A Survey on High Dimensional Statistics

Stathis Megas
Department of Physics and Astronomy
University of California, Los Angeles
Los Angeles, CA
stathismegas@gmail.com

Zhuyun Maggie Xiao
Electrical & Computer Engineering
University of California, Los Angeles
Los Angeles, CA
zxiao2015@g.ucla.edu

Hailin Yu
Electrical & Computer Engineering
University of California, Los Angeles
Los Angeles, CA
yhltc@g.ucla.edu

Abstract

This survey on high dimensional statistics will focus on Lasso linear model in noisy setting and graphical model in high-dimensional statistics. The first part provides a high level overview of high dimensional statistics and briefly talks about why studying high dimensional statistics is of increasing importance. Basic mathematical techniques including concentration inequalities and tail bounds are introduced. Next, theoretical results on Lasso linear model in noisy setting is discussed. Finally, we provide a brief introduction to the graphical models used for modeling high dimensional probability distributions. We start with the probabilistic point of view of graphical models, then move on to Gaussian and Ising models. In particular, we describe how learning is done under these two graphical models. We end the survey with a proposed step beyond what has been currently reported.

1 Introduction

In a statistics problem, there is a dataset X, and each data can be seen as a vector x_i. The dataset is always a two dimensional matrix with the number of data: n and the dimension of each data: d. Interestingly, the definition of high-dimensional in not only depend on d.

It is called classical case when $n \gg d$, and many probability theories can be used, such as the law of large numbers or the central limit theorem, to to get estimators with consistency and asymptotic normality. However, this is not always the reality, especially in modern time, there are plenty of data-set with $d \sim n$ or even $d \gg n$, which are the cases called high-dimensional. And the solutions to these cases are using the high-dimensional statistics.

There are a lot of real dataset in our life. For example, in stock market, each stock has a price every second, then it will have about 10000 different prices everyday which is larger than the number of stocks. As a result, if prices are used as features, it is a high-dimensional case. Also, in an objection detection problem, each image contains millions of pixels, which is far more larger than the number of images. Therefore, if all the pixels are used as features, it belongs to the high-dimensional case.

To find consistent results in high-dimensional cases, some assumptions such as sparse vectors, structured covariance matrices, low-rank matrices, structured regression functions, and some regularity conditions are applied. [1]

2 Difficulties

Classical method is not suitable for high-dimensional statistics mainly because the classical "large n, fixed d" theory fails to provide useful predictions.

This will be shown by using a linear regression model. For linear regression, the loss function is:

$$J(\theta) = \sum (y_n - \theta^T x_n)^2 / N \tag{1}$$

The optimal solution satisfies:

$$XX^T\theta = Xy, X = (x_1, ...x_n) \in R^{d*n} \tag{2}$$

Since the rank of XX^T and rank of X are the same, then:

$$rank(XX^T) = rank(X) \leq min(d, n) \tag{3}$$

For the classical case with $n \gg d$, if X is full rank, which means $rank(X) = d$, then XX^T is invertible and we can get a unique solution. Even if X is not full rank, which means features are linearly dependent, then a solution can be obtained by using part of the features.

However, when being applied to the high dimensional case: $d > n$, even if X is full rank,

$$rank(XX^T) = rank(X) = n < d \tag{4}$$

XX^T is not invertible any more. As a result, we cannot get a good solution. In addition, the testing error is bounded by $C\frac{d}{n}$. Therefore, the model is going to overfit under in high dimensional case. Last but not least, the algorithm complexity is $O(nd^2) + O(d^3)$. Therefore, it becomes difficult to solve the high dimensional case.

The main idea of solving the problem above is to reduce it to a low-dimensional structure. Such approach can always be carried out when there is sparsity in vectors, indicating some features are zero:

$$x_i = (x_{i1}, ..., x_{id}) = (x_{i1}, ..., x_{ip}, 0, 0...0), p << n \tag{5}$$

Fast decay in the eigenvalues of covariance matrices can also help, because dimensionality reduction methods such as PCA can be used to reduce the problem to a low-dimensional case while preserving almost all of the information. [8].

3 Tail Bounds

In this section, concentration inequalities and tail bounds, which are the basic techniques, will be introduced.

The meaning of tail bound is tail probability of a random variable $X : P(X \geq t)$. In classical models, tail probability is often bounded by the moments of X.

