Bernoulli **18**(3), 2012, 945–974
DOI: 10.3150/11-BEJ364

The log-linear group-lasso estimator and its asymptotic properties

YUVAL NARDI[1] and ALESSANDRO RINALDO[2]

[1]*Faculty of Industrial Engineering and Management, Technion–Israel Institute of Technology, Haifa 32000, Israel. E-mail: ynardi@ie.technion.ac.il*
[2]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213-3890, USA. E-mail: arinaldo@stat.cmu.edu*

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 solution of a convex penalized likelihood optimization problem based on the group-lasso penalty. We illustrate how it is possible to construct an estimator of the underlying log-linear model using the blocks of nonzero coefficients recovered by the group-lasso procedure. We investigate the asymptotic properties of the group-lasso estimator as a model selection method 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 model selection consistent, in the sense that, with overwhelming probability as the sample size increases, it correctly identifies all the sets of nonzero 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.

Keywords: consistency; group lasso; log-linear models; model selection.

1. Introduction

The theory of log-linear models has produced a variety of statistical methodologies and theoretical results for the analysis of categorical data that have found applications in numerous scientific areas, ranging from social and biological sciences, to medicine, disclosure limitation problems, data-mining, image analysis, finger-printing, language processing and genetics.

Inherently, log-linear modeling is a model selection procedure for contingency tables that encompasses testing a number of statistical models for the joint distribution of a set 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, that is, the case in which the sample size N is much larger than the number p of candidate parameters. It is applicable to a variety of goodness-of-fit measures, such as Pearson's χ^2, the likelihood ratio statistic and, more generally, any statistics belonging to the power-divergence family of Read and Cressie [29]. The applicability and validity of these methods demand the availability of large sample sizes and the existence of the maximum likelihood estimate (MLE).

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 inadequate. First off, the number of log-linear models grow extremely fast with the number of variables (for example, there are 7580 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, under a sparse scenario, the MLE is unlikely to exist. This implies that the information content present in the data is not sufficient to estimate all the parameters of the model, and, therefore, the possibility for inference is only limited to portions of the parameter space (see Rinaldo, Fienberg and Zhou [31] for details). As a result, traditional goodness-of-fit testing and model selection will produce very poor, if not completely erroneous, asymptotic approximations. It is quite clear that a more appropriate statistical formalization requires the consideration of a 'large p' setting.

In this article, we study 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 ℓ_1-regularization is possible, even if the number of parameters grows faster than the sample size. See, in particular, Meinshausen and Bühlmann [21], Zhao and Yu [41], Wainwright [37] and, for a different approach, Greenshtein [15] and Greenshtein and Ritov [16]. Existing work on penalized likelihood problems involving ℓ_1-regularization for discrete data include the nonasymptotic results about ℓ_2 consistency for estimation in high-dimensional generalized linear models via the lasso by van de Geer [35,36], and the analysis by Wainwright, Ravikumar and Lafferty [38] on the consistency of ℓ_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 maximum likelihood problem based on the group-lasso, a convex penalty function introduced by Yuan and Lin [40] in a nonasymptotic ANOVA setting and further analyzed by Nardi and Rinaldo [23]. The group-lasso regularization is an extension of the lasso, or ℓ_1, penalty function designed to penalized groups of coefficients simultaneously. It has been shown to be effective in logistic regression problems by Meier, van der Geer, and Bühlmann [20] and has been used in applications involving log-linear modeling of sparse contingency tables in Dahinden et al. [7].

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 arguments 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. Next, 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 nonzero blocks of the group-lasso estimates. Section 4 outlines our contribution by studying the consistency properties of the group-lasso estimator as a model selection procedure. We formulate a general double-asymptotic framework in which we allow both the sample size and the model complexity to grow. In Section 4.1, we derive conditions guaranteeing that the model estimates are consistent, that is, asymptotically, the group-lasso correctly identifies the set of interactions making up the underlying model. We

conclude our analysis with some central limit results in Section 4.2. The proofs appear in Section 6.

1.1. Notation

Let $X_1, \ldots, X_K$ denote K categorical variables, where each X_k takes values in $\mathcal{I}_k = \{1, \ldots, I_k\}$, with $I_k \geq 2$ an integer. Set $\mathcal{I} = \mathcal{I}_1 \times \cdots \times \mathcal{I}_K$. We denote by $\mathbb{R}^{\mathcal{I}}$ the class of real-valued functions on $\mathcal{I}$ which is the vector space of real-valued K-dimensional arrays indexed by the multi-index $\mathcal{I}$. The vector space $\mathbb{R}^{\mathcal{I}}$ can be naturally represented as a Euclidean space of dimension $I \equiv \prod_{k=1}^{K} I_k$. This identification can be realized in a straightforward fashion by ordering $\mathcal{I}$ as a linear list using any bi-jection between $\mathcal{I}$ and $\{1, 2, \ldots, I\}$. Each element i of $\mathcal{I}$, called a *cell*, is a multi-index $i = (i_1, \ldots, i_K)$. Using this coordinate vector representation, for any array $\mathbf{x} \in \mathbb{R}^{\mathcal{I}}$, $\mathbf{x}_i$ is the number indexed by the coordinate $i \in \mathcal{I}$. Also, the standard inner product $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i \in \mathcal{I}} \mathbf{x}_i \mathbf{y}_i$ and the induced Euclidean norm are well defined for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{\mathcal{I}}$.

We use the following notational conventions. Let $\{x_s, s \in \mathcal{S}\}$ be a set of vectors of possibly different dimensions, indexed by some finite set $\mathcal{S}$. We will denote by $\mathrm{vec}\{x_s, s \in \mathcal{S}\}$ the vector obtained by staking the x_s's one on top of each others in the same order as the elements of $\mathcal{S}$. For any d-dimensional Euclidean vector x, we write $\exp(x) = \mathrm{vec}\{\mathrm{e}^{x_i}, i = 1, \ldots, d\}$, and $\log x = \mathrm{vec}\{\log x_i, i = 1, \ldots, d\}$. For a linear subspace $\mathcal{A}$ of the d-dimensional Euclidean space, we denote with $\mathcal{A}^{\perp}$ the orthogonal complement of $\mathcal{A}$. If $\mathcal{B}$ is another linear subspace orthogonal to $\mathcal{A}$, we write $\mathcal{A} \oplus \mathcal{B}$ for the linear subspace obtained as their direct sum. Similarly, for matrices $\mathrm{U}_1, \ldots, \mathrm{U}_n$ with the same number of rows r and number of columns $c_1, \ldots, c_n$, respectively, we denote the operation of adjoining them into one matrix of dimension $r \times \sum_k c_k$ with $\bigoplus_{k=1}^{n} \mathrm{U}_k = [\mathrm{U}_1 \cdots \mathrm{U}_n]$.

Throughout the article, we will consider random vectors and functions of random vectors whose probability distributions will always be clear from the context. As a result, we will use the generic notation $\mathbb{P}(\mathcal{O})$ for the probability of an event $\mathcal{O}$ defined by such vectors and will write $\mathbb{E}$ for the corresponding expectation operator.

2. Log-linear models

We will be considering the usual log-linear modeling setting, which we now describe. See (see also Bishop, Fienberg and Holland [4], Haberman [17], Lauritzen [18]). The K categorical random variables $X_1, \ldots, X_K$ have an unknown joint distribution given by the strictly positive probability vector $\boldsymbol{\pi}$ in $\mathbb{R}^{\mathcal{I}}$ with coordinates

$$\boldsymbol{\pi}_{i_1, \ldots, i_K} = \mathbb{P}\big((X_1, \ldots, X_K) = (i_1, \ldots, i_K)\big), \qquad (i_1, \ldots, i_K) \in \mathcal{I}.$$

The positivity of $\boldsymbol{\pi}$ is a crucial assumption, ruling out the case of *structural zeros*, that is, cells that can never be observed.

We observe N independent and identically distributed realizations of the random vector $(X_1, \ldots, X_K)$. Their cross-classification results in a random integer-valued vector $\mathbf{n} \in \mathbb{R}^{\mathcal{I}}$, called a *contingency table*, whose ith coordinate entry $\mathbf{n}_i$ corresponds to the number of times the cell

combination $i = (i_1, \ldots, i_K)$ was observed in the sample. The table $\mathbf{n}$ has a Multinomial$(N, \boldsymbol{\pi})$ distribution.

Log-linear model theory is concerned with drawing inferences on $\boldsymbol{\pi}$ based on the observed table $\mathbf{n}$. Specifically, let $\mathbf{m} = \mathbb{E}\mathbf{n} = N\boldsymbol{\pi}$ denote the (necessarily positive) cell mean vector of $\mathbf{n}$ and set $\boldsymbol{\mu} = \log \mathbf{m}$. Notice that estimating $\boldsymbol{\mu}$ is equivalent to estimating $\boldsymbol{\pi}$. A log-linear model is specified by prescribing a linear subspace $\mathcal{M}$ of $\mathbb{R}^{\mathcal{I}}$ containing the constant functions and then requiring that $\boldsymbol{\mu}$ belongs to $\mathcal{M}$. Indeed, any point in $\mathcal{M}$ represents a different cell mean vector, and, therefore, a different probability distribution over $\mathcal{I}$.

Then, for a given table $\mathbf{n}$ the log-likelihood function ℓ^* at $\boldsymbol{\mu} \in \mathcal{M}$ is (see Haberman [17], page 11)

$$\ell^*(\boldsymbol{\mu}) = \begin{cases} \displaystyle\sum_{i \in \mathcal{I}} \mathbf{n}_i \log \frac{\mathbf{m}_i}{\langle \mathbf{m}, \mathbf{1} \rangle} + \log N! - \sum_{i \in \mathcal{I}} \log \mathbf{n}_i !, & \text{if } \langle \mathbf{m}, \mathbf{1} \rangle = N, \\ \text{undefined}, & \text{otherwise}, \end{cases}$$

where $\mathbf{m} = \exp(\boldsymbol{\mu})$ and $\mathbf{1} \in \mathbb{R}^{\mathcal{I}}$ is the I-dimensional vector containing ones. Indeed, because of the Multinomial sampling assumption, ℓ^* is only defined over the nonconvex set $\widetilde{\mathcal{M}} \subsetneq \mathcal{M}$ given by

$$\widetilde{\mathcal{M}} = \{\boldsymbol{\mu} \in \mathcal{M} \colon \langle \mathbf{m}, \mathbf{1} \rangle = N\}.$$

This parametrization is clearly quite inconvenient. Fortunately, it is possible to reparametrize the log-likelihood function as concave function defined over the entire $\mathbb{R}^k$, where k is the dimension of $\widetilde{\mathcal{M}}$. Specifically, let $\mathcal{R}(\mathbf{1})$ be the one-dimensional subspace of $\mathbb{R}^{\mathcal{I}}$ spanned by $\mathbf{1}$ and consider the linear subspace $\mathcal{M} \cap \mathcal{R}(\mathbf{1})^{\perp} \subset \mathbb{R}^{\mathcal{I}}$ of dimension $k = \dim(\mathcal{M}) - 1$.

Lemma 2.1. *Let* U *be any full-rank matrix whose columns span* $\mathcal{M} \cap \mathcal{R}(\mathbf{1})^{\perp}$ *and consider the function*

$$\ell(\theta) = \langle \mathrm{U}^{\top}\mathbf{n}, \theta \rangle - N \log \langle \exp(\mathrm{U}\theta), \mathbf{1} \rangle + \log N! - \sum_{i \in \mathcal{I}} \log \mathbf{n}_i !, \quad \theta \in \mathbb{R}^k. \tag{1}$$

Then, for each $\widetilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}$ *there exists one* $\theta \in \mathbb{R}^k$ *such that*

$$\exp(\widetilde{\boldsymbol{\mu}}) = \frac{N}{\langle \exp(\mathrm{U}\theta), \mathbf{1} \rangle} \exp(\mathrm{U}\boldsymbol{\theta}) \quad \textit{and} \quad \ell(\theta) = \ell^*(\widetilde{\boldsymbol{\mu}}) \qquad \textit{for each } \mathbf{n}, \tag{2}$$

and, conversely, for each $\theta \in \mathbb{R}^k$ *there exists one* $\widetilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}$ *satisfying the above identities.*

This reparametrization 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 $\mathcal{M}$ specifies a full, regular exponential family of dimension $k = \dim(\mathcal{M}) - 1$, natural sufficient statistic $\mathrm{U}^{\top}\mathbf{n}$ and natural parameter space $\mathbb{R}^k$ (see, e.g., Brown [6]). Throughout the article, we take (1) as the operative definition of log-likelihood function. Notice that the log-likelihood function depends on the choice of the design matrix U.

