Ordinal Data Analysis via Graphical Models

Arun Sai Suggala

Sunday 5th November, 2017

Abstract

Background. Undirected graphical models or Markov random fields (MRFs) are very popular for modeling multivariate probability distributions. A considerable amount of work on MRFs has focused on modeling continuous variables and unordered categorical variables also called as *nominal* variables. However, data from many real world applications involve ordered categorical variables also called as *ordinal* variables (*e.g.,* movie ratings on Netflix which can be ordered from 1 to 5 stars). While one can model ordinal variables using models designed for continuous or nominal variables, this can result in incorrect inferences about the variables. While, recent work has designed graphical models for modeling ordinal data, the proposed estimator for learning this model involves optimization of a difficult non-convex problem which is both computationally expensive and doesn't come with statistical guarantees.
Aim. Given multivariate ordinal data, we aim to estimate the joint probability distribution and the conditional dependency structure in the data. To this end, we provide a new estimator for ordinal probit model (a graphical model for ordinal data), that is computationally efficient and which comes with statistical guarantees.
Data. We analyze HINTS-FDA dataset, which is a survey on how people access and use smoking and cancer related information and how they perceive risks of smoking.
Results. We apply our estimator on HINTS-FDA data to understand the smoking behavior of people and their perceptions of smoking risks. Our analysis suggests that people who smoke, perceive smoking as less harmful than people who don't smoke and the lack of awareness of smoking risks could be a reason why many people smoke.
Conclusion. We have proposed a new estimator for learning ordinal probit model that is computationally tractable and can be easily scaled to large datasets. We empirically corroborated the superior performance of our estimator for the probit model.

Keywords: ordinal variable, graphical model, MRF.

DAP Committee members: Pradeep Ravikumar; Jeremy Weiss

1 Introduction

With the modern big data era, there has been considerable interest in learning joint distributions over a large number of variables. As a primary tool for handling joint distributions, graphical models (Lauritzen, 1996) not only provide us an efficient way of representing joint distributions compactly, but also use graphs to represent the interactions among the random variables, which is an object of additional interest in many scientific disciplines. Undirected graphical models, also known as Markov random fields (MRFs), represent joint distributions as the product of compatibility functions on cliques of a graph. This factorization of a distribution encodes the conditional independences among random variables, identified with reachability in the corresponding graph structure.

MRFs are extensively used in a variety of fields, including natural language processing (Manning and Schütze (1999)), biology (Friedman (2004)) and medicine (Allen and Liu (2012)). Given their wide applicability, MRFs have been extensively studied by researchers. The two most popular instances in the family of undirected graphical models are Gaussian graphical models (Speed and Kiiveri, 1986; Rue and Held, 2005) for continuous (and bell-shaped) data, discrete graphical models such as Ising model (Ising, 1925; Jalali et al., 2011) for nominal data (an example of a nominal variable is the religious affilication: *Catholic, Muslim, Jewish, other*), or mixed cases of these two instances (Lauritzen and Wermuth, 1989; Yang et al., 2014). However, variables that occur in many real world applications have ordered categorical scales. For example, in medical data, diseases are graded from *mild* to *fatal*, severity of an injury is rated from *mild injury* to *death*, stages of a disease is rated from *I* to *III*. Ordinal variables also occur very commonly in data collected from surveys. For example, each subject taking a survey could be asked to respond to a question using categories such as *strongly disagree, disagree, undecided, agree, strongly agree*, users of an online service could be asked to rate their experience from *one star* to *five star*. These examples clearly show that ordinal data is pervasive in many real world applications.

While there has been considerable work on both learning and inference with discrete graphical models, there has been very limited work on designing graphical models for ordinal variables. In a recent work Guo et al. (2015) introduced a graphical model for ordinal variables, called *probit* graphical model, which is based on the assumption that the ordinal variables are generated by discretizing a latent multivariate Gaussian random vector. They proposed a Maximum Likelihood (ML) Estimator for the probit model. However, learning the estimator involves optimization of a difficult non-convex problem. The authors propose an approximate EM algorithm for learning the estimator from the data. Consequently, their estimator doesn't come with optimization and statistical guarantees. Moreover, the EM algorithm presented in the paper is computationally expensive. In this work we present a new estimator for learning the probit model. Instead of solving the global Maxi-

mum Likelihood Estimation (MLE) problem, we solve multiple local MLE problems which can be solved very efficiently. Moreover, these local problems can be solved in parallel and as a result our estimator can easily scale to large datasets.

We apply our estimator to analyze HINTS-FDA dataset, which is a survey on how people access and use smoking and cancer related information and how they perceive risks of smoking. Through this analysis, we aim to identify the key social indicators that are associated with smoking, understand how the perceptions of smoking risks vary from people who smoke to who do not smoke and how people who smoke access health information. Such an understanding can be helpful in designing strategies to communicate smoking related health information more effectively.

2 Background and related work

As pointed out in Section 1, ordinal data is common in practice, especially in applications throughout biomedical and social sciences. One possible approach for modeling multivariate ordinal data is to ignore the order in categories and treat it as nominal data and use the models that were developed for nominal data. However ignoring the structure in the data has many disadvantages : a) this requires us to estimate complex models with huge number of parameters b) this can lead to incorrect inferences about the variables. Another possible approach for modeling ordinal data is to treat the ordinal variables as continuous variables and use the models that were developed for continuous data. This approach requires us to first convert the ordinal scale to a continuous scale. Unfortunately, there is no unique way to perform this conversion. For instance, consider the movie ratings example in which users rate movies as one of *awful, bad, not bad, good, excellent*. Here is one possible mapping of categories to the continuous scale: *awful-1, bad-2, not bad-3, good-4, excellent-5*. Another possible mapping is as follows: *awful-0, bad-2, not bad-5, good-8, excellent-10*. However there is no reason to prefer one mapping over the other and both these mappings can result in different inferences about the variables. This shows that there is a need for graphical models that can model ordinal data by taking into consideration the structure in the data.

