Recovering the sparse network of gene expression: solving Gaussian covariance selection with proximal gradient methods

PB18000072 PB18000388

January 30, 2021

Abstract: Long has it been plaguing researchers that through what pattern of a network genes interact with each other. Motivated by a Kaggle dataset of breast-cancer-ralated genes' expression levels, we set out tackling this problem by fitting a Gaussian graphical model with sparse edges, which culminates in solving a l-1 penalized Gaussian covariance selection problem. After a failed trial on vanilla gradient descent with Huber surrogate, we comprehensively study the proximal gradient methods, especially ISTA, FISTA and MFISTA, proposed by preceding researchers. Incremental theoretical analysis and numerical experiments are presented in our work to justify these algorithms' feasibility in Gaussian covariance selection cases. We also compare these algorithms with their most prevalent counterpart Glasso, and provoke further discussions on motivations and improvements.

1 Motivation: issues on genetics

1.1 The network of gene expression

Highlighted by the biologists, on top of the tall column of genetics stands the issue of intergene correlation. The so-called gene correlation refers to the fact that the expression level (activity or amout of expression products, usually mRNA) of one gene can exert a positive or negative domino effect on that of the one other. To represent such direct or indirect correlation, reserchers have embarked on depicting a gene network. As opposed to accurate but time-and-fund-consuming biological techniques such as isotopic tracer and QPCR, statisticians have come up with a handy method invoking Gaussian graphical models (GGM), which we will elaborate in the section 2.

1.2 The virtue of sparsity

The correlation between genes is in essence the reaction between their expression products, or products of expression products. In all terrestrial organisms, biochemical reactions are conducted along the cycles or chains of metabolism. Consequently, most chemicals do not react with each other directly, resulting in a sparse network of gene expression.

2 Graphical models and conditional independences

2.1 conditional independences

Definition 2.1 (Conditional independence) Here we have three random variables X_1, X_2, X_3, we say X_1 is conditionally independent of X_2 given X_3 if

$$P(X_1, X_2|X_3) = P(X_1|X_3)P(X_2|X_3)$$

We denote this by $X_1 \perp X_2 \,|\, X_3$. In particular, provided a set of random variables $\{X_i,\ i \in I\}$, we simply say that X_j and X_k are conditionally independent if they are conditionally independent given all the others.

2.2 Undirected graphical models

In an undirected graph, all edges are undirected. Now we have an finite set of random variables

$$X = (X^{(1)}, \ldots, X^{(p)}) \sim P$$

and we want to visualize it via an undirected graph G with vertices V and edges E. This time each vertex represents a random variable. Edges essentially mean nothing, while the absence of edge between two vertices stands for the conditional independence between the two corresponding random variables. Indexing the set $V = 1, 2, \ldots, p$ with $|V| = p$, we can simply denote this by

$$(j,k)\, and\, (k,j) \notin E \iff X^{(j)} \perp X^{(k)} \,|\, X^{(V\setminus\{j,k\})}$$

.

The model discribed above is called an (undirected) graphical model. A graphical model visualizes the joint distribution of the entire set of random variables and remarkably faciliates the research of conditional independence.

2.3 Gaussian graphical models

We specify the undirected graphical models to the assumption

$$X = (X^{(1)}, \ldots, X^{(p)}) \sim N_p(\mu, \mathbf{\Sigma})$$

The edges in a GGM are given by the inverse of the covariance matrix, namely the precision matrix (or concentration matrix):

Theorem 2.2 In Gaussian graphical model, the absence of edge (j,k) is equivalent to zeros in the i,j-th and j,i-th entries of precision matrix.

$$(j,k)\, and\, (k,j) \notin E \iff X^{(j)} \perp X^{(k)} \,|\, X^{(V\setminus\{j,k\})} \iff \Sigma^{-1}_{j,k} = 0. \tag{1}$$

which means if the ij-th component of Σ^{-1} is zero, then variables i and j are conditionally independent, given the other variables. The proof of this and all other theorems are left in section 7.

3 Penalized estimation for covariance matrix and edge set

We suppose data $X_1, X_2, \ldots, X_n \, i.i.d \sim N_p(\mu, \boldsymbol{Sigma})$, we then have

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$$
$$\hat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})(X_i - \hat{\mu})^T \triangleq \boldsymbol{S}$$

Although the explicit expression for $\boldsymbol{\Sigma}^{-1}$'s MLE is $\boldsymbol{S}^{-1}$, we do not employ a direct matrix inversion, which we will discuss in detail in section 6. We turn to another method as follows.

We notice that the negative log-likelihood function for $\boldsymbol{\Sigma}^{-1}$ is

$$-l(\boldsymbol{\Sigma}^{-1}; \boldsymbol{S}) = \; -logdet\boldsymbol{\Sigma}^{-1} + trace(\boldsymbol{S}\boldsymbol{\Sigma}^{-1}) + Const \tag{2}$$

The estimation of an underlying precision matrix over samples is given the name *covariance selection* (actually precision selection) by Dempster in 1972 [1]. Here we accept that term while restricting our problem to Gaussian distribution, and we attach to it the pursuit of a sparse solution. Throughout this paper, we refer to the problem as *Gaussian covariance selection*.

