Notes of Statistical Digital Signal Processing and Modelling (TODO)

Gralerfics

December 18, 2025

Contents

1. Linear Algebra

（TODO）主要关于线性方程组的解、秩、逆、特征值、二次型与正定性、矩阵微积分等。

2. Probability Theory

2.1. Random Variable

一些现象（如投骰子）因具有随机性而难以预先知道结果，我们称为**随机现象**。随机现象不断发生，我们每观察并记录一次其结果称为一次**随机试验**，记录的结果称为一个样本。

一次随机试验可能得到各种不同的结果（样本），所有样本的集合称为该随机现象的样本空间，其中包含了该随机现象可能出现的所有结果。

随机事件是样本空间的子集，它包含了所有可能发生的结果（样本）中的一部分结果（样本）。如果一次随机试验得到的结果（样本）ω 属于某个随机事件 A 包含的样本子空间，则称随机事件 A 发生了，否则称没有发生。

随机事件可以是各种陈述，例如“明天下雨”是“明天下雨与否”这一随机现象的一个随机事件，明天下雨与否的样本空间是 $\{下雨, 不下雨\}$，而“明天下雨”就是其子集 $\{下雨\}$。又如，“投掷一个均匀二十面骰子的结果大于十八”是“投掷一个均匀二十面骰子所得结果”这一随机现象的一个随机事件，随机现象的样本空间是 $\{1, 2, ..., 20\}$，而该随机事件可以记为 $\{19, 20\}$。

随机变量 x 用于描述一些结果可以数量化的随机现象的结果，例如前述“投掷一个二十面骰子所得的结果是 1 到 20 共二十个整数。而“明天下雨与否”这种不显式包含数量的样本空间也可以通过人为规定“下雨为 1，不下雨为 0”来实现数量化的表达。

> 实际上大部分讲义中用大写字母表示随机变量，但这只是记号的区别。为了同之后随机过程中的记号统一，这里就用小写字母表示，并在必要的时候注明此为随机变量。相应的，参数等自变量（例如后面概率密度函数的参数）就用希腊字母 α 等表示，以免混淆。

一个随机事件 A 发生的**概率**记为 $\Pr(A)$，表示事件发生的可能性。包含样本空间全集的随机事件的概率为 1，因为随机试验的结果一定被包含在其中，表示该随机事件必然发生。其他关于概率、条件概率、全概率和样本空间集合的内容就不再赘述。

2.2. Probability Density Function (PDF)

一个随机变量的不同取值具有不同的概率，如果考虑离散随机变量（可能取的值的个数是有限的），那么一个描述其所有可能取值到各取值概率的映射称为概率质量函数（Probability Mass Function，PMF），它包含了关于这个分布的所有信息。

如果考虑连续随机变量，可能取值的个数是无限的，那么取到某个具体值的概率都近似为 0，不能再用类似 PMF 的方式为每个取值赋予概率值的方式来构建函数了。我们首先为随机变量 x 定义累计分布函数（Cumulative Distribution Function，CDF）：

$$F_x(\alpha) = \Pr\{x \leq \alpha\} \tag{1}$$

大括号或是小括号只是记号，不影响。总之它的意思是 x 小于或等于参数 α 的概率。显然，对于取值为实数的连续随机变量 x，α 在负无穷时 CDF 趋于 0，α 在正无穷时 CDF 趋于 1。我们通过 CDF 定义概率密度函数（Probability Density Function，PDF）：

$$f_x(\alpha) = \frac{\mathrm{d}}{\mathrm{d}\alpha} F_x(\alpha) \tag{2}$$

即 PDF 是 CDF 的微分，由微积分基本定理可以知道，对 PDF 从 α_1 到 α_2 的积分（即这一段曲线下的面积）就等于随机变量 x 取值落在这一段上的概率：

$$\begin{aligned} \int_{\alpha_1}^{\alpha_2} f_x(\alpha)\,\mathrm{d}\alpha &= F_x(\alpha_2) - F_x(\alpha_1) \\ &= \Pr\{x \le \alpha_2\} - \Pr\{x \le \alpha_1\} \\ &= \Pr\{\alpha_1 < x \le \alpha_2\} \end{aligned} \tag{3}$$

同样地，PDF 包含了一个随机变量的分布的所有信息。

2.3. Joint Distribution

联合分布是关于多个随机变量协同分布关系的描述。例如对于两个随机变量 x 和 y，定义联合累积密度函数为：

$$F_{x,y}(\alpha, \beta) = \Pr\{x \le \alpha, y \le \beta\} \tag{4}$$

非常直白，就是同时对两个随机变量做出限制。其联合概率密度函数为：

$$f_{x,y}(\alpha, \beta) = \frac{\partial^2}{\partial\alpha\partial\beta} F_{x,y}(\alpha, \beta) \tag{5}$$

