Fundamentals of Mathematical and Computing Sciences:
Applied Mathematical Science

# Part III: Low rank matrix estimation (Lecture 1) From vector to matrix

2015-6-26

Taiji Suzuki (Room W707, post W8-46)
e-mail: suzuki.t.ct@m.titech.ac.jp

## Outline of the Lecture

This course introduces several basic concepts of mathematical optimization, probability and statistics, and is intended to provide key knowledge necessary for advanced study in Mathematical and Computing Sciences.

## Outline of this part (3rd part)

This part gives basic knowledges of low rank matrix estimation problems. Low rank matrix estimation has various applications such as computer vision, recommendation system, and reduced rank regression. In the series of lectures, problem formulation, methodologies, computational method and statistical properties are shown.

Lecture plan:

1. *From vector to matrix: Introduction to sparse estimation and low rank matrix estimation.
2. Estimation method: Statistical methodologies for estimating low rank matrix.
3. Computational method: Optimization method and sampling method.
4. Statistical property: Estimation accuracy, measure concentration of matrix valued random variables.
5. *Advanced topics.

(* indicates that the topic will be covered only by the support documents).
Evaluation: report.

## References

1. Tropp, J. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12, 389–434.
2. Rohde, A., and Tsybakov, A. B. (2011). Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39, 887–930.
3. Negahban, S., and Wainwright, M. J. (2012). Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13, 1665–1697.
4. Babacan, S. D., Luessi, M., Molina, R., and Katsaggelos, A. K. (2012). Sparse bayesian methods for low-rank matrix. IEEE Transactions on Signal Processing, 60, 964–3977.

# 1 Linear regression

Before we are going into the low rank matrix estimation, we briefly review the vector estimation problem.

Given fixed covariates $X = \begin{bmatrix} x_1^\top \\ x_2^\top \\ \vdots \\ x_n^\top \end{bmatrix} \in \mathbb{R}^{n \times p}$, we observe

$$Y = X\beta^* + \epsilon, \qquad \text{(regression)}$$

where $Y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$ (dependent variable, response) and $\epsilon = [\epsilon_1, \ldots, \epsilon_n]^\top \in \mathbb{R}^n$ (noise). We assume that $\{\epsilon_i\}_{i=1}^n$ is i.i.d. random variable with mean 0 and variance $\sigma^2$ ($\mathrm{E}[\epsilon_i] = 0, \mathrm{E}[\epsilon_i^2] = \sigma^2$). We observe $\{(x_i, y_i)\}_{i=1}^n$, and want to estimate $\beta^*$ (or $X\beta^*$) from the observed data.

There are many methods to estimate $\beta^*$, for example

- Least squares estimator.
- Ridge regression.
- Stein's shrinkage estimator.
- Lasso.

# 2 Least squares estimator

## 2.1 Definition of least squares estimator

$$\hat{\beta}_{\mathrm{LS}} := \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \|Y - X\beta\|^2 = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n (y_i - x_i^\top \beta)^2.$$

For simplicity, we assume that $X^\top X \succ O$. Then $\hat{\beta}_{\mathrm{LS}}$ can be expressed as

$$\hat{\beta}_{\mathrm{LS}} = (X^\top X)^{-1} X^\top Y.$$

($\because$) $\hat{\beta}_{\mathrm{LS}}$ satisfies

$$\begin{aligned} &\nabla_\beta \|Y - X\beta\|^2|_{\beta = \hat{\beta}_{\mathrm{LS}}} = 0 \\ \Leftrightarrow &X^\top (X\hat{\beta}_{\mathrm{LS}} - Y) = 0 \\ \Leftrightarrow &\hat{\beta}_{\mathrm{LS}} = (X^\top X)^{-1} X^\top Y. \end{aligned}$$

## 2.2 Statistical properties of least squares estimator

- $\hat{\beta}_{\mathrm{LS}}$ is an unbiased estimator:

$$\mathrm{E}_{Y|X}[\hat{\beta}_{\mathrm{LS}}] = \beta^*.$$

($\because$) $\mathrm{E}_{Y|X}[\hat{\beta}_{\mathrm{LS}}] = \mathrm{E}_{Y|X}[(X^\top X)^{-1} X^\top Y] = (X^\top X)^{-1} X^\top X \beta^* = \beta^*$.