Before presenting the related work on ordinal MRFs, we first review relevant literature on *univariate* ordinal distributions. A popular category of univariate ordinal distributions are based on the natural generative assumption that the ordinal variable is a quantization of a real-valued latent variable. Common distributions imposed on the latent variable include the logistic distribution, in which case it reduces to the classical cumulative ratio model (Agresti, 2010), as well as the more popular standard normal distribution, in which case it is called the ordered probit model (Becker and Kennedy, 1992).

Extensions of the univariate latent quantized model to multivariate distributions have been considered in the literature. Here the ordinal random vector is naturally modeled as a quantization of a real-valued latent random vector. Here, the efforts

have focused on the use of the multivariate normal distribution for the latent random vector, due plausibly to its more convenient mathematical nature; the resulting model is also known as the multivariate probit model (Ashford and Sowden, 1970; Amemiya, 1974). But even with this modeling assumption, the likelihood of the observed ordinal random vector is not available in closed-form, is considerably complex due to the presence a multi-dimensional integral, and in particular is non-convex, so that learning the model given just the ordinal observations is typically computationally intractable. There have been some efforts to propose computationally amenable approximations of the MLE, including MCMC based estimates (Chib and Greenberg, 1998), and expectation maximization Guo et al. (2015). However these approximations are still computationally expensive and do not come with statistical guarantees.

3 Probit Graphical Model

In this section, we formally describe the Probit graphical model and present our estimator for estimating the model from data. Suppose $Y := (Y_1, \ldots, Y_p)$ is a p-dimensional ordinal random vector, with each variable Y_s taking values from an ordinal set $\mathcal{Y}_s := \{0, \ldots, M\}$ (Note that to simplify the notation we use the same number of categories for all variables. The algorithm and analysis we present here are valid even with differing numbers of categories). In the probit model, the random vector Y is assumed to be generated from a latent multivariate Gaussian random vector $Z = (Z_1, \ldots, Z_p)$, where $Z \sim \mathcal{N}(0, \Sigma^*)$ and $Z_i \sim \mathcal{N}(0, 1)\ \forall i \in [1, p]$ (i.e, the diagonal entries of Σ are equal to 1). Each Y_i is obtained through discretization of Z_i as follows: $Y_i = k, \quad \text{iff} \quad Z_i \in [\theta_{k-1}^{(i)}, \theta_k^{(i)})$, where $\{\theta_k^{(i)}\}_{k=-1}^{M}$ is the set of thresholds, $\theta_{-1}^{(i)} = -\infty$, $\theta_M^{(i)} = \infty$. Then the density function of Y, $\mathbb{P}(Y; \Sigma^*, \Theta^*)$, is given by:

$$\begin{aligned} &\mathbb{P}(\theta_{Y_1-1}^{(1)} \leq Z_1 < \theta_{Y_1}^{(1)}, \ldots, \theta_{Y_p-1}^{(p)} \leq Z_p < \theta_{Y_p}^{(p)}; \Sigma) \\ &= \int_{z \in C(Y, \Theta^*)} \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} z \Sigma^{-1} z^T\right) dz \end{aligned} \tag{1}$$

where $\Theta^* = \{\theta_k^{(j)} : j \in [1, p], k \in [-1, M]\}$ and $C(Y, \Theta^*)$ is the hypercube defined by $[\theta_{Y_1-1}^{(1)}, \theta_{Y_1}^{(1)}) \times \ldots \times [\theta_{Y_p-1}^{(p)}, \theta_{Y_p}^{(p)})$.

Let $\{y_i\}_{i=1}^n$ be n i.i.d realizations of the random vector Y. Then the ℓ_1-regularized MLE estimator to learn the parameters Σ, Θ from observed data $\{y_i\}_{i=1}^n$ takes the form:

$$\underset{\Sigma, \Theta}{\text{minimize}} - \sum_{i=1}^{n} \log \mathbb{P}(y_i; \Sigma, \Theta) + \lambda_n \|\Sigma^{-1}\|_{1,\text{off}} \tag{2}$$

where $\|\cdot\|_{1,\text{off}}$ is the element-wise ℓ_1 norm excluding diagonal entries. It can be seen that the objective is non-convex, and intractable to optimize in general. Accordingly,

approximate estimates such as those based on MCMC (Chib and Greenberg, 1998), and expectation maximization Guo et al. (2015) have been proposed for learning the model parameters, but these are still relatively computationally demanding, but also does not come with the strong statistical guarantees of the actual regularized MLE solutions.

3.1 A Direct Estimation Method

We now propose an alternative estimator of the parameters in the probit graphical model distribution in (1). Given $\{y_i\}_{i=1}^n$, our aim is to estimate the unknown parameters Σ^*, Θ^*.

3.1.1 Estimation of Θ

We define $\widehat{\Theta}$, our estimator of Θ as follows:

$$\widehat{\Theta}_k^{(j)} = \begin{cases} -\infty & \text{if } k = -1 \\ \Phi^{-1}(\frac{1}{n}\sum_{i=1}^n \mathcal{I}(y_{i,j} \leq k)) & \text{if } k = 0, \ldots, M-1 \\ \infty & \text{if } k = M \end{cases} \tag{3}$$

