ESTIMATING HETEROGENEOUS GRAPHICAL MODELS FOR DISCRETE DATA WITH AN APPLICATION TO ROLL CALL VOTING

By Jian Guo, Elizaveta Levina, George Michailidis and Ji Zhu

Department of Statistics, University of Michigan

We consider the problem of jointly estimating a collection of graphical models for discrete data, corresponding to several categories that share some common structure. An example of this setting is voting records of senators on different issues, such as defense, energy, and healthcare. We develop a Markov graphical model to characterize the heterogeneous dependence structures arising from such data. The model is fitted via a joint estimation method that preserves the underlying common graph structure, but also allows for differences between the networks. The method employs a group penalty that targets the common zero interaction effects across all the networks. We apply the method to describe the internal networks of the US Senate on several important issues. Our analysis reveals individual structure for each issue, distinct from the underlying well-known bipartisan structure common to all categories which we are able to extract separately. We also establish consistency of the proposed method both for parameter estimation and model selection, and evaluate its numerical performance on a number of simulated examples.

1. Introduction. The analysis of roll call data of legislative bodies has attracted a lot of attention both in the political science and statistical literature. For political scientists, such data allow to study broad issues such as party cohesion as well as more specific ones such as coalition formation (see for example the books [7, 15, 18, 20]). For statisticians the challenge is how to best model and present the data, so that interesting patterns become apparent and informative for subsequent analyses. A number of techniques have been employed including principal components analysis (PCA) [5], multidimensional scaling (MDS) [6], Bayesian models [4] and graphical models [2]. Dimension reduction techniques such as PCA and MDS aim at constructing a “map”, with the members of the legislative body positioned relative to their peers according to their voting pattern. A typical example of such a map of the US Senate members in the 109th Congress (2005-06) using multidimensional scaling for selected votes is shown in Figure 1; for a detailed

AMS 2000 subject classifications: Primary 62H99
Keywords and phrases: graphical models, group penalty, high-dimensional data, $\ell_1$ penalty, Markov network, binary data
description of the data see Section 4. A clear separation between members of the two parties is seen (Republicans to the left of the map and Democrats to the right), together with some members exhibiting a voting pattern deviating from their ideological peers, e.g., Nelson (Democrat of Nebraska), and Collins and Snow (Republicans of Maine), while the independent Jeffords (shown in purple) votes like a Democrat. More interestingly, the voting patterns within the two parties exhibit a clustering structure, which a closer inspection of the votes and subsequent analysis showed was mainly driven by votes on defense/security and healthcare.

This finding suggests that treating all votes as homogeneous, i.e., assuming that they represent the same underlying relationship between senators, may mask more subtle patterns which depend on the issues being voted upon. Therefore, treating votes as heterogeneous is more accurate, and can provide further insight into the voting behavior of different groups of senators on different issues. In this paper, we focus on voting records on three types of bills: defense and national security, environment and energy, and healthcare issues. Voting on the latter category is typically more partisan than voting on defense and national security, and thus we expect to see different connections in different categories.

To model this heterogeneity, we use Markov network models to capture the dependence structure of binary and/or categorical random variables. Similarly to Gaussian graphical models, nodes in the network correspond
to variables, while edges represent dependence between nodes. Graphical models have been used in a number of application areas, including bioinformatics [1], natural language processing [11] and image analysis [13]. In the case of Gaussian graphical models, the structure of the underlying graph can be fully determined from the corresponding inverse covariance (precision) matrix, the off-diagonal elements of which are proportional to partial correlations between the variables. A number of methods have been recently proposed in the literature to fit sparse Gaussian graphical models (see for example, Banerjee, El Ghaoui and d’Aspremont [2], Meinshausen and Bühlmann [16], Peng et al. [19], Ravikumar et al. [22], Rothman et al. [23], Yuan and Lin [25] and references therein).

To accomplish the analysis allowing for heterogeneity, we develop a framework for fitting different Markov models for each category that are nevertheless linked, sharing nodes and some common edges across all categories, while other edges are uniquely associated with a particular category. Asymptotic properties of the proposed estimator are also established. Note that the Gaussian case was studied by Guo et al. [8], who proposed a joint likelihood based estimation method that borrowed strength across categories.

The advantage of using a Markov graphical model is that it quantifies the degree of dependence between the senators based on their voting record and hence the obtained network and is directly interpretable. Techniques like multidimensional scaling and principal components analysis represent relative similarities between senators’ voting records on the map, and hence the distance between any two senators can be interpreted as a quantitive measure of similarity between their voting records. However, unlike in a Markov network, these distances are not interpretible in the context of a generative probability model.

The remainder of the paper is organized as follows. Section 2 introduces the Markov network and addresses algorithmic issues, and Section 3 briefly illustrates the performance of the joint estimation method on simulated data. A detailed analysis of the US Senate’s voting record from 109th Congress is presented in Section 4. Section 5 presents asymptotic results, and some concluding remarks are drawn in Section 6. The Appendix contains the proofs.

2. Model and Estimation Algorithm. In this section, we present the Markov model for heterogeneous data, focusing on the special case of binary variables (also known as the Ising model). The extension to general categorical variables is briefly discussed in Section 6. We start by discussing estimation of separate models for each category and then develop a model
for joint estimation.

Notice that the main technical challenge when estimating the likelihood of Markov graphical models is its computational intractability due to the normalizing constant. To overcome this difficulty, different methods employing computationally tractable approximations to the likelihood have been proposed in the literature; these include methods based on surrogate likelihood [2, 12] and pseudo-likelihood [9, 10, 21]. Höefling and Tibshirani [10] also proposed an iterative algorithm that successively approximates the original likelihood through a series of pseudo-likelihoods, while Ravikumar, Wainwright and Lafferty [21] and Guo et al. [9] established asymptotic consistency of their respective methods.

2.1. Problem Setup and Separate Estimation. Suppose that data have been collected on $p$ binary variables for $K$ categories, with $n_k$ observations for the $k$-th category. Let $x_1^{(k)} = (x_{i,1}, \ldots, x_{i,p})$ denote a $p$-dimensional row vector containing the data for the $i$-th observation in the $k$-th category and assume that it is drawn independently from an exponential family with the density function

$$f_k(x_1, \ldots, x_p) = \frac{1}{Z(\Theta^{(k)})} \exp \left( \sum_{j=1}^{p} \theta_{j,j}^{(k)} X_j + \sum_{1 \leq j < j' \leq p} \theta_{j,j'}^{(k)} X_j X_{j'} \right),$$

