Introduction

My notes from the legendary book of Papoulis and Pillai [1]. The book assumes that the reader does already have fairly strong background in calculus and linear algebra. I used two books [2, 3] as assistant for evaluating some of the mathematical identities or integrals.

This book made me realize that a good technical book is not necessarily one that's easy to follow – one that shows you very easily how each identity or result is arrived at. Some of the results in the book are presented without much explanation, and this required me to show a lot of effort to understand them; this effort seems to lead to a staying power.

Chapter 1

Sequences of Random Variables

1.1 Notes

- Multivariate Transformation
- How to integrate some RVs from a multi-variate distro to obtain ... (Sec 7.2)
- How to compute the mean conditioned on a subset of RVs
- Characteristic function leads to so much interesting applications, such as:
- Computing the PDF of Bernoulli from Binomials (#256)
- Computing the PDF of Poisson from Binomials (#256)
- The sum of jointly normal RVs is normal (#257)
- The sum of the squares of independent normal variables is chi-square (#259)
- The sum of two chi-square distros is also chi-square (#260)
- The optimal single-value estimation (in the MS sense), c, of a future value of a RV $\mathbf{y}$ is $c = E(\mathbf{y})$.
- The optimal functional estimation of $\mathbf{y}$ *i.t.o.* a (dependent) RV $\mathbf{x}$ is $c(x) = E(\mathbf{y}|x)$. In general, this estimation is *non-linear*.
- The optimal linear f. estimation of $\mathbf{y}$ *i.t.o.* $\mathbf{x}$ is $c(x) = Ax+B$ where $B = \eta_y - A\eta_x$ and $A = r\sigma_x/\sigma_y$.
- *For Gaussian RVs, linear and non-linear MS estimators are identical (#264).*
- *The orthogonality principle*: The error between of linear estimator $\hat{\mathbf{y}} = A\mathbf{x} + B$ of $\mathbf{y}$ is orthogonal to the data: $E\{(\mathbf{y} - \hat{\mathbf{y}})\mathbf{x}\} = 0$
- Generalizes to the linear estimate of $\mathbf{s}$ *i.t.o.* multi RVs $\mathbf{x}_1, \ldots, \mathbf{x}_n$: $E\{(\mathbf{s} - \hat{\mathbf{s}})\mathbf{x}_i\}$ for $i = 1, \ldots, n$.
- Generalizes to *non-linear* estimation: $E\{[\mathbf{s} - g(\mathbf{X})]w(\mathbf{X})\}$, where $g(\mathbf{X})$ is the non-linear MS estimator and $w(\mathbf{X})$ *any* function of data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. (#269)
- Computing the linear MS estimator is *much easier* if the RVs $\mathbf{x}_i$ are orthogonal to one another (*i.e.* $R_{ij} = 0$ for $i \neq j$). That is why it is often we perform *whitening* (see #271-272).
- Convergence modes for a random sequence (RS) $\mathbf{x}_1, \ldots, \mathbf{x}_n, \ldots$
  - Everywhere (e), almost everywhere (a.e.), in the MS sense (MS), in prob (p), in distro (d).
  - Ordered in relaxedness (except last two): d > p > (a.e. | MS)

- Cauchy criterion (CC): We typically think of convergence as converging into a sequence x. With CC, we can eliminate this need. That is, we ask that $|x_{n+m} - x_n| \to 0$ as $n \to \infty$.

- All the below are technically applications of RS convergence:
  - The law of large numbers: If p is the prob of event A in a single experiment and k is number of successes in n trials, then we can show that p tends to k/n in *probability.*
  - Strong law of large numbers: We can even show that p tends to k/n almost everywhere (but proof is more complicated).

