

Lecture 8: Basic estimation methods

Lecturer: Yiqiao Zhong

Last update: 17:45:07, 2024-10-07

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

We will start with (a review of) two classical estimation methods, method of moments and maximum likelihood estimation. We will introduce some modern statistical applications and implications.

Cultural difference: estimation vs. prediction. As we progress in this course, we will notice significant differences between stats and CS communities, in terms of the approaches for data analysis.

- Statistics: emphasis on modeling, assumption on data generating processes, interpretation, and assessment of uncertainty, heavy involvement of domain expertise
- Computer science: priority on prediction accuracy, okay with black-box models, automation and user interface.

An excellent article by Leo Brieman [2] expounds the differences more than two decades ago. *Q: Is there a diverging trend in recent year?*

1 Method of Moments

1.1 Classical recipe

Consider $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta \in \mathcal{P}$, where $\theta \in \Theta \subset \mathbb{R}^k$. Assume the k -th (population) moment exists, namely $\mathbb{E}[|X_1|^k] < \infty$. Denote the j -th moment of P and j -th empirical moment by, respectively,

$$\mu_j = \mathbb{E}[X_i^j], \quad \hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j.$$

Suppose that we can find certain (Borel) functions $h_1, \dots, h_k : \mathbb{R}^k \rightarrow \mathbb{R}$,

$$\mu_j = h_j(\theta), \quad j = 1, \dots, k.$$

Then, solve for $\hat{\theta}$ in the equations with empirical substitutes of $(\mu_j)_{j \leq k}$,

$$\hat{\mu}_j = h_j(\hat{\theta}), \quad j = 1, \dots, k.$$

If $h = (h_1, \dots, h_k) : \mathbb{R}^k \rightarrow \mathbb{R}^k$ has inverse h^{-1} , then

$$\hat{\theta} = h^{-1}(\hat{\mu}).$$

Note: moments estimators may not exist or may not be unique, depending on h . We will soon learn in Lecture 9 that MoM estimators are generally consistent, i.e., when $n \rightarrow \infty$, the estimator $\hat{\theta}$ will be very close to θ .

Example 1 (Example 3.24 of [6]). Suppose that $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta$ have finite second moments, where the parameters $\theta = (\mu, \sigma^2)$ contain mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$. We do not make additional assumption on P_θ such as assuming a specific distribution family. We have

$$\mu_1 = \mathbb{E}[X_i] = \mu, \quad \mu_2 = \mathbb{E}[X_i^2] = \mu^2 + \sigma^2.$$

Replacing population moments with empirical moments $\hat{\mu}_1, \hat{\mu}_2$, we solve for $\hat{\mu}$ and $\hat{\sigma}^2$

$$\begin{cases} \hat{\mu}_1 = \hat{\mu} \\ \hat{\mu}_2 = \hat{\mu}^2 + \hat{\sigma}^2. \end{cases}$$

Often we denote

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i, \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

which are called sample mean and sample variance respectively. We obtain

$$\hat{\theta} = (\bar{X}, \underbrace{\frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2}_{\text{empirical variance}}) = (\bar{X}, \underbrace{\frac{n-1}{n} S^2}).$$

Remark 1. This example is simple yet already informative.

- Notice that our derivation only requires finite second moments; we do not assume specific family of distributions.
- If P_θ is the normal distribution, then $\hat{\theta}$ is the sufficient statistics. Since $\mathbb{E}[S^2] = \sigma^2$, we know that $\hat{\theta}$ is a (slightly) biased estimator, and actually has slightly smaller MSE (*exercise*). This is another example of bias—variance tradeoff.
- If P_θ is the double exponential or logistic distribution, then $\hat{\theta}$ is not sufficient and the estimator is suboptimal in terms of MSE.