- Variance (variance and covariance matrix) of $\hat{\beta}_{\mathrm{LS}}$ is given by

$$\mathrm{Var}(\hat{\beta}_{\mathrm{LS}}) = \mathrm{E}_{Y|X}[(\hat{\beta}_{\mathrm{LS}} - \beta^*)(\hat{\beta}_{\mathrm{LS}} - \beta^*)^\top] = (X^\top X)^{-1}\sigma^2.$$

$$\begin{aligned}(\because)\quad & \mathrm{E}_{Y|X}[(\hat{\beta}_{\mathrm{LS}} - \beta^*)(\hat{\beta}_{\mathrm{LS}} - \beta^*)^\top] \\ =&\mathrm{E}_{Y|X}[\{(X^\top X)^{-1}X^\top(X\beta^* + \epsilon) - \beta^*\}\{(X^\top X)^{-1}X^\top(X\beta^* + \epsilon) - \beta^*\}^\top] \\ =&\mathrm{E}_{Y|X}[\{(X^\top X)^{-1}X^\top\epsilon\}\{(X^\top X)^{-1}X^\top\epsilon\}^\top] = (X^\top X)^{-1}X^\top X(X^\top X)^{-1}\sigma^2 \\ =&(X^\top X)^{-1}\sigma^2.\end{aligned}$$

This is minimum variance among all unbiased estimator (discussed in the following).

## 2.3 Least squares estimator as an maximum likelihood estimator

Here assume that $\epsilon_i$ is generated from Gaussian distribution ($N(0, \sigma^2)$). Remind that the probability density function of $y_i$ for $\beta^* = \beta$ is given by

$$p(y_i|\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - x_i^\top\beta)^2}{2\sigma^2}\right),$$

because of the normality of the noise. Thus the log-likelihood of $\beta$ is given by

$$\log\prod_{i=1}^n p(y_i|\beta) = -\sum_{i=1}^n \frac{(y_i - x_i^\top\beta)^2}{2\sigma^2} - n\log(\sqrt{2\pi\sigma^2}).$$

Therefore, by maximizing the log-likelihood, we obtain the maximum likelihood estimator $\hat{\beta}_{\mathrm{MLE}} = (X^\top X)^{-1}X^\top Y$. One can observe that

$$\hat{\beta}_{\mathrm{LS}} = \hat{\beta}_{\mathrm{MLE}}.$$

**Theorem 1** (Cramer-Rao's Inequality)**.** *For all unbiased estimator $\hat{\beta}$, we have*

$$\mathrm{Var}(\hat{\beta}) = \mathrm{E}_{Y|X}[(\hat{\beta} - \beta^*)(\hat{\beta} - \beta^*)^\top] \succeq \mathrm{E}_{Y|X}[\nabla_\beta \log p(Y|\beta)\nabla_\beta^\top \log p(Y|\beta)]^{-1}|_{\beta=\beta^*} \qquad (1)$$

Here, the right hand side of Eq. (1) is the inverse of Fisher information matrix.
Notice that

$$\begin{aligned}\mathrm{E}_{Y|X}[\nabla_\beta \log p(Y|\beta)\nabla_\beta^\top \log p(Y|\beta)]|_{\beta=\beta^*} &= \mathrm{E}_{Y|X}\left[\frac{X^\top(X\beta^* - Y)}{\sigma^2}\frac{(X\beta^* - Y)^\top X}{\sigma^2}\right] \\ &= \mathrm{E}_{Y|X}\left[\frac{X^\top\epsilon\epsilon^\top X}{\sigma^4}\right] = X^\top X\sigma^{-2}.\end{aligned}$$