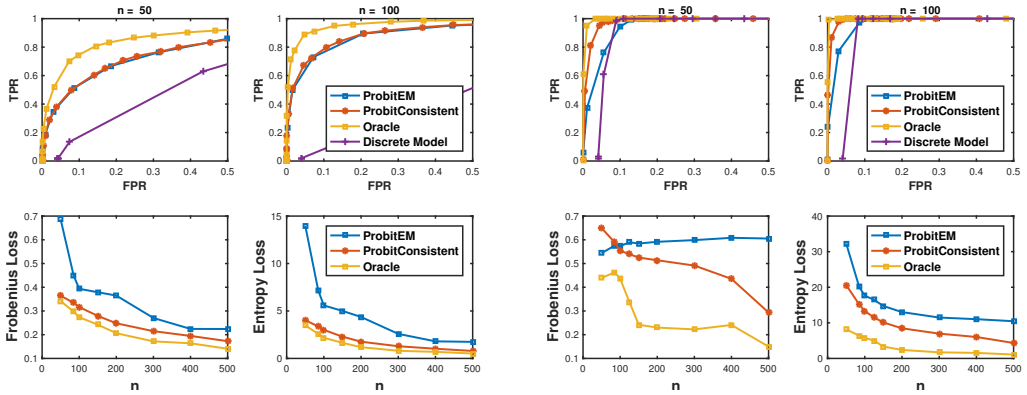
where $\Phi(.)$ is the CDF of standard normal distribution, $\mathcal{I}(.)$ is the indicator function, $y_{i,j}$ is the j^{th} coordinate of vector y_i. It can be seen that $\widehat{\Theta}$ consistently estimates Θ^*.

3.1.2 Estimation of Σ and latent graph structure

We present a two step approach for estimation of covariance matrix Σ. In the first step, we compute a raw estimate $\widetilde{\Sigma}$ of Σ using the approach we describe below. In the next step we plugin the estimated covariance matrix $\widetilde{\Sigma}$ into the graphical lasso estimator (Friedman et al., 2008) to estimate the sparse latent graph structure and a smoothed estimate $\widehat{\Sigma}$ of Σ^*.

Step 1: To estimate each entry of $\widetilde{\Sigma}$ we solve an independent optimization problem. Lets suppose we want to estimate Σ_{jk}, for $j \neq k$. The joint distribution of (Y_j, Y_k) is multinomial with probabilities: $\mathbb{P}(Y_j, Y_k; \Theta, \Sigma_{jk}) = \mathbb{P}(\theta_{Y_j-1}^{(j)} \leq Z_j \leq \theta_{Y_j}^{(j)}, \theta_{Y_k-1}^{(k)} \leq Z_k \leq \theta_{Y_k}^{(k)}; \Sigma_{jk})$. Note that the joint distribution of random variables Z_j, Z_k is bivariate normal with mean $[0, 0]$ and covariance $\begin{bmatrix} 1 & \Sigma_{jk} \\ \Sigma_{jk} & 1 \end{bmatrix}$. For a fixed Θ, one could estimate the unknown parameter Σ_{jk} by maximizing the log likelihood function, which has the following form:

$$\begin{aligned} \ell_{jk}(\sigma; \{y_i\}_{i=1}^n, \Theta) &= \sum_{a=0}^{M}\sum_{b=0}^{M} \frac{n_{ab}}{n} \log \mathbb{P}(Y_j = a, Y_k = b; \Theta, \sigma) \\ &= \sum_{a=0}^{M}\sum_{b=0}^{M} \frac{n_{ab}}{n} \log \phi_{a,b}(\sigma; \Theta), \end{aligned} \tag{4}$$



(a) Data sampled from a Probit model with chain graph structure, with $\omega = -0.3$.

(b) Data sampled from a Probit model with chain graph structure, with $\omega = -0.9$.

Figure 1: The top plots show the ROC curves for $n = 50, 100$. The bottom plots show the performance on Entropy Loss and Frobenius Loss metrics.

where $n_{ab} = \sum_{i=1}^{n} \mathcal{I}(y_{i,j} = a, y_{i,k} = b)$ and $\phi_{a,b}(\sigma; \Theta)$ is defined as: $\mathbb{P}(\theta_{a-1}^{(j)} \leq Z_j \leq \theta_a^{(j)}, \theta_{b-1}^{(k)} \leq Z_k \leq \theta_b^{(k)}; \sigma)$. However, the thresholds Θ^* are unknown. So to estimate Σ_{jk}, we replace Θ^* with its estimator $\widehat{\Theta}$ and maximize the following log likelihood:

$$\widetilde{\Sigma}_{jk} = \underset{\sigma \in \mathcal{M}}{\arg\max} \ \ell_{jk}(\sigma; \{y_i\}_{i=1}^{n}, \widehat{\Theta}).$$

where $\mathcal{M}$ is the domain of σ, which is $(-1, 1)$ unless no additional constraint on covariance is placed.

Step 2: In this step we plug-in $\widetilde{\Sigma}$ into a parametric Gaussian graphical model estimator to obtain the sparse graph structure and the final covariance matrix. While any consistent parametric Gaussian estimator (e.g., graphical lasso estimator (Friedman et al., 2008), CLIME (Cai et al., 2011), graphical Dantzig selector (Yuan, 2010)) can be used to estimate the latent graph structure, in this work we focus on the graphical lasso estimator (glasso), which involves solving the following optimization problem:

$$\widehat{\Sigma} = \underset{\Sigma^{-1} \succ 0}{\arg\min} \ \langle\langle \Sigma^{-1}, \widetilde{\Sigma} \rangle\rangle - \log\det\left(\Sigma^{-1}\right) + \lambda_n \|\Sigma^{-1}\|_{1,\text{off}} \tag{5}$$

where $\langle\langle A, B \rangle\rangle$ denotes the trace inner product of A and B. Note that this step also acts a model selection step. The ℓ_1 penalty in the objective gives us sparse graphs in the high dimensional setting. For complete details about the consistency properties of glasso, refer to (Ravikumar et al., 2011).

4 Synthetic Experiments

In this section we compare the performance of our new estimator for Probit Graphical model (which we call *ProbitConsistent*) with various baselines, on synthetic

datasets. We present more comparison results on HINTS-FDA dataset in Section 5.
Baselines: In these experiments we compare the performance of our estimator with the following estimators:

- *ProbitEM:* Expectation Maximization algorithm proposed by Guo et al. (2015) for learning Probit model.
- *Discrete model:* This model treats each ordinal variable as a nominal variable. For learning this model, we use the approach proposed by Jalali et al. (2011), which learns the graph structure by estimating the neighborhood at each node.
- *Oracle:* When the data is generated from a Probit model, we also compare with an oracle estimator which has access to the latent variables of the model. Here we run graphical lasso on the latent variables to estimate the graph structure.

