On Sparse, Spectral and Other Parameterizations of Binary Probabilistic Models

David Buchman[1] **Mark Schmidt**[2] **Shakir Mohamed**[1] **David Poole**[1] **Nando de Freitas**[1]
{davidbuc, shakirm, poole, nando}@cs.ubc.ca, mark.schmidt@inria.fr

[1] Department of Computer Science, University of British Columbia, Vancouver, B.C., Canada
[2] INRIA – SIERRA Team, Laboratoire d'Informatique de l'École Normale Supérieure, Paris, France

Abstract

This paper studies issues relating to the parameterization of probability distributions over binary data sets. Several such parameterizations of models for binary data are known, including the Ising, generalized Ising, canonical and full parameterizations. We also discuss a parameterization that we call the "spectral parameterization", which has received significantly less coverage in existing literature. We provide this parameterization with a spectral interpretation by casting log-linear models in terms of orthogonal Walsh-Hadamard harmonic expansions. Using various standard and group sparse regularizers for structural learning, we provide a comprehensive theoretical and empirical comparison of these parameterizations. We show that the spectral parameterization, along with the canonical, has the best performance and sparsity levels, while the spectral does not depend on any particular reference state. The spectral interpretation also provides a new starting point for analyzing the statistics of binary data sets; we measure the magnitude of higher order interactions in the underlying distributions for several data sets.

1 Introduction

Log-linear models are used as efficient parameterizations for probability distributions in a wide variety of applications (Bishop et al., 1975; Whittaker, 1990; Lauritzen, 1996; Wasserman, 2004; Koller and Friedman, 2009). Due to their relatively small number of parameters, pairwise log-linear models have sometimes been advocated in scenarios where limited data is available (Whittaker, 1990, §9.3). However, pairwise models only focus on unary and pairwise statistical properties of the data; a pairwise assumption can be restrictive if higher-order moments of the data are important and we have sufficient training examples available to reliably estimate these higher-order statistics. Despite this fact, almost all previous work on structure learning with ℓ_1-regularization has made the pairwise assumption, with a few exceptions. Dahinden et al. (2007) consider log-linear models of discrete data where all potentials up to a fixed order are considered with (group) ℓ_1-regularization to learn the structure, Schmidt and Murphy (2010) address hierarchical log-linear models with (overlapping-group) ℓ_1-regularization, and Ding et al. (2011) further consider hierarchical log-linear models with covariates.

This paper makes two contributions. First, it develops a spectral interpretation of an existing parameterization of full log-linear models in terms of orthogonal Walsh-Hadamard bases (Beer, 1981). Although this parameterization has been used occasionally, to the best of our knowledge, its interpretation in terms of orthogonal expansions is new. We should also point out that a spectral expansion of probability measures in terms of Walsh basis functions has been studied in the context of univariate and bivariate probability density functions (Maqusi, 1981), and that orthogonal Walsh-Hadamard bases were used for analyzing binary factorial designs (Rockmore, 1997, §2.2). Our focus on log-linear models enables us to use this parameterization to study the "natural statistics", or spectrum, of several popular binary data sets, in the same fashion that researchers have investigated the natural statistics of images and other signals. Our results show that in this spectral domain, lower-order potentials tend to have much more weight than higher-order potentials. This result is very intuitive, but it is not obvious. For example, one can construct distributions which only have high-order potentials, such as the "parity distri-

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

bution":

$$p_{parity}(\mathbf{x}) \propto e^{\prod_i x_i}, \qquad x_i \in \{-1,+1\}.$$

The parameterization we describe here, which we coin the *spectral parameterization*, was suggested previously by Bishop et al. (1975). That work examined the 'full' parameterization of general discrete probabilistic models, and suggested adding constraints to the model parameters in order to obtain a minimal parameterization. For binary variables, the constraints leave a single degree of freedom for each potential, and can thus be modeled using a single parameter – essentially leading to the spectral parameterization. This initial work failed to notice the harmonic properties of the parameterization and does not make a connection to Walsh-Hadamard expansions.

The spectral parameterization has received minimal attention in the machine learning and statistics literature. Most recently, both the spectral representation and the Hadamard transform appeared in a paper in the field of haplotype inference (Kato et al., 2010). Kato et al. (2010), however, do not use the Hadamard transform to describe the harmonic structure of the parameterization, but rather use it as a computational tool to geometrically average approximated marginal distributions in a computationally efficient manner, and to find maximum-probability states given a model with a specific structure. This work, which is the one we believe is closest to ours, does not address the issue of learning. However, when used for learning, the spectral parameterization implicitly defines a new category of priors – priors over the spectral parameters of the distribution.

This spectral interpretation is important for several reasons. First, it allows us to conduct empirical analyses similar to the ones that are carried out for other types of data using the Fourier and wavelet transforms. Second, it provides an important bridge between harmonic analysis and related fields such as compressed sensing, and the problem of learning the parameters and structure of discrete probabilistic graphical models. Although we do not explore this theoretical connection in this paper, we believe it could potentially lead to new theoretical results about the sample complexity of sparse undirected probabilistic models (Candes et al., 2006; Abbeel et al., 2006; Ravikumar et al., 2010).

