Matrix Algebra and Optimization for Statistics and Machine Learning

Yiyuan She

Department of Statistics, Florida State University

- Some basic decompositions

- A motivation from regression: $\min_\beta \|y - X\beta\|_2^2$, where $\hat{\beta} = (X^TX)^{-1}X^Ty$, when $rank(X) = p$ (and so $n \geq p$)
- Ridge regression: $\min_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$. Here, $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$ (arbitrarily large p)
  - Bias-variance tradeoff: Sacrifice a little bit of bias to reduce the variance significantly
- Cost of matrix inversion: $\mathcal{O}(p^3)$!
- In computation, matrix decompositions are very useful

SVD

- Given any $X \in \mathbb{R}^{n \times p}$ and $rank(X) = r$, we have

$$X = UDV^T$$

  where $U \in \mathbb{R}^{n \times r}$: $U^T U = I$, $V \in \mathbb{R}^{p \times r}$: $V^T V$, and $D = \mathrm{diag}\{d_1, \ldots, d_r\}$ with $d_1 \geq \cdots \geq d_r > 0$.
- d_i: singular values. U, V: orthogonal matrices
- Assume $n \geq p$. We can also write $X = UDV^T$ with $U \in \mathbb{R}^{n \times p}$, $V \in \mathbb{R}^{p \times p}$ and $D = \mathrm{diag}\{d_1, \ldots, d_p\}$. Here, V is orthonormal $VV^T = V^T V = I$, $d_1 \geq \cdots d_p \geq 0$.

Some intuition

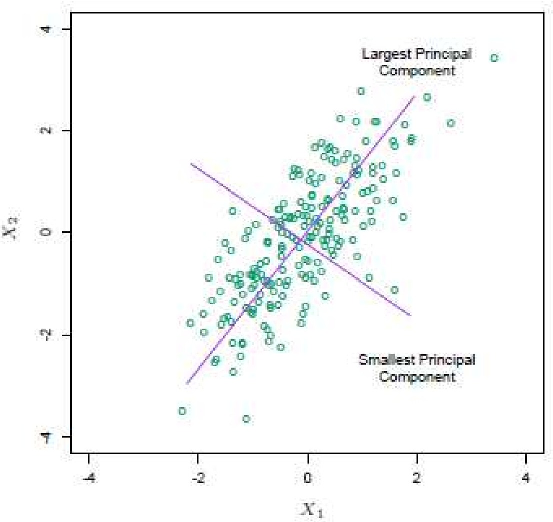
- Think of X as a linear transformation on a data vector $\alpha \in \mathbb{R}^p$: rotation, scaling (separable), rotation
- Think of X as a data matrix. With $U = [u_1, \ldots, u_r]$, $V = [v_1, \ldots, v_r]$ and $d_1 \geq \ldots, d_r > 0$, we have

$$X = UDV^T = \sum_{k=1}^{r} d_k u_k v_k^T = d_1 u_1 v_1^T + \cdots + d_r u_r v_r^T$$

  - Each component is of rank 1. u_k, v_k characterize the coordinate systems, and d_k represent the 'energy'
- A rank-k truncation leads to **principal component analysis**: $\min_B \|X - B\|_F^2$ s.t. $rank(B) \leq k$

- $\|X\|_F^2 = \sum x_{ij}^2$
- So $\|X\|_F^2 = trace(X^T X) = \sum d_i^2$ by SVD.
- Separation theorems for singular values of matrices and their applications in multivariate analysis, Rao, *Journal of Multivariate Analysis*, 362-377, 1979

- To visualize V, plot the n data points in $\mathbb{R}^p$ (in an r-dimensional subspace). (U? Plot p columns in $\mathbb{R}^n$)



Who cares?

