On Learning Discrete Graphical Models using Group-Sparse Regularization

Ali Jalali
ECE, UT Austin
alij@mail.utexas.edu

Pradeep Ravikumar
CS, UT Austin
pradeepr@cs.utexas.edu

Vishvas Vasuki
CS, UT Austin
vvasuki@cs.utexas.edu

Sujay Sanghavi
ECE, UT Austin
sanghavi@mail.utexas.edu

Abstract

We study the problem of learning the graph structure associated with a general discrete graphical models (each variable can take any of $m > 1$ values, the clique factors have maximum size $c \geq 2$) from samples, under high-dimensional scaling where the number of variables p could be larger than the number of samples n. We provide a quantitative consistency analysis of a procedure based on node-wise multi-class logistic regression with group-sparse regularization.

We first consider general m-ary pairwise models – where each factor depends on at most two variables. We show that when the number of samples scale as $n > K(m-1)^2 d^2 \log((m-1)^2(p-1))$– where d is the maximum degree and K a fixed constant – the procedure succeeds in recovering the graph with high probability. For general models with c-way factors, the natural multi-way extension of the pairwise method quickly becomes very computationally complex. So we studied the effectiveness of using the pairwise method even while the true model has higher order factors. Surprisingly, we show that under slightly more stringent conditions, the pairwise procedure *still* recovers the graph structure, when the samples scale as $n > K(m-1)^2 d^{\frac{3}{2}c-1} \log((m-1)^c (p-1)^{c-1})$.

1 Introduction

Markov Random Fields and Structure Learning. Undirected graphical models, also known as Markov random fields, are used in a variety of domains, including statistical physics [14], natural language processing [19], image analysis [35, 13, 6], and spatial statistics [27], among others. A Markov random field (MRF) over a p-dimensional discrete random vector $X = (X_1, X_2, \ldots, X_p)$ is specified by an undirected graph $G = (V, E)$, with vertex set $V = \{1, 2, \ldots, p\}$ – one for each variable – and edge set $E \subset V \times V$. The structure of this graph encodes certain conditional independence assumptions among subsets of the variables. In this paper, we consider the task of structure learning, i.e. estimating the underlying graph structure associated with a general discrete Markov random field from n independent and identically distributed samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$.

High-dimensional setting and Group sparsity. We are interested in structure learning in the setting where the dimensionality p of the data is larger than the number of samples n. While classical procedures typically break down under such high-dimensional scaling, an active line of recent research has shown it is still possible to obtain practical consistent procedures by leveraging low-dimensional structure. The most popular example is that of leveraging sparsity using ℓ_1-regularization (e.g., [4, 12, 21, 23, 31, 34, 37]). For MRF structure learning, such ℓ_1-regularization has been successfully used for Gaussian [21] and discrete binary pairwise (i.e. Ising) models [26, 17]. In these instances, there is effectively only one parameter per edge, so that a sparse graph corresponds to a sparse set of parameters. In this paper, we are interested in more general discrete graphical models – where each variable can take m possible values, and factors can be of order higher than two. We now have multiple parameters per edge, and thus the relevant low-dimensional structure is that of *group sparsity*: all parameters of an edge form a group, and a sparse graph now corresponds to certain *groups of parameters* being non-zero. The counterpart of ℓ_1 regularization for such group-sparse structure is ℓ_1/ℓ_q regularization for $q > 1$, where we collate the ℓ_q norms of the groups, and compute their overall ℓ_1 norm. Recent work on group and block-sparse linear

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15. Copyright 2011 by the authors.

regression [32, 8, 22, 18, 24, 25, 2] show that under such group-sparse settings, group-sparse regularization outperforms the use of ℓ_1 penalization.

Our Results: Pairwise m-ary models. In this paper, we provide a quantitative consistency analysis of group-sparse regularized structure recovery for general discrete graphical models. We first consider the case of pairwise but otherwise m-ary discrete graphical models, and analyze a group-sparse variant of the procedures in [26, 21]: for each vertex $r \in V$, we estimate its neighborhood set using ℓ_1/ℓ_2-regularized maximum conditional likelihood. This reduces to multi-class logistic regression, for which we characterize the number of samples needed for *sparsistency i.e.* consistent recovery of the group-support-set with high probability. This analysis extends recent high-dimensional analyses for linear models to logistic models, and is of independent interest even outside the context of graphical models. We then combine the neighborhood sets across vertices to form the graph estimate. There has been a strong line of work on developing fast algorithms to solve these sparse multiclass logistic regression programs including Meier et al. [20], Krishnapuram et al. [15]. Indeed, [9, 10] show good empirical performance using such ℓ_1/ℓ_q regularization even with the joint likelihood over all variables.

Our Results: General m-ary models. One (natural, but expensive) extension to graphical models with higher-order factors is to again use group-sparse regularization but with higher order factors as groups. However, this leads to prohibitive computational complexity – e.g. there are $\mathcal{O}(p^c)$ possible factors of order c. Indeed, in their empirical study of such regularizations, Dahinden et al. [9, 10] could scale up to small graph sizes, even while using some intelligent heuristics. This motivates our second main result. Suppose we solve the pairwise graphical model estimation problem, even when the true model has higher order factors. What is the relationship of this estimate with the true underlying graph? We investigate this for *hierarchical* graphical models where the absence of any lower-order factor also implies the absence of factors over supersets of the lower-order factor variables. Higher-order factors could, in principle, cause our pairwise estimator to include spurious edges. Surprisingly, we obtain the result that under slightly more stringent assumptions on the scaling of the sample size (dependent on the size of the higher-order factors) the pairwise estimator excludes the irrelevant edges, and includes all "dominant" pairwise edges whose parameters are larger than a certain threshold that depends on the size of the parameters values of higher-order factors. As a consequence, if all pairwise effects are dominant enough, we recover the graph exactly even while using a simple pairwise estimator. But even otherwise, the guaranteed false edge exclusion could be used for further greedy procedures, though we defer further discussion in the sequel.

