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 $n \to \infty$).
  - Good n values: if $\mathbf{x}_i$ are i.i.d. , then $n = 30$ is adequate for most applications. If the functions $f_i(x)$ are smooth, even $n = 5$ can be enough.
  - Error corr: The errors of CLT can be corrected by matching higher order moments using Hermite polynomials (see#281 and also Section 1.8.1).
  - CLT for products $\mathbf{y} = \mathbf{x}_1\mathbf{x}_2 \ldots \mathbf{x}_n$ (assuming $\mathbf{x}_i > 0$) is *lognormal*. (Easily seen by letting $\mathbf{z} = \ln \mathbf{y} = \ln \mathbf{x}_1 + \ldots + \ln \mathbf{x}_n$ and $\mathbf{y} = e^{\mathbf{z}}$.

- Sampling random sequences for a target distribution $F(x)$

- The general idea is to generate a uniform RS $u_1, \ldots, u_n$, and then use some method to convert it to a RS $x_1, \ldots, x_{n'}$ (where $n' \leq n$) that matches the distribution of $F(x)$.

- There are several methods for sampling
  - Percentile Transformation Method: This is a straightforward method ($x_i = F_x^{-1}(u_i)$) but we can apply it only if we can compute the inverse of the target distro, $F_x^{-1}(u)$. We can use this also to transform a sequence y_i with a non-uniform distribution: $x_i = F_x^{-1}(F_y(y_i))$ (see #292).
  - Rejection method: Allows sampling without F_x^{-1}. The idea is to find an event M such that $f_x(x|M) = f_y(x)$. Turns out that this event is $M = \{\mathbf{u} \leq r(\mathbf{x})\}$ where $r(\mathbf{x}) = af_y(x)/f_x(x)$. To implement this, one needs to find the upper bound (i.e. a; see Example 7-23).
  - Mixing method: $f_x(x) = p_1 f_1(x) + \ldots + p_m f_m(x)$. See proof in #294 for explanation.

- Some transformations:
  - Binomial RV: Sum of m i.i.d. Bernoulli's equals Binomial (#294).
  - Erlang: Sum of m i.i.d. exponentials equals Erlang
  - Chi-square: Sum an appropriate Erlang and a Normal
  - Student-t: Use a normal and a chi-square
  - Log-normal: Use a normal
- Generating Normal RSs
  - Clearly, we can use CLT and sum some uniform RSs
  - Rejection and mixing: We can mix two Truncated normal RVs obtained via rejection
  - Polar coordinates: Use a Rayleigh and a uniform to obtain two independent normals
  - Use two uniforms to obtain two normals (see Example 6-15 to follow the equation right after 7-172)

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.
- Berry-Esseen Theorem: Gives an upper bound for the approx. error of CLT, given that the third order of the unknown distro is finite and the variables $\mathbf{x}_i$ are i.i.d.

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)

Sketches to solutions

1.8 Solutions to equations in the book

1.8.1 Error correction for CLT

It is really not obvious how the error correction for CLT is arrived at in #281. The general idea is to match the higher order moments of the approximation with that of the data. The approximation error between the (unknown) density $f(x)$ and the approximated normal density $f_n(x;0,\sigma)$ is:

$$\epsilon(x) = f(x) - f_n(x;0,\sigma). \tag{1.1}$$

The idea is to write the error (and thus $f(x)$) *i.t.o.* an orthogonal set of polynomials, viz. Hermite Polynomials $H_k(x)$ (see 7-126). Since they are orthogonal, they form an orthogonal set on the real line:

$$\int_{-\infty}^{\infty} e^{-x^2/2} H_n(x) H_m(x) dx = \begin{cases} n!\sqrt{2\pi} & n = m \\ 0 & n \neq m \end{cases} \tag{1.2}$$

Using those polynomials, one can approximate any function (including $\epsilon(x)$ or $f(x)$) *i.t.o.* an infinite series:

$$\epsilon(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \sum_{k=3}^{\infty} C_k H_k\left(\frac{x}{\sigma}\right) \tag{1.3}$$

The sum starts from 3, as the moments of $\epsilon(x)$ up to order 2 are zero (I guess because the Gaussian approx. is sufficient to make those moments zero?) The book limits itself up to $4th$ order approx.

The idea is to find the coefficients C_3 and C_4 by matching the 3^{rd} and 4^{th} order moments of the unknown distro with that of the data (*** Be more pecise here). The third order moment of $\mathbf{x}$ is:

$$E_{f(x)}\{x^3\} = \int x^3 f(x) dx \tag{1.4}$$

The third order moment of the (approximated) normal density is zero (see 5-73), and therefore the error in terms of the third moments is:

$$E_{f(x)}\{x^3\} - E_{f_n(x)}\{x^3\} = \int x^3 f(x) dx - \int x^3 f_n(x;0,\sigma) dx = \int x^3 \left[f(x) - f_n(x;0,\sigma)\right] dx. \tag{1.5}$$

Since the content of the last brackets equals $\epsilon(x)$, and since $E_{f_n(x)}\{x^3\}$ is zero, the above equals $m_3 = E_{f(x)}\{x^3\}$ identity can be written as:

$$m_3 = \frac{1}{\sigma\sqrt{2\pi}} \int x^3 e^{-x^2/2\sigma^2} \sum_{k=3}^{\infty} C_k H_k\left(\frac{x}{\sigma}\right) dx \tag{1.6}$$

Now here is the tricky part that is not obvious in the book. Our goal is to find the coefficients C_k using the above equality. This is achieved by using the orthogonality of the Hermite polynomials. The third

order Hermite Polynomial is $H_3(x) = x^3 - 3x$, and using (1.2) and doing a change of variables we see that:

$$\int e^{-x^2/2\sigma^2} \left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) \left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) dx = \tag{1.7}$$

$$= \int e^{-x^2/2\sigma^2} \frac{x^3}{\sigma^3} \left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) dx - 3\int e^{-x^2/2\sigma^2} \frac{x}{\sigma} \left(\frac{x^3}{\sigma^3} - 3\frac{x}{\sigma}\right) dx \tag{1.8}$$

$$= 3!\sigma\sqrt{2\pi} \tag{1.9}$$

The second integral in (1.8) equals zero because $H_1(x) \propto x$ and $H_3(x) \perp H_1(x)$, and the first integral in (1.8) is proportional to m_3. More specifically, using (1.8) and (1.6), we arrive at the obscure derivation in the book:

$$m_3 = 3!\sigma^3 C_3. \tag{1.10}$$

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.