Chapter 11: Graphical Models for High-Dimensional Data:

11.1 $S^{d\times d}$: Set of symmetric matrices

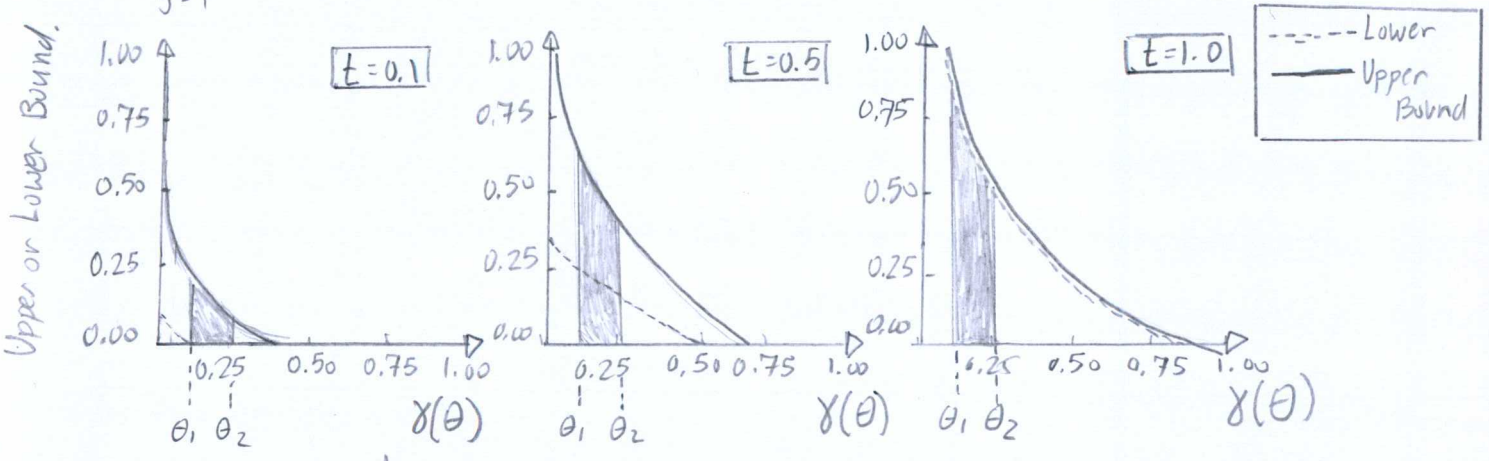
$S_+^{d\times d}$: Cone of symmetric and strictly positive matrices

$F: S^{d\times d} \to \mathbb{R}$ given by $F(\theta) = \begin{cases} -\sum_{j=1}^{d} \log \gamma_j(\theta) & \text{if } \theta \in S_+^{d\times d} \\ +\infty & \text{otherwise} \end{cases}$

where $\gamma_j(\theta) > 0$ are eigenvalues of θ

a) F is strictly convex in the $S_+^{d\times d}$ domain by Jensen's Inequality: $f(tX_1+(1-t)X_2) \le tF(X_1)+(1-t)f(X_2)$

$$-\sum_{j=1}^{d} \log \gamma(t\theta_1+(1-t)\theta_2) \le t\left(-\sum_{j=1}^{d} \log \gamma(\theta_1)\right) + (1-t)\left(-\sum_{j=1}^{d} \log \gamma(\theta_2)\right)$$



b) $\nabla F(\theta) = \nabla\left(-\sum_{j=1}^{d} \log \gamma_j(\theta)\right)$

$= -1/\theta$

$= -\theta^{-1}$

c) $\nabla F^2(\theta) = \nabla\left(-\sum_{j=1}^{d} \log \gamma_j(\theta)\right)^2$

$= 1/\theta^2$

$= \theta^{-1} \otimes \theta^{-1}$

11.2 $\hat{\theta}_{MLE} = \begin{cases} \Sigma^{-1} & \text{if } \hat{\Sigma} > 0 \\ \text{not defined} & \text{otherwise} \end{cases}$ where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} X_i X_j^T$