The partition function $Z(\Theta^{(k)}) = \sum_{X_j \in \{0,1\}, j} \exp(\theta_{j,j}^{(k)} X_j + \sum_{j < j'} \theta_{j,j'}^{(k)} X_j X_{j'})$ ensures that the density function in (2.1) integrates to one. The parameters $\theta_{j,j}^{(k)}$, $1 \leq j \leq p$ correspond to the main effect for variable $X_j$ in the $k$-th category, while $\theta_{j,j'}^{(k)}$, $1 \leq j < j' \leq p$ to the interaction effect between variables $X_j$ and $X_{j'}$. The underlying network associated with the $k$-th category is determined by the symmetric matrix $\Theta^{(k)} = (\theta_{j,j'}^{(k)})_{p \times p}$. Specifically, if $\theta_{j,j'}^{(k)} = 0$, then $X_j$ and $X_{j'}$ are conditionally independent in the $k$-th category given all the remaining variables, and hence their corresponding nodes are not connected. For each category, criterion (2.1) is referred to as the Markov network in the machine learning literature, and as the log-linear model in the statistics literature, where $\theta_{j,j'}^{(k)}$ is also interpreted as the conditional log-odds-ratio between $X_j$ and $X_{j'}$ given the other variables. Although general Markov networks allow higher order interactions (3-way, 4-way, etc), Ravikumar, Wainwright and Lafferty [21] pointed out that one can consider only the pairwise interaction effects without loss of generality, since higher order interactions can be converted to pairwise ones by introducing additional variables [24].
The simplest way to deal with heterogenous data is to estimate \( K \) separate Markov models, one for each category. If one further assumes sparsity for the \( k \)-th category, the structure of the underlying graph can be estimated by regularizing the log-likelihood using an \( \ell_1 \) penalty:

\[
\max_{\Theta^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \left\{ \sum_{j=1}^{p} \theta_{j,j}^{(k)} x_{i,j}^{(k)} + \sum_{j < j'} \theta_{j,j'}^{(k)} x_{i,j}^{(k)} x_{i,j'}^{(k)} \right\} - \log Z(\Theta^{(k)}) - \lambda \sum_{j < j'} |\theta_{j,j'}^{(k)}|.
\]

The \( \ell_1 \) penalty shrinks some of the interaction effects \( \theta_{j,j'}^{(k)} \) to zero and \( \lambda \) controls the degree of sparsity. However, estimating (2.2) directly is computationally infeasible due to the nature of the partition function. To overcome this difficulty, we adopt a pseudo-likelihood estimation method Guo et al. [9], Höefling and Tibshirani [10], based on

\[
\max_{\Theta^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left[ x_{i,j}^{(k)} \left( \theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) - \log \left\{ 1 + \exp \left( \theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right\} \right] - \lambda \sum_{j < j'} |\theta_{j,j'}^{(k)}|,
\]

(2.3)

where \( \Theta^{(k)} \) is restricted to be symmetric. Criterion (2.3) can be efficiently maximized using the modified coordinate descent algorithm of Höefling and Tibshirani [10].

2.2. Joint Estimation of Heterogeneous Networks. We start by reparameterizing each \( \theta_{j,j'}^{(k)} \) as

\[
\theta_{j,j'}^{(k)} = \phi_{j,j'} \gamma_{j,j'}^{(k)}, \quad 1 \leq j \neq j' \leq p; 1 \leq k \leq K.
\]

(2.4)

To avoid sign ambiguities between \( \phi_{j,j'} \) and \( \gamma_{j,j'}^{(k)} \), we restrict \( \phi_{j,j'} \geq 0, 1 \leq j < j' \leq p \). To preserve the symmetry of \( \Theta^{(k)} \), we also require \( \phi_{j,j'} = \phi_{j',j} \) and \( \gamma_{j,j'}^{(k)} = \gamma_{j,j'}^{(k)} \) for all \( 1 \leq j < j' \leq p \) and \( 1 \leq k \leq K \). Moreover, for identifiability reasons, we restrict the diagonal elements \( \phi_{j,j} = 1 \) and \( \gamma_{j,j}^{(k)} = \omega_{j,j}^{(k)} \). Note that \( \phi_{j,j'} \) is a common factor across all \( K \) categories that controls the occurrence of common links shared across categories, while \( \gamma_{j,j'}^{(k)} \) is an individual factor specific to the \( k \)-th category. The proposed joint estimation
method maximizes the following penalized criterion:

$$
\max_{\{\Phi^{(k)}, \Gamma^{(k)}\}} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left[ x_{i,j}^{(k)} \left( \theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right. \\
- \log \left\{ 1 + \exp \left( \theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right\} \\
- \eta_1 \sum_{j<j'} \phi_{j,j'} - \eta_2 \sum_{j<j'} \sum_{k=1}^{K} \gamma_{j,j'}^{(k)}
$$

(2.5)

where $\Phi^{(k)} = (\phi_{j,j'})_{p \times p}$ and $\Gamma^{(k)} = (\gamma_{j,j'})_{p \times p}$. The tuning parameter $\eta_1$ controls sparsity of the common structure across the $K$ networks. Specifically, if $\phi_{j,j'}$ is shrunk to zero, all $\gamma_{j,j'}^{(1)}, \ldots, \gamma_{j,j'}^{(K)}$ are also zero, and hence there is no link between nodes $j$ and $j'$ in any of the $K$ graphs. Similarly, $\eta_2$ is a tuning parameter controlling sparsity of links in individual categories. Due to the nature of the $\ell_1$ penalty, some of $\gamma_{j,j'}^{(k)}$’s will be shrunk to zero, resulting in a collection of graphs with individual differences. Note that this two-level penalty was originally proposed by Zhou and Zhu [26] for group variable selection in linear regression.

To simplify estimation, we convert the criterion (2.5) to an equivalent criterion with only one tuning parameter:

$$
\max_{\{\Theta^{(k)}\}} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left[ x_{i,j}^{(k)} \left( \theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right. \\
- \log \left\{ 1 + \exp \left( \theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right\} \\
- \lambda \sum_{1 \leq j < j' \leq p} \sqrt{\sum_{k=1}^{K} |\theta_{j,j'}^{(k)}|},
$$

(2.6)

where $\lambda = 2\sqrt{\eta_1 \eta_2}$. The equivalence between (2.5) and (2.6) can be formalized as follows (here $A \cdot B$ denotes the Schur-Hadamard element-wise product of two matrices).

**Proposition 1.** Let $\{\hat{\Theta}^{(k)}\}_{k=1}^{K}$ be a local minimizer of (2.6). Then there exists a local minimizer of (2.5), $\hat{\Phi}, \hat{\Gamma}^{(k)} \{\hat{\Theta}^{(k)}\}_{k=1}^{K}$, such that $\hat{\Theta}^{(k)} = \hat{\Phi} \cdot \hat{\Gamma}^{(k)}$, for all $1 \leq k \leq K$. On the other hand, if $\hat{\Phi}, \hat{\Gamma}^{(k)} \{\hat{\Theta}^{(k)}\}_{k=1}^{K}$ is a local minimizer of (2.5), then there also exists a local minimizer of (2.6), $\{\hat{\Theta}^{(k)}\}_{k=1}^{K}$, such that $\hat{\Theta}^{(k)} = \hat{\Phi} \cdot \hat{\Gamma}^{(k)}$, for all $1 \leq k \leq K$. 

The proof of this proposition is similar to the proofs of Lemma 1 and Theorem 1 in [26] and is omitted here.

