The Log-linear Group Lasso Estimator and Its Asymptotic Properties

Yuval Nardi*  
Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890 USA

Alessandro Rinaldo†  
Department of Statistics  
Carnegie Mellon University  
Pittsburgh, PA 15213-3890 USA

Abstract

We define the group lasso estimator for the natural parameters of the exponential families of distributions representing hierarchical log-linear models under multinomial sampling scheme. Such estimator arises as the unique solution of a convex penalized likelihood program using the group lasso penalty. We illustrate how it is possible to construct, in a straightforward way, an estimator of the underlying log-linear model based on the blocks of non-negative coefficients recovered by the group lasso procedure. We investigate the asymptotic properties of the group lasso estimator and of the associated model selection criterion in a double-asymptotic framework, in which both the sample size and the model complexity grow simultaneously. We provide conditions guaranteeing that the group lasso estimator is norm consistent and that the group lasso model selection is a consistent procedure, in the sense that, with overwhelming probability as the sample size increases, it will correctly identify all the sets of non-zero interactions among the variables. Provided the sequences of true underlying models is sparse enough, recovery is possible even if the number of cells grows larger than the sample size. Finally, we derive some central limit type of results for the log-linear group lasso estimator.

1 Introduction

Log-linear model analysis of categorical data is a widespread and important set of statistical methodologies that have found applications in very diverse scientific areas, ranging from social and biological sciences, to medicine, disclosure limitation problems, data-mining, image analysis, fingerprinting, language processing and genetics. Inherently, log-linear modeling is a model selection methodology for contingency tables that encompasses testing a number of statistical models for the joint distribution of a group of categorical variables. The classical asymptotic theory of model selection and goodness-of-fit testing is well developed and understood for the “small \( p \) and large \( N \)” case. It is applicable to a variety of goodness-of-fit measures, such as Pearson’s \( \chi^2 \), the likelihood ratio statistic and, more generally, the power-divergence family of Cressie and Read (1988). The applicability and validity of these methods demand the availability of large sample sizes and the existence of the maximum likelihood estimate.

*Email: yuval@stat.cmu.edu  
†Email: arinaldo@stat.cmu.edu
In recent years, the importance and usage of log-linear modeling methodologies have increased dramatically with the compilation and diffusion of large databases in the form of sparse contingency tables. In such instances, the number of sampled units is not much different, in fact often smaller, than the number of cells, so that most of the cell entries are very small or zero counts. In high-dimensional settings, the traditional methodologies indicated above are simply inadequate. First off, the number of log-linear models grow extremely fast with the number of variables (for example, there are 7,580 hierarchical models for a 5-way table!), and selecting an optimal model involves exploring a space of models of virtually infinite dimension. Secondly, for a given model of even moderate complexity, the MLE is unlikely to exist: the small information content in the data limits the possibility for inference only to a portion of the parameter space (see Rinaldo, 2006a, for details). As a result, traditional goodness-of-fit testing and model selection will produce very poor, if not completely erroneous, asymptotic approximation. It is quite clear that a more appropriate statistical formalization requires the consideration of a “large $p$” setting.

In this article, we propose a methodology for log-linear model selection that is particularly suited to high-dimensional tables, and we describe some of its asymptotic properties. Our results are akin to the asymptotic optimality of the lasso estimator in high dimensional least squares problems, where the recovery of the sparsity pattern of an unknown set of parameters in noisy settings via $\ell_1$ regularization is possible, even if the number of parameters grows faster than the sample size. See, in particular, Meinshausen and Bühlmann (2006), Zhao and Yu (2006), Wainwright (2006) and, for a different approach, see Greenshtein (2006) and Greenshtein and Ritov (2006). Existing work on penalized likelihood problems involving $\ell_1$ regularization for discrete problems include the non-asymptotic analyses of estimation in high-dimensional generalized linear models via lasso by van de Geer (2006a,b), and the sufficient conditions of Wainwright et al. (2006) for consistency of $\ell_1$ regularized logistic regression with binary variables under a double asymptotic framework. In section 5, we discuss in detail the differences between our problem and solutions and the existing results.

We formulate the log-linear model selection problem as a convex penalized likelihood problem based on the group lasso, a convex penalty function introduced by Yuan and Lin (2006) in a non-asymptotic ANOVA setting. The group lasso regularization is an extension of the lasso $\ell_1$ penalty designed to penalized groups of coefficients simultaneously. It has been shown to be effective in logistic regression problems by Meier et al. (2006) and has been used in applications involving log-linear modeling of sparse contingency tables in Dahinden et al. (2006).

The paper is organized as follows. In section 2 we describe the log-linear model settings we will be considering. The direct sum decomposition of the natural parameter space by log-linear subspaces defines a partition of the parameters in blocks of different dimensions, which are utilized as argument of the group penalty function. In Section 3 we describe the group lasso estimator for log-linear models, which can be computed by solving a convex program and can be interpreted as a smoothed MLE (see Section 3.1). Taking advantage of the combinatorial properties of log-linear models, we show that the group lasso estimator produces, in turn, an estimator of the underlying log-linear model, which is constructed simply by isolating the non-zero blocks of the group-lasso estimates. In Section 4, we study the consistency properties of the group lasso estimator and of the associated model selection procedure. We formulate a rather general double-asymptotic framework in which we allow both the sample size and the model complexity to grow. We break down our analysis into different steps, each step establishing progressively stronger results and, accordingly, requiring stronger assumptions, than the previous one. In Section 4.2, we derive conditions...
guaranteeing that the group lasso estimator is norm consistent. Our assumptions rely on an extension of local approximations by quadratic mean differentiability of regular models to the double-asymptotic settings we consider. In Section 4.3, we strengthen our assumptions to guarantee that the model estimates are consistent, i.e. that, asymptotically, the group lasso procedures correctly identifies the set of interactions making up the underlying model. We conclude our analysis with some central limit results in Section 4.4, which prove, in particular, that the group lasso estimator is asymptotically biased and inefficient. The proofs appear in Section 6 and in the Appendices.

2 Log-linear Models

We adopt the usual log-linear modeling setting, which we formalize below. We consider $K$ categorical random variables $X_1, \ldots, X_K$, each $X_k$ taking values on a finite set, which, without loss of generality, can always be assumed to be $I_k \equiv \{1, \ldots, I_k\}$. Letting $\mathcal{I} = \bigotimes_{k=1}^K I_k$, $\mathbb{R}^\mathcal{I}$ is the vector space of $K$-dimensional arrays of format $I_1 \times \ldots \times I_K$, i.e. the vector space of real-valued functions defined on $\mathcal{I}$. Each element of $\mathcal{I}$, a cell, is a multi-index $(i_1, \ldots, i_K)$, whose $k$-th coordinate indicates the value taken on by the $k$-th variable. For convenience, we identify $\mathbb{R}^\mathcal{I}$ with the Euclidean space $\mathbb{R}^{I}$, where $I = \prod_k I_k$, so that standard inner product $\langle x, y \rangle = \sum_{i \in \mathcal{I}} x_i y_i$ on $\mathbb{R}^\mathcal{I}$ is well defined. (This identification can be easily made by ordering $\mathcal{I}$ as a linear list using any bi-jection between $\mathcal{I}$ and the set $\{1, 2, \ldots, I\}$.) Therefore, each cell can be represented by a single index $i$ between 1 and $I$, rather than by a multi-index.

The cross-classification of $N$ independent and identically distributed realizations of $(X_1, \ldots, X_K)$ produces a random integer-valued vector $\mathbf{n} \in \mathbb{R}^I$, a contingency table, whose coordinate entry $n_{i_1, \ldots, i_K}$ corresponds to the number of times the cell combination $(i_1, \ldots, i_K)$ was observed in the sample. The table $\mathbf{n}$ has a Multinomial$(N, \pi)$, distribution, where $\pi$ is a strictly positive probability vector with coordinates

$$
\pi_{i_1, \ldots, i_K} = \mathbb{P}((X_1, \ldots, X_K) = (i_1, \ldots, i_K)).
$$

In log-linear modeling, the joint distribution of $(X_1, \ldots, X_K)$ is fully specified by representing the cell mean vector $\mathbf{m} = \mathbb{E}\mathbf{n} = N \pi$ by means of certain linear subspaces $\mathcal{M}$ of $\mathbb{R}^I$ containing $\log \mathbf{m}$, to the extent that log-linear models themselves are defined by such subspaces. Namely, by fixing $\mathcal{M}$, it follows that the logarithms of the cell mean vectors must satisfy specific linear constraints, to be specified below, which completely characterize the underlying distribution. The log-likelihood function at a point $\mu \in \mathcal{M}$ is

$$
\ell(\mu) = \sum_{i \in \mathcal{I}} n_i \log \frac{m_i}{\langle \mathbf{m}, \mathbf{1} \rangle} + \log N! - \sum_{i \in \mathcal{I}} \log n_i!,
$$

where $\mathbf{m} = \exp \mu$ and $\mathbf{1}$ is $I$-dimensional vector containing ones. Because of the Multinomial sampling assumption, $\ell$ is only defined over the subset $\bar{\mathcal{M}} \subseteq \mathcal{M}$ given by

$$
\bar{\mathcal{M}} = \{\mu \in \mathcal{M} : \langle \mathbf{m}, \mathbf{1} \rangle = N\},
$$

which is neither a vector space nor a convex set. Instead, it is convenient to explicitly discard the subspace of $\mathcal{M}$ that is fixed by design and to work with the smaller linear subspace $\mathcal{M} \cap \mathcal{R}(\mathbf{1})^\perp$, where $\mathcal{R}(\mathbf{1})$ is the one-dimensional subspace spanned by $\mathbf{1}$. For each $\beta \in \mathcal{M} \cap \mathcal{R}(\mathbf{1})^\perp$, set

$$
\ell^\ast(\beta) = \langle \mathbf{n}, \beta \rangle - N \log \langle \exp \beta, \mathbf{1} \rangle + \log N! - \sum_{i \in \mathcal{I}} \log n_i!.
$$
Then, there exists a bijection between \( \tilde{\mathcal{M}} \) and \( \mathcal{M} \cap \mathcal{R}(1) \) in the sense that, for each \( \tilde{\mu} \in \tilde{\mathcal{M}} \) there exists one \( \beta \in \mathcal{M} \cap \mathcal{R}(1) \) such that

\[
\tilde{\mu} = \frac{N}{(\exp^{\beta}, 1)} \exp^{\beta} \quad \text{and} \quad \ell^*(\beta) = \ell(\tilde{\mu}), \text{ for each } n,
\]

and, conversely, for each \( \beta \in \mathcal{M} \cap \mathcal{R}(1) \) there exists one \( \tilde{\mu} \in \tilde{\mathcal{M}} \) satisfying the above identities (for a proof of this result in more generality, see Lemma 2.2 in Rinaldo, 2006b).

Therefore, if \( U \) is any full-rank matrix whose columns span \( \mathcal{M} \cap \mathcal{R}(1) \), then \( \beta = U \theta \) for some \( \theta \in \mathbb{R}^k \), with \( k = \dim(\mathcal{M}) - 1 \), so that the log-likelihood function can re-written as

\[
\ell(\theta) = (U^T n, \theta) - N \log(\exp^{U \theta}, 1) + \log N! - \sum_{i \in \mathcal{I}} \log n_i!, \quad \theta \in \mathbb{R}^k. \tag{1}
\]

This re-parametrization is essentially equivalent to reduction to minimal form of the underlying exponential family of distributions for the cell counts via sufficiency. In fact, the previous display shows that each log-linear model correspond to a full, regular exponential family of dimension \( \dim(\mathcal{M}) - 1 \) and natural sufficient statistic \( U^T n \).

The gradient and Hessian matrix for \( \ell(\theta) \) are easily derivable. Letting \( b = \exp^{U \theta} \), one can see that

\[
\nabla \ell(\theta) = U^T (n - \left( \frac{N}{(b, 1)} \right) b) = U^T (n - m) \tag{2}
\]

and

\[
\nabla^2 \ell(\theta) = - \left( \frac{N}{(b, 1)} \right) U^T D_b - \frac{1}{(b, 1)} b b^T U = -U^T \left( D_m - \frac{m m^T}{N} \right) U, \tag{3}
\]

where \( m = \frac{N}{(b, 1)} b = E_{\theta} n \) and \( D_m \) denote the diagonal matrix with diagonal \( m \). It is worth pointing out that, because these models are exponential families, the negative Hessian is the covariance matrix of the natural sufficient statistics \( U^T n \) and also the Fisher information multiplied by the sample size.

### 2.1 Log-Linear Subspaces

In this section we construct the log-linear subspaces we will be considering. Although log-linear models are defined by generic linear manifolds of \( \mathbb{R}^I \), in practice it is customary to consider only very specific classes of linear subspaces, which are also characteristic of ANOVA models and experimental design. These subspaces present considerable advantages in terms of interpretability and ease of computation and can be constructed easily by exploiting various correspondences between combinatorial structure of the power set of \( \mathcal{K} = \{1, \ldots, K\} \) and a certain direct sum decomposition of \( \mathbb{R}^I \), to be described below.

A rather intuitive way of specifying a certain dependence structure among the \( K \) variables of interest is to provide a list of the interactions among them. Then, the associated statistical model is representable as a class of subsets of \( \mathcal{K} = \{1,2,\ldots,K\} \), each one indicating a different type of interaction. In fact, every subset \( h \) of \( \mathcal{K} \) can be given a straightforward ANOVA-type of interpretation, based on its cardinality \( |h| \), so that \( h \) identifies an interaction of order \( |h| - 1 \) among the variables \( \{i: i \in h\} \). For example, if \( |h| = 1 \), then \( h \) is a main effect, if \( h = \emptyset \), then \( h \) is the grand mean, and so on.
Formally, let $2^K$ be the power set of $K$, which we view as a boolean lattice with respect to the partial order induced by the operation of taking subset inclusion. An abstract simplicial complex $\Delta$ on $K$ is a class of subsets of $K$ such that $h \subseteq d$ for some $d \in \Delta$ implies $h \in \Delta$. A simplicial complex is uniquely determined by its elements that are maximal with respect to inclusion, known as its facets, which represents the highest order interactions. Therefore, $\Delta$ can be identified with the set of its facets, a convention we will use throughout the article. By construction, once an interaction term is part of the model, all lower order interactions are included, i.e. the model is hierarchical. Then, hierarchical log-linear model can be defined in a purely combinatoric form as a subset of $2^K$.

**Definition 2.1.** A hierarchical log-linear model is a simplicial complex $\Delta$ on $K$.

This definition provides a formal justification of the traditional notation (see, e.g., Bishop et al., 1975, Chapter 3) of identifying hierarchical log-linear models with classes of maximal subsets of $K$, sometimes denoted as generating classes, indicating the maximal order interactions. Notice also that Definition 2.1 includes as special case the class of graphical models, defined as follows. For every complex $\Delta$, one can construct its interaction graph, the graph with vertex set $K$ and edge set consisting of all unordered pairs $(i, j) \subseteq K$ such that $(i, j) \subseteq d$ for some $d \in \Delta$. A simplicial complex $\Delta$ is called graphical if its facets are the cliques of its interaction graph. From the probabilistic point of view, lack of an edge between two nodes or sets of nodes is a formal representation of various Markov properties of conditional independence among the corresponding variables (see, e.g., Lauritzen, 1996).