Gaussian: $f(X|0,\Sigma) = \frac{1}{\sqrt{2\pi\Sigma}} e^{\frac{-X^TX}{2\Sigma}}$

$$\frac{df(x|0,\Sigma)}{d\Sigma} = \frac{d}{d\Sigma}\left(\frac{1}{\sqrt{2\pi\Sigma}} e^{\frac{-X^TX}{2\Sigma}}\right)$$

$$= \frac{-1}{2}\left(\frac{1}{\sqrt{2\pi\Sigma}\,\Sigma}\right) e^{\frac{-X^TX}{2\Sigma}} + \frac{1}{2}\left(\frac{1}{\sqrt{2\pi\Sigma\,\Sigma^2}}\right) e^{\frac{-X^TX}{2\Sigma}}$$

$$= 0$$

$$\Sigma^* = \hat{\theta}_{MLE}$$

$$\hat{\theta}_{MLE} = \Sigma^{-1}$$

11.3 $X \in \mathbb{R}^d$, Σ^X, $Z_j = (X_j \mid X_{1 \le j \le 3})$

a) $\mathbb{E}[Z] = \mathbb{E}[X_j \mid X_{1 \le j \le 3}]$

$$= W_j + \theta \cdot X_{1 \le j \le 3} - X_j$$

$$= 0$$

$$X_j = \langle \theta, X_{1 \le j \le 3} \rangle + W_j$$

b) $\mathbb{E}\left[\left(X_{1 \le j \le 3} - \theta X_j\right) X_{1 \le j \le 3}^T\right] = \mathbb{E}\left[X_{1 \le j \le 3} \cdot X_{1 \le j \le 3}\right] - \theta\,\mathbb{E}\left[X_{1 \le j \le 3} \cdot X_j\right]$

$$= \Sigma_{1 \le j \le 3,\, 1 \le j \le 3} - \theta\,\Sigma_{1 \le j \le 3,\, j}$$

$$= 0$$

$$\theta = \frac{\Sigma_{1 \le j \le 3,\, j}}{\Sigma_{1 \le j \le 3,\, 1 \le j \le 3}}$$

c) The "just" in the problem relates a covariance matrix to zero-mean Gaussians with zero off-diagnol elements e.g. no variance between the x-, and y-directions.

The matrix is a "diagnolly dominant" covariance.

$$a_{ii} > \sum_{j \neq i} |a_{ij}| = a_{i1} + a_{i2} + \cdots + a_{ij}$$

$$\sum = \frac{\sum_{21} \sum_{12}}{\sum_{11} \cdot \sum_{11}} = I \qquad \sum^{1/2} = \left(\frac{\sum_{21} \sum_{12}}{\sum_{11} \sum_{11}} \right)^{1/2}$$

$$= \begin{bmatrix} I & A - \frac{A_{12}}{A_{11}} \\ -\frac{A_{21}}{A_{11}} & I \end{bmatrix} \qquad = \begin{bmatrix} I & 0 \\ -\frac{A_{21}}{A_{11}} & I \end{bmatrix} \begin{bmatrix} I & -\frac{A_{21}}{A_{11}} \\ 0 & I \end{bmatrix}$$

$$\theta^*_{jk} = \sum^{1/2} \cdot A \cdot \sum^{1/2}$$

$$= \begin{bmatrix} I & 0 \\ -\frac{A_{21}}{A_{11}} & I \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} I & -\frac{A_{12}}{A_{11}} \\ 0 & I \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} - A_{21} \cdot A_{11}^{-1} A_{12} \end{bmatrix}}$$

$$= \text{"Strictly Diagonal"}$$

11.4 $\hat{\Theta} = \underset{\Theta \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \{ \|\Theta\|_1 \}$ such that $\| \hat{\sum} \Theta - I_d \|_{max} \leq \lambda_n$

$T = \underset{T \in \mathbb{R}^{d \times d}}{\operatorname{argmin}} \{ \|T_j\|_1 \}$ such that $\| \hat{\sum} T_j - e_j \|_{max} \leq \lambda_n$

