Joint Gaussian Graphical Model Review Series – IV

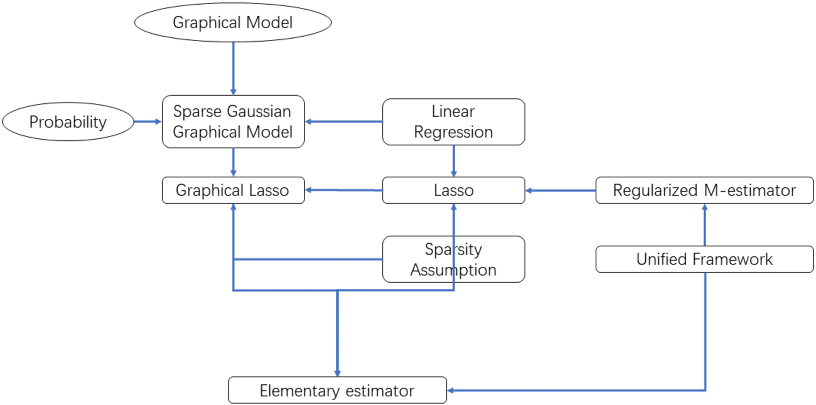
A Unified Framework for M-estimaotr and Elementary Estimators

Beilun Wang
Advisor: Yanjun Qi

[1]Department of Computer Science, University of Virginia
http://jointggm.org/

July 21st, 2017

Road Map


Graphical Model
Probability
Sparse Gaussian Graphical Model
Linear Regression
Graphical Lasso
Lasso
Regularized M-estimator
Sparsity Assumption
Unified Framework
Elementary estimator

Outline

Notation

Notation

- $\mathcal{L}$ The loss function.
- $\mathcal{R}$ The Regularization function (norm).
- $\mathcal{R}^*$ The Dual norm of $\mathcal{R}$.

Review

Review from last talk

- Likelihood of the precision matrix in the Gaussian case
- Graphical Model Basics

Regularized M-estimator

Example

We want to buy a TV.

Target:

Constrains: 4K, 65 inch

Result:

Regularized M-estimator

M-estimator

In statistics, M-estimators are a broad class of estimators, which are obtained as the minima of sums of functions of the data.
The parameters are estimated by argmin the sums of functions of the data.

target

$\mathcal{L}(X, \theta)$ the loss function

Conditions

$\mathcal{R}(\theta)$ the Regularization function

Therefore, the whole objective function is:

$$\underset{\theta}{\operatorname{argmin}}\, \mathcal{L}(X, \theta) + \lambda_n \mathcal{R}(\theta) \tag{3.1}$$

Example: Linear Model

Let's use the linear regression model as an example.

Target

Find β, such that $X\beta = y$.

Constrains: Sparsity

- **Prediction Accuracy:** Sacrifice a little bias and reduce the variance. Improve the overall performance.
- **Interpretation:** With a large number of predictors, we often would like to determine a smaller subset that exhibits the strongest effect.

$$\underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2 \tag{3.2}$$

$$\text{Subject to: } ||\beta||_0 \leq t \tag{3.3}$$

Example: Lasso

Since ℓ_0-norm is not a convex function, we need the closest convex function of ℓ_0-norm.

$$\underset{\beta}{\operatorname{argmin}} ||y - X\beta||_2 \tag{3.4}$$

$$\text{Subject to: } ||\beta||_1 \leq t \tag{3.5}$$

Lasso

$$\underset{\beta}{\operatorname{argmin}} \, ||y - X\beta||_2 + \lambda_n ||\beta||_1$$