**Example 2.2 (Hierarchical log-linear models).** $\Delta = \{\{1\}, \{2\}, \{3\}\}$ is the model of mutual independence of the three factors and $\Delta = \{\{1, 2\}, \{2, 3\}\}$ denotes the model of conditional independence of factor “1” and “3” given factor “2”, a decomposable model. The simplest example of a graphical non-decomposable (and non-reducible) model is the 4-cycle model on 4 factors, $\Delta = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{1, 4\}\}$. The simplest non-graphical model is the model of no-3-factor effect $\Delta = \{\{1, 2\}, \{2, 3\}, \{1, 3\}\}$. In general, for a $K$-way table, the largest hierarchical log-linear model is the model of no-$K$-factor effect, represented by the simplicial complex on $K$ nodes whose $K - 1$ facets form the set of all possible distinct subsets of $K$ with cardinality $K - 1$. ■

There is a remarkable correspondence between the combinatorics of simplicial complexes and of certain orthogonal subspaces of $\mathbb{R}^I$. In fact, for a given simplicial complex $\Delta$, a log-linear subspace $\mathcal{M}_\Delta$ can be constructed in a natural way as the direct sums of orthogonal subspaces indexed by subsets of $K$. Specifically,

$$\mathcal{M}_\Delta = \bigoplus_{\{h \subseteq d: d \in \Delta\}} \mathcal{U}_h,$$  \hspace{1cm} (4)

where $\mathcal{U}_h \perp \mathcal{U}_{h'}$ for $h, h' \subseteq K$ with $h \neq h'$. Below, we summarize the main features of this construction that are relevant to our problem. Notice that the resulting log-linear subspaces are precisely the factor-interaction subspaces and the subspaces of interactions, as described by Darroch and Speed (1983).

For any subset $h \subseteq K$ and cell index $i = (i_1, \ldots, i_K) \in I$, let $i_h$ denote its coordinate projection $\{i_k: k \in h\}$ onto $\bigotimes_{k \in h} \mathcal{I}_k$. Define the equivalence relation $\sim_h$ on $I$ given by

$$i \sim_h j \iff i_h = j_h,$$
for all $i, j \in \mathcal{I}$ and associate to each $h \subseteq \mathcal{K}$ the subspace $\mathcal{W}_h \subset \mathbb{R}^\mathcal{I}$ consisting of all functions on $\mathcal{I}$ that depend on $i \in \mathcal{I}$ only through $i^h$, i.e.

$$\mathcal{W}_h = \left\{ f \in \mathbb{R}^\mathcal{I} : f(i) = f(j) \text{ if } i^h \sim j \right\}, \quad (5)$$

Letting $\mathcal{W}^\mathcal{K} = \{ \mathcal{W}_h : h \in 2^\mathcal{K} \}$, where the subspaces are defined as in (5), the posets $\mathcal{W}^\mathcal{K}$ and $2^\mathcal{K}$ are isomorphic lattices because

$$\mathcal{W}_h \subseteq \mathcal{W}_h \iff h' \subseteq h$$

for all $h, h' \in 2^\mathcal{K}$, with the $\emptyset$ and $\hat{1}$ elements of $2^\mathcal{K}$ and $\mathcal{W}^\mathcal{K}$ being $\emptyset$ and $\mathcal{K}$, and $\mathcal{R}(1)$ and $\mathbb{R}^\mathcal{I}$, respectively. The full extent of this correspondence is explained in the next theorem. For a proof see Lauritzen (1996, Appendix B.2) and, for further details and alternative derivations, Rinaldo (2006b, Section 2).

**Theorem 2.3.** For any $h, h_1, h_2 \in 2^\mathcal{K}$,

$$h = h_1 \cup h_2 \iff \mathcal{W}_h = \mathcal{W}_{h_1} + \mathcal{W}_{h_2}$$

$$h = h_1 \cap h_2 \iff \mathcal{W}_h = \mathcal{W}_{h_1} \cap \mathcal{W}_{h_2}$$

Also, letting

$$\mathcal{U}_h = \mathcal{W}_h \cap \left( \sum_{\{ h' \in 2^\mathcal{K} : h' \subseteq h \}} \mathcal{W}_{h'} \right) \downarrow,$$

then

1. for any $h, h' \in 2^\mathcal{K}$, with $h \neq h'$ the subspaces $\mathcal{U}_h$ and $\mathcal{U}_{h'}$ are orthogonal to each other;

2. for each $h \in 2^\mathcal{K}$

$$\mathcal{W}_h = \bigoplus_{\{ h' \in 2^\mathcal{K} : h' \subseteq h \}} \mathcal{U}_{h'}.$$

In particular, for $h = \mathcal{K}$,

$$\mathbb{R}^\mathcal{I} = \bigoplus_{h' \subseteq \mathcal{K}} \mathcal{U}_{h'} \quad (6)$$

**Remark**

Throughout the document, we will be assuming that the elements of $2^\mathcal{K}$ are ordered in some predefined way, and that any indexing by subsets of $\mathcal{K}$ is done accordingly.

**Theorem 2.3** is of great practical value, as it provides the linear algebra tools needed to construct the log-linear subspaces (4). Then, any hierarchical log-linear can be equivalently specified either combinatorially, using Definition 2.1, or by the vector subspace defined in (4). Furthermore, the dimension of the log-linear subspaces $\mathcal{M}_\Delta$ and of all of its subspaces of interactions can also be computed directly, according to the next statement.

**Proposition 2.4.** Let $\mathcal{H}$ be a class of subsets of $\mathcal{K}$ and $\mathcal{M}_\mathcal{H} = \bigoplus_{h \in \mathcal{H}} \mathcal{U}_h$. Then

$$\dim(\mathcal{M}_\mathcal{H}) = \sum_{h \in \mathcal{H}} \prod_{k \in h} (I_k - 1). \quad (7)$$

6
In particular, for any log-linear model $\Delta$,
\[
\dim(\mathcal{M}_\Delta) = \left( \prod_{k=1}^{K} I_j \right) - \sum_{\{ h \in 2^K : h \subseteq d, d \in \Delta \}} \prod_{k \in h} (I_k - 1),
\]
with the convention that, for $h = \emptyset$, $\prod_{k \in h} (I_k - 1) = 1$.

The appendix contains the proof of Proposition 2.4, along with an algorithmically simple way of generating the design matrix $U_h$ spanning the subspace $\mathcal{U}_h$, for each $h$. More generally, one may assume the columns of each $U_h$ to be an orthonormal system, although the matrices constructed in the appendix would work just as well in our results.

For matrices $U_1, \ldots, U_n$ with the same number of rows $r$ and number of columns $c_1, \ldots, c_n$, respectively, we will denote the operation of adjoining them into one matrix of dimension $r \times \sum_{k=1}^{n} c_k$ with
\[
\bigoplus_{k=1}^{n} U_k = [U_1 \ldots U_n].
\]
Then, using this notation and with $U_h$ a full-rank matrix spanning $\mathcal{U}_h$, the columns of
\[
U_\Delta = \bigoplus_{h \in 2^K, h \neq \emptyset, h \subseteq d, d \in \Delta} U_h
\]
span $\mathcal{M}_\Delta$, and, therefore, $U$ is a full-rank design matrix for the log-linear model $\Delta$.

3 The Group Lasso Estimator for Log-Linear Models

Following the results in the previous section, the columns of the matrix
\[
U = \bigoplus_{h \in 2^K, h \neq \emptyset} U_h
\]
span $\mathbb{R}^{I-1}$, where
\[
\text{rank}(U_h) = \dim(\mathcal{U}_h) \equiv d_h,
\]
is given in (7). Accordingly, for any vector $\theta \in \mathbb{R}^{I-1}$, we can write
\[
\theta = \text{vec}\left\{ \theta_h, h \in 2^K, h \neq \emptyset \right\},
\]
where $\theta_h$ denotes the $d_h$-dimensional vector of $\theta$ corresponding to the sub-matrix $U_h$. Then, using (1), the log-likelihood function for the saturated $(I-1)$-dimensional log-linear model becomes
\[
\ell(\theta) = \sum_{h, h \neq \emptyset} \langle U_h^\top n, \theta_h \rangle - N \log(\exp\left\{ \sum_{h, h \neq \emptyset} U_h \theta_h \right\} \cdot 1) + \log N! - \sum_i \log n_i!, \quad \theta \in \mathbb{R}^{I-1}. \tag{9}
\]

Notice that the one-dimensional sub-space $\mathcal{R}(1)$ corresponding to the empty set is not included, because of the multinomial sampling restriction.
For any non-trivial (i.e. different from the uniform distribution) model \( \Delta \), with corresponding log-linear subspace \( \mathcal{M}_\Delta \), let
\[
\mathcal{H} = \left\{ h : h \neq \emptyset, h \subseteq d, \text{ some } d \in \Delta \right\},
\]
be the collections of sets representing all the interactions in \( \Delta \), or equivalently, the collections of factor interaction subspaces of \( \mathcal{M}_\Delta \), so that
\[
\dim(\mathcal{M}_\Delta) - 1 = \sum_{h \in \mathcal{H}} d_h \equiv d_\mathcal{H}.
\]
If \( \mathcal{H} \) is not empty (i.e. \( \Delta \) is different than \( \{\emptyset\} \)), it is is clear that the natural parameter space for \( \Delta \), i.e. \( \mathbb{R}^{d_\mathcal{H}} \), can be embedded as a linear subspace of \( \mathbb{R}^{I-1} \) consisting of all vectors such that
\[
\left\{ \begin{array}{ll}
\|\theta_h\| > 0, & h \in \mathcal{H} \\
\|\theta_h\| = 0, & h \notin \mathcal{H},
\end{array} \right.
\]
with \( \| \cdot \| \) being any norm on \( \mathbb{R}^{I-1} \). The log-likelihood function for this model is still given by Equation (9), where the summations are now taken over the sets \( h \) in the class \( \mathcal{H} \).

Then, the model selection problem of recovering the underlying model \( \Delta \) from an observed table \( n \) can be cast as an estimation problem for the block components of \( \theta \in \mathbb{R}^{I-1} \) that have positive norms, based on the likelihood (9) of the saturated model. To this end, one is naturally led to consider penalized maximum likelihood estimation procedures of the form
\[
\max_{\theta \in \mathbb{R}^{I-1}} \left\{ \ell(\theta) - \text{pen}(\theta) \right\},
\]
where \( \ell(\theta) \) is defined by (9) and \( \text{pen}(\theta) \) assigns a penalty to every block \( \theta_h \) that is non-zero. Ideally, the function \( \text{pen} \) should satisfy two requirements. First, it should act as a thresholding function by either keeping or killing (i.e. setting to zero) each block \( \theta_h, h \subset K, h \neq \emptyset \). Secondly, it should be reasonably well behaved (e.g. be convex) so that the problem (11) is computationally feasible.

Yuan and Lin (2006) propose the group lasso procedure for Gaussian models, based on a class of convex penalty functions which are specifically design to produce sparsity in the vector of estimated coefficients at the block level. These penalty functions are obtained as compositions of the \( \ell_1 \) norm over quadratic norms of the individual blocks. The group lasso penalty results from applying first the quadratic penalty to individual blocks, to promote non-sparsity, and then from applying the \( \ell_1 \) norm to the resulting block norms, to promote block sparsity. The group lasso methodology of Yuan and Lin (2006), originally developed for linear Gaussian models under ANOVA settings, was further extended to logistic regression models by Meier et al. (2006) and to log-linear models by Dahinden et al. (2006), which inspired our work. Specifically, the group lasso estimator for log-linear model we consider arises as the solution of the concave optimization problem
\[
\arg\max_{\theta \in \mathbb{R}^{I-1}} P_\Lambda(\theta),
\]
where
\[
P_\Lambda(\theta) = \frac{1}{N} \ell(\theta) - \lambda \sum_{h,h \neq \emptyset} \lambda_h \|\theta_h\|_2,
\]
with \( \ell(\cdot) \) defined as in (9) and \( \Lambda = \{\lambda, \{\lambda_h, h \neq \emptyset\}\} \) is a set of given tuning parameters. The parameter \( \lambda \) controls the overall effect of the penalty and should be a function of the sample
1. Obtain the log-linear group lasso estimator,
\[ \hat{\theta} = \operatorname{argmax}_{\theta \in \mathbb{R}^{I-1}} P_N(\theta). \]

2. Extract the set of non-zero blocks from \( \hat{\theta} \),
\[ \hat{\mathcal{H}} = \{ h : \| \hat{\theta}_h \|_2 > 0 \}. \]

3. Recover the hierarchical log-linear model from \( \hat{\mathcal{H}} \),
\[ \hat{\Delta} = \{ d : d \text{ is a maximal element of } \hat{\mathcal{H}} \} \]

Table 1: The group lasso model selection for hierarchical log-linear models.

size, while the block parameters \( \lambda_h \) allows for specific penalties depending on the sizes of the individual blocks. A reasonable choice for these tuning parameters is \( \lambda_h = \sqrt{d_h} \), so that each block of coefficients is penalized proportionally to its dimension, with larger blocks penalized more heavily.

**Lemma 3.1.** The program (12) admits a unique optimizer \( \hat{\theta} \in \mathbb{R}^{I-1} \) whose \( h \)-block component satisfies
\[
\begin{align*}
-\frac{1}{N} U_h^\top (n - \hat{m}) + \lambda \lambda_h \frac{\hat{\theta}_h}{\| \hat{\theta}_h \|_2} &= 0 & \text{if } \hat{\theta}_h \neq 0 \\
\frac{1}{N} \| U_h^\top (n - \hat{m}) \|_2 &\leq \lambda \lambda_h & \text{if } \hat{\theta}_h = 0,
\end{align*}
\]

where \( \hat{m} = N \langle \exp^{U_h \hat{\theta}}, 1 \rangle \exp^{U \hat{\theta}} = \mathbb{E}_\theta n. \)

Having obtained the group lasso estimator \( \hat{\theta} \), the model selection step entails building an estimate of the true model \( \Delta \) by extracting the blocks of \( \hat{\theta} \) with positive norm and then build a simplicial complex \( \hat{\Delta} \) as illustrated in Table 3. One may say that this procedure is effective at recovering the underlying set of interactions if, with high probability, \( \hat{\Delta} \) is sufficiently close to \( \Delta \). We call this property model selection consistency. Notice that, since we are only concerned with finding good estimators of \( \Delta \), it is not required of \( \hat{\theta} \) to satisfy any optimality criteria as an estimator of \( \theta^0 \), besides the ones leading to model selection consistency. In fact, we will see in Section 4.4 that \( \hat{\theta} \) is far from being optimal.

The main advantage of using the group lasso estimator for estimating \( \Delta \) rather than traditional methods of model selection based on sequential testing of a potentially very large number of competing models, is the computational ease. In fact the methodology described in Table 3 only involves determining a penalized maximum likelihood estimator of \( \theta^0 \) and thus require solving only one convex optimization problem.

We conclude this section remarking that one could make different choices for the group penalty function. In particular, following Tropp (2005), one could consider a penalty term which is built using the \( \ell_1 \) norm over the \( \ell_\infty \) norms of individual blocks. This particular choice would assign a milder penalty for complexity than our choice, for the same set of tuning parameters \( \Lambda \).
3.1 The Group Lasso Estimator as a Smoothed MLE

The proof of Lemma 3.1 shows that the group lasso penalty implies that the selected model \( \hat{\Delta} \) is one for which the maximum likelihood estimates exist. In fact, comparing the the group lasso estimator with the MLE for the model \( \hat{\Delta} \) one gets a better insight on how the group penalty works. We assume here some familiarity with theory of exponential families. See Brown (1986), or, for results specific to log-linear models, Rinaldo (2006a).