Where $e_j \in \mathbb{R}^d$ is the j^{th} canonical vector

a) $\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \langle \hat{\Theta}^2, \sum \rangle - tr(\Theta) + \lambda \|\Theta\|_1 \right\} = 0$

$$\lambda^* = \frac{1}{2} (\Theta \hat{\sum} + \hat{\sum} \Theta) - I$$

$\hat{T} = \underset{T}{\operatorname{argmin}} \left\{ \frac{1}{2} \langle \hat{T}^2, \sum \rangle - e_j tr(T) + \lambda \|T\|_1 \right\} = 0$

$$\lambda^* = \frac{1}{2} (T \sum + \sum T) - e_j$$

An optimal solution depends on the j^{th} vector.

b) When $\lambda \geq |||\theta^*|||_1 \, \|\hat{\Sigma} - \Sigma\|_{max}$

$$\|\hat{T}_j\|_1 = \underset{T}{\operatorname{argmax}} \left\{ \frac{1}{2} T \Sigma T - e T_j + \lambda_n |T_j| \right\}$$

$$= \underset{T}{\operatorname{argmin}} \left\{ \frac{1}{2} T \Sigma T - e T_j + |||\theta^*|||_1 \|\hat{\Sigma} - \Sigma\|_{max} |T_j| \right\}$$

$$\leq \underset{\theta}{\operatorname{argmin}} \left\{ \frac{1}{2} \theta \Sigma \theta - I + |||\theta^*|||_1 \|\hat{\Sigma} - \Sigma\|_{max} \|\theta_1\| \right\}$$

$$\leq \|\theta_j\|$$

c) $\max_{j=1 \ldots d} \Sigma_{ij} \leq 1$; $\|\hat{\Sigma} - \Sigma^*\| \, |||\theta^*|||_1 \leq \lambda$

$$\|\hat{\Sigma} - \Sigma^*\|_{max} \leq \frac{\lambda_{max}}{|||\theta^*|||_1}$$

where λ_{max} is the maximum eigenvalue

Two or three papers support the bound by Corollary or Propositions:

1) Eigenvalue bounds are necessary
2) A data dimension relationship: $\sigma^2 \log d = O(n)$
3) Exponential tails on the graph

Outcomes from the model in part d.

A) Continuous, increasing and a concave plot from 0 to inf.
B) Zero at the origin, singularity.
C) Constants in the coefficient for fits at all inputs.

d) If $\lambda = c_1 \sqrt{\frac{\log d}{n}}$

$$|||\hat{\theta} - \theta^*||| \leq c_1 \cdot |||\theta|||_2^2 \sqrt{\frac{\log d}{n}}$$

11.5. (Equation 11.37)

$$\hat{\theta}_{j+} = \underset{\theta_{j+}}{\operatorname{argmin}} \left\{ -\frac{1}{n} \sum_{i=1}^{n} \log P_{\theta_{j+}} \{ X_{ij} | X_{i\setminus j} \} + \lambda \sum_{k \in V \setminus \{j\}} |||\theta_{jk}||| \right\}$$

$$= \underset{\theta_{j+}}{\operatorname{argmin}} \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} V_{ij} \right\|_2^2 + \lambda \sum |||\theta_{jk}||| \right\}$$

$$= \underset{\theta_{j+}}{\operatorname{argmin}} \left\{ \underbrace{\left\| \frac{1}{2n} \sum_{i=1}^{n} \{ X_{ij} - X_{i\setminus j} \theta \} \right\|^2}_{\mathcal{L}_n(\theta_{j+}; X_j, X_{\setminus j})} + \lambda_n \|\theta\|_1 \right\}$$

The regression term is a Lagrangian from Equation 9.66

11.6. (Equation 11.32) $p_\theta(X_1, \ldots, X_d) \propto \left\{ \sum_{j \in V} \phi_j(X_j; \theta_j^*) + \sum \phi_{jk}(X_j, X_k; \theta_{jk}^*) \right\}$

where $\phi_j(X_j; \theta_j^*) = \theta_j^* X_j$

$\phi_{jk}(X_j, X_k; \theta_{jk}^*) = \theta_{jk}^* X_j X_k$

$$\theta_{jk} = \begin{cases} \theta_{jk}^* X_j X_k & (j,k) \in E \quad \text{"points in a graphs vector set"} \\ 0 & (j,k) \notin E \quad \text{"points not in a graphs vector set"} \end{cases}$$

