Appendix

A. Proof of Theorem 1

Following the notations in Besag (1974); Yang et al. (2012; 2015), we define $Q(\mathbf{Y})$ as

$$Q(\mathbf{Y}) = \log\big(\mathbb{P}[\mathbf{Y}]/\mathbb{P}[\mathbf{0}]\big),$$

for any $\mathbf{Y} = (Y_1, \ldots, Y_p) \in \{0, 1, \ldots, M\}^p$ where $\mathbb{P}[\mathbf{0}]$ denotes the probability that all random variables in $\mathbf{Y}$ are identically zero. In this proof, we focus only on the pairwise MRF, however note that even with the higher order dependencies the theorem still holds. Now, lets consider the most general pairwise form of $Q(\mathbf{Y})$

$$Q(\mathbf{Y}) = \sum_{1\le s\le p} Y_s G_s(Y_s) + \sum_{1\le s<t\le p} Y_s Y_t G_{st}(Y_s, Y_t), \tag{14}$$

where G_s, G_{st} can be arbitrary functions. In the proof, we will connect this definition of $Q(\mathbf{Y})$ to the node-conditional distributions $\mathbb{P}[Y_s|Y_{\setminus s}]$ and investigate how given $\mathbb{P}[Y_s|Y_{\setminus s}]$ effects the forms of $G_s(\cdot)$ and $G_{st}(\cdot)$ in (14).

Note that $Q(\mathbf{Y})$ and $\mathbb{P}[Y_s|Y_{\setminus s}]$ are related as

$$\exp\big(Q(\mathbf{Y}) - Q(\bar{Y}_s)\big) = \mathbb{P}[Y]/\mathbb{P}[\bar{Y}_s] = \mathbb{P}[Y_s|Y_{\setminus s}]/\mathbb{P}[0|Y_{\setminus s}], \tag{15}$$

where $\bar{Y}_s := (Y_1, \ldots, Y_{s-1}, 0, Y_{s+1}, \ldots, Y_p)$.

The probability in (2) can be represented as

$$\begin{aligned}
&\log \mathbb{P}[Y_s|Y_{\setminus s}] \\
&= \log\Big\{\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big) - \exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big)\Big\} \\
&\quad - \log\Big\{1+\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big)\Big\} - \log\Big\{1+\exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big)\Big\}.
\end{aligned}$$

Substituting this and Equation (14) in Equation (15) we get

$$\begin{aligned}
&Y_s G_s(Y_s) + Y_s \sum_{t\in\setminus s} Y_t G_{st}(Y_s, Y_t) \\
&= \log\Big\{\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big) - \exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big)\Big\} \\
&\quad - \log\Big\{1+\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big)\Big\} - \log\Big\{1+\exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s=j] - \mu_s(Y_{\setminus s})\Big)\Big\} \\
&\quad - \theta_{s;0} + \mu_s(Y_{\setminus s}) + \log\Big\{1+\exp\big(\theta_{s;0} - \mu_s(Y_{\setminus s})\big)\Big\}.
\end{aligned} \tag{16}$$

By setting $Y_t = 0, \forall t \neq s$ in the above equation, we obtain the first order function $Y_s G_s(Y_s)$:

$$\begin{aligned}
Y_s G_s(Y_s) &= \log\Big\{\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s=j]\Big) - \exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s=j]\Big)\Big\} \\
&\quad - \log\Big\{1+\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s=j]\Big)\Big\} - \log\Big\{1+\exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s=j]\Big)\Big\} \\
&\quad - \theta_{s;0} + \log\Big\{1+\exp\big(\theta_{s;0}\big)\Big\}
\end{aligned} \tag{17}$$

where we assume that $\mu_s(\mathbf{0}) = 0$ without loss of generality; if $\mu_s(\mathbf{0}) = c$ for some nonzero c, then we simply replace $\theta_{s;j}$ with $\theta'_{s;j}$ where $\theta'_{s;j} = \theta_{s;j} + c$.