Example 2. (Example 3.25 of [6]) Let $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} \text{Unif}([\theta_1, \theta_2])$, where $\theta_1 < \theta_2$ are both unknown.

$$\mu_1 = \mathbb{E}[X_i] = \frac{\theta_1 + \theta_2}{2}, \quad \mu_2 = \mathbb{E}[X_i^2] = \frac{\theta_1^2 + \theta_2^2 + \theta_1 \theta_2}{3}.$$

Replace population moments by the empirical version, and solve for θ_1, θ_2

$$\begin{cases} \hat{\mu}_1 = (\hat{\theta}_1 + \hat{\theta}_2)/2 \\ \hat{\mu}_2 = \frac{\hat{\theta}_1^2 + \hat{\theta}_2^2 + \hat{\theta}_1 \hat{\theta}_2}{3}, \end{cases}$$

which gives

$$\begin{cases} \hat{\theta}_1 = \bar{X} - \sqrt{\frac{3(n-1)}{n} S^2} \\ \hat{\theta}_2 = \bar{X} + \sqrt{\frac{3(n-1)}{n} S^2}. \end{cases}$$

Note that $\hat{\theta}$ is not based on sufficient statistics $(X_{(1)}, X_{(n)})$. In terms of MSE, $\hat{\theta}$ is much worse than the unbiased estimator constructed from $(X_{(1)}, X_{(n)})$ (Homework assignment).

Remark 2. For simple problems we can derive h by hand, but in more complicated problems, we often resort to existing numerical computation or devising new algorithms.

1.2 Generalized Method of Moments (GMM)

We start with generalization of moments. Suppose $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_\theta$, $X_i \in \mathbb{R}^p$, $\theta \in \Theta$, where Θ is compact for simplicity. Let $g : \mathbb{R}^{p+1} \rightarrow \mathbb{R}^m$ be a continuous function, such that

$$\mathbb{E}[g(X_i, \theta)] = 0.$$

Note that m is allowed to be larger than p . This is different from MoM and is quite useful because we can collect many generalized moments.

Example of g : We may consider different functions of data $g = (g_1, \dots, g_m)$, where

- $g_k(x, \theta) = x^k - \mu_k(\theta)$, where $\mu_k(\theta)$ is the k -th moment of X_i .
- $g_k(x, \theta) = \cos(kx) - V_k(\theta)$.
- $g_k(x, \theta) = \text{LeakyReLU}(kx) - \xi_k(\theta)$
- ...

Let $W \in \mathbb{R}^{m \times m}$ be a positive definite (weighting) matrix. Define generalized method of moments (GMM) estimator as

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right)^T W \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right).$$

We also write $\hat{\theta}(W, X)$ or simply $\hat{\theta}(W)$ to emphasize the dependence on W .

Example 3. GMM includes MoM as a special case. Let $W = I_m$, $m = k$, $g_k(x, \theta) = x^k - \mu_k(\theta)$. Then

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right\|^2 = \{\theta \in \Theta : \hat{\mu}_k - \mu_k(\theta) = 0\},$$

if the set on the RHS is nonempty. Thus, we recovered the regular method of moments estimator.

Remark 3. It is advantageous to choose $m > p$, so we combine many more moments information in $g(X_i, \theta)$ using a weighting matrix W . We hope to pick W so that useful moments receive a large weight instead of equal weights (i.e. $W = I_m$).

Later we will revisit GMM and understand that GMM not only inherits consistency property of MoM under mild conditions, but also is more efficient by selecting a good weighting matrix W to combine generalized moments in g .

2 Examples

Moments are simple yet useful statistics. In modern applications, the data dimension can be prohibitively large and fast computation is crucial. Low-order moments (1st, 2nd, 3rd) together with, for instance, spectral decomposition reveals structure behind data.

2.1 Second moments and spectral methods

Factor model. As an example, recall the factor model (under different assumptions): we have n i.i.d. observations $Y_1, \dots, Y_n \in \mathbb{R}^p$, each following the structure

$$Y_i = Bf_i + u_i, \quad \text{where } \mathbb{E}[f_i] = 0, \text{ Cov}(f_i) = I_r, \mathbb{E}[u_i] = 0, \text{ Cov}(u_i) = \Sigma$$

where the unknown parameters are the matrix $B \in \mathbb{R}^{p \times K}$. We know that $\text{Cov}(Y_i) = BB^\top + \Sigma$. Since there is potential identifiability issue of estimating B , we may want to estimate $L = BB^\top$ instead. A natural way is to use the covariance matrix—which is the second moments of observations—to estimate L :