The second contribution of this paper is to present a comprehensive comparison of different parameterizations of undirected discrete probabilistic models, including the full, Ising, canonical and spectral parameterizations. This comparison is done in the context of learning with several types of standard and group ℓ_1 regularizers. These experiments can be seen as an extension of the ones conducted by Schmidt and Murphy (2010) to a much broader range of parameterizations. A comparative analysis of such a wide range has not been previously undertaken in the literature. We believe researchers will find it useful as it sheds light on the choice of parameterization and variant of ℓ_1 regularization that are used when learning discrete probabilistic models.

2 Log-linear Models

Given n binary random variables $\mathbf{x} \in \{-1,+1\}^n$, we can express a positive joint probability as a globally normalized product of potential functions $\exp(\phi_A(\mathbf{x}_A))$ defined for each possible subset A of $S \triangleq \{1,2,\ldots,n\}$:

$$p(\mathbf{x}) \triangleq \frac{1}{Z}\prod_{A\subseteq S}\exp(\phi_A(\mathbf{x}_A)).$$

The normalizing constant Z ensures that the distribution sums to one. When the logarithm of each potential is linear in the parameters of the potential, and the normalizing constant is encoded as $-\log(Z) = \phi_\emptyset(\mathbf{x})$, we can express the model in the standard log-linear form (Bishop et al., 1975):

$$\log p(\mathbf{x}) = \sum_{A\subseteq S}\phi_A(\mathbf{x}_A) = \sum_{A\subseteq S}\mathbf{w}_A^T\mathbf{f}_A(\mathbf{x}_A), \qquad (1)$$

where $\mathbf{f}_A(\mathbf{x}_A)$ is a feature vector derived from $\mathbf{x}_A$. The vector $\mathbf{w}$ is the set of log-linear parameters. We use the short-hand $\mathbf{w}_A$ to refer to all the parameters associated with the function $\phi_A(\mathbf{x}_A)$, and use $\mathbf{w}$ to refer to the concatenation of all $\mathbf{w}_A$.

Undirected probabilistic graphical models can be derived from log-linear models by connecting node i to j if variables x_i and x_j co-occur in some ϕ term. More precisely, a log-linear model is *graphical* if there is a non-zero ϕ term for every clique in the graph (Wasserman, 2004). Conversely, $\phi_A(\mathbf{x}) = 0$ if $\{s,t\} \subseteq A$ for $t \neq s$, and (s,t) is not an edge. Second, a log-linear model is *hierarchical* if $\phi_A = 0$ and $A \subset B$ implies that $\phi_B = 0$. Typically, attention is restricted to the class of hierarchical log-linear models due to the interpretability of their sparsity pattern in terms of conditional independence (Whittaker, 1990).

In practice, it is typically not feasible to include a potential $\phi_A(\mathbf{x}_A)$ for all 2^n subsets. Removing the potential $\phi_A(\mathbf{x}_A)$ from the model is equivalent to setting it to zero for all values of $\mathbf{x}_A$, or equivalently setting all elements of $\mathbf{w}_A$ to zero. For example, we obtain the class of *pairwise models* if we enforce $\mathbf{w}_A = \mathbf{0}$ for all A with a cardinality greater than two. This effectively nullifies the effects of higher-order statistics present in the data on the model.

3 Parameterizations

3.1 Full Parameterization

With the *full parameterization* of log-linear models, pairwise potentials have the form:

$$\phi_{ij}(x_i, x_j) = \sum_{s_1}\sum_{s_2} \mathbb{I}_{<s_1,s_2>}(x_i, x_j) w_{ijs_1s_2},$$

where $s_1, s_2 \in \{-1, +1\}$ and the indicator $\mathbb{I}_{<s_1,s_2>}(x_i, x_j)$ is one if $x_i = s_1$ and $x_j = s_2$, and zero otherwise. For three-way potentials we have:

$$\phi_{ijk}(x_i, x_j, x_k) = \sum_{s_1,s_2,s_3} \mathbb{I}_{<s_1,s_2,s_3>}(x_i, x_j, x_k) w_{ijks_1s_2s_3}$$

and similarly for higher-order potentials. In general, if A contains k elements that can each take 2 values, $\phi_A(\mathbf{x}_A)$ will have 2^k parameters $\mathbf{w}_A$.

3.2 Ising Parameterizations

The *Ising parameterization* allows us to reduce the model complexity and consider potentials with a single parameter:

$$\phi_{ij}(x_i, x_j) = \sum_{s} \mathbb{I}_{<s,s>}(x_i, x_j) w_{ij}.$$

Generalized Ising models allow potentials to take c parameters rather than a single one, where c is the number of values that variables can take. For binary variables, $c = 2$, which we assume throughout. For pairwise potentials we have:

$$\phi_{ij}(x_i, x_j) = \sum_{s} \mathbb{I}_{<s,s>}(x_i, x_j) w_{ijs}.$$

And similarly, for three-way potentials:

$$\phi_{ijk}(x_i, x_j, x_k) = \sum_{s} \mathbb{I}_{<s,s,s>}(x_i, x_j, x_k) w_{ijks}.$$