Suppose nodes s and t are neighbors in graph G, i.e. $Y_sY_tG_{st}(Y_s, Y_t) \neq 0$. Setting $Y_r = 0$ for all $r \notin \{s, t\}$, we obtain

$$
\begin{aligned}
&Y_sG_s(Y_s) + Y_sY_tG_{st}(Y_s, Y_t) = \\
&\log\left\{\exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s = j] - \mu_s(0,\ldots,Y_t,\ldots,0)\Big) - \exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s = j] - \mu_s(0,\ldots,Y_t,\ldots,0)\Big)\right\} \\
&- \log\left\{1 + \exp\Big(\sum_{j=0}^{M}\theta_{s;j}\,\mathcal{I}[Y_s = j] - \mu_s(0,\ldots,Y_t,\ldots,0)\Big)\right\} \\
&- \log\left\{1 + \exp\Big(\sum_{j=0}^{M}\theta_{s;j-1}\,\mathcal{I}[Y_s = j] - \mu_s(0,\ldots,Y_t,\ldots,0)\Big)\right\} \\
&- \theta_{s;0} + \mu_s(0,\ldots,Y_t,\ldots,0) + \log\left\{1 + \exp\big(\theta_{s;0}\big)\right\}. \qquad (18)
\end{aligned}
$$

Combining (17) and (18) yields

$$
\begin{aligned}
&Y_sY_tG_{st}(Y_s, Y_t) = \\
&\log\left\{\sum_{j=0}^{M}\frac{\exp\Big(\theta_{s;j} - \mu_s(0,\ldots,Y_t,\ldots,0)\Big) - \exp\Big(\theta_{s;j-1} - \mu_s(0,\ldots,Y_t,\ldots,0)\Big)}{\exp\big(\theta_{s;j}\big) - \exp\big(\theta_{s;j-1}\big)}\,\mathcal{I}[Y_s = j]\right\} \\
&- \log\left\{\sum_{j=0}^{M}\frac{1 + \exp\Big(\theta_{s;j} - \mu_s(0,\ldots,Y_t,\ldots,0)\Big)}{1 + \exp\big(\theta_{s;j}\big)}\,\mathcal{I}[Y_s = j]\right\} \\
&- \log\left\{\sum_{j=0}^{M}\frac{1 + \exp\Big(\theta_{s;j-1} - \mu_s(0,\ldots,Y_t,\ldots,0)\Big)}{1 + \exp\big(\theta_{s;j-1}\big)}\,\mathcal{I}[Y_s = j]\right\} \\
&+ \mu_s(0,\ldots,Y_t,\ldots,0) + \log\left\{\frac{1 + \exp\big(\theta_{s;0} - \mu_s(0,\ldots,Y_t,\ldots,0)\big)}{1 + \exp\big(\theta_{s;0}\big)}\right\}. \qquad (19)
\end{aligned}
$$

Similarly, we can also obtain $Y_sY_tG_{st}(Y_s, Y_t)$ by considering the difference $Q(Y) - Q(\bar{Y}_t)$, instead of $Q(Y) - Q(\bar{Y}_s)$ in (15). Using $Q(Y) - Q(\bar{Y}_t)$, we obtain

$$
\begin{aligned}
&Y_sY_tG_{st}(Y_s, Y_t) = \\
&\log\left\{\sum_{j=0}^{M}\frac{\exp\Big(\theta_{t;j} - \mu_t(0,\ldots,Y_s,\ldots,0)\Big) - \exp\Big(\theta_{t;j-1} - \mu_t(0,\ldots,Y_s,\ldots,0)\Big)}{\exp\big(\theta_{t;j}\big) - \exp\big(\theta_{t;j-1}\big)}\,\mathcal{I}[Y_t = j]\right\} \\
&- \log\left\{\sum_{j=0}^{M}\frac{1 + \exp\Big(\theta_{t;j} - \mu_t(0,\ldots,Y_s,\ldots,0)\Big)}{1 + \exp\big(\theta_{t;j}\big)}\,\mathcal{I}[Y_t = j]\right\} \\
&- \log\left\{\sum_{j=0}^{M}\frac{1 + \exp\Big(\theta_{t;j-1} - \mu_t(0,\ldots,Y_s,\ldots,0)\Big)}{1 + \exp\big(\theta_{t;j-1}\big)}\,\mathcal{I}[Y_t = j]\right\} \\
&+ \mu_t(0,\ldots,Y_s,\ldots,0) + \log\left\{\frac{1 + \exp\big(\theta_{t;0} - \mu_t(0,\ldots,Y_s,\ldots,0)\big)}{1 + \exp\big(\theta_{t;0}\big)}\right\}. \qquad (20)
\end{aligned}
$$

At this point, (19) and (20) should be the same for all possible pairs of Y_s and Y_t.

