# Convex Structure Learning in Log-Linear Models: Beyond Pairwise Potentials

**Mark Schmidt and Kevin Murphy**
Department of Computer Science
University of British Columbia
{schmidtm,murphyk}@cs.ubc.ca

## Abstract

Previous work has examined structure learning in log-linear models with $\ell_1$-regularization, largely focusing on the case of pairwise potentials. In this work we consider the case of models with potentials of arbitrary order, but that satisfy a hierarchical constraint. We enforce the hierarchical constraint using group $\ell_1$-regularization with *overlapping* groups. An active set method that enforces hierarchical inclusion allows us to tractably consider the exponential number of higher-order potentials. We use a spectral projected gradient method as a subroutine for solving the overlapping group $\ell_1$-regularization problem, and make use of a sparse version of Dykstra's algorithm to compute the projection. Our experiments indicate that this model gives equal or better test set likelihood compared to previous models.

## 1 Introduction

The statistical learning community has devoted a substantial amount of recent effort towards parameter estimation in graphical models with $\ell_1$-regularization. The appeal of these methods is that they formulate structure learning (selecting the set of edges present in the model) as a convex optimization problem. This approach was initially examined for the class of Gaussian graphical models (Banerjee et al., 2008), but these techniques were then extended to pairwise log-linear models of discrete data.

The discrete case is much harder than the Gaussian case, because of the potentially intractable normalizing constant. To address this problem, one can use pseudo-likelihood (Wainwright et al., 2006; Höfling and Tibshirani, 2009) or other approximate methods (Lee et al., 2006; Banerjee et al., 2008). Another complicating factor in the discrete case is that each edge may have multiple parameters. This arises in multi-state models as well as conditional random fields. In these scenarios, we can use group $\ell_1$-regularization to ensure all the parameters associated with an edge are set to zero simultaneously (Dahinden et al., 2007; Schmidt et al., 2008).

A natural extension is to use group $\ell_1$-regularization to learn the structure of log-linear models with higher-order factors. Dahinden et al. (2007) considered a generalization of pairwise models where all potentials up to a fixed order are considered. However, this approach is only practical when the number of nodes $p$ or the maximum size of the factors $m$ is very small, since if we allow for $m$-way factors there are $\binom{p}{m}$ possible subsets of size $m$ to examine. Further, if we allow factors of arbitrary size then there are $2^p$ factors to consider.

In this paper, we consider using group $\ell_1$-regularization for convex structure learning in the special case of hierarchical log-linear models, where a factor is only included if all its subsets are also included. Similar to (Bach, 2008), we develop an active set method that can incrementally add higher order factors. This method uses the hierarchical property to rule out most of the possible supersets, and converges to a solution satisfying a set of necessary optimality conditions. Key to the convex parameterization of the space of hierarchical log-linear models is that we allow the groups to overlap. However, this results in a more difficult optimization problem. We tackle this by using a spectral projected gradient method, where the projection step is computed using R. Dykstra's (1983) algorithm. We show that allowing for such higher order interactions can result in improved prediction accuracy over pairwise models.

Appearing in Proceedings of the $13^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. 

## 2 Log-Linear Models

In log-linear models (Bishop et al., 1975), we can write the probability of a vector $\mathbf{x} \in \{1, 2, \dots, k\}^p$ as a globally normalized product of potential functions $\phi_A(\mathbf{x})$ defined for each possible subset $A$ of $S \triangleq \{1, 2, \dots, p\}$,

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

where the normalizing constant $Z$ enforces that the distribution sums to one,

$$Z \triangleq \sum_{\mathbf{x}'} \prod_{A \subseteq S} \exp(\phi_A(\mathbf{x}')).$$

Each potential function $\phi_A$ is only allowed to depend on the variables in the subset $A$. We consider a full parameterization of these potential functions. For example, if $A$ is the set $\{3, 4\}$ and both $x_3$ and $x_4$ are binary, then

$$\begin{aligned}\phi_{3,4}(\mathbf{x}) = &\mathbb{I}(x_3 = 1, x_4 = 1)w_{1,1} \\ &+ \mathbb{I}(x_3 = 1, x_4 = 2)w_{1,2} \\ &+ \mathbb{I}(x_3 = 2, x_4 = 1)w_{2,1} \\ &+ \mathbb{I}(x_3 = 2, x_4 = 2)w_{2,2},\end{aligned}$$

where each potential $\phi_A$ has as its own set of parameters[1]. We will use the short-hand $\mathbf{w}_A$ to refer to all the parameters associated with the potential $\phi_A$, and we will use $\mathbf{w}$ to refer to the concatenation of all $\mathbf{w}_A$. In general, when $A$ contains $m$ elements that can each take $k$ values, $\phi_A(\mathbf{x})$ will have $k^m$ indicator functions and $\mathbf{w}_A$ will contain the $k^m$ corresponding parameter values.