Firstly, some classical bounds are introduced ($t > 0$), [8]:

1. Markov's inequality: $P(X \geq t) \leq E[X]/t$.

2. Chebyshev's inequality: $P(|X - \mu| \geq t) \leq Var(X)/t^2$.

The simplest type of growth rate of the moment generating function in the Chebyshev's inequality is known as sub-Gaussian. A random variable X with mean μ is sub-Gaussian if there is a σ s.t. for any $\lambda \in R, E[exp(\lambda(X - \mu))] \leq exp(\sigma^2\lambda^2/2)$.

3. Hoeffding bound: $P[\sum (X_i - \mu_i) \geq t] \leq exp(-t^2/2 \sum \sigma_i^2)$,$X_i$ has mean μ_i and sub-Gaussian parameter σ_i.

4.Sub-exponential tail bound: $P(X - \mu \geq t) \leq exp(-t^2/2v^2), t \leq v^2/a$, and $P(X - \mu \geq t) \leq exp(-t/2a), t \geq v^2/a$, X is sub-exponential with parameter (v, a).

Secondly, martingale decomposition methods are introduced which are used to deal with tail bounds of functions, martingale is a sequences (Y_k) with same finite expected value, for example, sums of an i.i.d. sequence with zero mean is a martingale. Also, its difference sequence $D_k (= Y_k - Y_{k-1})$ is a martingale difference sequence, refer to [8].

1. Azuma–Hoeffding: Let D_k be a martingale difference sequence and $D_k \in [a_k, b_k]$ almost surely for all k, then $P[|\sum D_k| \geq t] \leq exp(-2t^2/\sum(b_k - a_k)^2)$.

2. Bounded differences inequality: if function f satisfies the bounded difference property with parameters $(L_1, ..., L_n)$, and the random vector X has independent components, then $P(|f(X) - E[f(X)]| \geq t) \leq 2exp(-2t^2/\sum L_k^2)$.

Lastly, Lipschitz functions will be introduced, which exhibit a particularly attractive form of dimension-free concentration. [8]

1. A function $f : R \to R^n$ is L-Lipschitz if for all $x, y \in R$, $|f(x) - f(y)| \leq L|x - y|$.

2. For L-Lipschitz function and i.i.d. Gaussian variables X: $P(|f(X) - E[f(X)]| \geq t) \leq 2exp(-2t^2/L^2)$, As a result, $f(X) - E[f(X)]$ is a sub-Gaussian with parameter at most L.

This result is useful: it guarantees that regardless of the dimension, any L-Lipschitz function of a standard Gaussian variables behaves like a Gaussian variable with variance L^2.

3. If a function $f : R \to R^n$ is differentiable, then for any convex function ϕ, $E[\phi(f(X) - E[f(X)])] \leq E[\phi(\pi/2 < \nabla f(X), Y >)]$, where $X, Y \sim N(0, I_n)$ are standard multivariate Gaussian.

4 Lasso, estimation in noisy setting

We have already mentioned that for $d > n$ generically we don't have convergence to something useful and this is why we look for low dimensional structures in our data because, when those structures are present, the predictions follow some universal law. One hopes that we don't need to restrict to too special cases in order to have a useful universal behavior. Indeed, the notion of sparsity in data is strong enough to lead to universal behavior in data that have it and yet it is found empirically that it is not too restrictive to exclude interesting real world problems.

4.1 Noisy Linear Regression

Let's first introduce the problem.

Definition 1 (Noisy Linear Regression). *Our deterministic Model is that for data (y, X), where $y \in \mathbb{R}$ and $X \in \mathbb{R}^d$, and noise $w \in \mathbb{R}$, we have $y = X\theta^* + w$. The task is to estimate θ^* given samples (y_i, X_i).*

Indeed this is a non trivial task since for $d > n$ the matrix X would not be invertible so we wouldn't be able to use the methods we learned in class. To solve the problem in a useful (ie universal) way we sacrifice breadth of scope. We decide to focus only on cases where θ^* is sparse or weakly sparse.

Definition 2 (Support Set). *The support of a vector $\theta \in \mathbb{R}^N$ is*

$$S(\theta) := \{j : \theta_j \neq 0\} \tag{6}$$

Definition 3 (Sparsity). *A vector θ is said to be hard sparse iff*

$$s := |S(\theta)| \ll d\ . \tag{7}$$

It is called weakly sparse iff $\ell_p(\theta) \ll d$.

One proposal for a θ that approximates θ^* is given from the LASSO program.