- Low-rank matrix approximation, as shown earlier
- The **Procrustes** problem for data alignment (assuming $A \in \mathbb{R}^{n \times m}$, $B \in \mathbb{R}^{n \times q}$, $m \geq q$):

$$\min_{T \in \mathbb{R}^{m \times q}} \|A - BT^T\|_F \text{ s.t. } T^T T = I$$

- $\hat{T} = UV^T$ from the SVD: $A^T B = UDV^T$
  - It is perhaps worth mentioning that $\min_T \|AT - B\|_F$ s.t. $T^T T = I$ has **no** explicit-form solution

- A main usage of SVD is to simplify expressions (often due to the cancelations of $U^T U, V^T V$)
- Regression with X of full column rank:
  $\hat{\beta} = (X^T X)^{-1} X^T y = (V D^{-1} U^T) y$ (pseudoinverse),
  $\hat{y} = X(X^T X)^{-1} X^T y = U U^T y$ (orthogonal projection)
- An interesting result is that the fit in ridge regression/OLS does not rely on V (**kernel** property)
  - In fact, by **SVD**, $X\hat{\beta} = (XX^T + \lambda I)^{-1} XX^T y$

Risk of ridge regression

- As another example, let's calculate the risk $\mathbb{E}[\|X\beta^* - X\hat{\beta}\|_2^2]$ of the ridge estimator $\hat{\beta} = (X^TX + \lambda I)^{-1}X^Ty$ under $y = X\beta^* + \epsilon$ (arbitrary p)
- First, it is easy to obtain Risk $= \|X\beta^* - \mathbb{E}[X\hat{\beta}]\|_2^2$ $+ Var(X\hat{\beta}) = \|X\beta^* - X(X^TX + \lambda I)^{-1}X^TX\beta^*]\|_2^2 +$ $Tr\{X(X^TX + \lambda I)^{-1}X^TX(X^TX + \lambda I)^{-1}X^T\}\sigma^2$
- To simplify this, we apply the skinny (compact) SVD with $V = [v_1, \cdots, v_r] \in \mathbb{R}^{p\times r}$, $r = rank(X)$.

Tradeoff

- We get the risk as $\|UD\{I-(D^2+\lambda I)^{-1}D^2\}V^T\beta^*\|_2^2$ $+\,Tr\{(D^2+\lambda I)^{-2}D^4\}\sigma^2$ or

$$\sum_{i=1}^{r} \frac{\lambda^2 d_i^2}{(d_i^2+\lambda)^2}(v_i^T\beta^*)^2 + \sum_{i=1}^{r} \frac{d_i^4}{(d_i^2+\lambda)^2}\sigma^2$$

- Bias^2 is an *increasing* function of λ, while Var is a *decreasing* function of λ.
- $\lambda \to 0+$ (limit, not $\lambda = 0$): $0 + r\sigma^2 = r\sigma^2$
- $\lambda \to +\infty$: $\sum d_i^2(v_i^T\beta^*)^2 + 0 = \|X\beta^*\|_2^2$ (and $\hat{\beta} = 0$)

Effective degrees of freedom

- Similarly, at a independent copy y' of y, we can obtain

$$\begin{aligned} \mathbb{E}[\|y' - X\hat{\beta}\|_2^2] &= \mathbb{E}[\|y - X\hat{\beta}\|_2^2] + 2\sigma^2 \sum_{i=1}^{r} \frac{d_i^2}{(d_i^2 + \lambda)} \\ &= \mathbb{E}[\text{Trn-Err} + 2\sigma^2 \text{DF}(\lambda)] \end{aligned}$$

- DF $= \sum_{i=1}^{r} \frac{d_i^2}{(d_i^2+\lambda)} = Tr\{X(X^TX + \lambda I)^{-1}X^T\}$
- If $\lambda = 0$, DF $= r$

- Our calculation applies to any p. Note that when X has full rank, $r = n$ or p.
- The optimal shrinkage amount is a function of the true signal and the noise level.
- This suggests the need of data-dependent tuning for ℓ_2