2.3. Algorithm and Model Selection. Criterion (2.6) leads to an efficient estimation algorithm based on the local linear approximation. Specifically, letting \( \hat{\theta}_{j,j'}^{(k)}[t] \) denote the estimates from the \( t \)-th iteration, we approximate

\[
\sqrt{\sum_{k=1}^{K} |\theta_{j,j'}^{(k)}|} \approx \sum_{k=1}^{K} |\theta_{j,j'}^{(k)}|/\sqrt{\sum_{k=1}^{K} |(\theta_{j,j'}^{(k)})[t]|},
\]

when \( \theta_{j,j'}^{(k)} \approx (\theta_{j,j'}^{(k)})[t] \). Thus, at the \( (t + 1) \)-th iteration, problem (2.6) is decomposed into \( K \) individual optimization problems:

\[
\max_{\Theta^{(k)}} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left[ x_{i,j}^{(k)} \left( \hat{\theta}_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) - \log \left\{ 1 + \exp \left( \hat{\theta}_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right\} \right]
\]

\[
- \lambda \sum_{1 \leq j < j' \leq p} \left( \sum_{k=1}^{K} |(\theta_{j,j'}^{(k)})[t]| \right)^{-1/2} |\theta_{j,j'}^{(k)}|.
\]

(2.7)

Note that criterion (2.7) is a variant of criterion (2.3) with a weighted \( \ell_1 \) penalty and hence can be solved by the algorithm of [10]. For numerical stability, we threshold \( \sqrt{\sum_{k=1}^{K} |(\theta_{j,j'}^{(k)})[t]|} \) at \( 10^{-10} \). The algorithm is summarized as follows:

**Step 1.** Initialize \( \hat{\theta}_{j,j'}^{(k)} \)'s \( (1 \leq j, j' \leq p; 1 \leq k \leq K) \) using the estimates from the separate estimation method;

**Step 2.** For each \( 1 \leq k \leq K \), update \( \hat{\theta}_{j,j'}^{(k)} \)'s by solving (2.7) using the JOSE algorithm in Guo et al. [9];

**Step 3.** Repeat Step 2 until convergence.

The tuning parameter \( \lambda \) in (2.6) controls the sparsity of the resulting estimator and can be selected using cross-validation. Specifically, for each \( 1 \leq k \leq K \), we randomly split the data in the \( k \)-th category into \( D \) subsets of similar sizes and denote the index set of the observations in the \( d \)-th subset as \( T_d^{(k)} \), \( 1 \leq d \leq D \). Then \( \lambda \) is selected by maximizing

\[
\frac{1}{D} \sum_{d=1}^{D} \sum_{k=1}^{K} \frac{1}{|T_d^{(k)}|} \sum_{i \in T_d^{(k)} j=1}^{p} x_{i,j}^{(k)} \left\{ (\hat{\theta}_{j,j}^{(k)})[-d](\lambda) + \sum_{j' \neq j} (\hat{\theta}_{j,j'}^{(k)})[-d](\lambda) x_{i,j'}^{(k)} \right\}
\]

\[
- \log \left[ 1 + \exp \left\{ (\hat{\theta}_{j,j}^{(k)})[-d](\lambda) + \sum_{j' \neq j} (\hat{\theta}_{j,j'}^{(k)})[-d](\lambda) x_{i,j'}^{(k)} \right\} \right],
\]

(2.8)
where $|T_d^{(k)}|$ is the cardinality of $T_d^{(k)}$ and $(\hat{\theta}^{(k)}_{j,j'})[^{-d}]$ is the joint estimate of $\theta^{(k)}_{j,j'}$ based on all observations except those in $T_d^{(1)} \cup \ldots \cup T_d^{(K)}$, as well as the tuning parameter $\lambda$.

3. Simulation Study. Before turning our attention to examining the US Senate voting patterns, we evaluate the performance of the joint estimation method on three synthetic examples, each with $p = 50$ variables and $K = 3$ categories. The network structure in each example is composed of two parts: the common structure across all categories and the individual structure specific to a category. The common structures in these examples are a chain graph, a nearest neighbor graph and a scale-free graph. These graphs are generated as follows:

**Example 1: Chain Graph.** A chain graph is generated by connecting nodes $1$ to $p$ in increasing order, as shown in Figure 2 (A1).

**Example 2: Nearest Neighbor Graph.** The data generating mechanism of the nearest neighbor graph is adapted from Li and Gui [14]. Specifically, we generate $p$ points randomly on a unit square, calculate all $p(p-1)/2$ pairwise distances, and find three nearest neighbors of each point in terms of these distances. The nearest neighbor network is obtained by linking any two points that are nearest neighbors of each other. Figure 2 (B1) illustrates a nearest-neighbor graph.

**Example 3: Scale-free Graph.** A scale-free graph has a power-law degree distribution and can be simulated by the Barabasi-Albert algorithm [3]. A realization of a scale-free network is depicted in Figure 2 (C1).

In each example, the network for the $k$-th category ($k = 1, \ldots, K$) is created by randomly adding links to the common structure. The individual links in different categories are disjoint and have the same degree of sparsity, measured by $\rho$, the ratio of the number of individual links to the number of common links. In particular, $\rho = 0$ corresponds to identical networks for all three categories. In the simulation study, we consider $\rho = 0, 1/4$ and $1$, gradually increasing the proportion of individual links (Figure 2). Given the graphs, the symmetric parameter matrix $\Theta^{(k)}$ is generated as follows. Each $\theta^{(k)}_{j,j'} = \theta^{(k)}_{j',j}$ corresponding to an edge between nodes $j$ and $j'$ is uniformly drawn from $[-1, -0.5] \cup [0.5, 1]$, whereas all other elements are set to zero. Then we generate the data using Gibbs sampling. Specifically, suppose the $i$-th iteration sample has been drawn and is denoted as $(x_1^{(k)})^{|t|}, \ldots, (x_p^{(k)})^{|t|}$; then, in the $(t + 1)$-th iteration, we draw $(x_j^{(k)})^{|t+1|}, 1 \leq j \leq p$, from the
Bernoulli distribution:

\[(x_{j}^{(k)})_{[t+1]} \sim \text{Bernoulli}\left(\frac{\exp(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)}(x_{j'}^{(k)})_{[t]})}{1 + \exp(\theta_{j,j}^{(k)} + \sum_{j' \neq j} \theta_{j,j'}^{(k)}(x_{j'}^{(k)})_{[t]})}\right).\]

To ensure that the simulated observations are close to i.i.d. samples from the target distribution, the first 1,000,000 rounds are discarded (burn-in) and the data are collected every 100 iterations from the sampler. In the simulation study, we consider a balanced scenario and an unbalanced scenario. The former consists of \(n_k = 200\) observations in each category, whereas the latter has three unbalanced categories with sample sizes \(n_1 = 150\), \(n_2 = 300\) and \(n_3 = 450\).

We compared the structure estimation results of the joint estimation method and the separate estimation method using ROC curves, which dynamically characterize the sensitivity (proportion of correctly identified links) and the specificity (proportion of correctly excluded links) by varying the tuning parameter \(\lambda\). Figure 3 shows the ROC curves averaged over 50 replications from the three examples in the balanced scenario, where the joint estimation method dominates separate estimation when the proportion of individual links is low. As \(\rho\) increases, the structures become more different, and the joint and separate methods move closer together. This is expected, since the joint estimation method is designed to take advantage of common structure. The results in the unbalanced scenario exhibit a similar pattern (Figure 4).

4. Analysis of the U.S. Senate voting records. We applied the proposed joint estimation method to the voting records of the U.S. Senate from the 109th Congress covering the period 2005-2006. The data were obtained directly from the Senate’s website (www.senate.gov). The variables correspond to the 100 senators, and the observations to the 645 votes that the Senate deliberated and voted on during that period, which include bills, resolutions, motions, debates and roll call votes. The votes are recorded as “yes” (encoded as “1”) and “no” (encoded as “0”). Missing observations were replaced with the majority vote of the senator’s party on that particular vote. The bills with a “yes/no” proportion greater than 90% or less than 10% were excluded from the analysis. Three categories of votes were extracted from bills, resolutions and motions: 1) defense and security issues (133); 2) environment and energy issues (34); 3) health and medical care issues (46). The tuning parameter for the proposed method was selected through cross-validation. Following Li and Gui [14], we used a bootstrap
procedure with the proposed estimator to evaluate the confidence of the estimated edges. Specifically, we estimated the network for multiple bootstrap samples of the same size, and only retained the edges that appeared more than $\alpha$ percent of the time. This procedure is similar to stability selection [17].