2.1. Log-linear subspaces for hierarchical log-linear models

Although log-linear models are defined by generic linear manifolds of $\mathbb{R}^{\mathcal{I}}$, in practice it is customary to consider only very specific classes of linear subspaces, which are also characteristic of ANOVA models and experimental design, yielding hierarchical log-linear. In this section, we briefly describe such subspaces. See Darroch, Lauritzen and Speed [8], Appendix B in Lauritzen [18] and Rinaldo [30] for details.

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 $\mathcal{K} \equiv \{1, 2, \ldots, K\}$, each one indicating a different type of interaction. In fact, every subset h of $\mathcal{K}$ can be given a straightforward ANOVA-type of an interpretation, based on its cardinality $|h|$, so that h identifies an interaction of order $|h| - 1$ among the variables $\{i\colon\ i \in h\}$. For example, if $|h| = 1$, then h is a main effect, if $h = \varnothing$, then h is the grand mean, and so on.

Formally, let $2^{\mathcal{K}}$ be the power set of $\mathcal{K}$, which we view as a lattice with respect to the partial order induced by the operation of taking subset inclusion.

Definition 2.2. *A hierarchical log-linear model Δ is a collection of subsets of $2^{\mathcal{K}}$ such that $h \in \Delta$ and $h' \subset h$ implies $h' \in \Delta$. An interaction model $\mathcal{H}$ is just a subset of $2^{\mathcal{K}}$.*

By definition, once an interaction term is part of Δ, all lower order interactions are included. Notice that Definition 2.2 includes as special case the class of graphical and hierarchical models (see, e.g., Lauritzen [18]). Though our analysis is valid also for the larger class of interaction models, we focus only on hierarchical log-linear models, primarily because the interpretability of interaction log-linear models is very limited.

To any given hierarchical model Δ, there corresponds one log-linear subspace $\mathcal{M}_\Delta \subset \mathbb{R}^{\mathcal{I}}$, constructed as the direct sums of subspaces of $\mathbb{R}^{\mathcal{I}}$ indexed by the subsets of $\mathcal{K}$ belonging to Δ. Specifically,

$$\mathcal{M}_\Delta = \bigoplus_{h \in \Delta} \mathcal{U}_h, \tag{3}$$

where $\{\mathcal{U}_h, h \in 2^{\mathcal{K}}\}$ are mutually orthogonal subspaces, called the *subspaces of interactions*. We refer the reader to Lauritzen [18], Appendix B, for details on these subspaces. In particular, $\mathcal{U}_\varnothing$ is the one-dimensional subspace $\mathcal{R}(\mathbf{1})$ and

$$\dim(\mathcal{U}_h) \equiv d_h = \prod_{k \in h} (I_k - 1), \qquad h \subseteq \mathcal{K}, h \neq \varnothing.$$

A design matrix for Δ can be constructed as follows. For each term $h \subseteq \mathcal{K}$ and factor $k \in \mathcal{K}$, define the matrix

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

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

$$\mathrm{Z}_k = \begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ -1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & 0 & \cdots & -1 \end{pmatrix}, \tag{4}$$

and $\mathbf{1}_k$ is the I_k-dimensional column vector of 1's. Let

$$\mathrm{U}_h = \bigotimes_{k=1}^{K} \mathrm{U}_k^h, \tag{5}$$

where $\otimes$ denotes the Kronecker product. Then, it is possible to show that, for each $h \in 2^{\mathcal{K}}$, U_h is a $(I \times d_h)$-dimensional full-rank matrix whose columns span $\mathcal{H}_h$. See Rinaldo [30], Section 3, for details. Thus, the columns of

$$\mathrm{U}_\Delta = \bigoplus_{h \in \Delta, h \neq \varnothing} \mathrm{U}_h \tag{6}$$

span $\mathcal{M}_\Delta \cap \mathcal{R}(\mathbf{1})^\perp$, and, therefore, U_Δ is a full-rank design matrix for the log-linear model Δ. By the same token, the columns of the matrix

$$\mathrm{U} = \bigoplus_{h \in 2^{\mathcal{K}}, h \neq \varnothing} \mathrm{U}_h$$

span the $(I - 1)$-dimensional subspace $\mathbb{R}^{\mathcal{I}} \cap \mathcal{R}^\perp$. As a result, any point $\boldsymbol{\mu} \in \mathbb{R}^{\mathcal{I}} \cap \mathcal{R}(\mathbf{1})^\perp$ can be written as

$$\boldsymbol{\mu} = \mathrm{U}\theta = \sum_{h \in 2^{\mathcal{K}}, h \neq \varnothing} \mathrm{U}_h \theta_h,$$

for some vector

$$\theta = \mathrm{vec}\{\theta_h, h \in 2^{\mathcal{K}}, h \neq \varnothing\} \in \mathbb{R}^{I-1}, \tag{7}$$

where θ_h denotes the d_h-dimensional sub-vector of θ corresponding to the sub-matrix U_h.

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. Then, using such ordering, any vector $\theta \in \mathbb{R}^{I-1}$ can be uniquely represented like in (7). Furthermore, we will use the notation $\sum_h$ to denote summation over all $h \subseteq \mathcal{K}$ with $h \neq \varnothing$.

3. The group-lasso estimator for log-linear models

In this section, we define the group-lasso estimator for the set of interactions of a given hierarchical log-linear model specified by a subspace $\mathcal{M}_\Delta$.

Using (1), the log-likelihood function for the saturated $(I-1)$-dimensional log-linear model is

$$\begin{aligned}\ell(\theta) = &\sum_{h\in 2^{\mathcal{K}}, h\neq\varnothing} \langle \mathrm{U}_h^\top \mathbf{n}, \theta_h\rangle - N\log\left\langle \exp\left(\sum_{h\in 2^{\mathcal{K}}, h\neq\varnothing} \mathrm{U}_h\theta_h\right), \mathbf{1}\right\rangle \\ &+ \log N! - \sum_i \log \mathbf{n}_i!, \qquad \theta\in\mathbb{R}^{I-1}.\end{aligned} \tag{8}$$

Notice that the one-dimensional sub-space $\mathcal{R}(\mathbf{1})$ corresponding to the empty set is not included, because of the multinomial sampling restriction.

For any nontrivial model Δ with corresponding log-linear subspace $\mathcal{M}_\Delta$ (i.e., a model different than $\{\varnothing\}$, which encodes the uniform distribution over $\mathcal{I}$), let

$$\mathcal{H} = \mathcal{H}(\Delta) = \{h\colon\ h\in\Delta, h\neq\varnothing\}, \tag{9}$$

be the collections of sets representing all the interactions in Δ, or, equivalently, the collections of factor interaction subspaces of $\mathcal{M}_\Delta$, so that $\dim(\mathcal{M}_\Delta)-1 = \sum_{h\in\mathcal{H}} d_h \equiv d_{\mathcal{H}}$. (Notice that $\mathcal{H}$ differs from Δ only because it does not contain the empty set). We will embed the natural parameter space of Δ, i.e., $\mathbb{R}^{d_{\mathcal{H}}}$, as a linear subspace of $\mathbb{R}^{I-1}$ consisting of all vectors such that

$$\begin{cases} \|\theta_h\| > 0, & h\in\mathcal{H}, \\ \|\theta_h\| = 0, & h\notin\mathcal{H}. \end{cases}$$

The log-likelihood function for this model is still given by equation (8), where the summations are now taken over the sets h in the class $\mathcal{H}$.

Let Δ_0 denote the true underlying log-linear model. Thus, there exists a vector of parameters $\theta^0\in\mathbb{R}^{I-1}$ such that $\|\theta_h^0\|$ is positive for all $h\in\mathcal{H}(\Delta_0)$ and zero otherwise. Having observed a contingency table $\mathbf{n}$, we seek to recover Δ_0. That is, our goal is to identify those block components of θ^0 having positive norms. To this end, we define the group-lasso estimator for log-linear models to be the solution of the concave optimization problem

$$\max_{\theta\in\mathbb{R}^{I-1}} P_\Lambda(\theta) \equiv \max_{\theta\in\mathbb{R}^{I-1}} \left\{\frac{1}{N}\ell(\theta) - \lambda\sum_h \lambda_h\|\theta_h\|\right\}, \tag{10}$$

with $\ell(\cdot)$ defined as in (8) and $\Lambda = \{\lambda, \{\lambda_h, h\neq\varnothing\}\}$ a set of given tuning parameters. The parameter λ controls the overall effect of the penalty and should be a function of the sample size, while the block parameters λ_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.

The group-lasso penalty appearing in (10) was first proposed by Yuan and Lin [40] in the context of linear Gaussian models under ANOVA settings (see also Nardi and Rinaldo [23]). It is specifically designed to produce sparsity in the vector of estimated coefficients at the block level. It is obtained as compositions of the ℓ_1 norm over quadratic norms of the individual blocks. The quadratic norms of individual blocks promote non-sparsity, whereas the ℓ_1 norm applied to the resulting block norms, promotes block sparsity. The group-lasso methodology of Yuan and Lin [40] was further extended to logistic regression models by Meier, van der Geer and Bühlmann [20] and to log-linear models by Dahinden et al. [7], which inspired our work.

Lemma 3.1. *The vector $\widehat{\theta} \in \mathbb{R}^{I-1}$ is an optimizer of (10) if and only if there exists a vector $\widehat{\eta} \in \mathbb{R}^{I-1}$ such that, for any h,*

$$-\frac{1}{N}\mathrm{U}_h^\top(\mathbf{n} - \widehat{\mathbf{m}}) + \lambda\lambda_h\widehat{\eta}_h = 0, \tag{11}$$

where

$$\widehat{\eta} = \begin{cases} \dfrac{\widehat{\theta}_h}{\|\widehat{\theta}_h\|_2}, & \text{if } \widehat{\theta}_h \neq 0, \\ \lambda_h\widehat{z}_h, & \text{if } \widehat{\theta}_h = 0 \end{cases}$$

with $\|\widehat{z}_h\| \leq 1$, and $\widehat{\mathbf{m}} = \frac{N}{\langle \exp(\mathrm{U}\widehat{\theta}), \mathbf{1}\rangle} \exp(\mathrm{U}\widehat{\theta})$. The solution is unique if $\|\widehat{z}_h\| < 1$ for each h for which $\widehat{\theta}_h = 0$.

Having obtained the group-lasso estimator $\widehat{\theta}$, the model selection step entails building an estimate of the true model Δ_0 by extracting the blocks of $\widehat{\theta}$ with positive norm and then build a hierarchical model $\widehat{\Delta}$ as illustrated in Table 1. There are two advantages in using the group-lasso estimator for estimating Δ_0 rather than traditional methods of model selection based on sequential testing of a potentially very large number of competing models. The first advantage is that the methodology described in Table 1 only involves determining a penalized maximum likelihood estimator of θ^0 and thus requires solving only one convex optimization problem (albeit a hard one, see the discussion below). In contrast, classical model selection procedure requires fitting and comparing a number of different models which, even for tables with a small number of variables, can be unfeasible. The second advantage is that the group-lasso estimator always return a model for which the maximum likelihood estimate exists, a fact that is not guaranteed by the computational procedures currently used in practice. See Fienberg and Rinaldo [12] for more details.