Definition 4 (LASSO program). *Given empirical data (y, X), pick*

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^d} \left[\frac{1}{2n}||y - X\theta||_2^2 + \lambda_n||\theta||_1\right]\ . \tag{8}$$

For the noiseless channel, w=0, the modelling assumption of sparsity would guarantee convergence of the LASSO recipe to the correct value $\hat{\theta} \to \theta^*$. However, this is no longer true in the presence of noise. For that we also need a regularity assumption regarding our data.

Definition 5 (Restricted Eigenvalue Condition). *The matrix X satisfies the restricted eigenvalue (RE) condition over the (support set) S with parameters (κ, α) iff*
$\forall \Delta \in \mathbb{C}_\alpha(S),\ \frac{1}{n}||X\Delta||_2^2 \geq \kappa||\Delta||_2^2$,

where we defined

Definition 6 (Cone of support set)**.** *For a support set S, we define*
$\mathbb{C}_\alpha(S) := \{\Delta \in \mathbb{R}^d; ||Proj_{S^c}\Delta||_1 \leq \alpha||Proj_S\Delta||_1\}$.

For intuition, the cone of a support set is all those vectors that can be represented well within the support, i.e. have a bigger projection on the support than on its complement (See figure 1(b)) . Moreover, the RE condition tells us that the Hessian of the LASSO cost function, $\nabla^2\mathcal{L}_n(\theta) = \frac{1}{n}X^TX$, has supported eigenvalues which are bounded below, so that the loss function is sharply peaked around its minimum (at least in the supported directions). Of course inevitably, for $n < d$, there are bound to be flat directions of the loss function, which is after all the source of the difficulty of non-asymptotic problems (see figure 1(a)).

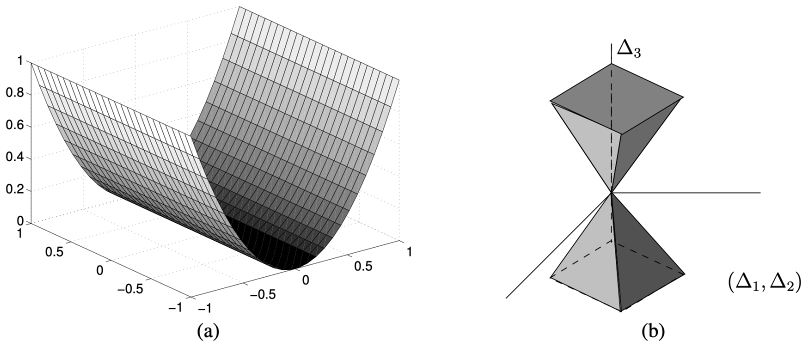


Figure 1: (a) The loss function plotted against two features valued in the horizontal plane. The existence of one flat direction tells us that the loss function has been calculated for just one data sample. (b) Illustration of the cone of a sparsity set which is supported only along the Δ_3 direction.

What this condition buys us in the proof is that, due to the sharp peak at the minimum of the cost function, convergence to the optimal θ^* speeds up as the value of the cost function is reduced.

Theorem 1. *Any solution of the Lagrangian Lasso with regularization parameter lower bounded as* $\lambda_n \geq 2||\frac{X^Tw}{w}||_\infty$ *satisfies the bound*

$$\left\|\hat{\theta} - \theta^*\right\|_2 \leq \frac{3}{\kappa}\sqrt{s}\lambda_n \ . \tag{9}$$

In other words, under those conditions, the Lasso correctly identifies θ^* and therefore can be legitimately used for predictions. Due to the usefulness of the RE condition, a very active field of research consists in precisely coming up with sufficient conditions for when it holds.
In fact under a more relaxed (probabilistic) version of RE, we can prove the following.

Theorem 2. *For* $\lambda_n \geq 2||X^Tw/n||_\infty$ *and any* $\theta^* \in \mathbb{R}^d$*, there exist* $c_1,\ c_2$*, such that any LASSO optimal solution* $\hat{\theta}$ *satisfies the bound*
$||\hat{\theta} - \theta^*||_2^2 \leq \frac{144}{c_1^2}\frac{\lambda_n^2}{\kappa}|S| + \frac{16}{c_1}\frac{\lambda_n}{\kappa}||\theta^*_{s^c}||_1 + \frac{32c_2}{c_1}\frac{\rho^2(\Sigma)}{\kappa}\frac{\log d}{n}||\theta^*_{S^c}||_1^2$.

Inequalities like that are called oracle inequalities because they are explicit about both our approximation and estimation error. The last two terms that know about the absolute truth $\theta^*_{S^c}$ specify the approximation error and the first term, which depends only on properties of the Hypothesis class I am learning, is the estimation error.