Now, consider the case of $Y_s = 1$ and $Y_t = 1$. For this fixed setting, both $\mu_s(0, \ldots, 1, \ldots, 0)$ and $\mu_t(0, \ldots, 1, \ldots, 0)$ are fixed constants; let us call them c_1 and c_2, respectively. Then equating (19) and (20) we get

$$\begin{aligned}
&\log\left\{\frac{\exp\left(\theta_{s;1} - c_1\right) - \exp\left(\theta_{s;0} - c_1\right)}{\exp\left(\theta_{s;1}\right) - \exp\left(\theta_{s;0}\right)}\right\} - \log\left\{\frac{1 + \exp\left(\theta_{s;1} - c_1\right)}{1 + \exp\left(\theta_{s;1}\right)}\right\} + c_1 \\
&= \log\left\{\frac{\exp\left(\theta_{t;1} - c_2\right) - \exp\left(\theta_{t;0} - c_2\right)}{\exp\left(\theta_{t;1}\right) - \exp\left(\theta_{t;0}\right)}\right\} - \log\left\{\frac{1 + \exp\left(\theta_{t;1} - c_2\right)}{1 + \exp\left(\theta_{t;1}\right)}\right\} \\
&- \log\left\{\frac{1 + \exp\left(\theta_{t;0} - c_2\right)}{1 + \exp\left(\theta_{t;0}\right)}\right\} + c_2 + \log\left\{\frac{1 + \exp\left(\theta_{s;0} - c_2\right)}{1 + \exp\left(\theta_{s;0}\right)}\right\}. \qquad (21)
\end{aligned}$$

Trivially, this equality cannot hold for all values of $\theta_{s;0}, \theta_{s;1}, \theta_{t;0}, \theta_{t;1} \in \mathbb{R}$. This shows that there can't exist a $Q(\mathbf{Y})$ that is consistent with node conditional distributions in Equation (2), which in turn entails that there is no consistent joint distribution for all choices of the parameters.

B. Proof of Theorem 2

The same strategy as in the proof of Theorem 1 can be adopted here. To this end, we derive the form of equation: $\exp\left(Q(\mathbf{Y}) - Q(\bar{Y}_s)\right) = \mathbb{P}[Y_s|Y_{\setminus s}]/\mathbb{P}[0|Y_{\setminus s}]$

$$\begin{aligned}
Y_s G_s(Y_s) + Y_s \sum_{t \in \setminus s} Y_t G_{st}(Y_s, Y_t) &= \sum_{j=0}^{M-1} \Big(\left(\theta_{s;j} - \mu_s(Y_{\setminus s})\right) \mathcal{I}[Y_s = j] \\
&+ \log\left\{1 + \exp\left(\theta_{s;j} - \mu_s(Y_{\setminus s})\right)\right\} \mathcal{I}[Y_s < j] \Big) - \theta_{s;0} + \mu_s(Y_{\setminus s}) \\
&- \sum_{j=1}^{M-1} \log\left\{1 + \exp\left(\theta_j - \mu_s(Y_{\setminus s})\right)\right\} \\
&= \sum_{j=0}^{M-1} \left(\theta_{s;j} - \mu_s(Y_{\setminus s})\right) \mathcal{I}[Y_s = j] - \sum_{j=1}^{\min\{Y_s, M-1\}} \log\left\{1 + \exp\left(\theta_{s;j} - \mu_s(Y_{\setminus s})\right)\right\} \\
&- \theta_{s;0} + \mu_s(Y_{\setminus s}), \qquad (22)
\end{aligned}$$

where we used the definition of $\mathbb{P}[Y_s|Y_{\setminus s}]$ given in Equation (4).

We can obtain the first order function $Y_s G_s(Y_s)$ by setting $Y_t = 0$ for all $t \neq s$ in (22):

$$Y_s G_s(Y_s) = \sum_{j=0}^{M-1} \theta_{s;j} \mathcal{I}[Y_s = j] - \sum_{j=1}^{\min\{Y_s, M-1\}} \log\left\{1 + \exp(\theta_{s;j})\right\} - \theta_{s;0} \qquad (23)$$

where we use the fact that $\mu_s(\mathbf{0}) = 0$.

Suppose nodes s and t are neighbors in graph G, i.e. $Y_s Y_t G_{st}(Y_s, Y_t) \neq 0$. Setting $Y_r = 0$ for all $r \notin \{s, t\}$, we obtain

$$\begin{aligned}
Y_s G_s(Y_s) + Y_s Y_t G_{st}(Y_s, Y_t) &= \sum_{j=0}^{M-1} \Big(\theta_{s;j} - \mu_s(0, \ldots, Y_t, \ldots, 0)\Big) \mathcal{I}[Y_s = j] \\
- \sum_{j=1}^{\min\{Y_s, M-1\}} &\log\left\{1 + \exp\left(\theta_{s;j} - \mu_s(0, \ldots, Y_t, \ldots, 0)\right)\right\} - \theta_{s;0} + \mu_s(0, \ldots, Y_t, \ldots, 0) \qquad (24)
\end{aligned}$$