For convenience, and without loss of generality, we replace \( P_\Lambda \) with \( NP_\Lambda \) in (12). The condition of optimality for \( \hat{\theta} \) implies that, for each \( h \in \hat{\mathcal{H}} \),
\[
\hat{\theta}_h^T U_h^T (n - \hat{m}) = N \lambda \lambda_h \| \hat{\theta}_h \|^2.
\]
Let \( \tilde{\Delta} \) be the simplicial complex derived by \( \hat{\mathcal{H}} \), as described in Table 3 and \( U_{\tilde{\Delta}} \) be defined as in (8). Consider the vector
\[
\hat{\theta}_{GL} = \bigoplus_{\{h : h \neq \emptyset, \emptyset \subseteq d, d \in \tilde{\Delta}\}} \hat{\theta}_h.
\]
At the optimum, the objective function becomes, disregarding an irrelevant additive constant,
\[
\langle n, U \hat{\theta} \rangle - N \log \langle \exp \hat{U} \theta, 1 \rangle - \sum_{h \in \hat{\mathcal{H}}} N \lambda \lambda_h \| \hat{\theta}_h \|^2 = \langle n, U \hat{\theta} \rangle - N \log \langle \exp \hat{U} \theta, 1 \rangle - \sum_{h \in \hat{\mathcal{H}}} \hat{\theta}_h^T U_h^T (n - \hat{m})
\]
\[
= \langle \hat{m}, U \hat{\theta} \rangle - N \log \langle \exp U \hat{\theta}, 1 \rangle
\]
\[
= \langle U_{\tilde{\Delta}}^T \hat{m}, \hat{\theta}_{GL} \rangle - N \log \langle \exp \{ U_{\tilde{\Delta}} \hat{\theta}_{GL} \}, 1 \rangle.
\] (14)

On the other hand, the optimal value of the log-likelihood function under the model \( \tilde{\Delta} \) is achieved at the MLE \( \hat{\theta}_{MLE} \), and is equal to
\[
\langle U_{\tilde{\Delta}}^T n, \hat{\theta}_{MLE} \rangle - N \log \langle \exp \{ U_{\tilde{\Delta}} \hat{\theta}_{MLE} \}, 1 \rangle.
\] (15)

Equation (15) elucidates a fundamental fact from the theory of extended exponential families, namely that the MLE, \( \hat{\theta}_{MLE} \), and the minimal sufficient statistics, \( U_{\tilde{\Delta}}^T n \), are in one-to-one correspondence with each other, through mean value parametrization, in the sense that the observed minimal sufficient statistics is the expected value of the minimal sufficient statistics with respect to the distribution identified by the MLE itself. In fact, the MLE is determined as the inverse of the mean value parametrization evaluated at the sufficient statistics. Furthermore, the mean value parametrization is, in fact, a homeomorphism between the natural parameter space and the cone generated by the columns of \( U_{\tilde{\Delta}} \), called the marginal cone.

The clear similarity between Equations (14) and (15) reveals that the penalized estimator \( \hat{\theta}_{GL} \) arises in a very similar fashion as the ordinary MLE, the crucial difference being that the mean value homeomorphism is no longer evaluated at the minimal sufficient statistics but at the different point \( U_{\tilde{\Delta}}^T \hat{m} \). Because, unlike the observed table \( n \), the vector of fitted values \( \hat{m} \) is strictly positive by construction, the point \( U_{\tilde{\Delta}}^T \hat{m} \) belongs to the relative interior of the marginal cone and, thus, can be seen as a smoothed version of the sufficient statistics \( U_{\tilde{\Delta}}^T n \). Geometrically, the penalty function pulls the sufficient statistics away from the boundary of the marginal cone. This forced amount of smoothness injected by regularization is controlled by the tuning parameters in \( \Lambda \), and affects the asymptotic properties of \( \hat{\theta} \), which, according to our results in Section 4.4, is asymptotically biased and inefficient. Nonetheless, the model estimator \( \hat{\Delta} \) is consistent.
4 Asymptotic Analysis

4.1 Introduction

We will provide an asymptotic analysis of the model selection procedure described in Table 3 by studying the properties of the group lasso estimator.

We will consider a rather general double-asymptotic framework, in which we allow both the sample size and the complexity of the statistical model to grow simultaneously. In particular, we will be assuming a sequence of statistical experiments consisting of log-linear models over an increasingly large set of cell combinations, implied by both a growing number of categorical variables and a growing number of levels for the variables, and with increasing sample size. To formally represent this sequence of experiments, we will introduce a “time” variable $n$, which serves merely as an index and is not necessarily a quantification of the rate of increase of the sample size. Intuitively, the larger the index $n$, the bigger the contingency table, the larger the sample size and the more complex the model selection problem.

To be specific, at time $n$,

- it is available a multinomial sample of size $N_n$ from the joint distribution of $K_n$ categorical variables, each defined over a finite set $I_{jn} = \{1, \ldots, I_{jn}\}$, $j_n = 1 \ldots, K_n$; the support of this distribution is the set $I_n = \bigotimes_{j_n} I_{jn}$ of all cell combinations, of cardinality $I_n = \prod_{j_n} I_{jn}$;

- the true underlying distribution is defined by a hierarchical log-linear model $\Delta_n$, as described in Section 2.1: the observed cell counts come from an exponential family distributions with log-likelihood function (9) and true natural parameter $\theta_0^n \in \mathbb{R}^{I_n-1}$, such that $\|\theta_{h_n}^n\|_2 > 0$ for $h_n \in H_n$ and $\|\theta_{h_n}^n\|_2 = 0$ for $h_n \notin H_n$, with $H_n$ defined as in (10);

- the vector of true parameters $\theta_0^n$ is estimated by solving the program (12) with tuning parameters $\Lambda_n = \{\lambda_n, \{\lambda_{h_n}, h_n \neq \emptyset\}\}$;

- the group lasso estimate $\hat{\theta}_n$ is then used to estimate $\Delta_n$ as described in Table 3, leading to the optimal selected model $\hat{\Delta}_n$.

In the rest of the article we will use the notation $\{t_n\} \in \bigotimes \mathbb{R}^{k_n}$ to denote a sequence of vectors such that $t_n \in \mathbb{R}^{k_n}$, for every $n$.

We remark that the true model at each “time point” $n$ needs not be related with the true models at different values of $n$. The sequential setting we adopt is a convenient device for representing very generally an asymptotic framework for log-linear model selection with a diverging number of parameters; in fact, there are many factors that may increase the complexity of a log-linear model (e.g., number of variables, number of interactions in the model, number of levels for each variable) that we found it convenient to just allow each of them to change at every $n$.

In our sequential setting, the probability spaces are allowed to change with $n$ and, when we speak of convergence in probability to a constant or of tightness with respect to the index $n$, we explicitly refer to a sequence of different probability measures. Accordingly, we will use the stochastic small and large order notation $o_{P_n}$ and $O_{P_n}$ respectively with an index $n$ for the probability measures. This notation is well-defined: see Definition 7.11 and Lemma 7.12 in Schervish (1998).

We will embed the true parameter $\theta_0^n$ in $\mathbb{R}^{d_{H_n}}$ and denote it with $\theta_0^{H_n}$. We will indicate with $\{\pi_0^n\}_n$ the sequence of true probability vectors and with $\{m_0^n\}_n$ the sequence of mean vectors, with
\[ m_n^0 = N_n \pi_n^0, \text{ for each } n. \]

One may take note that the Fisher information matrix at \( \theta_{H_n}^0 \) is

\[ F_{H_n} = U_{H_n}^\top \left( D \pi_n^0 \pi_n^0 \right) U_{H_n}, \]

with maximal and minimal eigenvalues denoted by \( l_n^{\text{max}} \) and \( l_n^{\text{min}} \), respectively. The negative Hessian of the log-likelihood function is

\[ \Sigma_{H_n} = U_{H_n}^\top \left( D m_n^0 - m_n^0 (m_n^0)^\top \right) U_{H_n} = N_n F_{H_n}, \]

which is the covariance matrix of the natural sufficient statistics \( U_{H_n}^\top n_n \).

In the remainder of the article we will study some of the asymptotic properties of the sequence of lasso estimates \( \{\hat{\theta}_n\}_n \) generated according to the previous scheme. Although framed in a more general setting, our results carry over to standard asymptotic framework in which only the sample size \( N_n \) and the set of penalty parameters \( \Lambda_n \) change with \( n \). In our analysis we will establish a series of progressively stronger results, which, naturally, demand increasingly stronger assumptions. In Section 4.2, we show that \( \hat{\theta}_n \) is a norm-consistent estimator of \( \theta_n^0 \) and in Section 4.3 we prove the stronger property of model selection consistency, i.e.

\[ \lim_{n} P(\hat{\Delta}_n = \Delta_n) \to 1. \tag{16} \]

Finally, in Section 4.4 we give a central limit theorem for \( \hat{\theta}_n \).

### 4.2 Estimation consistency

In this section we establish a rather weak, but nevertheless essential, consistency property for the group lasso estimator. Suppose we knew the sequence of true models \( \Delta_n \) and we estimated \( \theta_{H_n}^0 \) with the group lasso estimate \( \tilde{\theta}_n \) by solving (12), where the parameter space is now \( \mathbb{R}^{d_{H_n}} \). We give sufficient conditions to establish that \( \tilde{\theta}_n \) is a \( \ell_2 \)-consistent sequence of estimators, in the sense that, for each \( \epsilon > 0 \),

\[ \lim_{n} P\left( \|\tilde{\theta}_n - \theta_{H_n}^0\|_2 > \epsilon \right) = 0. \tag{17} \]

Because we are taking norms over Euclidian spaces of arbitrarily large dimensions, and using the chain of inequalities \( \ell_\infty \leq \ell_2 \leq \ell_1 \), the previous result implies \( \ell_\infty \)-consistency but not \( \ell_1 \)-consistency. For the same reasons, model selection consistency (16) does not follow from (17). In fact, according to our method of proof, estimation consistency (17) is a necessary condition for model consistency, for which we need additional conditions.

We show (17) by establishing a more refined consistency property for \( \tilde{\theta}_n \). Under the assumption of Theorem 4.1, the lasso estimator enjoys the same convergence rate as the maximum likelihood estimator for exponential families with diverging number of parameters found by Portnoy (1988). Following a similar remark in Portnoy (1988), we point out that the conclusion of the theorem holds only for individual sequences of true parameters \( \{\theta_{H_n}^0\}_n \) and may not be expected in general to hold uniformly over subsets of \( \times_n \mathbb{R}^{d_{H_n}} \).

**Theorem 4.1.** Consider the settings described in Section 4.1. Assume

1. [NC.1] \( d_{H_n} = o(N_n); \)

12
2. [NC.2] the uan condition (19) is satisfied;

3. NC.3] \( \sqrt{l_{\max}^n / l_{\min}^n} = O(1) \) where \( l_{\min}^n, l_{\max}^n \) are the minimal and maximal eigenvalue of the Fisher information matrix, respectively;

4. [NC.4] \( \lambda_n = O(\sqrt{d_n} N_n) \).

Then,

\[
\|\hat{\theta}_n - \theta^0_{H_n}\|_2 = O_{P_{\theta^0_{H_n}}}(\sqrt{d_n N_n}) \tag{18}
\]

The proof of Theorem 4.1 relies on a more general result about sequences of regular statistical experiments with diverging number of parameters, which is of independent relevance and is described in the next sub-section.

Remarks

Theorem 4.1 should be compared with Theorem 2.1 Portnoy (1988), establishing norm consistency of the MLE for exponential families with an increasing number of parameters. In fact, the conclusion of our theorem still holds if we replace [NC.2] with the condition expressed in Equation (2.4) in Portnoy (1988, page 359), which constraints the rate of growth of the expectations of third order moments over compact neighborhood of \( \theta^0_{H_n} \) (see also Ghosal, 2000, for similar conditions). Both our uan condition and Equation (2.4) in Portnoy (1988) are used to control the order of magnitude of the remainder term in the local quadratic approximation of the log-likelihood function around \( \theta^0_{H_n} \), uniformly over compact neighborhoods. This essentially guarantees that, with high probability and for large enough \( n \), the log-likelihood function behaves around \( \theta^0_{H_n} \) like a concave function. We conjecture that condition (2.4) in Portnoy (1988, page 359) is milder than our uan condition, but, as stated, it is only applicable to exponential families. Similarly, condition [NC.4] is needed to guarantee that, as \( n \) grows, the penalty function would not disrupt this local quadratic approximation. Therefore, by (18), the group lasso estimator will eventually lie, with high probability, within the same neighborhoods of the true parameter sequence as the MLE.

Theorem 4.1 is also similar to Theorem 1 in Fan and Peng (2004), whose assumptions result from a direct adaptation to the present settings of the classical regularity conditions for efficient likelihood estimation involving uniform (with respect to \( n \)) boundedness of the third-order derivatives of the density functions in neighborhoods of the true parameter vector (see, e.g. Lehman and Casella, 1998, Theorem 6.5.1). We remark that these assumptions lead to a much weaker control over the remainder term of the quadratic approximation, which must be compensated by requiring a faster rate of growth of the sample size compared to the number of parameters, namely \( d_n^2 N_n \rightarrow 0 \).

The condition [NC.3] on the eigenvalues of the Fisher information matrix is also used in Portnoy (1988, Theorem 2.1). Geometrically, it prevents the ellipsoids \( t_n^T I(\theta^0_{H_n}) t_n \) (see definition of \( I(\cdot) \) below) from getting over-stretched along some directions, thus destroying concavity. Statistically, it preserves identifiability of the true models, as \( n \rightarrow \infty \).

4.2.1 A quadratic approximation result for regular models in the double asymptotic setting

Consider a sequence of statistical models \( \{ (P_n, X_n, F_n) \} \), where, for each \( n \), \( P_n = \{ P_{\theta_n}, \theta_n \in \Theta_n \} \) is a family of probability distributions over some Borel space \( (X_n, F_n) \) that is dominated by some
Proposition 4.2. Consider the settings described above and assume further that, for all sequences \( \theta_n \) such that \( k_n \to \infty \) as \( n \to \infty \). We further assume that, for each \( n \), \((P_n, \mathbb{X}_n, \mathcal{F}_n)\) is regular at \( \theta_0^0 \) that is (see, e.g. Bickel et al., 1993), \( \theta_0^0 \in \text{int}(\Theta_n) \), \((P_n, \mathbb{X}_n, \mathcal{F}_n)\) is quadratic mean differentiable at \( \theta_0^0 \) with quadratic mean derivative \( \eta_{\theta_0^0} \) and the Fisher information matrix

\[
I(\theta_n) = 4 \int \eta_{\theta_n}^T \eta_{\theta_n} d\mu_n
\]

exists (see, e.g., Lehman and Romano, 2005, Chapter 12), is non-singular and continuous at \( \theta_0^0 \).

Associate to the sequence of models \((P_n, \mathbb{X}_n, \mathcal{F}_n)\) a sequence of experiments consisting, for each \( n \), of an iid sample \((X_1, \ldots, X_{N_n})\) of size \( N_n \) drawn according to a distribution indexed by an unknown parameter \( \theta_0^0 \in \text{int}(\Theta_n) \), where \( N_n \to +\infty \) in such a way that \( \sqrt{\frac{k_n}{N_n}} \to 0 \). Let \( \{\ell_n\}_n \) denote the sequence of likelihood functions and let

\[
Z_n = \frac{1}{\sqrt{N_n}} \sum_{i=1}^{N_n} 2 \eta_{\theta_0^0}(X_i) \sqrt{p_{\theta_0^0}(X_i)}
\]

be the score function.

Proposition 4.2. Consider the settings described above and assume further that, for all sequences \( \{t_n\}_n \in \mathbb{R}^{k_n} \) such that \( t_n = \sqrt{\frac{k_n}{N_n}} x_n \), with \( \|x_n\|_2 \leq C \) for each \( n \), the following uan condition is satisfied

\[
\max_{i_n} |T_{i_n}| = o_{P_{\theta_0^0}}(1), \quad (19)
\]

where

\[
T_{i_n} = \left( \frac{\sqrt{p_{\theta_0^0} + \frac{1}{\sqrt{N_n}} t_n}}{\sqrt{p_{\theta_0^0}}}(X_{i_n}) - 1 \right).
\]

Then, for any finite \( C > 0 \),

\[
\ell_n \left( \theta_0^0 + \frac{t_n}{\sqrt{N_n}} \right) - \ell_n(\theta_0^0) = \langle t_n, Z_n \rangle - \frac{1}{2} t_n^T I(\theta_n^0) t_n(1 + o_{P_{\theta_0^0}}(1))
\]

as \( n \to \infty \), uniformly over all sequences \( \{t_n\}_n \) defined above. Furthermore, sufficient conditions for (19) are that \( \frac{\|x_n\|_2^2}{N_n} \to 0 \), with \( \|x_n\|_2^2 \) the largest eigenvalue of \( I(\theta_n^0) \), and that, for each sequence \( \{t_n\}_n \) such that \( \sum_{i=1}^{N_n} \|x_n\|_2^2 \to 0 \),

\[
\int \left( \sqrt{p_{\theta_0^0} + \frac{t_n}{\sqrt{N_n}}} - \sqrt{p_{\theta_0^0}} - t_{\eta_{\theta_0^0}}^T \eta_{\theta_0^0} \right)^2 \ d\mu_n = o \left( \frac{\|x_n\|_2^2}{k_n} \right), \quad (20)
\]

as \( n \to \infty \).