Evaluation Metrics: To evaluate the accuracy of the estimated graph structure, we generate ROC plots for all the approaches. TPR in ROC plots is the proportion of correctly detected edges and FPR is the proportion of the misidentified non edges. Since *Oracle*, *ProbitEM* and *ProbitConsistent* estimate the same model, we compare them using two other metrics, namely Frobenius Loss and Entropy Loss which are defined as: Frobenius Loss $= \frac{\|\Sigma^{-1}-\widehat{\Sigma}^{-1}\|_F}{\|\Sigma^{-1}\|_F}$, Entropy Loss $= \mathrm{tr}(\Sigma\widehat{\Sigma}^{-1}) - \log\det(\Sigma\widehat{\Sigma}^{-1}) - p$, where Σ is the true covariance matrix and $\widehat{\Sigma}$ is the estimated covariance matrix.

Model Selection: To compare *Oracle*, *ProbitEM* and *ProbitConsistent* using Frobenius Loss and Entropy Loss, we need a criteria to pick the tuning parameter λ used in each of these estimators. For *ProbitEM* we use the cross validation technique proposed in Guo et al. (2015) and for *Oracle* we use the standard cross validation for glasso.

For *ProbitConsistent* we use the following k-fold cross validation technique. We partition the data set into k subsets. Each time, we use one of the k subsets as the validation set and the remaining k-1 subsets as training set. Let $\widehat{\Sigma}_{-i}$ be the covariance matrix output by Step 2 of *ProbitConsistent*, when trained using all the subsets except i^{th} subset. And let $\tilde{\Sigma}_i$ be the raw estimate of Σ obtained from Step 1 of *ProbitConsistent*, using i^{th} subset. We pick a λ which maximizes the following score: $\sum_{i=1}^{k} \log\det(\widehat{\Sigma}_{-i}^{-1}) - \langle\langle \widehat{\Sigma}_{-i}^{-1}, \tilde{\Sigma}_i \rangle\rangle$.

Experiment Settings: In all our experiments we fix the number of nodes in the graph to 50 and set number of categories for each random variable to 5. To reduce the variance, we average results over 10 repetitions.

4.1 Results

In our experiments we generate ordinal data from a Probit model. We simulate data from a chain graph. The inverse covariance matrix of the latent variables is chosen

as follows:

$$\Sigma^{-1}_{j,k} = \begin{cases} \omega^{|j-k|} & \text{if } |j-k| \leq 1 \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

We pick an $\omega \in (-1, 1)$ in our experiments and set the thresholds (θ) at node j as : $\theta^{(j)} = [-\text{Inf}, -10, -0.7, 0.7, 10, \text{Inf}]$. Finally we scale the covariance matrix so that all the variances are equal to 1. Figures 1a, 1b show the results obtained using $\omega = -0.3, -0.9$ respectively. More results for large n and grid and random graphs can be found in Appendix A.

4.2 Discussion

When the correlations between latents are high (Figure 1b), it can be seen that *ProbitConsistent* performs much better than *ProbitEM*. In this case the Frobenius norm of *ProbitEM* doesn't go down as n increases. This could be because of the mean field approximation that is made by Guo et al. (2015) to speed up the EM algorithm. In the E-step of their algorithm they approximate $\mathbb{E}[Z_j Z_k | Y; \widehat{\Theta}, \widehat{\Sigma}]$ as $\mathbb{E}[Z_j | Y; \widehat{\Theta}, \widehat{\Sigma}] \times \mathbb{E}[Z_k | Y; \widehat{\Theta}, \widehat{\Sigma}]$. This decouples the interactions between any two random variables. In the presence of high correlations, this turns out to be a poor approximation. In general, we can also see that *ProbitConsistent* has better Frobenius and Entropy losses than *ProbitEM*, when the sample complexity is low.

5 HINTS-FDA Data Analysis

5.1 Dataset Description

The Health Information National Trends Survey (HINTS)[1] is a nationally representative survey conducted by the National Cancer Institute (NCI) every few years. HINTS collects data on how American public accesses and uses cancer related information. The survey evaluates public's cancer related knowledge and perception of cancer risks.

In this work we analyze HINTS-FDA data which is a special data collected by NCI in partnership with the Food and Drug Administration (FDA) and is made publicly available by NCI. This survey collected data on the public's use of tobacco products and assessed public's knowledge of perceptions of tobacco product harm, tobacco product claims. This survey was conducted by mail from May 29 through September 8, 2015. A total of 13,001 households have been selected for the survey, out of which 3738 households have mailed back the completed questionnaire (a response rate of 28.7%). For complete details on survey methodology and sampling strategy please refer to (National-Cancer-Institute, 2017b).

The survey questionnaire has approximately 350 questions. It collected detailed information on the following topics:

[1] https://hints.cancer.gov/

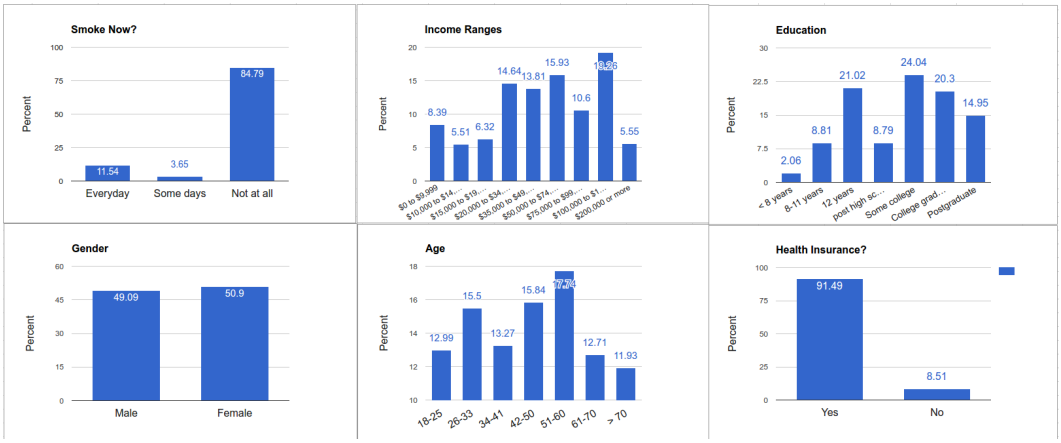


Figure 2: Summary statistics of the HINTS-FDA dataset.