Other equivalent formulation

$$\underset{\beta}{\operatorname{argmin}} ||\beta||_1 \tag{3.6}$$

$$\text{Subject to: } y = X\beta \tag{3.7}$$

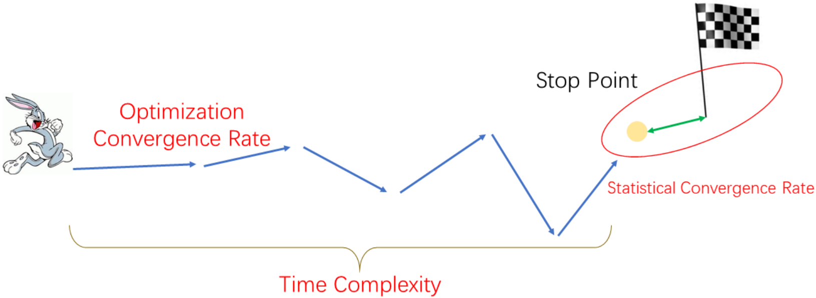
Dantzig selector

$$\underset{\beta}{\operatorname{argmin}} ||\beta||_1 \tag{3.8}$$

$$\text{Subject to: } ||X^T(X\beta - y)||_\infty \leq \lambda_n \tag{3.9}$$

A unified framework

Three major Criteria



Three major Criteria

- Statistical Convergence Rate: How close is between your estimated parameter and the true parameter. It corresponds to estimation error and approximation error.
- Computational Complexity: How fast the algorithm is with respect to certain parameters, e.g., n and p.
- Optimization Rate of Convergence: How fast each optimization step move to the estimated parameter, such as linear or quadratic.

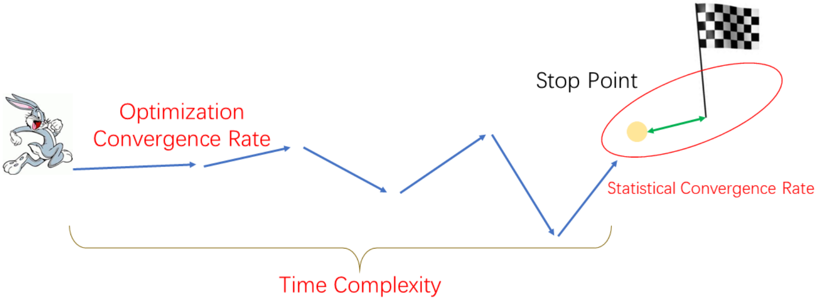
Traditional statisticians focus on the statistical convergence rate (Accuracy).

High dimension vs low dimension

- low dimension: when n is large, the error is asymptotic 0 by the law of large number.
- high dimension (i.e.,$p/n \rightarrow c \neq 0$): the error is not asymptotic 0.

High dimensional analysis is relative hard. Traditionally, we need carefully proof for every estimator.

Three major Criteria



A unified framework for M-estimator [Negahban et al.(2009)Negahban, Yu, Wainwright, and Ravi

Decomposability of $\mathcal{R}$

Suppose a subspace $\mathcal{M} \subset \mathbb{R}^p$, a norm-based regularizer $\mathcal{R}$ is decomposable with respect to $(\mathcal{M}, \bar{\mathcal{M}}^\perp)$ if

$$\mathcal{R}(\theta + \gamma) = \mathcal{R}(\theta) + \mathcal{R}(\gamma)$$

for all $\theta \in \mathcal{M}$ and $\gamma \in \bar{\mathcal{M}}^\perp$, where $\bar{\mathcal{M}}^\perp := \{v \in \mathbb{R}^p | < u, v >= 0 \forall u \in \bar{\mathcal{M}}\}$.

Subspace compatibility constant

$$\Phi(\mathcal{M}) := \sup_{u \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(u)}{||u||}$$

with respect to the pair $(\mathcal{R}, || \cdot ||)$.

A unified framework for M-estimator [Negahban et al.(2009)Negahban, Yu, Wainwright, and Ravi

Example: ℓ_1

ℓ_1 is decomposable and the $\Phi(\mathcal{M}) = \sqrt{s}$ with respect to (ℓ_1, ℓ_2).