Given a set of $n$ sample data vectors $\mathbf{x}^i$, the gradient of the average log-likelihood has a simple form. For example, consider the parameter $w_{1,2}$ in the potential $\phi_A(\mathbf{x})$ in the example above. The gradient is

$$\nabla_{w_{1,2}} \frac{1}{n} \sum_{i=1}^{n} \log p(\mathbf{x}^i|\mathbf{w}) =$$

$$\sum_{i=1}^{n} \frac{1}{n} \mathbb{I}(x_3^i = 1, x_4^i = 2) - p(x_3 = 1, x_4 = 2).$$

where $p(x_3, x_4)$ is the marginal according to the model. Thus we see that at a maximum likelihood solution (where the gradient is zero), we must match the empirical marginals to the model marginals.

In practice, it is typically not feasible to include a potential $\phi_A(\mathbf{x})$ for all $2^p$ subsets. Under our parameterization, removing the potential $\phi_A(\mathbf{x})$ from the model is equivalent to setting all elements of $\mathbf{w}_A$ to zero (or any other constant value). For example, we obtain the class of pairwise models used in prior work when $\mathbf{w}_A = \mathbf{0}$ for all $A$ with a cardinality greater than two (removing the restriction that higher-order marginals match the empirical frequencies). The prior work on structure learning in pairwise log-linear models with $\ell_1$-regularization assigns each pairwise set of parameters $\mathbf{w}_A$ to a single group, and optimizes the log-likelihood function subject to $\ell_1$-regularization of the norms of the groups (ie. group $\ell_1$-regularization), encouraging sparsity in terms of the groups.

We can extend this prior work to the general case by by solving the optimization problem (for $\lambda_A \geq 0$)

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

This is the approach taken in Dahinden et al. (2007), who also consider a variant where only factors up to a certain order are considered. A problem with this formulation is that sparsity in the variable groups $A$ does not directly correspond to conditional independencies in the model (except in the pairwise case). In particular, in a log-linear model variable sets $B$ and $C$ are independent given all other variables if and only if all elements of $\mathbf{w}_A$ are zero for all $A$ that contain at least one element from $B$ and at least one element from $C$, see Whittaker (1990, Proposition 7.2.1).

Further, it should be quite clear that this optimization problem can become very difficult if the number of variables is non-trivial and we don't enforce a cardinality restriction. For example, in our experiments we consider a model with 32 variables and 4 states, so the above optimization problem would contain $4^{32}$ groups, each containing up to $4^{32}$ parameters.

## 3 Hierarchical Log-Linear Models

As an alternative to using an explicit cardinality constraint, we consider fitting general log-linear models subject to the following constraint:

**Hierarchical Inclusion Restriction**:
If $\mathbf{w}_A = \mathbf{0}$ and $A \subset B$, then $\mathbf{w}_B = \mathbf{0}$.

This is the class of *hierarchical* log-linear models (Bishop et al., 1975; Whittaker, 1990, §7). While a subset of the space of general log-linear models, the set of hierarchical log-linear is much larger than the set of pairwise models, and can include interactions of any order. Further, group-sparsity in hierarchical models directly corresponds to conditional independence.

The hierarchical inclusion restriction imposes constraints on the possible sparsity pattern of $\mathbf{w}$, beyond

[1] This model is over-parameterized and unidentifiable, but the log-likelihood is convex and a unique optimum exists if we add a strictly convex regularizer.

that obtained using (disjoint) group $\ell_1$-regularization. In the context of linear regression and multiple kernel learning, several authors have recently shown that group $\ell_1$-regularization with *overlapping* groups can be used to enforce hierarchical inclusion restrictions (Zhao et al., 2009; Bach, 2008). As an example, if we would like to enforce the hierarchical inclusion restriction on the sets $A \subset B$, we can do this using two groups: The first group simply includes the variables in $B$, while the second group includes the variables in both $A$ and $B$. Regularization using these groups encourages $A$ to be non-zero whenver $B$ is non-zero, since when $B$ is non-zero $A$ is not locally penalized at zero, see Zhao et al. (2009, Theorem 1).

Generalizing this basic idea, to enforce that the solution of our regularized optimization problem satisfies the hierarchical inclusion restriction, we can solve the convex optimization problem

$$\min_{\mathbf{w}} -\sum_{i=1}^{n} \log p(\mathbf{x}^i|\mathbf{w}) + \sum_{A \subseteq S} \lambda_A \Big( \sum_{\{B|A \subseteq B\}} ||\mathbf{w}_B||_2^2 \Big)^{1/2}.$$

If we define the set of parameters $\mathbf{w}_A^*$ as the concatenation of the parameters $\mathbf{w}_A$ with all parameters $\mathbf{w}_B$ such that $A \subset B$, we can write this as

$$\min_{\mathbf{w}} -\sum_{i=1}^{n} \log p(\mathbf{x}^i|\mathbf{w}) + \sum_{A \subseteq S} \lambda_A ||\mathbf{w}_A^*||_2. \qquad (2)$$

This is very similar to (1), except that the parameters of higher-order terms are added to the corresponding lower-order groups. By Theorem 1 of Zhao et al. (2009) we can show that under reasonable assumptions a minimizer of (2) will satisfy hierarchical inclusion (we give details in the appendix). Unfortunately, there are several complicating factors in solving (2), we address these in the next three sections, beginning with the exponential number of groups to consider.