Remarks

The above result is a natural extension to our settings of the standard local asymptotic approximation of the log-likelihood function for regular statistical models with fixed-size parameter space. The intuition for considering sequences of parameters living within a \( O(\sqrt{k_n}) \) neighborhood comes from simple high-dimensional geometry. In fact, for fixed-size parameter space, it is well known that the local approximation holds uniformly for values of the local parameter within a blow-up.
versions by a factor of \(\sqrt{N_n}\) of sets of order \(O(1)\). That is, as more data become available, the local view gets finer and the quadratic approximation gets better, uniformly for values of the original parameter over compact balls of radii shrinking as \(1/\sqrt{N_n}\). When the dimension of the parameter space \(k_n\) increases with \(n\) as described above, the positive effect of a larger sample over the quality of the local approximation, measured by \(\sqrt{N_n}\), is diminished by the increase in complexity, measured by \(\sqrt{k_n}\). To see this geometrically, notice that, in order for a sequence of closed balls in Euclidean spaces of different dimensions to look proportionally the same, each ball must be a rescaled version of the unit-volume ball in its ambient space. That is, the radii need to grow like \(\sqrt{k_n}\). If no such adjustment is made, any compact ball of fixed radius would very rapidly become minuscule as \(n\) grows, merely due to the increase in dimensionality. Consequently, the local parameters are now free to vary within balls that grow slower with \(N_n\), at the rate \(\sqrt{N_n/k_n}\).

Besides weaker requirements on the amount of smoothness, there is one more reason why quadratic mean differentiability seems particularly convenient. In fact, for regular models, the Hellinger distance induces on \(\Theta_n\) approximately the same topology as the Euclidian distance, at least locally, so that quadratic mean differentiability represents quite naturally the underlying geometric changes induced by a simultaneous increase in dimensionality and sample size.

When the parameter space has fixed dimension, the condition (20) reduces to quadratic mean differentiability and the uan condition (19) is always satisfied in regular models. However, with increasing dimensions, the uan condition guarantees an error of order \(o_P(\theta_0/k_n)\) and is a much weaker requirement than (20), which can be shown to imply an error of order \(o_P(1)\), just as in the fixed dimensional case. In fact, assumption (20) is strong enough to eliminate the effect of increasing dimensionality, by having the rate of increase in the sample size \(N_n\), amplified by the rate of increase of the parameter space, i.e. \(k_n\), so that the norm of \(1/\sqrt{N_n}t_n\) is dimension independent. This can be achieved if \(k_n = o(N_n^\alpha)\), for some \(1/2 \leq \alpha < 1\), the case \(\alpha = 1/2\) being what is needed for some of the central limit results we give in Section 4.4. Finally, it is worth pointing out that quadratic mean differentiability at each \(n\) implies (20) only along subsequences \(\{(P_{n_j}, \mathcal{X}_{n_j}, \mathcal{F}_{n_j})\}_{j}\) of models.

4.3 Model Selection Consistency

Having established norm consistency for the group lasso solution, we then proceed to derive sufficient conditions for the stronger property of model selection consistency (16). Our method of analysis is based on linearizing the sub-gradient optimality conditions (13) via a Taylor expansion around the sequence of true parameters \(\theta_0\). Norm consistency results turn out to be necessary to guarantee enough stochastic control over the remainder term of that expansion. The conditions we develop are quite similar in spirit to the ones arising from the study of sparse recovery of a linear signals under Gaussian or white noise using the lasso penalty (see, in particular, Wainwright, 2006; Zhao and Yu, 2006).

Recall the definition of \(\mathcal{H}_n\) from (10) and let \(\mathcal{H}_n^c = 2^{K_n} \setminus (\mathcal{H}_n \cap \emptyset)\), so that \(\|\theta_0\|_2 > 0\) for each \(h_n \in \mathcal{H}_n\) and \(\|\theta_0\|_2 = 0\) for each \(w_n \in \mathcal{H}_n^c\). Consider the sequence of events

\[
O_n = \left\{\|\hat{\theta}_{h_n}\|_2 > 0, \forall h_n \in \mathcal{H}_n\right\} \cap \left\{\|\hat{\theta}_{w_n}\|_2 = 0, \forall w_n \in \mathcal{H}_n^c\right\}.
\]  

(21)

Then, the model selection consistency property (16) of the group lasso solutions \(\hat{\theta}_n\) is equivalent to
convergence in probability of \( O_n \), namely
\[
\lim_n \mathbb{P}(O_n) = 1. \tag{22}
\]

Define the two random vectors
\[
\hat{\eta}_{H_n} = \text{vec}\{ \lambda_{h_n} \frac{\hat{\theta}_{h_n}}{\| \hat{\theta}_{h_n} \|_2}, h_n \in \mathcal{H}_n \}
\]
and
\[
\hat{\zeta}_{H_n} = \text{vec}\{ \lambda_{w_n} \hat{z}_{w_n}, w_n \in \mathcal{H}^c_n \}, \tag{23}
\]
where \( \| \hat{z}_{w_n} \|_2 \leq 1 \).

The vector \( \hat{\eta}_{H_n} \) is an explicit function of the group lasso estimator (it is not related to the quadratic mean derivative \( \eta_{\theta_0} \)). Using the optimality conditions (13) and applying a Taylor expansion of \( \hat{m}_n \) around \( m_0 \), it can be verified that \( O_n \) holds if and only if both the equations
\[
\hat{\theta}_{H_n} = \theta_{\mathcal{H}_n}^0 + N_n \Sigma_{\mathcal{H}_n}^{-1} \left( \frac{1}{N_n} U_{\mathcal{H}_n}^T (n_n - m_n^0) - \frac{1}{N_n} U_{\mathcal{H}_n}^T R_n - \lambda_n \hat{\eta}_{H_n} \right) \tag{24}
\]
and
\[
\lambda_n \hat{\zeta}_{H_n} = \frac{1}{N_n} U_{\mathcal{H}_n}^T (n_n - m_n^0) - \frac{1}{N_n} U_{\mathcal{H}_n}^T R_n - W_n \left( \frac{1}{N_n} U_{\mathcal{H}_n}^T (n_n - m_n^0) - \frac{1}{N_n} U_{\mathcal{H}_n}^T R_n - \lambda_n \hat{\eta}_{H_n} \right), \tag{25}
\]
hold, where \( \| R_n \|_2 = o(\| \hat{\theta}_{H_n} - \theta_{\mathcal{H}_n}^0 \|_2) \), and
\[
W_n = U_{\mathcal{H}_n}^T \left( D_{m_n^0} \frac{m_n^0}{N_n} \frac{m_n^0}{N_n}^T \right) U_{\mathcal{H}_n} \Sigma_{\mathcal{H}_n}^{-1}.
\]

We rely on Equations (24) and (25) to derive sufficient conditions guaranteeing that
\[
\lim_n \mathbb{P}(\| \hat{\theta}_{h_n} \|_2 > 0, \forall h_n \in \mathcal{H}_n) = 1 \tag{26}
\]
and
\[
\lim_n \mathbb{P} \left( \max_{w_n \in \mathcal{H}^c_n} \| \hat{z}_{w_n} \|_2 \leq 1 \right) = 1. \tag{27}
\]

In turn, (26) and (27) together imply that (22) holds.

Remarks
We point out that equation (27) in particular entails an asymptotic evaluation of the probability of \( \ell_2 \)-unit balls in high dimensional settings, whereas for the model selection consistency of the usual lasso procedure it is enough to consider only the \( \ell_{\infty} \)-unit ball. This difference is rather substantial and demands a much stronger control of the asymptotic behavior of the group lasso estimator than it is needed for the ordinary lasso estimator. One can see this geometrically, for example, by noticing that, in \( \mathbb{R}^k \), the volume of the \( \ell_{\infty} \)-unit ball (i.e. the unit cube \( \text{conv}\{-1, +1\}^k \)) is \( 2^k \), while the volume of the \( \ell_2 \)-unit ball (i.e. the unit sphere) is roughly \( \left( \sqrt{\frac{2\pi e}{k}} \right)^k \), so that, for large \( k \), the unit sphere is just an extremely tiny fragment of the cube. A more refined geometric argument can
be made using Dvoretsky’s Theorem (see, e.g., Pisier, 1999), according to which the $\ell_2$-unit ball can be approximated arbitrary well by almost spherical sections of the $\ell_\infty$-unit ball of dimension about $\log k$. That is, when seen from inside the $k$-dimensional unit cube, the unit sphere looks approximately like a slice of a sub-space of dimension $\log k$. Probabilistically, one can also observe that for a $k$-dimensional vector $X = (X_1, \ldots, X_k)$ of i.i.d. random variables with sub-gaussian tails, $\mathbb{E}\|X\|_\infty = O(\sqrt{\log k})$, while $\mathbb{E}\|X\|_2 = O(\sqrt{k})$, as $k \to \infty$, so that control over the $\ell_\infty$ unit ball can be achieved by smaller sample sizes.

The restrictions imposed by (27) are due to the choice of the quadratic norm for the penalty function and cannot be avoided. By the same token, using an $\ell_\infty$ norm in the penalty function would require dealing with $\ell_1$–unit balls, which are even smaller than $\ell_2$–unit balls and in general harder to control. In this respect, the $\ell_1$ lasso penalty, which results in a $\ell_\infty$ norm condition for the sub-gradient, appears to be optimal.

We will specialize the assumption [NC.3] about the eigenvalues of the Fisher information by requiring that

$$[\text{NC.3′}] 0 < D_{\min} < l_{\min}^0 \leq l_{\max}^0 < D_{\max} < \infty.$$ 

Note that this condition does not preclude $\max_{i_n} \pi_{i_n}^0$ from decreasing to zero and, in fact, this assumption will be used in Section 4.4.

Theorem 4.3. Assume the conditions of norm consistency [NC.1], [NC.2], [NC.3′] and [NC.4]. Then, Equation (26) holds if

1. [MSC.1] letting $\alpha_n = \min_{h_n \in \mathcal{H}_n} \|\theta^0_{h_n}\|_2$,

$$\frac{1}{\alpha_n} \left( \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} + \lambda_n \sqrt{\sum_{h_n \in \mathcal{H}_n} \lambda_{h_n}^2} \right) \to 0,$$

which, for $\lambda_{h_n} = \sqrt{d_{h_n}}$, simplifies to

$$\frac{\sqrt{d_{\mathcal{H}_n}}}{\alpha_n} \left( \sqrt{\frac{1}{N_n}} + \lambda_n \right) \to 0.$$

Equation (27) holds if

1. [MSC.2] (‘almost’ parameter orthogonality) for each $w_n \in \mathcal{H}_c^n$,

$$\left\| U_{w_n}^\top \left( D_{m_n}^0 - \frac{m_n^0(m_n^0)^\top}{N_n} \right) U_{\mathcal{H}_n} \Sigma^{-1}_{\mathcal{H}_n} \right\|_2 < \frac{(1 - \epsilon)}{|\mathcal{H}_c^n|};$$

2. [MSC.3]

$$\lim_n \frac{|\mathcal{H}_n| \max_{h_n \in \mathcal{H}_n} \lambda_{h_n}}{|\mathcal{H}_c^n| \min_{w_n \in \mathcal{H}_c^n} \lambda_{w_n}} \leq 1;$$

3. [MSC.4]

$$\max_{w_n \in \mathcal{H}_c^n} \frac{\sqrt{d_{w_n}}}{\sqrt{N_n} \lambda_{w_n}} \to 0$$

which, for the choice $\lambda_{h_n} = \sqrt{d_{h_n}}$, requires $\lambda_{h_n}^2 N_n \to \infty$. 

17
Remarks

The condition [MSC.2] implies that
\[
\|W_n\|_2 = \left\| U_{\mathcal{H}_n}^\top \left( D_{m_0} - \frac{m_0 m_0^\top}{N_n} \right) U_{\mathcal{H}_n} \Sigma^{-1}_{\mathcal{H}_n} \right\|_2 < (1 - \epsilon),
\]
which is the equivalent of the irreducibility condition used in Wainwright (2006) and Zhao and Yu (2006), using the Euclidian norm. In fact, under the Gaussian homoschedastic ensemble settings, the irreducibility condition on the design matrices is precisely an ‘almost’ parameter orthogonality condition on the Fisher information matrix for the saturated model. The notable difference in our [MSC.2] is the need to account for the number of blocks of parameter which are zero.

MSC.3 is a sparsity condition that imposes bounds on the model complexity in terms of the number of interactions that can be recovered.

Condition MSC.4 indirectly provides some information on the rates of increase for the dimensions for the subspaces of interactions not included in the true models, i.e. of \(I_n\). In fact, setting \(\lambda_{h_n} = 1\) for all \(h_n\) and \(n\), and noting that
\[
\max_{w_n \in \mathcal{H}_n} d_{w_n} = \prod_{j_n \in \mathcal{K}_n} (I_{j_n} - 1),
\]
[MSC.4] requires
\[
\frac{\prod_{j_n \in \mathcal{K}_n} (I_{j_n} - 1)}{\lambda_{h_n}^2 N_n} \rightarrow 0.
\]

Furthermore, since,
\[
d_{\mathcal{H}_n} = \sum_{w_n \in \mathcal{H}_n} d_{w_n} = (I_n - 1) - d_{\mathcal{H}_n},
\]
a more stringent sufficient conditions implied by [MSC.4] (compare, e.g., the bound in Equation (15) a) in Wainwright, 2006) is
\[
\frac{I_n - d_{\mathcal{H}_n}}{\lambda_{h_n}^2 N_n} \rightarrow 0.
\]
For \(\lambda_{h_n} = \sqrt{d_{h_n}}\), [MSC.4] reduces to \(\lambda_{h_n}^2 N_n \rightarrow \infty\), so that the block penalties are of the exact order of magnitude needed to eliminate the effect of the direct dependence of \(I_n\) on \(N_n\) displayed above.

The proof of the Theorem shows that it is possible to make more refined assumptions, so that [MSC.1] may be specialized as follow.

**Addendum 4.4.** Let \(h_n^c = \mathcal{H}_n \setminus h_n\), then Equation (26) holds if

1. [MCS.1’]
\[
\left\| U_{h_n}^\top \left( D_{m} - \frac{m m^\top}{N_n} \right) U_{h_n} \left( U_{h_n}^\top \left( D_{m} - \frac{m m^\top}{N_n} \right) U_{h_n} \right)^{-1} \right\|_2 \leq \frac{(1 - \epsilon)}{|\mathcal{H}_n|} \tag{30}
\]
and
\[
\frac{1}{\alpha_n} \left( \frac{1}{N_n} \max_{h_n \in \mathcal{H}_n} \sqrt{d_{h_n}} + \lambda_{h_n} \max_{h_n \in \mathcal{H}_n} \lambda_{h_n} \right) \rightarrow 0, \tag{31}
\]
which, for \(\lambda_{h_n} = \sqrt{d_{h_n}}\), simplifies to
\[
\frac{\max_{h_n \in \mathcal{H}_n} \sqrt{d_{h_n}}}{\alpha_n} \left( \frac{1}{N_n} + \lambda_{h_n} \right) \rightarrow 0.
\]
This condition (30) requires the Fisher information matrix to behave approximately as a block diagonal matrix, with the blocks given by the covariance matrices $\Sigma_{h_n}$, $h_n \in \mathcal{H}_n$. If this is the case, then (26) is verified under (31), which is a much weaker requirement than (28).

### 4.4 A Central Limit Theorem for the Log-linear Group Lasso Estimator

Our final results concern the large sample properties of the distribution of lasso group estimates $\{\hat{\theta}_n\}_n$. In addition to the conditions guaranteeing both norm and model selection consistency, we need to impose further restrictions guaranteeing some form of asymptotic normality under our double asymptotic framework. The main rationale behind retaining the set of assumptions for consistency is that they allow us to work only with the simpler and well-behaved sequence of events $O_n$ defined in (22), which converges in probability.