Pseudo-inverse

- Given any $X \in \mathbb{R}^{n \times p}$, let $X = UDV^T$ be its (compact) SVD where $D \in \mathbb{R}^{r \times r}$ with positive diagonal elements.
- The Pseudo-inverse or Moore-Penrose inverse of X is defined as $X^{+} = VD^{-1}U^T$ (which is **unique**)
  - $X = UDV^T$ (**any** SVD) $\Rightarrow X^{+} = VD^{+}U^T$
  - $X^{+} = \lim_{\epsilon \to 0}(X^TX + \epsilon I)X^T = \lim_{\epsilon \to 0} X^T(XX^T + \epsilon I)$
- When $p > n$, OLS does not have a unique solution but $(X^TX)^{+}X^Ty = X^{+}y$ is the one with minimum ℓ_2-norm
- Orthogonal projection on $\mathcal{R}(X)$: $X(X^TX)^{+}X^T$ (UU^T)

- The SVD definition of MP inverse is perhaps easier than the standard definition: $XX^+X = X$, $X^+XX^+ = X^+$, $(XX^+)^T = XX^+$, $(X^+X)^T = X^+X$
- The MP inverse has many properties (not shown here). The SVD perspective helps to derive all of them.
- MP inverse is just an example of *generalized* inverses

- There are other forms in addition to the skinny SVD.
- For example, when $n \geq p$ (WLOG), most software packages will deliver <u>orthonormal</u> U and V. (D?)
- Once the SVD is available, inverting (D) is "effortless"
- But calculating the SVD is expensive
  - $\mathcal{O}(np^2 + p^3)$
- We will introduce some related decompositions

Spectral decomposition

- If X is symmetric ($X^T = X$, and so $n = p$), the SVD becomes the spectral decomposition: $X = UDU^T$.
- D provides all eigenvalues of X. U: orthonormal.
- From the SVD $X = UDV^T \in \mathbb{R}^{n \times p}$, $XX^T = UD^2U$ and $X^TX = VD^2V^T$. [What does this imply?]
- A square matrix does not have to be symmetric to be **diagonalizable**

Diagonalizable

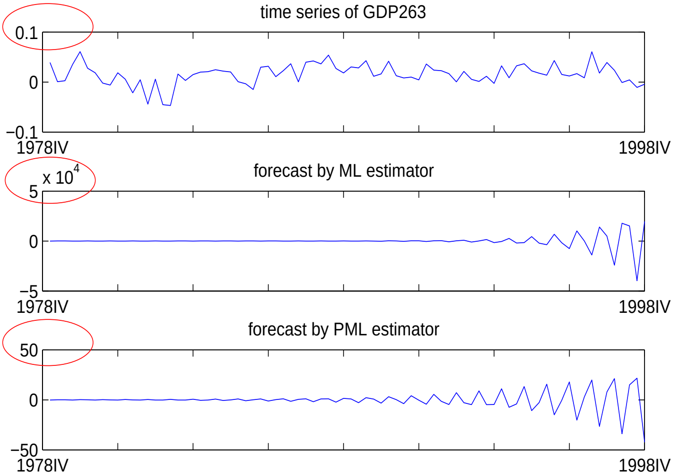
- Given $X \in \mathbb{R}^{n \times n}$, there exists a nonsingular matrix $A = [\alpha_1, \ldots, \alpha_n] \in \mathbb{R}^{n \times n}$ such that

$$X = ADA^{-1}$$

- A is not necessarily unitary, and X may not be symmetric.
- From $XA = AD$ or $X\alpha_i = d_i\alpha_i,\ 1 \leq i \leq n$, α_i are eigenvalues of X.