The network representation, depicting both the common and the individual structures with a cut-off value for inclusion $\alpha$ of 0.4 and 0.6, is given in Figures 5 and 6, respectively. Note that unlike techniques such as principal components analysis and multidimensional scaling that directly embed the senators in a two-dimensional map, the proposed method estimates the edges and constructs the adjacency matrix of the graph of senators; subsequently, we employed a graph drawing program to position the senators on a two-dimensional map. The common networks estimated by the joint estimation method are shown in the top left panels of Figures 5 and 6. For the individual categories, we only plot the edges associated with the category to enhance the visual reading of the graphs. As expected, members of the two political parties are clearly separated. For both cutoff inclusion values, there are strong positive associations between senators of the same party and equally strong negative associations between senators of opposite parties. Obviously, at the higher cutoff value the common dependence structure becomes sparser. Of particular interest is the finding that at both cutoff values, there are many more associations between Democratic senators than Republican ones and this pattern holds for both the common and individual structures. One possible explanation may be that the Democrats were in the opposition, thus voting more like a block. Further, the Independent Senator Jeffords is associated with the Democrats, while the moderate Republicans Collins, Snowe, Chafee and Specter (who switched to the Democratic party in early 2009) and the conservative Democrat Nelson (Nebraska) are not strongly associated with the Republican colleagues, thus confirming results of previous analyses by Clinton, Jackman and Rivers [4] and de Leeuw [5] (albeit based on data from the 105th Congress). Somewhat more surprisingly, a similar finding holds for the Republican senators McCain, Graham and DeMint, a pattern also hinted at by the multidimensional scaling analysis presented in the Introduction, but more clearly captured by the Markov model. Also, the analysis suggests that Senator Lieberman had a solid Democratic voting record before switching to an independent in 2008.

Other interesting patterns emerging from the analysis are that the more moderate members of two parties are located closer to the center of their respective “clouds” (e.g. Warner, Voinovich, Smith on the Republican side and Levin, Reid, Mikulski, Rockefeller on the Democratic side), the close
ties of the liberal Democrats Kennedy, Boxer and Nelson (Florida), the close voting records of senators from the same state (Murkowski and Stevens from Alaska and Cantwell and Murray from Washington).

Examining the individual networks for the three categories shown in Figures 5 and 6, we note that additional positive associations among Democrats emerge, primarily for defense and healthcare categories, thus indicating a stronger ideological cohesion on these issues. Further, a number of stable negative associations emerge in the environment and healthcare categories, indicating a stronger ideological divide between senators.

Other patterns of interest include a strong dependence between Durbin, Corzine, Lincoln, Harkin, Dodd and Dayton on defense and health issues, and between Schumer, Clinton, Murray, and Lautenberg, on the same issues. There is also a cluster of positive associations around Democratic senator Nelson (Florida) on environmental issues, which is of interest given his position on offshore oil drilling.

For comparison purposes, separate multidimensional scaling analyses are shown in Figure 7 for all the votes together and for the three individual categories. As alluded in the Introduction, the clustering within the two parties is driven to a large extent by the corresponding clustering in the defense and health categories. On the other hand, voting on environmental issues creates a clear separation between the two parties, although the analysis reveals that the moderate Republicans Chafee, Collins and Snowe are shown to have a voting record similar to the Democrats, while the Democrats Nelson (Nebraska) and Landrieu are closer to the Republicans.

5. Asymptotic Properties. In this section, we study the asymptotic properties of the proposed joint estimation method. Since the structure of the underlying network only depends on the interaction effects, we focus on a variant of the model without main effects. Specifically, we solve

\[
\max_{\{\Theta^{(k)}\}_{k=1}^{K}} \left\{ \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} \left[ x_{i,j}^{(k)} \left( \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) - \log \left( 1 + \exp \left( \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right) \right] \right\}
\]

\[
- \lambda \sum_{j<j'} \sqrt{\sum_{k=1}^{K} |\theta_{j,j'}^{(k)}|}.
\]

(5.1)

We will show that the estimator in criterion (5.1) is consistent in terms of both parameter estimation and model selection, when \(p\) and \(n\) go to infinity and the tuning parameter \(\lambda\) goes to zero at some appropriate rate.

Before stating the main results, we introduce necessary notation and regularity conditions. For each \(k = 1, \ldots, K\), denote \(\Theta^{(k)} = (\theta_{1,2}^{(k)}, \ldots, \theta_{j,j'}^{(k)}, \ldots, \theta_{p-1,p}^{(k)})\).
as a \(p(p-1)/2\)-dimensional vector, recording all upper triangular elements in \(\Theta^{(k)}\). Let \(\bar{\theta}^{(k)}\) be the true value of \(\theta^{(k)}\). Let \(\mathbf{Q}^{(k)}\) be the population Fisher information matrix of the model in criterion (5.1) (see the Appendix for a precise definition) and let \(\mathbf{X}^{(i)}\) be a matrix with \(p\) rows and \(p(p-1)/2\) columns, whose \((j,j')\)-th column is composed of zeros except the \(j\)-th \((j'-\)th component being \(x_{i,j'}\)). In addition, we define \(\mathbf{U}^{(k)} = E[\mathbf{X}^{(k)}_i \mathbf{X}^{(k)}_i^T]\). To index the zero and nonzero elements, let \(S_k = \{(j,j') : \theta^{(k)}_{j,j'} \neq 0, 1 \leq j < j' \leq p\}\) and \(S_k^c = \{(j,j') : \theta^{(k)}_{j,j'} = 0, 1 \leq j < j' \leq p\}\), and let \(S_\cap = \bigcap_{k=1}^K S_k\) and \(S_\cup = \bigcup_{k=1}^K S_k\). The cardinalities of \(S_k\) and \(S_\cap\) are denoted by \(q_k\) and \(q\), respectively. For any matrix \(\mathbf{W}\) and subsets of row and column indices \(\mathcal{U}\) and \(\mathcal{V}\), let \(\mathbf{W}_{\mathcal{U},\mathcal{V}}\) be the matrix consisting of rows \(\mathcal{U}\) and columns \(\mathcal{V}\) in \(\mathbf{W}\). Finally, let \(\Lambda_{\min}(\cdot)\) and \(\Lambda_{\max}(\cdot)\) denote the smallest and largest eigenvalue of a matrix, respectively.

