA
{mccallum, xuerui, pal}@cs.umass.edu

Technical Report UM-CS-2005-053
Version 1.0, June 1, 2005
Version 2.0, July 21, 2005
Version 2.1, July 29, 2005

July 28, 2005

Abstract

We introduce *predictive random fields*, a framework for learning undirected graphical models based *not* on joint, generative likelihood, or on conditional likelihood, but based on a product of several conditional likelihoods each relying on common sets of parameters and predicting different subsets of variables conditioned on other subsets. When applied to models with latent variables, such as the Harmonium, this approach results in powerful clustering models that combine the advantages of conditional random fields with the unsupervised clustering ability of popular topic models, such as latent Dirichlet allocation and its successors. We present new algorithms for parameter estimation based on contrastive divergence. Experimental results show significant improvement in inferring hidden document categories, and learning models of authors, words, topics and time.

1 Introduction

There has been a great deal of recent interest in generative probabilistic models for discovering latent class structure from data. For example, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and multinomial Principal Component Analysis (PCA) (Buntine, 2002) have extended older Latent Semantic Indexing (Deerwester et al., 1990) methods in ways more suitable for document data, such as integer word counts. Many variations on LDA have been created for application to images and their captions (Blei & Jordan, 2003), documents and their authors (Rosen-Zvi et al., 2004), and email messages exchanged in a social network (McCallum et al., 2005), and visual object recognition (Sivic et al., 2004).

These unsupervised directed graphical models have several disadvantages, however. Inferring the posterior distribution over the latent variables is usually intractable, and often requires slow iterative procedures. Designing the model dependency structure with directed arcs can be constraining: one must be careful to avoid creating cycles; auto-correlation is difficult to model; the implications choices about causality can be tricky. As locally-normalized generative models, they cannot robustly use arbitrary, overlapping features of the data without capturing a prohibitive number of dependencies.

Conditional Random Field (CRF) models (Lafferty et al., 2001) have recently received significant attention since they do robustly handle many non-independent features, and can represent auto-correlation. Furthermore, since CRFs are obtained by maximizing a Conditional Likelihood objective function, they often yield improved accuracy over their jointly-trained generative counterparts. However a drawback of the CRF framework is that CRF models are optimized for only one particular conditional distribution. Thus, when some of the conditioned variables are hidden or missing, or some considerations of input density or outlier detection are required, problems can arise.

This paper presents an approach that combines the advantages of richly-structured unsupervised latent variable models with the advantages of conditional random fields. Drawing from these recent results, we construct and optimize undirected models with hidden variables conditionally, but for a number of different conditional distributions. We term this a *predictive* estimation criterion, and term the model *predictive random fields* (PRF). The PRF framework addresses the problems of CRFs by maximizing multiple marginal conditional likelihood objectives defined among groups of variables.

For our experiments here we are particularly interested in using these new types of probabilistic models to capture relationships among the following attributes of documents: words, authors, time and "hidden" factors or topics within scientific research papers. As such, here we develop a PRF factorization structure taking the form of a restricted Boltzmann machine (Hinton, 2002), also described as a Harmonium in (Smolensky, 1986). We compare our approach with a traditional Maximum Likelihood Harmonium (Smolensky, 1986; Welling et al., 2005) optimized using the contrastive divergence algorithm of (Hinton, 2002). To achieve these results we also construct a new algorithm inspired by the contrastive divergence approach and use this algorithm for predictive likelihood learning.

We begin our discussion in Section 2 with a simple example involving a Gaussian mixture model with hidden sub-classes and observed classes. We then demonstrate the effects of optimizing such a model under the joint likelihood, the conditional likelihood and under a predictive likelihood. In Section 3 we briefly review exponential family models, Markov Random Fields and Conditional Random Fields. We then present our new model, Predictive Random Fields. We review restricted Boltzmann machines and the Harmonium factorization, and then develop a multi-attribute exponential family harmonium model for documents. In Section 4 we further develop our new optimization criterion, the Predictive Likelihood (PL) through contrast with Maximum (*Joint*) Likelihood (ML) and Maximum Conditional Likelihood (CL) learning. Correspondingly, we present new algorithms for CL learning with hidden variables and PL learning, extending the contrastive divergence approach of (Hinton, 2002). Finally, in Section 7 we present results illustrating the utility of the PRF approach through constructing multi-attribute harmonium models of documents and comparing the utility of the latent representations obtained under different optimization schemes. We compare the traditional model optimized under ML with treating the model as a CRF (optimizing using Maximum CL) and treating the model as a PRF (optimizing using Maximum PL).

2 Intuition and a Simple Example with a Gaussian Mixture Model

Consider a Gaussian mixture model for real valued random observed variables $\mathbf{x}$ (e.g., observed 2D values) with an unobserved sub-class, s associated with each observed class label c. We will use the notation $\tilde{\mathbf{x}}$ and $\tilde{c}$ to denote observations or instantiations of continuous and discrete random variables. We can write a model for the joint distribution of these random variables as

$$P(\mathbf{x}, c, s) = p(\mathbf{x}|s)P(s|c)P(c), \tag{1}$$

where $P(s|c)$ is a sparse matrix associating a number of sub-classes with a given class and $p(\mathbf{x}|s)$ is given by

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_s, \boldsymbol{\Sigma}_s) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_s|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_s)^T \boldsymbol{\Sigma}_s^{-1}(\mathbf{x} - \mu_s)\right) \tag{2}$$

and we shall use Θ to denote all the parameters of the model defined by (1). It is possible to optimize the model defined by (1) in a number of different ways. First, consider the log *joint likelihood* $\mathcal{L}_{\mathbf{x},c}$ of such a model, which can be expressed as:

$$\mathcal{L}_{\mathbf{x},c}(\Theta; \{\tilde{\mathbf{x}}_i\}, \{\tilde{c}_i\}) = \sum_i \log P(\tilde{\mathbf{x}}_i, \tilde{c}_i|\Theta) = \sum_i \log \sum_{s_i} P(\tilde{\mathbf{x}}_i, s_i, \tilde{c}_i|\Theta) = \mathcal{L}_{\mathbf{x},c}(\Theta) \tag{3}$$