In general, asymptotic normality under the double asymptotic settings obtains under stricter assumptions than under standard (i.e. with fixed-dimensional parameter space) asymptotic problems. Below, we provide a series of conditions, each providing a sense that for large enough $n$, the group lasso estimates (appropriately rescaled and translated) are close to a standard Normal distribution. We point out that only the first two results in the theorem statement amount to a full central limit statement, while the third result only offers necessary condition for asymptotic normality.

To state our result, we need to formulate some notation. Let $J^0_{\mathcal{H}_n}$ be a $d_{\mathcal{H}_n} \times d_{\mathcal{H}_n}$ block-diagonal matrix whose $h_n$-block is the $d_{h_n} \times d_{h_n}$ matrix

$$
\lambda_{h_n} \frac{1}{\|\theta_{h_n}^0\|_2} \left( I_{d_{h_n}} - \frac{\theta_{h_n}^0}{\|\theta_{h_n}^0\|_2} \left( \frac{\theta_{h_n}^0}{\|\theta_{h_n}^0\|_2} \right)^\top \right),
$$

with $h_n \in \mathcal{H}_n$ and with $I_{d_{h_n}}$ denoting the $d_{h_n}$-dimensional identity matrix. In the following, $C_n$ will denote a $k \times d_{\mathcal{H}_n}$ matrix, where $k$ is an arbitrary fixed number, such that

$$
\lim_{n} C_n C_n^\top = G
$$

for some $k \times k$ nonnegative and symmetric matrix $G$.

**Theorem 4.5.** Assume the conditions for norm and model selection consistency and let

$$
X_n = \sqrt{N_n} F_{\mathcal{H}_n}^{-1/2} \left( \left( F_{\mathcal{H}_n} + \lambda_n J^0_{\mathcal{H}_n} \right) (\hat{\theta}_n - \theta_0^0) + \lambda_n \eta_{\mathcal{H}_n} \right).
$$

1. For each sequence $\{C_n\}$ of $k \times d_{\mathcal{H}_n}$ matrices satisfying (32), $C_n X_n$ converges weakly to the $N_k(0, G)$ distribution if either the [CLT] condition

$$
d_{\mathcal{H}_n} = o(N_n^{1/2})
$$

or both the [CLT.Ma] condition

$$
d_{\mathcal{H}_n} = o(N_n)
$$

and [CLT.Mb] condition

$$
\max_{i \in \mathcal{I}_n} \pi_i^0 = O \left( \frac{1}{\sqrt{N_n d_{\mathcal{H}_n}}} \right).
$$

hold.
2. If the [CLT.BE] condition
\[ d_{H_n} = o\left(\frac{N_n^{2/7}}{n}\right), \]
holds, then
\[ \sup_{A_n} |\mathbb{P}(X_n \in A_n) - \mathbb{P}(Z_n \in A_n)| \to 0, \]
where \( Z_n \) has a \( N_{d_{H_n}}(0, I_{d_{H_n}}) \) distribution, the supremum is taken over the convex sets \( A_n \) in \( \mathbb{R}^{d_{H_n}} \) and convergence occurs at the rate \( O\left(\frac{d_{H_n}^{1/2}}{N_n}\right) \).

3. Let \( \psi_n \) and \( \phi_n \) denote the characteristic functions of \( X_n \) and \( Z_n \), respectively. Then, for each \( \epsilon > 0 \) and \( T > 0 \), there exists a \( n_0(\epsilon, T) \) such that, for all \( n > n_0(\epsilon, T) \),
\[ \sup \left\{ |\psi_n(t) - \phi_n(t)| : \|t\|_2 \leq T \right\} < \epsilon. \tag{34} \]
if the [CLT.LF] condition
\[ d_{H_n} = o\left(\frac{N_n^{1/3}}{n}\right) \]
holds.

Remarks

1. The theorem is mainly of theoretical interest, and it indicates that the group lasso estimate is asymptotically unbiased and inefficient. In fact, Equation (33) demonstrates that the asymptotic behavior of the group lasso estimator is affected by two terms. One is the bias term \( \lambda_n \eta_{H_n}^0 \) which depends on the gradient of the penalty function at the true parameter. The other term \( J_{H_n}^0 \) is the Hessian at the true parameter of the penalty function, a positive definite matrix which inflates the inverse Fisher information. Both these terms are asymptotically significant and indicate that the group lasso estimates may lack asymptotic optimality. Note that this phenomenon is probably quite general. In fact, additional terms of this sort appear also in a similar result in (Fan and Peng, 2004, Theorem 2).

2. Condition [CLT] is a rather weak one but nonetheless it guarantees, via a simple Lindberg-Feller argument, the asymptotic normality of a fixed number of linear combinations of the coordinates of \( \hat{\theta}_n \). In particular, it includes that case of
\[ C_n = \left[ I_k \quad O \right] \]
where \( O \) is a \( k \times (d_{H_n} - k) \) matrix of zeros. For this particular choice, [CLT] guarantees the marginal asymptotic normality of any fixed number of coordinates of \( \hat{\theta}_n \).

3. Condition [CLT.BE] is a full central limit type of results for the group lasso estimator and is based on a multivariate Barry-Esseen type of bound found in Bentkus (2003). As it is usual with uniform results of this type, it is necessary to control the fluctuations of third order moments, and, consequently, to have a rather large sample size. To our knowledge, this is the best rate available. See also Portnoy (1986) for a similar result requiring only a rate \( d_{H_n}^2 = o\left(\frac{N_n^{1/2}}{n}\right) \), whose applicability and relevance to our problem is however unclear.
4. The last set of conditions, [CLT.LF], and [CLT.Ma] and [CLT.Mb], establish control over the second moments and are thus quite mild. They both lead to a double-asymptotic version of Lindberg-Feller condition, with [CLT.Ma] and [CLT.Mb] stemming from on a more elaborated method of proof relying on a generalization of some results by Morris (1975). The Lindberg-Feller condition [CLT.LF] requires a slower rate of increase of \( d_{H_n} \) with respect to the sample size \( N_n \) than the other conditions based on [CLT.Ma] and [CLT.Mb], which however demands some control over speed at which the maximal cell probability tends to 0. However, they do not produce a central limit type of result. They only guarantee that, for \( n \) large enough, the characteristic function of an appropriate affine function of the group lasso estimate is very close to the characteristic function of a standard Gaussian, within a large compact balls. Although it is well known that, for multidimensional problems, closeness of characteristic function by itself is not enough to guarantee closeness of multivariate distributions (Senatov, 1998, see, e.g.), nonetheless, this result provides some sense that the group lasso estimate may behave, for large \( n \), like a Gaussian vector.

5. Instead of [CLT.Mb], one may want to enforce the more specialized assumption \( \max_{i \in I_n} \pi_i^0 \leq CI_n^{-1} \), for some positive constant \( C \), (see, e.g. Quine and Robinson, 1984, Theorem 1). Then, in order for the Theorem to hold, one has to further assume that

\[
\frac{I_n}{\sqrt{d_{H_n} N_n}} = o(1),
\]

which is compatible with the conditions for norm consistency.

5 Conclusions

Our problem and results differ from existing analyses of \( \ell_1 \) regularized least square problems in at least three aspects. Firstly, unlike the case of regularized least squares or Gaussian error problems, the first order optimality conditions for the group lasso program are non-linear in the parameters. As model selection consistency hinges upon establishing appropriate bounds for the norms of the differences between the blocks of true and estimated parameters, our strategy in Section 4.2 was to linearize the sub-gradient equations via a first order Taylor expansion. This expansion, in turn, is valid provided one has enough control over the remainder term, which we achieved by proving the norm consistency property for the group lasso estimate. Thus, in our settings, norm consistency is necessary for model selection consistency. In contrast, for quadratic problems, whose first order conditions are linear in the parameters, norm consistency does not appear to be needed, although, it may still be important for central limit results, like in our case.

Secondly, the sub-gradient optimality conditions for the group penalty function are formulated in terms of the \( \ell_2 \) norms of groups of parameters, and not in term of the \( \ell_\infty \) norm for the whole parameter vector, which is the case for lasso-based procedure. As discussed in section 4.3, this difference is crucial and is the main reason why our convergence rates require a much larger sample size than for the \( \ell_1 \) penalty.

Thirdly, in our problem we do not need to concern ourselves with random design. This is a consequence of the contingency table settings and simplifies the analysis. More generally, as
we are working with exponential families of distribution, the Fisher information matrix is data-

independent. Consequently, unlike for example the case of Gaussian ensembles, for model selection

consistency it is sufficient to impose analytic, and not stochastic, conditions on the asymptotic

behavior of the Fisher information. These conditions (namely, the almost parameter orthogonality’

condition [MSC.2]) correspond to the various irreducibility condition used in the lasso literature,

that we equivalently formulate in terms of the Fisher information.

Except for the irreducibility conditions, we did not take advantage of the exponential nature of

multinomial distribution, so that the norm consistency and central limit results of sections 4.2 and

4.4, respectively rely on rather general properties that may hold for other families of distributions.

6 Proofs

Proof of Lemma 3.1. The first order optimality conditions for a vector \( \theta \in \mathbb{R}^{I-1} \) is \( 0 \in \partial P_\Lambda (\theta) \),

the sub-differential set of \( P_\Lambda (\theta) \). The gradient of \( \ell(\theta) \) was already given in Equation (2). As for the

penalty term, which is not differentiable when some of the blocks are zero, we use the fact that if

\( \| \cdot \| \) is a norm in an Banach space \( X \), then the subdifferential at a point \( x \) is

\[
\partial \| x \| = \begin{cases} 
\{ x^* \in X^*: \| x^* \| \leq 1 \} & \text{if } x = 0 \\
\{ x^* \in X^*: \| x^* \| = 1, \langle x, x^* \rangle = \| x \| \} & \text{if } x \neq 0,
\end{cases}
\]

where \( X^* \) denotes the dual space of \( X \). Then, since the dual space of \( L_2 \) is \( L_2 \), we conclude that for any \( x \in \mathbb{R}^{(I-1)} \), the subdifferential of \( \sum_{h,h \neq \emptyset} \lambda_h \| x_h \|_2 \) at \( \theta \) is a subset of \( \mathbb{R}^{(I-1)} \) comprised by vectors whose \( h \) block component is

\[
\{ \lambda_h x^h \in \mathbb{R}^h: \| x^h \|_2 \leq 1 \} \quad \text{if } \theta = 0 \\
\lambda_h \| \theta_h \|_2 \quad \text{if } \theta \neq 0.
\]

Equation (2) and (35) implies (13). As for uniqueness, the results in Rinaldo (2006a, Section 3)

show that the solution to the likelihood equations is always unique, unless the maximum likelihood

estimator does not exist, in which case for every sequence of parameters \( \{ \theta_n \}_n \) such that

\[
\| \theta_n \| \to \infty. \quad \text{However, the penalty term would prevent this from happening. This, combined with}
\]

the strict convexity of the \( \ell_2 \) norm, guarantees uniqueness. \( \blacksquare \)

Proof of Proposition 4.2. We follow the proof of Lehman and Romano (2005, Theorem 12.3) (see

also Bickel et al., 1993, 509-513). Using a Taylor expansion of \( \log(1 + x) \), we can write

\[
\ell_n \left( \theta_n^0 + \frac{t_n}{\sqrt{N_n}} \right) - \ell_n (\theta_n^0) = 2 \sum_i T_{i_n} - \sum_i T_{i_n}^2 \left( 1 - 2r(T_{i_n}) \right),
\]

where \( r(T_{i_n}) \to 0 \) as \( T_{i_n} \to 0 \). Because of the uan assumption (19),

\[
r(T_{i_n}) = o_P(1).
\]

22
Next, by quadratic mean differentiability, for each \( \{t_n\}_n \),

\[
\int \left( \sqrt{p_{\theta_0}^{(t_n)}} - \sqrt{p_{\theta_0}^{(t_n)}} - \frac{1}{\sqrt{N_n}} t_n^\top \eta_{\theta_0}^{(t_n)} \right)^2 d\mu_n = o \left( \frac{k_n}{N_n} \right),
\]

which implies that

\[
N_n \int \left( \sqrt{p_{\theta_0}^{(t_n)}} - \sqrt{p_{\theta_0}^{(t_n)}} \right)^2 d\mu_n = t_n^\top I(\theta_0^{(t_n)}) t_n + o (k_n).
\]

Then,

\[
\sum_{i_n} E_{\theta_0^{(t_n)}} T_{i_n} = -\frac{N_n}{2} \int \left( \sqrt{p_{\theta_0}^{(t_n)}} - \sqrt{p_{\theta_0}^{(t_n)}} \right)^2 d\mu_n = -\frac{1}{8} t_n^\top I(\theta_0^{(t_n)}) t_n + o (k_n).
\]

Next, because

\[
T_{i_n} = \frac{1}{2} \langle t_n, Z_{i_n} \rangle + \frac{R_n}{\sqrt{p_{\theta_0}^{(t_n)}}(X_{i_n})}
\]

we have

\[
\sum_{i_n} (T_{i_n} - E_{\theta_0^{(t_n)}} T_{i_n}) = \frac{1}{2} \langle t_n, Z_n \rangle + \sum_{i_n} \left( \frac{R_n}{\sqrt{p_{\theta_0}^{(t_n)}}(X_{i_n})} - E_{\theta_0^{(t_n)}} \frac{R_n}{\sqrt{p_{\theta_0}^{(t_n)}}(X_{i_n})} \right),
\]

where \( \int R_n^2 d\mu_n = o \left( \frac{k_n}{N_n} \right) \), and, since

\[
N_n \mathbb{E}_{\theta_0^{(t_n)}} \frac{R_n^2}{p_{\theta_0}^{(t_n)}(X_{i_n})} = o(k_n),
\]

we arrive at

\[
\sum_{i_n} (T_{i_n} - E_{\theta_0^{(t_n)}} T_{i_n}) - \frac{1}{2} \langle t_n, Z_n \rangle = o_{\mathbb{P}_{\theta_0}} (k_n).
\]

Finally, noting that

\[
E_{\theta_0^{(t_n)}} T_{i_n}^2 = \int \left( \sqrt{p_{\theta_0}^{(t_n)}} - \sqrt{p_{\theta_0}^{(t_n)}} \right)^2 d\mu_n = \frac{1}{4N_n} t_n^\top I(\theta_0^{(t_n)}) t_n + o \left( \frac{k_n}{N_n} \right).
\]

and using (19), we conclude, by the weak law of large numbers, that

\[
\sum_{i_n} T_{i_n}^2 = \frac{1}{4} t_n^\top I(\theta_0^{(t_n)}) t_n + o_{\mathbb{P}_{\theta_0}} (k_n).
\]

Combining (36), (38), (40) and (41), we get

\[
\ell_n \left( \theta_0^{(t_n)} + \frac{t_n}{\sqrt{N_n}} \right) - \ell_n (\theta_0^{(t_n)}) = \langle t_n, Z_n \rangle - \frac{1}{2} t_n^\top I(\theta_0^{(t_n)}) t_n + o_{\mathbb{P}_{\theta_0}} (k_n),
\]

which gives the desired result, since \( t_n^\top I(\theta_0^{(t_n)}) t_n = O(k_n^{\max}) \).
To show that (20) implies (19), write for convenience $S_{in} = 2 \frac{p_{\theta_0}(X_{in})}{\sqrt{p_{\theta_0}(X_{in})}}$ and notice that $t_n^TS_{in}$ is of order $O(l_n^{\max}k_n)$ and that, under (20),

$$
\mathbb{E}_{\theta_n} \left( T_{in} - \langle t_n \frac{1}{\sqrt{N_n}}, S_{in} \rangle \right)^2 = 4 \int \left( \frac{p_{\theta_n} + t_n}{\sqrt{p_{\theta_n}}} - \frac{1}{\sqrt{N_n}} t_n^T \eta_{\theta_n} \right)^2 d\mu_n = o \left( \frac{1}{N_n} \right).
$$

Then, following Bickel et al. (1993, page 510), for each $\epsilon > 0$,

$$
P \left( |T_n| > \epsilon \right) \leq P \left( |T_n - \langle t_n \frac{1}{\sqrt{N_n}}, S_{in} \rangle| \geq \frac{\epsilon}{2} \right) + P \left( |\langle t_n \frac{1}{\sqrt{N_n}}, S_{in} \rangle| \geq \frac{\epsilon}{2} \right)
$$

$$
\leq \frac{4}{\epsilon^2} \mathbb{E}_{\theta_n} \left( T_{in} - \langle t_n \frac{1}{\sqrt{N_n}}, S_{in} \rangle \right)^2 + \frac{4}{\epsilon^2} \mathbb{E} \left\| \frac{1}{\sqrt{N_n}} t_n^T S_{in} \right\|^2 I \left\{ \left\| \frac{1}{\sqrt{N_n}} t_n^T S_{in} \right\|_2 > \epsilon \right\}
$$

$$
= o \left( \frac{1}{N_n} \right),
$$

which implies (19).