The asymptotic properties of the joint estimation method rely on the following regularity conditions:

(A) **Nonzero elements bounds:** There exist positive constants \(\gamma_{\min}\) and \(\gamma_{\max}\) such that

\[
\begin{align}
(i) & \quad \min_{1 \leq k \leq K} \min_{(j,j') \in S_k} |\theta^{(k)}_{j,j'}| \geq \gamma_{\min}; \\
(ii) & \quad \max_{1 \leq k \leq K} \max_{(j,j') \in S_k \setminus S_\cap} |\theta^{(k)}_{j,j'}| \leq \gamma_{\max}.
\end{align}
\]

(B) **Dependency:** There exist positive constants \(\tau_{\min}\) and \(\tau_{\max}\) such that for any \(k = 1, \ldots, K\),

\[
(5.2) \quad \Lambda_{\min}(\mathbf{Q}^{(k)}_{S_k^c,S_k}) \geq \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\mathbf{U}^{(k)}_{S_k,S_k}) \leq \tau_{\max}.
\]

(C) **Incoherence:** There exists a constant \(\tau \in (1 - \sqrt{\gamma_{\min}/4\gamma_{\max}}, 1)\) such that for any \(k = 1, \ldots, K\),

\[
(5.3) \quad \|\mathbf{Q}^{(k)}_{S_k^c,S_k}(\mathbf{Q}^{(k)}_{S_k,S_k})^{-1}\|_{\infty} \leq 1 - \tau.
\]

Condition (A) enforces a lower bound on the magnitudes of all nonzero elements, as well as an upper bound on the magnitudes of those nonzero elements associated with individual links. Conditions (B) and (C) bound the amount of dependence and the influence that the non-neighbors can have on a given node, respectively. Conditions similar to (B) and (C) were also assumed by Meinshausen and Buhlmann [16], Ravikumar, Wainwright and Lafferty [21], Peng et al. [19] and Guo et al. [9]. Our conditions are most closely related to those of Guo et al. [9], but here they are extended to the heterogenous data setting.
Theorem 1. (Parameter estimation). Suppose all regularity conditions hold. If the tuning parameter \( \lambda = C_\lambda \sqrt{\log p/n} \) for some constant \( C_\lambda > (8 - 4\tau)\sqrt{\gamma_{\min}/(1 - \tau)} \) and if \( \min\{n/q^3, n_1/q^3_1, \ldots, n_K/q^3_K\} > (4/C)\log p \) for some constant \( C = \min\{\tau^2/288(1 - \tau)^2, \tau^2/72, \tau_{\min}\tau/48\} \), then there exists a local maximizer of the criterion (5.1), \( \{\hat{\theta}^{(k)}\}_{k=1}^K \), such that, with probability tending to 1,

\[
\sum_{k=1}^K \|\hat{\theta}^{(k)} - \theta^{(k)}\|_2 \leq M \sqrt{\frac{q\log p}{n}},
\]

for some constant \( M > (2KC_\lambda/\tau_{\min}\sqrt{\gamma_{\min}})(3 - 2\tau)/(2 - \tau) \).

Theorem 2. (Structure selection). Under conditions of Theorem 1, with probability tending to 1, the maximizer \( \{\hat{\theta}^{(k)}\}_{k=1}^K \) from Theorem 1 satisfies

\[
\hat{\theta}^{(k)}_{j,j'} \neq 0, \quad \text{for all} \quad (j, j') \in S_k, k = 1, \ldots, K;
\]

\[
\hat{\theta}^{(k)}_{j,j'} = 0, \quad \text{for all} \quad (j, j') \in S^c_k, k = 1, \ldots, K.
\]

Theorems 1 and 2 establish the consistency in terms of parameter estimation and structure selection, respectively. The proofs are given in the Appendix.

6. Concluding Remarks. We have proposed a joint estimation method for the analysis of heterogeneous Markov networks motivated by the need to jointly estimate heterogeneous networks, such as those of the Senate voting patterns. The method improves estimation of the networks’ common structure by borrowing strength across categories, and allows for individual differences. Asymptotic properties of the method have been established. In particular, we show that the convergence rate is similar to the rate for Gaussian graphical models in a similar context [9]. The proposed method can be extended to deal with general categorical data with more than two levels using the strategy described in Ravikumar, Wainwright and Lafferty [21] and Guo et al. [9]. The most interesting feature emerging from the analysis of the Senate voting records is the existence of more stable associations for the Democrats, both in terms of the common structure and in the healthcare and defense categories.

There are other techniques suitable for analyzing roll call data. Dimension reduction techniques create maps, where the relative positioning of the senators allows one to infer similarity in their voting patterns. They provide a useful visual tool to capture broad patterns and relationships. On the other
hand, a Markov network model aims directly at estimating the associations between the senators and thus provides an alternative view of the voting patterns, which together with the thresholding technique employed gives a measure of the stability of such associations. Further, the joint estimation method allows one to separately study the overall voting patterns and those driven by specific issues. In our view, both set of techniques are useful, with dimension reduction providing a global perspective, and the Markov model revealing more nuanced patterns.

Acknowledgements. E. Levina’s research is partially supported by NSF grant DMS-0805798. G. Michailidis’s research is partially supported by NIH grant 1RC1CA145444-0110. J. Zhu’s research is partially supported by NSF grants DMS-0705532 and DMS-0748389.

References.
\[ \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j=1}^{p} x_{i,j}^{(k)} \left( \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) - \log \left( 1 + \exp \left( \sum_{j' \neq j} \theta_{j,j'}^{(k)} x_{i,j'}^{(k)} \right) \right), \]

whose first derivative and second derivative are denoted by \( \nabla l(\theta^{(k)}) \) and \( \nabla^2 l(\theta^{(k)}) \), respectively. Note that \( \nabla l(\theta^{(k)}) \) is a \( p(p-1)/2 \)-dimensional vector and \( \nabla^2 l(\theta^{(k)}) \) is a \( p(p-1)/2 \times p(p-1)/2 \) matrix. Then, the population Fisher information matrix of the model in (5.1) at \( \theta^{(k)} \) can be defined as \( \mathcal{Q}^{(k)} = -\mathbb{E}[\nabla^2 l(\theta^{(k)})] \), and its sample counterpart is \( \hat{\mathcal{Q}}^{(k)} = -\nabla^2 l(\theta^{(k)}) \).

We also write \( \hat{U}^{(k)} = 1/n \sum_{i=1}^{n} X_{(i)}^{(k)} X_{(i)}^{(k)T} \) for the sample counterpart of

Appendix. The appendix presents the proofs of Theorems 1 and 2. The main idea of the proof is closely related to [9], and some strategies for dealing with the joint estimation are borrowed from [8].
\( \tilde{U}^{(k)} \) defined in Section 5. Let \( \theta^{(k)} = (\theta^{(k)}_{1,2}, \ldots, \theta^{(k)}_{j,j'}, \ldots, \theta^{(k)}_{p-1,p}) \) be the same as \( \theta^{(k)} \) except that all elements in \( S^c_k \) are set to zero and write \( \delta^{(k)} = \theta^{(k)} - \bar{\theta}^{(k)} \) and \( \hat{\delta}^{(k)} = \tilde{\theta}^{(k)} - \bar{\theta}^{(k)} \). Finally, let \( W \) be a subset of the index set \( \{1, 2, \ldots, p(p-1)/2\} \). For a \( p(p-1)/2 \)-dimensional vector \( \beta \), we define \( \beta_W \) as the vector consisting of the elements of \( \beta \) associated with \( W \).

Next, we introduce a variant of criterion (5.1) by restricting all true zeros in \( \{\theta^{(k)}\}_{k=1}^K \) to be estimated as zero. Specifically, the restricted criterion is formulated as follows:

\[
\max_{\{\theta^{(k)}\}_{k=1}^K} \sum_{k=1}^{K} l(\theta^{(k)}) - \lambda \sum_{1 \leq j < j' \leq p} \sqrt{\sum_{k=1}^{K} |\theta^{(k)}_{j,j'}|},
\]

and its maximizer is denoted by \( \{\hat{\theta}^{(k)}\}_{k=1}^K \). In addition, we consider the sample versions of regularity conditions (B) and (C).

