High Dimensional Probability Primer

Gordon Lichtstein

1 Introduction

High dimensional probability studies random objects (vectors) in $\mathbb{R}^n$, where n is large, sometimes larger than the number of samples. High dimensional probability has applications across data science, algorithms, and statistics. This talk will focus on the basics of HDP and its applications, namely covariance estimation of high dimensional data.

2 Random Variables

Basic Definitions

Let X be a real-valued random variable on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$.

- **Expectation:** $\mathbb{E}[X]$.
- **Covariance:** $\mathrm{Cov}(X,Y) = \mathbb{E}\big[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\big]$.
- **Variance:** $\mathrm{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}X)^2\big] = \sigma^2 = \mathrm{Cov}(X,X)$.
- **L_p norm:** $\|X\|_{L_p} := (\mathbb{E}|X|^p)^{1/p}$ for $p \geq 1$.

Common distributions

- **Bernoulli:** $X \sim \mathrm{Ber}(p)$ if $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$. Often $p = 0.5$.
- **Rademacher (symmetric Bernoulli):** $\mathbb{P}(X = 1) = \mathbb{P}(X = -1) = 1/2$.
- **Exponential:** $X \sim \mathrm{Exp}(\lambda)$ if $\mathbb{P}(X > t) = e^{-\lambda t}$ for $t \geq 0$.
- **Gaussian:** $g \sim N(\mu, \sigma^2)$ with density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big),$$

Sub-gaussian random variables

A random variable X is sub-gaussian if it has Gaussian-type tails. This allows us to use concentration inequalities for Gaussian random variables in a more general context.

A random variable is sub-gaussian if there exist constants $C, c > 0$ such that

$$\mathbb{P}(|X| \geq t) \ \leq \ C\exp(-ct^2) \text{ for all } t \geq 0.$$

Equivalently, a variable is subgaussian if its Orlicz ψ_2 norm (also called the sub-gaussian norm) is finite:

$$\|X\|_{\psi_2} := \inf\Big\{t > 0 : \ \mathbb{E}\exp\big(X^2/t^2\big) \leq 2\Big\}.$$

Bounded variables, Rademacher, and Gaussians are sub-gaussian. Exponential distributions are not, as their tails are too heavy.

Sub-exponential random variables

A random variable X is sub-exponential if it has exponentially decaying tails. This is a weaker condition than a random variable being sub-gaussian, and sub-exponential random variables frequently appear when working with squares and products of sub-gaussian random vars.

A random variable is sub-exponential if there exist constants $C, c > 0$ such that

$$\mathbb{P}(|X| \geq t) \ \leq \ C\exp(-ct) \text{ for all } t \geq 0.$$

Equivalently, X is sub-exponential if its Orlicz ψ_1 norm (also called the sub-exponential norm) is finite:

$$\|X\|_{\psi_1} := \inf\Big\{t > 0 : \ \mathbb{E}\exp\big(|X|/t\big) \leq 2\Big\}.$$

Sub-exponential random variables satisfy Bernstein-type concentration inequalities.

Useful Properties

- If X is sub-gaussian, then X^2 is sub-exponential, and $\|X^2 - \mathbb{E}[X^2]\|_{\psi_1} \ \leq \ C\|X\|_{\psi_2}^2$
- If X and Y are sub-gaussian, then XY is sub-exponential, and $\|XY\|_{\psi_1} \ \leq \ C\|X\|_{\psi_2}\|Y\|_{\psi_2}$.
- If X is sub-gaussian, then $X - \mathbb{E}[X]$ is sub-gaussian, and $\|X - \mathbb{E}[X]\|_{\psi_2} \ \leq \ C\|X\|_{\psi_2}$.
- If X is sub-exponential, then $X - \mathbb{E}[X]$ is sub-exp, and $\|X - \mathbb{E}[X]\|_{\psi_1} \ \leq \ C\|X\|_{\psi_1}$.

3 Random Vectors

Basic definitions and concepts

Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector.

- **Expectation:** $\mathbb{E}[X] \in \mathbb{R}^n$, defined componentwise.
- **Second moment matrix:** $\Sigma := \Sigma(X) = \mathbb{E}[XX^T]$.
- **Covariance matrix:** $\mathrm{Cov}(X) := \mathbb{E}\big[(X-\mathbb{E}X)(X-\mathbb{E}X)^T\big] = \mathbb{E}\big[XX^T\big] - \mathbb{E}[X]E[X]^T$.
- **Principle component analysis:** PCA collects the eigenvectors of the largest eigenvalues of the covariance matrix, describing the data's most important dimensions.

High-dimensional Gaussian and sub-gaussian random vectors

A random vector $X \sim N(\mu, \Sigma)$ in $\mathbb{R}^n$ is a multivariate Gaussian if every one-dimensional projection $\langle X, u\rangle$ is a Gaussian random variable.