Combining (23) and (24) yields

$$Y_sY_tG_{st}(Y_s,Y_t) = \mu_s(0,\ldots,Y_t,\ldots,0)\,\mathcal{I}[Y_s = M] - \sum_{j=1}^{\min\{Y_s,M-1\}} \log\left\{\frac{1+\exp\left(\theta_{s;j} - \mu_s(0,\ldots,Y_t,\ldots,0)\right)}{1+\exp(\theta_{s;j})}\right\} \tag{25}$$

Similarly, we can also obtain $Y_sY_tG_{st}(Y_s,Y_t)$ by considering the difference $Q(Y) - Q(\bar{Y}_t)$, instead of $Q(Y) - Q(\bar{Y}_s)$ in (15). Using $Q(Y) - Q(\bar{Y}_t)$, we obtain

$$Y_sY_tG_{st}(Y_s,Y_t) = \mu_t(0,\ldots,Y_s,\ldots,0)\,\mathcal{I}[Y_t = M] - \sum_{j=1}^{\min\{Y_t,M-1\}} \log\left\{\frac{1+\exp\left(\theta_{t;j} - \mu_t(0,\ldots,Y_s,\ldots,0)\right)}{1+\exp(\theta_{t;j})}\right\} \tag{26}$$

At this point, (25) and (26) should be the same for all possible pairs of Y_s and Y_t.

As in the proof of Theorem 1, let us consider the case of $Y_s = 1$ and $Y_t = 1$ where $M \geq 1$. Again, since both $\mu_s(0,\ldots,1,\ldots,0)$ and $\mu_t(0,\ldots,1,\ldots,0)$ are fixed constants, by the equality of (25) and (26) we have

$$\log\left\{\frac{1+\exp\left(\theta_{s;1} - c_1\right)}{1+\exp(\theta_{s;1})}\right\} - c_1\,\mathcal{I}[1 = M] = \log\left\{\frac{1+\exp\left(\theta_{t;1} - c_2\right)}{1+\exp(\theta_{t;1})}\right\} - c_2\,\mathcal{I}[1 = M] \tag{27}$$

If $M = 1$, then (27) can be reduced as

$$\log\left\{\frac{1+\exp\left(\theta_{s;1} - c_1\right)}{1+\exp(\theta_{s;1})}\right\} = \log\left\{\frac{1+\exp\left(\theta_{t;1} - c_2\right)}{1+\exp(\theta_{t;1})}\right\} + c_1 - c_2 \tag{28}$$

and if $M > 1$, (27) can be reduced as

$$\log\left\{\frac{1+\exp\left(\theta_{s;1} - c_1\right)}{1+\exp(\theta_{s;1})}\right\} = \log\left\{\frac{1+\exp\left(\theta_{t;1} - c_2\right)}{1+\exp(\theta_{t;1})}\right\} \tag{29}$$

In any case, the equality cannot hold for all values in $\theta_{s;1}, \theta_{t;1} \in \mathbb{R}$. This shows that there can't exist a $Q(\mathbf{Y})$ that is consistent with node conditional distributions in Equation (4), which in turn entails that there is no consistent joint distribution for all choices of the parameters.

C. Proof of Theorem 4

To show that our final estimate of Σ is consistent, we first show that our estimate $\tilde{\Sigma}$ from Step 1 concentrates well around Σ^*, by appealing to the result in (Mei et al., 2017).

Lemma 1. *Under conditions (C-2), (C-3), there exists some known quantities C_1 and C_2 depending on L_1, L_2, L_3, M, α, γ, δ, and $c > 2$ such that if $n \geq 4C_1 \log p \log n$, then*

$$|\widetilde{\Sigma}_{jk} - \Sigma^*_{jk}| \leq C_2\sqrt{\frac{\log p \log n}{n}} \tag{30}$$

with at least probability $1 - p^{-c}$.

Proof. To prove the Lemma we use Theorem 2 of Mei et al. (2017) which shows that, under certain regularity conditions, there is a one-to-one mapping between the critical points of the empirical risk and the population risk. Moreover Mei et al. (2017) obtain a bound for the gap between any critical point of the empirical risk and the *corresponding critical point* of population risk. We first note that the three required assumptions for Theorem 2 of (Mei et al., 2017) hold for the sample loss defined in Equation (10), on the domain $[-1+\delta, 1-\delta]$:

- The derivative of $\ell_{jk}(\sigma; \Theta^*, \mathbb{Y}_n)$ is equal to $\sum_{a=0}^{M}\sum_{b=0}^{M} \frac{n_{ab}}{n} \frac{\phi'_{ab;jk}(\sigma;\Theta^*)}{\phi_{ab;jk}(\sigma;\Theta^*)}$, hence $|\ell'(\sigma)|$ is upper bounded by γL_1. Therefore, $\ell'_{jk}(\sigma) - \ell'_{jk}(\Sigma^*_{jk})$ is bounded between $[-2\gamma L_1, 2\gamma L_1]$ and sub-Gaussian with a parameter $\rho_1^2 := (2\gamma L_1)^2$.
- Similarly, the absolute of second derivative $|\ell''_{jk}(\sigma)|$ is bounded by $(\gamma L_2 + \gamma^2 L_1^2)$, and $|\ell''_{jk}(\sigma) - \ell''_{jk}(\Sigma^*_{jk})|$ is sub-Gaussian with a parameter $\rho_2^2 := \left(\gamma L_2 + \gamma^2 L_1^2\right)^2$.
- Finally, $|\ell'''_{jk}(\sigma)| \leq \left(\gamma L_3 + 12\gamma^2 L_1 L_2 + 2\gamma^3 L_1^3\right)$, hence $\ell''_{jk}(\cdot)$ is Lipschitz with $\rho_3 := \left(\gamma L_3 + 12\gamma^2 L_1 L_2 + 2\gamma^3 L_1^3\right)$.

From an application of Theorem 2 of (Mei et al., 2017), it follows that the global maximizer of our empirical log-likelihood in (10) converges to a corresponding critical point of the expected log-likelihood $\bar{\ell}_{jk}$. But since the global maximizer of our empirical log-likelihood in (10) is precisely the MLE, which is consistent, it follows that the corresponding critical point of the expected log-likelihood is precisely the true covariance parameter Σ^*_{jk}. □

Now, we can directly appeal to the recent results on the sparsistency of graphical lasso:

Lemma 2. *Suppose that with probability at least* $1 - c_1 p^{-c_2}$,

$$\|\widetilde{\Sigma} - \Sigma^*\|_\infty \leq c_3 \sqrt{\frac{\log p}{n}}. \tag{31}$$

Consider our estimator (11) *with regularization parameter* $\lambda_n = (8c_3/\alpha)\sqrt{\log(p)/n}$. *Then, if the sample size* n *is lower bounded as* $n \geq \frac{2}{c_2} \max\left\{\frac{2K_{\Gamma^*}(1+\frac{8}{\alpha})}{\beta_{\min}}, (1+8/\alpha)d\max\{K_{\Sigma^*}K_{\Gamma^*}, K^3_{\Sigma^*}K^2_{\Gamma^*}\}\right\}^2 (\log p + \log c_1)$, *then the inverse of estimate* $\widehat{\Sigma}$ *satisfies the bound as*

$$\left\|\widehat{\Sigma}^{-1} - (\Sigma^*)^{-1}\right\|_\infty \leq 2c_3(1 + 8\alpha^{-1})K_{\Gamma^*}\sqrt{\frac{\log p}{n}}$$

and, moreover, the graph structure of latent Gaussian encoded in $(\Sigma^*)^{-1}$ *is consistently recovered by* $\widehat{\Sigma}^{-1}$ *as long as* $\beta_{\min} := \min_{ij} \left|[(\Sigma^*)^{-1}]_{ij}\right| \geq 2c_3(1 + 8\alpha^{-1})K_{\Gamma^*}\sqrt{\frac{\log p}{n}}$, *with probability at least* $1 - c_1 p^{-c_2+2}$.

The lemma follows from an application of Theorems 1 and 2 of (Ravikumar et al., 2008).

D. Experiments

D.1. Simulations

Data from Probit Model: Here we present results from simulations when the data is generated from a Probit model with grid and random graph structures. We first describe the graphs and exact model parameters that were used in these simulations.

- **Grid Graph:** We select a 10×5 grid graph, with 10 rows and 5 columns. For all the vertical edges we set the corresponding entries in inverse covariance matrix as -0.25 and for all the horizontal edges we set the corresponding entries as 0.25. We set the thresholds (θ) at node j as : $\theta^{(j)} = [-\text{Inf}, -10, -0.7, 0.7, 10, \text{Inf}]$. Figure 5 presents the results from this simulation.
- **Random Graph:** We use the same graph generation procedure as (Liu et al., 2012). For each node j in the graph we associate a bivariate random variable $U_j = (U_{1,j}, U_{2,j}) \in [0,1]^2$ uniformly sampled from a unit square. An edge is included between (j,k) with probability:

  $$\frac{1}{\sqrt{2\pi}} \exp - \frac{\|U_j - U_k\|_2^2}{0.15}.$$

  If an edge is added between (j,k) then the corresponding entry in the inverse covariance matrix is set to $\omega \in (-1,1)$. We use the same thresholds (θ) as in grid graph, to convert the latent variables to ordinal variables. Figure 6 presents the results for $\omega = 0.8, -0.65$.