\( (B') \) **Sample dependency:** There exist positive constants \( \tau_{\min} \) and \( \tau_{\max} \) such that for any \( k = 1, \ldots, K \),

\[
\Lambda_{\min}(\hat{Q}^{(k)}_{S_k,S_k}) \geq \tau_{\min} \quad \text{and} \quad \Lambda_{\max}(\hat{U}^{(k)}_{S_k,S_k}) \leq \tau_{\max}.
\]

\( (C') \) **Sample incoherence:** There exists a constant \( \tau \in (1-\sqrt{\gamma_{\min}/4\gamma_{\max}}, 1) \) such that for any \( k = 1, \ldots, K \),

\[
\|\hat{Q}^{(k)}_{S_k,S_k} (\hat{Q}^{(k)}_{S_k,S_k})^{-1}\|_\infty \leq 1 - \tau.
\]

For convenience of the readers, the proof of our main result is divided into two parts: Part I presents the main idea of the proof by listing the important propositions and the proofs of Theorems 1 and 2, whereas Part II contains additional technical details and proofs of propositions in Part I.

**Part I: Propositions and Proof of Theorems 1 and 2.** The proof consists of the following steps. Proposition 2 shows that, under sample regularity conditions \( (B') \) and \( (C') \), the conclusions of Theorems 1 and 2 hold for the local maximizer of the restricted problem (6.1). Next, Proposition 3 proves that the population regularity conditions \( (B) \) and \( (C) \) give rise to their sample counterparts \( (B') \) and \( (C') \) with probability tending to one; hence the conclusions of Proposition 2 also hold with the population regularity conditions. Lastly, we show that the local maximizer of (6.1) is also a local maximizer of the original model (5.1). This is established via Proposition 4, which sets out the Karush-Kuhn-Tucker (KKT) conditions for the local
maximizer of criterion (5.1), and Proposition 5, which shows that, with probability tending to one, the local maximizer of (6.1) satisfies these KKT conditions.

**Proposition 2.** Suppose condition (A) and the sample conditions (B') and (C') hold. If the tuning parameter \( \lambda = C_\lambda \sqrt{(\log p) / n} \) for some constant \( C_\lambda > (8 - 4\tau)\sqrt{\tau_{\text{min}} / (1 - \tau)} \) and \( q \sqrt{(\log p) / n} = o(1) \), then with probability tending to one, there exists a local maximizer of the restricted criterion, \( \{\hat{\theta}^{(k)}\}_{k=1}^K \), satisfying

(i) \( \sum_{k=1}^K \|\hat{\theta}^{(k)} - \theta^{(k)}\|_2 \leq M \sqrt{q (\log p) / n} \) for some constant \( M > (2KC_\lambda / \tau_{\text{min}} \sqrt{\tau_{\text{min}}})[(3-2\tau)/(2 - \tau)] \);

(ii) For each \( k = 1, \ldots, K \), \( \hat{\theta}_{j,j'}^{(k)} \neq 0 \) for all \( (j,j') \in S_k \) and \( \hat{\theta}_{j,j'}^{(k)} = 0 \) for all \( (j,j') \in S_k^c \).

**Proposition 3.** Suppose the regularity conditions (B) and (C) hold, then for any \( \epsilon > 0 \), the following inequalities hold with probability tending to one for all \( k = 1, \ldots, K \):

(i) \( P\{\lambda_{\min}(\hat{Q}_{S_k,S_k}^{(k)}) \leq \tau_{\min} - \epsilon\} \leq 2 \exp\{-\epsilon^2 / 2(n_k/q_k^2) + 2 \log q_k\} \);

(ii) \( P\{\lambda_{\max}(\hat{U}_{S_k,S_k}^{(k)}) \geq \tau_{\max} + \epsilon\} \leq 2 \exp\{-\epsilon^2 / 2(n_k/q_k^2) + 2 \log q_k\} \);

(iii) \( P\{||\hat{Q}_{S_k,S_k}^{(k)}(\hat{\theta}_{S_k,S_k}^{(k)})||_\infty \geq 1 - \tau / 2\} \leq 12 \exp\{-Cn_k/q_k^3 + 4 \log p\} \), for some constant \( C = \min\{\tau_{\min}^{-1} \tau^{-2} / 288(1 - \tau)^2, \tau_{\min}^{-2} \tau^{-2} / 72, \tau_{\min} \tau / 48\} \).

**Proposition 4.** \( \{\theta\}_{k=1}^K \) is a local maximizer of problem (5.1) if and only if the following conditions hold for all \( k = 1, \ldots, K \):

\[
\nabla_{j,j'} l(\hat{\theta}^{(k)}) = \lambda \text{sgn}(\hat{\theta}_{j,j'}^{(k)})/(\sum_{k=1}^K |\hat{\theta}_{j,j'}^{(k)}|)^{1/2}, \quad \text{if } \hat{\theta}_{j,j'}^{(k)} \neq 0;
\]

\[
|\nabla_{j,j'} l(\hat{\theta}^{(k)})| < \lambda / (\sum_{k=1}^K |\hat{\theta}_{j,j'}^{(k)}|)^{1/2}, \quad \text{if } \hat{\theta}_{j,j'}^{(k)} = 0.
\]

**Proposition 5.** Under all conditions of Proposition 2, with probability tending to one, we have, for each \( k = 1, \ldots, K \),

\[
\nabla_{j,j'} l(\hat{\theta}^{(k)}) = \lambda \text{sgn}(\hat{\theta}_{j,j'}^{(k)})/(\sum_{k=1}^K |\hat{\theta}_{j,j'}^{(k)}|)^{1/2}, \quad \text{for all } (j,j') \in S_k;
\]

\[
|\nabla_{j,j'} l(\hat{\theta}^{(k)})| < \lambda / (\sum_{k=1}^K |\hat{\theta}_{j,j'}^{(k)}|)^{1/2}, \quad \text{for all } (j,j') \in S_k^c.
\]

**Proof of Theorems 1 and 2.** The condition \( \min\{n/q^3, n_1/q_1^3, \ldots, n_K/q_K^3\} > (4/C) \log p \) implies that, for each \( k = 1, \ldots, K \), we have \( -Cn_k/q_k^3 + 4 \log p < 0 \) and \( -(\epsilon^2 / 2)(n_k/q_k^2) + 2 \log q_k < 0 \) when \( q_k \) is large enough. This condition also implies \( q \sqrt{(\log p) / n} = o(1) \). In addition, by Proposition 3, the
sample conditions \((B')\) and \((C')\) hold with probability tending to one when regularity conditions \((B)\) and \((C)\) hold. Therefore, by Proposition 2, with probability tending to one, the solution of the restricted problem \(\{\hat{\theta}^{(k)}\}_{k=1}^{K}\) satisfies both parameter estimation consistency and structure selection consistency. On the other hand, by Proposition 5, with probability tending to one, \(\{\hat{\theta}^{(k)}\}_{k=1}^{K}\) also satisfies the KKT conditions in Proposition 4, thus it is a local maximizer of criterion \((5.1)\). This proves Theorems 1 and 2. \(\Box\)

**Part II: Proofs of Propositions.** Before proving the propositions, we state a few lemmas which will be used in the proofs. These lemmas are variants of Lemmas 1, 2 and 5 in Guo et al. \cite{9}, adapted to the settings of the heterogenous model and thus the proofs are omitted here. Likewise, the proof of Proposition 3 is very similar to the proof of Propositions 3 and 4 in Guo et al. \cite{9} and is omitted.