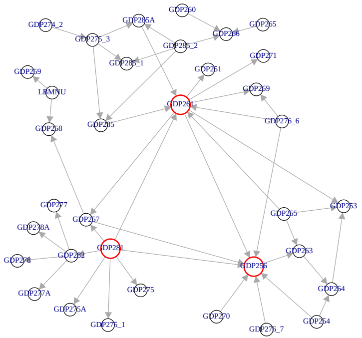
Largest singular value vs. largest eigenvalue

- Let $X \in \mathbb{R}^{n \times n}$. Its largest singular value d_1 defines the spectral norm $\|X\|_2$.
- Let $\lambda_1, \cdots, \lambda_n$ be its eigenvalues (real or complex). The spectral radius of X is defined as $\max_{1 \leq i \leq n} |\lambda_i|$
- Although they are equal in the symmetric case, in general they are not.
- $\rho(X) \leq \|X\|$ (and $\|X\|_2$ is not a bad bound)

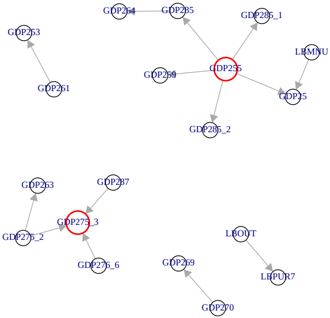

time series of GDP263
0.1
0
−0.1
1978IV
1998IV
forecast by ML estimator
x 10^4
5
0
−5
1978IV
1998IV
forecast by PML estimator
50
0
−50
1978IV
1998IV

- Assume multiple time series $x_t \in \mathbb{R}^p$, $0 \leq t \leq T$.
- The simplest model might be the first order vector autoregression (VAR): $x_t | x_{t-1} \sim \mathcal{N}(Ax_{t-1}, \sigma^2 I)$.
- For large p, a sparse A is desired (Granger causality)
- Stationarity of the system is guaranteed by $\rho(A) < 1$.
- With a **convex** relation, we can formulate a problem

$$\min_{B \in \mathbb{R}^{p \times p}} \sum_{t=1}^{T} \|x_t - Ax_{t-1}\|_2^2 + \lambda \|A\|_1 \text{ s.t. } \|A\|_2 \leq 1$$



(a) Pre-Great Moderation



(b) Post-Great Moderation

- Pre-Great Moderation
  - Gross private domestic investment indices (Serven, 1992)
- Post-Great Moderation
  - GDP255: real personal consumption expenditure. GDP275-3: energy goods price index (Jorgenson & Wilcoxen 1990, Halkos & Tzeremes, 2011)

- There are similar concerns in recurrent neural networks
- On the difficulty of training recurrent neural networks Pascanu, Mikolov & Bengio, *Proceedings of the 30th International Conference on Machine Learning*, 28(3): 1310-1318, 2013.

QR factorization

- Let's turn to another popular decomposition that is more efficient than SVD.
- Any $X \in \mathbb{R}^{n \times p}$ of full column rank can be factorized as

$$X = QR$$

- Here, $n \geq p$, $Q \in \mathbb{R}^{n \times p}$ is orthogonal, and R is upper triangular with nonzero diagonal elements.
- $Q\alpha$ and $Q^T\beta$ can be efficiently computed, and inverting R (or R^T) is easy [Why?]
- Regression: $\hat{\beta} = (R^T Q^T Q R)^{-1} R^T Q^T y = R^{-1} Q^T y$

- QR can be obtained efficiently
    - Householder reflections, Givens rotations, or Grand-Schmidt orthogonalization (modified)
- Cost: $2np^2 - (2/3)p^3$
- QR cannot exploit sparsity well

LU factorization

- Any $X \in \mathbb{R}^{n\times p}$ of full column rank can be factored as

$$X = PLU$$

- P: a permutation matrix, $L \in \mathbb{R}^{n\times p}$: unit lower triangular, $U \in \mathbb{R}^{p\times p}$: upper triangular & nonsingular
- [Sometimes, to exploit the sparsity of X, we permute its columns as well (full pivoting)—$X = PLUQ$]
- Computation: Gaussian elimination (with pivoting)
- Cost: $(2/3)p^3 + p^2(n-p)$ flops