11.7 $X \in \{-1, 1\}^d$; $p_\theta(X_1, \ldots X_d) = \exp \left\{ \sum_{(j,k) \in E} \theta_{jk} \cdot X_j \circ X_k - \Phi(\theta) \right\}$

a) For each edge $(j,k) \in E$, $\frac{d\Phi(\theta)}{d\theta_{jk}} = \frac{1}{d} \sum_{(j,k) \in E} X_j X_j$

$$= \mathbb{E}[X_j X_k]$$

b) $P(X_1 = 1, \theta) = \exp \left\{ \sum_{(j,k) \in E} \theta_j X_k \right\}$

$P(X_1 = -1, \theta) = \exp \left\{ -\sum_{(j,k) \in E} \theta_j X_k \right\}$

$$p(X_k \mid X_{-k}) = \frac{p(X_k = 1)}{p(X_k = 1) + p(X_k = -1)}$$

$$= \frac{\exp\left\{\sum_{(j,k)\in E} \theta X_k\right\}}{\exp\left\{\sum_{(j,k)\in E} \theta X_k\right\} + \exp\left\{-\sum_{(j,k)\in E} \theta X_k\right\}}$$

$$= \frac{\exp\left\{\sum_{(j,k)\in E} \theta X_k\right\}}{\exp\left\{\sum_{(j,k)\in E} \theta X_k\right\} + 1} \left(\frac{1}{\exp\left\{-\sum_{(j,k)\in E} \theta X_k\right\}}\right)$$

$$= \frac{\exp\left(2\sum_{(j,k)\in E} \theta X_k\right)}{\exp\left(\sum_{(j,k)\in E} \theta X_k\right) + 1}$$

$$\frac{p(X_k \mid X_{-k})}{1 - p(X_k \mid X_{-k})} = \exp\left(\sum_{(j,k)\in E} \theta X_k\right)$$

$$\sum_{(j,k)\in E} \theta X_k = \log \frac{p(X_k \mid X_{-k})}{1 - p(X_k \mid X_{-k})}$$

11.8. $X = (X_1, \ldots, X_d) = N(X \mid 0, \sigma^2)$

$V = N(X \mid 0, \sigma^2 I_d)$

$Z = X + V$

If $\sigma^2 \|\theta^*\|_2 \le 1$, then $Z = X + V$

$= N(X \mid 0, \sigma_x^2 I_d) + N(X \mid 0, \sigma_y^2 I_d)$

$= N(X \mid 0, (\sigma_x^2 + \sigma_y^2) I_d)$

$$\frac{1}{\sigma_z^2} = \frac{1}{\sigma_x^2} + \frac{1}{\sigma_y^2}$$

"Expand out in powers"

1) $e^{-x^2/2\sigma_z^2} = \sum_{n=0}^{\infty} \frac{1}{n!}\left(\frac{-x^2}{2\sigma_z^2}\right)^n$

$= 1 - \frac{x^2}{2\sigma_z^2} + \frac{x^4}{8\sigma_z^4} - \frac{x^6}{48\sigma_z^6} + \cdots$

$= 1 - \left(\frac{1}{2}\right)\left(\frac{\sigma_x^2\sigma_y^2}{\sigma_x^2+\sigma_y^2}\right)x^2 + \left(\frac{1}{8}\right)\left(\frac{\sigma_x^4\sigma_y^4}{(\sigma_x^2+\sigma_y^2)^2}\right)x^4 - \cdots$

2) At $x=1$, $\sigma_z^2 = \begin{pmatrix} 1 & \sqrt{\frac{\sigma_x^2\sigma_y^2}{2(\sigma_x^2+\sigma_y^2)}} \\ \sqrt{\frac{\sigma_x^2\sigma_y^2}{2(\sigma_x^2+\sigma_y^2)}} & 1 \end{pmatrix}$

Cholesky Decomposition.

$= \begin{pmatrix} \sqrt{1} & 0 \\ \sqrt{\frac{\sigma_x^2\sigma_y^2}{2(\sigma_x^2+\sigma_y^2)}} & \sqrt{1-\frac{\sigma_x^2\sigma_y^2}{2(\sigma_x^2+\sigma_y^2)}} \end{pmatrix}\begin{pmatrix} \sqrt{1} & \sqrt{\frac{\sigma_x^2\sigma_y^2}{2(\sigma_x^2+\sigma_y^2)}} \\ 0 & \sqrt{1-\frac{\sigma_x^2\sigma_y^2}{2(\sigma_x^2+\sigma_y^2)}} \end{pmatrix}$

$= \sigma_z' \circ \sigma_z'$

In any case, whatever the expansion, the exponsion represents an accumulative deviation from each graph vertex. The standard deviation is amplitude in a Gaussian function and spread too.