Second, in contrast to the joint likelihood, the log *conditional likelihood* $\mathcal{L}_{c|\mathbf{x}}$ can be expressed as:

$$\begin{aligned}\mathcal{L}_{c|\mathbf{x}}(\Theta; \{\tilde{\mathbf{x}}_i\}, \{\tilde{c}_i\}) &= \sum_i \log P(\tilde{c}_i|\tilde{\mathbf{x}}_i, \Theta) = \sum_i \log \sum_{s_i} P(\tilde{c}_i, s_i|\tilde{\mathbf{x}}_i, \Theta) \\ &= \sum_i \log \sum_{s_i} P(\tilde{c}_i, s_i, \tilde{\mathbf{x}}_i|\Theta) - \sum_i \log \sum_{s_i} \sum_{c_i} P(c_i, s_i, \tilde{\mathbf{x}}_i|\Theta) \\ &= \mathcal{L}_{\mathbf{x},c}(\Theta) - \mathcal{L}_{\mathbf{x}}(\Theta)\end{aligned} \tag{4}$$

The criterion defined by (3) is of course very widely used in many communities, while the criterion of (4) has only recently begun to generate increased attention in the Machine Learning community (Jebara & Pentland, 2000; Lafferty et al., 2001) and in the statistics community (Edwards & Lauritzen, 2001). In this paper we introduce a third *class of criteria* we call *predictive likelihoods*. There are a number of possible ways to construct a predictive likelihood objective function. For the model here, one possible and intuitively useful construction encoding our desire to have a joint model with parameters set so as to be equally good at predicting real values $\mathbf{x}$ given discrete classes $\tilde{c}$ *and* predicting discrete classes c given real values $\tilde{\mathbf{x}}$ can be written as $\mathcal{L}_{c|\mathbf{x},\mathbf{x}|c}$ which we express as:

$$\begin{aligned}\mathcal{L}_{c|\mathbf{x},\mathbf{x}|c}(\Theta; \{\tilde{\mathbf{x}}_i\}, \{\tilde{c}_i\}) &= \sum_i \log P(\tilde{c}_i|\tilde{\mathbf{x}}_i, \Theta) + \sum_i \log P(\tilde{\mathbf{x}}_i|\tilde{c}_i, \Theta) \\ &= 2\mathcal{L}_{\mathbf{x},c}(\Theta) - \mathcal{L}_{\mathbf{x}}(\Theta) - \mathcal{L}_c(\Theta)\end{aligned} \tag{5}$$

2.1 Comparing Optimizations

Consider the following simple example data set which is similar to the example presented in Jebara's work (Jebara & Pentland, 2000) to illustrate his Conditional Expectation Maximization (CEM) approach. Similarly, we generate data from two classes, each with four sub-classes drawn from 2D isotropic Gaussian. The data are illustrated by red $\circ$'s and blue $\times$'s in Figures 1 and 2 . In contrast to (Jebara & Pentland, 2000), here we fit models with diagonal covariance matrices and we use the conditional expected gradient (Salakhutdinov et al., 2003) optimization approach to update parameters. To illustrate the effects of the different optimization criteria we have fit models with two subclasses for each class. We run each algorithm with 30 random initializations using gradient based optimization for the three objective functions and choose the best model under each of the metrics, ML (3), CL (4) and PL (5).

In Figure 1 we show the result of the best model under CL optimization for two different batches of 30 random initializations. These experiments illustrate the fact that there are a variety of very closely optimal solutions under this criterion. Further, the setting of the parameters optimized under the CL metric is somewhat arbitrary as there are a wide variety of settings for cluster means that will produce good conditional likelihoods.

In Figure 2 (left) we show the best model under joint likelihood optimization and (right) the best model under our predictive likelihood optimization. Importantly, under the ML objective the clusters modeling the $\circ$ data are not affected by either the $\times$ data or their associated clusters in the model and vice versa. In contrast, as illustrated in Figure 2 (right) under the PL objective we are optimizing the model so as to produce both a good distribution for predicting data vectors $\mathbf{x}$ given $\tilde{c}$ *and* a good distribution for predicting c given $\tilde{\mathbf{x}}$.

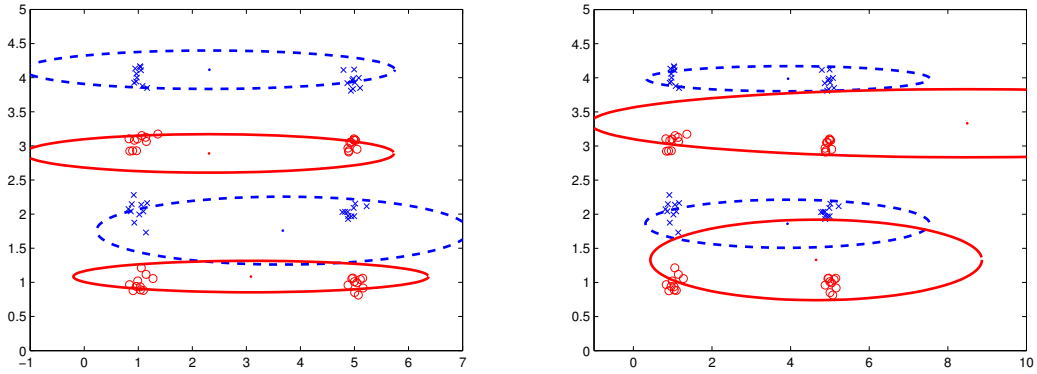

Figure 1: (left) One of the optimal solutions found by conditional likelihood optimization. (right) Another near optimal CL solution.

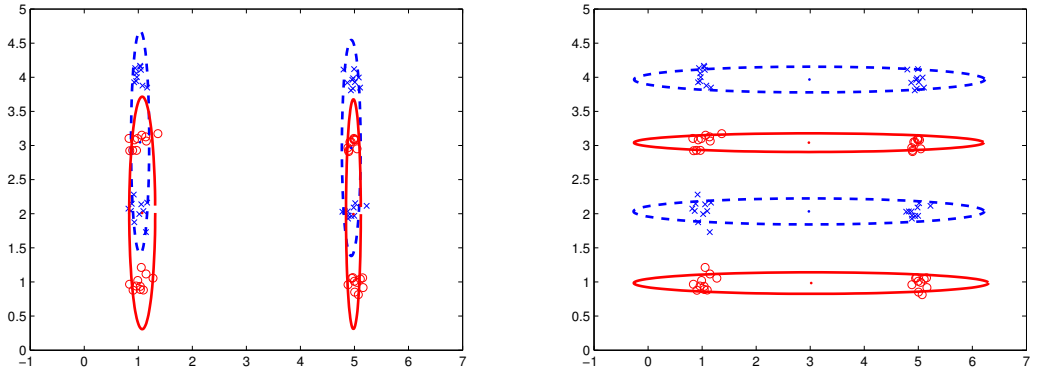

Figure 2: (left) Joint likelihood optimization. (right) Predictive likelihood optimization