As we do in the general likelihood problems for the Lasso, we add an l-1 penalty to off-diagnol elements, leading to

$$\hat{\boldsymbol{\Sigma}}^{-1}(\lambda) = argmin_{\boldsymbol{X}\succ 0} \; - logdet\boldsymbol{X} + trace(\boldsymbol{S}\boldsymbol{X}) + \lambda||\boldsymbol{X}||_{1,off} \tag{3}$$

in which λ>0 and

$$||\boldsymbol{X}||_{1,off} = \sum_{j\neq k} |\boldsymbol{X}_{j,k}| \tag{4}$$

where the minimization is over positive definite matrices. The following proposition justifies the well-posedness of the optimization problem.

Proposition 2.3 (Convexity and well-posedness) Assume that the diagonal entries of empirical covariance matrix $\boldsymbol{S}$ are strictly positive. Let $F(\boldsymbol{X}) = -logdet\boldsymbol{X} + trace(\boldsymbol{S}\boldsymbol{X}) + \lambda||\boldsymbol{X}||_{1,off}$ be the objective function defined on the positive-definite cone, then $F(\boldsymbol{X})$ is strictly convex and bounded from the below. Moreover, the optimization problem (3) is well-posed, i.e., the solution to (3) exists finitely and is unique.

So far a Gaussian graphical model has been embeded on gene expression network, and the Gaussian covariance selection problem culminates in a convex optimization problem. We will give it a final strike, if we resolve the convex optimization (3). At the very beginning we employed the vanilla gradient descent scheme with differentiable Huber

surrogate, yet only to find its inefficiency in recovering a sparse structure. Before we elaborate the drawbacks of this algorithm in Section 6, we will first introduce our final solution: three proximal gradient methods, ISTA, FISTA and MFISTA.

Proximal gradient methods are at first devised to overcome the indifferentiable points or weird domains of some special object functions. This group of algorithms are, before long, warmly embraced and widely studied because of their superior performance in various optimization issues. The iterative shrinkage-thresholding algorithm (ISTA), which is summarized from [2], specializes in tackling l-1 penalized optimization problems. Fast iterative shrinkage-thresholding algorithm (FISTA) and its monotonic version monotonic FISTA (MFISTA), proposed by Beck and Teboulle [3, 4], performs as upgraded versions of ISTA, yielding sparse solutions at a faster rate.

What has kept us in hunger is the feasibility of those proximal gradient methods in capturing a sparse Gaussian precision matrix, and especially on that breast cancer data from kaggle. Oztropak et al. have already had a glimpse on FISTA in their comparison between an inclusive range of Gaussian covariance selection algorithms [5]; yet proximal gradient algorithms' capability in this issue has not been specifically looked into. In the rest of this paper, we present theoretical justification for ISTA, FISTA and MFISTA's feasibility by adjusting Beck and Teboulle's proofs to (no. of our problem). Numerical experiments on both synthetic and real data also provide further insights.

Organization of the rest of this paper: In next section, we provide a comprehensive description from general proximal gradient methods to MFISTA toghther with their convergence properties. Also do we present numerical results on synthetic and real-world data in section 5. Section 6 is devoted to further discussions in a Q&A style to elaborate our motivations and improvements. Proofs of theorems and propositions are deferred till the last section.

4 Proximal gradient methods: from general case to MFISTA

4.1 Definition of proximal mapping

Before moving on to proximal gradient method itself, it is necessary to introduce the definition of proximal mapping

Definition 4.1 (proximal mapping) Given an Euclid space $\mathbb{E}$ and a function $f : \mathbb{E} \to \mathbb{R}$, the proximal mapping $\mathbb{E} \to \mathbb{E}$ with respect to f is given by

$$prox_f(\boldsymbol{x}) = argmin_{\boldsymbol{u}\in\mathbb{E}} \quad f(\boldsymbol{u}) + \frac{1}{2}||\boldsymbol{u} - \boldsymbol{x}||^2 \tag{5}$$

Here we assume that the minima taken in (5) is unique, for it is indeed the case in later discussions.

4.2 General proximal gradient method

Gradient is perheps the most commonly used measurement in numerical optimization. However, under the circumstance that unsmoothness occurs locally, we are forced to make a concession. Consider the following problem

$$min_{\boldsymbol{x}\in\mathbb{E}} \quad F(\boldsymbol{x}) \triangleq f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{6}$$

with f smooth while g maybe not (say, the l-1 penalty term). The essence of proximal gradient method is an iterative process at each step of which the smooth part f is approximated pseudo-quadratically. A minimizer of the proximate function is then drawn as the next point on the path towards the global minima, that is

$$\begin{aligned} \boldsymbol{x}^{k+1} &= argmin_{\boldsymbol{x}\in\mathbb{E}} \quad f(\boldsymbol{x}_k) + \langle \nabla f(\boldsymbol{x}_k), f(\boldsymbol{x}) - f(\boldsymbol{x}_k)\rangle + \frac{1}{2t_k}||\boldsymbol{x} - \boldsymbol{x}_k||^2 + g(\boldsymbol{x}) \\ &= argmin_{\boldsymbol{x}\in\mathbb{E}} \quad t_k g(\boldsymbol{x}) + \frac{1}{2}||\boldsymbol{x} - (\boldsymbol{x}^k - t_k \nabla f(\boldsymbol{x}^k))||^2 \end{aligned} \tag{7}$$

where t_k is a positive constant named as the stepsize chosen at the k-th step.

Alternatively, (7) can be written in the proximal mapping form

$$\boldsymbol{x}^{k+1} = prox_{t_k g}(\boldsymbol{x}^k - t_k \nabla f(\boldsymbol{x}^k)) \tag{8}$$

which leads to the prototype of general proximal gradient method.