11.9. (Equation 11.44) $\hat{\Theta} = \underset{\Theta \in S_+^{d\times d}}{\operatorname{argmin}} \{\langle\langle \Theta, \hat{\Gamma} \rangle\rangle - \log\det\Theta + \lambda |||\Theta|||_1\}$

a) $\sum_s \doteq \operatorname{cov}(X)$, $\lambda_n > \|\hat{\Gamma} - \Sigma_x\|_{max}$

$\hat{\Theta} = \underset{\Theta \in S_+^{d\times d}}{\operatorname{argmin}} \{\langle\langle \Theta, \Gamma \rangle\rangle - \log\det\Theta + \|\hat{\Gamma} - \Sigma_x\|_{max} |||\Theta|||_1\}$

$= 0$

$$\theta^* = \frac{1}{\Gamma + \|\Gamma - \Sigma\|}$$

$$\tilde{\theta} = \log(\Gamma + \|\Gamma - \Sigma\|) + 1$$

b) When λ has no bound, $\tilde{\theta}$ has a singularity at zero i.e. no value.

11.10. $y = X\theta^* + w$; rowspan$(X) = N(x|0, \Sigma)$

$Z = X + V$; rowspan$(V) = N(x|0, \sigma^2 I_d)$

$$\tilde{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - Z\theta\|_2^2 \right\}$$

Two incosistncies exist in the problem:

1) Z's variance is potentially imaginary.

$$Z = X + V$$
$$= N(x|0, \Sigma) + N(x|0, \sigma^2 I_d)$$
$$= N(x|0, \Sigma + \sigma^2 I_d)$$

From 11.8, $\frac{1}{\sigma_Z^2} = \frac{1}{\Sigma} + \frac{1}{\sigma^2 I_d}$

$$\sigma_Z^2 = \frac{\Sigma \sigma^2 I_d}{\Sigma + \sigma^2 I_d}$$

Taylor Exansion to second-order

$$= \begin{pmatrix} 1 & \sqrt{\frac{\Sigma \sigma^2 I_d}{\Sigma + \sigma^2 I_d}} \\ \sqrt{\frac{\Sigma \sigma^2 I_d}{\Sigma + \sigma^2 I_d}} & 1 \end{pmatrix}$$

Cholesky Decomposition

$$= \begin{pmatrix} 1 & 0 \\ \sqrt{\frac{\Sigma \sigma^2 I_d}{\Sigma + \sigma^2 I_d}} & \sqrt{1 - \frac{\Sigma \sigma^2 I_d}{\Sigma + \sigma^2 I_d}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \sqrt{1 - \frac{\Sigma \sigma^2 I_d}{\Sigma + \sigma^2 I_d}} \end{pmatrix}$$

The root $\sqrt{1 - \frac{\sum \sigma^2 I_d}{\sum t\sigma^2 I_d}}$ needs a bound in the problem, such as $\frac{\sum \sigma^2 I_d}{\sum t\sigma^2 I_d} \leq 1$.

2) A rare case when rowspan(V) draws from rowspan(X) as in this problem - a zero determinant in the denominator of argmin{}.

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - Z\theta\|_2^2 \right\}$$

$$= \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|X\theta + w - (X+V)\theta\|_2^2 \right\}$$

$$= \frac{1}{n} \|w - V\theta\| (\in V)$$

$$= 0$$

$$\theta^* = \frac{w}{V}$$

The denominator has a possible zero division

A solution involves a "regulizer" term, λ.

$$\tilde{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \frac{1}{2n} \|y - Z\theta\|_2^2 + \lambda \|\theta\|_1 \right\}$$

11.11 What is a high-dimensional dataset?

n-rows (measurements) and d-columns (data points)

where $n < d$.

Although, the upper limit $\left(\sqrt{\frac{n}{\log d}}\right)$ bounds the maximum size.