Existing approaches. Methods for estimating such graph structure include those based on constraint and hypothesis testing [29], and those that estimate restricted classes of graph structures such as trees [5], polytrees [11], and hypertrees [30]. Another class of approaches estimate the local neighborhood of each node via exhaustive search for the special case of bounded degree graphs. Abbeel et al.[1] propose a method for learning factor graphs based on local conditional entropies and thresholding, but the computational complexity grows at least as quickly as $\mathcal{O}(p^{d+1})$, where d is the maximum neighborhood size in the graphical model. Bresler et al. [3] describe a related local search-based method, and prove under relatively mild assumptions that it can recover the graph structure with $\Theta(\log p)$ samples. However, in the absence of additional restrictions, the computational complexity of the method is $\mathcal{O}(p^{d+1})$. Csiszár and Talata [7] show consistency of a method that uses pseudo-likelihood and a modification of the BIC criterion, but this also involves a prohibitively expensive search.

2 Problem Setup and Notation

MRFs and their Parameterization. We consider the task of estimating the graph structure associated with a general discrete Markov random field. Let $X = (X_1, \ldots, X_p)$ be a random vector, each variable X_i taking values in a discrete set $\mathcal{X} = \{1, 2, \ldots, m\}$ of cardinality m. Let $G = (V, E)$ denote a graph with p nodes, corresponding to the p variables $\{X_1, \ldots, X_p\}$. Let $\mathcal{C}$ be a set of cliques (fully-connected subgraphs) of the graph G, and let $\{\phi_C : \mathcal{X}^{|C|} \mapsto \mathbb{R},\ C \in \mathcal{C}\}$ be a set of "clique potential" functions. With this notation, the distribution of X takes the form

$$\mathbb{P}(x) \quad \propto \quad \exp\left\{\sum_{C\in\mathcal{C}} \phi_C(x_C)\right\}. \tag{1}$$

Since $\mathcal{X}$ is discrete, each potential function ϕ_C can be parameterized as linear combinations of $\{0,1\}$-valued indicator functions – one for each configuration of x_C. For each $s \in V$ and $j \in \{1, \ldots, m-1\}$, we can define node-wise indicators,

$$\mathcal{I}[x_s = j] \quad = \quad \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{otherwise.} \end{cases}$$

Note that we omit an indicator for $x_s = m$ from the list, since it is redundant given the indicators for $j = 1, \ldots, m-1$. In a similar fashion, we can define the

$|C|$-way clique-wise indicator functions $\mathcal{I}[x_C = v]$, for $v \in \{1, 2, \ldots, m-1\}^{|C|}$.

With this notation, any set of potential functions can then be written as

$$\phi_C(x_C) \quad = \sum_{v \in \{1,\ldots,m-1\}^{|C|}} \theta^*_{C;v}\, \mathcal{I}[x_C = v] \quad \text{for } C \in \mathcal{C}$$

Thus, (1) can be rewritten as,

$$\mathbb{P}_{\theta^*}(x) \propto \exp\bigg\{ \sum_{C \in \mathcal{C}; v \in \{1,\ldots,m-1\}^{|C|}} \theta^*_{C;v}\, \mathcal{I}[x_C = v] \bigg\}. \tag{2}$$

Thus, the Markov random field can be parameterized in terms of the collection of tensors $\theta^* := \{\theta^*_{C;v}\ C \in \mathcal{C}; v \in \{1, \ldots, m-1\}^{|C|}\}$. In the sequel, it will be useful to collate these into vectors $\theta^*_C \in \mathbb{R}^{(m-1)^{|C|}}$ associated with the cliques $C \in \mathcal{C}$.

Pairwise Markov Random Fields. Here the set of cliques consists of the set of nodes V and the set of edges E. Thus, using nodewise and pairwise indicator functions as before, any pairwise MRF over $(X_1, \ldots, X_p)$ can be expressed as

$$\mathbb{P}(x) \propto \exp\bigg\{ \sum_{s \in V; j \in \{1,\ldots,m-1\}} \theta_{s;j} \mathcal{I}[x_s = j] + \sum_{(s,t) \in E; j,k \in \{1,\ldots,m-1\}} \theta_{st;jk} \mathcal{I}[x_s = j,\, x_t = k] \bigg\}, \tag{3}$$

for a set of parameters $\theta^* := \{\theta^*_{s;j}, \theta^*_{st;jk} : s, t \in V;\, (s,t) \in E;\, j,k \in \{1, \ldots, m-1\}\}$. It will be useful to collate these into vectors $\theta^*_s \in \mathbb{R}^{m-1}$ for each $s \in V$, and the vectors $\theta^*_{st} \in \mathbb{R}^{(m-1)^2}$ associated with each edge.

Graphical Model Selection. Suppose that we are given a collection $D := \{x^{(1)}, \ldots, x^{(n)}\}$ of n samples, where each p-dimensional vector $x^{(i)} \in \{1, \ldots, m\}^p$ is drawn i.i.d. from a distribution $\mathbb{P}_{\theta^*}$ of the form (2), for parameters θ^* and graph $G = (V, E^*)$ over the p variables. The goal of *graphical model selection* is to infer the edge set E^* of the graphical model defining the probability distribution that generates the samples. Note that the true edge set E^* can also be expressed as a function of the parameters as

$$E^* = \{(s,t) \in V \times V : \exists C \in \mathcal{C}; \{s,t\} \in C;\ \theta^*_C \neq 0\}. \tag{4}$$

In this paper, we focus largely on the special case of pairwise Markov random fields.

2.1 Pairwise Model Selection

We now describe the graph selection procedure we study for the m-ary pairwise model. It is the natural generalization of the procedures for binary graphical models [26] and Gaussian graphical models [21]. Specifically, we first focus on recovering the neighborhood of a fixed vertex $r \in V$, and then combine the neighborhood sets across vertices to form the graph estimate.

Let us define the vector $\Theta^*_{\backslash r} \in \mathbb{R}^{(m-1)^2(p-1)}$, which is the concatenation of $(p-1)$ *groups* – i.e. one (short) vector $\theta^*_{rt} \in \mathbb{R}^{(m-1)^2}$ for each $t \in V \backslash \{r\}$. Note that r having a small neighborhood is equivalent to many of these vectors θ^*_{rt} being zero; in particular, the problem of neighborhood estimation for vertex r corresponds to the recovery of the set

$$\mathcal{N}(r) \quad = \quad \Big\{ u \in V \backslash \{r\} \mid \|\theta^*_{ru}\|_0 \neq 0 \Big\}.$$

This is precisely the structure captured by *group-sparsity.* In particular, each θ^*_{rt}, with $t \in V\backslash\{r\}$, corresponds to a group; if r has a small neighborhood,only few of these groups will be non-zero.