ROC plots on large n: Figure 7 provides the ROC plots for $n = 200, 400$, when the data is generated from probit model.

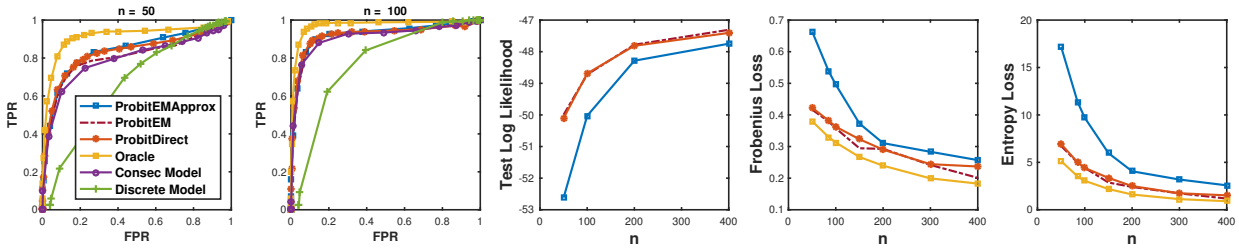

Figure 5. Comparison of various estimators when the data is generated from a probit model with grid graph structure. The left two plots show ROC curves for $n = 50, 100$. The right three plots show performance on test log likelihood, frobenius and entropy losses.

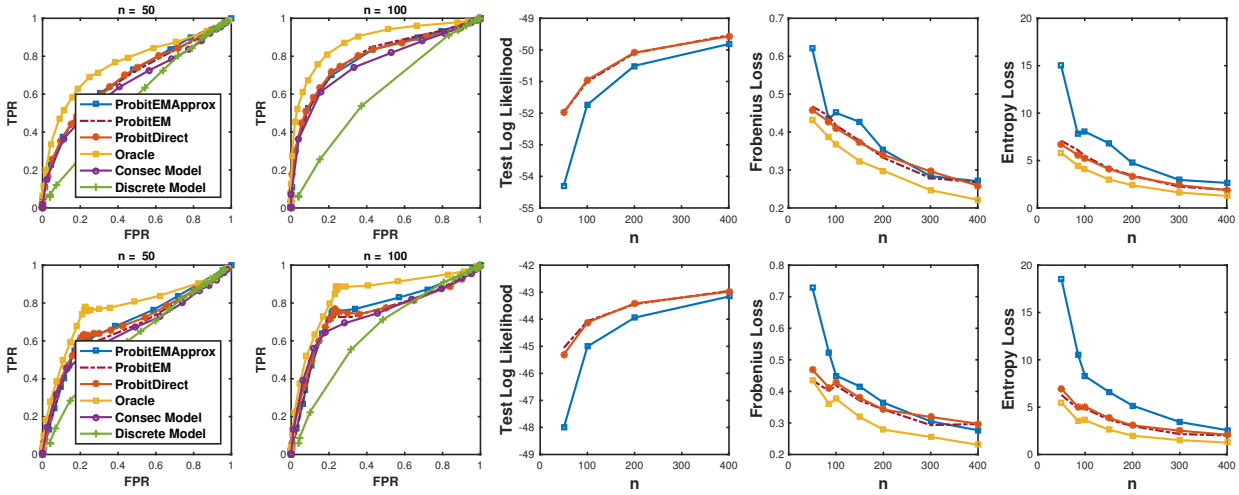

Figure 6. Comparison of various estimators when the data is generated from a probit model with chain graph structure. Top and bottom rows correspond to $\omega = 0.8$ and $\omega = -0.65$ respectively. The left two columns show ROC curves for $n = 50, 100$. The right three columns show performance on test log likelihood, frobenius and entropy losses.

D.2. Train Time

We now compare the training time of our estimators with *ProbitEMApprox*. For a comparison of training times of *ProbitEM* and *ProbitEMApprox* see Guo et al. (2015), where the authors show that *ProbitEM* is ≈ 1 order of magnitude slower than *ProbitEMApprox*.

In this experiment, we fix $p = 200$ and sample data from a probit model with chain graph structure (with ω in Equation (13) set to 0.3). Note that the choice of regularization parameter can effect the training time of each of these estimators. So, we report the training time of these methods averaged over different choices of regularization parameters. Table 1 shows the results from this experiment[1].