**Lemma 1.** For each \(k=1,\ldots,K\), with probability tending to 1, we have 
\[
\|\nabla l(\bar{\theta}^{(k)})\|_{\infty} \leq C_{\nabla} \sqrt{(\log p)/n}
\]
for some constant \(C_{\nabla} > 4\).

**Lemma 2.** If the sample dependency condition \((B')\) holds and \(q\sqrt{(\log p)/n} = o(1)\), then for any \(\alpha_k \in [0, 1], k = 1,\ldots,K\), the following inequality holds with probability tending to 1:
\[
(6.6) \quad -\sum_{k=1}^{K} \delta_{S_k}^{(k)T} \left[\nabla^2 l(\bar{\theta}^{(k)}) + \alpha_k \delta_{S_k}^{(k)}\right] S_k \delta_{S_k}^{(k)} \geq \frac{1}{2} \tau_{\min} \sum_{k=1}^{K} \|\delta^{(k)}\|_2^2.
\]

**Lemma 3.** Suppose the sample dependency condition \((B)\) holds. For any \(\alpha_k \in [0, 1], k = 1,\ldots,K\), the following inequality holds with probability tending to 1:
\[
(6.7) \quad \|\nabla^2 l(\bar{\theta}^{(k)}) + \alpha_k \delta^{(k)} - \nabla^2 l(\bar{\theta}^{(k)})\|_{\infty} \leq \tau_{\max} \|\delta^{(k)}\|_2^2.
\]

**Proof of Proposition 2.** The main idea of the proof was first introduced in this context in Rothman et al. \cite{23} and has since been used by many authors. Define
\[
(6.8) \quad G(\{\delta^{(k)}\}_{k=1}^{K}) = -\sum_{k=1}^{K} (l(\bar{\theta}^{(k)}) + l(\bar{\theta}^{(k)})) - \lambda \sum_{1 \leq j < j' \leq p} \left\{ \left( \sum_{k=1}^{K} |\bar{\theta}^{(k)}_{j,j'} + \delta_{S_k}^{(k)}|^{1/2} \right) - \left( \sum_{k=1}^{K} |\bar{\theta}^{(k)}_{j,j'}|^{1/2} \right) \right\}.
\]

It can be seen from \((6.1)\) that, \(\{\hat{\delta}^{(k)}\}_{k=1}^{K}\) minimizes \(G(\{\delta^{(k)}\}_{k=1}^{K})\) and \(G(\{0\}_{k=1}^{K}) = 0\). Thus we must have \(G(\{\hat{\delta}^{(k)}\}_{k=1}^{K}) \leq 0\). If we take a closed set \(A\) which
contains \( \{0\}_{k=1}^K \), and show that \( G \) is strictly positive everywhere on the boundary \( \partial \mathcal{A} \), then it implies that \( G \) has a local minimum inside \( \mathcal{A} \), since \( G \) is continuous and \( G(\{0\}_{k=1}^K) = 0 \). Specifically, we define \( \mathcal{A} = \{ \{\bar{\theta}^{(k)}\}_{k=1}^K : \sum_{k=1}^K \|\bar{\theta}^{(k)}\|_2 \leq M a_n \} \), with boundary \( \partial \mathcal{A} = \{ \{\bar{\theta}^{(k)}\}_{k=1}^K : \sum_{k=1}^K \|\bar{\theta}^{(k)}\|_2 = M a_n \} \), for some constant \( M > (2 K C \lambda / \tau_{\text{min}} \sqrt{\gamma_{\text{min}}})((3 - 2 \tau)/(2 - \tau)) \) and \( a_n = \sqrt{q(\log p)/n} \). For any \( \{\bar{\theta}^{(k)}\}_{k=1}^K \in \partial \mathcal{A} \), the Taylor series expansion gives \( G(\{\bar{\theta}^{(k)}\}_{k=1}^K) = I_1 + I_2 + I_3 \), where

\[
I_1 = - \sum_{k=1}^K (\nabla l(\bar{\theta}^{(k)}))_{S_k}^T \delta^{(k)}_{S_k},
\]

\[
I_2 = - \sum_{k=1}^K \delta^{(k)}_{S_k}^T \left[ \nabla^2 l(\bar{\theta}^{(k)}) + \alpha_k \delta^{(k)}_{S_k} \right]_{S_k} \delta^{(k)}_{S_k}, \text{ for some } \alpha_k \in [0, 1],
\]

\[
I_3 = \lambda \sum_{(j,j') \in S_\cup} \sum_{k=1}^K \left\{ \sum_{k=1}^K \left| \bar{\theta}^{(k)}_{j,j'} + \delta^{(k)}_{j,j'} \right| - \left( \sum_{k=1}^K \left| \bar{\theta}^{(k)}_{j,j'} \right| \right) \right\}^{1/2} - \left( \sum_{k=1}^K \left| \bar{\theta}^{(k)}_{j,j'} \right| \right)^{1/2}.
\]

Since \( C \lambda > (8 - 4 \tau) \sqrt{\gamma_{\text{min}}}/(1 - \tau) \), we have \( [(1 - \tau)/(2 - \tau)] C \lambda / \sqrt{\gamma_{\text{min}}} > 4 \). By Lemma 1,

\[
|I_1| \leq \sum_{k=1}^K \|\nabla l(\bar{\theta}^{(k)}))_{S_k}\|_\infty \|\delta^{(k)}_{S_k}\|_1 \leq [(1 - \tau) C \lambda M \gamma_{\text{min}}^{-1/2}/(2 - \tau)] (q \log p)/n .
\]

In addition, by condition \( q \sqrt{\log p}/n = o(1) \), Lemma 2 holds and thus

\[
I_2 \geq (\tau_{\text{min}}/2) \sum_{k=1}^K \|\bar{\theta}^{(k)}\|_2^2 \geq [(\tau_{\text{min}}/(2 K)] M^2 q (\log p)/n .
\]

Finally, by the triangular inequality and regularity condition (A),

\[
|I_3| \leq \lambda \sum_{(j,j') \in S_\cup} \sum_{k=1}^K \frac{\left| \bar{\theta}^{(k)}_{j,j'} + \delta^{(k)}_{j,j'} \right| - \left| \bar{\theta}^{(k)}_{j,j'} \right|}{\left( \sum_{k=1}^K \left| \bar{\theta}^{(k)}_{j,j'} \right| + \delta^{(k)}_{j,j'} \right)^{1/2} + (\sum_{k=1}^K \left| \bar{\theta}^{(k)}_{j,j'} \right|)^{1/2}} \leq (\lambda \gamma_{\text{min}}^{-1/2}) \sum_{k=1}^K \sum_{(j,j') \in S_\cup} \delta^{(k)}_{j,j'} \leq (\lambda q^{1/2} \gamma_{\text{min}}^{-1/2}) \sum_{k=1}^K \|\delta^{(k)}\|_2
\]

\[
(6.12) \leq (M C \lambda \gamma_{\text{min}}^{-1/2}) \{q(\log p)/n\}
\]

Then we have

\[
(6.13) G(\{\bar{\theta}^{(k)}\}_{k=1}^K) \geq M^2 q \log p/n \left( \frac{\tau_{\text{min}}}{2 K} - \frac{(1 - \tau) C \lambda}{(2 - \tau) M \gamma_{\text{min}}^{1/2}} - \frac{C \lambda}{M \gamma_{\text{min}}^{1/2}} \right) > 0.
\]
The last inequality uses the condition \( M > (2KC\lambda/\tau_{\min}\sqrt{\gamma_{\min}})(3-2\tau)/(2-\tau) \). Therefore, with probability tending to 1, we have \( \sum_{k=1}^{K} \| \hat{\theta}^{(k)} - \bar{\theta}^{(k)} \|_2 \leq M\sqrt{q(\log p)/n} \), and consequently claim (i) in Proposition 2 holds.