$$\hat{L}^0 = \frac{1}{n} \sum_{i=1}^n Y_i Y_i^\top - \Sigma$$

which satisfies $\mathbb{E}[\hat{L}^0] = L$. One structural property we have not used in the rank constraint: we must have $\text{rank}(L) \leq K$. We can use eigendecomposition (aka spectral decomposition) to improve $\hat{L}^0$. Let the spectral decomposition of $\hat{L}^0$ be $\hat{L}^0 = U \Lambda U^\top$ where $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_p\}$ contains eigenvalues in descending order and $U = [u_1, \dots, u_p]$. Let

$$\hat{L} = \sum_{j \leq K} \lambda_j u_j u_j^\top$$

be the new estimator, which satisfies $\text{rank}(\hat{L}) \leq K$ by definition. This estimator is generally better (less noise).

Other examples.

- In natural language processing (NLP), a classical technique of representing words/tokens is **word embedding**. For example, a method named *GloVe* [5] calculates second-moments count data, known as co-occurrence matrix X . Each entry X_{ij} tallies the number of times word j occurs in the context of word i . Then an optimization problem is used to extract the low-dimensional embeddings of words.

2.2 Latent variable model and tensor decomposition

Topic models. Latent variables are commonly used in structured modeling [1]. As an example of topic models, assume that there are K topics in a corpus (collection of text documents). Imagine that for each document, a latent (unobserved) random variable h representing the topic is drawn from a multinomial distribution:

$$P(h = j) = w_j, \quad j = 1, 2, \dots, K$$

and then a word is drawn from a distribution over the vocabulary (finite discrete set) of size d . Denote by $\mu_j \in \mathbb{R}^d$ the pmf in vector form of the distribution if the topic is j . Suppose that we represent each observed word as an vector in the canonical basis

$$x_t = e_i \quad \text{if and only if the } t\text{-th word in the document is } i$$

Let x_1, x_2, x_3 be word vectors in the same document. We calculate the population moments and obtain

$$\begin{aligned} \mathbb{E}[x_1 \otimes x_2] &= \sum_{k=1}^K w_k \mu_k \otimes \mu_k \\ \mathbb{E}[x_1 \otimes x_2 \otimes x_3] &= \sum_{k=1}^K w_k \mu_k \otimes \mu_k \otimes \mu_k \end{aligned}$$

Here the first quantity denoted by $M_2 \in \mathbb{R}^{d \times d}$ is a matrix, also known as a 2nd-order tensor; and the second quantity denoted by $M_3 \in \mathbb{R}^{d \times d \times d}$ is 3rd-order tensor. The notation $\otimes$ is tensor product: $[a \otimes b \otimes c]_{ijk} = a_i b_j c_k$.

Our goal is to estimate $(w_k)_{k=1}^K$ and $(\mu_k)_{k=1}^K$ using MoM. The crucial part is to find a mapping from (M_2, M_3) to $(w_k, \mu_k)_{k=1}^K$ with efficient computation. Once the mapping is found, we just replace the population moments with empirical moments calculated from the data.

Note that the case $K = 1$ is obvious since we must have $w_1 = 1$ and it is similar to the previous factor model.

Orthogonal tensor decomposition. It turns out that if $K \leq d$, then there is a nice algorithm to find the unknown parameters. An outline of the algorithm is the following: first we use M_2 to determine a linear transformation $\bar{\mu}_k = B\mu_k \in \mathbb{R}^K$ such that $\bar{\mu}_1, \dots, \bar{\mu}_K$ are orthonormal (**exercise**: how do we find such B ?) This step is known as **whitening**. We use B to transform tensor M_3 and obtain a new tensor with the form

$$\bar{M}_3 = \sum_{i=k}^K \lambda_k \bar{\mu}_k \otimes \bar{\mu}_k \otimes \bar{\mu}_k$$

for certain $(\lambda_k)_{k=1}^K$. Now our goal is to find a mapping from $(M_2, \bar{M}_3)$ to $(\lambda_k)_{k=1}^K$ and $(\bar{\mu}_k)_{k=1}^K$ since we can use B to get the original parameters in the end. To this end, we start from a random initial unit vector $u \in \mathbb{S}^{d-1}$ apply an iterative algorithm called the **power method** or **power iteration**—which is exactly the basic idea behind numerical computation of spectral decomposition.