In order to estimate the neighborhood $\mathcal{N}(r)$, we thus perform a regression of X_r on the rest of the variables $X_{\backslash r}$, using the group-sparse regularizer $\left\|\Theta_{\backslash r}\right\|_{1,2} := \sum_{u \in V\backslash\{r\}} \|\theta_{ru}\|_2$. The conditional distribution of X_r given the other variables $X_{\backslash r} = \{X_t \mid t \in V\backslash\{r\}\}$ takes the form

$$\mathbb{P}_{\Theta^*}\left[X_r = j \mid X_{\backslash r} = x_{\backslash r}\right] = \frac{\exp\left(\theta^*_{r;j} + \sum_{t \in V\backslash\{r\}} \sum_k \theta^*_{rt;jk} \mathcal{I}[x_t = k]\right)}{1 + \sum_\ell \exp\left(\theta^*_{r;\ell} + \sum_{t \in V\backslash\{r\}} \sum_k \theta^*_{rt;\ell k} \mathcal{I}[x_t = k]\right)}, \tag{5}$$

for all $j \in \{1, \ldots, m-1\}$. Thus, X_r can be viewed as the response variable in a multiclass logistic regression, in which the indicator functions associated with the other variables

$$\Big\{ \mathcal{I}[x_t = k],\, t \in V\backslash\{r\},\ k \in \{1, 2, \ldots, m-1\} \Big\},$$

play the role of the covariates.

Thus, we study the following convex program as an estimate for Θ^*_r

$$\hat{\Theta}_{\backslash r} \in \min_{\Theta_{\backslash r} \in \mathbb{R}^{(m-1)^2(p-1)}} \Big\{ \ell(\Theta_{\backslash r}; D) + \lambda_n \left\|\Theta_{\backslash r}\right\|_{1,2} \Big\}, \tag{6}$$

where $\ell(\Theta_{\backslash r}; D) = \frac{1}{n}\sum_{i=1}^n \ell^{(i)}(\Theta_{\backslash r}; D) := \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}_\Theta\left[X_r = x_r^{(i)} \mid X_{\backslash r} = x_{\backslash r}^{(i)}\right]$ is the rescaled multiclass logistic likelihood defined by the conditional distribution (5), and $\lambda_n > 0$ is a regularization parameter. The convex program (6) is an ℓ_1/ℓ_2-regularized multiclass logistic regression problem, and is thus the multiclass logistic analog of the group Lasso [36].

The solution to the program (6) yields an estimate $\widehat{\mathcal{N}}(r)$ of the neighborhood of node r by

$$\widehat{\mathcal{N}}(r) = \{t \in V : t \neq r; \|\hat{\theta}_{rt}\|_2 \neq 0\}.$$

We are interested in the event that all the node neighboorhoods are estimated exactly, $\{\widehat{\mathcal{N}}(r) = \mathcal{N}(r);\ \forall r \in V\}$, which we also write as $\{\hat{E} = E^*\}$ since it entails that the the full graph is estimated exactly.

Sparsistency. Our main result is a high-dimensional analysis of the estimator (6), where allow the problems dimensions such as the number of nodes p, the maximum node degree d, the size of the state space m (and in the case of higher-order MRFs, the maximum clique size c) to vary with the number of observations n. Our goal is to establish sufficient conditions on the scaling of (n, p, d, m, c) such that our proposed estimator is consistent in the sense that

$$\mathbb{P}\left[\widehat{E}_n = E^*\right] \quad \rightarrow \quad 1 \qquad \text{as } n \rightarrow +\infty.$$

We sometimes call this property sparsistency, as a shorthand for consistency of the sparsity pattern of the parameters.

2.2 Higher-order Model Selection

Natural, high-complexity Extension. Let us first see what this model selection recipe of node-wise regression with group-sparse regularization, would entail when extended to the general higher-order Markov random fields (2) case. Recall that such a higher-order MRF is parameterized by vectors $\theta^*_C \in \mathbb{R}^{(m-1)^{|C|}}$ for $C \in \mathcal{C}$. Let c be the maximum clique size. It would be convenient to view the parameters as a collection of $\sum_{j=1}^{c} \binom{p}{j}$ vectors indexed by a cliques C of size less than or equal to c, but non-zero if and only if the clique $C \in \mathcal{C}$.

Again, we fix a node r, and define the long vector $\Theta^*_{\backslash r} \in \mathbb{R}^{\sum_{j=1}^{c-1} \binom{p-1}{j} (m-1)^{j+1}}$ as the concatenation of the parameter vectors θ^*_{rC} for all $C \subseteq V \backslash r$; $|C| < c$. Note that recovery of the neighborhood of a vertex r corresponds to the recovery of the set

$$\mathcal{N}(r) \quad = \quad \left\{ u \in V \backslash \{r\} \ \mid\ \exists\, C \subseteq V \backslash \{r, u\};\ \|\theta^*_{ruC}\|_0 \neq 0 \right\}.$$

Thus, we could again make use of group sparsity where in this case, the groups of parameters are the parameter vectors θ^*_{rC} for different $C \subseteq V \backslash r$; $|C| < c$. We can then see that a small neighborhood $\mathcal{N}(r)$ for node r entails that $\Theta^*_{\backslash r}$ will have many of these groups be zero. The group-structured penalty would then take the form $\|\Theta^*_{\backslash r}\|_{1,2} := \sum_{\{C \subseteq V \backslash r\, |C| < c\}} \|\theta^*_{rC}\|_2$.

Thus we would solve:

$$\min_{\Theta_{\backslash r} \in \mathbb{R}^{\sum_{j=1}^{c-1} \binom{p-1}{j} (m-1)^{j+1}}} \left\{ \ell(\Theta_{\backslash r}; D) + \lambda_n \left\|\Theta_{\backslash r}\right\|_{1,2} \right\}, \tag{7}$$

where $\ell(\Theta_{\backslash r}; D)$ is the likelihood of the data as before. Dahinden et al. [9, 10] studied the related program of ℓ_1/ℓ_2 regularized maximum likelihood over the complete graph (instead of node-wise regressions) but showed good empirical performance of discrete graphical model structure recovery. The caveat with the higher-order group-sparse approach is the prohibitive computational complexity of this procedure. Note that the number of parameters is $\sum_{j=1}^{c-1} \binom{p-1}{j} (m-1)^{j+1}$ which scales prohibitively even for moderate c. Indeed, even the computations in the pairwise case are not inexpensive.