3.3 Canonical Parameterizations

Another strategy for decreasing the number of parameters is the *canonical parameterization* (Koller and Friedman, 2009; Lauritzen, 1996). To understand this parameterization, consider the set of features:

$$f_A(\mathbf{x}_A) = \begin{cases} 1 & \text{iff } \mathbf{x}_A = \mathbf{1} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{1}$ denotes a vector of ones. It is a simple exercise to verify that for this choice one has:

$$\log p(\mathbf{x}_A = \mathbf{1}, \mathbf{x}_{\overline{A}} = -\mathbf{1}) = \sum_{C \subseteq A} w_C,$$

where $\overline{A}$ is the set complement of A, and all f_A and w_A are scalars. Solving the linear system of 2^n equations for $\mathbf{w}$ yields:

$$w_A = \sum_{C \subseteq A} (-1)^{|A-C|} \log p(\mathbf{x}_C = \mathbf{1}, \mathbf{x}_{\overline{C}} = -\mathbf{1}). \quad (2)$$

This is a special case of the Möbius inversion lemma. We have chosen here the specific reference state $\mathbf{1}$, which may be replaced with an arbitrary reference state (see Koller and Friedman (2009) for discussion).

In this canonical parameterization we need only one parameter per potential, hence 2^n parameters are needed to represent the entire distribution, one of which is the normalizing constant. The Ising parameterization also requires 2^n parameters, but it is not complete (it can't represent all positive distributions). The other parameterizations require more parameters: the full parameterization requires 3^n, since each parameter corresponds to an assignment of one of $\{\text{not in } A, -1, +1\}$ to each variable, and generalized Ising requires $2^n c = 2^{n+1}$ ($c = 2$ for binary variables).

3.4 Spectral Parameterization

We consider an alternative parameterization that has received less attention in the literature, but which has important properties of theoretical and practical significance. We refer to this fourth parameterization as the *spectral parameterization*. The motivation for considering such a parameterization was made aptly by Koller and Friedman (2009, §4.4.2.1, pg. 130): "*... canonical parameters are not very intuitive, highlighting yet again the difficulties of constructing a reasonable parameterization of a Markov network by hand.*"

With $n = 3$, the spectral parameterization of $\log p(\mathbf{x})$ is as follows:

$$\begin{aligned}\sum_{A \subseteq S} \phi_A(\mathbf{x}_A) \;&=\; w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_{12}x_1x_2 \\ &\quad + w_{13}x_1x_3 + w_{23}x_2x_3 + w_{123}x_1x_2x_3.\end{aligned}$$

That is, the spectral parameterization has only one parameter per potential. In general, we have:

$$\log p(\mathbf{x}) \;=\; \sum_{A \subseteq S} w_A \prod_{i \in A} x_i. \quad (3)$$

Again, we only need 2^n parameters to represent the entire distribution. We are required to assess $2^n - 1$ parameters for all w_A, where $A \neq \emptyset$; $w_\emptyset$ can be used to ensure values sum to 1.

Although this spectral parameterization is not widely discussed in existing literature, it has been known (for general discrete log-linear models) for several decades (Bishop et al., 1975). Here, we show that the spectral representation provides a map between log-linear models and orthogonal expansions in terms of Walsh-Hadamard bases. As an example, consider a log-linear model with two random variables: $q(x_1, x_2) = w_0 + w_1x_1+w_2x_2+w_{12}x_1x_2$, where $q(x_1, x_2) \triangleq \log p(x_1, x_2)$. We can ground this expression for all realizations of the random variables to obtain:

$$\begin{aligned} q(+1,+1) &= w_0 + w_1 + w_2 + w_{12} \\ q(-1,+1) &= w_0 - w_1 + w_2 - w_{12} \\ q(+1,-1) &= w_0 + w_1 - w_2 - w_{12} \\ q(-1,-1) &= w_0 - w_1 - w_2 + w_{12}, \end{aligned}$$

which written in matrix notation is: $\mathbf{q} = 2\mathbf{H}_2\mathbf{w}$. $\mathbf{H}_2$ is the *Hadamard matrix* for two variables:

$$\mathbf{H}_2 = 2^{-1}\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix} \tag{4}$$

For n variables, the Hadamard matrix has entries $(\mathbf{H}_n)_{ij} = 2^{-n/2}(-1)^{i\cdot j}$, where $i \cdot j$ is the bitwise dot product of the binary representations of the numbers i and j indexing the 2^n possible realizations of $\mathbf{x}$ (e.g., Pratt et al. (1969)). The rows (and columns) of the Hadamard matrix are orthogonal *Walsh functions* $h(\cdot)$ (Beer, 1981). Let $\mathbf{x}^{(t)}$ denote the t-th realization of $\mathbf{x}$ for $t = 0, \ldots, 2^n - 1$. Then, we can rewrite our linear system in terms of these basis functions as follows:

$$q(\mathbf{x}^{(t)}) = \sum_{A\subseteq S} h_A(\mathbf{x}^{(t)})w_A. \tag{5}$$

This is the forward Walsh-Hadamard transform. Since the Walsh functions are orthogonal, that is $\sum_{t=0}^{2^n-1} h_A(\mathbf{x}^{(t)})h_B(\mathbf{x}^{(t)}) = 2^n\mathbb{I}_{A=B}$, we can multiply both sides of the forward transform by h_A and sum over t to obtain the reverse transform:

$$w_A = 2^{-n}\sum_{t=0}^{2^n-1} q(\mathbf{x}^{(t)})h_A(\mathbf{x}^{(t)}). \tag{6}$$