3 Undirected or Globally Normalized, Factored Probability Models

3.1 Exponential Family Models, MRFs, CRFs and PRFs

A probability model is said to be an exponential family model if a density can be written as

$$P(\mathbf{x}|\boldsymbol{\theta}) = \boldsymbol{h}(\mathbf{x}) \exp(\boldsymbol{\eta}^T \boldsymbol{T}(\mathbf{x}) - A(\boldsymbol{\theta})) = \frac{1}{Z(\boldsymbol{\theta})} \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x})), \tag{6}$$

Where[1] $\exp(A(\boldsymbol{\theta})) = Z(\boldsymbol{\theta}) = \int \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x}))\mathrm{d}\mathbf{x}$, $\boldsymbol{\theta}^T = [\boldsymbol{\eta}^T \ \ \mathbf{1}^T]$ and $\boldsymbol{f}(\mathbf{x})^T = [\boldsymbol{T}(\mathbf{x})^T \ \ \log \boldsymbol{h}(\mathbf{x})^T]$. Exponential family models as expressed in (6) are equivalent to Markov Random Fields (MRFs) and can be expressed in a more standard (MRF) form, where the joint distribution of random variables is given by the normalized product of potential functions operating only on *subsets* of variables $x_s \subset \mathbf{x} = \{x_1, \ldots, x_D\}, s \in \mathcal{S}$, where s is an index for the family of subsets $\mathcal{S}$ and D is the number of variables or dimension of $\mathbf{x}$. Importantly, under this defintion, subsets could consist of a single variable or multiple variables. Accordingly $\psi_s(x_s) = \exp(\boldsymbol{\theta}_s^T \boldsymbol{f}_s(x_s))$ and $P(\mathbf{x}|\boldsymbol{\theta}) = Z(\boldsymbol{\theta})^{-1} \prod_s \psi_s(x_s)$.

[1] To compute $Z(\boldsymbol{\theta})$ for discrete $\mathbf{x}$ one can replace the integral by a summation.

Consider now partitioning variables into two groups, "observations" $\mathbf{x}$ and hidden factors or unobserved variables $\mathbf{z}$. We use these descriptions to denote that observations will always be observed $\tilde{\mathbf{x}}$ and hidden factors $\mathbf{z}$ will never be observed. It is possible to construct an exponential family model for the joint distribution of such a model with the form $P(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) = Z(\boldsymbol{\theta})^{-1} \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x}, \mathbf{z}))$. In some cases it is possible to optimize such a model by integrating out the unobserved variables and optimizing the log marginal likelihood of the observed variables. This is the procedure taken in the exponential family harmonium model of (Welling et al., 2005). Consider now partitioning variables in our exponential family model into two groups, observations or "explanatory variables" $\mathbf{x}$ and "labels" $\mathbf{y}$. It is possible to define an exponential family model for the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ as

$$P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\mathbf{x}; \boldsymbol{\theta})} \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x}, \mathbf{y})), \tag{7}$$

where $Z(\mathbf{x}; \boldsymbol{\theta}) = \int \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x}, \mathbf{y}))\mathrm{d}\mathbf{y}$. Examples of this type of model include CRFs (Lafferty et al., 2001) and graphical models in the sense of (Lauritzen & Wermuth, 1989) derived by conditioning on observations or explanatory variables in a corresponding family of joint distributions.

Often times when modeling data we are particularly concerned with modeling certain conditional distributions among subsets of variables. Consider now constructing a model such that

$$P(x_s, \mathbf{z}|x_{\bar{s}}; \boldsymbol{\theta}) = \frac{1}{Z(x_{\bar{s}}; \boldsymbol{\theta})} \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x}, \mathbf{z})), \tag{8}$$

where $x_{\bar{s}} = \{x \in \mathbf{x} | x \notin x_s\}$ and $Z(x_{\bar{s}}; \boldsymbol{\theta}) = \iint \ldots \int \exp(\boldsymbol{\theta}^T \boldsymbol{f}(\mathbf{x}, \mathbf{z}))\mathrm{d}x_s\mathrm{d}\mathbf{z}$. In such a model we will not observe the random variables $\mathbf{z}$, but we will be interested in modeling them. Further we will also be particularly concerned with modeling the marginal conditional distributions defined by $P(x_s|x_{\bar{s}}; \boldsymbol{\theta}) = \int P(x_s, \mathbf{z}|x_{\bar{s}}; \boldsymbol{\theta})\mathrm{d}\mathbf{z}$. A Predictive Random Field is a Markov Random Field where we have optimized the MRF so as to achieve the best possible prediction for a group of subsets of variables. Thus, the quantities we wish to focus on predicting in a PRF define the PRF and are given formally by a set of marginal conditional likelihoods that we wish to optimize. For a data set consisting of $i = 1 \ldots N$ observation instances and $j = 1 \ldots M$ subsets of observed variables, $\tilde{x}_{i,j}$ thus represents the ith example of the variables in subset j. Under these definitions, the optimal or maximum predictive likelihood parameter settings for a data set are given by

$$\underset{\theta}{\operatorname{argmax}} \prod_i \prod_j \int P(\tilde{x}_{i,j}, \mathbf{z}_{i,j}|\tilde{x}_{i,\bar{j}}; \theta)\mathrm{d}\mathbf{z}_{i,j}, \tag{9}$$

which represents the optimal parameter settings for an MRF under multiple marginal conditional likelihoods, or equivalently under a predictive likelihood. A PRF is thus constructed from an MRF by selecting a set of conditional distributions of an MRF, one for each observed random variable (or subset of observed variables) in the model conditioned on all of the other random variables in the model. A PRF becomes useful when a single set of consistent parameters $\boldsymbol{\theta}$ have been obtained for the underlying MRF through optimizing the Predictive Likelihood. We will return to the predictive likelihood again in Section 4.2 in the context of exponential family models and Harmoniums. However, first we shall review the exponential family Harmonium.