- *Tobacco Product Use, Beliefs about Tobacco Products, Beliefs About Cigarette Claims, Beliefs About Cancer, How do people access Health Information?, Socio Demographic Indicators.*

Information related to other topics such as *use of dietary supplements* was also collected through the survey. However, we do not use that information in our analysis. Almost all the questions in the survey (except for $\sim$ 10 questions) have either ordinal or categorical responses. For the complete survey questionnaire please refer to (National-Cancer-Institute, 2017a). Figure 2 shows some summary statistics of the data.

5.2 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.*

5.3 Comparison with Baseline Models

Before we move onto the data analysis, we present more experimental results comparing the performance of our estimator with other baselines, on HINTS-FDA dataset.

Tasks: In this experiment we compare the performance of estimators on various ordinal regression tasks (note that although we train a graphical model, we can use the trained model for regression tasks):

- *Predict IncomeRange*: In this task, the goal is to predict the Income Range (an ordinal variable with 9 categories) of an individual given various *sociodemographic* indicators of that individual. The specific explanatory variables used are as follows: *Education, Marital Status, Have Health Insurance?, Occupation Status*.
- *Predict SmokeNow*: Here the goal is to predict whether an individual smokes or not (an ordinal variable with 3 categories : *smoke very often, smoke very rarely, don't smoke*) using explanatory variables related to sociodemographic indicators (described above).
- *Predict SmokeNow2*: Here the goal is to predict whether an individual smokes or not from the individual's smoking perceptions. The specific explanatory variables used are as follows: *CigarettesHarmHealth, TobaccoSaferNow, LowNicotineHarmful, LowNicotineAddictive, NicotineCauseCancer*.
- *Predict PreventionNotPossible*: The goal of this task is to predict if an individual thinks prevention of cancer is possible or not (which is an ordinal variable with 4 categories: from *Strongly Agree* to *Strongly Disagree*). We again use the sociodemographic indicators as the explanatory variables for this task.

Models: We fit a Probit model for HINTS-FDA data using both our approach and the EM approach of Guo et al. (2015) (we call these estimators *ProbitConsistent* and *ProbitEM* respectively as in Section 4) for each of the tasks described above. We treat the explanatory variables and the dependent variable as nodes in the graph and individuals as samples drawn from the graph. We learn the graphical model from training data and use the same cross validation strategy described in Section 4 to pick the best tuning parameter. In the testing phase, given the explanatory variables we perform MAP inference to predict the dependent variable.

We also compare performance of our estimator with two other baseline regression models: Multinomial Logistic regression model(*MultLogi*), Proportional Odds Logit model (*PropOdds*) Peterson and Harrell Jr. (1990). *MultLogi* is a model for classification which ignores the order of categories in the dependent variable and

treats it as nominal variable. Note that we do *not* compare with Multinomial Probit model because, the probit link in Probit model and logit link in *MultLogi* model have similar shapes and in practice, both usually have similar performance. Denoting $y \in \{0, \dots, M\}$ as the dependent variable and $\mathbf{x}$ as the vector of explanatory variables, the model can be summarized as follows:

$$\log \frac{P(y=j)}{P(y=M)} = \alpha_j - \beta_j^T \mathbf{x}, \quad \forall j \in \{0, M-1\},$$

where $\alpha, \{\beta_j\}_{j=0}^{M-1}$ are the parameters of the model. *PropOdds* is a popular model for ordinal regression . The model can be summarized as follows:

$$\log \frac{P(y \leq j)}{P(y > j)} = \alpha_j - \beta^T \mathbf{x}, \quad \forall j \in \{0, M-1\}.$$

Note that *PropOdds* is a more frugal model that *MultLogi* model. More importantly, *PropOdds* takes the ordering of categories in the dependent variable y into consideration.

Performance Metric: Note that accuracy, which is commonly used to measure performance of classification tasks, is not a good performance metric for ordinal regression. So we use *Kendall's Tau-b* correlation measure between predicted and true responses, a popular metric for measuring the performance of ordinal regression. It is defined as follows. Let $(x_1, y_1) \dots (x_n, y_n)$ be a set of observations of the joint random variables X and Y. Then *Kendall's Tau-b* between $\{x_i\}_{i=1}^n, \{y_i\}_{i=1}^n$ is proportional to:

$$\tau_B \propto \sum_{i<j} \text{sign}[(x_i - x_j)(y_i - y_j)].$$

Note that $\tau_B \in [-1, 1]$, with 1 corresponding to perfect positive association.

We generated a random 80/20 train-test split from the overall dataset and fit the models described above using the training split. Table 1 shows the τ_B correlation between predicted and true responses on test dataset. It can be seen that *ProbitConsistent* has a better performance than *ProbitEM* in almost all the tasks. *PropOdds*, *MultLogi* seem to have a slightly better performance than *ProbitConsistent*. This is expected, because *PropOdds*, *MultLogi* are specifically trained for the regression task, whereas *ProbitConsistent* is not.