**Proof of Theorem 4.1.** For ease of notation we write $O_{P_n}$ and $o_{P_n}$ for $O_{P_{\theta_0}}$ and $o_{P_{\theta_0}}$, respectively. Also, for simplicity, we multiply $P_{\Lambda_n}$ by $N_n$ in (12). The conditions of proposition 4.2 are satisfied for the sequence of exponential families under considerations, with $\eta_{\theta_n} = \nabla p_{\theta_n} \frac{1}{\sqrt{p_{\theta_n}}}, \theta_n = \theta_0$ and $k_n = d_{\mathcal{H}_n}$. Then, for any finite $C > 0$ and for each sequence $\{x_n\}_n \in \bigotimes_n \mathbb{R}^{d_{\mathcal{H}_n}}$ with $\|x_n\|_2 \leq C$, the term $\ell_n \left( \theta_0_{\mathcal{H}_n} + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n \right) - \ell_n(\theta_0_{\mathcal{H}_n})$ is equal to

$$
\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n^T \mathbb{U}_{\mathcal{H}_n} (m_n - m_0_n) - \frac{1}{2} \frac{d_{\mathcal{H}_n}}{N_n} x_n^T \Sigma_{\mathcal{H}_n} x_n (1 + o_{P_n}(1)).
$$

For the first term, we notice that

$$
\mathbb{E}_{\theta_n} \left\| \mathbb{U}_{\mathcal{H}_n}^T (m_n - m_0_n) \right\|^2 = \text{tr} \left( \mathbb{U}_{\mathcal{H}_n}^T \left( D m_0_n - \frac{m_0_n(m_0_n)^T}{N_n} \right) \right) \leq N_n k_n d_{\mathcal{H}_n}.
$$

Thus, by Chebyshev and Cauchy-Swartz inequalities,

$$
\left| \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n^T \mathbb{U}_{\mathcal{H}_n} (m_n - m_0_n) \right| = O_{P_n} \left( d_{\mathcal{H}_n} \sqrt{k_n} \|x_n\|_2 \right). \quad (42)
$$

Using the fact that the Fisher information matrix is positive definite for each $n$,

$$
-\frac{1}{2} \frac{d_{\mathcal{H}_n}}{N_n} x_n^T \Sigma_{\mathcal{H}_n} x_n (1 + o_{P_n}(1)) \leq -\frac{1}{2} \frac{d_{\mathcal{H}_n}}{N_n} N_n \|x_n\|^2 \|x_n\|_2 (1 + o_{P_n}(1)) \leq -\frac{1}{2} O_{P_n} \left( d_{\mathcal{H}_n} \|x_n\|_2 \right) (1 + o_{P_n}(1)). \quad (43)
$$

Combining (42) and (43), by choosing $C$ large enough, for each $\epsilon > 0$, and all $n$ large enough

$$
P \left( \sup_{\|x_n\|_2 = C} \ell_n \left( \theta_0_n + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n \right) - \ell_n(\theta_0_n) < 0 \right) > 1 - \epsilon.
$$

24
Next we consider the difference in the penalty terms between $\hat{\theta}_H^0$ and $\theta_0^H + \sqrt{\frac{d_{H_n}}{N_n}} x_n$. By a first order Taylor expansion,

$$N_n \lambda_n \left( \sum_{h \subseteq H_n} \lambda_h \| \theta_h^0 \|^2_{2} - \sum_{h \subseteq H_n} \lambda_h \| \theta_h^0 \|^2_{2} \right) = N_n \lambda_n \sum_{h \subseteq H_n} \lambda_h \sqrt{\frac{d_{H_n}}{N_n}} (x_n^* \bar{\theta}_h^0)$$

where $x_n^*$ lies between 0 and $x_n$. Using the Cauchy-Stwartz inequality, the absolute value of the last quantity is bounded by

$$\lambda_n \sqrt{N_n d_{H_n}} \| x_n \|_2 \left( \sum_{h \subseteq H_n} \lambda_h \right)$$

Under the above condition, we see that $\| \tilde{\theta}_n - \theta_0^H \|_2 = O_{P_n} \left( \sqrt{\frac{d_{H_n}}{N_n}} \right)$, as required.

Proof of Theorem 4.3. We will deal with Equations (26) and Equation (27) separately.

Proof of Equation (26) It is enough to show

$$\mathbb{P} \left( \left\| N_n \Sigma^{-1}_{H_n} \left( \frac{1}{N_n} U_{H_n}^T (n_n - m_n^0) - \frac{1}{N_n} U_{H_n}^T R_n - \lambda_n \bar{\theta}_n \right) \right\|_2 \leq \alpha_n \right) \rightarrow 1,$$

where $\alpha_n = \min_{h_n \in H_n} \| \theta_h^0 \|_2$. In fact, the former condition implies that the $h_n$-block of the vector inside the norm sign in the previous display is less than $\| \theta_h^0 \|_2$, $\forall h_n \in H_n$, which, by the triangle inequality, will produce the desired result.

First we consider the term $\Sigma^{-1}_{H_n} U_{H_n}^T (n_n - m_n^0)$. The vector $U_{H_n}^T (n_n - m_n^0)$ has mean zero and covariance matrix $\Sigma_{H_n}$. Furthermore, because of NC.1', letting $\gamma_{H_n}^{\min} = \lambda_{min}(\Sigma_{H_n})$, we have

$$\gamma_{H_n}^{\min} \frac{1}{N_n} \geq D_{min} > 0 \quad \text{for all } n.$$ (45)

Combining these observations, and using the formula for the expected value of a quadratic form, we arrive at

$$\mathbb{E} \| \Sigma^{-1}_{H_n} U_{H_n}^T (n_n - m_n^0) \|_2^2 = \text{tr} \Sigma_{H_n}^{-1} \leq \frac{d_{H_n}}{\gamma_{H_n}^{\min}} \leq \frac{d_{H_n}}{D_{min} N_n}$$

where $d_{H_n} = \sum_{h \in H_n} d_h$. Then, Chebyshev inequality implies

$$\left\| \Sigma^{-1}_{H_n} U_{H_n}^T (n_n - m_n^0) \right\|_2 = O_P \left( \frac{d_{H_n}}{N_n} \right).$$ (46)

Next, using (45) for the operator norm of $\Sigma_{H_n}$, we get the upper bound

$$\left\| N_n \Sigma^{-1}_{H_n} \lambda_n \bar{\theta}_n \right\|_2 \leq \frac{1}{D_{min}} \lambda_n \sqrt{\sum_{h_n \in H_n} \lambda_h^2},$$ (47)

which, for $\lambda_{h_n} = \sqrt{d_{h_n}}$, simplifies to $\frac{1}{D_{min}} \lambda_n \sqrt{d_H}$. 

25
Finally, the norm of

\[ \Sigma_{\mathcal{H}_n}^{-1} U_{\mathcal{H}_n}^T R_n \]

is no larger than

\[ \frac{1}{\sqrt{D_{\min}N_n}} \sqrt{d_{\mathcal{H}_n} o(||\hat{\theta}_{\mathcal{H}_n} - \theta^0_{\mathcal{H}_n}||_2)} = oP \left( \frac{d_{\mathcal{H}_n}}{N_n} \right), \tag{48} \]

because \( ||U_{\mathcal{H}_n}||_2 \leq \sqrt{d_{\mathcal{H}_n}}. \)

Using equations (46), (47) and (48), condition (44) is satisfied if MSC.1 holds.

**Proof of Addendum 4.4.** Using the properties of the inverse of a block-matrix, we can write

\[
\Sigma_{\mathcal{H}_n}^{-1} = \begin{bmatrix} A & B \\ B^T & C \end{bmatrix}^{-1} = \begin{bmatrix} (A - BC^{-1}B^T)^{-1} & A^{-1}B(\sigma^2 A^{-1}B - C)^{-1} \\ (B^TA^{-1}B - C)^{-1}B^TA^{-1} & (C - B^TA^{-1}B)^{-1} \end{bmatrix}.
\]

If

\[ ||A^{-1}B||_2 < 1, \tag{49} \]

then, because the positive definiteness, we obtain the bounds (see, e.g., Horn and Johnstone, 1990, for more details)

\[ ||(A - BC^{-1}B^T)^{-1}x||^2 \leq ||A^{-1}x||^2 \]

and

\[ ||A^{-1}B(B^TA^{-1}B - C)^{-1}x||^2 \leq ||A^{-1}B||_2 ||(B^TA^{-1}B - C)^{-1}x||^2 \leq ||C^{-1}x||^2. \tag{51} \]

For any \( h_n \in \mathcal{H}_n \) and with \( h_n^c = \mathcal{H}_n \setminus h_n \), letting \( A = U_{h_n}^T \left( D_m - \frac{mm^T}{N_n} \right) U_{h_n}, B = U_{h_n}^T \left( D_m - \frac{mm^T}{N_n} \right) U_{h_n}^c \), and \( C = U_{h_n}^T \left( D_m - \frac{mm^T}{N_n} \right) U_{h_n}^c \), we may decompose \( \Sigma_{\mathcal{H}_n} \) accordingly. Then, (30) guarantees that (49) is satisfied, for any choice of \( h_n \), and we obtain the bounds (50) and (51). On can proceed recursively by picking any set \( s_n \in h_n^c \) and applying the same arguments to \( U_{h_n}^c \left( D_m - \frac{mm^T}{N_n} \right) U_{h_n}^c \), so that we exhaust all sets in \( \mathcal{H}_n \). In the end, we see that, using (30) repeatedly, the norm any block of vector in left hand side of (44) is bounded by

\[
\frac{1}{|\mathcal{H}_n|} \sum_{h_n \in \mathcal{H}_n} \left\| N_n \left( U_{h_n}^T \left( D_m - \frac{mm^T}{N_n} \right) U_{h_n} \right)^{-1} \left( \frac{1}{N_n} U_{h_n}^T (n - m) - \frac{1}{N_n} U_{h_n}^T R_n - \lambda_n \hat{\eta}_{h_n} \right) \right\|_2,
\]

so that Equation (26) is satisfied if

\[ \mathbb{P} \left( \left\| N_n \left( U_{h_n}^T \left( D_m - \frac{mm^T}{N_n} \right) U_{h_n} \right)^{-1} \left( \frac{1}{N_n} U_{h_n}^T (n - m) - \frac{1}{N_n} U_{h_n}^T R_n - \lambda_n \hat{\eta}_{h_n} \right) \right\|_2 \leq \alpha_n \right) \rightarrow 1, \tag{52} \]

for each \( h_n \in \mathcal{H}_n \).

Noting that assumption NC.1’ also guarantees that the minimal eigenvalue of \( \Sigma_{h_n} \) is bounded by \( D_{\min}N_n \), the same arguments used in the proof of (44) show that (52) is true for each \( h_n \in \mathcal{H}_n \) if both conditions in MSC.1’ are verified.
Proof of Equation (25).

In equation (25) write
\[ \hat{\zeta}_{Hc} = \Lambda_{Hcn} \hat{z}_{Hcn}, \]
where \( \Lambda_{Hcn} \) is a \( d_{Hcn} \)-dimensional diagonal matrix whose diagonal is \( \text{vec}\{1_{wn} \lambda_{wn}, w_n \in H_n^c\} \), with \( 1_{hn} \) denoting the \( d_{hn} \)-dimensional vector with entries all equal to 1. Then, letting
\[ \Sigma_n^0 = Dm_n^0 - \frac{m_n^0 (m_n^0)^\top}{N_n}, \]
(25) becomes
\[ \hat{z}_{Hn}^c = \frac{1}{N_n \lambda_n} \Sigma_n^0 U_{Hn}^\top \left( \frac{1}{N_n \lambda_n} U_{Hn}^\top (n_n - m_n^0) - \frac{1}{N_n \lambda_n} U_{Hn}^\top R_n - \eta_{Hn} \right). \]

For any \( w_n \in H_n^c \), consider the corresponding block in the vector \( \Lambda_{Hn}^{-1} W_n \hat{\eta}_{Hn} \), i.e. the vector
\[ \frac{1}{\lambda_{wn}} U_{wn}^\top \Sigma_n^0 U_{Hn} \left( U_{Hn}^\top \Sigma_n^0 U_{Hn} \right)^{-1} \hat{\eta}_{Hn}. \]

Because of assumption MSC.2, the Euclidian norm of (53), for any choice of \( w_n \in H_n^c \), is bounded by
\[ (1 - \epsilon) \frac{\sum_{h_n \in H_n} \lambda_{hn}}{|H_n^c|} \min_{w_n \in H_n^c} \lambda_{wn}, \]
which, in turn, is smaller that
\[ (1 - \epsilon) \frac{|H_n| \max_{h_n \in H_n} \lambda_{hn}}{|H_n^c| \min_{w_n \in H_n^c} \lambda_{wn}}. \]

Then, under MSC.3 (53) will be eventually less than \((1 - \epsilon)\), uniformly over \( w_n \in H_n^c \).

Next, for each block \( w_n \in H_n^c \), we study the vector
\[ \frac{1}{N_n \lambda_n \lambda_{wn}} \left[ U_{wn}^\top (n_n - m_n^0) - W_{wn} U_{Hn}^\top (n_n - m_n^0) \right]. \]

After some algebra, the covariance matrix of the term inside the parenthesis can be shown to be the matrix
\[ U_{wn}^\top (\Sigma_n^0)^{1/2} \left[ I_{d_{Hn}} - (\Sigma_n^0)^{1/2} U_{Hn}^\top \Sigma_n^0 U_{Hn} \right]^{-1} U_{Hn}^\top (\Sigma_n^0)^{1/2} U_{wn}, \]
whose largest eigenvalue is smaller than \( N_n \lambda_{wn} \). Therefore, by Chebyshev inequality, the term (54) is of order no bigger than
\[ \frac{1}{\lambda_{wn} \lambda_{wn}} O_{P_n} \left( \sqrt{\frac{d_{wn}}{N_n}} \right). \]

Under the condition [MSC.4], uniformly over \( w_n \in H_n^c \), the expression (54) is \( o_{P_n}(1) \).

As for the reminder term it is easy to see that it converges in probability to 0, so that (27) holds true.

\[ \blacksquare \]
**Proof of Theorem 4.5.** All the claims in the proof are made inside the event $O_n$. Because the norm consistency assumptions are in force, $O_n$ occurs in probability and, therefore, our claims hold true within a set or probability converging to 1. In particular, $\|\hat{\theta}_{\mathcal{H}_n} - \theta^0_{\mathcal{H}_n}\|_2 = O_P\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right)$ (1 + $o_p(1)$) = $O_P\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right)$. Reorganize equation (24) as

$$\Sigma_{\mathcal{H}_n}^{1/2} (\hat{\theta}_{\mathcal{H}_n} - \theta^0_{\mathcal{H}_n}) = \Sigma_{\mathcal{H}_n}^{1/2} U_{\mathcal{H}_n}^T (n_n - m^0_n) - \Sigma_{\mathcal{H}_n}^{1/2} N_n \lambda_n \eta_{\mathcal{H}_n} - \Sigma_{\mathcal{H}_n}^{1/2} U_{\mathcal{H}_n}^T R_n$$

(55)

By similar arguments used in the proof of Theorem 4.3, the term

$$\Sigma_{\mathcal{H}_n}^{-1/2} U_{\mathcal{H}_n}^T R_n$$

is of order

$$\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} O_P\left(\|\tilde{\theta}_n - \theta^0_n\|_2\right) = o_P\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right),$$

and therefore converges in probability to 0.