Algorithm 1 General Proximal Gradient Method

Initialization: pick $x^0 \in dom(f),\ k = 0$
While: stopping criteria not met
1: pick t_k>0
2: $\boldsymbol{x}^{k+1} = prox_{t_k g}(\boldsymbol{x}^k - t_k \nabla f(\boldsymbol{x}^k))$
3: $k = k + 1$

We present the following convergence property of general proximal methods in convex cases.

Theorem4.2 (Convergence of general proximal gradient method) Assume that f is L-smooth and convex, g is continuous and convex. Let $\boldsymbol{x}^*$ be the optima of (6). A constant stepsize schedule $t_k \equiv \frac{1}{L}$ generates a sequence $\left\{\boldsymbol{x}^k\right\}$ satisfying

$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) \leq \frac{L||\boldsymbol{x}^0 - \boldsymbol{x}^*||^2}{2k} \tag{9}$$

In particular, when $g(\boldsymbol{X}) = \lambda||\boldsymbol{X}||_{1,off}$ and $f(\boldsymbol{X}) = -logdet\boldsymbol{X} + trace(\boldsymbol{S}\boldsymbol{X})$, the proximal mapping turns out to be

$$\begin{aligned} \boldsymbol{X}^{k+1} &= \mathcal{T}_{\lambda t_k}(\boldsymbol{X}^k - t_k \nabla f(\boldsymbol{X}^k)) \\ &= \mathcal{T}_{\lambda t_k}(\boldsymbol{X}^k - t_k(-\boldsymbol{X}_k^{-1} + \boldsymbol{S})) \end{aligned} \tag{10}$$

in which

$$\mathcal{T}_t(\boldsymbol{Y})_{ij} \triangleq \begin{cases} \boldsymbol{Y}_{ij} & i = j \\ \boldsymbol{Y}_{ij} + t & i \neq j \quad \boldsymbol{Y}_{ij} < -t \\ 0 & i \neq j \quad |\boldsymbol{Y}_{ij}| < -t \\ \boldsymbol{Y}_{ij} - t & i \neq j \quad \boldsymbol{Y}_{ij} > \quad t \end{cases} \tag{11}$$

In this case, Algotithm 1 is called an interative shrinkage-thresholding algorithm (ISTA) and writen as follows.

Algorithm 2 ISTA

Initialization: pick $\boldsymbol{X}^0 \in dom(f),\ k = 0$

While: stopping criteria not met

1: pick $t_k > 0$

2: $\boldsymbol{X}^{k+1} = \mathcal{T}_{\lambda t_k}(\boldsymbol{X}^k - t_k \nabla f(\boldsymbol{X}^k))$

3: $k = k + 1$

4.3 Accelerated variants: FISTA and MFISTA

The preceding subsection has witnessed an $O(1/k)$ rate of convergence. Yet for greedy researchers, it is too slow to meet their satisfaction. Beck and Teboulle come up with an upgraded version named FISTA with the following description

Algorithm 3 FISTA

Initialization: pick $\boldsymbol{x}^0 \in dom(f)$, $k = 0$, $r_0 = 1$, set $\boldsymbol{y}^0 = \boldsymbol{x}^0$

While: stopping criteria not met

1: pick $t_k > 0$

2: $\boldsymbol{x}^{k+1} = prox_{t_k g}(\boldsymbol{y}^k - t_k \nabla f(\boldsymbol{y}^k))$

3: $r_{k+1} = \frac{1+\sqrt{1+4r_k^2}}{2}$

4: $\boldsymbol{y}^{k+1} = \boldsymbol{x}^{k+1} + \frac{r_k - 1}{r_{k+1}}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k)$

5: $k = k + 1$

Remark 4.3 Although FISTA, as its name implies, was first devised to solve penalized optimization, it can be readily fitted to a bunch of situations in which the object funtion F can be seperated into f and g subject to certain restrictions. Hence in Algorithm 3 we demonstrate FISTA in a general proximal gradient form.

FISTA as is illustrated above guarantees a convergence rate of $O(1/k^2)$, which we summarizes as

Theorem 4.4 ($O(1/k^2)$ convergence of FISTA) Suppose that the same assumptions in Theorem 4.2 hold. The sequence $\boldsymbol{x}^k$ generated by FISTA with constant stepsize $t_k \equiv \frac{1}{L}$ satisfies

$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) \leq \frac{2L||\boldsymbol{x}^0 - \boldsymbol{x}^*||^2}{(k+1)^2} \tag{12}$$

The sequence of function values generated by FISTA is not necessarily nonincreasing. In fact FISTA may suffer from an unwanted fluctuation, which has been remedied by Beck and Teboulle with their trump card: the MFISTA. Equiped with a slight revision, MFISTA yields a sequence with nonincreasing function values and meanwhile preserves an $O(1/k^2)$ rate.