3 Maximum likelihood estimation

Maximum likelihood estimation (MLE) is introduced in most undergraduate-level courses and will be discussed in detail in STAT 710. Here we briefly recap MLE. Suppose that we have observations Y as random variables drawn from a distribution from a parametric family (P_θ) where $\theta \in \Theta$ in the unknown parameter(s). Let the (joint) pdf or pmf of Y be f_θ .

At each value $Y = y$, the likelihood function is a function of θ namely $L(\theta; y) = f_\theta(y)$. The MLE $\hat{\theta}(Y)$ or simply $\hat{\theta}$ is defined by any

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta; y).$$

Finding $\hat{\theta}$ basically boils down to an maximization problem. It is often easier to work with log-likelihood function $\ell(\theta; y) = \log L(\theta; y)$. For non-constrained problems ($\Theta = \mathbb{R}^d$), we simply look for candidate solutions using the **first-order condition**

$$\frac{\partial \ell(\theta)}{\partial \theta} = 0.$$

For constrained problems (e.g., the boundary of Θ is determined by some equations), we may resort to KKT conditions.

Example 4. *Least square estimator is the maximum likelihood estimator in the normal distribution family. Suppose that the n observations $Y = (Y_1, \dots, Y_n)$ has a linear form*

$$Y_i = x_i^\top \theta + \varepsilon_i$$

where $X = [x_1, \dots, x_n]^\top \in \mathbb{R}^{n \times d}$ is fixed and known, and $\theta \in \mathbb{R}^d$ is the unknown parameter, and $\varepsilon_i \stackrel{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ where σ^2 is known. Equivalently, we can write $Y \sim N(X\theta, \sigma^2 I_n)$. The log-likelihood function is

$$\ell(\theta; Y) = C - \frac{1}{2} \sum_{i=1}^n (Y_i - x_i^\top \theta)^2 \tag{1}$$

where C is a constant not dependent on θ . Thus maximizing $\ell(\theta; y)$ is the same as minimizing the squared loss. Thus, MLE gives exactly the least square estimator.

In classical statistics, MLE is widely used and studied. MLE enjoys good asymptotic properties such as consistency and asymptotic efficiency under mild conditions, which are the topics in the second part of “Mathematical Statistics”.

Here we mention a drawback of MLE compared with MoM in the modern setting.

3.1 MLE and computational complexity

Solving a maximization problem can be challenging in high dimensions.

Community detection. In statistical analysis of networks, we want to identify communities of n nodes (e.g., students) based on an observed network/graph, e.g., whether or not two nodes interact with each other.

As one simple model, we assume that there are n nodes in total consisting of two communities of equal number of nodes. For $1 \leq i < j \leq n$,

$$\begin{aligned} Y_{ij} &\stackrel{\text{i.i.d.}}{\sim} \text{Bern}(p^*) && \text{if node } i \text{ and } j \text{ are in the same community} \\ Y_{ij} &\stackrel{\text{i.i.d.}}{\sim} \text{Bern}(q^*) && \text{if node } i \text{ and } j \text{ are in the different communities} \end{aligned}$$

For simplicity we assume $0 < q^* \leq p^* \leq 1/2$. Let $z_i^* \in \{\pm 1\}$ representing the community membership: $z_i^* = 1$ if node i is in the first community and $z_i^* = -1$ if node i is in the second community. We can interpret our observation $(Y_{ij})_{i < j}$ as a graph: there are n nodes and there is an edge iff $Y_{ij} = 1$.

Let us formulate this problem using a statistical estimation framework. Our parameter space is

$$\Theta = \{(p, q, (z_i)_{i=1}^n) : 0 < q \leq p \leq 1/2, z_i \in \{\pm 1\}\};$$

an unknown $\theta^* \in \Theta$ is selected and used to generate the observation, which is a collection of binary-valued $(Y_{ij})_{i < j}$. Our goal is to estimate θ^* using MLE.