Sparsistency of a Simpler Estimate. But as we show in Section 4, even when the underlying model is a higher order MRF, surprisingly just solving the pairwise program (6) is *sufficient* to recover the true edges, under certain conditions. Thus, in our second main result, we again analyze the sparsistency of the estimator in (6), but for the case where the underlying graph is a higher-order MRF.

2.3 Notation

We use the following notation for group-structured norms. For any vector $u \in \mathbb{R}^p$ where $\{1, \ldots, p\}$ is partitioned into a set of T disjoint groups $\mathcal{G} = \{G_1, \ldots, G_T\}$, we define $\|u\|_{\mathcal{G},a,b} = \|(\|u_{G_1}\|_a, \ldots, u_{G_T}\|_a)\|_b$. In our case, for the pairwise model, the nodewise regression has the parameter vector $\Theta^*_{\backslash r} \in \mathbb{R}^{(m-1)^2(p-1)}$. Its groups are collated on the edges: $\mathcal{G} = \{\mathcal{G}_{rs}; s \in V \backslash r\}$ where $\mathcal{G}_{rt}$ is the index set of parameters on the (r, t) edge, $\{\theta_{rt;jk};\ j, k \in \{1, \ldots, m-1\}\}$.

Similarly, suppose $\Theta^*_{\backslash r}$ is the nodewise regression parameter for the higher-order model case. Then its groups are collated on the cliques: $\mathcal{G} = \{\mathcal{G}_{rC}; C \subseteq V \backslash r\, |C| < c\}$, where $\mathcal{G}_{rC}$ is the index set of parameters on the $r \cup C$ clique, $\{\theta_{rC;jv};\ j \in \{1, \ldots, m-1\}; v \in \{1, \ldots, m-1\}^{|C|}\}$. In the sequel, we will suppress the dependence of the group norms on these group partitions $\mathcal{G}$ when it is clear from context, so that we will simply use $\|\Theta^*_{\backslash r}\|_{a,b}$ for $\|\Theta^*_{\backslash r}\|_{\mathcal{G},a,b}$.

We will be focusing on the choice $a = 1, b = 2$ which yields the group-lasso penalty [36]. For a matrix $M \in \mathbb{R}^{p \times p}$, and denoting the i-th row of M by M^i, we can define the analogs of the group-structured norms on matrices: $\|M\|_{(a,b),(c,d)} := \|(\|M^1\|_{c,d}, \ldots, \|M^p\|_{c,d})\|_{a,b}$. In our analysis, we will always use $b = d = 2$, so that we use the minimized notation: $\|M\|_{a,c}$ to denote $\|M\|_{(a,2),(c,2)}$.

3 Pairwise Discrete Graphical Models

Let $S_r = \{u \in V : (r,u) \in E\}$ be the set of all neighbors of the node r in the graph and $S_r^c = V \backslash S_r$. Notice that $\|\theta^*_{ru}\|_0 = 0$ for all $u \in S_r^c$. Fixing $r \in V$, and defining $\Theta^*_{\backslash r}$ as before, let $S_r^{(ex)}$ be the index set of parameters $\{\theta^*_{rt;jk} \neq 0\}$ in $\Theta^*_{\backslash r}$. When clear from context, we will overload notation and again use S_r for this index set.

Let $Q^* = \mathbb{E}\left[\nabla^2 \log\left(\mathbb{P}_{\Theta^*_{\backslash r}}\left[X_r | X_{\backslash r}\right]\right)\right]$ be the population Fisher information matrix. Note that $Q^* \in \mathbb{R}^{(m-1)^2(p-1)\times(m-1)^2(p-1)}$. Similarly, let $Q^n = \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell^{(i)}\left(\Theta_{\backslash r}; D\right)$ be the sample Fisher information matrix.

Define $\mathcal{J}^* = \mathbb{E}\left[\mathcal{I}\left[x_{t_2} = k_2\right]\mathcal{I}\left[x_{t_1} = k_1\right]^T\right] \in \mathbb{R}^{(m-1)(p-1)\times(m-1)(p-1)}$. Accordingly, define $\mathcal{J}^n$ to be the empirical mean of the same quantity over n drawn samples. In the proofs (specifically in analyzing the derivative of the Hessian of the log-likelihood function), we will actually need control over $\Im^* := \mathcal{J}^* \otimes \mathbf{1}_{(m-1)\times(m-1)}$, the Kronecker product of $\mathcal{J}^*$ and matrix of all ones (which would be of the size of Q^*). But by properties of Kronecker products, we have $\Lambda_{\max}(\Im^*) = \Lambda_{\max}(\mathcal{J}^*)$, so that it suffices to impose assumptions on the maximum eigen values of $\mathcal{J}^*$ and $\mathcal{J}^n$.

3.1 Assumptions

We begin by stating the assumptions imposed on the true model. We note that similar sufficient conditions have been imposed in papers analyzing Lasso [33] and block-regularization methods [22, 24].

(A1) **Invertibility:** $\Lambda_{\min}\left(Q^*_{S_rS_r}\right) \geq C_{\min} > 0$.

(A2) **Incoherence:** $\left\|Q^*_{S_r^cS_r}\left(Q^*_{S_rS_r}\right)^{-1}\right\|_{\infty,2} \leq \frac{1-2\alpha}{\sqrt{d_r}}$ for some $\alpha \in \left(0, \frac{1}{2}\right)$.

(A3) **Boundedness:** $\Lambda_{\max}\left(\mathcal{J}^*\right) \leq D_{\max} < \infty$.

The next lemma states that imposing these assumptions on the population quantities implies analogous conditions on the sample statistics with high probability.

Lemma 1. *Assumptions (A1)-(A3) on the population Fisher information matrix yield the following (analogous) properties on the empirical Fisher information matrix:*

(B1) $\mathbb{P}\left[\Lambda_{\min}\left(Q^n_{S_rS_r}\right) < C_{\min} - \epsilon\right]$
$\leq 2\,\exp(-\frac{1}{8}(\epsilon\sqrt{n} - \sqrt{d_r})^2 + \log((m-1)^2 d_r))$.