As for $\Sigma_{\mathcal{H}_n}^{-1/2} N_n \lambda_n \eta_{\mathcal{H}_n}$, notice that, on $O_n$, the vector $\tilde{\eta}_{\mathcal{H}_n}$ is a differentiable function of $\hat{\theta}_{\mathcal{H}_n} \in \mathbb{R}^{d_{\mathcal{H}_n}}$. Then, using a Taylor expansion around $\theta^0_{\mathcal{H}_n}$,

$$\Sigma_{\mathcal{H}_n}^{-1/2} N_n \lambda_n \eta_{\mathcal{H}_n} = \Sigma_{\mathcal{H}_n}^{-1/2} N_n \lambda_n \left(\tilde{\eta}_{\mathcal{H}_n} + J_{\mathcal{H}_n} (\hat{\theta}_{\mathcal{H}_n} - \theta^0_{\mathcal{H}_n}) + o_P\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right)\right).$$

(56)

The remainder term in Equation (56) is of order

$$\lambda_n o_P(\sqrt{d_{\mathcal{H}_n}}),$$

which become negligible for $\lambda_n = O\left(\frac{1}{\sqrt{d_{\mathcal{H}_n}}}\right)$ (obviously, $\lambda_n = O\left(1/\sqrt{T_n}\right)$ will do). Then using (55), we obtain

$$\Sigma_{\mathcal{H}_n}^{-1/2} U_{\mathcal{H}_n}^T (n_n - m^0_n) = \Sigma_{\mathcal{H}_n}^{-1/2} \left((\Sigma_{\mathcal{H}_n} + N_n \lambda_n J_{\mathcal{H}_n}) (\hat{\theta}_n - \theta^0_n) + N_n \lambda_n \eta^0_{\mathcal{H}_n}\right) + o_P(1).$$

(57)

Thus, we only need to consider the term $\Sigma_{\mathcal{H}_n}^{-1/2} U_{\mathcal{H}_n}^T (n_n - m^0_n)$. For $1 \leq j_n \leq N_n$, let

$$Y_{j_n} = \frac{1}{\sqrt{N_n}} F_{n}^{-1/2} U_{\mathcal{H}_n} (X_{j_n} - \pi^0_{n}),$$

where the variables $X_{j_n}$ are iid Multinomials with size 1 and probability vector $\pi^0_{n}$. Then,

$$\Sigma_{\mathcal{H}_n}^{-1/2} U_{\mathcal{H}_n}^T (n_n - m^0_n) = \sum_{j_n} Y_{j_n},$$

where $\mathbb{E} Y_{j_n} = 0$, $\text{Cov} Y_{j_n} = \frac{1}{N_n} I_{d_{\mathcal{H}_n}}$ and $\sum_{j_n} \text{Cov}(Y_{j_n}) = I_{d_{\mathcal{H}_n}}$.

The result for part 1. under assumption [CLT] is based on standard arguments and entails checking the multivariate Lindberg-Feller conditions. We omit this proof because it is almost identical to the proof we produce below for the [CLT.LF] conditions, the only difference being a rate $O(N_n^{-1})$ in equation (66). See also the proof of Theorem 2 in Fan and Peng (2004).
Part 2. of the theorem follows in a straightforward way from the main theorem in Bentkus (2003) and the fact that \( \mathbb{E}\|F_n^{-1/2}U^n(\mathbb{H}_n, \mathbb{X}_n - \pi_n^0)\|^2 \) is of order \( O\left(a_n^{3/2}\right) \), by the same arguments used in the proof of part 3., given below.

Next we prove the result of part 1. under both [CLT.Ma] and [CLT.Mb]. We relax the assumption [CLT] by allowing the dimension of the parameter space to grow faster. To this end, we derive multi-dimensional analogs of Lemma 2.1 and 2.2 and Theorem 2.1 in Morris (1975). In particular, our proof follows closely the proof of Morris (1975, Lemma 2.2). We first obtain joint limit law by using Lemma 9.1, and then establish the conditional limit law by using a multi-dimensional version of condition (2.9) in Morris (1975). Note that the result in Steck (1957) about conditional limit laws is actually a multi-dimensional one, but somehow was formulated in Morris (1975, Th. 2.1.) as one-dimensional. The conditional law we are interested is the distribution of \( Z_n \), defined below in (60).

Let \( \gamma_n = N_n^{-1}C_n\Sigma_{\mathbb{H}_n}^{-1/2}U^n \mathbf{m}_n^0 \), and set \( A_n = C_n\Sigma_{\mathbb{H}_n}^{-1/2}U^n \). Note that \( \mathbf{m}_n^0 = N_n\pi_n^0 \), thus \( \gamma_n = A_n\pi_n^0 \). Denote the \( i \)-th column of \( A_n \) by \( a_i, i = 1, \ldots, I_n \). Then, the left hand side of (57), pre-multiplied by \( C_n \), can be written as

\[
Z_n = \sum_{i \in I_n} f_i(n_i),
\]

where \( f_i(n_i) = (a_i - \gamma_n)(n_i - m_i^0) \). Let \( \{X_i; i = 1, \ldots, I_n\} \) be independent Poisson random variables with mean \( m_i^0 = N_n\pi_i^0 \), so that \( \mathbb{E}f_i(X_i) = 0 \) and \( \sum_i \text{cov}(f_i(X_i), X_i) = 0 \), by construction. Next, define

\[
V_n = N_n^{-1/2} \sum_i (X_i - m_i^0) \quad \text{(58)}
\]
\[
U_n = \Xi_n^{-1/2} \sum_i f_i(X_i) \quad \text{(59)}
\]

where \( \Xi_n = \sum_i \text{cov}(f_i(X_i)) \). A simple calculation though shows that \( \Xi_n = C_nC_n^\top \), a square matrix of fixed dimensions \( k \times k \). The goal is to prove the asymptotic normality of \( U_n \) given \( \{V_n = 0\} \), and then use the fact (underlying Morris’ method) that

\[
\mathcal{L}(\Xi_n^{-1/2}Z_n) = \mathcal{L}(U_n|V_n = 0), \quad \text{(60)}
\]

where \( \mathcal{L} \) stands for law.

The random variables \( V_n \) have zero means and unit variances. Furthermore, by the same arguments used in the early parts of (Morris, 1975, Lemma 2.2), assumption [CLT.Mb] guarantees that the uan condition is satisfied, so the sequence \( V_n \) converge in distribution to a Gaussian variable. Similarly, the random vector \( U_n \) satisfies \( \mathbb{E}U_n = 0 \), \( \text{cov}(U_n) = I_k \), the identity matrix of dimensions \( k \times k \), and, by construction, \( \text{cov}(V_n, U_n) = 0 \). We argue below that \( U_n \) satisfies the multi-dimensional Lindeberg condition. By Lemma (9.1), this will imply the asymptotic normality of the joint limit law of \( (V_n, U_n) \).

By Schwartz inequality, for any \( \epsilon > 0 \),

\[
\sum_i \mathbb{E}\left[ \|f_i(X_i)\|^2; \|f_i(X_i)\|_2 > \epsilon \right] \leq \sum_i \left[ \mathbb{E}\|f_i(X_i)\|^2 \mathbb{P} (\|f_i(X_i)\|_2 > \epsilon) \right]^{1/2}. \quad \text{(61)}
\]
We will show that, for each $\epsilon > 0$, the right hand side of (61) tends to zero. Recall that $f_i(X_i) = (a_i - \gamma_n)(X_i - m_i^0)$. The length of $\gamma_n$ can be bounded as follows, $\|\gamma_n\|_2 \leq \||C_n\|_2\Sigma_{H_n}^{-1/2}U_{H_n}^\top \pi_n^0\|_2 \leq O(1)D_{\min}N_n^{-1/2}\|U_{H_n}^\top \pi_n^0\|_2$. Elements of $U_{H_n}^\top \pi_n^0$ are absolutely bounded by a constant $D_1$, thus $\|\gamma_n\|_2 \leq DN_n^{-1/2}d_{H_n}^{-1/2}$. Similarly, $\|a_i\|_2 = \|A_ne_i\|_2 \leq DN_n^{-1/2}d_{H_n}^{-1/2}$, where $e_i$ is the standard unit vector in $\mathbb{R}^I_n$ with $i$-th coordinate equal to 1. Adding up, $\|a_i - \gamma_n\|_2 = O\left(N_n^{-1/2}d_{H_n}^{-1/2}\right)$, which tends to zero by assumption [CLT.Ma].

Next, we use the following large deviation result for Poisson random variables, due to Bobkov and Ledoux (1998) and based on a modified logarithmic Sobolev inequality:

**Theorem 6.1.** Let $X$ be a poisson random variable with parameter $\lambda$. Then, for every $h : \mathbb{N} \to \mathbb{R}$, with $\sup_{x \in \mathbb{N}} |h(x + 1) - h(x)| \leq 1$,

$$\mathbb{P}\left(h(X) - \mathbb{E}h(X) \geq b\right) \leq \exp\left\{-\frac{b}{4} \log \left(1 + \frac{b}{2\lambda}\right)\right\}, \quad (62)$$

for all $b \geq 0$.

Then, using Theorem 62, for some constant $D$,

$$\mathbb{P}\left(X_i - m_i^0 \geq \epsilon \|a_i - \gamma_n\|_2^{-1}\right) \leq \exp\left\{-\frac{\epsilon}{4}\|a_i - \gamma_n\|_2^{-1} \log \left(1 + \frac{1}{2\epsilon} \|a_i - \gamma_n\|_2\right)\right\}$$

$$\leq \exp\left\{-\epsilon DN_n^{-1/2}d_{H_n}^{-1/2} \log \left(1 + \frac{1}{\epsilon D} \frac{1}{\sqrt{N_n}d_{H_n} \max_i \pi_i}\right)\right\}$$

$$= \exp\left(-\frac{b}{4} \log \left(1 + \frac{b}{2\lambda}\right)\right), \quad (63)$$

as $n \to \infty$. The last inequality follows by condition CLT.Mb. The same result may be achieved by applying a modified logarithmic Sobolev inequality to the left tail.

Finally, $\sum_i(\mathbb{E}\|f_i(X_i)\|^2)^{1/2} = \sum_i \|a_i - \gamma_n\|_2^2(m_i^0 + 3(m_i^0)^2)^{1/2}$, which is of the order of magnitude of $O\left(d_{H_n}\right)$. This, together with (61) and (63) and assumption CLT.Ma, shows that $U_n$ satisfies the Lindeberg condition, as stated.

We turn now to consider the conditional limit law. As mentioned above, Theorem 2.1. in Morris (1975) holds true also for multi-dimensional variables. We only need to replace condition (2.9) in Morris (1975) by a multi-dimensional version. Specifically, we show that

$$\lim_{r \to 0} \sup_n \sup_v \mathbb{E}\left\|\sum_i [f_i(L_i + M_i) - f_i(L_i)]\right\|_2^2 = 0, \quad (64)$$

where $L_n = (L_1, \ldots, L_n)$ and $M_n = (M_1, \ldots, M_n)$ are Multinomial random variables with probability vector $\pi_n^0$, and sample sizes $N_n + v_nN_n^{1/2}$ and $rN_n^{1/2}$, respectively, where the parameters $v_n = O(1)$ and $r$ are specified as in Morris (1975, Lemma 2.2). Notice that $f_i(L_i + M_i) - f_i(L_i) = (a_i - \gamma_n)M_i$. Thus,

$$\left\|\sum_i (a_i - \gamma_n)M_i\right\|_2^2 = \left\|A_nM_n - rN_n^{-1/2}A_n\pi_n^0\right\|_2^2 = (M_n - \mathbb{E}M_n)^\top B_n(M_n - \mathbb{E}M_n),$$

and
where $\mathbb{E}M_n = r N_n^{1/2} \pi^0$, and $B_n = A_n^\top A_n$. Taking expectation yields
\[
\mathbb{E}(M_n - \mathbb{E}M_n)^\top B_n (M_n - \mathbb{E}M_n) = \frac{1}{\sqrt{N_n}} \mathbf{tr} \left( B_n \left( D \pi^0_n - \pi^0_n (\pi^0_n)^\top \right) \right) \\
= \frac{1}{\sqrt{N_n}} \mathbf{tr} \left( B_n \left( D m^0_n - \frac{m^0_n (m^0_n)^\top}{N_n} \right) \right) \\
= O(1) r \frac{1}{\sqrt{N_n}},
\]

since
\[
\mathbf{tr} \left( B_n \left( D m^0_n - \frac{m^0_n (m^0_n)^\top}{N_n} \right) \right) = \mathbf{tr}(C_n C_n^\top) = O(1) .
\]

Therefore,
\[
\mathbb{E} \left\| \sum_i [f_i(L_i + M_i) - f_i(L_i)] \right\|^2_2 = O(1) r \frac{1}{\sqrt{N_n}} \to 0,
\]

which shows that condition (64) holds, and the statement in part 1. is proved.

Finally, we prove part 3. of the theorem statement, by showing that the Lindberg-Feller conditions of Lemma 8.1 are satisfied. Under assumption [CLT.LF] we only need to show that
\[
\sum_{j_n} \mathbb{E}_{\mathcal{H}_n} \|Y_{j_n}\|_2^4 P(\|Y_{j_n}\|_2 \geq \epsilon) \to 0
\]
as $n \to \infty$. Using Cauchy Stwartz inequality and the fact that the $Y_{j_n}$ are identically distributed, it is sufficient to show that
\[
N_n (\mathbb{E}\|Y_{j_n}\|_2^4 P(\|Y_{j_n}\|_2 \geq \epsilon))^{1/2} \to 0. \tag{65}
\]

By Chebychev inequality, for each $\epsilon > 0$,
\[
P(\|Y_{j_n}\|_2 \geq \epsilon) \leq \frac{\mathbf{tr}(I_{\mathcal{H}_n})}{\epsilon^2 N_n} = O \left( \frac{d_{\mathcal{H}_n}}{N_n} \right) . \tag{66}
\]

Next, using the the assumption on the minimal eigenvalue of $F_n$,
\[
\mathbb{E}\|Y_{j_n}\|_2^4 \leq O \left( \frac{1}{N_n^2} \right) \mathbb{E}\|U_{\mathcal{H}_n}^\top (X_{j_n} - \pi^0_n)\|_2^4 = O \left( \frac{d_{\mathcal{H}_n}^2}{N_n^2} \right) ,
\]

where in the last step we use the fact that the entries of $U_{\mathcal{H}_n}^\top (X_{j_n} - \pi^0_n)$ are bounded, uniformly over $n$.

Combining the last two displays, the left hand side of (65) is of order
\[
N_n O \left( \sqrt{\frac{d_{\mathcal{H}_n}}{N_n} \frac{d_{\mathcal{H}_n}}{N_n}} \right) = O \left( \frac{d_{\mathcal{H}_n}^{3/2}}{N_n^{1/2}} \right) ,
\]

which, in virtue of assumption [CLT.LF], vanishes, as desired.

\[\blacksquare\]
7 Appendix A: Bases for $\mathcal{U}_h$

Given a log-linear model $\mathcal{H}$, bases for the subspaces $\mathcal{U}_h = \mathcal{M}_h$, with $h \in \mathcal{H}$ will be defined and computed. The term contrast bases is appropriate because they indeed correspond to contrasts in models of analysis of variance. Using Birch’s notation (see, in particular, Bishop et al., 1975), the design matrix for $\mathcal{U}_h$ will encode to the $u$-terms corresponding to the $|h|$-order interactions among the factors in $h$.