Equations 5 and 6 define the Walsh-Hadamard transform. $\mathbf{H}_n$ is symmetrical, and as with the Fourier transform, the Walsh-Hadamard transform pair can be expressed as:

$$\mathbf{q} = 2^{n/2}\mathbf{H}_n\mathbf{w} \qquad \text{and} \qquad \mathbf{w} = 2^{-n/2}\mathbf{H}_n\mathbf{q}. \tag{7}$$

Likewise, it can be computed using the *Fast Walsh-Hadamard transform* (FWHT) in $O(n2^n)$ (that is, $m \log m$ instead of m^2, where $m = 2^n$).

3.5 Comparing different parameterizations

Table 1 summarizes the properties of the parameterizations under consideration. We consider the generalization of the parameterizations to variables with c values, however the spectral parameterization is considered only for $c = 2$. "Canonical" refers to the canonical parameterization with a general reference state, while **C1** and **C2** use reference states $\mathbf{1}$ and $-\mathbf{1}$ respectively.

A parameterization is *complete* if it can represent any positive distribution. A parameterization is *minimal* if it has no redundant parameters. The next column corresponds to symmetry with respect to the values of individual variables. Assume we flip the values of a specific variable. How would the parameters change, in order to reflect the modification to the distribution? Symmetric parameterizations would only need to make trivial modifications to their parameters, such as exchanging some of them with each other, or flipping their signs. Non-symmetric parameterizations require much more complex calculations. In particular, a sparse $\mathbf{w}$ may then become dense. The next column offers a similar analysis, for exchanging the values of two different variables. Finally, "*uniquely* defined" means the distribution is uniquely defined as a function of the parameter vector $\mathbf{w}$, without additional external information such as an identity of a reference state.

Generalized Ising is complete only for binary variables, for which it has two times the minimal number of parameters. Ising has the minimal number, but with redundancy, as it is not complete. The canonical parameterization requires the specification of a reference state. Each possible reference state can be seen as giving rise to a different parameterization, therefore the canonical parameterization is not uniquely defined. The spectral parameterization is the only one that is minimal and symmetric w.r.t. values. In addition, it is also complete, symmetric w.r.t. variables, and it is unique.

4 Sparse Regularization for Structure Learning

All complete parameterizations share the same maximum-likelihood (ML) estimate. Once regularization is introduced, each combination of parameterization and regularizer defines a different prior over the space of distributions. Aside from computational issues, this choice also impacts prediction performance. In our experiments, we consider four different forms of sparse regularization. The first is the standard ℓ_1-regularization, used in the context of log-linear models by Lee et al. (2006):

$$||\mathbf{w}_A||_1 \triangleq \sum_j |\mathbf{w}_A^{(j)}|, \quad ||\mathbf{w}_A||_2 \triangleq \big(\sum_j (\mathbf{w}_A^{(j)})^2\big)^{1/2},$$

$$\max_{\mathbf{w}} \sum_{i=1}^n \log p(\mathbf{x}_i|\mathbf{w}) - \sum_{A\subseteq S} \lambda_A||\mathbf{w}_A||_1. \tag{8}$$

The ℓ_1-regularization encourages the parameter vector $\mathbf{w}$ to be sparse. To illustrate why one has to devote careful thinking to the pairing of parameterization and

Table 1: Properties of parameterizations.

Parameterization	**#Parameters**	**Complete**	**Minimal**	**Symm w.r.t. values**	**Symm w.r.t. variables**	**Uniquely defined**
Full	$(c+1)^n$	Yes	No	Yes	Yes	Yes
Ising	2^n	No	No	No	Yes	Yes
Generalized Ising	$2^n c$	For $c = 2$	No	No	Yes	Yes
Canonical	c^n	Yes	Yes	No	No	No
Canonical (C1/C2)	c^n	Yes	Yes	No	Yes	Yes
Spectral (for $c = 2$)	2^n	Yes	Yes	Yes	Yes	Yes

sparse regularizer let us look at a very simple example where undesirable behavior can easily arise. Assume we have a model with two binary random variables described by the following joint probability table: $p(x_1, x_2) = \begin{bmatrix} \delta_1 & \delta_2 \\ \delta_3 & 1-\delta_1-\delta_2-\delta_3 \end{bmatrix}$. If we place an ℓ_1 regularizer on the parameters δ_i, then while the first three cells are forced to shrink to zero, the fourth one is forced to grow to 1. Thus, a naive choice of parameterization with ℓ_1-regularization can induce arbitrary and potentially unwanted behavior.