Remark. Though motivated by model selection with hierarchical models, our analysis below will actually show the log-linear group lasso estimator can asymptotically recover any log-linear interaction subspace. This is the reason why in equation (9) we used the more general notation $\mathcal{H}$ to encode the interactions of a hierarchial log-linear model Δ. Thus, the last step in the algorithm described in Table 1, which forces the estimated model to be hierarchical, should not be used if interested in general interaction models. Furthermore, while asymptotically this step is unnecessary for a hierarchical model, we nonetheless believe it would improve the finite sample performance of the algorithm.

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

1. Obtain the log-linear group-lasso estimator, $\widehat{\theta} = \text{argmax}_{\theta \in \mathbb{R}^{I-1}} P_\Lambda(\theta)$.
2. Extract the set of non-zero blocks from $\widehat{\theta}$, $\widehat{\mathcal{H}} = \{h\colon \|\widehat{\theta}_h\| > 0\}$.
3. Recover the hierarchical log-linear model from $\widehat{\mathcal{H}}$, $\widehat{\Delta} = \{h'\colon h' \subseteq h, \text{ for some } h \in \widehat{\mathcal{H}}\}$.

Model complexity and computational considerations

We now make some remarks on the complexity of the class of log-linear models. The model selection problem for log-linear models is characterized by a combinatorial explosion in the number of possible models that is much larger than for linear and many generalized-linear models. Combined with the fact that the notion of sample size is also quite different, as it refers to total number of counts N in our settings and to the total number of observed counts $\prod_k I_k$ in the linear and generalized linear model settings, a direct comparison between the computational burden of our estimator versus the lasso or group-lasso procedure for those models is not entirely adequate.

It follows from our combinatorial definition of a general log-linear model from Section 2.1 that, for a K-way table, log-linear models can be represented as subsets of the set of all the $2^K - 1$ possible interactions, except the one specified by the empty set. Table 2 displays the number of log-linear models of different types as a function of K. For a given K, the number of possible unrestricted log-linear model is 2^{2^K-1}, which super-exponential in K. The number of hierarchical models, for which no closed form expression is available, is much smaller, but still appears to grow extremely fast (roughly exponentially) in K. For the even smaller subclasses of graphical[1] and decomposable models, the space of possible models is still extremely large, for small values of K. For example, for a binary 5-dimensional table with a total of 32 observed counts, there are 1233 decomposable models! Due to to this tremendous combinatorial complexity, model selection by exhaustive model search is computationally prohibitive and, in fact, unfeasible even for very small tables. Indeed, model selection techniques for log-linear models must rely necessarily on very greedy algorithms and are often designed to consider only graphical or decomposable models. See, for instance, the Baysian procedure of Dobra and Massam [9], and the frequentis model selection search based on asymptotic χ^2 testing implemented in the software `MIM`, described in Edwards [10].

In contrast, the group-lasso procedure we consider is parametrized, in both the likelihood and the penalty part, by only $2^K - 1$ terms, which correspond to the direct sum decomposition of $\mathbb{R}^{\mathcal{I}}$ into the orthogonal interaction subspaces. Although this number grows exponentially in K, it is still order of magnitudes smaller than the possible number of models. Indeed, based on Table 2, it appears to be loosely logarithmic in the number of possible models. As a result, even though the

[1]The number of possible graphical models is $\sum_{i=0}^{K} \binom{K}{i} 2^{\binom{i}{2}}$.

Table 2. Number of log-linear models of different types as a function of K. Source: Lauritzen [19]

	K				
Type	1	2	3	4	5
Unrestricted	2	8	128	32,768	2,147,483,648
Hierarchical	2	5	19	167	7580
Graphical	2	5	18	113	1450
Decomposable	2	5	18	110	1233

computational complexity of (10) may be high, it is significantly smaller than exhaustive model selection and many greedy model selection algorithms.

To our knowledge, two algorithms for computing the group-lasso estimates are currently available: the block-coordinate descent method of Meier, van der Geer and Bühlmann[20], developed for the specific case of logistic regression with grouped variables but immediately extendible to our settings, and the path-following algorithm of Dahinden et al. [7] for general log-linear models on binary variables. Both methods showed good performance on simulated data and have been applied successfully to real-life datasets, with the tuning parameters chosen by cross-validation. Those results corroborate our theoretical findings that group-lasso estimator possess good theoretical properties and is a valuable alternative to greedy model selection procedures. For completeness, we reference the more recent works by Roth and Fisher [32], Puig, Wiesel and Hero [27], Yuan, Roshan, and Zou [39] and Friedman, Hastie and Tibshirani [13].

Nonetheless, we remark that the development of efficient computational methods for calculating the group-lasso solution (for log-linear as well as for linear model) with proven performance particularly in very high-dimensional settings still remains an open problem.

4. Asymptotic analysis

In this section, we provide the main results of the paper. We perform here a 'large p and large N' type of an asymptotic analysis of the model selection procedure described in Table 1 and of the properties of the group-lasso estimator. We 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 assume 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 $\mathcal{I}_{k_n} = \{1, \ldots, I_{k_n}\}$, $k_n = 1, \ldots, K_n$; the support of this distribution is the set $\mathcal{I}_n = \bigotimes_{k_n} \mathcal{I}_{k_n}$ of all cell combinations, of cardinality $I_n = \prod_{k_n} I_{k_n}$;

- the true underlying distribution is defined by a hierarchical log-linear model Δ_n, as described in Section 2.1: the observed cell counts come from an exponential family distributions with log-likelihood function (8) and true natural parameter $\theta_n^0 \in \mathbb{R}^{I_n - 1}$, such that $\|\theta_{h_n}^0\| > 0$ for $h_n \in \mathcal{H}_n$ and $\|\theta_{h_n}^0\| = 0$ for $h_n \notin \mathcal{H}_n$, with $\mathcal{H}_n$ defined as in (9); the corresponding vector of cell probabilities is denoted with $\boldsymbol{\pi}_n^0 \in \mathbb{R}^{\mathcal{I}_n}$ and the mean vector $N_n \boldsymbol{\pi}_n^0$ with $\mathbf{m}_n^0$;
- the vector of true parameters θ_n^0 is estimated by solving the program (10) with tuning parameters $\Lambda_n = \{\lambda_n, \{\lambda_{h_n}, h_n \neq \varnothing\}\}$;
- the group-lasso estimate $\widehat{\theta}_n$ is then used to estimate Δ_n as described in Table 1, leading to the optimal selected model $\widehat{\Delta}_n$.

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, for instance, Schervish ([33], Definition 7.11 and Lemma 7.12).

Projecting down the true parameter θ_n^0 into $\mathbb{R}^{d_{\mathcal{H}_n}}$ we write $\theta_{\mathcal{H}_n}^0$. One may take note that, for a single observation, the Fisher information matrix at $\theta_{\mathcal{H}_n}^0$ is

$$\mathrm{F}_{\mathcal{H}_n} = \mathrm{U}_{\mathcal{H}_n}^\top \big(\mathrm{D}_{\boldsymbol{\pi}_n^0} - \boldsymbol{\pi}_n^0 (\boldsymbol{\pi}_n^0)^\top\big) \mathrm{U}_{\mathcal{H}_n},$$

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

$$\Sigma_{\mathcal{H}_n} = \mathrm{U}_{\mathcal{H}_n}^\top \left(\mathrm{D}_{\mathbf{m}_n^0} - \frac{\mathbf{m}_n^0 (\mathbf{m}_n^0)^\top}{N_n} \right) \mathrm{U}_{\mathcal{H}_n} = N_n \mathrm{F}_{\mathcal{H}_n}, \tag{12}$$

which is the covariance matrix of the natural sufficient statistics $\mathrm{U}_{\mathcal{H}_n}^\top \mathbf{n}_n$ and the Fisher information based on a i.i.d. sample of size N.

In the reminder of the article, we will study some of the asymptotic properties of the sequence of group-lasso estimates $\{\widehat{\theta}_n\}_n$ generated according to the previous scheme. In Section 4.1, we prove the that the group-lasso estimator is model selection consistent, that is,

$$\lim_n \mathbb{P}(\widehat{\Delta}_n = \Delta_n) = 1. \tag{13}$$

Finally, in Section 4.2 we give a central limit theorem for $\widehat{\theta}_n$.

4.1. Model selection consistency

Here, we derive sufficient conditions for the property of model selection consistency (13). Our method of analysis is based on linearizing the sub-gradient optimality conditions (11) via a Taylor expansion around the sequence of true parameters θ_n^0. As it turns out, norm (or l_2) consistency is necessary to guarantee enough stochastic control over the remainder term of that expansion. To that end, we first establish norm consistency (in Lemma 4.1) concerning a related optimization problem (see (17) below). The conditions we develop for model selection consistency 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 [37], Zhao and Yu [41]).

Recall the definition of $\mathcal{H}_n$ from (9) and let $\mathcal{H}_n^c = 2^{\mathcal{K}_n} \setminus (\mathcal{H}_n \cup \varnothing)$, so that $\|\theta_{h_n}^0\| > 0$ for each $h_n \in \mathcal{H}_n$ and $\|\theta_{w_n}^0\| = 0$ for each $w_n \in \mathcal{H}_n^c$. Consider the sequence of events indexed by n

$$\mathcal{O}_n = \{\|\widehat{\theta}_{h_n}\| > 0, \forall h_n \in \mathcal{H}_n\} \cap \{\|\widehat{\theta}_{w_n}\| = 0, \forall w_n \in \mathcal{H}_n^c\}. \tag{14}$$

Then, the model selection consistency property (13) of the group-lasso solutions $\widehat{\theta}_n$ is equivalent to convergence in probability of $\mathcal{O}_n$, namely $\lim_n \mathbb{P}(\mathcal{O}_n) = 1$. This in turn occurs if and only if

$$\lim_n \mathbb{P}(\|\widehat{\theta}_{h_n}\| > 0, \forall h_n \in \mathcal{H}_n) = 1 \tag{15}$$

and

$$\lim_n \mathbb{P}(\|\widehat{\theta}_{w_n}\| = 0, \forall w_n \in \mathcal{H}_n^c) = 1. \tag{16}$$

In this section, we will provide sufficient conditions for (15) and (16).

Our method of analysis relies on the primal-dual witness construction of Wainwright [37], which we summarize below.

1. Solve a restricted group-lasso problem

$$\widetilde{\theta}_n = \underset{\theta \in \mathbb{R}^{d\mathcal{H}_n}}{\operatorname{argmax}} P_\Lambda(\theta), \tag{17}$$

   where $\Lambda = \{\lambda_n, \lambda_h, h \in \mathcal{H}_n(\Delta_0)\}$.
2. Choose a vector $\widehat{\eta}_{\mathcal{H}_n}$ that belongs to the subdifferential of group lasso penalty $\lambda_n \sum_{h,h\in\mathcal{H}_n} \lambda_h \|x_h\|$ evaluated at $\widetilde{\theta}_n$, and solve for a $(I - 1 - d_{\mathcal{H}_n})$-dimensional vector

$$\widehat{\eta}_{\mathcal{H}_n^c} = \operatorname{vec}\{\lambda_{w_n} \widehat{z}_{w_n}, w_n \in \mathcal{H}_n^c\}$$

   the optimality conditions (11). Check that $\|\widehat{z}_{w_n}\| < 1$, for all $w_n \in \mathcal{H}_n^c$.
3. The vector $\hat{\theta}_n = \operatorname{vec}\{\widetilde{\theta}_n, 0\} \in \mathbb{R}^{I-1}$ is the unique solution (see Lemma 3.1) of (11) and the event in equation (14) holds.