## 4 Active Set Method

Using $f(\mathbf{w})$ to denote the objective in (2), the subdifferential of $f(\mathbf{w})$ is

$$\partial f(\mathbf{w}) = -\nabla \sum_{i=1}^{n} \log p(\mathbf{x}^i|\mathbf{w}) + \sum_{A \subseteq S} \lambda_A \text{sgn}(\mathbf{w}_A^*),$$

where we use $\text{sgn}(\mathbf{y})$ to denote a set-valued map that for a real-vector $\mathbf{y} \neq \mathbf{0}$ returns $\mathbf{y}/||\mathbf{y}||_2$ and for $\mathbf{y} = \mathbf{0}$ returns all values such that $||\mathbf{y}||_2 \leq \mathbf{1}$ (we pad the output of this signum function with zeroes so that it has the right dimension). A vector $\tilde{\mathbf{w}}$ is a minimizer of a convex function iff $\mathbf{0} \in \partial f(\tilde{\mathbf{w}})$ (Bertsekas et al., 2003, §4).

We call $A$ an *active group* if $\mathbf{w}_B \neq \mathbf{0}$ for some $B$ such that $A \subseteq B$. If $A$ is not an active group and $\mathbf{w}_B = \mathbf{0}$ for some $B \subset A$, we call $A$ an *inactive group*. Finally, we define a *boundary group* as a group $A$ satisfying $\mathbf{w}_B \neq \mathbf{0}$ for all $B \subset A$ and $\mathbf{w}_C = \mathbf{0}$ for all $A \subseteq C$. (ie. the boundary groups are the groups that can be made non-zero without violating hierarchical inclusion).

The optimality conditions with respect to an active group $A$ reduce to

$$\nabla_{\mathbf{w}_A} \sum_{i=1}^{n} \log p(\mathbf{x}^i|\mathbf{w}) = \sum_{B \subseteq A} \lambda_A \mathbf{w}_A / ||\mathbf{w}_B^*||_2. \qquad (3)$$

If we treat all inactive groups as fixed, the optimality conditions with respect to a boundary group $A$ become

$$||\nabla_{\mathbf{w}_A} \sum_{i=1}^{n} \log p(\mathbf{x}^i|\mathbf{w})||_2 \leq \lambda_A. \qquad (4)$$

The combination of (3) and (4) constitute necessary and sufficient conditions for a minimizer of (2) under the constraint that inactive groups are fixed at zero. These also comprise necessary (but not necessarily sufficient) conditions for global optimality of (2).

We now consider an active set method that incrementally adds variables to the problem until (3) and (4) are satisfied, that uses hierarchical inclusion to exclude the possibility of adding most variables. The method alternates between two phases:

- Find the set of active groups, and the boundary groups violating (4).
- Solve the problem with respect to these variables.

We repeat this until no new groups are found in the first step, and at this point we have (by construction) found a point satisfying (3) and (4). The addition of boundary groups has an intuitive interpretation; we only add the zero-valued group $A$ if it satisfies hierarchical inclusion and the difference between the model marginals and the empirical frequencies exceeds $\lambda_A$.

Consider a simple 6-node hierarchical log-linear model containing non-zero potentials on (1)(2)(3)(4)(5)(6) (1,2)(1,3)(1,4)(4,5)(4,6)(5,6)(4,5,6). Though there are 20 possible threeway interactions in a 6-node model, only one satisfies hierarchical inclusion, so our method would not consider the other 19. Further, we do not need to consider any fourway, fiveway, or sixway interactions since none of these can satisfy hierarchical inclusion. In general, we might need to consider more higher-order interactions, but we will never need to consider more than a polynomial number of groups more than the number present in the final model. That

is, hierarchical inclusion and the active set method can save us from looking at an exponential number of irrelevant higher-order factors[2]. Further, to stop us from considering overly complicated models that do not generalize well, to set the regularization parameter(s) we start with the unary model and incrementally decrease the regularization until a measure of generalization error starts to increase.

## 5 Projected Gradient Method

In step 1 of the active set method we must solve (2) with respect to the active set. This comprises a group $\ell_1$-regularization problem with overlapping groups, and we note that the objective function is non-differentiable when any group $\mathbf{w}_A^*$ has all elements zero. Besides a special case discussed in (Zhao et al., 2009) where the solution can be computed directly, previous approaches to solving group $\ell_1$-regularization problems with overlapping groups include a boosted LASSO method (Zhao et al., 2009) and a re-formulation of the problem as a smooth objective with a simplex constraint (Bach, 2008). Unfortunately, applying these methods to graphical models would be relatively inefficient since it might require evaluating the normalizing constant in the model (and its derivatives) a very large number of times.