5.4 Analysis Methodology

In this section we outline the methodology we used for analyzing HINTS-FDA dataset. We treat each question in the survey as a node in the graph and responses of individuals to these questions as samples drawn from the graph. We selected 114 questions from the dataset,that are relevant for our analysis, on the following topics: *Tobacco Product Use, Beliefs about Tobacco Products, Beliefs About Cigarette*

Task	*ProbitEM*	*ProbitConsistent*	*PropOdds*	*MultLogi*
Predict IncomeRange	0.487	0.499	0.503	0.494
Predict SmokeNow	0.176	0.341	0.347	0.343
Predict SmokeNow2	0.117	0.126	0.1	0.084
Predict PreventionNotPossible	0.239	0.393	0.411	0.415

Table 1: Performance comparison of *ProbitEM*, *ProbitConsistent*, *PropOdds*, *MultLogi* models on various regression tasks. The numbers in the cells represent the *Kendalls's Tau-b* correlation between the predicted and the true responses on test dataset.

Claims, How do people access Health Information?, Socio Demographic Indicators. We fit the probit model using *ProbitConsistent* on the selected questions. To choose the best tuning parameter we use 10 fold cross validation technique described in Section 4. We obtain 95% confidence intervals for the edge strengths (*i.e.,* partial correlations) of the latent graph through jackknife re-sampling technique. The HINTS-FDA dataset comes with jackknife replicates, which we use to compute confidence intervals. We place an edge in the graph only if its confidence interval doesn't intersect with $[-0.1, 0.1]$.

5.5 Results

Next, we present the results from our analysis. We first consider the question of how various variables related to *sociodemographic* indicators are associated with smoking behavior of a person. Figure 3 shows the estimated graph structure for these variables. Table 2 describes some of the relevant nodes in the graph. For a more complete list of variables, please refer to Table 3 in Appendix B. Of particular interest to us is the variable *SmokeNow*, which asks people how often they smoke. Specifically, we are interested in how this variable is associated with other variables. From the graphs, we can see that the variable that has a very significant association with *SmokeNow* is *Education*. This indicates that if a person is well educated then conditioned on all the other variables, there is lower chance that the person smokes. The other two significant associations of *SmokeNow* are with *Mexican* and *White*. If a person is Mexican, it can be seen that conditioned on rest of the variables, there is a lower chance that the person smokes. The opposite is true if the person is White.

Next, we try to understand how the perceptions of smoking risks vary with smoking behavior. Figure 4 presents the estimated graph. Table 3 describes some of the relevant nodes in the graph. It can be seen that *SmokeNow* and *FewCi-*

Node Name	Question	Possible Responses
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
Mexican	Are you a Mexican?	1-'Yes', 2-'No'
NoticeHealthInfoInternet	Have you read such health information on the Internet?	1-'Yes', 2-'No'
SmokeNow	Do you now smoke cigarettes every day, some days or not at all?	1-Everyday, 2-Some days, 3-Not at all
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'
UseInternet	Do you ever go on-line to access the Internet?	1-'Yes', 2-'No'
White	Are you a White?	1-'Yes', 2-'No'

Table 2: Description of questions corresponding to key nodes in Figures 3, 4, 5.

garettesHarmHealth have a positive partial correlation between them. It indicates that, conditioned on the rest of the variables, people who smoke, perceive smoking as less harmful than people who don't smoke. The observations from Figures 3, 4 suggest that, it is the lack of proper awareness about the risks of smoking that causes more people to smoke (which also agrees with our natural belief). So, strategies for communicating smoking related health information should focus more on the less educated stratum of the population.

We now study how people access their health information, to see if people who smoke use a different medium than people who don't smoke to access health information. Figure 5 shows the graph with variables relevant to how people get their health info. It can be seen that their is a strong negative relation between *Education* and *UseInternet*, *NoticeHealthInfoInternet* indicating that people who are less educated don't use internet to access health information. The negative edges between *TobaccoEffects_TV* and *Education*, *SmokeNow* suggests that people who smoke and who are less educated get their tobacco related health information through TV more than others (the opposite is true with *TobaccoEffects_Newspaper*). All these observations possibly suggest that TV is a more effective way than Newspapers and Internet to communicate health information to less educated stratum of the population. Note that this last statement needs further analysis, as our analysis only focused on identifying associations and not causal relations.

5.6 Limitations

In this analysis we only focused on identifying associations between various variables of interest. However, an analysis that identifies causal relationships would be much more helpful in understanding what factors cause people to smoke. Such an analysis can help us in designing schemes to effectively reduce smoking prevalence.

The probit model we used in the analysis, assumes that the latent variables are

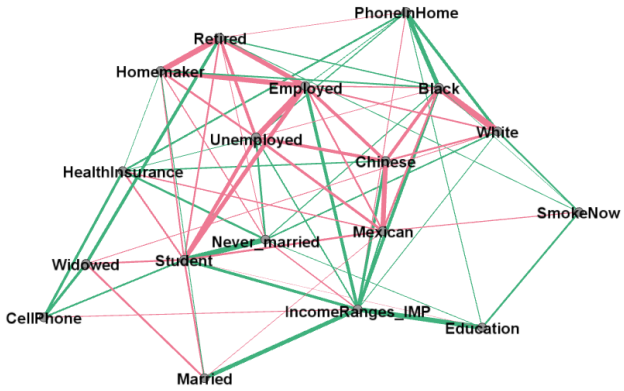


Figure 3: The estimated latent graph structure corresponding to *SmokeNow* and sociodemographic indicators. The graph is generated from the marginal distribution of the corresponding variables. Green edges represent positive partial correlations and red edges represent negative partial correlations. Edge thickness is proportional to the magnitude of the partial correlation.