3.2 Exponential Family Harmoniums

A Harmonium model (Smolensky, 1986) is a Markov Random Field consisting of observed variables and hidden variables. A Harmonium is also a type of restricted Boltzmann machine (Hinton, 2002) that can be written as an exponential family model. Such a model can be written as

$$P(\mathbf{x}, \mathbf{y}|\boldsymbol{\Theta}) = \exp\left\{ \sum_i \boldsymbol{\theta}_i^T \boldsymbol{f}_i(\mathbf{x}_i) + \sum_j \boldsymbol{\theta}_j^T \boldsymbol{f}_j(\mathbf{y}_j) + \sum_i \sum_j \boldsymbol{\theta}_{ij}^T \boldsymbol{f}_{ij}(\mathbf{x}_i, \mathbf{y}_j) - A(\boldsymbol{\Theta}) \right\}, \tag{10}$$

where $\mathbf{y}$ is a vector of hidden variables, $\mathbf{x}$ is a vector of observations, $\boldsymbol{\theta}_i$ represents parameter vectors (or weights), $\boldsymbol{\theta}_{ij}$ represents a parameter vector on a cross product of states, $\boldsymbol{f}_i$ denotes potential functions, $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_{ij}, \boldsymbol{\theta}_i, \boldsymbol{\theta}_j\}$ is the set of all parameters and A is the log-partition function or normalization constant. A Harmonium model factorizes the third term of (10) into $\boldsymbol{\theta}_{ij}^T \boldsymbol{f}_{ij}(\mathbf{x}_i, \mathbf{y}_j) = \boldsymbol{f}_i(\mathbf{x}_i)^T \mathbf{W}_{ij}^T \boldsymbol{f}_j(\mathbf{y}_j)$, where $\mathbf{W}_{ij}^T$ is a parameter matrix with dimensions $a \times b$, i.e., with rows equal to the number of states of $\boldsymbol{f}_i(\mathbf{x}_i)$ and columns equal to the number of states of $\boldsymbol{f}_j(\mathbf{y}_j)$. Figure 3 illustrates a Harmonium model as a factor graph (Kschischang et al., 2001). Importantly, a Harmonium describes the factorization of a joint distribution for observed and hidden variables into a globally normalized product of local functions. In our experiments here we shall use the Harmonium's factorization structure to define a MRF and we will then define sets of conditionals as given by (9) so as to construct PRFs.

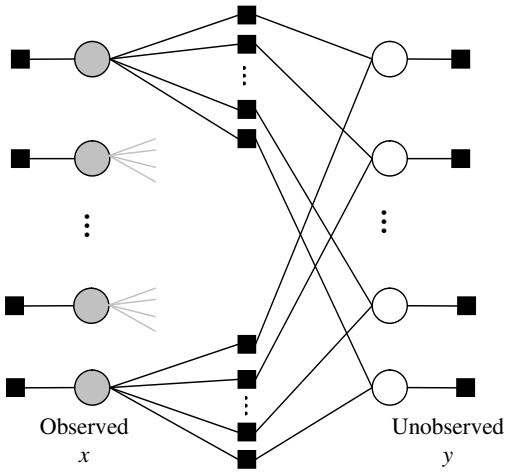

Figure 3: A factor graph for a Harmonium model. We will be particularly concerned with models where there are fewer hidden variables than observed variables.

4 Objective Functions and Learning using Contrastive Divergence

Here we review Maximum Likelihood (ML) and Conditional Likelihood (CL) in the context of learning for exponential family models. We then develop our new method which we call Predictive Likelihood (PL) learning and we contrast the approach with Bayesian Learning (BL) and prediction. In the following subsections we present the straightforward learning equations for gradient based methods which result for exponential family models for ML, CL and our new PL learning methods.

4.1 Maximum Likelihood Learning

We can optimize parameter settings $\tilde{\theta}$ of an undirected model with observed variables $\tilde{\mathbf{x}}$ and unobserved variables $\mathbf{z}$ under a maximum likelihood objective by computing

$$\underset{\theta}{\text{argmax}} \prod_i P(\tilde{\mathbf{x}}_i; \theta) = \underset{\theta}{\text{argmax}} \prod_i \int P(\tilde{\mathbf{x}}_i, \mathbf{z}_i; \theta) \mathrm{d}\mathbf{z}_i \tag{11}$$

It is well known that minimizing the KL divergence between the empirical distribution of the data and the marginal model distribution is equivalent to maximizing the likelihood as

$$KL(\tilde{P}(\tilde{\mathbf{x}}) || P(\mathbf{x}; \theta)) = -H(\tilde{P}(\tilde{\mathbf{x}})) - E_{\tilde{P}(\tilde{\mathbf{x}})} \langle \log P(\mathbf{x}; \theta) \rangle \tag{12}$$

and the log-likelihood can be expressed as $\mathcal{L}(\theta;\tilde{\mathbf{x}}) = N \cdot E_{\tilde{P}(\tilde{\mathbf{x}})}\langle \log P(\mathbf{x};\theta)\rangle$, where $E_{\tilde{P}(\tilde{\mathbf{x}})}$ denotes the expectation under the empirical distribution $\tilde{P}(\tilde{\mathbf{x}})$ and N is the number of data elements. In a (globally normalized) exponential family model with exponential function $\mathbf{F}(\mathbf{x};\theta)$, the gradient of the log-likelihood can be expressed as:

$$\frac{\partial \mathcal{L}(\theta;\tilde{\mathbf{x}})}{\partial \theta} = E_{\tilde{P}(\tilde{\mathbf{x}})}\left\langle \frac{\partial \mathbf{F}(\mathbf{x};\theta)}{\partial \theta} \right\rangle - E_{P(\mathbf{x};\theta)}\left\langle \frac{\partial \mathbf{F}(\mathbf{x};\theta)}{\partial \theta} \right\rangle \tag{13}$$

To compute the expectation in (13) in the models we investigate here, we use the contrastive divergence algorithm of (Hinton, 2002) whereby the model expectation is approximated by a Gibbs sample of hidden variables initialized with the input data followed by a single Gibbs sample of the observed variables for the expectation after on or a few iterations.