(B2) $\mathbb{P}\left[\left\|Q^n_{S_r^cS_r}\left(Q^n_{S_rS_r}\right)^{-1}\right\|_{\infty,2} > \frac{1-\alpha}{\sqrt{d_r}} + \epsilon\right]$
$\leq 6\,\exp(-\frac{1}{8}(\bar{C}_{\min}(\frac{\alpha}{3\sqrt{d_r}} + \epsilon)\sqrt{n}$
$-\,(1 + \frac{\sqrt{d_r}}{C^2_{\min}\sqrt{n}})\sqrt{d_r})^2 + \log((m-1)^2(p-1)))$.

(B3) $\mathbb{P}\left[\Lambda_{\max}\left(\mathcal{J}^n\right) > D_{\max} + \epsilon\right] \;\leq\; 2\exp\Big(-\frac{1}{8}(\epsilon\sqrt{n} - \sqrt{d_r})^2 + \log\left((m-1)^2 d_r\right)\Big)$.

3.2 Main Theorem

We can now state our main result on the sparsistency of the group-sparse regularized estimator.

Theorem 1. *Consider a discrete graphical model of the form* (3) *with parameters* Θ^* *and associated edge set* E *such that conditions (A1)-(A3) are satisfied. Suppose the regularization parameter satisfies*

$$\lambda_n \geq \frac{8(2-\alpha)}{\alpha}\left(\sqrt{\frac{\log(p-1)}{n}} + \frac{m-1}{4\sqrt{n}}\right). \qquad (8)$$

Then, there exist positive constants K, c_1 *and* c_2 *such that if the number of samples* n *scales as*

$$n \geq K(m-1)^2 d_r^2 \log\left((m-1)^2(p-1)\right), \qquad (9)$$

then with probability $1 - c_1\exp(-c_2\lambda_n^2 n)$ *we are guaranteed*

(a) For each node $r \in V$, *the* ℓ_1/ℓ_2 *regularized logistic regression* (6) *has a unique solution and hence specifies a neigborhood* $\widehat{\mathcal{N}}(r)$.

(b) For each node $r \in V$ *correctly excludes all edges not in the true neighborhood* $\mathcal{N}(r)$. *Moreover, it includes all edges* (r,t) *such that* $\left\|\theta^*_{rt;jk}\right\|_2 \geq \frac{10}{C_{\min}}\lambda_n$.

Before sketching the proof outline, we first state some lemmas characterizing the solution of (3).

Lemma 2 (Optimality Conditions). *Any optimal primal-dual pair* $(\hat{\Theta}_{\backslash r}, \hat{Z}_{\backslash r})$ *of* (3) *satisfies*

1. **(Stationary Condition).**

$$\nabla\ell\left(\hat{\Theta}_{\backslash r}\right) + \lambda_n\hat{Z}_{\backslash r} = 0. \qquad (10)$$

2. **(Dual Feasibility).** *$\hat{Z}_{\backslash r}$ is equal to the subgradient $\partial\|\hat{\Theta}_{\backslash r}\|_{1,2}$ so that for any $u \in V\backslash r$,*

   (a) if $(\hat{\Theta}_{\backslash r})_{u;jk} \neq 0$ for some j, k then

   $$(\hat{Z}_{\backslash r})_u = \frac{(\hat{\Theta}_{\backslash r})_u}{\|(\hat{\Theta}_{\backslash r})_u\|_2}.$$

   (b) if the entire group $(\hat{\Theta}_{\backslash r})_u = 0$, then $\|(\hat{Z}_{\backslash r})_u\|_2 \leq 1$.

The next lemma states that structure recovery is guaranteed if the dual is *strictly* feasible.

Lemma 3 (Strict Dual Feasibility). *Suppose that there exists an optimal primal-dual pair $\left(\hat{\Theta}_{\backslash r}, \hat{Z}_{\backslash r}\right)$ for (6) such that $\left\|\left(\hat{Z}_{\backslash r}\right)_{S_r^c}\right\|_{\infty,2} < 1$. Then, any optimal primal solution $\widetilde{\Theta}_{\backslash r}$ satisfies $\left(\widetilde{\Theta}_{\backslash r}\right)_{S_r^c} = \mathbf{0}$. Moreover, if the Hessian sub-matrix $\left[\nabla^2 \ell\left(\hat{\Theta}_{\backslash r}\right)\right]_{S_r S_r} \succ 0$ then $\hat{\Theta}_{\backslash r}$ is the unique optimal solution.*

We are now ready to sketch the proof of Theorem 1.

Proof. Part (a). The proof proceeds by a primal-dual witness technique, and consists of the construction of a feasible primal-dual pair in the following two steps:

(i) **Primal Candidate using an oracle subproblem**: Let $\hat{\Theta}_{\backslash r}$ be the optimal solution of the restricted problem

$$\hat{\Theta}_{\backslash r} = \arg \min_{\left(\Theta_{\backslash r}\right)_{S_r^c} = \mathbf{0}} \left\{ \ell(\Theta_{\backslash r}; D) + \lambda_n \left\|\Theta_{\backslash r}\right\|_{1,2} \right\}. \tag{11}$$

(ii) **Dual Candidate from Stationary Optimality Condition**: For any column $u \in S_r$ set $\left(\hat{Z}_{\backslash r}\right)_u = \frac{1}{\left\|\left(\hat{\Theta}_{\backslash r}\right)_u\right\|_2} \left(\hat{\Theta}_{\backslash r}\right)_u$.
Set $\left(\hat{Z}_{\backslash r}\right)_{S_r^c}$ from the stationary condition (2).

Showing Strict Dual Feasibility. By construction, the $(\hat{\Theta}_{\backslash r}, \hat{Z}_{\backslash r})$ pair satisfies the stationary condition (10). It remains to show that the the dual $\hat{Z}_{\backslash r}$ is strictly feasible. We show that this holds, and also that the solution is unique, with high probability in Lemma 5.

Part (b). By uniqueness of the solution shown in part [(a)], the method excludes all edges that are not in the set of edges. To show that all correct edges are included, i.e., to show the correct correct sign recovery, it suffices to show that

$$\left\|\widehat{\Theta}_{S_r} - \widehat{\Theta}^*_{S_r}\right\|_{\infty,2} \leq \frac{\theta_{\min}}{2},$$

where, $\theta_{\min} = \min_{t \in V\backslash\{r\}} \|\theta_{rt;jk}\|_2$.