- The ones below are more specifically applications for estimating $E\{\bar{\mathbf{x}}_n\}$ (and optionally $\bar{\sigma}_n^2$): the sample mean of a RS of n RVs $\bar{\mathbf{x}}_n = \frac{\mathbf{x}_1 + \ldots + \mathbf{x}_n}{n}$ (the variance of the sample mean)
  - Markov's thm: If the RVs $\mathbf{x}_i$ are s.t. the mean of $\bar{\eta}_n$ of $\bar{\mathbf{x}}_n$ tends to a limit η and its variance $\bar{\sigma}_n$ tends to 0 as $n \to \infty$, then $E\{(\bar{\mathbf{x}}_n - \eta)^2\} \to 0$ as $n \to \infty$. Convergence in MS sense.
    (Note that $\mathbf{x}_i$ do not have to be uncorrelated or independent)
  - Corollary: if $\mathbf{x}_i$ are uncorrelated and $\frac{\sigma_1^2 + \ldots + \sigma_n^2}{n} \to 0$ as $n \to \infty$, then $\bar{\mathbf{x}}_n \to \eta = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} E\{\mathbf{x}_n\}$ as $n \to \infty$. Convergence in MS sense.
    (Note that now we do require uncorrelated RVs but in exchange, compared to Markov's thm, the condition on the mean is removed and we *do* have a way of computing the mean η.)
  - Khinchin's thm: the above two required us to know *something* about the variance. According to Khinchin, if $\mathbf{x}_i$ are i.i.d. (stricter condition), then we $\bar{\mathbf{x}}_n$ tends to η even if we know nothing about the variance of $\mathbf{x}_i$'s. However, now we have convergence in probability only.

- The Central Limit Theorem (CLT) is also application of RS conv.—conv. in *distribution*:
  - Given n independent RVs $\mathbf{x}_i$, we form their sum (not sample mean!) $\mathbf{x} = \mathbf{x}_1 + \ldots + \mathbf{x}_n$. This is an RV with mean η and σ. CLT states that the distro $F(x)$ of $\mathbf{x}$ approaches a *normal distro* with the same mean and variance: $F(x) \approx G(\frac{x-\eta}{\sigma})$.
  - If the RVs are continuous, then $f(x)$, the *density* of x, also approaches a normal density.
  - The approximations become *exact* asymptotically (*i.e.* as $n \to \infty$).

1.2 Interesting identities/lemmas/theorems

- The unbiased linear estimator with minimum variance is the one shown in (7-17).
- Theorem 7.1: The correlation matrix is nonnegative definite
- Sum of jointly normal RVs is normal
- Sample mean and variance of n RVs $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n}$ and $\bar{\mathbf{v}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{x} - \bar{\mathbf{x}})^2$.
- Goodman's theorem: The statistics of a real zero-mean n-dimensional normal RV are completely determined i.t.o. the n^2 parameters—the elements of its covariance matrix. However, a similar *complex* RV requires $2n^2 + n$ elements. Goodman's theorem (#259) gives a special class of normal complex RVs that are determined completely with n^2 parameters only.

1.3 Important concepts

- Group independence
- Correlation and covariance matrices
- The orthogonality principle
- Mean square estimation (Linear and non-linear)

1.4 Some terminology

- Homogeneous linear estimation: Linear estimation without a bias term
- Nonhomogeneous linear estimation: Linear estimation with a bias term

1.5 Redo in future

- Poisson
- Chi square
- Show why sample variance is divided by $n-1$.
- Prove why the generalized orthogonality principle (7-92) leads to optimal MS estimators (linear or non-linear).
- Order convergence modes
- Prove the law of large numbers

Towards a motivation guide

1.6 Why transformations are useful

- By applying an orthonormal transformation (*a.k.a.* whitening) to a set of RVs we can easily compute the optimal

Towards a cheatsheet

1.7 Interesting RV Transformations

- If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are jointly normal, then $\mathbf{z} = \mathbf{x}_1 + \ldots \mathbf{x}_n$ is also normal (#257)
- If $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are independent and each $\mathbf{x}_i$ is $N(0, 1)$, then $\mathbf{z} = \mathbf{x}_1 + \ldots \mathbf{x}_n$ is $\chi^2(n)$ (#259)
- If $\mathbf{x}$ is $\chi^2(n)$ and $\mathbf{y}$ is $\chi^2(m)$ and $\mathbf{x}, \mathbf{y}$ independent, then $\mathbf{z} = \mathbf{x} + \mathbf{y}$ is $\chi^2(n + m)$ (#260)

Bibliography

[1] A. P. anb S. Unnikrishna Pillai, *Probability, Random Variables and Stochastic Processes.* McGraw - Hill, 2002.

[2] I. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series and Products.* Academic Press, 1980.

[3] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions.* Dover, 1970.