We solve this problem by writing it as an equivalent differentiable but constrained problem. In particular, we introduce a scalar auxiliary variable $g_A$ to bound the norm of each group $\mathbf{w}_A^*$, leading to a smooth objective with second-order cone constraints,

$$\min_{\mathbf{w},\mathbf{g}} -\log p(\mathbf{x}|\mathbf{w}) + \sum_{A \subseteq S} \lambda_A g_A, \quad (5)$$
$$s.t.\ \ \forall_A,\ \ g_A \geq ||\mathbf{w}_A^*||_2.$$

The objective function in this constrained optimization is differentiable and convex, and when the constraints are satisfied it represents an upper bound on the regularized likelihood (2). Further, at a minimizer it will be the case that $g_A = ||\mathbf{w}_A^*||_2$ for all $A$, since otherwise we could decrease the objective function while remaining feasible by decreasing $g_A$ to $||\mathbf{w}_A^*||_2$.

### 5.1 Spectral Projected Gradient

To solve (5), we use the spectral projected gradient (SPG) method introduced by Birgin et al. (2000). This has been shown to give good performance in a variety of related applications (Figueiredo et al., 2007; van den Berg and Friedlander, 2008; Schmidt et al., 2008). SPG is an enhancement of the basic gradient projection method, where to improve the convergence rate we use use the Barzilai-Borwein step-size $\alpha$ within a non-monotonic backtracking-Armijo line search. For minimizing a function $f(\mathbf{w})$ over a convex set $\mathcal{C}$, SPG uses simple iterations of the form

$$\mathbf{w}_{k+1} = \mathcal{P}_{\mathcal{C}}(\mathbf{w}_k - \alpha \nabla f(\mathbf{w}_k)).$$

The function $\mathcal{P}_{\mathcal{C}}(\mathbf{w})$ computes the Euclidean projection of a point $\mathbf{w}$ onto the convex set $\mathcal{C}$,

$$\mathcal{P}_{\mathcal{C}}(\mathbf{w}) = \arg\min_{\mathbf{x} \in \mathcal{C}} ||\mathbf{x} - \mathbf{w}||_2.$$

In our problem, the projection for a single constraint only affects the corresponding variables $\mathbf{w}_A^*$ and $g_A$, and can be written as the solution of

$$\arg\min_{\mathbf{x},y} ||[\mathbf{x}\ \ y] - [\mathbf{w}_A^*\ \ g_A]||_2, \ \ s.t.\ \ y \geq ||\mathbf{x}||_2.$$

This is isomorphic to Exercise 7.3(c) of (Boyd and Vandenberghe, 2004), whose solution is

$$\mathcal{P}_{\mathcal{C}}(\mathbf{w}_A^*, g_A) = \quad (6)$$
$$\begin{cases} (\mathbf{0}, 0), & \text{if } ||\mathbf{w}_A^*||_2 \leq -g_A, \\ (\mathbf{w}_A^*, g_A), & \text{if } ||\mathbf{w}_A^*||_2 \leq g_A, \\ \frac{1+g_A/||\mathbf{w}_A^*||_2}{2}(\mathbf{w}_A^*, ||\mathbf{w}_A^*||_2), & \text{if } ||\mathbf{w}_A^*||_2 > |g_A|. \end{cases}$$

Thus, it is simple to analytically compute the projection onto a single constraint.

### 5.2 Dykstra's Algorithm

If the groups were disjoint, we could simply compute the projection onto each constraint individually. But because the groups overlap this will not work. Thus, we would like to solve the problem of computing the projection onto a convex set defined by the intersection of sets, where we can efficiently project onto each individual set.

One of the earliest results on this problem is due to von Neumann (1950, §13), who proved that the limit of cyclically projecting a point onto two closed linear sets is the projection onto the intersection of the sets. Bregman (1965) proposed to cyclically project onto a series of general convex sets in order to find a point in their intersection, but this method will not generally converge to the projection. The contribution of Dykstra (1983) was to show that by taking the current iterate and removing the difference calculated from the previous cycle, then subsequently projecting this value, that the cyclic projection method converges to the optimal solution for general convex sets. Deutsch and Hundal (1994) have shown that Dykstra's algorithm converges at a geometric rate for polyhedral sets. Algorithm 1 gives pseudo-code for Dykstra's algorithm (we obtain Bregman's method if we fix $I_i$ at $\mathbf{0}$).

[2]We could apply the same procedure to solve (1), where we treat all groups $A$ with $\mathbf{w}_A = \mathbf{0}$ as boundary groups.