Since the optimality conditions (11) are non-linear functions of $\theta \in \mathbb{R}^{I-1}$, step 2. entails first a linearization step to bound the difference between $\widetilde{\theta}_n$ and $\theta_{\mathcal{H}_n}^0$ by a Taylor series expansion, and then showing that the resulting remainder term vanished in probability. This, in turn, follows from the next result, which established that $\widetilde{\theta}_n$ is a norm consistent estimator of $\theta_{\mathcal{H}_n}^0$.

Lemma 4.1. *Assume*

1. [NC.1] $d_{\mathcal{H}_n} = o(N_n)$;
2. [NC.2] $0 < D_{\min} < l_n^{\min} \leq l_n^{\max} < D_{\max} < \infty$, *here* $l_n^{\min}, l_n^{\max}$ *are the minimal and maximal eigenvalue of the Fisher information matrix, respectively*;
3. [NC.3] *for any* $D > 0$,

$$\sup\left\{|\mathbb{E}_{\theta_n}[\langle a, \mathrm{U}_{\mathcal{H}_n}^\top X\rangle]|\colon \|\theta_n - \theta_{\mathcal{H}_n}^0\| \leq D\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}, \|a\| = 1\right\} = O\left(\sqrt{\frac{N_n}{d_{\mathcal{H}_n}}}\right), \tag{18}$$

where, for $\theta_n \in \mathbb{R}^{d_{\mathcal{H}_n}}$, $\mathbb{E}_{\theta_n}$ *denotes the expectation operator with respect to the distribution* Multinomial$(1, \boldsymbol{\pi}_n)$, *with*

$$\boldsymbol{\pi}_n = \frac{\exp(\mathrm{U}_{\mathcal{H}_n}\theta_n)}{\langle \exp(\mathrm{U}_{\mathcal{H}_n}\theta_n), \mathbf{1}\rangle};$$

4. [NC.4] $\lambda_n = O(\frac{1}{\sum_{h_n \in \mathcal{H}_n} \lambda_{h_n}})$.

Then,

$$\|\widetilde{\theta}_n - \theta_{\mathcal{H}_n}^0\| = O_{P_n^0}\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right) = o_{P_n^0}(1). \tag{19}$$

The proof of Lemma 4.1 relies on results of Portnoy [26]. Indeed, conditions [NC.1] and [NC.2] are essentially derived from Portnoy [26], Theorem 2.1. The more technical condition [NC.3], also derived from Portnoy [26], equation (2.4), is needed to establish some control the order of magnitude of the remainder term in the local quadratic approximation of the log-likelihood function around $\theta_{\mathcal{H}_n}^0$, uniformly over compact neighborhoods (see also Ghosal [14], for similar conditions).

Armed with (19), we now proceed to prove the property of model selection consistency for the group-lasso.

Theorem 4.2. *Assume the conditions of Lemma* 4.1. *Then, equation* (15) *holds if*

- [MSC.1] *letting* $\alpha_n = \min_{h_n \in \mathcal{H}_n} \|\theta_{h_n}^0\|$,

$$\frac{1}{\alpha_n}\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} + \lambda_n\sqrt{\sum_{h_n \in \mathcal{H}_n} \lambda_{h_n}^2}\right) \to 0, \tag{20}$$

which, for $\lambda_{h_n} = \sqrt{d_{h_n}}$, *simplifies to* $\frac{\sqrt{d_{\mathcal{H}_n}}}{\alpha_n}(\sqrt{\frac{1}{N_n}} + \lambda_n) \to 0$.

Equation (16) *holds if*

- [MSC.2] (*'almost' parameter orthogonality*) *for some* $\varepsilon \in (0, 1)$ *and for each* $w_n \in \mathcal{H}_n^c$,

$$\left\|\mathrm{U}_{w_n}^\top\left(D_{\mathbf{m}_n^0} - \frac{\mathbf{m}_n^0(\mathbf{m}_n^0)^\top}{N_n}\right)\mathrm{U}_{\mathcal{H}_n}\Sigma_{\mathcal{H}_n}^{-1}\right\| < \frac{(1-\varepsilon)}{|\mathcal{H}_n^c|}; \tag{21}$$

- [MSC.3]

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

- [MSC.4]

$$\left(\max_{w_n \in \mathcal{H}_n^c} \frac{d_{w_n}}{\lambda_{w_n}^2} \right) \frac{\log(I_n - 1 - d_{\mathcal{H}_n})}{N\lambda_n^2} \to 0,$$

which, for the choice $\lambda_{h_n} = \sqrt{d_{h_n}}$, *becomes* $\frac{\log(I_n - 1 - d_{\mathcal{H}_n})}{N\lambda_n^2} \to 0, \to \infty$.

Condition [MSC.2] implies that

$$\|\mathrm{W}_n\| = \left\| \mathrm{U}_{\mathcal{H}_n^c}^\top \left(D_{\mathbf{m}_n^0} - \frac{\mathbf{m}_n^0 (\mathbf{m}_n^0)^\top}{N_n} \right) \mathrm{U}_{\mathcal{H}_n} \Sigma_{\mathcal{H}_n}^{-1} \right\| < (1 - \varepsilon),$$

which is the equivalent of the so-called *irreducibility condition* appearing in the growing lasso literature (e.g., Wainwright [37] and Zhao and Yu [41]). Condition [MSC.3] is a sparsity condition and condition [MSC.4] provides some information on the rates of increase for the dimensions for the subspaces of interactions not included in the true models (see equation (15) (a) in Wainwright [37]). Inspection of the proof will reveal that, using a simpler argument based on Chebyshev's inequality only, and assuming $\lambda_{h_n} = \sqrt{d_{h_n}}$ for all h, [MSC.4] reduces to $\lambda_n^2 N_n \to \infty$, so that model selection consistency is obtained under conditions that do no depend on I_n.

4.2. 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 $\{\widehat{\theta}_n\}_n$. 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 $\mathcal{O}_n$ defined in (14), 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.

To state our result, we need to formulate some notation. Let $\mathrm{J}_{\mathcal{H}_n}^0$ 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

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

with $h_n \in \mathcal{H}_n$ and with $\mathrm{I}_{d_{h_n}}$ denoting the d_{h_n}-dimensional identity matrix. Below, G_n will denote a $k \times d_{\mathcal{H}_n}$ matrix, where k is an arbitrary fixed number, such that

$$\lim_n \mathrm{G}_n \mathrm{G}_n^\top = \mathrm{G} \tag{22}$$

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

Theorem 4.3. *Assume the conditions for norm and model selection consistency and let*

$$X_n = \sqrt{N_n} \mathrm{F}_{\mathcal{H}_n}^{-1/2} \big((\mathrm{F}_{\mathcal{H}_n} + \lambda_n \mathrm{J}_{\mathcal{H}_n}^0)(\widehat{\theta}_n - \theta_{\mathcal{H}_n}^0) + \lambda_n \eta_{\mathcal{H}_n}^0 \big), \tag{23}$$

where

$$\eta_{\mathcal{H}_n}^0 = \mathrm{vec} \left\{ \lambda_{h_n} \frac{\theta_{h_n}^0}{\|\theta_{h_n}^0\|}, h_n \in \mathcal{H}_n \right\}.$$

1. *For each sequence* $\{\mathrm{G}_n\}$ *of* $k \times d_{\mathcal{H}_n}$ *matrices satisfying* (22), $\mathrm{G}_n X_n$ *converges weakly to the* $N_k(0, \mathrm{G})$ *distribution if either the* [CLT.LF] *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*

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

hold.

2. *If the* [CLT.BE] *condition*

$$d_{\mathcal{H}_n} = o(N_n^{2/7}),$$

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* $N_{d_{\mathcal{H}_n}}(0, \mathrm{I}_{d_{\mathcal{H}_n}})$ *distribution, the supremum is taken over the convex sets* A_n *in* $\mathbb{R}^{d_{\mathcal{H}_n}}$ *and convergence occurs at the rate* $O(\frac{d_{\mathcal{H}_n}^{7/2}}{N_n})$.

The theorem indicates that the group-lasso estimate is asymptotically unbiased and inefficient. In fact, equation (23) demonstrates that the asymptotic behavior of the group-lasso estimator is affected by two terms. One is the bias term $\lambda_n \eta_{\mathcal{H}_n}^0$ which depends on the gradient of the penalty function at the true parameter. The other term $\mathrm{J}_{\mathcal{H}_n}^0$ 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 (see also Fan and Peng [11], Theorem 2).

Both Condition [CLT.LF] and conditions [CLT.Ma] and [CLT.Mb] guarantee the asymptotic normality of a *fixed* number of linear combinations of the coordinates of $\widehat{\theta}_n$. In particular, it includes that case of $\mathrm{G}_n = [\mathrm{I}_k \ \ \mathrm{O}]$, where O is a $k \times (d_{\mathcal{H}_n} - k)$ matrix of zeros. For this choice, the marginal asymptotic normality of any fixed number of coordinates of $\widehat{\theta}_n$ is guaranteed. Condition [CLT.LF] results from a simple Lindberg–Feller argument, whereas conditions [CLT.Ma] and [CLT.Mb] follow by adapting and generalizing some proofs in Morris [22]. We note that [CLT.Mb] may be replaced by $\max_{i \in \mathcal{I}_n} \boldsymbol{\pi}_i^0 \leq C I_n^{-1}$, for some positive constant C (see, e.g., Quine and Robinson [28], Theorem 1). Then, in order for the theorem to hold, one has to further assume that $I_n = o(\sqrt{d_{\mathcal{H}_n} N_n})$, which is compatible with the conditions for norm consistency.

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 [1]. 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 [25] for a similar result requiring only a rate $d_{\mathcal{H}_n}^2 = o(N_n^{1/2})$, whose applicability and relevance to our problem is however unclear.

5. Conclusions

In this article, we studied some asymptotic properties of the group-lasso estimator. Our results show that this estimator can be used to recover asymptotically the true underlying model under conditions that allow for a model complexity increasing with the sample size and also for a number of cells larger than N.

Our setting, analysis and results differ from existing analyses of ℓ_1 regularized least square problems in a few aspects. Firstly, unlike the case of regularized least squares or Gaussian error problems, the first order optimality conditions for the group-lasso program are nonlinear 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 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, we did not concern ourselves with any form of consistency other than the model selection consistency. However, other forms of consistency are also relevant for the class of models presented here. In particular, we mention the general risk consistency and the ℓ_2 consistency of the penalized estimators for generalized linear model and logistic regression models by van de Geer [35,36] and Meier, van der Geer and Bühlmann [20], respectively, where non-asymptotic bounds and oracle inequalities are available. Finally, a rather general framework for proving

norm consistency of penalized maximum likelihood estimators under decomposable regularizers which may be applicable to our problem is presented in Negahban et al. [24].

Finally, in our problem we do not need to worry about random design. In fact, 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.

6. Proofs

Proof of Lemma 2.1. We only provide a sketch of the proof and refer to Haberman [17], page 11, and Rinaldo [30], Lemma 2.2, for more details. It is possible to show that the linear subspace $\mathcal{M} \cap \mathcal{R}(\mathbf{1})^{\perp}$ and $\widetilde{\mathcal{M}}$ are homeomorphic sets, and the one-to-one mapping between $\widetilde{\boldsymbol{\mu}} \in \widetilde{\mathcal{M}}$ and $\boldsymbol{\beta} \in \mathcal{M} \cap \mathcal{R}(\mathbf{1})^{\perp}$ is given by

$$\widetilde{\boldsymbol{\mu}} = \boldsymbol{\beta} + \mathbf{1} \log\left(\frac{N}{\langle \exp(\boldsymbol{\beta}), \mathbf{1} \rangle} \right).$$

Furthermore, for any $\mathbf{n}$, some algebra shows that

$$\langle \mathbf{n}, \boldsymbol{\beta} \rangle - N \log\langle \exp(\boldsymbol{\beta}), \mathbf{1} \rangle + \log N! - \sum_{i \in \mathcal{I}} \log \mathbf{n}_i ! = \ell^*(\widetilde{\boldsymbol{\mu}}),$$

for each pair of homeomorphic points $\boldsymbol{\beta}$ and $\widetilde{\boldsymbol{\mu}}$. Then, for any full-rank matrix U with $\mathcal{M} \cap \mathcal{R}(\mathbf{1})^{\perp}$ as its column span, (1) and (2) both follow from the previous displays by noting that there exists also a one-to-one correspondence between $\mathcal{M} \cap \mathcal{R}(\mathbf{1})^{\perp}$ and $\mathbb{R}^{I-1}$, given by $\boldsymbol{\beta} = \mathrm{U}\theta$.