Algorithm 4 MFISTA

Initialization: pick $\boldsymbol{x}^0 \in dom(f)$, $k=0$, $r_0=1$, set $\boldsymbol{y}^0=\boldsymbol{x}^0$
While: stopping criteria not met
1: pick $t_k>0$
2: $\boldsymbol{z}^k = prox_{t_k g}(\boldsymbol{y}^k - t_k \nabla f(\boldsymbol{y}^k))$
3: pick $\boldsymbol{x}^{k+1} \in dom(f)$ such that $F(\boldsymbol{x}^{k+1}) \leq min\left\{F(\boldsymbol{x}^k), F(\boldsymbol{z}^k)\right\}$
4: $r_{k+1} = \frac{1+\sqrt{1+4r_k^2}}{2}$
5: $\boldsymbol{y}^{k+1} = \boldsymbol{x}^{k+1} + \frac{r_k - 1}{r_{k+1}}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k) + \frac{r_k}{r_{k+1}}(\boldsymbol{z}^k - \boldsymbol{x}^{k+1})$
6: $k = k+1$

Theorem 4.5 ($O(1/k^2)$ convergence of MFISTA) Suppose that the same assumptions in Theorem 4.2 hold. The sequence $\boldsymbol{x}^k$ generated by MFISTA with constant stepsize $t_k \equiv \frac{1}{L}$ satisfies

$$F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*) \leq \frac{2L||\boldsymbol{x}^0 - \boldsymbol{x}^*||^2}{(k+1)^2} \tag{13}$$

Remark 4.6 Step 3 in Algorithm 4 can be realized through different approaches. For example, we may set

$$\boldsymbol{x}^{k+1} = \begin{cases} \boldsymbol{z}^k & F(\boldsymbol{z}^k) < F(\boldsymbol{x}^k) \\ \boldsymbol{x}^k & \text{otherwise} \end{cases} \tag{14}$$

Throughout the rest of this paper, we employ (12) as step 3 in the loop of MFISTA algorithm.

Rigorous readers may argue that theorem 4.2, 4.4 and 4.5 indicate merely the convergence in function values. It is indeed a crucial issue which has not been clearly stated in [3, 4, 11]. However in our optimization problem (3), two modes of convergnece are equivalent. The following proposition resolves the controversy.

Proposition 4.7 In optimization problem (3), the convergence in $\left\{F(\boldsymbol{X}^k)\right\}$ implies the convergence of $\left\{\boldsymbol{X}^k\right\}$ in the matrix space.

Finally, it is of great neccesity to justify the compatibility of our objective function in (3) with the above proximal gradient methods. We state as follows.

Propersition 4.8 (Compatibility) The objective function $F(\boldsymbol{X})$ in (3) together with its decomposition into f and g satisfies the assumptions in theorem 4.2.

5 Numerical Experiments

5.1 Preparations

Before numerical tests, we clarify some details of our setting.

(a) Stopping criterion: We choose dual gap $< 1e-8$ as stopping criterion. The following proposition justifies our choice.

Proposition 5.1 Strong duality holds in optimization problem (). Moreover, the dual gap $d(\boldsymbol{X})$ is given by

$$d(\boldsymbol{X}) = tr(\boldsymbol{S}\boldsymbol{X}) + \lambda||\boldsymbol{X}||_{1,off} - p \tag{15}$$

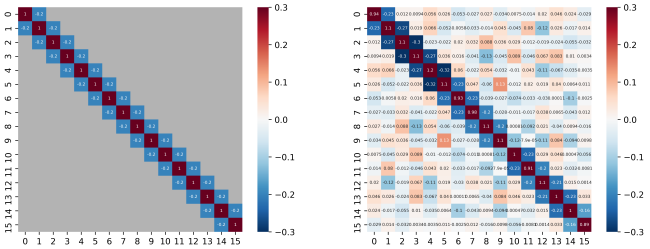
(b) λ scaling: We go through a range of λ with five fold cross-validation with respect to MLE. In later session, more model assessment criteria will be discussed.

(c) Stepsize schedule: We apply a constant stepsize $t_k \equiv 0.5$, which is small enough. The reason for that choice is indicated in the proof of proposition 4.8.

(d) Data processing We z-score each feature of the raw data. Z-scoring is in essence a linear transformation and will not alternate the network structure.

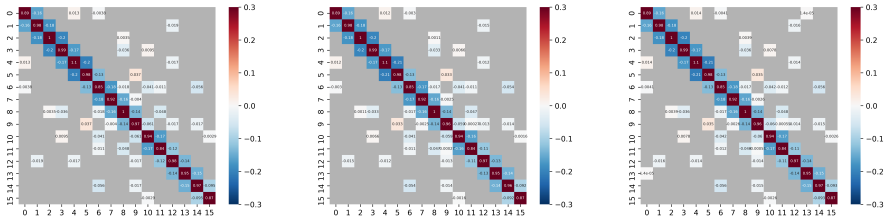
5.2 Experiments on synthetic data

We synthesize different multivariate Gaussian distributions to test ISTA, FISTA and MFISTA.



(a) true precision (b) noisy $\boldsymbol{S}^{-1}$

Figure 1: Precision matrix and inverse of empirical covariance matrix



(a) ISTA (b) FISTA (c) MFISTA

Figure 2: Precision matrix selected by proximal gradient methods

Example 1: chain model with $p = 16$, $n = 400$

What lies in the heart of covariance selection problem is the recovery of a sparse network pattern. In some mood, the bias between the selected precision matrix and the underlying real one is not of our greatest concern. By contrast, the capability to capture true edges and exclude hollow ones makes an upmost criterion for algorithm evaluation. In a word, *to be zero or not to be zero, is the question.*

Consequently, to further evaluate the quality of our covariance selection, we introduce true negative rate (TNR) and true positive rate (TPR) to reveal the specificity and sensitivity of discussed algorithms.