**Algorithm 1**: Dykstra's Cyclic Projection Algorithm

Input: Point $\mathbf{w}$ and sets $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_q$
Output: Projection of $\mathbf{w}$ onto $\mathcal{C} = \bigcap_{i=1}^{q} \mathcal{C}_i$
$\mathbf{w}_0 \leftarrow \mathbf{w}$
$\forall_i, I_i \leftarrow \mathbf{0}$
**while** $\mathbf{w}_j$ is changing by more than $\epsilon$ **do**
  **for** $i = 1$ to $q$ **do**
    $\mathbf{w}_j \leftarrow \mathcal{P}_{\mathcal{C}_i}(\mathbf{w}_{j-1} - I_i)$
    $I_i \leftarrow \mathbf{w}_j - (\mathbf{w}_{j-1} - I_i)$
    $j \leftarrow j + 1$
  **end for**
**end while**

Desipte its simplicity, Dykstra's algorithm is not widely used because of its high storage requirements. In its unmodified form, applying Dykstra's algorithm to compute the projection in (5) would be impractical, since for each group we would need to store a copy of the entire parameter vector. Fortunately, in (5) each constraint only affects a small subset of the variables. By taking advantage of this it is straightforward to derive a sparse variant of Dykstra's algorithm that only needs to store a copy of each variable for each group that it is associated with. This leads to an enormous reduction in the memory requirements. Further, although using Dykstra's algorithm rather than an analytic update leads to a relatively high iteration cost for the SPG method, SPG requires fewer function (and gradient) evaluations than previous methods for overlapping group $\ell_1$-regularization. This leads to a net computational gain in our framework since the function and graident evaluations will typically be much more expensive than applying Dykstra's algorihtm.

# 6 Large-Scale Applications

Thus far, we have ignored the potential intractability of evaluating the normalizing constant of the model, and the issue that the size of the factors grows exponentially with their order. In practice, both of these issues might need to be addressed in order to apply the methodology described here. This section outlines one possible solution for each of these issues.

## 6.1 Pseudo-Likelihood

In general, calculating the negative log-likelihood and its gradient will be intractable, since they require computing the logarithm of the normalizing constant and its gradient (respectively). To address this, we considered using a variant on Besag's pseudo-likelihood (Besag, 1975). Specifically, we considered optimizing the (regularized) product of the conditional distributions rather than the joint distribution,

$$\min_{\mathbf{w}} -\sum_{i=1}^{n}\sum_{j=1}^{p} \log p(x_j^i | \mathbf{x}_{-j}^i, \mathbf{w}) + \sum_{A \subseteq S} \lambda_A ||\mathbf{w}_{A^*}||_2.$$

This optimization takes the form of a set of multinomial logistic regression problems, each with overlapping group $\ell_1$-regularization. However, note that these multinomial logistic regression problems are not independent, since several parameters are shared across the problems. Because calculating the local normalizing constants in these individual problems is straightforward, the (convex) pseudo-likelihood approximation can be applied when the number of variables (or states) is large. Optimizing the regularized pseudo-likelihood instead of the joint likelihood involves only a trivial modification of the objective function passed to the SPG routine. Applying a variational approximation to the joint likelihood (see Wainwright and Jordan, 2008) would be an equally straightforward extension.

## 6.2 Weighted Ising Parameterization

One way to reduce the number of parameters present in the model is with a more parsimonious parameterization of the factors. In our experiments, we considered a weighted Ising-like parameterization. Here, each factor contains a weight for configurations where all nodes take the same state, but their is no dinstinction between other configurations. For the example in Section 2, the (log-)potentials would have the form

$$\phi_{3,4}(\mathbf{x}) = \mathbb{I}(x_3 = 1, x_4 = 1)w_1 + \mathbb{I}(x_3 = 2, x_4 = 2)w_2.$$

These potentials are far more parsimonious since each factor has only $k$ parameters (rather than $k^m$), but in using these potentials we lose the ability to model arbitrary discrete distributions. Nevertheless, these potentials capture the most important statistical regularities present in many data sets.

# 7 Experiments

Our experiments considered building generative models of the following data sets:

| Name | n | p | k | Source |
|---|---|---|---|---|
| VOC10 | 9963 | 10 | 2 | Everingham et al. |
| Jokes | 24983 | 10 | 2 | Goldberg et al. |
| Yeast | 2417 | 14 | 2 | Elisseeff and Weston |
| AWMA | 2602 | 16 | 2 | Qazi et al. |
| Flow | 5400 | 11 | 3 | Sachs et al. |
| VOC20 | 9963 | 20 | 2 | Everingham et al. |
| Traffic | 4413 | 32 | 4 | Chechetka and Guestrin |

For the VOC data set we concentrated on a model of the 10 most frequent labels (VOC10) as well as the full label set (VOC20). For the Jokes data set we used the 10 jokes rated by all users. We concentrated only on