2.4. Mathematic Expectation

数学期望可以认为是进行无穷多次随机试验后，将所得结果取平均后的值。从定义上讲，是所有可能结果用其概率加权的平均值，例如连续随机变量 x 的期望是：

$$E(x) = \int_{-\infty}^{\infty} \alpha f_x(\alpha)\,\mathrm{d}\alpha \tag{6}$$

若是离散随机变量则更清晰一些（α 为任意可能取值）：

$$E(x) = \sum_{\alpha} \alpha f_x(\alpha) \tag{7}$$

很简单，但举一个不是求某个随机变量期望的例子，如随机变量 x 的方差被定义为 $E\{(x - E(x))^2\}$，其积分式应为：

$$E\{(x - E(x))^2\} = \int_{-\infty}^{\infty} (\alpha - E(x))^2 f_x(\alpha)\,\mathrm{d}\alpha \tag{8}$$

这里用的分布函数仍然是 $f_x(\alpha)$ 而不是类似 $f_x((\alpha - E(x))^2)$ 或者 $f_{(x-E(x))^2}(\alpha)$ 的什么东西——我想表达的是，我们应该给 E 加一个下标，表示是对哪个随机变量的分布求期望：

$$E_x(x) = \int_{-\infty}^{\infty} \alpha f_x(\alpha)\,\mathrm{d}\alpha \tag{9}$$

因为期望的求解是需要两个要素的：一是分布，平均的过程是在这个分布上进行的；二是表达式，也就是括号里的东西。期望的定义式中，分布函数 $f_x(\alpha)$ 下面的 x 是来自于 E_x 的 x，代表的是第一要素的分布，而不是 $E(x)$ 括号里表达式的 x；而分布函数前乘的表示取值的 α，是将参数 α 带入表达式中的 x 后得到的表达式，详见例子 Equation 10。

这其实是一个简单的问题，但有时没转过弯就可能在这里卡一下，因为期望的符号中对分布的描述往往是被省略了，默认认为是表达式中存在的随机变量的分布，这里仅作辨析提醒。

对于存在多个随机变量的情况，我们更需要明确到底是求什么分布下的期望，例如关于两个随机变量的表达式的期望，积分应改为二重积分以遍历所有取值组合，分布函数也应改为联合概率密度函数，例如：

$$E_{x,y}(xy^*) = \int_{\alpha=-\infty}^{\infty}\int_{\beta=-\infty}^{\infty} \alpha\beta^* f_{x,y}(\alpha,\beta)\,\mathrm{d}\alpha\,\mathrm{d}\beta \tag{10}$$

2.5. Moments, Mean and Variance

接下来就是基于上述内容纯定义性质的一些统计量了。

首先定义随机变量 x 的 **k 阶原点矩**为 $E\big(x^k\big)$，而一阶原点矩就是 x 的均值（Mean），记为 $m_x = E(x)$。

再定义 **k 阶中心矩**为 $E\big\{(x-E(x))^k\big\}$，可见二阶中心矩就是 x 的方差，记为 $\sigma_x^2 = \mathrm{Var}(x)$，$\sigma_x$ 称为标准差，为方差的平方根。

其他一些统计量及其对应的物理含义就不赘述了。

2.6. Independence

独立性描述两个随机变量是否互不影响。如果两个随机变量 x 和 y 独立，那就说明 x 取某个值 α 的概率乘以 y 取某个值 β 的概率相乘就能直接得到它们分别取这两个值的概率，这是因为独立事件的概率相乘等于它们同时发生的概率，参考乘法原理。