Definition 5.2 Denote the truth of sparsity pattern by $\mathbf{\Sigma}^{-1}$ and the selected pattern by $\hat{\boldsymbol{X}}$. Definitions of TNR (specificity) and TPR (sensitivity) are as follows

$$TNR = \frac{\#\left\{(i,j) : \mathbf{\Sigma}_{ij}^{-1} = \hat{\boldsymbol{X}}_{ij} = 0\right\}}{\#\left\{(i,j) : \mathbf{\Sigma}_{ij}^{-1} = 0\right\}} \tag{16}$$

$$TPR = \frac{\#\left\{(i,j) : \mathbf{\Sigma}_{ij}^{-1} \neq 0,\ \hat{\boldsymbol{X}}_{ij} \neq 0\right\}}{\#\left\{(i,j) : \mathbf{\Sigma}_{ij}^{-1} \neq 0\right\}} \tag{17}$$

Here we compare the three proximal gradient algorithms with graphical lasso (GLasso) [7] which appears to be the most prevalent covariance selection algorithm.

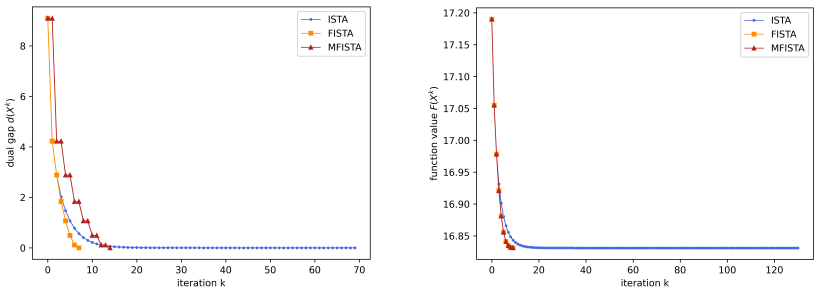
algorithm	TNR	TPR	average run time(ms)
ISTA	83.81%	100.0%	48.97 (143 iter)
FISTA	82.86%	100.0%	4.02 (9 iter)
MFISTA	81.90%	100.0%	5.60 (10 iter)
Glasso	78.10%	100.0%	15.11

Table 1: Performance on example 1

It is commonly witnessed that covariance selection algorithms seldom miss true edges in the network (non-zero entries in the precision matrix), yet tend to add false connections between vertices. In a word, those algorithms are highly sensitive but not fully specilized, which is verified by the full TPR and relatively low TNR in our experimental outcome (see Table 1).

When it comes to time complexity, we notice that the prototype ISTA lags behind Glasso. Its counterparts with $O(1/k^2)$ rate, however, possess superior performances beyond Glasso. More results are provided below in Figure 3 to elaborate the evolution from ISTA to MFISTA.

For more results on synthetic data, see supplementary materials.

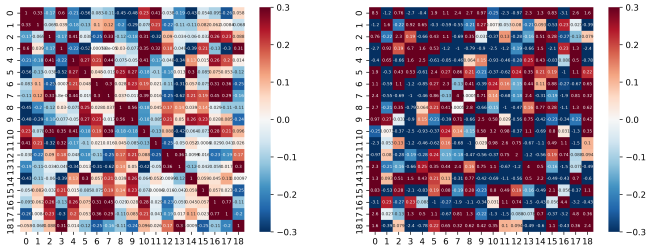


(a) dual gap curve (b) function value curve

Figure 3: Curves of dual gap and function value under $\lambda = 0.5$, $t = 0.25$

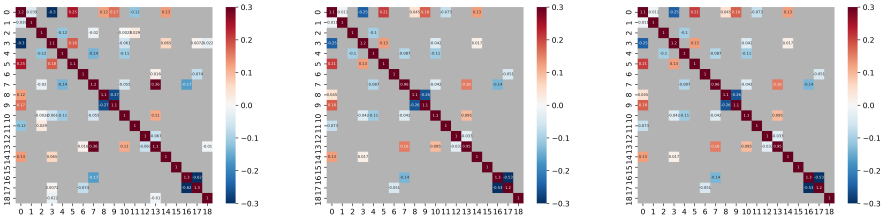
5.3 Experiments on real world data

Example 2 We begin with the breast cancer data from Kaggle mentioned in abstract. Due to limitations in devices, we have simplified the data by picking only 19 out of 10000+ breast-cancer-related genes with 40 samples, i.e., $p = 19$ and $n = 40$. Settings in section 5.1 are maintained. Although, in the realm of biology, the network embeded in this 19-gene-group remains an open problem, our statistical trials may provide helpful clues. The precision selected by ISTA, FISTA, and MFISTA are as follows.



(a) empirical covariance $\boldsymbol{S}$ (b) noisy $\boldsymbol{S}^{-1}$

Figure 4: Empirical covariance matrix and its inverse



(a) ISTA (b) FISTA (c) MFISTA

Figure 5: Selected precision matrix for breast cancer data

As is shown in Figure 6, we notice that FISTA and MFISTA capture exactly the same network, while ISTA's outcome is slightly different with 6 more edges (noted with dash lines).

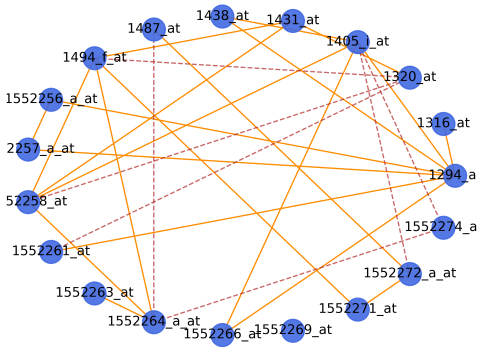


Figure 6: Gene network selected by proximal gradient methods

6 Further Discussions

This section is devoted to supplementary talks. In this section, we answer several questions to reveal some of our motivations and come up with feasible improvements.