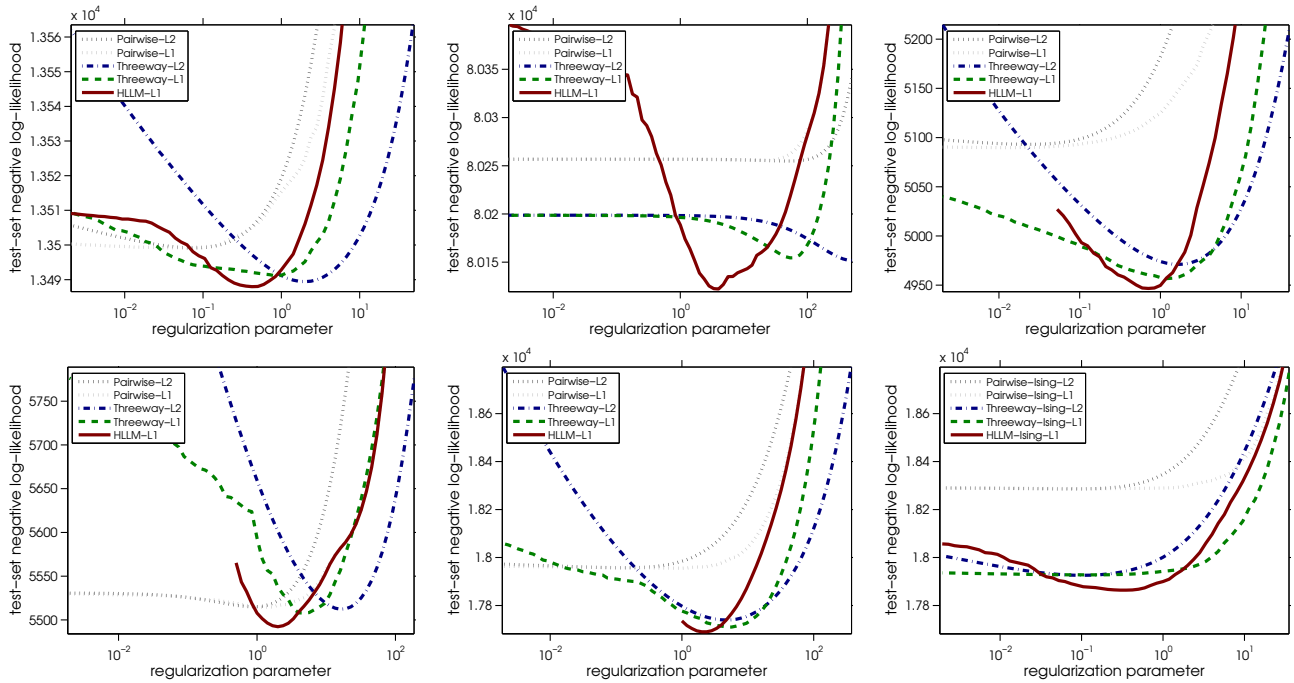


Figure 1: Test set negative log likelihood versus strength of regularizer for various methods and datasets. Top row, from left to right, the data sets are VOC10, Jokes, and Yeast. Bottom row, from left to right, the data sets are AWMA, Flow (full potentials), and Flow (Ising potentials).

the labels in the Yeast data set, while we ignored the interventions in the Flow data.

Our first experiment considered the first five data sets, where we used exact likelihood calculation and considered both the full and Ising parameterizations. On each data set we compared our hierarchical log-linear model with overlapping group $\ell_1$-regularization (labeled as HLLM-L1 in the figures) to fitting log-linear models restricted to both pairwise and threeway potentials with both $\ell_2$-regularization and group $\ell_1$-regularization. While the $\ell_2$-regularized models do not yield a sparse structure, we tested these models because they may still perform well at prediction. Note that unlike the pairwise and threeway models, an $\ell_2$-regularized version of the hierarchical log-linear model is infeasible. We trained on a random half of the data set, and tested on the remaining half as the regularization parameter $\lambda$ was varied. For the pairwise and threeway models, we set $\lambda_A$ to the constant $\lambda$. For the hierarchical model, we set $\lambda_A$ to $\lambda 2^{|A|-2}$, where $|A|$ is the cardinality of $A$. For all the models, we did not regularize the unary weights and we fixed the unary weight of one state to zero for each node.

The results of these experiments are plotted in Figure 1. For the Flow data set, we plot the results using both Ising and full potentials. For the other data sets the difference between parameterizations was very small so we only plotted the results with full potentials.

We see that threeway interactions seem to improve performance over pairwise models on many data sets, and that sometimes higher order interactions can offer a further advantage. The highest-order factor selected by the HLLM-L1 model with the lowest prediction error for each of the data sets was 4 (for Yeast), 5 (for Jokes, AWMA, and Sachs), and 6 (for VOC10). For the VOC10 and AWMA data sets, there was only a single factor of the highest order selected.