For parameterizations where each $\mathbf{w}_A$ is a scalar (Ising, canonical, and spectral), Equation 8 encourages the removal of factors from the model. However, for parameterizations where each $\mathbf{w}_A$ has more than one element (generalized Ising, full), ℓ_1-regularization does not encourage entire factors $\mathbf{w}_A$ to be equal to the zero vector simultaneously. If this is desired, one can use group ℓ_1-regularization:

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{w}) - \sum_{A \subseteq S} \lambda_A ||\mathbf{w}_A||_2, \qquad (9)$$

which was first used in the context of log-linear models by Dahinden et al. (2007). In cases where w_A is a scalar, Equations 8 and 9 are equivalent. Although group ℓ_1-regularization encourages the removal of entire factors even for parameterizations that have more than one parameter per factor, it still does not directly encourage conditional independencies in the learned distribution. This is because it may estimate a sparse but non-hierarchical model, where $\mathbf{w}_A = \mathbf{0}$ in the solution of Equation 8 or 9 but this is not true under a re-parameterization of the same distribution (for non-minimal parameterizations). To encourage the removal of factors *and* the underlying model to be hierarchical, we also use the hierarchical group ℓ_1-regularization proposed by Schmidt and Murphy (2010):

$$\max_{\mathbf{w}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{w}) - \sum_{A \subseteq S} \lambda_A \left(\sum_{B \supseteq A} ||\mathbf{w}_B||_2^2 \right)^{1/2}. \qquad (10)$$

The fourth choice, which we refer to as the "flat" regularizer, is simply ℓ_p^p:

$$\begin{aligned} & \max_{\mathbf{w}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{w}) - \lambda ||\mathbf{w}||_p^p \\ = & \max_{\mathbf{w}} \sum_{i=1}^{n} \log p(\mathbf{x}_i|\mathbf{w}) - \lambda \sum_j |\mathbf{w}^{(j)}|^p. \qquad (11) \end{aligned}$$

We use this regularizer only for the spectral parameterization, and for $p = 1, 2$. For $p = 1$, this is identical to the standard ℓ_1 regularizer with a constant $\lambda_A = \lambda$. This regularizer penalizes all parameters independently and identically, a fact we rely on in Section 6.

5 Experiment 1: Comparing Parameterizations

Here, we compare different parameterizations and regularizers on several data sets in terms of test-set negative log-likelihood. We followed the same training protocol as Schmidt and Murphy (2010). For cross-validation of the hyperparameter, we divided the data into equal training, validation and testing sets. We repeated this with 10 different splits to control for variability in the results. We refer the reader to Schmidt and Murphy (2010) for further details. We experimented with a wide range of benchmark data sets, including the Yeast (Elisseeff and Weston, 2002), USPS (Roweis), Jokes (Goldberg et al., 2001), Flow (Sachs et al., 2005), Rochdale (Whittaker, 1990), Czech (Edwards and Havranek, 1985) and NLTCS (Erosheva et al., 2007) data sets. We found the same trends in the results across these data sets, so for clarity of presentation we focus only on the NLTCS, Yeast (labels only), and USPS (central 16 pixels). We purposely chose data sets with a limited number of variables, for which it is possible to compute the partition function by enumeration, so as to avoid introducing artifacts in the comparison due to approximations.

We studied the Ising (**I**), canonical (**C1**, **C2** and **CR**), spectral (**S**), generalized Ising (**GI**), and full (**F**) parameterizations. We use **C1** and **C2** to denote the parameterizations using reference states **1** and **−1**, respectively. **CR** uses a random reference state. Note that Schmidt and Murphy (2010) consider the **GI** and

F parameterizations, while Ding et al. (2011) consider the **C1** parameterization. Each parameterization was assessed using the standard ℓ_1 regularization (Equation 8). The over-complete parameterizations (generalized Ising and full) were also tested with the group ℓ_1 regularization (Equation 9), where each group refers to the parameters $\mathbf{w}_A$ associated with a particular set of variables A. For the other parameterizations group ℓ_1 regularization is identical to the standard ℓ_1 regularization. In addition, we tested all parameterizations using the hierarchical group ℓ_1 regularization (Equation 10), where each group refers to the set of parameters associated with a particular A and all its supersets. We set $\lambda_A = \mathbb{I}_{|A|\geq 2} 2^{|A|-2}\lambda$. The hyperparameter λ was chosen using the validation sets. We used the optimization software and hierarchical-search strategy of Schmidt and Murphy (2010). Finally, we examined the spectral parameterization with the flat regularizer of Equation 11 with $p = 1$ ("flat ℓ_1") and $p = 2$ ("flat ℓ_2^2"), along with a pseudo-counts (**PC**) estimator. The PC estimator uses Dirichlet smoothing to estimate the log-probabilities from data.

Figure 1 shows a comparison of the predictive performance for several combinations of regularizers and parameterizations, in which lower values mean better performance. Figure 2 shows the corresponding sparsity levels of the learned models. In the figures, we use '-group' to indicate that the parameterization is subject to the group ℓ_1 regularizer and '-h' for the hierarchical regularizer. Due to space considerations, not all combinations are shown, but the ones shown do capture most of the important information.