Thus, $\mathrm{Var}(\hat{\beta}) \succeq (X^\top X)^{-1}\sigma^2$ holds for all unbiased estimator $\hat{\beta}$. As we have see, $\mathrm{Var}(\hat{\beta}_{\mathrm{LS}}) = (X^\top X)^{-1}\sigma^2$. Therefore, it holds that

$$\mathrm{Var}(\hat{\beta}) \succeq \mathrm{Var}(\hat{\beta}_{\mathrm{LS}}) \ \text{ for all unbiased estimator } \hat{\beta}\ .$$

In that sense, the least squares estimator is called Best Unbiased Estimator (BUE).

## 2.4 Mean Squared Error (MSE) of the least squares estimator

**Question**: How accurate is the LS estimator?

MSE is defined as

$$\mathrm{MSE} = \mathrm{E}_{Y|X}[\|\hat{\beta}_{\mathrm{LS}} - \beta^*\|^2].$$

MSE can be evaluated as

$$\mathrm{E}_{Y|X}[\|\hat{\beta}_{\mathrm{LS}} - \beta^*\|^2] = \sigma^2 \mathrm{Tr}[(X^\top X)^{-1}],$$

because

$$\mathrm{E}_{Y|X}[\|\hat{\beta}_{\mathrm{LS}} - \beta^*\|^2] = \mathrm{E}_{Y|X}\{\mathrm{Tr}[(\hat{\beta}_{\mathrm{LS}} - \beta^*)(\hat{\beta}_{\mathrm{LS}} - \beta^*)^\top]\} = \sigma^2 \mathrm{Tr}[(X^\top X)^{-1}].$$

Now, we evaluate how MSE is dependent on the dimension $p$. To do so, we assume that $x_i$ is i.i.d. random variable generated from a distribution that satisfies $\mathrm{E}_x[xx^\top] = S(\in \mathbb{R}^{p\times p})$. By the low of large numbers, we have that

$$\frac{X^\top X}{n} \to S \quad \text{(in probability)}.$$

This implies that

$$\mathrm{E}_{Y|X}[\|\hat{\beta}_{\mathrm{LS}} - \beta^*\|^2] = \frac{\sigma^2}{n}\mathrm{Tr}[(X^\top X/n)^{-1}] \to \frac{\sigma^2}{n}\mathrm{Tr}[S^{-1}] \quad \text{(in probability)},$$

by the continuity of the inverse operation of a matrix (Slutsky's lemma).

If $S \succeq \lambda_{\min} I_p \succ O$ for some $\lambda_{\min} > 0$, then

$$\frac{\sigma^2}{n}\mathrm{Tr}[S^{-1}] \leq \frac{\sigma^2}{n}\mathrm{Tr}[(\lambda_{\min} I_p)^{-1}] \leq \frac{p}{n}\frac{\sigma^2}{\lambda_{\min}}.$$

This is linear to $p$ (the dimension of the parameter).

Predictive accuracy is also an important performance measure. That (more precisely the in-sample predictive accuracy) is defined as

$$\text{Predictive accuracy} = \mathrm{E}_{Y|X}\left[\frac{1}{n}\sum_{i=1}^n (x_i^\top \beta^* - x_i^\top \hat{\beta}_{\mathrm{LS}})\right].$$