Format:

$$\min_{\|\theta\|_1 \le \sqrt{\frac{n}{\log d}}} \left\{ f_n(\theta) + \lambda \|\theta\|_1 \right\}$$

Functional limit
- avoids trivial solutions

Regularization term
- prevents zero division

Three examples explain the functional limit below.
Note: Papers online study beyond $\sqrt{\frac{n}{\log d}}$ in datasets. The documents suggest autoregressive patterns in outputs, such as overfit tails, oscillitory data, and "waves" in the image's fit.

Example #1: Additive Noise

A function category, $\min_{\|\theta\|_1 \le \sqrt{\frac{n}{\log d}}} \left\{ \frac{1}{2}\theta^T \hat{\Gamma} \theta - \langle \hat{\gamma}, \theta \rangle + \lambda_n \|\theta\|_1 \right\}$

$$\Gamma = \mathrm{cov}(X) = \frac{1}{n} X^T X - \Sigma$$

$$\gamma = \mathrm{cov}(X, y) = \frac{1}{n} X^T y$$

- When $\Sigma \succeq 0$, Γ is not always positive, semidefinite.
- A Negative Γ means negative covariance, such as

$$\Gamma = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \sigma^2 I & 0 \\ 0 & \sigma^2 I \end{bmatrix} = \begin{bmatrix} -\sigma^2 I_d & 0 \\ 0 & -\sigma^2 I_d \end{bmatrix}$$

- Γ has negative eigenvalues, $-\sigma^2 I_d$
- A regulary avoided saddle point in the eigenvalue space, $\lambda < 0$.
- The solution θ has a singularity at zero

Example #2: Missing Data

A function category, $\min_{\|\theta\|_1 \le \sqrt{\frac{n}{\log d}}} \{\frac{1}{2}\theta^T \Gamma \theta - \langle \bar{\gamma}, \theta \rangle + \lambda_n \|\theta\|_1\}$

$$\Gamma = \text{cov}(X) = \frac{X^T X}{n} - \rho \, \text{diag}\left(\frac{X^T X}{n}\right)$$

$$\gamma = \text{cov}(X, y) = \frac{1}{n} X^T y$$

- When ρ shows irregular behavior on the diagonal, specifically $\rho \neq 0$, but $\rho \in (0, 1)$.
- Γ has potential non-convex forms, negative eigenvalues, and bad outcomes for θ.

Example #3: Multiplicative Noise:

A function class; $\min_{\|\theta\|_1 \le \sqrt{\frac{n}{\log d}}} \{\frac{1}{2}\theta^T \Gamma \theta - \langle \bar{\gamma}, \theta \rangle + \lambda_n \|\theta\|_1\}$

$$\Gamma = \text{cov}(X) = \frac{1}{n} X^T X / \mathbb{E}[u^T u]$$

$$\bar{\gamma} = \text{cov}(X, Y) = \frac{1}{n} X^T Y / \mathbb{E}[u]$$

- u is non-negative noise
- $\mathbb{E}[u^T u]$, covariance in cases generates columns without noise.
- Null columns in a denominator cause singularity, division by zero.
- A case is a multivariate Bernoulli where $\mathbb{E}[u^T u] = \text{cov}(\text{Bernoulli}(p_i, p_j)) = \begin{cases} (1-p_i)(1-p_j) & i \neq j \\ (1-p_i) & i \neq j \end{cases}$

	Equation	Basic Idea	Examples
Biased Estimator	$\mathbb{E}_{X\|\theta}[\hat{\theta}] - \theta \neq 0$	• Observation • subjective	• variance • Lower or upper bound • Fahrenheit [Relative]
Unbiased Estimator	$\mathbb{E}_{X\|\theta}[\hat{\theta}] - \theta = 0$	• consistent • objective	• Shrinkage Estimation • Distance • Kelvin [Absolute]

11.12.

a) From Example 11.6, $Z = X + V$, $y = X^T\theta + w$, V is independent of X

$$V = N(x|0, \Sigma_V), \quad \Sigma_V = cov(V)$$

$$X = N(x|0, \Sigma_X), \quad \Sigma_X = cov(X)$$

$$\Gamma = \frac{1}{n} Z^T Z - \Sigma_V - \Sigma_X, \quad \gamma = \frac{Z^T y}{n}$$