The flat and PC models under-perform (Figures 1(b), 1(c)) and produce dense models (Figures 2(b), 2(c)). We do not consider them attractive for practical modeling purposes, but we added them to the comparison to facilitate the analysis of the next section. For all parameterizations, we have found no significant difference between the standard, group and hierarchical ℓ_1 regularizers in terms of predictive performance, as shown in Figure 1(a) for the NLTCS data set. An exception is the full parameterization, for which standard ℓ_1 under-performed relative to the other regularizers (Figures 1(a-c)). A possible explanation could be the much larger number of parameters in the full parameterization: 3^n, versus 2^n for the Ising, canonical and spectral parameterizations, and 2^{n+1} for the generalized Ising parameterization. More importantly, a much smaller percentage of its parameters correspond to the lower-cardinality potential functions, compared to the other parameterizations. For all parameterizations, standard ℓ_1 produced sparser models than group ℓ_1 and hierarchical group ℓ_1 (see Figure 2).

Figure 1 shows that the Ising parameterization performs poorly. This is not surprising, considering the fact that it is not complete. The canonical, spectral and generalized Ising parameterizations seem to have performance similar to each other. However, both the canonical and spectral parameterizations seem to do best in terms of both sparsity and predictive performance. These are also the only parameterizations that are both complete and minimal. The performance of the canonical parameterization may depend on the particular reference state. This is particularly evident in Figure 2(c).

For problems that can naturally be described as having a base state with rare deviations from this state, the canonical parameterization with this state as the reference state may be an appealing choice. In other cases, our experiments suggest either trying different reference states for the canonical parameterization, or using the spectral parameterization instead. Our experiments also suggest that standard ℓ_1 regularization is preferred due to its simplicity and sparser resulting models.

6 Experiment 2: The Statistics (Spectrum) of Binary Data Sets

The spectral decomposition of binary distributions enables us to answer the question: What are the statistics of commonly used binary data sets? Researchers have used tools such as the Fourier and wavelet transforms to understand the properties of natural data, such as images. By casting binary distributions in terms of the Walsh-Hadamard transform, we can now develop a method for estimating the spectral coefficients of the underlying distribution. This experiment reveals that low-order statistics are much more significant than higher-order ones for binary data sets routinely used in practice.

We trained the spectral parameterization with the flat ℓ_1 and ℓ_2^2 regularizers on all the data sets mentioned in the previous experiment. We also computed the pseudo-counts estimate, and used FWHT to transform the log-probabilities from the pseudo-counts model to the spectral parameterization. Readers may wonder, in light of the previous experiment, why we use the flat regularizers. Flat regularizers do not impose a prior bias for smaller-cardinality factors, so it is interesting to see that without this bias there is still a preference for smaller factors: indeed this is what our experiments indicate. This then provides a rationale for a prior bias towards considering only low-order factors. The parameter λ was estimated using 10-fold cross-validation. We used the size of the factors $|A|$ to group parameters. For each group, we calculated the mean magnitude $E(|w_A|)$ of the parameters in the group (for example, the average magnitude of the parameters for all factors of size 3). For the flat ℓ_1 estimates, we also calculated the density (percentage of non-zero parameters).

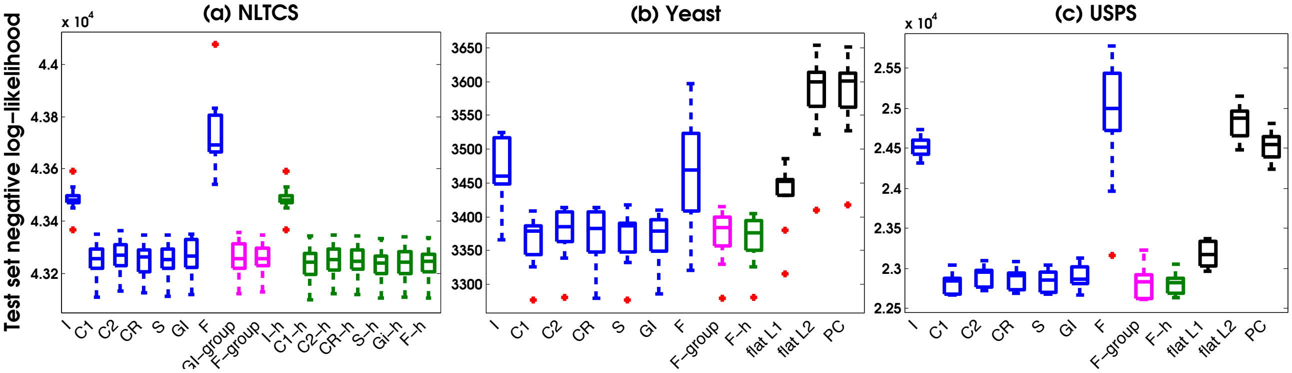

Figure 1: Test set negative log-likelihood for different parameterizations and regularizers.