We provide an $\|\cdot\|_{\infty,2}$ bound on the error in (21), from which

$$\begin{aligned} \frac{2}{\theta_{\min}} \left\|\widehat{\Theta}_{S_r} - \widehat{\Theta}^*_{S_r}\right\|_{\infty,2} &\leq \frac{2}{\theta_{\min}} \frac{5}{C_{\min}} \lambda_n \\ &\leq 1, \end{aligned}$$

provided that $\theta_{\min} > \frac{10}{C_{\min}} \lambda_n$.

□

4 Higher-Order Discrete Graphical Models

Consider the general higher-order MRF from (2)

$$\mathbb{P}(x) \propto \exp\left\{ \sum_{C \in \mathcal{C}; v \in \{1,\ldots,m-1\}^{|C|}} \theta^*_{C;v} \mathcal{I}[x_C = v] \right\},$$

parameterized by the collection of vectors $\theta^*_C \in \mathbb{R}^{(m-1)^{|C|}}$ associated with the cliques $C \in \mathcal{C}$.

As before, we fix a node r, and define the long vector $\Theta*_{\backslash r} \in \mathbb{R}^{\sum_{j=1}^{c-1} \binom{p-1}{j} (m-1)^{j+1}}$ as the concatenation of the parameter vectors θ^*_{rC} for all $C \subseteq V\backslash r$; $|C| < c$. Let $\bar{\Theta}^*_P \in \mathbb{R}^{(m-1)^2 d_r}$ be the vector containing only neighbor-pairwise parameters $\bar{\theta}^*_{rt;jk}$ for all $t \in \mathcal{N}(r)$. Accordingly, let $\bar{\Theta}^*_{P^c}$ represent all non-zero non-pairwise entries.

Hierarchical Models. A common assumption imposed on such higher-order MRFs is that they be hierarchical models [16]. Specifically, any MRF of the form (2) is hierarchical if for any clique C, $\theta^*_C = 0$ implies that $\theta^*_B = 0$ for any clique $B \supseteq A$ containing A. This has an importance consequence: the set of pairwise effects

$$\mathcal{N}(r) \quad = \quad \left\{ u \in V\backslash\{r\} \mid \|\theta^*_{ru}\|_0 \neq 0 \right\},$$

completely characterizes the set of edges.

Thus, if we are able to estimate just the pairwise parameters of the entire higher-order model, we would still be able to recover the edge-set. Thus, we study the estimator in (6) but now when the observations are actually drawn from $\bar{\Theta}^*_{\backslash r}$. The hope is that this solution would still estimate the pairwise parameters of the underlying higher-order model well.

4.1 Assumptions

For fixed positive values $C_{\min}$, $D_{\max}$ and $\alpha \in \left(0, \frac{1}{2}\right)$, let $\gamma := \frac{D_{\max}}{C_{\min}} \left\|\bar{\Theta}^*_{P^c}\right\|_1$ and $\tau = \frac{\alpha+\gamma(\sqrt{d_r}+1)}{1+\gamma}$. We impose the following assumptions on the truth:

(C0) **Mismatch Factor**: $\gamma \leq \left(\frac{\alpha}{2-\alpha}\right)^2 \frac{C_{\min}}{100\sqrt{2}(m-1)d_r}$.

This condition is required because of the mismatch of the true underlying model and our pairwise model. In other words, we have a nonzero mean noise, caused by model mismatch, that needs to be small. Moreover, since $C_{\min} \leq (m-1)\sqrt{d_r}$ (see section 7.2), this condition ensures that $\tau \in \left(0, \frac{1}{2}\right)$ for suitable choice of α.

(C1) **Invertibility**:

$$\Lambda_{\min}\left(\mathbb{E}\left[\nabla^2 \log\left(\mathbb{P}_{\bar{\Theta}^*_{\backslash r}}\left[X_r \mid X_{\backslash r}\right]\right)\right]_{S_r S_r}\right) \geq C_{\min}(1+\gamma).$$

(C2) **Incoherence**:

Let $\bar{Q}^* := \mathbb{E}\left[\nabla^2 \log\left(\mathbb{P}_{\bar{\Theta}^*_{\backslash r}}[X_r|X_{\backslash r}]\right)\right]$. Then

$$\left\|\bar{Q}^*_{S_r^c S_r}(\bar{Q}^*_{S_r S_r})^{-1}\right\|_{\infty,2} \leq \frac{1-2\tau}{\sqrt{d_r}}.$$

(C3) **Boundedness**: $\Lambda_{\max}\left(\mathcal{J}^*\right) \leq D_{\max} < \infty$,

where $\mathcal{J}^* = \mathbb{E}\left[\mathcal{I}[X_{S_1} = x_{S_1}]\mathcal{I}[X_{S_2} = x_{S_2}]^T\right]$ for any subset of nodes S_1 and S_2, and c is the size of the maximum clique in the true graphical model. Note that $\mathcal{J}^* \in \mathbb{R}^{\sum_{i=1}^{c-1}(m-1)^j(p-1)^j \times \sum_{i=1}^{c-1}(m-1)^j(p-1)^j}$.

As in the pairwise case, in the proofs (to control the derivative of the Hessian of the log-likelihood function), we need to bound the maximum eigen value of matrix $\Im^* = \mathcal{J}^* \otimes \mathbf{1}_{\sum_{j=1}^{c-1}(m-1)^j \times \sum_{j=1}^{c-1}(m-1)^j}$. But again by properties of Kronecker products, $\Lambda_{\max}\left(\Im^*\right) = \Lambda_{\max}\left(\mathcal{J}^*\right)$, so that it suffices to impose assumptions on $\mathcal{J}^*$.

The next lemma states that imposing these assumptions on the population quantities implies analogous conditions on the sample statistics with high probability. Define $\mathcal{D} = \sum_{j=1}^{c-1} d_r^j$.