4.2 Current Research on LASSO and REC

The success of LASSO in the asymptotic regime has been great and ways in which it can be expanded and improved are still being investigated. In the next section we will thoroughly describe one such research area that uses graphical methods. But, for illustration, we mention here two other aspects intensely investigated : computationally effective ways in which to certify that a matrix X obeys REC, and ways to determine the LASSO penalty in a data driven way, as opposed to using the broad-strokes option $\lambda_n||\theta||_1$. In Raskutti et al. (2010) [4], they proved a lower bound on $||X\Delta||_2$ using just the maximum diagonal entry $\rho^2(\Sigma)$ of a covariance matrix Σ. On the other front, in Velten et al. (2018) [7], they tried to exploit the fact that many covariates can be qualitatively different and hence the non-sparsity penalty should take that extra structure into account.

5 Graphical Models for High-Dimensional Data & Ising Model

In this section, we move on to discuss graphical models, which combine probability theory and graph theory, and are useful for modeling high-dimensional probability distributions. Here, we focus exclusively on the undirected graphs G models (a.k.a, Markov random fields). These models are used in many different domains, including statistical physics for modeling interactions in a magnetic field, natural language processing, image analysis, spatial statistics [5], finding relationship between variables, relating nutrients and food names (Figure. 2(a)) [6]. They are also used in medical applications to learn a breast cancer genetic network of mutations (binary) and gene expression (counts via RNA-sequencing) (Figure. 2(b)). Their attractiveness is due to the computational reduction that one obtains by using a sparse graph in computing conditional and marginal probabilities, in addition to the type of insights that graphical models can give with regards to the relationship between different random variables, see examples on [9, 6].

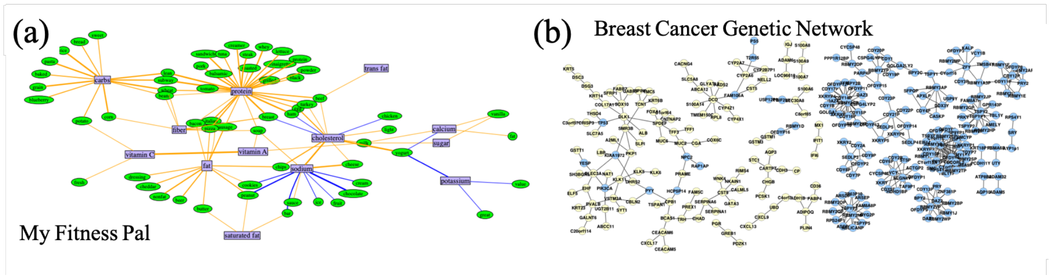


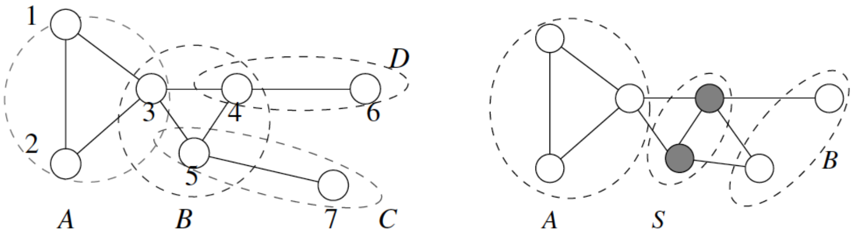
Figure 2: (a) Graphical model describing the relationships between different types of variables. The purple nodes represent the nutrients, and the green circles are Bernoulli distributed, representing food entry. There are 20k different words in the food entry (d), and there are thousands of micro-nutrients (n). The orange lines indicate the L2 sizes of weights between two nodes. [6] (b) The graph demonstrates the connected components of a breast cancer genetic network estimated by the Truncated Poisson and Ising mixed graphical model for gene expression. Yellow nodes represent RNA-sequencing, and blue nodes indicate genomic mutations (d) of the 697 breast cancer subjects (n). [9] In both graphical models, d » n.