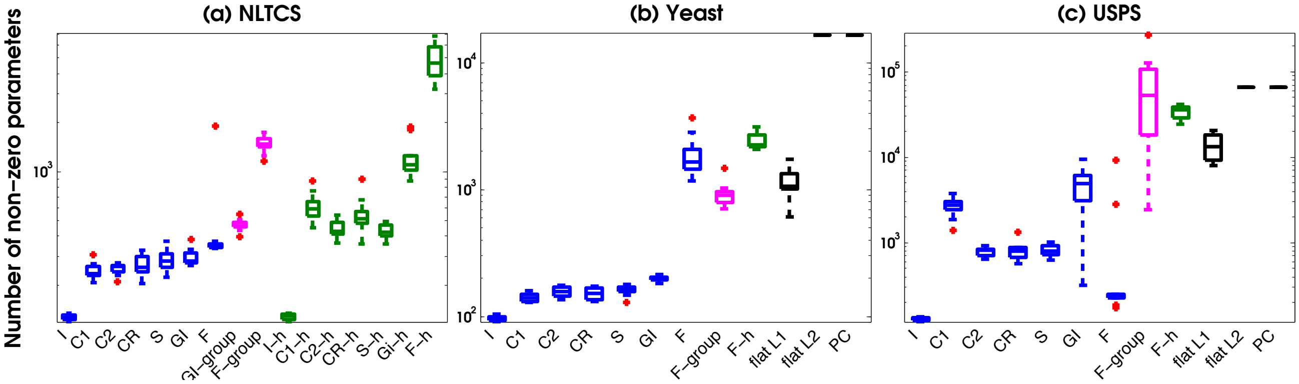

Figure 2: Model sparsity for different parameterizations and regularizers.

The results in Figure 3 show that the average magnitude of the spectral parameters diminishes with increasing factor size $|A|$. This behavior is especially pronounced for the ℓ_1 regularizer. According to our results in the previous experiment, the flat ℓ_1 regularizer allows for a much better fit than the flat ℓ_2^2 and PC strategies. It is therefore a more reliable estimator of the underlying distribution. Putting this fact together with the fact that ℓ_1 predicts that $|w_A|$ declines rapidly with increasing $|A|$ provides empirical support for the statement that real-world distributions tend to be sparse (with most parameters close to zero). Figures 3(b), 3(d) show the sparsity levels for different values of $|A|$ and for different data sets, when training took place with the flat ℓ_1 regularizer. We can see that parameters associated with a large $|A|$ are sparse, especially for the data sets that have more parameters (in Figure 3(b), Jokes has 100% density for $|A| = 10$, however, there is only one such parameter, so this is not significant). In addition, the "spectra" (Figures 3(a), 3(c)) illustrate that pair-wise and three-way models can capture a large portion of the variability in these data sets. This probably accounts for their popularity. The idea of ignoring higher-order potentials was also recently considered by Jalali et al. (2010), who analyzed the effect of fitting a pairwise model to data in terms of consistency of estimating the presence of higher-order interactions in hierarchical models. Nonetheless, our results also show that while adding higher order potentials may lead to an improvement in performance, this potential improvement comes with an additional computational cost that might not be justifiable in practical domains. Our spectral analysis of binary distributions provides justification for the standard machine learning and statistical practices of ignoring higher-order potentials. Importantly, it enables researchers to assess the cost of ignoring these terms.

7 Conclusions and Future Work

We have presented a comparison of different parameterizations for discrete probabilistic log-linear models, when learning their parameters and structure with sparsity-promoting regularizers. We found that the spectral parameterization is one of the best performing, and were able to interpret it as a harmonic series expansion of orthogonal Walsh-Hadamard bases. Since this interpretation brings closer the fields of harmonic analysis and discrete probabilistic modeling, it opens many doors for future research. We have already used the spectral parameterization to study the statistics, or the spectrum, of popular binary data sets. However, we believe this theoretical connection