□

Proof of Lemma 3.1. The first order optimality conditions for a vector $\theta \in \mathbb{R}^{I-1}$ is $\mathbf{0} \in \partial P_{\Lambda}(\theta)$, the subdifferential set of $P_{\Lambda}(\theta)$. The gradient of ℓ at a point $\theta \in \mathbb{R}^{I-1}$ is

$$\nabla \ell(\theta) = \mathrm{U}^{\top} \left(\mathbf{n} - \left(\frac{N}{\langle \mathbf{b}, \mathbf{1} \rangle} \right) \mathbf{b} \right) = \mathrm{U}^{\top} (\mathbf{n} - \mathbf{m}), \tag{24}$$

where $\mathbf{b} = \exp(\mathrm{U}\theta)$. As for the penalty term, which is not differentiable when some of the blocks are zero, standard subgradient calculus (see, e.g., Bertsekas [2]) yields that for any $\theta \in \mathbb{R}^{I-1}$, the subdifferential of the function $x \mapsto \sum_h \lambda_h \|x_h\|$ at θ is a subset of $\mathbb{R}^{I-1}$ comprised by vectors whose h-block component is

$$\begin{cases} \lambda_h \dfrac{\theta_h}{\|\theta_h\|_2}, & \text{if } \theta_h \neq 0, \\ \lambda_h z_h, & \text{if } \theta_h = 0, \end{cases} \tag{25}$$

where $\|z_h\| \leq 1$ for each h such that $\theta_h = 0$. Equations (24) and (25) imply (11).

As for uniqueness, we follow the proof of Lemma 2 in Wainwright [37]. Suppose $\widehat{\theta}$ is an optimal solution to (10). Then, by duality theory, given a subgradient $\widehat{\eta} \in \mathbb{R}^{I-1}$ any optimal solution $\breve{\theta}$ must satisfy $\widehat{\eta}^\top \breve{\theta} = \sum_h \lambda_h \|\breve{\theta}_h\|$. This holds only if $\breve{\theta}_h = 0$ for all h for which $\|\widehat{\eta}_h\| < \lambda_h$. Thus, if there exists a solution $\breve{\theta}$ to the problem (10) different than $\widehat{\theta}$, it must satisfy $\breve{\theta}_h = 0$ for all h such that $\widehat{\theta}_h = 0$. Finally, uniqueness follows since ℓ is strictly concave, though not necessarily strongly concave. □

Proof of Lemma 4.1. With some abuse of notation we write $O_{P_n^0}$ and $o_{P_n^0}$ to refer to probabilistic statements for the sequence of probability distributions indexed by $\{\theta^0_{\mathcal{H}_n}\}_n$, with $\theta^0_{\mathcal{H}_n} \in \mathbb{R}^{d_{\mathcal{H}_n}}$ for each n. We will first analyze the asymptotic behavior of $\ell(\theta^n_{\mathcal{H}_n}) - \ell_n(\theta^0_{\mathcal{H}_n})$, uniformly over sequences of the form $\theta^n_{\mathcal{H}_n} = \theta^0_{\mathcal{H}_n} + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n$ with $\|x_n\| \leq D$, for all n and some $D > 0$. To this end, we follow the arguments used in Portnoy [26], Theorem 2.4. First off, notice that we can write

$$\mathbf{n} = \sum_{j=1}^{N_n} X_j,$$

where $X_1, \ldots, X_{N_n}$ are i.i.d. vectors in $\mathbb{R}^{\mathcal{I}}$ distributed like a Multinomial$(1, \boldsymbol{\pi}^0_n)$, with

$$\boldsymbol{\pi}^0_n = \frac{\exp(\mathrm{U}_{\mathcal{H}_n}\theta^0_{\mathcal{H}_n})}{\langle \exp(\mathrm{U}_{\mathcal{H}_n}\theta^0_{\mathcal{H}_n}), \mathbf{1}\rangle}.$$

By a Taylor series expansion, the term $\ell_n(\theta^0_{\mathcal{H}_n} + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n) - \ell_n(\theta^0_{\mathcal{H}_n})$ is equal to

$$\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n^\top \mathrm{U}_{\mathcal{H}_n}^\top (\mathbf{n}_n - \mathbf{m}^0_n) - \frac{1}{2}\frac{d_{\mathcal{H}_n}}{N_n} x_n^\top \Sigma_{\mathcal{H}_n} x_n + \frac{N_n}{6}\left(\frac{d_{\mathcal{H}_n}}{N_n}\right)^{3/2} \mathbb{E}_{\theta^*_n}[(\langle x_n, \mathrm{U}_{\mathcal{H}_n}^\top X_1\rangle)^3], \tag{26}$$

where θ^*_n is on the line joining $\theta^n_{\mathcal{H}_n}$ and $\theta^0_{\mathcal{H}_n}$. For the first term in (26), we have

$$\begin{aligned}\mathbb{E}\|\mathrm{U}_{\mathcal{H}_n}^\top(\mathbf{n}_n - \mathbf{m}^0_n)\|_2^2 &= \mathrm{tr}\left(\mathrm{U}_{\mathcal{H}_n}^\top\left(\mathrm{D}_{\mathbf{m}^0_n} - \frac{\mathbf{m}^0_n(\mathbf{m}^0_n)^\top}{N_n}\right)\mathrm{U}_{\mathcal{H}_n}\right)\\ &\leq N_n l_n^{\max} d_{\mathcal{H}_n}.\end{aligned}$$

Thus, by Markov and Cauchy–Schwarz inequalities,

$$\left|\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}} x_n^\top \mathrm{U}_{\mathcal{H}_n}^\top (\mathbf{n}_n - \mathbf{m}^0_n)\right| = O_{P_n^0}\big(d_{\mathcal{H}_n}\sqrt{l_n^{\max}}\|x_n\|\big) = O_{P_n^0}(d_{\mathcal{H}_n}\|x_n\|), \tag{27}$$

where the last identity follows from the eigenvalue assumption [NC.2]. Next, using the fact that the Fisher information matrix is positive definite for each n, and once again [NC.2],

$$-\frac{1}{2}\frac{d_{\mathcal{H}_n}}{N_n}x_n^\top \Sigma_{\mathcal{H}_n} x_n \leq -\frac{1}{2}\frac{d_{\mathcal{H}_n}}{N_n} N_n \|x_n\|^2 l^n_{\min} x \leq -\frac{1}{2} O(d_{\mathcal{H}_n}\|x_n\|^2), \tag{28}$$

which bounds the second term (26). Finally, the assumption [NC.3] yields (see Portnoy [26], Theorem 2.4)

$$\frac{N_n}{6}\left(\frac{d_{\mathcal{H}_n}}{N_n}\right)^{3/2} \mathbb{E}_{\theta_n^*}\left[\left|\left\langle \frac{x_n}{\|x_n\|}, \mathrm{U}_{\mathcal{H}_n}^\top X_1\right\rangle\right|^3\right] = O(d_{\mathcal{H}_n}\|x_n\|).$$

Combining the previous display with (27), (28) and with (26), we obtain, by choosing D large enough that, for each $\varepsilon > 0$, and all n large enough

$$\mathbb{P}\left(\sup_{\{x_n, \|x_n\|=C\}} \ell_n\left(\theta_n^0 + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}x_n\right) - \ell_n(\theta_n^0) < 0\right) > 1-\varepsilon.$$

The strict concavity of ℓ (warranted by [NC.2]) further guarantees that, for all n large enough, with probability tending to one, there are no other maximizers of ℓ outside the ball $\{\theta^n_{\mathcal{H}_n}\colon \theta^n_{\mathcal{H}_n} = \theta^0_{\mathcal{H}_n} + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}x_n\}$.

Next, we consider the difference in the penalty terms between $\theta^0_{\mathcal{H}_n}$ and $\theta^n_{\mathcal{H}_n} \equiv \theta^0_{\mathcal{H}_n} + \sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}x_n$. By a first order Taylor expansion,

$$N_n\lambda_n\left(\sum_{h\in\mathcal{H}_n}\lambda_h\|\theta_h^n\|_2 - \sum_{h\in\mathcal{H}_n}\lambda_h\|\theta_h^0\|_2\right) = N_n\lambda_n\sum_{h_n\in\mathcal{H}_n}\lambda_{h_n}\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}(x^*_{h_n})^\top\frac{\theta^0_{h_n}}{\|\theta^0_{h_n}\|},$$

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

$$\lambda_n\sqrt{N_n d_{\mathcal{H}_n}}\|x_n\|\left(\sum_{h\in\mathcal{H}_n}\lambda_h\right).$$

Under the assumed conditions, we see that $\|\widetilde{\theta}_n - \theta^0_{\mathcal{H}_n}\| = O_{P_n^0}(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}})$, as required. □

Proof of Theorem 4.2. We follow the primal-dual witness method of Wainwright [37] as described in Section 4.1. Let $\widetilde{\theta}_n \in \mathbb{R}^{d_{\mathcal{H}_n}}$ be the restricted group-lasso estimator (17) and define the vector $\hat{\theta}_n \in \mathbb{R}^{I-1}$, whose h_n block is given by

$$\hat{\theta}_{h_n} = \begin{cases} \widetilde{\theta}_{h_n}, & \text{if } h_n \in \mathcal{H}_n, \\ 0, & \text{otherwise.} \end{cases}$$

Then, by construction, $\widehat{\theta}_{\mathcal{H}_n} = \widetilde{\theta}_n$. Set also

$$\widehat{\mathbf{m}}_n = N_n \frac{\exp(\mathrm{U}\widehat{\theta}_n)}{\langle \exp(\mathrm{U}\widehat{\theta}_n), \mathbf{1}\rangle} = N_n \frac{\exp(\mathrm{U}_{\mathcal{H}_n}\widetilde{\theta}_n)}{\langle exp(\mathrm{U}_{\mathcal{H}_n}\widetilde{\theta}_n), \mathbf{1}\rangle}.$$

Next, consider the random vector $\widehat{\eta} \in \mathbb{R}^{I-1} = \mathrm{vec}(\widehat{\eta}_{\mathcal{H}_n}, \widehat{\eta}_{\mathcal{H}_n^c})$, where

$$\widehat{\eta}_{\mathcal{H}_n} = \mathrm{vec}\left\{ \lambda_{h_n} \frac{\widehat{\theta}_{h_n}}{\|\widehat{\theta}_{h_n}\|}, h_n \in \mathcal{H}_n \right\}$$

and

$$\widehat{\eta}_{\mathcal{H}_n^c} = \mathrm{vec}\{\lambda_{w_n} \widehat{z}_{w_n}, w_n \in \mathcal{H}_n^c\}, \tag{29}$$

with the sub-vectors $\{\widehat{z}_{w_n}, w_n \in \mathcal{H}_n^c\}$ to be chosen in an appropriate way as described below. Notice also that $\widehat{\eta}_{\mathcal{H}_n}$ belongs to the subdifferential of $\lambda_n \sum_{h,h\in\mathcal{H}_n} \lambda_h \|x_h\|$ evaluated at $\widetilde{\theta}_n$.