对所有可能取值都满足这一条件，也就意味着二者的分布相乘就可以直接得到联合分布，即独立性的充要条件：

$$f_{x,y}(\alpha,\beta) = f_x(\alpha)f_y(\beta) \tag{11}$$

2.7. Covariance Function

协方差（Covariance）用于衡量随机变量之间的相关程度，两个随机变量 x 和 y 的协方差定义为：

$$c_{xy} = \mathrm{Cov}(x,y) = E\Big\{(x-m_x)\big(y-m_y\big)^*\Big\} \tag{12}$$

顺便，注意到 $c_{xx} = E\Big\{\big(x-m_x\big)^2\Big\}$ 就是 x 的方差。

可以认为，协方差衡量的是随机变量之间的线性关系显著程度。具体地，从定义上看它可以理解为：当 x 高于均值时，y 倾向于同时高于（或低于）自己均值的程度。这个程度

越大，通俗地说就是说明 x 越大时 y 也倾向于越大（或越小），二者的线性关系就较强。因为减去均值免去了其影响，它衡量的就主要是变量的相对变化趋势。

2.8. Correlation Function

相关函数（Correlation Function）定义为两个随机变量内积的期望：

$$r_{xy} = E\{xy^*\} \tag{13}$$

它可以看作未去除均值的协方差函数。可以推导二者关系为：

$$\begin{aligned} c_{xy} &= E\Big\{(x - m_x)\big(y - m_y\big)^*\Big\} \\ &= E\{xy^*\} + E\big\{m_x m_y^*\big\} - m_x E\{y^*\} - E\{x\}m_y^* \\ &= E\{xy^*\} + m_x m_y^* - m_x m_y^* - m_x m_y^* \\ &= r_{xy} - m_x m_y^* \end{aligned} \tag{14}$$

它们之间只是相差一个常数 $m_x m_y^*$，所以它们表达的物理意义可以是类似的。

2.9. Correlation Coefficient and Orthogonality

当然，协方差没有去除量纲，在不同的随机变量组合中不具备普遍性。通常我们利用归一化等手段定义相关系数来衡量相关性（Correlation），例如最常用的皮尔逊相关系数：

$$\rho_{xy} = \frac{c_{xy}}{\sigma_x \sigma_y} \tag{15}$$

可以证明其取值范围为 $[-1, 1]$，用以统一地衡量随机变量间线性关系的显著程度。这个相关系数如果为零，就称两个随机分布是正交（Orthogonal）的。

当相关系数为 0 时，也即 $c_{xy} = 0$ 时，代表两个随机变量不相关。相关系数为正时代表正相关，为负时代表负相关。可以推得不相关的充要条件：

$$c_{xy} = 0 \quad \Rightarrow \quad r_{xy} - m_x m_y^* = 0 \quad \Rightarrow \quad E\{xy^*\} = E\{x\}E^*\{y\} \tag{16}$$

即随机变量相乘的期望等于随机变量的期望相乘。

注意，相关性和独立性不是一回事。

首先，不相关不一定代表独立。例如，两个随机变量满足 $y = x^2$ 时，虽然二者完全不独立，但相关系数仍然是 0。如前所述，相关性主要衡量的是线性关系的显著程度，在这个例子中，x 为负时 y 随 x 增长而减小，x 为正时 y 随 x 增长而增长，对称地抵消了两部分关系，结果上看没有线性关系的成分，但其实存在二次关系。

但独立就一定不相关，因为独立时两个随机变量的分布完全没有关联，各自的密度函数直接相乘就能得到联合分布密度函数，由定义也可推得 $E\{xy^*\} = E\{x\}E^*\{y\}$。

总结来说就是，独立 $\Rightarrow$ 不相关，不是充要关系。

3. Random Process

一个随机过程 $\{x(n)\}$ 实际就是一串随机变量拼成的序列。

这些随机变量之间可以是独立同分布（i.i.d）的，例如白噪声，但也可以不是，或者说现实世界里大部分都不是；这个序列的索引可以代表时间，也可以不是，方便起见我们后面默认在研究时间序列，索引 n 代表不同的时间点。

同前，对随机性的研究重点在于研究变化中不变的东西，如数据的统计特征，故我们还是从提供一系列统计特征的定义开始。

3.1. Statistic

3.1.1. Mean

不是说随机过程会有一个统一的均值，这个均值还是分别定义给每个随机变量的，得到的是一个序列：

$$m_x(n) = E\{x(n)\} \tag{17}$$

3.1.2. Auto-correlation

对于一个随机过程 $x(n)$，其自相关函数即为指定两个索引 k 和 l 的随机变量的相关函数：

$$r_x(k,l) = r_{x(k),x(l)} = E\{x(k)x^*(l)\} \tag{18}$$

就该公式来看，这里的每一个 r_x 只是和两个随机变量有关，和随机过程整体没有什么关联，不过后续介绍宽平稳（WSS）的时候就有别的说法了。

3.1.3. Auto-covariance

类比协方差定义自协方差：

$$c_x(k,l) = E\big\{[x(k) - m_x(k)][x(l) - m_x(l)]^*\big\} = r_x(k,l) - m_x(k)m_x^*(l) \tag{19}$$

3.1.4. Cross-correlation

互相关则是关于两个随机过程的，例如 $\{x(n)\}$ 和 $\{y(n)\}$ 的互相关函数定义为：

$$r_{xy}(k,l) = r_{x(k),y(l)} = E\{x(k)y^*(l)\} \tag{20}$$

3.1.5. Cross-covariance

同样地，互协方差定义为：

$$c_{xy}(k,l) = E\Big\{[x(k) - m_x(k)]\big[y(l) - m_y(l)\big]^*\Big\} = r_{xy}(k,l) - m_x(k)m_y^*(l) \tag{21}$$