		ProbitDirect	*Consec Model*	*ProbitEMApprox*
p = 200	$n = 100$	20.66	167.05	81.96
	$n = 200$	19.48	196.33	63.81
	$n = 400$	17.80	246.65	51.08

Table 1. Training time (in seconds) of *ProbitEMApprox*, *Consec Model*, *ProbitDirect*.

ProbitEMApprox solves glasso in each iteration of EM, whereas *ProbitDirect* only solves glasso once. As a result *ProbitEMApprox* is much slower than *ProbitDirect*. Although *Consec Model* is slower than other estimators, its training can be performed in a distributed fashion. So it can be used to learn very large networks.

[1] *ProbitEMApprox* is implemented R and *ProbitDirect*, *Consec Model* are implemented in MATLAB. *ProbitEMApprox* was run for a maximum of 25 iterations. Step 1 in estimation of Σ of *ProbitDirect* was run only once for different choices of regularization parameter.

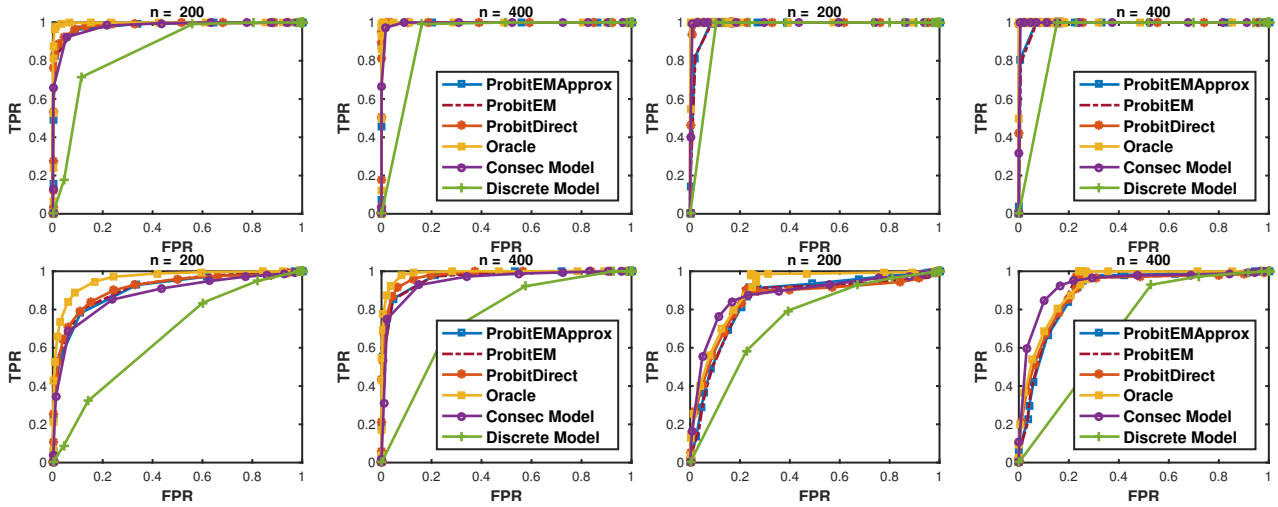

Figure 7. ROC plots for $n = [200, 400]$, when the data is generated form a probit model. The two plots on the top left are for the chain graph described in the main part of the paper with $\omega = -0.3$ and the two plots on top right are for the chain graph with $\omega = -0.9$. The bottom left plots correspond to the random graph described in the appendix with $\omega = 0.8$ and bottom right plots correspond to the random graph with $\omega = -0.65$.

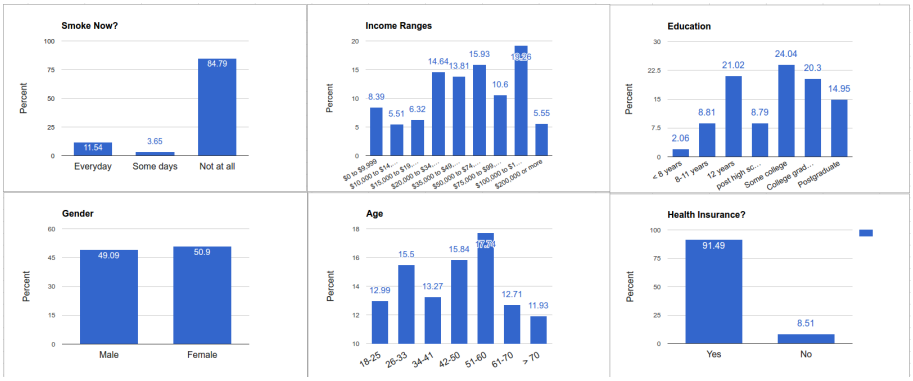

Figure 8. Summary statistics of the HINTS-FDA dataset.