4.2 Conditional Likelihood Learning

When we have observations $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ for random variables $\mathbf{x}$ *and* $\mathbf{y}$, we can define a maximum conditional likelihood objective by computing

$$\underset{\theta}{\operatorname{argmax}} \prod_i P(\tilde{\mathbf{y}}_i|\tilde{\mathbf{x}}_i;\theta) = \underset{\theta}{\operatorname{argmax}} \prod_i \frac{P(\tilde{\mathbf{y}}_i,\tilde{\mathbf{x}}_i;\theta)}{\int P(\tilde{\mathbf{y}}_i,\tilde{\mathbf{x}}_i;\theta)\mathrm{d}\mathbf{y}_i} \tag{14}$$

The conditional log-likelihood can be expressed as $\mathcal{L}_{\mathbf{y}|\mathbf{x}}(\theta;\tilde{\mathbf{y}},\tilde{\mathbf{x}}) = N \cdot E_{\tilde{P}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}\langle \log P(\mathbf{y}|\mathbf{x};\theta)\rangle$, where $E_{\tilde{P}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}$ denotes the expectation under the empirical conditional distribution $\tilde{P}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})$ and N is the number of data elements. In a (globally normalized) exponential family model with exponential function $\mathbf{F}(\mathbf{y},\mathbf{x};\theta)$, the gradient of the log-likelihood can be expressed as:

$$\frac{\partial \mathcal{L}_{\mathbf{y}|\mathbf{x}}(\theta;\tilde{\mathbf{y}},\tilde{\mathbf{x}})}{\partial \theta} = E_{\tilde{P}(\tilde{\mathbf{y}}|\tilde{\mathbf{x}})}\left\langle \frac{\partial \mathbf{F}(\mathbf{y},\mathbf{x};\theta)}{\partial \theta} \right\rangle - E_{P(\mathbf{y}|\tilde{\mathbf{x}};\theta)}\left\langle \frac{\partial \mathbf{F}(\mathbf{y},\tilde{\mathbf{x}};\theta)}{\partial \theta} \right\rangle \tag{15}$$

To compute the expectation in (15) when there are additional hidden variables, we use a variation of the contrastive divergence sampling scheme for the conditional marginal model likelihood.

4.3 Predictive Likelihood Learning

Consider now the idea of maximizing the product of conditional distributions with the unobserved variables integrated out of the model. Such a procedure can we expressed as

$$\underset{\theta}{\operatorname{argmax}} \prod_i \prod_j P(\tilde{x}_{i,j}|\tilde{x}_{i,\bar{j}};\theta) = \underset{\theta}{\operatorname{argmax}} \prod_i \prod_j \frac{\int P(\tilde{x}_{i,j},\tilde{x}_{i,\bar{j}},\mathbf{z}_{i,j};\theta)\mathrm{d}\mathbf{z}_{i,j}}{\iint P(x_{i,j},\tilde{x}_{i,\bar{j}},\mathbf{z}_{i,j};\theta)\mathrm{d}x_{i,j}\mathrm{d}\mathbf{z}_{i,j}} \tag{16}$$

We call this optimization procedure Predictive Likelihood learning. If we can perform the integration over $\mathbf{z}$ in (16) or in cases when we have no hidden variables $\mathbf{z}$, the predictive log-likelihood can be expressed as

$$\mathcal{L}_{x_j|x_{\bar{j}},x_{\bar{j}}|x_j}(\theta;\tilde{\mathbf{x}}) = N\sum_j \cdot E_{\tilde{P}(\tilde{x}_j|\tilde{x}_{\bar{j}})}\langle \log P(x_j|x_{\bar{j}};\theta)\rangle, \tag{17}$$

where $E_{\tilde{P}(\tilde{x}_j|\tilde{x}_{\bar{j}})}$ denotes the expectation under the empirical distribution $\tilde{P}(\tilde{x}_j|\tilde{x}_{\bar{j}})$ and N is the number of data elements. Importantly, in contrast with the pseudo-likelihood (Besag, 1975), the quantity defined by (17) and the optimization here involve an integration over (a large number of) hidden variables. In a

(globally normalized) exponential family model with exponential function $\mathbf{F}(\mathbf{y}, \mathbf{x}; \theta)$, the gradient of the log-likelihood can be expressed as:

$$\frac{\partial \mathcal{L}_{x_j|x_{\bar{j}}, x_{\bar{j}}|x_j}(\theta; \tilde{\mathbf{x}})}{\partial \theta} = \sum_j \left[E_{\tilde{P}(\tilde{x}_j|\tilde{x}_{\bar{j}})} \left\langle \frac{\partial \mathbf{F}(x_j, x_{\bar{j}}; \theta)}{\partial \theta} \right\rangle - E_{P(x_j|\tilde{x}_{\bar{j}}; \theta)} \left\langle \frac{\partial \mathbf{F}(x_j, \tilde{x}_{\bar{j}}; \theta)}{\partial \theta} \right\rangle \right] \tag{18}$$

Again for our experiments here we use an interleaved version of the contrastive divergence sampling approach for approximating expectations involving marginal conditional distributions.

4.4 From Bayesian Learning and Prediction to Bayesian PRFs

In a Bayesian framework one treats parameters as random quantities. Bayesian Learning refers to the task of finding the posterior over parameters given a data set. One makes Bayesian predictions using the *predictive distribution* for the "next" observed variable x_{n+1} using the full posterior over hidden variables *and* parameters. When we have both an unobserved random variable $\mathbf{y}$ and treat parameters θ as random quantities, the Bayesian predictive distribution is given by

$$P(x_{n+1}|\tilde{x}_{1:n}) = \iint P(x_{n+1}, \mathbf{y}, \theta|\tilde{x}_{1:n}) \mathrm{d}\mathbf{y} \mathrm{d}\theta \tag{19}$$

Predictive likelihood learning is closely related to Bayesian learning and Bayesian prediction but the reader should notice that in our previous discussion we did not focus on defining models with priors for parameters and integrating over the uncertainty associated with parameters for prediction. In contrast, in PL learning we optimize point estimates for parameters. As well, rather than exploiting the predictive distribution for predicting new observations, we define the predictive likelihood using conditionals between each variable of the observation given the other variables or for groups of variables given the remaining variables and we integrate out the uncertainty associated with hidden variables. Further, we have defined the Predictive likelihood independently for each observation vector.