On the other hand, by the definition of \( \hat{\theta}^{(k)} \), we have \( \hat{\theta}_{j,j'} = 0 \) for all \((j, j') \in S_k^c\). By regularity condition (A) and Proposition 2 (i), for any \((j, j') \in S_k, k = 1, \ldots, K\), we have \( |\hat{\theta}_{j,j'}| \geq |\theta_{j,j'} - \bar{\theta}_{j,j'}| \geq \gamma_{\min}/2 > 0 \), when \( n \) is large enough.

**Proof of Proposition 5.** By Proposition 2, with probability tending to one, we have \( \hat{\theta}_{j,j'} \neq 0 \) for all \((j, j') \in S_k\). Since \( \{\hat{\theta}^{(k)}\}_{k=1}^{K} \) is a local maximizer of the restricted problem (6.1), with probability tending to one, \( \nabla_{j,j'}l(\hat{\theta}^{(k)}) = \lambda \sgn(\hat{\theta}_{j,j'}) / (\sum_{k=1}^{K} \hat{\theta}_{j,j'}^2)^{1/2} \), for all \((j, j') \in S_k\).

To show the second claim, we apply the mean value theorem and write

\[
\nabla l(\bar{\theta}^{(k)}) = \nabla l(\hat{\theta}^{(k)}) + r^{(k)} - \bar{Q}^{(k)} \hat{\theta}^{(k)},
\]

where \( r^{(k)} = \{\nabla^2 l(\hat{\theta}^{(k)}) + \alpha_k \hat{\theta}^{(k)} \} - \nabla^2 l(\bar{\theta}^{(k)}) \hat{\theta}^{(k)} \). After some simplifications, we have

\[
(6.14) \quad [\nabla l(\bar{\theta}^{(k)})]_{S_k^c} = [\nabla l(\hat{\theta}^{(k)})]_{S_k^c} + r^{(k)}_{S_k^c} - [Q_{S_k^c,S_k}(Q_{S_k^c,S_k})^{-1}] [\nabla l(\bar{\theta}^{(k)})]_{S_k} + r^{(k)}_{S_k} - [\nabla l(\hat{\theta}^{(k)})]_{S_k}
\]

and thus,

\[
\| [\nabla l(\hat{\theta}^{(k)})]_{S_k^c} \|_\infty \leq \| [\nabla l(\hat{\theta}^{(k)})]_{S_k^c} \|_\infty + \| r^{(k)}_{S_k^c} \|_\infty \\
+ \| Q_{S_k^c,S_k}(Q_{S_k^c,S_k})^{-1} \|_\infty \{ \| [\nabla l(\hat{\theta}^{(k)})]_{S_k^c} \|_\infty + \| r^{(k)}_{S_k} \|_\infty + \| [\nabla l(\bar{\theta}^{(k)})]_{S_k} \|_\infty \}
\leq (2-\tau)\| [\nabla l(\hat{\theta}^{(k)})]_{S_k^c} \|_\infty + (2-\tau)\| r^{(k)}_{S_k^c} \|_\infty + (1-\tau)\| [\nabla l(\bar{\theta}^{(k)})]_{S_k} \|_\infty \\
\leq [((1-\tau)C\lambda/\sqrt{\gamma_{\min}}\sqrt{\log p}/n + (2-\tau)\tau_{\max}M^2q(\log p)/n \\
+ (1-\tau)\lambda/\min_{(j,j') \in S_k} \sum_{k=1}^{K} |\hat{\theta}_{j,j'}|^{1/2}]\| 1/2
\]

\((6.15)\quad \lambda/\sqrt{\gamma_{\min}} + o_p(\lambda).\)

On the other hand, \( \lambda/\sum_{k=1}^{K} |\hat{\theta}_{j,j'}|^{1/2} = +\infty \) when \((j, j') \in S_k^c\). Otherwise, if \((j, j') \in S_k \setminus S_k^c\), then

\[
\lambda/\sum_{k=1}^{K} |\hat{\theta}_{j,j'}|^{1/2} \geq \lambda/[\sum_{k=1}^{K} (|\hat{\theta}_{j,j'} - \bar{\theta}_{j,j'}| + |\bar{\theta}_{j,j'}|)]^{1/2} \geq \lambda/\sqrt{\gamma_{\max}} \geq (2-2\tau)\lambda/\sqrt{\gamma_{\min}}.
\]
Thus, for any \((j, j') \in S_k^c \,(k = 1, \ldots, K)\), we have

\[
|\nabla_{j,j'} l(\hat{\theta}^{(k)})| \leq \max_{1 \leq k \leq K} \max_{(j, j') \in S_k^c} |\nabla_{j,j'} l(\hat{\theta}^{(k)})| < \min_{1 \leq k \leq K} \min_{(j, j') \in S_k^c} \frac{\lambda}{\sqrt{\sum_{k=1}^{K} |\hat{\theta}^{(k)}_{j,j'}|}} \leq \frac{\lambda}{\sqrt{\sum_{k=1}^{K} |\hat{\theta}^{(k)}_{j,j'}|}}.
\]

(6.16)
Fig 2. The networks used in three simulated examples. The black lines represent the common structure, whereas the red, blue and green lines represent the individual links in the three categories. $\rho$ is the ratio of the number of individual links to the number of common links.
Fig 3. Results for the balanced scenario ($n_1 = n_2 = n_3 = 200$). The ROC curves are averaged over 50 replications. $\rho$ is the ratio between the number of individual links and the number of common links.
Fig 4. Results for the unbalanced scenario ($n_1 = 150$, $n_2 = 300$, $n_3 = 450$). The ROC curves are averaged over 50 replications. $\rho$ is the ratio between the number of individual links and the number of common links.
Fig 5. The common and individual structures for the Senate voting data with an inclusion cutoff value of 0.4. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively. A solid line corresponds to a positive interaction effect and a dashed line to a negative interaction effect. The width of a link is proportional to the magnitude of the corresponding overall interaction effect. For each individual network, the links that only appear in this category are highlighted in grey.
The common and individual structures for the Senate voting data with an inclusion cutoff value of 0.6. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively. A solid line corresponds to a positive interaction effect and a dashed line to a negative interaction effect. The width of a link is proportional to the magnitude of the corresponding overall interaction effect. For each individual network, the links that only appear in this category are highlighted in grey.
FIG 7. Multidimensional scaling analysis for all the votes together, and the three individual categories. The nodes represent the 100 senators, with red, blue and purple node colors corresponding to Republican, Democrat, or Independent (Senator Jeffords), respectively.