We begin discussing probabilistic and estimation aspects of the graphical models framework. Specifically, we start by connecting the distribution of $X = (X_1, \ldots, X_d) \in \bigotimes_{j=1}^{d} \mathcal{X}_j$ to an undirected graph G, whenever the probability density function of X can be written as

$$p(x_1, \ldots, x_d) = \prod_{C \in \mathcal{C}} \phi_C(X_C), \tag{10}$$

for some $\phi_C : \mathcal{X}^C \to [0, \infty)$, where $\mathcal{X}^C = \bigotimes_{j \in C} \mathcal{X}_j$, with $\mathcal{C}$ a collection of cliques of the graph G. Recall that a clique is a collection of nodes where every pair of nodes are connected by an edge. Refer to Figure. 3(a) for an example. There the joint probability function can be expressed as the product of four different functions, each associated with one of the cliques in the graph (Cliques: A, B, C and D).

$$p(x_1,\ldots,x_7) \propto \psi_{123}(x_1,x_2,x_3)\;\psi_{345}(x_3,x_4,x_5)\;\psi_{46}(x_4,x_6)\;\psi_{57}(x_5,x_7).$$



(a)

(b)

Figure 3: (a) Example of cliques. A and B are 3-cliques; C and D are 2-cliques. (b) S is a vertex cutset which breaks the graph into disconnected subgraphs A and B. [8]

Formally, graphical models consists of classes of probability distributions for which (10 holds for a graph G and functions ϕ_C defined over the cliques of G. Remarkably, there exists a connection between the form (10) and the conditional independence statements involving the distribution of X. Specifically, recall that X is Markov with respect to a graph G if for all vertex cutsets S breaking the graph G into disjoint pieces A and B, the conditional independence statement $X_A \perp X_B | X_S$ holds. An example of a cutset S is given in Figure 3(b), where we see that cutset S breaks the graph into two disconnected subgraphs, A and B. The following fundamental result holds:

Theorem 3. ***(Hammersley–Clifford).*** *For a given undirected graph and any random vector $X = (X_1, \ldots, X_d)$ with strictly positive density p, the following two properties are equivalent:*

The random vector X factorizes according to the structure of the graph G, as in (10).

The random vector X is Markov with respect to the graph G.

In real applications, the goal is to learn the graph structure G on the basis of n independent and identically distributed random draws from $\{X^{(i)}\}_{i=1}^n$ from p. The high-dimensional nature of this problem comes from the fact that d can be larger than n. Hence, to make estimation feasible, it is often imposed as an assumption that the graph G is sparse, in the sense that the numbers of edges in G is much smaller than d.

Next, we review some of the main examples of graphical models estimation, focusing on the Gaussian case, widely known as graphical lasso, see [3], and on the Ising model, which corresponds to the case in which X_j is binary, see [9, 6] for further generalizations.

5.1 Gaussian graphical model

Suppose we have N independent copies of a multivariate normal random vector X dimension p, with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^p \times \mathbb{R}^p$. In the Gaussian graphical model, the goal is to estimate the covariance matrix Σ. Notably, this problem is related to the graphical model framework in the following way. Specifically, by the properties of the normal distribution, if the ith and jth components of Σ^{-1} is zero, then variables X_i and X_j are conditionally independent given all others. Thus by the Hammersley–Clifford theorem, the inverse covariance matrix Σ^{-1} induces a graphical model for the distribution of X. Such graphical can be represented by graph G with the property that (i, j) is an edge in G if and only if $(\Sigma^{-1})_{i,j} \neq 0$.

Next we describe how to estimate the graph G induced by the unknown inverse covariance matrix σ^{-1} given the N independent copies $\{X_i\}_{i=1}^N$. First, notice that in low dimensions, $d << n$, a natural estimator is to simply take the sample variance

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^\top,$$

where $\bar{X} = \sum_{i=1}^n X_i/n$, an estimate Σ^{-1} with S^{-1}. However, the estimator S^{-1} will not have exactly zero components, and more importantly, it might not be well defined when $d >> n$. To tackle this problem, [3] proposed to exploit the fact that in real problems it is expected that the graph G will be sparse. Thus, [3] proposed the graphical lasso which is defined as

$$\hat{\theta} = \min_{\Theta} \left\{ -\log \det \Theta + \operatorname{tr}(S\Theta) + \rho \|\Theta\|_1 \right\}, \tag{11}$$

where S is the empirical covariance and $\rho > 0$ is a tuning parameter, and

$$\|\Theta\|_1 = \sum_{i=1}^{p} \sum_{j=1}^{p} |\Theta_{i,j}|.$$

Hence, $\hat{\Theta}$ is an estimator of Σ^{-1} by minimizing the log-likelihood of the implied Gaussian model that generates the data, plus an ℓ_1 penalty on the inverse covariance matrix to encourage sparsity. From $\hat{\Theta}$ we can estimate the graph G with $\hat{G}$, where (i, j) is an edge in $\hat{G}$ if and only if $\hat{\Theta}_{i,j} \neq 0$.