Question 1: According to the formal invariance property of maximum likelihood estimator, since S is the MLE of covariance matrix, S^{-1} is exactly the MLE of precision matrix. Why we turn to a rather intricate convex optimization problem as opposed to a direct matrix inversion operation?

The pursuit of sparsity is obviously one reason for our choice. Besides, the situations in which $p>n$ are common. Under such situations, the empirical covariance matrices are not ivertible. Even when $p \leq n$, if n is not much larger than p, a direct inversion may suffer from high variance.

Question 2: What can we benefit from 'proximal'? Do straightforward (sub)gradient methods work in this problem?

In section 3, we have mentioned our failed first trial with vanilla gradient descent method in which Huber function $H(\boldsymbol{X})$ is employed as a surrogate of l-1 loss. Thanks to the differentiability of Huber function, gradient descent method can be well fitted to it. This method can indeed compress matrix entries to infinitesimal values, but never exactly to 0. It is the intrinsic float error in numerical computation which hinders gradient descent from sweeping matrix entires to exact 0. The following picture is more than words.

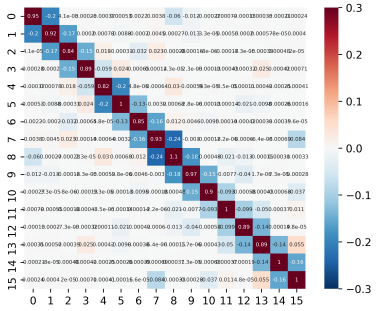
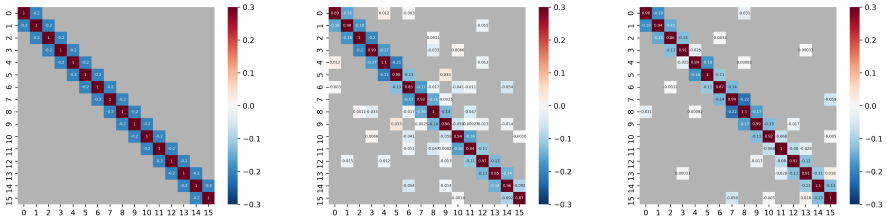


Figure 7: Example 1 solved by gradient methods with Huber δ=1e-8

Similar phenomena also occur in other sparse solution pursuit procesures such as l-1 penalized linear regression. In conclusion, only under the presence of an explicit formula that equating elements to zero (like ()) can an algorithm realize a sparse numerical solution.

Question 3: In Gaussian covariance selection problems, is cross-validation the best strategy for choosing hyper-parameter?

The topic on model selection is bound to be a long talk. We decide to take a mere glimpse on it by providing some comparisons on numerical results. With regard to example 1, we compare the precision matrix selected via 5-fold cross-validation with that captured by the well-known Akaike Information Criterion (AIC). The groud algorithm is FISTA.



(a) truth (b) 5-fold cross-validation (c) AIC

Figure 8: Comparison between cross-validation and AIC, based on example 1 with FISTA

Apparently in this example AIC performs better.

7 Theoretical Proof

Proof of theorem 2.2: It is nothing new in multivariate statistics. For covenience in notation we denote the precision matrix $\mathbf{\Sigma}^{-1}$ by $\boldsymbol{K}$. In the view of kernel function, the conditional distribution of $X^{(j)}$ and $X^{(k)}$ is bivariate Gaussian with kernel

$$exp\left\{-\frac{1}{2}(\boldsymbol{K}_{j,j}x_j^2+2\boldsymbol{K}_{j,k}x_jx_k+\boldsymbol{K}_{k,k}x_k^2+C_jx_j+C_kx_k)\right\}$$

According to the properties of bivariate Gaussian distribution, $X^{(j)}$ and $X^{(k)}$ are conditionally independent if and only if their kernel factorizes, i.e., $\boldsymbol{K}_{j,k}=0$.

Proof of proposition 2.3: Since the none smooth part $g(\boldsymbol{X})$ is convex, it suffices to show that $f(\boldsymbol{X})$ is strictly convex. Notice that the Hessian $\nabla^2 f(\boldsymbol{X}) = \boldsymbol{X}^{-1} \otimes \boldsymbol{X}^{-1}$. The collection of eigenvalues of a kronecker product $\{\gamma : \gamma \in spec(\boldsymbol{A} \otimes \boldsymbol{B})\}$ is $\{\lambda\mu : \lambda \in spec(\boldsymbol{A}),\ \mu \in spec(\boldsymbol{B})\}$, resulting in the positive-definiteness of $\boldsymbol{X}^{-1} \otimes \boldsymbol{X}^{-1}$ and subsequently the strict convexity of $f(\boldsymbol{X})$.