Lemma 4. *Assumptions (C0) - (C3) imply the following bounds on the pairwise parameters*

$$\textit{(D1)}\ \mathbb{P}\left[\Lambda_{\min}\left(\left[\nabla^2 \ell\left(\bar{\Theta}^*_P; D\right)\right]_{S_r S_r}\right) \leq C_{\min} - \epsilon\right] \leq 2\exp\left(-\tfrac{1}{8}\left(\epsilon\sqrt{n} - \sqrt{\mathcal{D}}\right)^2 + \log\left(\textstyle\sum_{j=2}^{c}(m-1)^j d_r^{j-1}\right)\right).$$

$$\textit{(D2)}\ \mathbb{P}\left[\left\|\nabla^2 \ell\left(\bar{\Theta}^*_P; D\right)_{S_r^c S_r} \nabla^2 \ell\left(\bar{\Theta}^*_P; D\right)^{-1}_{S_r S_r}\right\|_{\infty,2} > \tfrac{1-\alpha}{\sqrt{d_r}} + \epsilon\right] \leq 6\exp\Bigg(-\tfrac{1}{8}\left(\bar{C}_{\min}\left(\tfrac{\tau}{3\sqrt{\mathcal{D}}} + \epsilon\right)\sqrt{n} - \left(1 + \tfrac{\sqrt{\mathcal{D}}}{C_{\min}^2\sqrt{n}}\right)\sqrt{\mathcal{D}}\right)^2$$

$$. \qquad + \log\left(\textstyle\sum_{j=2}^{c}(m-1)^j(1-p)^{j-1}\right)\Bigg).$$

$$\textit{(D3)}\ \mathbb{P}\left[\Lambda_{\max}\left(\hat{\mathbb{E}}\left[\mathcal{I}[X_{t_1} = k_1]\mathcal{I}[X_{t_2} = k_2]^T\right]\right) \geq D_{\max} + \epsilon\right] \leq 2\exp\left(-\tfrac{1}{8}\left(\epsilon\sqrt{n} - \sqrt{\mathcal{D}}\right)^2 + \log\left(\textstyle\sum_{j=2}^{c}(m-1)^j d_r^{j-1}\right)\right).$$

4.2 Main Theorem

The following theorem shows that if the graphical model satisfies hierarchical assumption, then pairwise estimation exactly recovers the underlying graphical model provided that the higher order dependency parameters are not too large.

Theorem 2. *Consider an m-ary graphical model with parameter $\bar{\Theta}^*$ and associate edge set $\bar{E}$ such that conditions (C0)-(C3) and hierarchical assumption are satisfied. Suppose the size of the largest clique in the graph is c and the regularization parameter satisfies*

$$\lambda_n \geq \frac{8(2-\alpha)}{\alpha}\left(\sqrt{\frac{\log(p-1)}{n}} + \frac{m-1}{4\sqrt{n}} + \frac{1}{4}\left\|\bar{\Theta}^*_{P^c}\right\|_1\right).$$

Then, there exist positive constants K, c_1 and c_2 such that for

$$n \geq K(m-1)^2 d_r^{\frac{3}{2}c-1} \log\left((m-1)^c(p-1)^{c-1}\right),$$

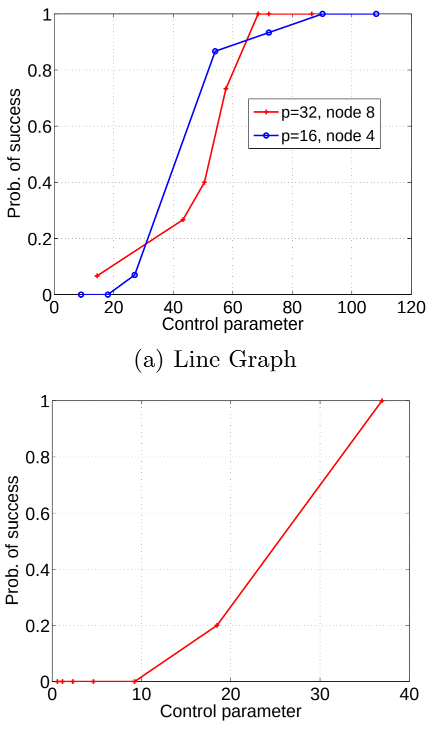
with probability $1 - c_1 \exp\left(-c_2\left(\lambda_n - 2\left\|\bar{\Theta}^*_{P^c}\right\|_1\right)^2 n\right)$ *we are guaranteed*

(a) For each node $r \in V$, the ℓ_1/ℓ_2 regularized logistic regression (6) *has a unique solution and hence specifies a neigborhood $\widehat{\mathcal{N}}(r)$.*

(b) For each node $r \in V$ correctly excludes all edges not in the true neighborhood $\mathcal{N}(r)$. Moreover, it includes all edges (r,t) such that $\left\|\bar{\theta}^*_{rt;jk}\right\|_2 \geq \frac{10}{C_{\min}}\lambda_n$.

Proof. Part [(a)]. The proof proceeds along the same lines as that of Theorem 1. We construct a primal-dual pair precisely as before using an oracle subproblem. However, showing strict-dual feasibility is more delicate when the true model has higher-order factors.

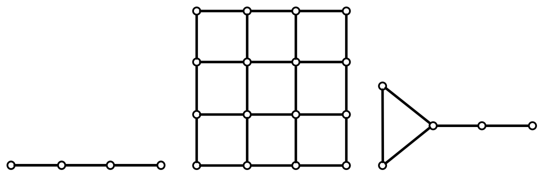
Showing Strict Dual Feasibility. By construction, the $(\hat{\Theta}_{\backslash r}, \hat{Z}_{\backslash r})$ pair satisfies the stationary condition (10), as before. We then show that the the dual $\hat{Z}_{\backslash r}$ is strictly feasible, and also that the solution is unique, with high probability in Lemma 7.



(a) Line Graph

(b) Grid

Figure 1: Probability of success $\mathbb{P}[\widehat{\mathcal{N}}(r) = \mathcal{N}(r)]$versus the control parameter $\beta(n,p,d) = \frac{n}{10d\log(p)}$ for discrete graphical models on a Line Graph and a Grid.



(a) Line Graph (b) Grid (c) Three-Way

Figure 2: Line graph (a) and Grid (b) are used in studying pairwise graphical model selection. Three-way graph (c) is used for studying higher-order graphical model selection.

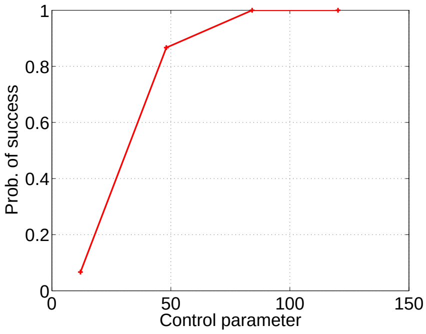


Figure 3: Probability of success $\mathbb{P}[\widehat{\mathcal{N}}(r) = \mathcal{N}(r))]$versus the control parameter $\beta(n,p,d) = \frac{n}{10d\log(p)}$ for a higher order discrete graphical model on a Three-way graph.