For each term $h \subseteq \mathcal{K}$ and factor $k \in \mathcal{K}$, define the matrix

$$
U^h_k = \begin{cases} 
Z_k & \text{if } k \in h \\
1_k & \text{if } k \notin h,
\end{cases}
$$

where $Z_k$ is a $I_k \times (I_k - 1)$ matrix with entries

$$
Z_k = \begin{pmatrix}
1 & 0 & 0 & 0 & \ldots & 0 \\
-1 & 1 & 0 & 0 & \ldots & 0 \\
0 & -1 & 1 & 0 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \ldots & 1 \\
0 & 0 & 0 & 0 & \ldots & -1
\end{pmatrix},
$$

(67)

and $1_k$ is the $I_k$-dimensional column vector of 1’s. Let

$$
U_h = \bigotimes_{k=1}^{\mathcal{K}} U^h_k.
$$

(68)

Since the elements of $U^h_k$ are $-1, 0$ and $1$, $U_h$ has entries that can only be $-1, 0$ and $1$. Additional properties of the design matrices $U_h$ and hence of the subspaces spanned by their columns are given in the next Lemma. The results in Proposition 7 follow immediately.

**Lemma 7.1.**

i. For every $h, h' \in 2^{\mathcal{K}}$, with $h \neq h'$, the columns of $U_h$ are linearly independent and $U_h^T U_{h'} = 0$;

ii. $\mathbb{R}^\mathcal{X} = \bigoplus_{h \in 2^{\mathcal{K}}} \mathcal{R}(U_h)$;

iii. for any $h \in 2^{\mathcal{K}}$, $\mathcal{R}(U_h) = \mathcal{U}_h$, where $\mathcal{U}_h$ is the subspace of interactions for the factors in $h$.

**Proof.** Part i.: the first statement follows from the fact that the columns of $Z^h_k$ are independent for each $k$ and $h$ and $U_h$ has dimension $\left(\prod_{k=1}^{\mathcal{K}} I_k\right)$. As for the second statement, without loss of generality, we can assume that there exists a factor $k$ such that $k \in h$ and $k \notin h'$. Then, $Z^h_k = C_k$ and $Z^h_{k'} = 1_k$, so $(Z^h_k)^T Z^h_{k'} = 0$, hence the result.

Part ii.: by i., the subspaces $\mathcal{R}(U_h)$, are orthogonal (hence the direct sum notation is well defined) and $\dim \mathcal{R}(U_h) = \prod_{j \in h} (I_j - 1)$. Therefore:

$$
\dim \left( \bigoplus_{h \in 2^{\mathcal{K}}} \mathcal{R}(U_h) \right) = \sum_{h \in 2^{\mathcal{K}}} \prod_{j \in h} (I_j - 1) = \prod_{k \in \mathcal{K}} I_k,
$$

32
where the last equality follows from the Möbius inversion formula (see, e.g., Lemma 3.13 in Rinaldo, 2006b) and the fact that, for \( h = \emptyset \), \( \dim(U_h) = 1 \).

Part iii. The matrix

\[
S_h = \bigotimes_{k \in K} S^h_k
\]

where

\[
S^h_k = \begin{cases} 
I_k - \frac{1}{T_j} I_{j_k} & \text{if } k \in h \\
\frac{1}{T_j} I_{j_k} & \text{otherwise}.
\end{cases}
\]

is the projector onto \( \mathcal{U}_h \) (see Corollary 3.10 Rinaldo, 2006b). Therefore it suffices to show \( S_h U_h = U_h \) and \( S_h U_{h'} = 0 \) for \( h \neq h' \). This implies \( \mathcal{R}(U_h) \subseteq \mathcal{U}_h \) and the results follow from the fact that the inclusion cannot be strict because of the orthogonal decompositions of \( \mathbb{R}^T \) as in ii. and Equation (6). It is easy to see that \( S_k^h U_k^h = U_k^h \) and, for any \( h \neq h' \) with \( k' \in h' \setminus h \), \( S_k^h U_{k'}^h = S_k^h U_{k'}^h = 0 \). Therefore,

\[
S_h U_h = \bigotimes_{k=1}^K S^h_k U_k^h = U_h \quad \text{and} \quad S_h U_{h'} = \bigotimes_{k=1}^K S^h_k U_{k'}^h = 0
\]

for any \( h \neq h' \).

\[\square\]

8 Appendix B

Lemma 8.1. Let \( X_{j_n} \), \( j_n = 1, \ldots, N_n \), be i.i.d. vectors in \( \mathbb{R}^{k_n} \) with \( \mathbb{E}_n X_{j_n} = 0 \) and \( \sum_{j_n} \text{cov}_n X_{j_n} = I_{k_n} \). Assume that \( k_n \to \infty \) and \( N_n \to \infty \) in such a way that

\[
\lim_n \frac{1}{N_n} \sum_{j_n} \mathbb{E}_n \|X_{j_n}\|^2 I\{\|X_{j_n}\| > \epsilon \sqrt{N_n}\} = 0.
\]

(69)

Let \( \psi_n \) be the characteristic function of \( \frac{1}{\sqrt{N_n}} \sum_{j_n} X_{j_n} \) and \( \phi_n \) the characteristic function of a \( k_n \)-dimensional standard Gaussian distribution. Then, for each \( \epsilon > 0 \) and \( T > 0 \), there exists a \( n^0(\epsilon, T) \) such that, for all \( n > n^0(\epsilon, T) \),

\[
\sup \left\{ |\psi_n(t) - \phi_n(t)| : \|t\|_2 \leq T \right\} < \epsilon.
\]

(70)

Proof. The result follows from reduction to the one-dimensional case by the Cramer-Wold device. Consider an arbitrary sequence \( \{t_n\} \in \otimes_n \mathbb{R}^{k_n} \). Let \( W_{j_n} = t_n^T X_{j_n} \) and set

\[
s^2_n = \sum_{j_n} \text{Var}_n W_{j_n} = N_n \|t_n\|^2_2,
\]

so that \( \frac{1}{s_n} \sum_{j_n} W_{j_n} \) has mean 0 and unit variance. Notice that the Lindberg-Feller condition holds for the sequence of random variables \( \frac{1}{s_n} \sum_{j_n} W_{j_n} \) since, for each \( \epsilon > 0 \),

\[
\lim_n \frac{1}{s^2_n} \sum_{j_n} \mathbb{E}_n W_{j_n}^2 I\{\|W_{j_n}\| > \epsilon s_n\} = \lim_n \frac{1}{N_n \|t_n\|^2_2} \mathbb{E}_n \left(t_n^T X_{j_n}\right)^2 I\{|t_n^T X_{j_n}| > \epsilon \sqrt{N_n} \|t_n\|_2\}
\]

\[
\leq \frac{1}{N_n} \sum_{j_n} \mathbb{E}_n \|X_{j_n}\|^2_2 I\{\|X_{j_n}\|_2 > \epsilon \sqrt{N_n}\} = 0,
\]

(71)
by (69), the Cauchy-Swartz inequality and the fact
\[ \{ |t_n^T X_{j_n}| > \epsilon \sqrt{N_n} \|t_n\|_2 \} \subseteq \{ \|X_{j_n}\|_2 > \epsilon \sqrt{N_n} \}. \]

Notice that (71) does not depend on the sequence \( \{t_n\} \) and the value at 1 of the characteristic function of \( \frac{1}{s_n} \sum_{j_n} W_{j_n} \) converges to \( \exp^{-1/2} \), uniformly over all sequences \( \{t_n\} \). From this, it follows that
\[ \sup \{ |\psi_n(u_n) - \phi_n(u_n)| : \|u_n\| = 1 \} \rightarrow 0. \]

The previous uniform convergence holds also for sequences \( \{t_n\} \) such that \( \|t_n\|_2 = T \) for each \( n \) and any \( T > 0 \), a fact that can be easily established once again from the Lindberg-Feller conditions for the one dimensional case in the chain of inequalities leading to (71). Formally, for any \( \epsilon > 0 \) and \( T > 0 \) there exists a \( n_0(\epsilon, T) \) such that for each \( n > n_0(\epsilon, T) \), we have \( \sup \{ |\psi_n(t) - \phi_n(t)| : \|t\| = T \} < \epsilon \). As \( n_0(\epsilon, T) \) is non-decreasing in \( T \) for fixed \( \epsilon \), the proof of (70) is complete.

\[ \blacksquare \]

9 Appendix C

The following Lemma is a multivariate analog of Lemma 2.1. in Morris (1975).

**Lemma 9.1.** Let \( S = (S_{ik}, R_k) = \sum_{i=1}^k X_{ik} \), where \( R_k = (S_{2k}, \ldots, S_{pk}) \), \( X_{ik} = (X_{i1k}, Y_{ik}) \), and \( Y_{ik} = (X_{i2k}, \ldots, X_{ipk}) \). Suppose that \( \{X_{ik}\}_{i=1}^k \) are independent random vectors, with \( \mathbb{E}X_{i1k} = 0, \mathbb{E}Y_{ik} = 0 \), and \( \text{Var}(S_k) = I_p \), the \( p \times p \) identity matrix. Suppose \( S_{1k} \) satisfies the lnc condition, i.e., \( \max_{1 \leq i \leq k} \text{Var}X_{i1k} = o(1) \) as \( k \rightarrow \infty \), and that \( S_{1k} \xrightarrow{w} N(0, I) \). Finally, suppose that \( R_k \) satisfies the (multi-dimensional) Lindeberg condition, i.e., for all \( \epsilon > 0 \),
\[ \sum_{i=1}^k \mathbb{E} \left[ \|Y_{ik}\|^2 ; \|Y_{ik}\|^2 > \epsilon \right] = o(1) \quad \text{,} \quad (k \rightarrow \infty). \]

Then \( S_k \xrightarrow{w} N_p(0, I_p) \).

**Proof.** As in Morris’ proof, \( S_{1k} \) satisfies the (one-dimensional) Lindeberg condition, i.e.,
\[ \sum_{i=1}^k \mathbb{E} \left[ X_{i1k}^2 ; X_{i1k}^2 > \epsilon \right] = o(1) \quad \text{,} \quad (k \rightarrow \infty). \]

Therefore,
\[ \sum_{i=1}^k \mathbb{E} \left[ \|X_{ik}\|^2 ; \|X_{ik}\|^2 > \epsilon \right] = \]
\[ \sum_{i=1}^k \mathbb{E} \left[ X_{i1k}^2 + \|Y_{ik}\|^2 ; X_{i1k}^2 + \|Y_{ik}\|^2 > \epsilon \right] \]
\[ \leq 2 \sum_{i=1}^k \mathbb{E} \left[ \max\{X_{i1k}^2, \|Y_{ik}\|^2\} ; \max\{X_{i1k}^2, \|Y_{ik}\|^2\} > \epsilon/2 \right] \]
\[ \leq 2 \sum_{i=1}^k \mathbb{E} \left[ X_{i1k}^2 ; X_{i1k}^2 > \epsilon/2 \right] + 2 \sum_{i=1}^k \mathbb{E} \left[ \|Y_{ik}\|^2 ; \|Y_{ik}\|^2 > \epsilon/2 \right] = o(1) \]
Thus, $S_k$ satisfies the (multi-dimensional) Lindeberg condition and the proof is complete (see, e.g., Bhattacharya and Rao, 1976, pp. 183-184).

10 Acknowledgments

This research was supported in part by NSF grant EIA0131884 to the National Institute of Statistical Sciences, by NSF grant DMS-0631589, Army contract DAAD19-02-1-3-0389 and a grant from the Pennsylvania Department of Health through the Commonwealth Universal Research Enhancement Program, all to Carnegie Mellon University. We thank Larry Wasserman for his valuable comments.

References

Bentkus, V. (2003). On the dependence of the Berry–Esseen bound on dimension, *Journal of Statistical Planning and Inference*, 113, 385–402.

Bhattacharya, R. N. and Ranga Rao, R. (1976). *Normal approximation and asymptotic expansions*, John Wiley & Sons.

Bickel, P.J., Klaassen, C.A.K, Ritov, Y. and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, CMS Books in Mathematics, Springer.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*, MIT Press, Cambridge, Massachusetts.

Bobkov, S. G. and Ledoux, M. (1998). On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures, *Journal of Functional Analysis*, 2, 347–365.

Brown, L.D. (1986). *Fundamentals of Statistical Exponential Families*, IMS Lecture Notes Monograph Series, Vol.9. Institute of Mathematical Statistics.

Cressie, N. A. C. and T. R. C. Read (1988). *Goodness-of-Fit Statistics for Discrete Multivariate Data*, Springer-Verlag.

Dahinden, C., Parmiggiani, G., Emerick, M.C. and Bühlmann, P. (2006). Sparse Contingency Tables and High-Dimensional Log-Linear Models for Alternative Splicing in Full-Length cDNA Libraries, Research Report 132, Swiss Federal Institute of Technology.

Darroch, J.N. and Speed, T. (1980). “Additive and multiplicative models and interactions,” *Annals of Statistics*, 8, 522–539.

Fan J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters, *Annals of Statistics*, 32(3), 928–961.

Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity, *Journal of Multivariate Analysis*, 74(1), 49–78.

Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional predictor selection and the virtue of overparametrization, *Bernoulli*, 10, 971–988.
Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under $\ell_1$ constraint, *Annals of Statistics*, 34(5), 2367–2386.

Horn, R.A. and Johnstone, C.R. (1990). *Matrix Analysis*, Cambridge University Press.

Lauritzen, S.L. (1996). *Graphical Models*, Oxford University Press.

Lehman, E. L. and Casella, G. (2005). *Theory of Point Estimation*, Springer.

Lehman, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, Springer.

Meier, L., van der Geer, S. and Bühlmann, P. (2006). The Group Lasso for Logistic Regression, Research Report 131, Swiss Federal Institute of Technology, Zurich.

Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso, *Annals of Statistics*, 34(3), 1436–1462.

Morris, C. (1975). Central Limit Theorems for Multinomial Sums, *Annals of Statistics*, 1(3), 165–188.

Pisier, G. (1999). *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press.

Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity, *Annals of Statistics*, 16(1), 356–366.

Portnoy, S. (1988). On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$, *Probability Theory and Related Fields*, 71, 571–583.

Quine, M. P. and Robinson, J. (1984). Normal Approximations to Sums of Scores Based on Occupancy Numbers, *Annals of Statistics*, 12(3), 794–804.

Rinaldo (2006). “On maximum likelihood estimation for log-linear models,” submitted for publication.

Rinaldo (2006). “Computing Maximum Likelihood Estimates in Log-Linear Models,” Technical report, Department of Statistics, Carnegie Mellon University.

Schervish, M.J. (1998). *Theory of Statistics*, Springer.

Senatov, V.V. (1998). *Normal Approximation: New Results, Methods and Problems*, Brill Academic Publishers.

Steck, G.P. (1957). Limit theorems for conditional distributions, *University California Publication in Statistics*, 2, 237–284.

Tropp, J.A. (2005). Algorithms for Simultaneous Sparse Approximation Part II: Convex Relaxation. Technical Report.

van de Geer, S. A. (2006). High-dimensional generalized linear models and the Lasso, Research Report 133, Swiss Federal Institute of Technology, Zurich.
van de Geer, S. A. (2006). On non-asymptotic bounds for estimation in generalized linear models with highly correlated design, Research Report 134, Swiss Federal Institute of Technology, Zurich.

van der Vaart, A.W. and Wellner, J.A. (1998). *Weak Convergence and Empirical Processes*, Springer.

Wainwright, M. J. (2006). “Sharp thresholds for high-dimensional and noisy recovery of sparsity, Technical Report 708, Department of Statistics, UC Berkeley.

Wainwright, M., Ravikumar P. and Lafferty, J. (2006). High dimensional graphical model selection using L1-regularized logistic regression, NIPS 2006.

Yuan, M. and Lin Y. (2006). Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society*, B, 68(1), 49–67.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of Lasso, *Journal of Machine Learning Research*, 7, 2541–2563.