(Check that $\frac{1}{n}\sum_{i=1}^n (x_i^\top \beta^* - x_i^\top \hat{\beta}_{\mathrm{LS}})$ is equivalent to $\mathrm{E}_{\tilde{Y}|X}[\frac{1}{n}\|\tilde{Y} - X\hat{\beta}_{\mathrm{LS}}\|^2]$ up to constant where $\tilde{Y}$ is an independent copy of $Y$). The predictive accuracy is evaluated as

$$\begin{aligned}\mathrm{E}_{Y|X}\left[\frac{1}{n}\sum_{i=1}^n (x_i^\top \beta^* - x_i^\top \hat{\beta}_{\mathrm{LS}})\right] &= \mathrm{E}_{Y|X}\left[\frac{1}{n}\|X\beta^* - X\hat{\beta}_{\mathrm{LS}}\|^2\right] \\ &= \frac{1}{n}\mathrm{Tr}[X\mathrm{Var}(\hat{\beta}_{\mathrm{LS}})X^\top] = \frac{\sigma^2}{n}\mathrm{Tr}[I_p] = \sigma^2\frac{p}{n}.\end{aligned}$$

The predictive accuracy is also linear to $p$. Therefore, if $p$ is large compared to $n$, we don't have favorable estimation accuracy.

Question: What happens if $\beta^*$ is **sparse**? Can we improve the accuracy?
$\Rightarrow$ Yes. Model selection.

# 3 Model Selection: AIC

**AIC (Akaike's Information Criterion)** invented by **Hirotugu Akaike** is a criterion to minimize the predictive accuracy. AIC is originally developed to specify the order of AR model. It can be applied to not only linear regression but also other statistical models.

Suppose that the number of non-zero component of $\beta^*$ is small (the explanatory variable contains a lot of redundant information). We want to estimate the index set of the non-zero components ($J := \{j | |\beta_j^*| \neq 0\}$).

**Note**: Just choosing the index set that minimizes the empirical risk is not a good idea. $\rightarrow$ **Overfitting**.

Let $\hat{\beta}_{\hat{J}}$ be the least squares estimator on the submodel $\hat{J}$:

$$\hat{\beta}_{\hat{J}} := \underset{\beta \in \mathbb{R}^p : \beta_{\hat{J}^c} = \mathbf{0}}{\arg\min} \|Y - X\beta\|^2.$$

Ideally if we know the true non-zero components, i.e. $\hat{J} = J$, then

$$\text{predictive accuracy of } \hat{\beta}_{\hat{J}} = \sigma^2 \frac{|J|}{n} \ll \sigma^2 \frac{p}{n},$$

under a sparse setting $|J| \ll p$. However, in practice, we don't know $J$. Thus we need to estimate that.

$$\mathrm{AIC}(\hat{J}) = \frac{1}{\sigma^2} \|Y - X\hat{\beta}_{\hat{J}}\|^2 + 2|\hat{J}|.$$

Choose $\hat{J} \subseteq \{1, \ldots, n\}$ that minimizes AIC.

AIC is an unbiased estimator of the predictive error up to constant (if $\hat{J}$ includes $J$). Minimizing AIC leads to a good predictive accuracy (indeed it is minimax optimal).

*Proof.* (Rough proof) Suppose that $\hat{J}$ includes $J$ and let $X_{\hat{J}} = (X_{i,j})_{i=1,;n;j \in \hat{J}}$ (submatrix of $X$ with column indices $\hat{J}$), then we have

$$\begin{aligned}
&\mathrm{E}_{Y|X}\left[\frac{1}{n}\|X\hat{\beta}_{\hat{J}} - X\beta^*\|^2\right] \\
=&\mathrm{E}_{Y|X}\left[\frac{1}{n}\|X\hat{\beta}_{\hat{J}} - Y - \epsilon\|^2\right] \\
=&\mathrm{E}_{Y|X}\left[\frac{1}{n}\|X\hat{\beta}_{\hat{J}} - Y\|^2 - \frac{2}{n}\langle X\hat{\beta}_{\hat{J}} - Y, \epsilon\rangle + \frac{1}{n}\|\epsilon\|^2\right].
\end{aligned}$$