Cholesky factorization

- Also known as a symmetric LU factorization.
- Suppose $\Sigma \in \mathbb{R}^{p \times p}$ is symmetric and positive definite
  - Let $\Sigma = X^T X$ and $X = QR$. What can you see?
- Σ can be factored as $\Sigma = LL^T (= R^T R)$ for some lower-triangular L with positive diagonal elements
- Cost: $(1/3)p^3$ flops
  - Again, for a **sparse** Σ, we can permute its rows and columns to save computational cost $(P\Sigma P^T = LL^T)$
- The Cholesky factor R is often a good substitute for the **square root** $X^{1/2} = UD^{1/2}U^T$ in computation

Example: multivariate Gaussian designs

- Model: $X = [\tilde{x}_1, \ldots, \tilde{x}_n]^T$ with $\tilde{x}_i$ i.i.d. $\sim \mathcal{N}(0, \Sigma)$. Equivalently, we can write $\mathrm{vec}(X) \sim \mathcal{N}(0, \Sigma \otimes I)$
- Generate Z with i.i.d. $\mathcal{N}(0, 1)$ elements. Construct $X = Z\Sigma^{1/2}$, or preferably $X = ZR$ where $R^T R = \Sigma$
- Cholesky factors can also be used to sphere the data
- Given the data matrix X, learning Σ is an important intriguing topic

Latent variable graphical model

- Assume $\tilde{x}_i \sim \mathcal{N}(0, \Sigma_{p\times p})$ and let $X = [\tilde{x}_1^T, \ldots, \tilde{x}_n^T]^T$.
- This gives a pretty challenging problem when $p^2 \gg n$. A popular assumption in Gaussian graph learning is that $\Sigma^{-1} =: \Omega$, the concentration matrix, is sparse.
  - Meaning: $\Omega_{j,j'} = 0 \Leftrightarrow j \perp j'$ **given** the rest variables
- From the loss $\log\det\Sigma + \langle \Sigma^{-1}, \hat{\Sigma}\rangle$ $(\hat{\Sigma} = \frac{X^T X}{n})$, we get

$$\min_{\Omega} -\log\det\Omega + \langle \Omega, \hat{\Sigma}\rangle + \lambda\|\Omega\|_1 \text{ s.t. } \Omega \succ 0 \quad \text{(convex!)}$$

- Conditional independence holds only in the presence of a small number of unobserved missing variables

- Assume $[X, Y] \sim \mathcal{N}(0, \tilde{\Sigma})$ and $\tilde{\Omega} = \tilde{\Sigma}^{-1}$ is sparse.
- Then what does $\Omega (= \Sigma^{-1})$ correspond to?
- Let $\tilde{\Omega} = \begin{bmatrix} \Omega_X & \Omega_{XY} \\ \Omega_{YX} & \Omega_Y \end{bmatrix}$. Then the *Schur complement* of Ω_Y gives

$$\Sigma^{-1} = \Omega_X - \Omega_{XY}\Omega_Y^{-1}\Omega_{YX}$$

- Ω_X is sparse, and positive definite!
- $\Omega_{YX}\Omega_Y^{-1}\Omega_{XY}$ is psd and has low rank (e.g., thin Y)

- Define $S = \Omega_X$, $L = \Omega_{YX}\Omega_Y^{-1}\Omega_{XY}$ so that $\Omega = S - L$.
- Latent variable graphical model (Chandrasekaran et al 12)

$$\begin{aligned} \min_{S,L} \; & \langle S - L, \hat{\Sigma} \rangle - \log\det(S - L) + \lambda \|S\|_1 + \lambda' tr(L) \\ \text{s.t.} \; & S - L \succ 0, L \succeq 0 \end{aligned}$$

- For any A, $tr(A) = \sum \sigma_i(A)$ and so enforces low rank

- $S - L \succ 0, L \succeq 0$ give *generalized* inequalities
- We get a **convex** optimization problem owing to the reparametrization and the ℓ_1-type regularization
- How to solve the problem?
  - SDP, proximal methods, ADMM, etc.
- Applications in machine learning and bioinformatics