Γ and γ are unbiased estimators.

$$\Gamma - \Sigma_X = \frac{1}{n} Z^T Z - \Sigma_V - \Sigma_X$$

$$= \frac{1}{n}(X^T + V^T)(X + V) - \Sigma_V - \Sigma_X$$

$$= \frac{X^T X}{n} + 2\frac{X^T V}{n} + \frac{V^T V}{n} - \Sigma_V - \Sigma_X = 0$$

Independent

$$\hat{\gamma} - \Sigma_{XY} = \frac{Z^T y}{n} - \Sigma_{XY}$$

$$= \frac{(X^T + V^T)}{n} y - \Sigma_{XY}$$

$$= \frac{X^T y}{n} + \frac{V^T y}{n} - \Sigma_{XY}$$

$$= \frac{X^T y}{n} + \frac{V^T(X\theta + w)}{n} - \Sigma_{XY}$$

$$= \frac{X^T y}{n} - \Sigma_{XY}$$

$$= 0$$

(Proposition 11.10)

$$\|\hat{\Gamma}\theta^* - \gamma\|_{max} \leq \Phi(Q, \sigma_w)\sqrt{\frac{\log d}{n}}$$

"Covariance subtraction for sub-gaussians becomes subgaussian."

11.13. $Z_{ij} = \begin{cases} Z_{ij}/1-v & \text{if } (i,j) \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$

$$\hat{\Gamma} = \frac{1}{n} Z^T Z - v\,\mathrm{diag}\left(\frac{Z^T Z}{n}\right) \; ; \; \Sigma_X = \mathrm{cov}(X)$$

a) $\hat{\Gamma}$ and $\hat{\gamma}$ are unbiased estimators:

$$\hat{\Gamma} - \Sigma_X = \frac{1}{n} Z^T Z - v\,\mathrm{diag}\left(\frac{Z^T Z}{n}\right) - \Sigma_X$$

$$= \frac{1}{n} X^T X - \Sigma_X$$

$$= 0$$

$$\hat{\gamma} - \Sigma_{XY} = \frac{1}{n} Z^T y - \Sigma_{XY}$$

$$= \frac{1}{n} X^T y - \Sigma_{XY}$$

$$= 0$$

b) X_{ij} and Z_{ij} are sub-Gaussian:

By Proposition 11.13, $\mathbb{P}[\|\hat{\Gamma}\theta - \hat{\gamma}\|] \leq \underbrace{\phi(Q, \sigma_W)}_{c(\sigma + \|\theta\|_2)} \sqrt{\frac{\log d}{n}}$

Critical Note The upper bound in high-dimensional datasets curtail empirical rules in 'normal' distributions.

High-dimensional limit	Empirical Statistics		
		value	z-table
$\phi(Q, \sigma_W)\sqrt{\frac{\log d}{n}}$	one-sigma:	σ	1.0
		1.7σ	1.7
	two-sigma:	2σ	2.0

coefficient

$\sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$

b) X_{ij} and V_{ij} are sub-Gaussian

By Proposition 11.18, $\mathbb{P}[\|\hat{\Gamma}\theta^* - \gamma\|_\infty \leq \underbrace{\phi(Q,\sigma_w)}_{= c(\sigma + \|\theta\|_2)} \sqrt{\frac{\log d}{n}}]$

$$\mathbb{P}[\|\hat{\Gamma}\theta^* - \gamma\|_\infty \leq c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}]$$

$$= \mathbb{P}[e^{1 + \lambda\mathbb{E}[\|\hat{\Gamma}\theta - \gamma\|_\infty] + \frac{\lambda^2\mathbb{E}[(\|\hat{\Gamma}\theta - \gamma\|_\infty)^2]^2}{2}} \leq e^{\lambda c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}}]$$

$$\leq \mathbb{P}[e^{\frac{\lambda^2\mathbb{E}[(\|\hat{\Gamma}\theta - \gamma\|_\infty)^2]^2}{2}} \leq e^{\lambda c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}}]$$