For the existence and uniqueness of a finite solution, we rephrase (3) in the constrainted optimization form

$$min_{\boldsymbol{X} \succ 0}\ f(\boldsymbol{X}) = -logdet(\boldsymbol{X}) + trace(\boldsymbol{S}\boldsymbol{X}) \tag{18}$$

$$s.t.\ \sum_{j \neq k} |\boldsymbol{X}_{j,k}| \leq C(\lambda) \tag{19}$$

Expanding f into three terms and invoking Hadamard inequality leads to

$$f(\boldsymbol{X}) \geq -log(\prod_i \boldsymbol{X}_{i,i}) + \sum_i \boldsymbol{S}_{i,i}\boldsymbol{X}_{i,i} + \sum_{j \neq k} \boldsymbol{S}_{j,k}\boldsymbol{X}_{j,k} \tag{20}$$

By the assumption that $\boldsymbol{S}$ possesses strictly positive diagonal entries, the sum of the first two terms is bounded from the below. Also, the final term is bounded. Thus the objective f is bounded from below subject to the constraint. Combine the conclusion with strict convexity, it is then crystal clear that a unique finite solution $\boldsymbol{X}^*$ exists.

The convergence analysis of proximal gradient methods is based on [2, 9]

Lemma 7.1 (Fundemental prox-grad inequality) Suppose that f and g satisfies the assumptions in theorem 4.2. Denote $prox_{tg}(\boldsymbol{x} - t\nabla f(\boldsymbol{x}))$ simply by $P_L(\boldsymbol{x})$ where $L = \frac{1}{t}$. For any $\boldsymbol{x} \in \mathbb{E}$ and $\boldsymbol{y} \in int(dom(f))$ satisfying

$$f(P_L(\boldsymbol{y})) \leq f(\boldsymbol{y}) + \langle \nabla f(\boldsymbol{y}),\ P_L(\boldsymbol{y}) - \boldsymbol{y} \rangle + \frac{L}{2}||P_L(\boldsymbol{y}) - \boldsymbol{y}||^2 \tag{21}$$

we have

$$F(\boldsymbol{x}) - F(P_L(\boldsymbol{x})) \geq \frac{L}{2}||\boldsymbol{x} - P_L(\boldsymbol{y})||^2 - \frac{L}{2}||\boldsymbol{x} - \boldsymbol{y}||^2 + \mathcal{L}_f(\boldsymbol{x} - \boldsymbol{y}) \tag{22}$$

where

$$\mathcal{L}_f(\boldsymbol{x} - \boldsymbol{y}) = f(\boldsymbol{x}) - f(\boldsymbol{y}) - \langle \nabla f(\boldsymbol{y}),\ \boldsymbol{x} - \boldsymbol{y} \rangle \tag{23}$$

Proof: It is quite a common conclusion which can be found in many books such as [11].

Lemma 7.2 (Monotonicity in function value) Suppose that the assumptions in theorem 4.2 hold. The sequence of function values $\left\{F(\boldsymbol{X}^k)\right\}$ generated by a general proximal gradient method is non-increasing.

Proof: By assumptions in theorem 4.2, we notice that $f(\boldsymbol{x}^{k+1}) \leq f(\boldsymbol{x}^k) + \left\langle \nabla f(\boldsymbol{x}^k),\ \boldsymbol{x}^{k+1} - \boldsymbol{x}^k \right\rangle + \frac{L}{2}||\boldsymbol{x}^{k+1} - \boldsymbol{x}^k||^2$. Then invoking lemma 7.1 soon leads to the desired result.

Proof of theorem 4.2 Denote $\frac{1}{t}$ by L as usual. Since a constant stepsize schedule satisfies (21), by the fundamental prox-grad inequality we have

$$\begin{aligned} F(\boldsymbol{x}^*) - F(\boldsymbol{x}^{\boldsymbol{k+1}}) \geq & \frac{L}{2}(||\boldsymbol{x}^* - \boldsymbol{x}^{k+1}||^2 - ||\boldsymbol{x}^* - \boldsymbol{x}^k||^2) + \mathcal{L}_f(\boldsymbol{x}^* - \boldsymbol{x}) \\ \geq & \frac{L}{2}(||\boldsymbol{x}^* - \boldsymbol{x}^{k+1}||^2 - ||\boldsymbol{x}^* - \boldsymbol{x}^k||^2) \end{aligned}$$

Hence we obtain

$$\sum_{j=0}^{k-1}(F(\boldsymbol{x}^{j+1}) - F(\boldsymbol{x}^*)) \leq \frac{L}{2}(||\boldsymbol{x}^* - \boldsymbol{x}^0||^2 - ||\boldsymbol{x}^* - \boldsymbol{x}^k||^2) \leq \frac{L}{2}||\boldsymbol{x}^* - \boldsymbol{x}^0||^2$$

Moreover, lemma 7.2 implies $k(F(\boldsymbol{x}^k) - F(\boldsymbol{x}^*)) \leq \sum_{j=0}^{k-1}(F(\boldsymbol{x}^{j+1}) - F(\boldsymbol{x}^*))$, which ends up to be the desired result.

Proof of theorem 4.4 and 4.5 Although more explicit operations are equipped in the cases of FISTA and MFISTA, the proof is quite similar to that of their general proximal gradient counterparts. To avoid redundant labor, we invite readers to [3, 4].

Proof of proposition 4.7 It is strict convexity and well-posedness that count on. For any sequence $\left\{\boldsymbol{X}^k\right\}$ that converges in function value to $\boldsymbol{X}^*$, if bounded, according to Bolzano-Weierstrass theorem, every subsequence has a further convergent subsequence that converges to a point $\boldsymbol{X}^{**}$. The convergence in function value gives $F(\boldsymbol{X}^{**}) = F(\boldsymbol{X}^*)$. Then by well-posedness we obtain $\boldsymbol{X}^{**} = \boldsymbol{X}^*$, which proves the convergence of $\left\{\boldsymbol{X}^k\right\}$ to $\boldsymbol{X}^*$.