D.3. HINTS-FDA Study

D.3.1. Data Preprocessing

Missing values: The original data collected through the survey has missing responses for a number of questions. Some of these missing responses have already been imputed in the data that was made publicly available through the HINTS website. In our analysis, we impute the rest of the missing responses using median. If a question has more than 50% missing responses then we don't use the responses for that question in our analysis.

Categorical Data: Some of the questions in the survey have categorical responses (*e.g.,* Marital Status). We use *one hot encoding* technique for such responses to convert them into binary format.

Count Data: For responses which are neither categorical nor ordinal (such as *age*, *how many hours does a person watch TV* etc.,) we binned the responses into a fixed number of categories and converted them into ordinal variables. For example, for *number of hours of TV watched per week* we created 5 buckets : *<1hr, 2-3hrs, 3-5hrs, 5-10hrs, >10hrs*.

Node Name	Question	Possible Responses
CigarettesHarmHealth	How long do you think someone has to smoke cigarettes before it harms their health?	1-'< 1 year', 2-'1 year' 3 - '5 years', 4 - '10 years' 5 - '20 years or more'
DailySmokelessHarm	How much do you think people harm themselves when they use smokeless tobacco every day?	1-No harm, 2-Little harm, 3-Some harm, 4 - A lot of harm
Education	What is the highest grade or level of schooling you completed?	1-'Less than 8 years', 2-'8 through 11 years' , 3-'12 years or completed high school', 4-'Post high school training' 5-'Some college', 6-'College graduate', 7-'Postgraduate'
FewCigarettesHarmHealth	How much do you think people harm themselves when they smoke a few cigarettes every day?	1-No harm, 2-Little harm, 3-Some harm, 4 - A lot of harm
HealthInsurance	Do you have any kind of health care coverage?	1-Yes, 2-No
IncomeRanges_IMP	what is the combined annual income of your family?	1-'$0-$9,999', 2-'$10,000-$14,999' , 3-'$15,000-$19,999', 4-'$20,000-$34,999' 5-'$35,000-$49,999', 6-'$50,000-$74,999', 7-'$75,000-$99,999', 8-'$100,000-$199,999' 9- '$200,000 or more'
Mexican	Are you a Mexican?	1-'Yes', 2-'No'
PhoneInHome	Is there at least one telephone inside your home?	1-Yes, 2-No
Retired	Occupation Status	1-Not Retired, 2-Retired
SmokeNow	Do you now smoke cigarettes every day, some days or not at all?	1-Everyday, 2-Some days, 3-Not at all
Student	Occupation Status	1-Not Student, 2-Student
TobaccoEffects_TV	how often have you seen, heard, or read a message about the health effects of tobacco use on TV?	1-'Never', 2-'A couple of times', 3-'Lot of times'
White	Are you a White?	1-'Yes', 2-'No'

Table 2. Table describing the questions corresponding to some of the nodes in Figures 4, 9.

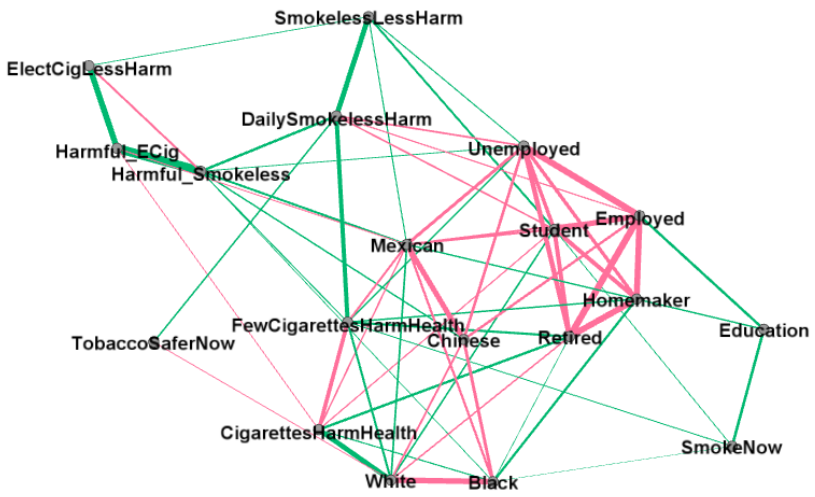

Figure 9. The estimated latent graph structure for variables corresponding to perceptions of smoking risks and *SmokeNow*. The graph is generated from the marginal distribution of the corresponding variables. Green edges represent positive partial correlations and red edges represent negative partial correlations. Edge thickness is proportional to the magnitude of the partial correlation.