However, when we treat parameters as random quantities in a graphical model, it is straightforward to see how one can define *Bayesian Predictive Likelihoods* by taking the product of marginal conditional distributions for some *observations* given others. One can thus optimize different variations of the Bayesian Predictive likelihood, by combing the marginal Bayesian conditional distributions of (19) using different orderings of the data $x_{1:n+1}$. Such criteria are similar to the idea of optimizing soft versions of the leave-one-out or leave-N-out cross-validation error.

5 A Multiple Attribute Exponential Family Harmonium for Documents

The Exponential Family Harmonium model of (Welling et al., 2005) modeled all the conditional distributions for observed variables as arising from Multinomial conditional distributions. There are a rich class of conditional distributions that can arise from using exponential family models with a Harmonium structure. In addition to modeling words as done in the model of (Welling et al., 2005), we are interested in modeling authorship and information concerning the time of a publication. To achieve these goals, here we will construct Harmonium models such that conditional distributions given observed or sampled values have the form of multivariate Bernoulli, Multinomial and Discrete distributions as we now define.

To model positive integer counts, represented as a vector $\mathbf{x}$ such that $x_i \in \mathbb{N}_0$ we can write a d-dimensional *Multinomial* distribution with N total counts across all dimensions as

$$P(\mathbf{x}|\mathbf{p}, N) = \frac{N!}{\prod_{i=1}^d x_i!} \prod_{i=1}^d p_i^{x_i} \propto \exp\left\{ \log\left[\frac{\mathbf{p}}{1 - \mathbf{p}^T\mathbf{1}}\right]^T \mathbf{x} + N \log(1 - \mathbf{p}^T\mathbf{1}) \right\} \tag{20}$$

where $\mathbf{1}$, $\mathbf{x}$ and $\mathbf{p}$ are vectors of length $d-1$, $\mathbf{1} = [1, 1, \ldots, 1]^T$ and it is important to note that $\mathbf{x}$ on the left hand side of (20) does not contain the last element of the original vector $\mathbf{x}$. When $N = 1$ we have the *Discrete* distribution $\mathcal{D}(\mathbf{x}; \mathbf{p})$ where categorical variables are encoded by associating each state with a dimension and encoding observations with a one in the corresponding dimension and setting all other dimensions to zero. If we let $\boldsymbol{\theta} = \log(\mathbf{p} \cdot (1 - \mathbf{p}^T\mathbf{1})^{-1})$, $\mathbf{p} = \exp(\boldsymbol{\theta})(1 + \exp(\boldsymbol{\theta})^T\mathbf{1})^{-1}$ we can then express the Multinomial distribution as

$$\mathcal{M}(\mathbf{x}; \boldsymbol{\theta}, N) = \exp\{\boldsymbol{\theta}^T\mathbf{x} - N\log(1 + \exp(\boldsymbol{\theta})^T\mathbf{1})\} \tag{21}$$

To model authors, here we use a multivariate *Bernoulli* distribution. If $\mathbf{x}$ is a vector with elements such that $x_i \in \{0, 1\}$, a d-dimensional multivariate Bernoulli is written as

$$P(\mathbf{x}|\mathbf{p}) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{(1-x_i)} = \exp\left\{\log\left[\frac{\mathbf{p}}{(\mathbf{1} - \mathbf{p})}\right]^T\mathbf{x} + \log(\mathbf{1} - \mathbf{p})^T\mathbf{1}\right\} \tag{22}$$

Thus if we let $\boldsymbol{\theta} = \log(\mathbf{p} \cdot (\mathbf{1} - \mathbf{p})^{-1})$, $p_i = \exp(\theta_i)(1 + \exp(\theta_i))^{-1}$ we can express Bernoulli distribution $\mathcal{B}$ as a function of $\mathbf{x}$ and $\boldsymbol{\theta}$

$$\mathcal{B}(\mathbf{x}; \boldsymbol{\theta}) = \exp\{\boldsymbol{\theta}^T\mathbf{x} - \log(\mathbf{1} + \exp(\boldsymbol{\theta}))^T\mathbf{1}\} \tag{23}$$

We construct our harmonium models such that multivariate Bernoulli conditional distributions for observed binary values are "full-rank" and are drawn from an independent multivariate Bernoulli. Thus, in contrast to (Welling et al., 2005), our models only have one parameter for each binary observation. However, as in (Welling et al., 2005), the conditional distribution for hidden variables $\mathbf{y}$ in our model take the form of independent multivariate Gaussian.

We now define a composite, multi-attribute Harmonium where we will use a composite variable for observations consisting of $\mathbf{x}^T = [\mathbf{x}_b^T\ \mathbf{x}_m^T\ \mathbf{x}_d^T]^T$. The conditional distributions for hidden variables $\mathbf{y}_n$ given the values of observed variables $\tilde{\mathbf{x}}_b$ and the conditional distribution for observed variables $\mathbf{x}_b$ given a sampled value of hidden variables $\tilde{\mathbf{y}}_n$ are given by

$$P(\mathbf{y}_n|\tilde{\mathbf{x}}) = \mathcal{N}(\mathbf{y}_n; \hat{\boldsymbol{\mu}}, \mathbf{I}), \quad \hat{\boldsymbol{\mu}} = \boldsymbol{\mu} + \mathbf{W}^T\tilde{\mathbf{x}} \quad \text{and} \tag{24}$$

$$P(\mathbf{x}_b|\tilde{\mathbf{y}}) = \mathcal{B}(\mathbf{x}_b; \hat{\boldsymbol{\theta}}_b), \quad \hat{\boldsymbol{\theta}}_b = \boldsymbol{\theta}_b + \mathbf{W}_b\tilde{\mathbf{y}} \tag{25}$$

respectively. In practice we will hold the "local potentials" fixed such that $\boldsymbol{\theta}_b = \mathbf{0}$ and $\boldsymbol{\mu} = \mathbf{0}$. In our models here, the conditional distributions are defined on vectors of word frequencies and take the form of an independent but multivariate Multinomial and Discrete distributions. The corresponding conditionals are given by

$$P(\mathbf{x}_m|\tilde{\mathbf{y}}) = \mathcal{M}(\mathbf{x}_m; \hat{\boldsymbol{\theta}}, N), \quad \hat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}_m + \mathbf{W}_m\tilde{\mathbf{y}} \quad \text{and} \tag{26}$$

$$P(\mathbf{x}_d|\tilde{\mathbf{y}}) = \mathcal{D}(\mathbf{x}_d; \hat{\boldsymbol{\theta}}), \quad \hat{\boldsymbol{\theta}}_d = \boldsymbol{\theta}_d + \mathbf{W}_d\tilde{\mathbf{y}}, \tag{27}$$