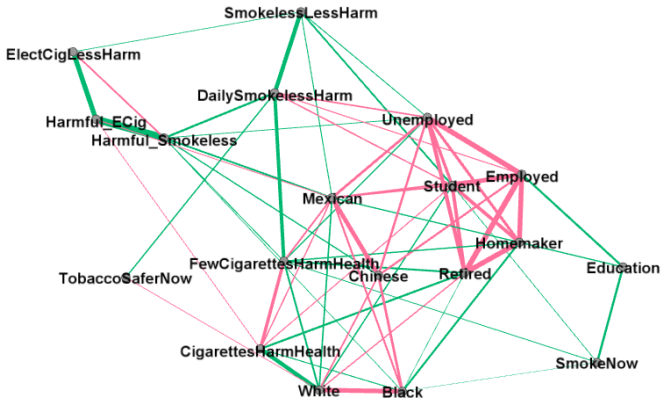


Figure 4: The estimated latent graph structure for variables corresponding to perceptions of smoking risks and *SmokeNow*. The graph is generated from the marginal distribution of the corresponding variables.

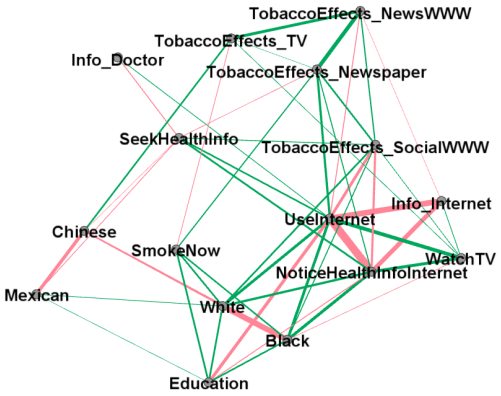


Figure 5: The estimated latent graph structure for variables corresponding to how people access health information.The graph is generated from the marginal distribution of the corresponding variables.

normally distributed. While we didn't verify the correctness of this assumption, we believe that any choice of latent distribution will lead to similar inferences about the data (as long as the latent distribution is flexible enough to model both negative and positive associations). This can be clearly seen in the univariate case, where any choice of latent distribution leads to the same class of models.

6 Conclusion

We have proposed a new estimator for learning ordinal probit model that is computationally tractable and can be easily scaled to large datasets. We empirically corroborated the superior performance of our estimator for the probit model. We applied our estimator on HINTS-FDA dataset to analyze the smoking behavior of people and their perceptions of smoking risks. We found that people who smoke, perceive smoking as less harmful than people who don't smoke and the lack of awareness of smoking risks could be a reason why many people smoke.

7 Future Work

The probit model we considered in this work can't be used to model mixed data (i.e., data in which different kinds of variables occur together). We don't often see a dataset that only has ordinal variables. So, one direction for future work would be to extend the probit model to handle mixed data efficiently. Missing data is another issue that is common in many real world datasets. In our analysis, we imputed the missing data using the median. Another direction for future work would be to handle the missing data problem in a more sophisticated way.

References

Agresti, A. (2010). *Analysis of ordinal categorical data*, volume 656. John Wiley & Sons.

Allen, G. I. and Liu, Z. (2012). A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pages 1–6. IEEE.

Amemiya, T. (1974). Bivariate probit analysis: Minimum chi-square methods. *Journal of the American Statistical Association*, 69(348):940–944.

Ashford, J. R. and Sowden, R. R. (1970). Multi-variate probit analysis. *Biometrics*, 26(3):535–546.

Becker, W. E. and Kennedy, P. E. (1992). A graphical exposition of the ordered probit. *Econometric theory*, 8(01):127–131.

Cai, T., Liu, W., and Luo, X. (2011). A constrained l1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, pages 347–361.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799–805.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2015). Graphical models for ordinal data. *Journal of Computational and Graphical Statistics*, 24(1):183–204.

Ising, E. (1925). Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik*, 31:253–258.

Jalali, A., Ravikumar, P., Vasuki, V., and Sanghavi, S. (2011). On learning discrete graphical models using group-sparse regularization. 14.

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

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *The Annals of Statistics*, pages 31–57.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). The nonparanormal skeptic. *arXiv preprint arXiv:1206.6488*.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*, volume 999. MIT Press.

National-Cancer-Institute (2017 (accessed April 11, 2017)a). Hints fda english annotated survey.

National-Cancer-Institute (2017 (accessed April 11, 2017)b). Hints-fda methodology report.

Peterson, B. and Harrell Jr., F. E. (1990). Partial proportional odds models for ordinal response variables. *Applied statistics*, pages 205–217.

Ravikumar, P., Wainwright, M. J., Raskutti, G., Yu, B., et al. (2011). High-dimensional covariance estimation by minimizing 1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC.

Speed, T. P. and Kiiveri, H. T. (1986). Gaussian Markov distributions over finite graphs. *Annals of Statistics*, 14(1):138–150.

Yang, E., Baker, Y., Ravikumar, P., Allen, G. I., and Liu, Z. (2014). Mixed graphical models via exponential families. 17.

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286.

Appendix

A Synthetic Experiments

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 6 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 7 presents the results for $\omega = 0.8, -0.65$.

B HINTS FDA

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
HealthInsurance	Do you have any kind of health care coverage?	1-Yes, 2-No
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
Homemaker	Occupation Status	1-Not Homemaker, 2-Homemaker
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'
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
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'
Mexican	Are you a Mexican?	1-'Yes', 2-'No'
White	Are you a White?	1-'Yes', 2-'No'

Table 3: Table describing the questions corresponding to some of the nodes in Figures 3, 4, 5.