We can work out the likelihood function. For any $\theta \in \Theta$, we can write the likelihood function

$$\begin{aligned} L(\theta; Y) &= \prod_{z_i z_j = 1} p^{Y_{ij}} (1-p)^{(1-Y_{ij})} \prod_{z_i z_j = -1} q^{Y_{ij}} (1-q)^{(1-Y_{ij})} \\ &= (1-p)^{\frac{n}{2}(\frac{n}{2}-1)} q^{\frac{n}{2}(\frac{n}{2}-1)} \left(\frac{p}{1-p}\right)^{\sum_{z_i z_j = 1} Y_{ij}} \left(\frac{q}{1-q}\right)^{\sum_{z_i z_j = -1} Y_{ij}} \end{aligned}$$

where the product and sum are over all pairs (i, j) such that $i < j$. Denote $C_n(p, q) = \frac{n}{2}(\frac{n}{2}-1) \log((1-p)q)$, we have

$$\begin{aligned} \log L(\theta; Y) &= C_n(p, q) + \log\left(\frac{p}{1-p}\right) \sum \mathbf{1}\{z_i z_j = 1\} Y_{ij} + \log\left(\frac{q}{1-q}\right) \sum \mathbf{1}\{z_i z_j = -1\} Y_{ij} \\ &= C_n(p, q) + \log\left(\frac{q}{1-q}\right) \sum Y_{ij} + \log\left(\frac{p(1-q)}{(1-p)q}\right) \sum \mathbf{1}\{z_i z_j = 1\} Y_{ij} \end{aligned}$$

where we used the simple fact $\mathbf{1}\{z_i z_j = 1\} + \mathbf{1}\{z_i z_j = -1\} = 1$ to replace $\mathbf{1}\{z_i z_j = -1\}$. Notice that $(z_i)_{i=1}^n$ only shows up in the third term in the last expression. For convenience, we write $z = (z_1, \dots, z_n)$ in vector form. So, in order to maximize $\log L(\theta; Y)$ over θ , we can separate the maximization problem into two sub-maximization problems

$$\begin{aligned} \max_{\substack{0 < q \leq p \leq 1/2 \\ z \in \{\pm 1\}^n : z^\top \mathbf{1}_n = 0}} \log L(\theta; Y) &= \max_{0 < q \leq p \leq 1/2} \left[C_n(p, q) + \log\left(\frac{p}{1-p}\right) \sum Y_{ij} \right. \\ &\quad \left. + \max_{z \in \{\pm 1\}^n : z^\top \mathbf{1}_n = 0} \log\left(\frac{p(1-q)}{(1-p)q}\right) \sum \mathbf{1}\{z_i z_j = 1\} Y_{ij} \right] \end{aligned}$$

Below we will only focus on the inner maximization problem. Since $p \geq q$, we have $\frac{p(1-q)}{(1-p)q} \geq 1$ so the log factor is nonnegative.

$$\max_{z \in \{\pm 1\}^n : z^\top \mathbf{1}_n = 0} \sum_{i < j} \mathbf{1}\{z_i z_j = 1\} Y_{ij} \tag{2}$$

where the constraint $z^\top \mathbf{1}_n = 0$ ensures the equal-sized condition.

Max-cut problem and spectral relaxation. The maximization problem without the linear constraint in Eqn. 2 is the well-known max-cut problem. It looks for the cut of a graph that crosses the maximum number of edges. It is a known NP-hard problem in computational complexity theory; it is very difficult to find the maximum when n is large. Of course, hope is not lost since we assumed a statistical model. In fact, we can solve Eqn. 2 easily and the maximizer recovers z^* exactly with high probability¹ under some conditions.

For convenience, let $Y \in \mathbb{R}^{n \times n}$ be a symmetric matrix by letting $Y_{ij} = Y_{ji}$ and $Y_{ii} = 0$. Since $z_i z_j \in \{\pm 1\}$, we have $\mathbf{1}\{z_i z_j = 1\} = (1 + z_i z_j)/2$, so Eqn. 2 is equivalent to

$$\max_{z \in \{\pm 1\}^n : z^\top \mathbf{1}_n = 0} z^\top Y z.$$

The constraints, especially the discrete constraint, are annoying. So if we drop the constraint, we consider a **relaxed problem**:

$$\max_{z : \|z\|_2 = \sqrt{n}} z^\top Y z.$$

This is exactly the variational form of the top eigenvector (up to a rescaling). So we have a natural algorithm based on the spectral method:

1. Calculate the top eigenvector of Y , denoted by $\tilde{z}(Y)$;
2. Apply a discretization: set $\hat{z}_i = 1$ if $[\tilde{z}(Y)]_i > 0$ and $\hat{z}_i = -1$ if $[\tilde{z}(Y)]_i < 0$ for all $i = 1, \dots, n$.