- $\mathbb{E}[X] = \mu$ and $\mathrm{Cov}(X) = \Sigma$.
- The density of X is

$$f(x) = \frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}} \exp\Big(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big).$$

- **Gaussian tail bound (one dimensional Gaussian)**: Let $g \sim N(0,1)$. Then for all $t > 0$,

$$\Big(\frac{1}{t} - \frac{1}{t^3}\Big)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \;\le\; \mathbb{P}(g \ge t) \;\le\; \frac{1}{t}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

  In particular, for $t \ge 1$,

$$\mathbb{P}(g \ge t) \le \frac{1}{\sqrt{2\pi}}e^{-t^2/2}.$$

A random vector $X \in \mathbb{R}^n$ is called sub-gaussian if all one-dimensional projections are sub-gaussian. Gaussian vectors and vectors with independent sub-gaussian entries are sub-gaussian. Equivalently, there exists $K > 0$ such that for all $u \in \mathbb{R}^n$,

$$\|\langle X, u\rangle\|_{\psi_2} \le K\|u\|_2.$$

The smallest such K is called the sub-gaussian norm of X and is denoted $\|X\|_{\psi_2}$.

Isotropic random vectors

A random vector $X \in \mathbb{R}^n$ is called isotropic if $\mathbb{E}[X] = 0$ and $\mathbb{E}[XX^T] = I_n$.

- Equivalently, $\mathbb{E}[\langle X, u\rangle^2] = \|u\|_2^2$ for all $u \in \mathbb{R}^n$.
- Any random vector with covariance Σ can be made isotropic by $Z = \Sigma^{-1/2}(X - \mathbb{E}[X])$.
- Intuitively, isotropic random vectors look the "same" in every direction.

4 Epsilon nets

An ε-net $N \subset S^{n-1}$ of the sphere is a finite subset such that for every $x_0 \in S^{n-1}$ there exists $x \in N$ with $\|x - x_0\|_2 \leq \varepsilon$.

Epsilon nets allow us to get a handle on a set (in our case, the sphere) by approximating with a much smaller space, which is easier to reason about. There exist ε-nets with cardinality at most $(1+2/\varepsilon)^n$.

5 Linear Algebra Review

- **Eigenvalues and trace:** For a symmetric matrix A with eigenvalues $\lambda_1, \dots, \lambda_n$, $\mathrm{tr}(A) = \sum_{i=1}^n \lambda_i$, and $\|A\| = \max_i |\lambda_i|$.
- **Covariance matrices:** If $\Sigma = \mathbb{E}[XX^T]$, then $\mathrm{tr}(\Sigma) = \mathbb{E}\|X\|_2^2$, and $\|\Sigma\| \leq \mathrm{tr}(\Sigma)$.
- **Operator norm of matrices:** $\|A\| = \sup_{\|x\|_2=1} \|Ax\|_2$. $\|A\|$ equals the largest singular value of A (s_1). The operator norm of a symmetric matrix is its largest absolute eigenvalue

6 Important Random Variable Inequalities

- **Union bound:** For events $A_1, \dots, A_n$,

$$\mathbb{P}\Big(\bigcup_{i=1}^n A_i\Big) \leq \sum_{i=1}^n \mathbb{P}(A_i).$$

- **Markov's inequality:** For any nonnegative random variable X and $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

- **Chernoff's inequality:** Let $X_1, \dots, X_N$ be independent Bernoulli random variables with parameters p_i. Let $S_N = \sum_{i=1}^N X_i$, and $\mu = \mathbb{E}[S_N]$. Then for any $t > \mu$,

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu}\Big(\frac{e\mu}{t}\Big)^t.$$

- **Hoeffding's inequality:** Let $X_1, \dots, X_n$ be independent mean-zero sub-gaussian random variables with $\|X_i\|_{\psi_2} \leq K$. Then for all $t \geq 0$,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^n X_i\Big| \geq t\Big) \leq 2\exp\Big(-c\,\frac{t^2}{nK^2}\Big).$$

7 Important Matrix Inequalities

- **Matrix Bernstein inequality (expectation):** Let $X_1,\dots,X_N$ be independent, mean-zero, symmetric $n \times n$ random matrices such that $\|X_i\| \le K \quad$ almost surely for all i. Then
$$\mathbb{E}\Big\|\sum_{i=1}^{N} X_i\Big\| \le C\left(\Big\|\sum_{i=1}^{N}\mathbb{E}[X_i^2]\Big\|^{1/2}\sqrt{1+\log n}+K(1+\log n)\right).$$

- **Davis–Kahan theorem:** Let S and T be symmetric matrices of the same dimension. Fix an index i and assume the i-th largest eigenvalue of S is separated from the rest of the spectrum:
$$\min_{j\neq i}|\lambda_i(S)-\lambda_j(S)| = \delta > 0.$$
Then the angle between the corresponding eigenvectors satisfies
$$\sin\angle\big(v_i(S), v_i(T)\big) \le \frac{2\|S-T\|}{\delta}.$$
In particular, there exists $\theta \in \{-1,1\}$ such that
$$\|v_i(S)-\theta\, v_i(T)\|_2 \le \frac{2^{3/2}\|S-T\|}{\delta}.$$
So the unit eigenvectors $v_i(S), v_i(T)$ are close to eachother up to a sign

8 References

- R. Vershynin, High-Dimensional Probability: An Introduction with Applications in Data Science, 2024.