where we hold $\boldsymbol{\theta}_m = \boldsymbol{\theta}_d = -\log(d - 1)$. For a simple model with only one type of observed variable, $\mathbf{W}^T \in \{\mathbf{W}_b^T, \mathbf{W}_m^T, \mathbf{W}_d^T\}$ in (24). In our composite model $\mathbf{W}^T = [\mathbf{W}_b^T\mathbf{W}_m^T\mathbf{W}_d^T]$. In a Harmonium structured model we have a marginal distribution for a vector of binary observations. We can represent that marginal distribution using an extended form[2] of the multivariate Bernoulli and Multinomial in which observations are no longer independent as was the case in (23) and (21). The following equation can can thus be used to represent the marginal as

$$P(\mathbf{x}|\boldsymbol{\theta}, \boldsymbol{\Lambda}) = \exp\{\boldsymbol{\theta}^T\mathbf{x} + \mathbf{x}^T\boldsymbol{\Lambda}\mathbf{x} - A(\boldsymbol{\theta}, \boldsymbol{\Lambda})\} \tag{28}$$

[2]To the best of our knowledge, there is no commonly agreed upon form for the non-independent multivariate Bernoulli.

where $\mathbf{\Lambda} = \frac{1}{2}\mathbf{W}\mathbf{W}^T$. We can compute the gradient of the log-likelihood under this construction using

$$\frac{\partial \mathcal{L}(\mathbf{W}^T; \tilde{\mathbf{X}})}{\partial \mathbf{W}^T} = \frac{1}{N_d} \sum_{i=1}^{N_d} \left(\mathbf{W}^T \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T - \frac{1}{N_s} \sum_{j=1}^{N_s} \mathbf{W}^T \tilde{\mathbf{x}}_{i,(j)} \tilde{\mathbf{x}}_{i,(j)}^T \right) \tag{29}$$

where N_d are the number of vectors of observed data, $\tilde{\mathbf{x}}_{i,(j)}$ are samples indexed by j and N_s are the number of MCMC samples used per data vector, which is either one or a small number of MCMC steps initialized from the data vector for the contrastive divergence approach.

6 Other Relevant Work

In different communities different variations of undirected models and methods for their optimization have been developed. For example, in the statistic literature there are a number of models satisfying these properties including: exponential family models, generalized linear models (GLIMs) (McCullagh & Nelder, 1989), and the types of log-linear graphical models described in (Edwards, 1995). Heckerman and collaborators have proposed an undirected alternative to widely used Bayesian Network models known has Dependency Networks (Heckerman et al., 2000). In this approach, they construct models for the conditional distributions of each variable in a model given the other variables in the model. In their approach, conditional distributions are constructed from *arbitrary* conditional models. The structure of the underlying Markov network can then be determined from feature selection procedures applied to the conditional models. However, this approach does not easily facilitate the construction of a joint distribution consistent with the different conditional models. In contrast, in the PRF approach we introduce here, we produce a model with similar properties but which does result in a consistent joint model.

A number of related methods have been introduced for Maximum Likelihood (ML) learning and are applicable in exponential family models. In (Frydenberg & Edwards, 1989), the authors present an algorithm called MIPS, a modified iterative scaling algorithm for finding maximum likelihood estimates in exponential family models. In the original CRF paper (Lafferty et al., 2001) an iterati6. Location A fresh copy will be always kept in xuerui/author-topic/, although it is updated pretty often.

7. Some Matlab scripts a. loadATWT.m : load author-topic and word-topic matrices into memory b. JSD.m : compute Jensen-Shannon Divergence of two distributions c. topAuthorsWords.m : compute top N authors and words for each topic d. compTopicDist.m : compute the JSD matrix for topics 76ve scaling algorithm was proposed for learning using the conditional likelihood. While more recent work (Sha & Pereira, 2003) has shown the benefits of using limited memory quasi-newton, gradient descent based methods. In the context of ML learning with hidden variables, the Expectation Maximization algorithm (Dempster et al., 1977) represents one way to find a local maximum marginal likelihood solution. Correspondingly, in (Edwards & Lauritzen, 2001) an algorithm for maximizing the (CL) with hidden variables is proposed consisting of computing a "tilted" version of the unconditional likelihood, called the T-step, followed by a maximization or M-step.

7 Results and Analysis

We are interested in examining the quality of the latent representations obtained when optimizing multi-attribute Harmonium structured models under ML, CL and PL objectives. We use a similar testing strategy to (Welling et al., 2005) but focus on comparing the different latent spaces obtained with the different optimization objectives. Our models below all use a continuous latent variable parameterization consistent with (24).

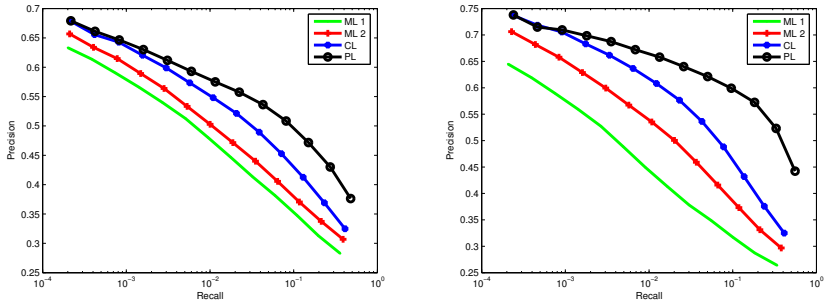

Figure 4: Precision-recall curves for the "20newsgroups" data using ML, CL and PL with (left) 10 latent variables and (right) 20 latent variables. Random guessing is a horizontal line at .25.

For our experiments, we use the reduced "20newsgroups" dataset prepared in MATLAB by Sam Roweis[3]. In this data set, 16242 documents are represented by 100 word vocabulary binary occurrences and are labeled as one of four domains. To evaluate the quality of our latent space, we retrieve documents that have the same domain label as a test document based on their cosine coefficient in the latent space when observing only binary occurrences. We randomly split data into a training set of 12,000 documents and a test set of 4242 documents. We use the full rank multivariate Bernoulli parameterization of (25) for binary word occurrences and the discrete parameterization of (27) for domains. Figure 4 shows precision-recall results. ML-1 is our model with no domain label information. ML-2 is optimized with domain label information. CL is optimized to predict domains from words and PL is optimized to predict both words from domains and domains from words.