Under some conditions, $\hat{z}(Y) = z^*$ holds with high probability.

3.2 Comparing MoM with MLE

Let us give a summary the pros and cons of MoM compared with MLE.

- Pros
 - (1) MoM is very flexible, easy to derive.
 - (2) MoM does not need to assume a parametric distribution family. With minimal knowledge of the distribution, consistency is easy to achieve.
- Cons
 - (1) MoM is often not statistical efficient, especially when P_θ is known, in which case MLE is better.
 - (2) High moments $\mathbb{E}[X_i^k]$ are unreliable, especially in the presence of heavy tail distributions.

To combine the best of the two worlds, sometimes it is possible to devise a hybrid method, or use the estimator from MoM as an initial value in an iterative algorithm for MLE.

Remark 4. *There is an interesting history of MoM v.s. MLE. In the study of measurement errors, early mathematicians and physicists including Laplace, Gaussian, Legendre established fundamental notations such as the least squares method, Gaussian distribution, central limit theorem, MLE. This historical work laid the foundations for estimation theory.*

4 Further reading

History about MoM and MLE can be found in many sources, such as Wikipage.

GMM. Lars Peter Hansen. A good reference includes [4], [3] and Wikipedia.

¹The success probability is very close to 1 when n is large.

Tensors. Let $A \in \mathbb{R}^{n_1 \times \dots \times n_k}$ be a k -th order tensor. Denote $[\mathbf{n}] = \prod_{j=1}^k \{1, 2, \dots, n_j\}$ and $\mathbf{j} = (j_1, \dots, j_k)$. We can represent A in terms of the basis

$$A = \sum_{\mathbf{j} \in [\mathbf{n}]} A_{\mathbf{j}} e_{j_1} \otimes \dots \otimes e_{j_k}$$

For matrices $V_1, \dots, V_k$ where $V_j \in \mathbb{R}^{n_j \times m_j}$, we may change the basis of A as

$$A[V_1, \dots, V_k] = \sum_{\mathbf{j} \in [\mathbf{n}]} \sum_{\mathbf{s} \in [\mathbf{m}]} A_{\mathbf{s}} [V_1]_{s_1, j_1} [V_2]_{s_2, j_2} \dots [V_k]_{s_k, j_k} e_{j_1} \otimes \dots \otimes e_{j_k}$$

In particular, in $k = 2$, we can use SVD to change the basis of A : if the SVD is given by $A = USV^\top$, then

$$(A[U, V])_{j_1, j_2} = \sum_{s_1, s_2} U_{s_1, j_1} A_{s_1, s_2} V_{s_2, j_2} = (U^\top AV)_{j_1, j_2} = S_{j_1, j_2} = \begin{cases} \sigma_{j_1} & \text{if } j_1 = j_2 \\ 0 & \text{if } j_1 \neq j_2 \end{cases}$$

where $S = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_{\min\{n_1, n_2\}}\}$.

The previous calculation shows that for $k = 2$, the 2nd-order tensor (matrix) can be viewed as a diagonal matrix under change of bases. So under the new bases, A as a linear map is simply rescaling. Further, if $n_1 = n_2 = n$, and we choose $\tilde{U} = US^{1/2}$ and $\tilde{V} = VS^{1/2}$. Then,

$$A[\tilde{U}, \tilde{V}] = \tilde{U}^\top A \tilde{V} = I_n.$$

Convex relaxation. A famous algorithm to **approximately** solve the max-cut problem is semidefinite programming (SDP) relaxation; link. In statistics, for nonconvex problems, a common practice is to solve a similar convex problem with less stringent constraints—which is known as convex relaxation.

References

- [1] Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, Matus Telgarsky, et al. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.
- [2] Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.
- [3] A.R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.
- [4] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.
- [5] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.
- [6] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.