A unified framework for M-estimator

Decomposable Regularizer

Restricted Strict Convex loss function

$\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}(\theta^*))$

$$\|\widehat{\theta}_{\lambda_n} - \theta^*\| \leq 9\frac{\lambda_n^2}{\kappa_{\mathcal{L}}}\Psi^2(\overline{\mathcal{M}})$$

$$\mathcal{R}(\widehat{\theta}_{\lambda_n} - \theta^*) \leq 12\frac{\lambda_n}{\kappa_{\mathcal{L}}}\Psi^2(\overline{\mathcal{M}}).$$

Example: Lasso

$$||\widehat{\theta}_{\lambda_n} - \theta^*||_2^2 \leq O(\frac{s \log p}{n})$$

In high dimensional setting, the sparsity assumption actually improves the convergence rate a lot.

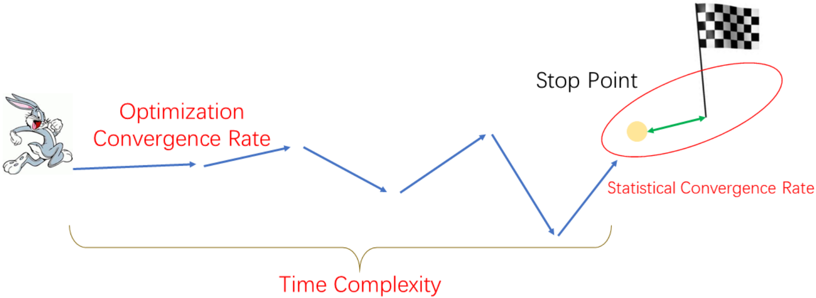
Elementary Estimator

We have a very powerful tool to easily prove the convergence rate. We can also follow the similar process to prove the convergence rate for estimators like Dantzig Selector.

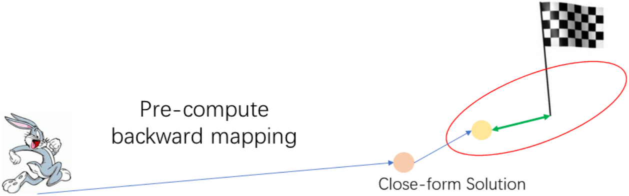
However, a lot of statistical method is slow when p and n are large and they are not scalable at all.

Are there any estimators with close form solution for the statistic problem, which also achieve the optimal convergence rate?

Three major Criteria



Three major Criteria



Elementary Estimator[Yang et al.(2014b)Yang, Lozano, and Ravikumar]

$$\underset{\theta}{\operatorname{argmin}} \mathcal{R}(\theta) \tag{5.1}$$

$$\text{Subject to: } \mathcal{R}^*(\theta - \mathcal{B}^*(\widehat{\phi})) \leq \lambda_n \tag{5.2}$$

Here $\mathcal{B}^*(\widehat{\phi})$ is a backward mapping for $\widehat{\phi}$.

Example: sparse linear regression[Yang et al.(2014a)Yang, Lozano, and Ravikumar]

$$\underset{\theta}{\operatorname{argmin}} ||\theta||_1 \tag{5.3}$$

$$\text{Subject to: } ||\theta - (X^T X + \epsilon I)^{-1} X^T y||_\infty \leq \lambda_n \tag{5.4}$$

Summary

- We review the unified framework for M-estimator, which can be applied to most regularized M-estimator problem
- Following the similar proof strategy, we have the set of elementary estimators.

References I

S. Negahban, B. Yu, M. J. Wainwright, and P. K. Ravikumar.
A unified framework for high-dimensional analysis of *m*-estimators with decomposable regularizers.
In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

E. Yang, A. Lozano, and P. Ravikumar.
Elementary estimators for high-dimensional linear regression.
In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 388–396, 2014a.

E. Yang, A. C. Lozano, and P. K. Ravikumar.
Elementary estimators for graphical models.
In *Advances in Neural Information Processing Systems*, pages 2159–2167, 2014b.