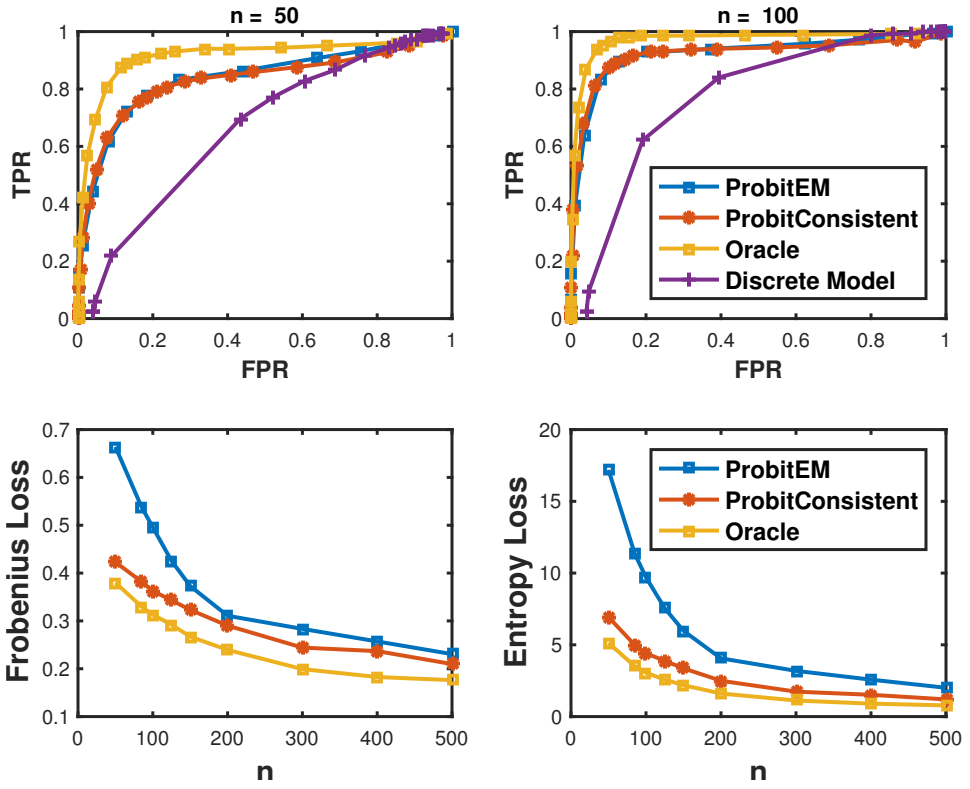


Figure 6: Data sampled from a Probit model with grid graph structure. The top plots show the ROC curves for $n = 50, 100$. The bottom plots show the performance on Entropy Loss and Frobenius Loss metrics.

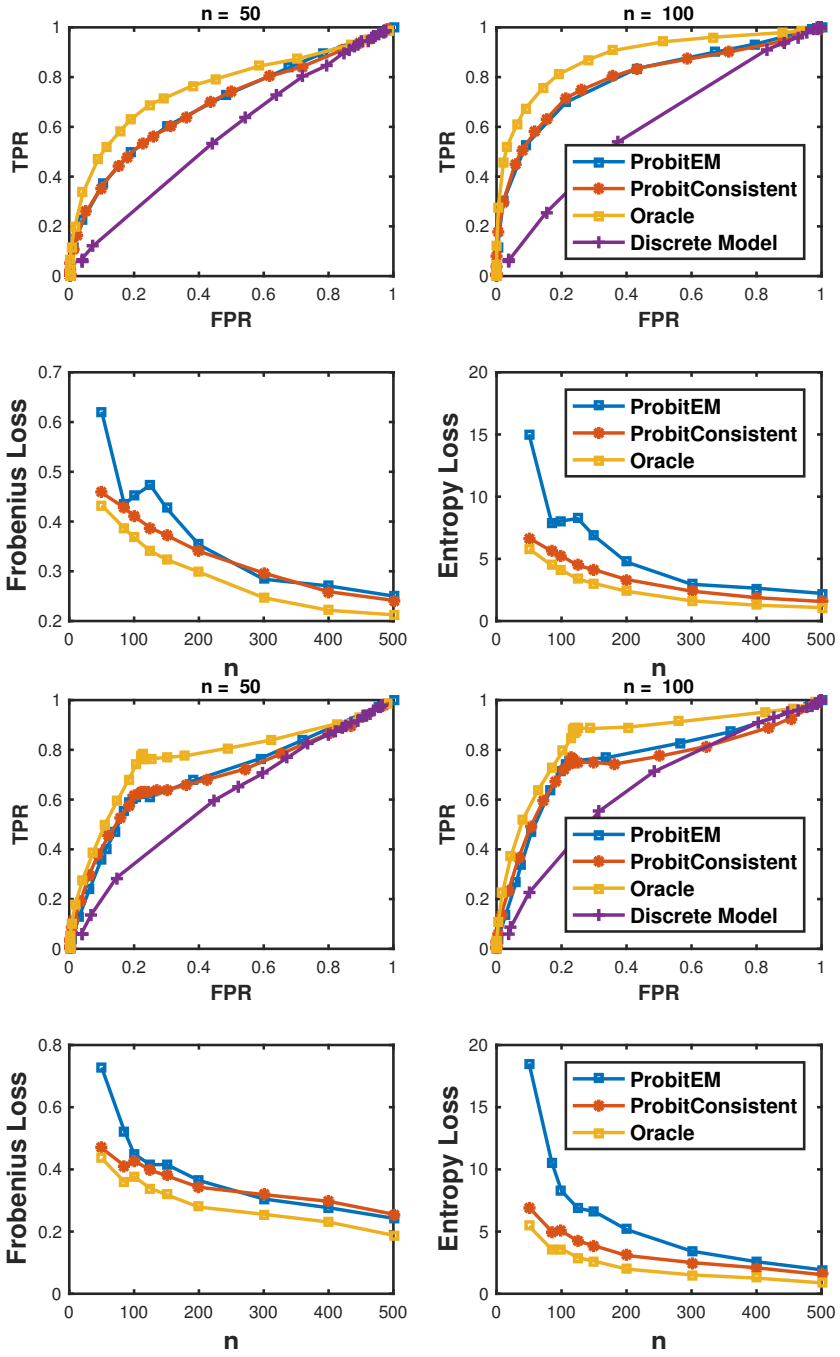


Figure 7: Data sampled from a Probit model with random graph structure. Left 4 plots correspond to $\omega = 0.8$ and the right 4 plots correspond to $\omega = -0.65$. The top plots show the ROC curves for $n = 50, 100$. The bottom plots show the performance on Entropy Loss and Frobenius Loss metrics.