To test the scalability of our method we next examined the VOC20 and Traffic data sets, using the pseudo-likelihood and the Ising parameterization discussed in §6 (here, we also enforced a cardinality restriction of 4). Note that the previous paper that used (disjoint) group $\ell_1$ for learning log-linear models (Dahinden et al., 2007) only considered up to 5 variables, while (Dobra and Massam, 2008) considered a log linear model with up to 16 variables, but used stochastic local search to identify the structure. The latter uses iterative proportional fitting to fit the parameters of each (non-decomposable) model, which is computationally very expensive. We plot the test-set pseudo-likelihood against the regularization parameter for the different models in Figure 2, where we again see that relaxing the pairwise assumption leads to better predictive results. Our method can in principle be used to learn models with higher-order interactions on even larger data sets.

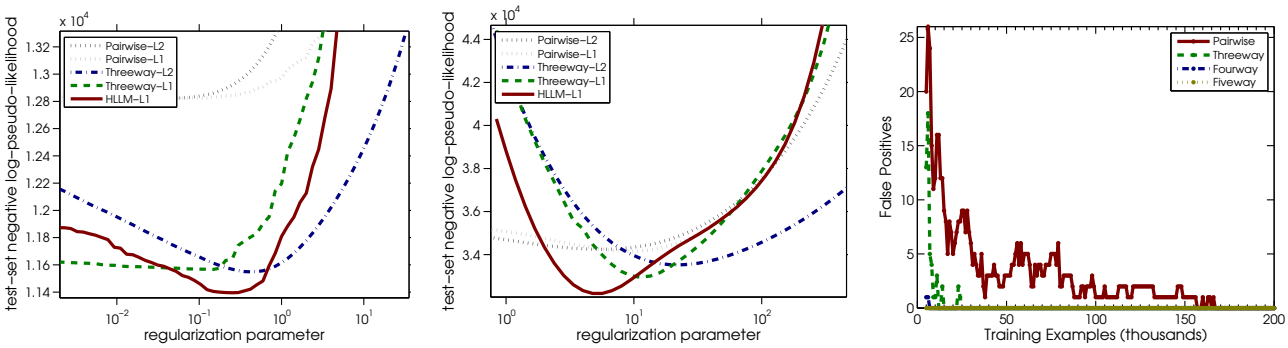


Figure 2: Test set negative log pseudo-likelihood versus strength of regularizer for VOC20 (left) and Traffic (middel) data sets. Right: False positives of different orders against training set size for the first model along the regularization path where the HLLM-L1 selects a superset of the true data-generating model.

We next sought to assess the performance of the HLLM-L1 model for structure estimation. We created a 10-node synthetic data set that includes all unary factors as well as the factors $(2,3)(4,5,6)(7,8,9,10)$ (a non-hierarchical model), where the model weights were generated from a $\mathcal{N}(0,1)$ distribution. In Figure 2 (right), we plot the number of false positives of different orders present in the first model along the regularization path that includes all three factors in the true structure against the number of training examples (we define a false positive as a factor where none of its supersets are present in the true model). For example, with 20000 samples the order of edge additions was (with false positives in square brackets) (8,10)(7,9)(9,10)(7,10)(4,5)(8,9)(2,3)(4,6)(8,9,10)(7,8) (7,8,9)(7,8,10)(5,6)[1,8][5,9][3,8][3,7](4,5,6)[1,7](7,9,10) (7,8,9,10) (at this point it includes all three factors in the true structure, with 5 pairwise false positives and no higher-order false positives). In the figure, we see that the model tends to include false positives before it adds all true factors, but the number decreases as the sample size increases. Further, there tend to be few higher-order false positives; although it includes spurious pairwise factors even with 150000 samples, the model includes no spurious threeway factors beyond 30000 samples, no spurious fourway factors beyond 10000 samples, and no fiveway factors for any sample size (the plot begins at 5000).

We next examined the coronary heart disease data set analyzed in (Edwards and Havránek, 1985). The first fifteen factors added along the HLLM-L1 regularization path on this data set are: (B,C)(A,C)(B,E)(A,E)(C,E)(D,E)(A,D)(B,F)(E,F) [C,D][A,F](A,D,E)(D,F)[D,E,F][A,B]. The first seven factors are the union of the minimally sufficient hierarchical models from the analysis by Edwards and Havránek. These are also the factors with posterior mode greater than 0.5 for a prior strength of 2 and 3 in the hierarchical models of (Dobra and Massam, 2008), while the first eight are the factors selected with a prior strength of 32 and 64. With a prior strength of 128 Dobra and Massam (2008) find the ninth factor introduced by our model, as well as the factor (D,F) introduced later. The remaining factor with this prior strength is the factor (B,C,F), that is not found until much later in the regularization path in our model. In contrast, the first three-way factor introduced by our model is (A,D,E). This factor is present in both of the accepted graphical model in (Edwards and Havránek, 1985), and is the only threeway factor with a posterior greater than 0.5 (under a Laplace approximation) in the graphical models of (Dobra and Massam, 2008) for a prior strength of 1, 2, 3, 32, and 64. The other factors (that we denoted with square brackets) are not recognized in the prior work, and may represent false positives due to the use of a point estimate with this small sample size.