Now, observe that

$$\begin{aligned}
&\mathrm{E}_{Y|X}\left[\langle X\hat{\beta}_{\hat{J}} - Y, \epsilon\rangle\right] \\
=&\mathrm{E}_{Y|X}\left[\langle X_{\hat{J}}(X_{\hat{J}}^\top X_{\hat{J}})^{-1} X_{\hat{J}}^\top Y - X\beta^* - \epsilon, \epsilon\rangle\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}_{Y|X}\left[\langle X_{\hat{J}}(X_{\hat{J}}^\top X_{\hat{J}})^{-1}X_{\hat{J}}^\top(X_{\hat{J}}\beta^*_{\hat{J}}+\epsilon) - X_{\hat{J}}\beta^*_{\hat{J}} - \epsilon, \epsilon\rangle\right] \quad (\because J \subseteq \hat{J}) \\
&= \mathrm{E}_{Y|X}\left[\|X_{\hat{J}}(X_{\hat{J}}^\top X_{\hat{J}})^{-1}X_{\hat{J}}^\top\epsilon\|^2\right] - n\sigma^2 = |\hat{J}|\sigma^2 - n\sigma^2.
\end{aligned}
$$

Then, we have that

$$
\begin{aligned}
\mathrm{E}_{Y|X}\left[\frac{1}{n}\|X\hat{\beta}_{\hat{J}} - X\beta^*\|^2\right] &= \mathrm{E}_{Y|X}\left[\frac{1}{n}\|X\hat{\beta}_{\hat{J}} - Y\|^2 + \frac{2|\hat{J}|\sigma^2}{n}\right] - \sigma^2 \\
&= \mathrm{E}_{Y|X}\left[\frac{\sigma^2}{n}\mathrm{AIC}(\hat{J})\right] - \sigma^2.
\end{aligned}
$$

□

Minimizing AIC is computationally much demanding ($\mathrm{O}(2^p)$).
⇒ NP-hard (submodular function maximization).
⇒ $L_1$-regularization (Lasso) [7]: Convex optimization, statistically nice properties[*1].

# 4 Estimation of low rank matrix: From vector to matrix

Model:

$$y_i = \langle X_i, A^*\rangle + \epsilon_i, \quad (i = 1, \ldots, n),$$

where $\langle X, A\rangle = \mathrm{Tr}[X^\top A]$, $X_i \in Real^{M\times N}$ is an explanatory variable, $A^* \in Real^{M\times N}$ is the true matrix (supposed to be low rank), and $\epsilon_i$ is i.i.d. noise.

**Basic idea:**

$$
\begin{aligned}
\min_{A\in\mathbb{R}^{M\times N}} \quad & \sum_{i=1}^{n}(y_i - \langle X_i, A\rangle)^2 \\
\text{s.t.} \quad & \mathrm{rank}(A) \le d.
\end{aligned}
$$

Analogous to AIC minimization. The cardinality of non-zero components is replaced by rank. Note that this is non-convex.

**Applications:**

- Computer vision
- Recommendation system [6] (NetFlix prize [3])
- Reduced rank regression [1, 4, 5]
- Multi-task learning [2]

# References


[1] T. Anderson. Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Mathematical Statistics*, 22:327–351, 1951.


[*1] Instead of $L_1$-regularization, a greedy method to minimize AIC is also useful. That corresponds to submodular function maximization, and there is a guarantee of the approximation error for the greedy method.

[2] A. Argyriou, C. A. Micchelli, M. Pontil, and Y. Ying. A spectral regularization framework for multi-task structure learning. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 25–32, Cambridge, MA, 2008. MIT Press.

[3] J. Bennett and S. Lanning. The netflix prize. In *Proceedings of KDD Cup and Workshop 2007*, 2007.

[4] G. R. Burket. *A study of reduced-rank models for multiple prediction*, volume 12 of *Psychometric monographs*. Psychometric Society, 1964.

[5] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, pages 248–264, 1975.

[6] N. Srebro, N. Alon, and T. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances in Neural Information Processing Systems (NIPS) 17*, 2005.

[7] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.