Otherwise we assume the unboundedness of $\left\{\boldsymbol{X}^k\right\}$. Then we may select a subsequence $\boldsymbol{X}^{k(n)}$ with matrix norms growing to infinity. WLOG, suppose that $||\boldsymbol{X}^{k(n)} - \boldsymbol{X}^*||>1$ for all n. Denote the intersection between the line segment $\overline{\boldsymbol{X}^{k(n)}\boldsymbol{X}^*}$ and the unit sphere $\mathcal{S}(\boldsymbol{X}^*, 1)$ by $\boldsymbol{Y}^n$. Strict convexity implies $limsup\ F(\boldsymbol{Y}^n) \leq F(\boldsymbol{X}^*)$. Once again by Bolzano-Weierstrass theorem, the boundedness of $\mathcal{S}(\boldsymbol{X}^*, 1)$ reveals a point $\boldsymbol{Y}^*$ on it with no greater function value than $\boldsymbol{X}^*$, which contradicts the well-posedness.

Proof of proposition 4.8 Given the above conclusions, it suffices to prove that $f(\boldsymbol{X})$ is L-smooth for some positive constant L. Despite that $f(\boldsymbol{X})$ is not L-smooth globally, we may assume that, with stepsizes small enough, the trajectory of optimization is contained in a close neighbourhood around $\boldsymbol{X}^*$ within $\mathbb{S}^p_{++}$, namely, $\boldsymbol{X}^k \succ \rho\boldsymbol{I}$ for some $\rho>0$. Then we deduce as follows.

$$\begin{aligned} ||\nabla f(\boldsymbol{Y}) - \nabla f(\boldsymbol{X})|| = & ||\boldsymbol{X}^{-1} - \boldsymbol{Y}^{-1}|| = ||\boldsymbol{X}^{-1}(\boldsymbol{Y} - \boldsymbol{X})\boldsymbol{Y}^{-1}|| \\ \leq & ||\boldsymbol{X}^{-1}||\ ||\boldsymbol{Y} - \boldsymbol{X}||\ ||\boldsymbol{Y}^{-1}|| \\ \leq & \frac{p^4}{\rho^2}||\boldsymbol{Y} - \boldsymbol{X}|| \end{aligned} \tag{24}$$

Proof of proposition 5.1 Both objective and constraint functions are convex, and there exists $\boldsymbol{X}^0 \in dom(f) = \mathbb{S}^p_{++}$ satisfying the constraints. Therefore, according to Slater's condition, the strong duality holds. Inspired by [12], we may

rewrite (18) and (19) as

$$max_{\boldsymbol{X}\succ 0}\, min_{||\boldsymbol{V}||_{\infty,off}\leq C(\lambda),\, \boldsymbol{V}_{i,i}=0}\; -logdet(\boldsymbol{X}) + trace(\boldsymbol{S}(\boldsymbol{X}+\boldsymbol{V})) \tag{25}$$

Intercharge the max and min. Simple tranformation yields the dual problem

$$max_{\boldsymbol{U}\succ 0}\, logdet(\boldsymbol{U}) + p \tag{26}$$

$$s.t.\; ||\boldsymbol{U}-\boldsymbol{S}||_{\infty,off} \leq C(\lambda),\; and\, \boldsymbol{U}_{i,i} = \boldsymbol{S}_{i,i},\; i = 1,2,\ldots,p \tag{27}$$

The dual objective function is restricted in a compact region and is hence well-posed. Moreover, optimality condition relates primal and dual solution with $\boldsymbol{X}\boldsymbol{U} = \boldsymbol{I}$. Substituting $\boldsymbol{U}$ with $\boldsymbol{X}^{-1}$ provides the explicit dual gap.

References

[1] Dempster A P. Covariance selection[J]. Biometrics, 1972: 157-175.

[2] Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint[J]. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 2004, 57(11): 1413-1457.

[3] Beck A, Teboulle M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems[J]. SIAM journal on imaging sciences, 2009, 2(1): 183-202.

[4] Beck A, Teboulle M. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems[J]. IEEE transactions on image processing, 2009, 18(11): 2419-2434.

[5] Oztoprak F, Nocedal J, Rennie S, et al. Newton-like methods for sparse inverse covariance estimation[J]. Advances in neural information processing systems, 2012, 25: 755-763.

[6] Ravikumar P, Raskutti G, Wainwright M J, et al. Model Selection in Gaussian Graphical Models: High-Dimensional Consistency of l1-regularized MLE[C]//NIPS. 2008: 1329-1336.

[7] Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso[J]. Biostatistics, 2008, 9(3): 432-441.

[8] Edwards D. Introduction to graphical modelling[M]. Springer Science & Business Media, 2012.

[9] Beck A. First-order methods in optimization[M]. Society for Industrial and Applied Mathematics, 2017.

[10] Wainwright M J. High-dimensional statistics: A non-asymptotic viewpoint[M]. Cambridge University Press, 2019.

[11] Boyd S P, Vandenberghe L. Convex optimization[M]. Cambridge university press, 2004.

[12] Scheinberg K, Ma S, Goldfarb D. Sparse inverse covariance selection via alternating linearization methods[J]. arXiv preprint arXiv:1011.0097, 2010.