$$\operatorname{argmin}_\lambda \left\{ \frac{\lambda^2\mathbb{E}[(\|\hat{\Gamma}\theta - \gamma\|_\infty)^2]^2}{2} - \lambda c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}} \right\} = 0$$

$$\lambda^* = \frac{c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}}{\mathbb{E}[(\|\Gamma\theta - \gamma\|_\infty)^2]^2}$$

$$\mathbb{P}[\|\hat{\Gamma}\theta^* - \gamma\|_\infty \leq c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}] \leq e^{\frac{-c(\sigma + \|\theta\|_2)^2 \frac{\log d}{n}}{2\,\mathbb{E}[(\|\hat{\Gamma}\theta - \gamma\|_\infty)^2]^2}}$$

$$\mathbb{E}[\|\hat{\Gamma}\theta - \gamma\|_\infty] = \frac{1}{\lambda^*}\mathbb{E}\left[\log e^{\frac{-c(\sigma + \|\theta\|_2)^2 \frac{\log d}{n}}{2\,\mathbb{E}[(\|\Gamma\theta - \gamma\|_\infty)^2]^2}}\right]$$

$$= \frac{c(\sigma + \|\theta\|_2)}{2}\sqrt{\frac{\log d}{n}}$$

Note: The $(1/2)$ goes away with a double-sided dist.

c) (Equation 11.50) $\langle \Delta, T\Delta \rangle \geq k\|\Delta\|_2^2 - c_0 \frac{\log d}{n}\|\Delta\|_1^2$

From definition 7.12, $\frac{1}{n}\|X\Delta\|_2^2 \geq k\|\Delta\|_2^2$ for all Δ

$$\langle \Delta, T\Delta \rangle = \Delta T \Delta$$

$$\geq k\|\Delta\|_2^2$$

$$\geq k\|\Delta\|_2^2 - c_0 \frac{\log d}{n}\|\Delta\|_2^2$$

$$\mathbb{P}\left[\|\hat{T}\theta^* - \gamma\|_\infty \le c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}\right]$$

$$:= \mathbb{P}\left[e^{1 + \lambda\mathbb{E}[\|\hat{T}\theta - \gamma\|_\infty] + \frac{\lambda^2\mathbb{E}[\|T\theta - \gamma\|_\infty^2]^2}{2} + \cdots} \ge e^{\lambda c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}}\right]$$

$$\le \mathbb{P}\left[e^{\frac{\lambda^2\mathbb{E}[\|T\theta - \gamma\|^2]^2}{2}} \le e^{\lambda c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}}\right]$$

$$\operatorname*{argmin}_\lambda \left\{ \frac{\lambda^2\mathbb{E}[\|T\theta - \gamma\|^2]^2}{2} - \lambda c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}} \right\} = 0$$

$$\lambda^* = \frac{c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}}{\mathbb{E}[\|T\theta - \gamma\|^2]^2}$$

$$\mathbb{P}\left[\|T\theta - \gamma\|_\infty \le c(\sigma + \|\theta\|_2)\sqrt{\frac{\log d}{n}}\right] \le e^{\frac{-c(\sigma + \|\theta\|_2)\frac{\log d}{n}}{2\mathbb{E}[\|T\theta - \gamma\|^2]^2}}$$

$$\mathbb{E}[\|\hat{T}\theta - \gamma\|_\infty] = \frac{1}{\lambda^*}\mathbb{E}\left[\log e^{\frac{-c(\sigma + \|\theta\|_2)\frac{\log d}{n}}{2\mathbb{E}[\|T\theta - \gamma\|^2]^2}}\right]$$

$$= \frac{c(\sigma + \|\theta\|_2)}{2}\sqrt{\frac{\log d}{n}}$$

c) (Equation 11.50) $\langle \Delta, T\Delta \rangle \ge k\|\Delta\|_2^2 - c_0\frac{\log d}{n}\|\Delta\|_1^2$

From definition 7.12, $\frac{1}{n}\|X\Delta\|_2^2 \ge k\|\Delta\|_2^2$ for all Δ

$$\langle \Delta, T\Delta \rangle = \Delta T\Delta$$

$$\ge k\|\Delta\|_2^2$$

$$\ge k\|\Delta\|_2^2 - c_0\frac{\log d}{n}\|\Delta\|_2^2$$