From Figure 4 we see that the latent space captured by the model is more relevant for domain classification when the model is optimized under the CL and PL objectives. Further, at low recall both the CL and PL derived latent spaces produced similar precisions. However, as recall increases the precision for comparisons made in the PL derived latent space is consistently better.

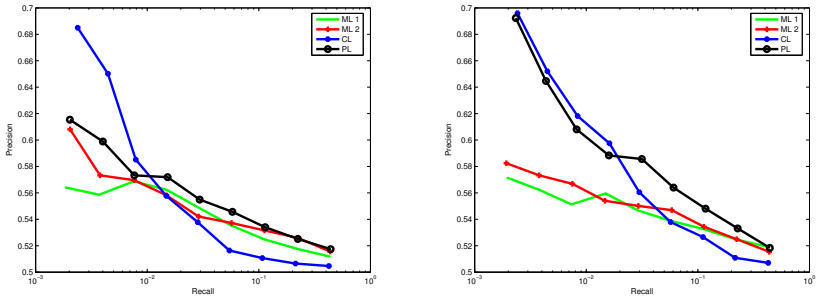

Figure 5: Precision-recall curves for the NIPS data set using ML, CL and PL with (left) 10 latent variables and (right) 20 latent variables. Random guessing is a horizontal line at .47

We have also tested our model on the NIPS Conference Papers data set obtained from Roweis' web site. We processed this data set via the following procedure: 1) only authors who published (by themselves or co-authored with someone else) more than 5 NIPS papers are considered; resulting in a set of 125 authors, 2) only papers authored by one or more of the above authors are considered, resulting in 873 papers, then 3) we

[3] http://www.cs.toronto.edu/∼roweis/data.html

select the top 150 words in terms of mutual information for authors. Papers are labeled by the NIPS volume number in which they were published. We use the parameterization of (26) for word counts, (25) for authors and (27) for time. We retrieve documents that have the same volume label as a test document based on the cosine coefficient between them in the latent space. For evaluation, we are not interested in the exact match between volume numbers, rather we consider a paper as relevant if a paper retrieved was published within ± 3 years of when the test document was published. In the training process, momentum and annealing are used to speed up convergence. We trained our model with 10 and 20 latent variables respectively, and report the precision and recall. Figure 5 illustrates the performance. Here CL optimization was defined to predict time from authors and words. PL optimization was defined so as to predict time from authors and words and the reverse, to predict words and authors from time.

From Figure 5 we see that at very low recall CL optimization produced a latent space resulting in higher precision. However, for both the 10 and 20 latent variable spaces as recall increases, the PL derived latent space produces precision results that are better than the CL and ML derived latent spaces.

8 Conclusions and Discussion

We have presented a new framework for optimizing Markov random fields based on an objective function we call the Predictive Likelihood. We refer to MRFs obtained through parameter optimization under this metric as Predictive Random Fields. We have developed an optimization approach and algorithms inspired by the contrastive divergence approach which was initially designed for Maximum Likelihood objectives. Our experiments show that latent class structure with improved relevance for classification tasks can be obtained via optimization under our PL objective when compared to Maximum Likelihood and Conditional Likelihood optimization. These results suggest that further exploration of PL optimization and the construction of PRFs with different structures represents a promising avenue of future research.

9 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, in part by The Central Intelligence Agency, the National Security Agency and National Science Foundation under NSF grant #IIS-0326249, and in part by the Defense Advanced Research Projects Agency (DARPA), through the Department of the Interior, NBC, Acquisition Services Division, under contract number NBCHD030010. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician*, *24*, 179–195.

Blei, D., & Jordan, M. (2003). Modeling annotated data. *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 127–134).

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA. *ECML 2002*.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & Harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, *41*, 391–407.

Dempster, A. J., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, *39*.

Edwards, D. (1995). *Introduction to graphical modelling*. Springer-Verlag, New York.

Edwards, D., & Lauritzen, S. (2001). The TM algorithm for maximizing a conditional likelihood function. *Biometrika*, *88*, 961–972.

Frydenberg, M., & Edwards, D. (1989). A modified iterative scaling algorithm for estimation in regular exponential families. *Comput. Stat. and Data Anal.*, *8*, 142–153.

Heckerman, D., Chickering, D., Meek, C., Rounthwaite, R., & Kadie, C. (2000). Dependency networks for density estimation, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, *1*, 49–75.

Hinton, G. (2002). Training products of experts by minimizing contrastive divergence. *Neural Computation*, *14*, 1771–1800.

Jebara, T., & Pentland, A. (2000). On reversing jensen's inequality. *In Neural Information Processing Systems 13, NIPS 13*.

Kschischang, F. R., Frey, B., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, *47*, 498–519.

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. 18th International Conf. on Machine Learning* (pp. 282–289). Morgan Kaufmann, San Francisco, CA.

Lauritzen, S., & Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, *17*, 31–57.

McCallum, A., Corrada-Emanuel, A., & Wang, X. (2005). Topic and role discovery in social networks. *Proceedings of IJCAI 2005*. Edinburgh, Scotland.

McCullagh, P., & Nelder, J. (1989). *Generalized linear models, 2nd edition*. Chapman and Hall, London.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. Banff, Alberta, Canada.

Salakhutdinov, R., Roweis, S., & Ghahramani, Z. (2003). Optimization with EM and expectation-conjugate-gradient. *Proceedings of the Twentieth International Conference on Machine Learning (ICML), Washington DC*.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of Human Language Technology-NAACL*. Edmonton, Canada.

Sivic, J., Russell, B., Efros, A., Zisserman, A., & Freeman, B. (2004). The couch potato project: learning about objects from looking at images. *In NIPS 2004 workshop on Structed Data and Representations in Probalistic Models for Categorization*.

Smolensky, P. (1986). *Information processing in dynamical systems: foundations of harmony theory*, chapter 2, 194–281. McGraw-Hill, New York.

Welling, M., Rosen-Zvi, M., & Hinton, G. (2005). Exponential family harmoniums with an application to information retrieval. In L. K. Saul, Y. Weiss and L. Bottou (Eds.), *Advances in neural information processing systems 17*, 1481–1488. Cambridge, MA: MIT Press.