The pair $(\widehat{\theta}_{\mathcal{H}_n}, \widehat{\eta}_{\mathcal{H}_n})$ must satisfy the optimality conditions (11) for the blocks indexed by $h_n \in \mathcal{H}_n$. Using this conditions along with a Taylor expansion of $\widehat{\mathbf{m}}_n$ around $\mathbf{m}_n^0$, we obtain the expression

$$\widehat{\theta}_{\mathcal{H}_n} = \theta^0_{\mathcal{H}_n} + N_n \Sigma^{-1}_{\mathcal{H}_n} \left(\frac{1}{N_n} \mathrm{U}^\top_{\mathcal{H}_n} (\mathbf{n}_n - \mathbf{m}_n^0) - \frac{1}{N_n} \mathrm{U}^\top_{\mathcal{H}_n} R_n - \lambda_n \widehat{\eta}_{\mathcal{H}_n} \right). \tag{30}$$

Employing a similar strategy, we now consider the optimality conditions (11) for the remaining blocks indexed by $w_n \in \mathcal{H}_n^c$ and solve for $\{\widehat{z}_{w_n}, w_n \in \mathcal{H}_n^c\}$ in terms of $\widehat{\theta}_{\mathcal{H}_n} - \theta^0_{\mathcal{H}_n}$. Eventually, we are led to the expression

$$\begin{aligned} \lambda_n \widehat{\eta}_{\mathcal{H}_n^c} = {} & \frac{1}{N_n} \mathrm{U}^\top_{\mathcal{H}_n^c} (\mathbf{n}_n - \mathbf{m}_n^0) - \frac{1}{N_n} \mathrm{U}^\top_{\mathcal{H}_n^c} R_n \\ & - \mathrm{W}_n \left(\frac{1}{N_n} \mathrm{U}^\top_{\mathcal{H}_n} (\mathbf{n}_n - \mathbf{m}_n^0) - \frac{1}{N_n} \mathrm{U}^\top_{\mathcal{H}_n} R_n - \lambda_n \widehat{\eta}_{\mathcal{H}_n} \right), \end{aligned} \tag{31}$$

where $\|R_n\| = o_{P_n^0}(\|\widehat{\theta}_{\mathcal{H}_n} - \theta^0_{\mathcal{H}_n}\|)$, and

$$\mathrm{W}_n = \mathrm{U}^\top_{\mathcal{H}_n^c} \left(D_{\mathbf{m}_n^0} - \frac{\mathbf{m}_n^0 (\mathbf{m}_n^0)^\top}{N_n} \right) \mathrm{U}_{\mathcal{H}_n} \Sigma^{-1}_{\mathcal{H}_n}.$$

Notice that, by Lemma 4.1, $\|R_n\| = o_{P_0^n}(1)$, so the remainder term in the above Taylor expansion is negligible.

We then rely on equations (30) and (31) to show that the assumed conditions are sufficient to guarantee that

$$\lim_n \mathbb{P}(\|\widehat{\theta}_{h_n}\| > 0, \forall h_n \in \mathcal{H}_n) = 1 \tag{32}$$

and

$$\lim_n \mathbb{P}\Big(\max_{w_n \in \mathcal{H}_n^c} \|\widehat{z}_{w_n}\| \le 1 \Big) = 1. \tag{33}$$

Because (32) is equivalent to (15) and (33) is equivalent to (16), model selection will follow. We will deal with equations (32) and (33) separately.

Proof of equation (32). It is enough to show that

$$\mathbb{P}\left(\left\| N_n \Sigma_{\mathcal{H}_n}^{-1}\left(\frac{1}{N_n}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0) - \frac{1}{N_n}\mathrm{U}_{\mathcal{H}_n}^{\top} R_n - \lambda_n \widehat{\eta}_{\mathcal{H}_n}\right)\right\| \leq \alpha_n\right) \to 1, \tag{34}$$

where $\alpha_n = \min_{h_n \in \mathcal{H}_n} \|\theta_{h_n}^0\|$. 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 $\|\theta_{h_n}^0\|$, $\forall h_n \in \mathcal{H}_n$, which, by the triangle inequality, will produce the desired result.

First, we consider the term

$$\Sigma_{\mathcal{H}_n}^{-1}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0).$$

The vector $\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)$ has mean zero and covariance matrix $\Sigma_{\mathcal{H}_n}$. Furthermore, because of [NC.2], letting $\gamma_{\mathcal{H}_n}^{\min} = \lambda_{\min}(\Sigma_{\mathcal{H}_n})$, we have

$$\gamma_{\mathcal{H}_n}^{\min}\frac{1}{N_n} \geq D_{\min} > 0 \qquad \text{for all } n. \tag{35}$$

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

$$\mathbb{E}\|\Sigma_{\mathcal{H}_n}^{-1}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)\|_2^2 = \operatorname{tr}\Sigma_{\mathcal{H}_n}^{-1} \leq \frac{d_{\mathcal{H}_n}}{\gamma_{\mathcal{H}_n}^{\min}} \leq \frac{d_{\mathcal{H}_n}}{D_{\min}N_n},$$

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

$$\|\Sigma_{\mathcal{H}_n}^{-1}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)\| = O_{P_n^0}\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right). \tag{36}$$

Next, using (35) for the operator norm of of $\Sigma_{\mathcal{H}_n}$, we get the upper bound

$$\|N_n \Sigma_{\mathcal{H}_n}^{-1}\lambda_n \widehat{\eta}_{\mathcal{H}_n}\| \leq \frac{1}{D_{\min}}\lambda_n\sqrt{\sum_{h_n \in H}\lambda_{h_n}^2}, \tag{37}$$

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

Finally, the norm of

$$\Sigma_{\mathcal{H}_n}^{-1}\mathrm{U}_{\mathcal{H}_n}^{\top} R_n$$

is no larger than

$$\frac{1}{\sqrt{D_{\min}N_n}}\sqrt{d_{\mathcal{H}_n}}\,o(\|\widehat{\theta}_{\mathcal{H}_n} - \theta_{\mathcal{H}_n}^0\|) = o_{P_n^0}\left(\frac{d_{\mathcal{H}_n}}{N_n}\right), \tag{38}$$

because $\|\mathrm{U}_{\mathcal{H}_n}\| \leq \sqrt{d_{\mathcal{H}_n}}$.

Using equations (36), (37) and (38), condition (34) is satisfied if MSC.1 holds.

Proof of equation (33). In equation (31) write $\widehat{\eta}_{\mathcal{H}_n^c} = \Lambda_{\mathcal{H}_n^c}\widehat{z}_{\mathcal{H}_n^c}$, where $\Lambda_{\mathcal{H}_n^c}$ is a $d_{\mathcal{H}_n^c}$-dimensional diagonal matrix whose diagonal is $\mathrm{vec}\{\mathbf{1}_{w_n}\lambda_{w_n}, w_n \in \mathcal{H}_n^c\}$, with $\mathbf{1}_{h_n}$ denoting the d_{h_n}-dimensional vector with entries all equal to 1. Then, (31) becomes

$$\begin{aligned}\widehat{z}_{\mathcal{H}_n^c} = {} & \frac{1}{N_n\lambda_n}\Lambda_{\mathcal{H}_n^c}^{-1}\mathrm{U}_{\mathcal{H}_n^c}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0) - \frac{1}{N_n\lambda_n}\Lambda_{\mathcal{H}_n^c}^{-1}\mathrm{U}_{\mathcal{H}_n^c}^{\top}R_n \\ & - \Lambda_{\mathcal{H}_n^c}^{-1}\mathrm{W}_n\left(\frac{1}{N_n\lambda_n}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0) - \frac{1}{N_n\lambda_n}\mathrm{U}_{\mathcal{H}_n}^{\top}R_n - \widehat{\eta}_{\mathcal{H}_n}\right).\end{aligned}$$

For any $w_n \in \mathcal{H}_n^c$, consider the corresponding block in the vector $\Lambda_{\mathcal{H}_n^c}^{-1}\mathrm{W}_n\widehat{\eta}_{\mathcal{H}_n}$, that is, the vector

$$\frac{1}{\lambda_{w_n}}\mathrm{U}_{w_n}^{\top}\Sigma_n^0\mathrm{U}_{\mathcal{H}_n}(\mathrm{U}_{\mathcal{H}_n}^{\top}\Sigma_n^0\mathrm{U}_{\mathcal{H}_n})^{-1}\widehat{\eta}_{\mathcal{H}_n}. \tag{39}$$

Because of assumption [MSC.2], the Euclidian norm of (39), for any choice of $w_n \in \mathcal{H}_n^c$, is bounded by

$$\frac{(1-\varepsilon)}{|\mathcal{H}_n^c|}\frac{\sum_{h_n\in\mathcal{H}_n}\lambda_{h_n}}{\min_{w_n\in\mathcal{H}_n^c}\lambda_{w_n}},$$

which, in turn, is smaller that

$$(1-\varepsilon)\frac{|\mathcal{H}_n|\max_{h_n\in\mathcal{H}_n}\lambda_{h_n}}{|\mathcal{H}_n^c|\min_{w_n\in\mathcal{H}_n^c}\lambda_{w_n}}.$$

Then, under MSC.3 (39) will be eventually less than $(1-\varepsilon)$, uniformly over $w_n \in \mathcal{H}_n^c$.

Next, for $w_n \in \mathcal{H}_n^c$, we consider the vector

$$\frac{1}{N_n\lambda_n\lambda_{w_n}}[\mathrm{U}_{w_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0) - \mathrm{W}_{w_n}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)]. \tag{40}$$

The covariance matrix of the term inside the parenthesis is

$$\mathrm{U}_{w_n}^{\top}(\Sigma_n^0)^{1/2}[I_{d_{\mathcal{H}_n}} - (\Sigma_n^0)^{1/2}\mathrm{U}_{\mathcal{H}_n}(\mathrm{U}_{\mathcal{H}_n}^{\top}\Sigma_n^0\mathrm{U}_{\mathcal{H}_n})^{-1}\mathrm{U}_{\mathcal{H}_n}^{\top}(\Sigma_n^0)^{1/2}](\Sigma_n^0)^{1/2}\mathrm{U}_{w_n}, \tag{41}$$

where

$$\Sigma_n^0 = D_{\mathbf{m}_n^0} - \frac{\mathbf{m}_n^0(\mathbf{m}_n^0)^{\top}}{N_n}.$$

Since the largest eigenvalue of the matrix in (41) is smaller than the largest eigenvalue of the covariance matrix of $\mathrm{U}_{w_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)$, by Chebyshev's inequality it is enough to show that the ℓ_2 norm of

$$\frac{1}{N_n\lambda_n\lambda_{w_n}}\mathrm{U}_{w_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)$$

vanishes in order to conclude that (40) has vanishing ℓ_2 norm as well. To this end, notice that

$$\frac{1}{N_n\lambda_n\lambda_{w_n}}\|\mathrm{U}_{w_n}^\top(\mathbf{n}_n-\mathbf{m}_n^0)\| \le \frac{\sqrt{d_{w_n}}}{N_n\lambda_n\lambda_{w_n}}\|\mathrm{U}_{w_n}^\top(\mathbf{n}_n-\mathbf{m}_n^0)\|_\infty$$
$$\le \frac{\sqrt{d_{w_n}}}{N_n\lambda_n\lambda_{w_n}}\|\mathrm{U}_{\mathcal{H}_n^c}^\top(\mathbf{n}_n-\mathbf{m}_n^0)\|_\infty.$$