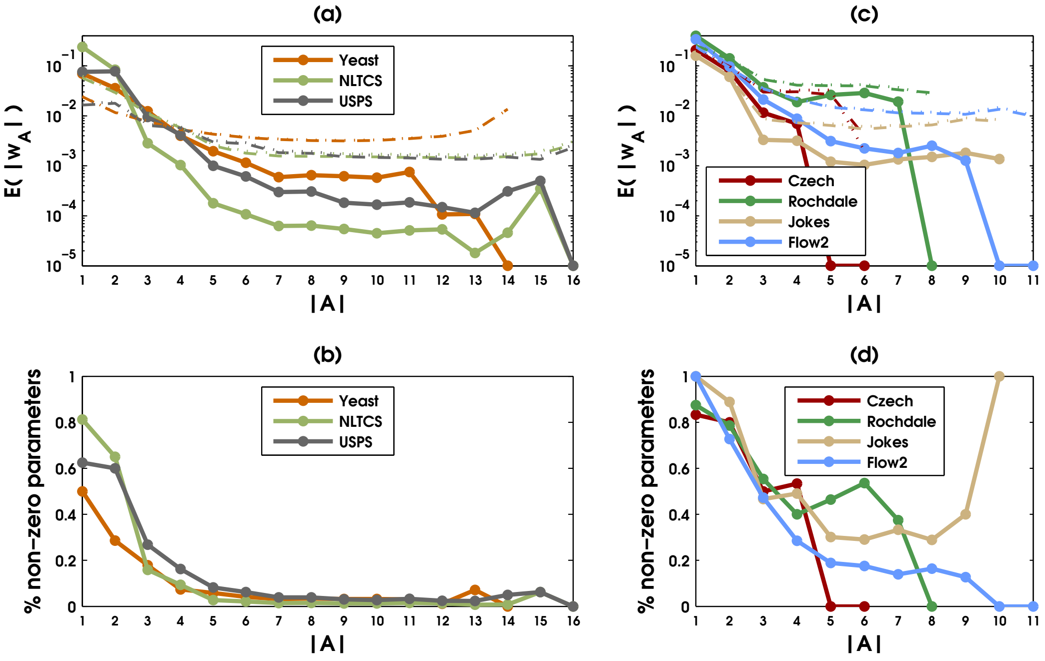

Figure 3: (a),(c) Mean parameter magnitude (log-scale), and (b),(d) percentage of non-zero parameters, for different factor sizes $|A|$. Data sets are split into two groups to allow for clearer plots. Solid lines correspond to the flat ℓ_1 regularizer, dashed lines to the flat ℓ_2^2 regularizer and dotted lines for the pseudo-counts estimator. The difference between the pseudo-counts and ℓ_2^2 regularizer is hardly discernible.

could be exploited in connection with compressed sensing to provide sample complexity theorems for learning discrete probabilistic models. For example, if we could measure log-probabilities directly, then the theorems of compressed sensing could be easily adapted to this domain to estimate the number of measurements needed to reconstruct the sparse probabilistic model. However, since log-probabilities cannot typically be measured directly, some theoretical challenges lie ahead. In future work, we also plan to build on the work of Bishop et al. (1975) to extend the results to distributions over random variables with more than two values.

Acknowledgements

This research was supported by NSERC grants to N. de Freitas and D. Poole. M. Schmidt was supported by the European Research Council (SIERRA-ERC-239993). S. Mohamed is supported by the Canadian Institute for Advanced Research (CIFAR).

References

Pieter Abbeel, Daphne Koller, and Andrew Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Journal of Machine Learning Research*, 7:1743–1788, 2006.

Tom Beer. Walsh transforms. *American Journal of Physics*, 49(5):466–472, 1981.

Yvonne M. Bishop, Stephen E. Fienberg, and Paul W. Holland. *Discrete multivariate analysis: Theory and practice.* MIT Press, 1975.

Emmanuel Candes, Justin Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

Corinne Dahinden, Giovanni Parmigiani, Mark C. Emerick, and P. Peter Bühlmann. Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics*, 8: 476, 2007.

Shilin Ding, Grace Wahba, and Jerry Xiaojin Zhu. Learning higher-order graph structure with features by structure penalty. *Neural Information Processing Systems*, 2011.

David Edwards and Tomas Havranek. A fast procedure for model search in multidimensional contingency tables. *Biometrika*, 72(2):339–351, 1985.

Andre Elisseeff and Jason Weston. Kernel methods for multi-labelled classification and categorical regression problems. In *Neural Information Processing Systems*, 2002.

Elena A. Erosheva, Stephen E. Fienberg, and Cyrille Joutard. Describing disability through individual-level mixture models for multivariate binary data. *Annals of Applied Statistics*, 1(2):502–537, 2007.

Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retrieval*, 4(2):133–151, 2001.

Ali Jalali, Pradeep Ravikumar, Vishvas Vasuki, and Sujay Sanghavi. On learning discrete graphical models using group-sparse regularization. In *AISTATS*. 2010.

Masashi Kato, Qian Ji Gao, Hiroshi Chigira, Hiroyuki Shindo, and Masato Inoue. A haplotype inference method based on sparsely connected multi-body Ising model. *Journal of Physics: Conference Series*, 233(1), 2010.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: Principles and techniques*. MIT Press, 2009.

Steffen L. Lauritzen. *Graphical models*. Oxford University Press, USA, 1996.

Su-In Lee, Varun Ganapathi, and Daphne Koller. Efficient structure learning of Markov networks using L_1-regularization. *Neural Information Processing Systems*, 2006.

Mohammad Maqusi. Walsh series expansions of probability distributions. *IEEE Transactions on Electromagnetic Compatibility*, 23(4):401–407, 1981.

W.K. Pratt, J. Kane, and H.C. Andrews. Hadamard transform image coding. *Proceedings of the IEEE*, 57(1):58–68, 1969.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

Daniel N. Rockmore. Some applications of generalized FFTs. In *Proceedings of the 1995 DIMACS Workshop on Groups and Computation*, pages 329–369. June, 1997.

Sam Roweis. USPS data. `http://cs.nyu.edu/∼roweis/data.html`.

Karen Sachs, Omar Perez, Dana Pe'er, Doug Lauffenburger, and Garry Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

Mark Schmidt and Kevin P. Murphy. Convex structure learning in log-linear models: Beyond pairwise potentials. *Artificial Intelligence and Statistics*, 2010.

Larry Wasserman. *All of statistics*. Springer, 2004.

Joe Whittaker. *Graphical models in applied multivariate analysis*. John Wiley and Sons, 1990.