## 8 Discussion

Jacob et al. (2009) consider a different notion of overlapping groups to encourage a sparsity pattern that is a union of groups. They represent each variable as a combination of auxiliary variables and penalize these (disjoint) variables. We could enforce hierarchical inclusion in this framework by adding to each group all *subsets* of the group, as opposed to all supersets in (2). An advantage of this is that the projection is given explicitly by (6), but a disadvantage is that it would be grossly over-parameterized (we would have an auxiliary variable for every subset of each non-zero factor).

We proposed an efficient way to learn sparse log-linear models with higher-order interactions using overlapping group $\ell_1$-regularization, that uses an active set method and Dykstra's algoirthm within an SPG routine. Our experiments indicate that the model gives improved predictive performance on several data sets.

The SPG algorithm may be useful for other problems with overlapping group $\ell_1$-regularization, while Dykstra's algorithm could alternately be used within the optimal method of Nesterov (2003) or projected quasi-Newton methods (Schmidt et al., 2009). The main outstanding issue in this work is deriving an efficient way to test (or bound) sufficient optimality conditions for (2) as in (Bach, 2008), and deriving an efficient search for sub-optimal inactive groups.

## Appendix

To show that minimizers of (2) satisfy hierarchical inclusion, assume we have a minimizer $\tilde{\mathbf{w}}$ of (2) that does not. Then there exists some $A$ such that $\tilde{\mathbf{w}}_A = \mathbf{0}$ and some $B$ such that $A \subset B$ and $\tilde{\mathbf{w}}_B \neq \mathbf{0}$. This implies group $A$ is active and must satisfy (3). Using $\tilde{\mathbf{w}}_A = \mathbf{0}$, we have that $||\nabla_{\mathbf{w}_A} \log p(\mathbf{x}|\tilde{\mathbf{w}})||_2$ is exactly 0, and assuming this does not happen by chance it contradicts that $\tilde{\mathbf{w}}_A$ is a minimizer.

## References


F. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. *NIPS*, 2008.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *JMLR*, 2008.

D. Bertsekas, A. Nedic, and A. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

E. Birgin, J. Martínez, and M. Raydan. Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim.*, 10(4):1196–1211, 2000.

Y. Bishop, S. Fienberg, and P. Holland. *Discrete Multivariate Analysis*. MIT Press, 1975.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

L. Bregman. The method of successive projection for finding a common point of convex sets. *Dokl. Akad. Nauk SSSR*, 162(3):487–490, 1965. English translation in Soviet Math. Dokl., 6:688–692, 1965.

A. Chechetka and C. Guestrin. Efficient principled learning of thin junction trees. *NIPS*, 2007.

C. Dahinden, G. Parmigiani, M. Emerick, and P. Bühlmann. Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinf.*, 8(476), 2007.

F. Deutsch and H. Hundal. The rate of convergence of Dykstra's cyclic projections algorithm: The polyhedral case. *Numer. Funct. Anal. Optim.*, 15(5):537–565, 1994.

A. Dobra and H. Massam. The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. Technical report, University of Washington, 2008.

R. Dykstra. An algorithm for restricted least squares regression. *JASA*, 78(384):837–842, 1983.

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

A. Elisseeff and J. Weston. Kernel methods for multi-labelled classification and categorical regression problems. In *NIPS*, 2002.

M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results, 2007.

M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE J. Sel. Top. Sign. Proces.*, 1(4):586–597, 2007.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retrieval*, 4(2):133–151, 2001.

H. Höfling and R. Tibshirani. Estimation of sparse binary pairwise Markov networks using pseudo-likelihoods. *JMLR*, 10:883–906, 2009.

L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009.

S. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using l1-regularization. *NIPS*, 2006.

Y. Nesterov. *Introductory Lectures on Convex Optimization*. Kluwer Academic Publishers, 2003.

M. Qazi, G. Fung, S. Krishnan, R. Rosales, H. Steck, R. Rao, D. Poldermans, and D. Chandrasekaran. Automated heart wall motion abnormality detection from ultrasound images using Bayesian networks. *IJCAI*, 2007.

K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308 (5721):523–529, 2005.

M. Schmidt, K. Murphy, G. Fung, and R. Rosales. Structure learning in random fields for heart motion abnormality detection. *CVPR*, 2008.

M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-Newton algorithm. *AISTATS*, 2009.

E. van den Berg and M. Friedlander. Probing the Pareto frontier for basis pursuit solutions. *SIAM J. Sci. Comput.*, 31(2):890–912, 2008.

J. von Neumann. *Functional Operators, vol. II, The Geometry of Orthogonal Spaces*, volume 22 of *Annals of Mathematical Studies*. Princeton University Press, 1950. This is a reprint of notes first distributed in 1933.

M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, 2008.

M. Wainwright, P. Ravikumar, and J. Lafferty. High-dimensional graphical model selection using l1-regularized logistic regression. *NIPS*, 2006.

J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, 1990.

P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics*, 37(6A):3468–3497, 2009.