Next, write

$$\mathrm{U}_{\mathcal{H}_n^c}^\top\frac{(\mathbf{n}_n-\mathbf{m}_n^0)}{N_n} = \sum_{j_n=1}^{N_n}\mathrm{U}_{\mathcal{H}_n^c}^\top\frac{(X_{j_n}-\boldsymbol{\pi}_n^0)}{N_n},$$

where the vectors $X_{j_n}, 1\le j_n\le N_n$ are i.i.d. Multinomial$(1,\boldsymbol{\pi}_n^0)$. Since the entries of $\mathrm{U}_{\mathcal{H}_n^c}$ are all -1, 0 or 1, by Bernstein's inequality followed by a union bound, we get

$$\mathbb{P}\left\{\left\|\frac{1}{N_n}\mathrm{U}_{\mathcal{H}_n^c}^\top(\mathbf{n}_n-\mathbf{m}_n^0)\right\|_\infty > c\frac{\lambda_n\lambda_{w_n}}{\sqrt{d_{w_n}}}\right\} \le 2\exp\left\{-\frac{N_nc^2\lambda_n^2\lambda_{w_n}^2/d_{w_n}}{1/8+(2/3)c\lambda_n\lambda_{w_n}/\sqrt{d_{w_n}}}+\log d_{\mathcal{H}_n^c}\right\},$$

which vanishes under [MSC.4]. As for the terms involving R_n, following the arguments used above, it is easy to see that they both converge in probability to 0, so that (33) holds true. □

Proof of Theorem 4.3. All the claims in the proof are made on the event $\mathcal{O}_n$. Because the norm consistency assumptions are in force, $\mathcal{O}_n$ occurs in probability and, therefore, our claims hold true within a set or probability converging to 1. In particular, $\|\widehat{\theta}_{\mathcal{H}_n}-\theta^0_{\mathcal{H}_n}\| = O_{P_n^0}(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}})(1+o_{P_n^0}(1)) = O_{P_n^0}(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}})$. Reorganize equation (30) as

$$\Sigma_{\mathcal{H}_n}^{1/2}(\widehat{\theta}_{\mathcal{H}_n}-\theta^0_{\mathcal{H}_n}) = \Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^\top(\mathbf{n}_n-\mathbf{m}_n^0) - \Sigma_{\mathcal{H}_n}^{-1/2}N_n\lambda_n\widehat{\eta}_{\mathcal{H}_n} - \Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^\top R_n. \tag{42}$$

By similar arguments used in the proof of Theorem 4.2, the term

$$\Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^\top R_n$$

is of order

$$\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}o_{P_n^0}(\|\widehat{\theta}_n-\theta_n^0\|) = o_{P_n^0}\left(\frac{d_{\mathcal{H}_n}}{N_n}\right),$$

and therefore converges in probability to 0.

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

$$\Sigma_{\mathcal{H}_n}^{-1/2}N_n\lambda_n\widehat{\eta}_{\mathcal{H}_n} = \Sigma_{\mathcal{H}_n}^{-1/2}N_n\lambda_n\left(\eta^0_{\mathcal{H}_n}+\mathrm{J}^0_{\mathcal{H}_n}(\widehat{\theta}_{\mathcal{H}_n}-\theta^0_{\mathcal{H}_n})+o_{P_n^0}\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\right)\right). \tag{43}$$

The remainder term in equation (43) is of order

$$\lambda_n o_{P_n^0}(\sqrt{d_{\mathcal{H}_n}}),$$

which become negligible for $\lambda_n = O(\frac{1}{\sqrt{d_{\mathcal{H}_n}}})$ (obviously, $\lambda_n = O(1/\sqrt{I_n})$ will do). Then using (42), we obtain

$$\Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0) = \Sigma_{\mathcal{H}_n}^{-1/2}\big((\Sigma_{\mathcal{H}_n} + N_n\lambda_n J_{\mathcal{H}_n}^0)(\widehat{\theta}_n - \theta_n^0) + N_n\lambda_n\eta_{\mathcal{H}_n}^0\big) + o_{P_n^0}(1). \quad (44)$$

Thus, we only need to consider the term $\Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0)$. For $1 \le j_n \le N_n$, let

$$Y_{j_n} = \frac{1}{\sqrt{N_n}}\mathrm{F}_n^{-1/2}\mathrm{U}_{\mathcal{H}_n}^{\top}(X_{j_n} - \boldsymbol{\pi}_n^0),$$

where the variables X_{j_n} are i.i.d. Multinomials with size 1 and probability vector $\boldsymbol{\pi}_n^0$. Then,

$$\Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^{\top}(\mathbf{n}_n - \mathbf{m}_n^0) = \sum_{j_n} Y_{j_n},$$

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

To show the result in part 1 it is sufficient to show that, under assumption [CLT.LF], the multivariate Lindberg–Feller conditions hold, namely

$$\sum_{j_n}\mathbb{E}_{\theta_{\mathcal{H}_n}^0}\|\mathrm{G}_n Y_{j_n}\|^2 I_{\{\|\mathrm{G}_n Y_{j_n}\|\ge\varepsilon\}} \to 0$$

as $n \to \infty$, where $I_{\{\cdot\}}$ denotes the indicator function. The proof is quite standard (see also the proof of Theorem 2 in Fan and Peng [11]) and we only sketch it.

Because the vectors Y_{j_n}'s are identically distributed, and invoking the Cauchy–Schwarz inequality, it is sufficient to show that

$$N_n\big(\mathbb{E}\|\mathrm{G}_n Y_{j_n}\|^4\mathbb{P}(\|\mathrm{G}_n Y_{j_n}\| \ge \varepsilon)\big)^{1/2} \to 0. \quad (45)$$

By Chebychev inequality, for fixed $\varepsilon > 0$,

$$\mathbb{P}(\|\mathrm{G}_n Y_{j_n}\| \ge \varepsilon) \le \frac{\mathrm{tr}(\mathrm{G}_n\mathrm{G}_n^{\top})}{\varepsilon^2 N_n} = O\left(\frac{1}{N_n}\right),$$

where $\mathrm{tr}(\mathrm{G}_n\mathrm{G}_n^{\top}) = O(1)$ because of (22). Similarly, using the fact that the minimal eigenvalue of F_n is bounded away from zero,

$$\mathbb{E}\|\mathrm{G}_n Y_{j_n}\|^4 \le O\left(\frac{1}{N_n^2}\right)\mathbb{E}\|\mathrm{U}_{\mathcal{H}_n}^{\top}(X_{j_n} - \boldsymbol{\pi}_n^0)\|^4 = O\left(\frac{d_{\mathcal{H}_n}^2}{N_n^2}\right),$$

where in the last step we use the fact that the entries of $\mathrm{U}_{\mathcal{H}_n}^{\top}(X_{j_n} - \boldsymbol{\pi}_n^0)$ are bounded, uniformly over n. Combining the last two displays, the left-hand side of (45) is of order

$$N_n O\left(\sqrt{\frac{d_{\mathcal{H}_n}}{N_n}}\frac{1}{N_n}\right) = O\left(\frac{d_{\mathcal{H}_n}}{N_n^{1/2}}\right),$$

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

Next, we prove the result of part 1 under both [CLT.Ma] and [CLT.Mb]. We relax the assumption [CLT.LF] by allowing the dimension of the parameter space to grow fatser. To this end, we derive multi-dimensional analogs of Lemmas 2.1 and 2.2 and Theorem 2.1 in Morris [22]. In particular, our proof follows closely the proof of Morris [22], Lemma 2.2. We first obtain joint limit law by using Lemma 6.2, and then establish the conditional limit law by using a multi-dimensional version of condition (2.9) in Morris [22]. Note that the result in Steck [34] about conditional limit laws is actually a multi-dimensional one, but somehow was formulated in Morris [22], Theorem 2.1, as one-dimensional. The conditional law we are interested is the distribution of Z_n, defined below in (48).

Let $\gamma_n = N_n^{-1}\mathrm{G}_n\Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^{\top}\mathbf{m}_n^0$, and set $\mathrm{A}_n = \mathrm{G}_n\Sigma_{\mathcal{H}_n}^{-1/2}\mathrm{U}_{\mathcal{H}_n}^{\top}$. Note that $\mathbf{m}_n^0 = N_n\boldsymbol{\pi}_n^0$, thus $\gamma_n = \mathrm{A}_n\boldsymbol{\pi}_n^0$. Denote the ith column of A_n by a_i, $i = 1, \ldots, I_n$. Then, the left-hand side of (44), premultiplied by G_n, can be written as

$$Z_n = \sum_{i \in \mathcal{I}_n} f_i(n_i),$$

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 $m_i^0 = N_n\boldsymbol{\pi}_i^0$, so that $\mathbb{E}f_i(X_i) = 0$ and $\sum_i \mathrm{cov}(f_i(X_i), X_i) = 0$, by construction. Next, define

$$V_n = N_n^{-1/2}\sum_i (X_i - m_i^0), \tag{46}$$

$$U_n = \Xi_n^{-1/2}\sum_i f_i(X_i), \tag{47}$$

where $\Xi_n = \sum_i \mathrm{cov}(f_i(X_i))$. A simple calculation though shows that $\Xi_n = \mathrm{G}_n\mathrm{G}_n^{\top}$, 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

$$\mathcal{L}(\Xi_n^{-1/2}Z_n) = \mathcal{L}(U_n | V_n = 0), \tag{48}$$

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 [22], 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 $\mathbb{E}U_n = 0$, $\mathrm{cov}(U_n) = \mathrm{I}_k$, the identity matrix of dimensions $k \times k$, and, by construction, $\mathrm{cov}(V_n, U_n) = 0$. We argue below that U_n satisfies the

multi-dimensional Lindeberg condition. By Lemma (6.2), this will imply the asymptotic normality of the joint limit law of (V_n, U_n).

By Schwartz inequality, for any $\varepsilon > 0$,

$$\sum_i \mathbb{E}[\|f_i(X_i)\|^2; \|f_i(X_i)\| > \varepsilon] \leq \sum_i \big[\mathbb{E}\|f_i(X_i)\|^4 \mathbb{P}\big(\|f_i(X_i)\| > \varepsilon\big)\big]^{1/2}. \tag{49}$$

We will show that, for each $\varepsilon > 0$, the right-hand side of (49) tends to zero. Recall that $f_i(X_i) = (a_i - \gamma_n)(X_i - m_i^0)$. The length of γ_n can be bounded as follows, $\|\gamma_n\| \leq \|\mathrm{G}_n\| \|\Sigma_{\mathcal{H}_n}^{-1/2} \mathrm{U}_{\mathcal{H}_n}^\top \boldsymbol{\pi}_n^0\| \leq O(1) D_{\min} N_n^{-1/2} \|\mathrm{U}_{\mathcal{H}_n}^\top \boldsymbol{\pi}_n^0\|$. Elements of $\mathrm{U}_{\mathcal{H}_n}^\top \boldsymbol{\pi}_n^0$ are absolutely bounded by a constant D_1, thus $\|\gamma_n\| \leq D N_n^{-1/2} d_{\mathcal{H}_n}^{1/2}$. Similarly, $\|a_i\| = \|\mathrm{A}_n e_i\| \leq D N_n^{-1/2} d_{\mathcal{H}_n}^{1/2}$, where e_i is the standard unit vector in $\mathbb{R}^{I_n}$ with ith coordinate equal to 1. Adding up, $\|a_i - \gamma_n\| = O(N_n^{-1/2} d_{\mathcal{H}_n}^{1/2})$, 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 [5] and based on a modified logarithmic Sobolev inequality:

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

$$\mathbb{P}\big(h(X) - \mathbb{E}h(X) \geq b\big) \leq \exp\left\{-\frac{b}{4} \log\left(1 + \frac{b}{2\lambda}\right)\right\}, \tag{50}$$

for all $b \geq 0$.