5.2 Ising model

Another prominent graphical model was studied in [5]. This model has a wide range of applications in biology, physics, and political sciences.

The Ising graphical model differs from the Gaussian graphical model in several ways. First, the components of X are all binary, for simplicity $X \in \{-1, 1\}^d$. Secondly, as we will see, the estimation of the graphical node associated is done node by node, unlike the Gaussian graphical which is based on a single optimization problem.

We now formalize the Ising graphical model. Suppose that we are given a graph $G = (V, E)$, with set of nodes $V = \{1, \ldots, d\}$. A random vector $X \in \{-1, 1\}^d$ is said to have an Ising model distribution based on G if

$$\mathbb{P}_{\theta^*}(x) = \frac{1}{Z(\theta^*)} \exp\left(\sum_{(s,t) \in E} \theta^*_{st} x_s x_t \right),$$

where $Z(\theta^*)$ is a normalization constant, that ensures that P_{θ^*} is a valid joint probability mass function.

In practice the graph G is unknown, and the goal is to estimate G given $\{x^{(1)}, \ldots, x^{(n)}\}$ independent copies of X, with d potentially larger than n, the parameter θ^*.

Given $\{x^{(1)}, \ldots, x^{(n)}\} \subset \{-1, 1\}^d$ independent copies of X, with d potentially larger than n, the parameter θ^*, and hence the graph structure G, is learned by solving for each node r the problem

$$\min_{\theta_{V\backslash r}} \; -\frac{1}{n} \sum_{i=1}^{n} \log \mathbb{P}_{\theta^*}\left(x_r^{(i)} | x_{\backslash r}^{(i)}; \theta_{V\backslash r} \right) + \lambda \|\theta_{V\backslash r}\|_1, \tag{12}$$

where $\lambda > 0$ is a tuning parameter. Here, $x_{\backslash r}^{(i)}$ consists of the vector $x^{(i)}$ after removing all the coordinates $j \in V\backslash\{r\}$. Furthermore, $\theta_{V\backslash r}$ is defined as

$$\theta_{V\backslash r} = \{\theta_{ru}, u \in V\backslash\{r\}\}.$$

The motivation behind (12) is to estimate the neighborhood of r, the set $\mathcal{N}_r = \{u : X_r X_u | X_{V\backslash\{u,r\}}\}$, under the assumption that $|\mathcal{N}_r|$ is small. Hence, (12) solves an optimization problem that balances the conditional likelihood of X_r given $\{X_j\}_{j \in V\backslash r}$, and a surrogate penalty (ℓ_1) on the size of the neighborhood.

6 Future Directions

There are plenty of different research directions in high dimensional statistics such as second-order cone programs, semidefinite programming relaxations, sparse PCA and low-rank matrix estimation.

With regards to graphical models, some future directions include nonparametric graphical models. In fact the vast majority of work has focused on parametric models such as exponential families. Only recently there has been work on model beyond these families, one of those works is [2], but the framework there does not model a full nonparametric distribution. Other directions of work include deep graphical models. While the modeling part can easily be implemented, an open theoretical question is whether is possible to provide sparsistency guarantees for deep graphical models, such as in the work of [5] for Ising graphical model.

References

[1] Alekh Agarwal, Sahand Negahban, and Martin J Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In *Advances in Neural Information Processing Systems*, pages 37–45, 2010.

[2] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. The multiple quantile graphical model. In *Advances in Neural Information Processing Systems*, pages 3747–3755, 2016.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[4] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

[5] Pradeep Ravikumar, Martin Wainwright, and John D. Lafferty. High-dimensional ising model selection using 1-regularized logistic regression. *Annals of Statistics (Ann. Statist.)*, 38, 2010.

[6] Wesley Tansey, Oscar Hernan Madrid Padilla, Arun Sai Suggala, and Pradeep Ravikumar. Vector-space markov random fields via exponential families. In *International Conference on Machine Learning*, pages 684–692, 2015.

[7] Britta Velten and Wolfgang Huber. Adaptive penalization in high-dimensional regression and classification with external covariates using variational bayes. *arXiv preprint arXiv:1811.02962*, 2018.

[8] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

[9] Eunho Yang, Yulia Baker, Pradeep Ravikumar, Genevera Allen, and Zhandong Liu. Mixed graphical models via exponential families. In *Artificial Intelligence and Statistics*, pages 1042–1050, 2014.