Part [(b)]. Here again, we can argue as in the proof of Theorem 1 to show that all correct edges are included given an $\|\cdot\|_{\infty,2}$ bound on the error provided in (24). □

5 Experiments

In this section, we report a set of synthetic experiments investigating the consequences of the main theorems. These results illustrate the behavior of the structure learning algorithm on various types of graphs. We fix the size of the alphabet $m = 3$. For a given graph type, we pick a pairwise parameter set Θ^*. We generate n samples according to the probability distribution corresponding to Θ^*. Then, we solve (6) and compare the graph corresponding to the solution with the original graph. If the two graphs are identical, we declare that the algorithm has succeeded.

Pairwise Model: We consider two different classes of graphs: line graphs and grids (Fig. 4.2). In particular, we consider line graphs of size $p = 16, 32$ and a grid of size $\sqrt{p} \times \sqrt{p} = 16$. In each of these cases, the parameter vector Θ^* is generated by setting each non-zero entry $\theta^*_{rt;jk} \in [-0.5, 0.5]$ for the line graphs and $\theta^*_{rt;jk} \in [0, 5]$ for the grid unfirmly at random. To estimate the probability of success, we use 15 batches of samples drawn from the distribution specified by Θ^*. We consider two types of simulations:

Neighborhood Recovery: Here, we focus on the recovery of the neighborhood of a particular node in a graph. Fixing a sample batch, for each pair (p, n), we set $\lambda_n = K\left(\sqrt{\frac{p-1}{n}} + \frac{m-1}{4\sqrt{n}}\right)$, where K is the constant chosen by cross validation. We compare the graph induced by $\widehat{\Theta}_{K^*}$ with the graph induced by Θ^* to get the probability of success. Fig 4.2 shows the probability of success in neighborhood recovery. Notice that for different values of n and p, the phase transition graphs stack on the top of each other; this shows that the scaling of the samples n is correct.

Higher-Order Model: In this case, we consider a graph with higher order dependencies and try to estimate it using the pairwise model. We consider the three-way graph (triangle + line graph of size $p-2$) shown in Fig 2(c). There is only one three-way factor involving three nodes. The rest of the graph is characterized by pairwise parameters. Solving (7), we investigate the probability of success for neighborhood recovery of the node that connects the line graph and the triangle. Fig. 4.2 illustrates the result.

References

[1] P. Abbeel, D. Koller, and A. Y. Ng. Learning factor graphs in polynomial time and sample complexity. *Jour. Mach. Learning Res.*, 7:1743–1788, 2006.

[2] F. Bach. Consistency of the group lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008.

[3] G. Bresler, E. Mossel, and A. Sly. Reconstruction of Markov random fields from samples: Some easy observations and algorithms. `http://front.math.ucdavis.edu/0712.1402`, 2008.

[4] E. Candes and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 2006.

[5] C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Trans. Info. Theory*, 14(3):462–467, 1968.

[6] G. Cross and A. Jain. Markov random field texture models. *IEEE Trans. PAMI*, 5:25–39, 1983.

[7] I. Csiszár and Z. Talata. Consistent estimation of the basic neighborhood structure of Markov random fields. *The Annals of Statistics*, 34(1):123–145, 2006.

[8] C.Zhang and J.Huang. Model selection consistency of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36:1567–1594, 2008.

[9] C. Dahinden, G. Parmigiani, M.C. Emerick, and P. Buhlmann. Penalized likelihood for sparse contingency tables with an application to full-length cdna libraries.

[10] C. Dahinden, M. Kalisch, and P. Buhlmann. Decomposition and model selection for large contingency tables. *Biometrical Journal*, 52(2):233–252, 2010.

[11] S. Dasgupta. Learning polytrees. In *Uncertainty on Artificial Intelligence*, pages 134–14, 1999.

[12] D. Donoho and M. Elad. Maximal sparsity representation via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100:2197–2202, March 2003.

[13] M. Hassner and J. Sklansky. The use of Markov random fields as models of texture. *Comp. Graphics Image Proc.*, 12:357–370, 1980.

[14] E. Ising. Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925.

[15] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(6):957–968, 2005.

[16] S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

[17] S.-I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of markov networks using l1-regularization. In *Neural Information Processing Systems (NIPS) 19*, 2007.

[18] K. Lounici, A. B. Tsybakov, M. Pontil, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In *22nd Conference On Learning Theory (COLT)*, 2009.

[19] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[20] L. Meier, S. van de Geer, and P. Buhlmann. The group lasso for logistic regression. 70:53–71, 2008.

[21] N. Meinshausen and P. Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3), 2006.

[22] S. Negahban and M. J. Wainwright. Joint support recovery under high-dimensional scaling: Benefits and perils of $\ell_{1,\infty}$-regularization. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[23] A. Y. Ng. Feature selection, ℓ_1 vs. ℓ_2 regularization, and rotational invariance. In *International Conference on Machine Learning*, 2004.

[24] G. Obozinski, M. J. Wainwright, and M. I. Jordan. Support union recovery in high-dimensional multivariate regression. *Annals of Statistics*, 2010.

[25] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. Sparse additive models. *Journal of the Royal Statistical Society, Series B*, .

[26] P. Ravikumar, M. J. Wainwright, and J. Lafferty. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, .

[27] B. D. Ripley. *Spatial statistics*. Wiley, New York, 1981.

[28] A. Rothman, P. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal on Statistics*, 2:494–515, 2008.

[29] P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction and search. *MIT Press*, 2000.

[30] N. Srebro. Maximum likelihood bounded tree-width Markov networks. *Artificial Intelligence*, 143(1):123–138, 2003.

[31] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Info. Theory*, 51(3):1030–1051, March 2006.

[32] B. Turlach, W.N. Venables, and S.J. Wright. Simultaneous variable selection. *Techno- metrics*, 27:349–363, 2005.

[33] M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55: 2183–2202, 2009.

[34] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Info. Theory*, To appear. Original version: UC Berkeley Technical Report 709, May 2006.

[35] J.W. Woods. Markov image modeling. *IEEE Transactions on Automatic Control*, 23:846–850, October 1978.

[36] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society B*, 1(68):49, 2006.

[37] P. Zhao and B. Yu. On model selection consistency of lasso. *J. of Mach. Learn. Res.*, 7:2541–2567, 2007.