Then, using Theorem 50, for some constant D,

$$\begin{aligned}
&\mathbb{P}(X_i - m_i^0 \geq \varepsilon \|a_i - \gamma_n\|^{-1}) \\
&\quad \leq \exp\left\{-\frac{\varepsilon}{4} \|a_i - \gamma_n\|^{-1} \log\left(1 + \frac{1}{2} \frac{\varepsilon}{m_i^0 \|a_i - \gamma_n\|}\right)\right\} \\
&\quad \leq \exp\left\{-\varepsilon D N_n^{1/2} d_{\mathcal{H}_n}^{-1/2} \log\left(1 + \varepsilon D \frac{1}{\sqrt{N_n d_{\mathcal{H}_n}} \max_i \boldsymbol{\pi}_i^0}\right)\right\} \\
&\quad = \exp\big(-O\big(\sqrt{N_n / d_{\mathcal{H}_n}}\big)\big),
\end{aligned} \tag{51}$$

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, $\sum_i (\mathbb{E}\|f_i(X_i)\|^4)^{1/2} = \sum_i \|a_i - \gamma_n\|^2 (m_i^0 + 3(m_i^0)^2)^{1/2}$, which is of the order of magnitude of $O(d_{\mathcal{H}_n})$. This, together with (49) and (51) 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 [22] holds true also for multi-dimensional variables. We only need to replace condition (2.9)

in Morris [22] by a multi-dimensional version. Specifically, we show that

$$\lim_{r\to 0}\sup_{n}\sup_{v}\mathbb{E}\left\|\sum_{i}[f_i(L_i+M_i)-f_i(L_i)]\right\|^2=0, \tag{52}$$

where $L_n=(L_1,\ldots,L_{I_n})$ and $M_n=(M_1,\ldots,M_{I_n})$ are Multinomial random variables with probability vector $\boldsymbol{\pi}_n^0$, 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 [22], Lemma 2.2. Notice that $f_i(L_i+M_i)-f_i(L_i)=(a_i-\gamma_n)M_i$. Thus,

$$\left\|\sum_{i}(a_i-\gamma_n)M_i\right\|^2=\|\mathrm{A}_nM_n-rN_n^{-1/2}\mathrm{A}_n\mathbf{m}_n^0\|=(M_n-\mathbb{E}M_n)^\top B_n(M_n-\mathbb{E}M_n),$$

where $\mathbb{E}M_n=rN_n^{1/2}\boldsymbol{\pi}^0$, and $\mathrm{B}_n=\mathrm{A}_n^\top\mathrm{A}_n$. Taking expectation yields

$$\begin{aligned}\mathbb{E}(M_n-\mathbb{E}M_n)^\top B_n(M_n-\mathbb{E}M_n)&=r\sqrt{N_n}\,\mathrm{tr}\big(\mathrm{B}_n\big(D_{\boldsymbol{\pi}_n^0}-\boldsymbol{\pi}_n^0(\boldsymbol{\pi}_n^0)^\top\big)\big)\\&=r\frac{1}{\sqrt{N_n}}\,\mathrm{tr}\left(\mathrm{B}_n\left(D_{\mathbf{m}_n^0}-\frac{\mathbf{m}_n^0(\mathbf{m}_n^0)^\top}{N_n}\right)\right)\\&=\mathrm{O}(1)r\frac{1}{\sqrt{N_n}},\end{aligned}$$

since

$$\mathrm{tr}\left(\mathrm{B}_n\left(D_{\mathbf{m}_n^0}-\frac{\mathbf{m}_n^0(\mathbf{m}_n^0)^\top}{N_n}\right)\right)=\mathrm{tr}(\mathrm{G}_n\mathrm{G}_n^\top)=O(1).$$

Therefore,

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

which shows that condition (52) holds, and the statement in part 1 is proved.

Part 2 of the theorem follows in a straightforward way from the main theorem in Bentkus [1] and the fact that $\mathbb{E}\|\mathrm{F}_n^{-1/2}\mathrm{U}_{\mathcal{H}_n}^\top(X_{j_n}-\boldsymbol{\pi}_n^0)\|^3$ is of order $O(d_{\mathcal{H}_n}^{3/2})$, by the same arguments used in the proof of part 1. □

The following lemma is a multivariate analog of Lemma 2.1. in Morris [22].

Lemma 6.2. *Let* $\mathbf{S}_k=(S_{1k},\mathbf{R}_k)=\sum_{i=1}^k\mathbf{X}_{ik}$, *where* $\mathbf{R}_k=(S_{2k},\ldots,S_{pk})$, $\mathbf{X}_{ik}=(X_{i1k},\mathbf{Y}_{ik})$, *and* $\mathbf{Y}_{ik}=(X_{i2k},\ldots,X_{ipk})$. *Suppose that* $\{\mathbf{X}_{ik}\}_{i=1}^k$ *are independent random vectors, with* $\mathbb{E}X_{i1k}=0,\mathbb{E}\mathbf{Y}_{ik}=\mathbf{0}$, *and* $\mathrm{Var}(\mathbf{S}_k)=I_p$, *the* $p\times p$ *identity matrix. Suppose* S_{1k} *satisfies the uan condition, i.e.,* $\max_{1\le i\le k}\mathrm{Var}\,X_{i1k}=o(1)$ *as* $k\to\infty$, *and that* $S_{1k}\xrightarrow{w}N(0,1)$. *Finally,*

suppose that $\mathbf{R}_k$ *satisfies the* (*multi-dimensional*) *Lindeberg condition*, *i.e.*, *for all* $\varepsilon > 0$,

$$\sum_{i=1}^{k} \mathbb{E}[\|\mathbf{Y}_{ik}\|^2; \|\mathbf{Y}_{ik}\|^2 > \varepsilon] = o(1) \qquad (k \to \infty).$$

Then $\mathbf{S}_k \xrightarrow{w} N_p(\mathbf{0}, I_p)$.

Proof. As in Morris' proof, S_{1k} satisfies the (one-dimensional) Lindeberg condition, i.e.,

$$\sum_{i=1}^{k} \mathbb{E}[X_{i1k}^2; X_{i1k}^2 > \varepsilon] = o(1) \qquad (k \to \infty).$$

Therefore,

$$\begin{aligned}\sum_{i=1}^{k} \mathbb{E}[\|\mathbf{X}_{ik}\|^2; \|\mathbf{X}_{ik}\|^2 \varepsilon] &= \sum_{i=1}^{k} \mathbb{E}[X_{i1k}^2 + \|\mathbf{Y}_{ik}\|^2; X_{i1k}^2 + \|\mathbf{Y}_{ik}\|^2 \varepsilon] \\ &\le 2 \sum_{i=1}^{k} \mathbb{E}[\max\{X_{i1k}^2, \|\mathbf{Y}_{ik}\|^2\}; \max\{X_{i1k}^2, \|\mathbf{Y}_{ik}\|^2\}\varepsilon/2] \\ &\le 2 \sum_{i=1}^{k} \mathbb{E}[X_{i1k}^2; X_{i1k}^2 > \varepsilon/2] + 2 \sum_{i=1}^{k} \mathbb{E}[\|\mathbf{Y}_{ik}\|^2; \|\mathbf{Y}_{ik}\|^2 > \varepsilon/2] = o(1).\end{aligned}$$

Thus, $\mathbf{S}_k$ satisfies the (multi-dimensional) Lindeberg condition and the proof is complete (see, e.g., Bhattacharya and Rao [3], pages 183–184). □

Acknowledgements

The authors thank Larry Wasserman for his valuable comments, and one anonymous reviewer and the associate editor for their suggestions, which greatly improved the exposition and the readability of the article. This research was supported in part by NSF Grant EIA-0131884 to the National Institute of Statistical Sciences, by NSF Grant DMS-06-31589, Army contract DAAD19-02-1-3-0389 and a Health Research Formula Fund Award granted by the Commonwealth of Pennsylvania's Department of Health.

References

[1] Bentkus, V. (2003). On the dependence of the Berry–Esseen bound on dimension. *J. Statist. Plann. Inference* **113** 385–402. MR1965117

[2] Bertsekas, D.P. (1995). *Nonlinear Programming*. Athena: Scientific.

[3] Bhattacharya, R.N. and Ranga Rao, R. (1976). *Normal Approximation and Asymptotic Expansions*. New York: Wiley. MR0436272
[4] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W. (2007). *Discrete Multivariate Analysis*: *Theory and Practice*. New York: Springer.
[5] Bobkov, S.G. and Ledoux, M. (1998). On modified logarithmic Sobolev inequalities for Bernoulli and Poisson measures. *J. Funct. Anal.* **156** 347–365. MR1636948
[6] Brown, L.D. (1986). *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. *Institute of Mathematical Statistics Lecture Notes—Monograph Series* **9**. Hayward, CA: IMS. MR0882001
[7] Dahinden, C., Parmiggiani, G., Emerick, M.C. and Bühlmann, P. (2007). Penalized likelihood for sparse contingency tables with an application to full-length cDNA libraries. *BMC Bioinformatics* **8** 476.
[8] Darroch, J.N., Lauritzen, S.L. and Speed, T.P. (1980). Markov fields and log-linear interaction models for contingency tables. *Ann. Statist.* **8** 522–539. MR0568718
[9] Dobra, A. and Massam, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600
[10] Edwards, D. (2000). *Introduction to Graphical Modelling*, 2nd ed. New York: Springer. MR1880319
[11] Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32** 928–961. MR2065194
[12] Fienberg, S.E. and Rinaldo, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267
[13] Friedman, J., Hastie, T. and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. Available at http://arxiv.org/abs/1001.0736.
[14] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68. MR1790613
[15] Greenshtein, E. (2006). Best subset selection, persistence in high-dimensional statistical learning and optimization under l_1 constraint. *Ann. Statist.* **34** 2367–2386. MR2291503
[16] Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039
[17] Haberman, S.J. (1974). *The Analysis of Frequency Data*. Chicago: Univ. Chicago Press. MR0408098
[18] Lauritzen, S.L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. New York: Oxford Univ. Press. MR1419991
[19] Lauritzen, S.L. (2002). Lectures on contingency tables. Available at http://www.stats.ox.ac.uk/~steffen/papers/cont.pdf.
[20] 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.
[21] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363
[22] Morris, C. (1975). Central limit theorems for multinomial sums. *Ann. Statist.* **3** 165–188. MR0370871
[23] Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electron. J. Stat.* **2** 605–633. MR2426104
[24] Negahban, S., Ravikumar, P., Wainwright, M.J. and Yu, B. (2010). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers, Available at http://arxiv.org/abs/1010.2731v1.
[25] Portnoy, S. (1986). On the central limit theorem in $\mathbf{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields* **73** 571–583. MR0863546
[26] Portnoy, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876

[27] Puig, A., Wiesel, A. and Hero, A. (2009). A multidimensional shrinkage-thresholding operator. In *Proceeding of the IEEE/SP 15th Workshop on Statistical Signal Processing*.
[28] Quine, M.P. and Robinson, J. (1984). Normal approximations to sums of scores based on occupancy numbers. *Ann. Probab.* **12** 794–804. MR0744234
[29] Read, T.R.C. and Cressie, N.A.C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate Data*. New York: Springer. MR0955054
[30] Rinaldo, A. (2006). Computing maximum likelihood estimates in log-linear models. Technical Report 835, Dept. Statistics, Carnegie Mellon Univ.
[31] Rinaldo, A., Fienberg, S.E. and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **3** 446–484. MR2507456
[32] Roth, V. and Fischer, B. (2008). The group-Lasso for generalized linear models: Uniqueness of solutions and efficient algorithms. In *Proceedings of the 25th International Conference on Machine Learning*.
[33] Schervish, M.J. (1995). *Theory of Statistics*. New York: Springer. MR1354146
[34] Steck, G.P. (1957). Limit theorems for conditional distributions. *Univ. California Publ. Statist.* **2** 237–284. MR0091552
[35] van de Geer, S.A. (2006). High-dimensional generalized linear models and the Lasso. Research Report 133, Swiss Federal Institute of Technology, Zurich.
[36] 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.
[37] Wainwright, M.J. (2009). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming. *IEEE Trans. Inform. Theory* **55** 2183–2202.
[38] Wainwright, M., Ravikumar, P. and Lafferty, J. (2011). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.* **38** 1287–1319.
[39] Yuan, M., Joseph, V.R. and Zou, H. (2009). Structured variable selection and estimation. *Ann. Appl. Stat.* **3** 1738–1757. MR2752156
[40] Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574
[41] Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7** 2541–2563. MR2274449

Received October 2009 and revised December 2010