这些后面其实用不太到，主要是列举一下。

3.2. Structural Invariance

可以注意到，前面定义的统计量主要还是关注分立的随机变量，而与随机过程的整体没有太大关联；而接下来我们要来关注整个随机过程的结构上的一些不变性，这些性质将会在分布未知而只能通过有限的样本进行估计时提供合理的依据。

具体地，随机变量是理想的，拥有一个概率密度函数来描述它的分布；但我们往往研究具体的信号而不知其真实分布，只能借助样本反过来估计真实分布的情况。我们可以将一个具体的信号 $x[n]$ 视为一个随机过程的一次实现（Realization），其每个时间点的样本都来自对这个随机过程中对应时间随机变量的一次采样。Figure 1 展示了对一个随机过程进行大量试验后得到的一系列可能实现，高亮的为其中一条。

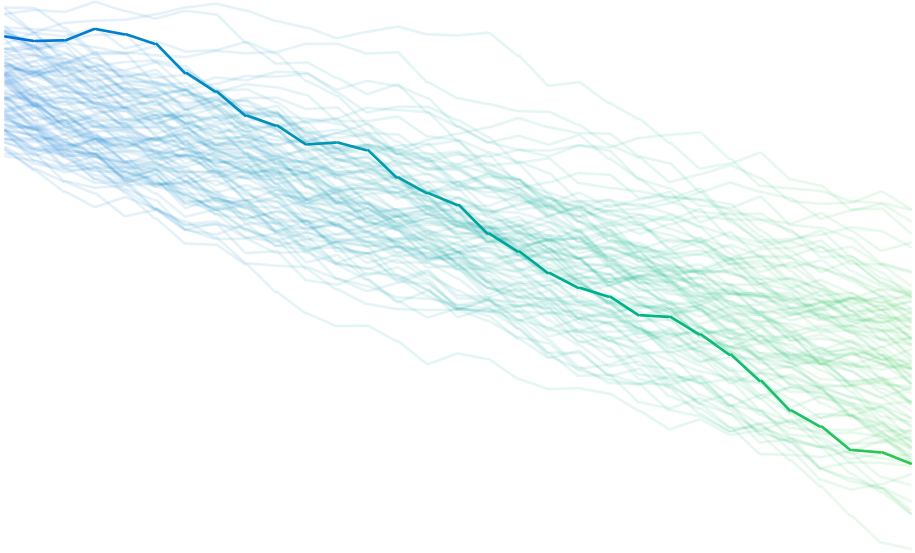

Figure 1: Different realizations of a random process

3.2.1. Stationarity

我们关注的**第一个问题**是：在一个随机过程中，一些统计特征会不会随时间而改变？

我们引入平稳性（Stationarity）的概念，主要思想是衡量任意一段子序列在延迟任意一段时间后，其统计特征是否还维持不变。要使统计特征一致，最简单的就是直接令分布完全一致，由此定义强平稳（Strict-Sense Stationary，SSS）：

$$\forall k, \{n_1, n_2, ..., n_m\}, f_{x(n_1),x(n_2),...,x(n_m)}(\cdot) = f_{x(n_1+k),x(n_2+k),...,x(n_m+k)}(\cdot) \tag{22}$$

强平稳的要求太高，大部分情况下我们也不需要如此强的条件，故我们定义宽平稳（Wide-Sense Stationary，WSS），只关注一阶和二阶统计量的一致性：

1. $$m_x(n) = m_x \tag{23}$$
2. $$r_x(k, l) = r_x(k - l) \tag{24}$$
3. $$c_x(0) < \infty \tag{25}$$

即对于宽平稳过程：均值与时间无关，自相关函数只与时间差有关而与绝对时间无关。

上述条件的第三条，即方差有限，在许多材料中会被省略。

方差有限等价于均方值有限，二者就差一个均值的平方，物理意义上可以认为它代表功率有限，而工程领域常忽略这一条件是由于大部分实际物理过程功率都有限。

> 数学上，这一条确保了二阶矩的存在。不过实际上第二条件中要求 $r_x(0)$ 存在，而期望的存在性隐含了其收敛到有限值的含义，已经隐含了第三条件。

这使得我们在一条实现上截取不同时间的子序列，其平均性质是一致的。在信号平稳的情况下，我们可以将一条够长的样本劈开劈成几份，并声称这劈出的几份背后隐含着一致的统计特征。

3.2.2. Ergodicity

第一个问题引申出的平稳性让我们可以去劈开一条足够长的样本当作不同的样本来用，那么我们关注的**第二个问题**是：一条足够长的实现能否代表无穷多条可能实现的总体（样本空间）？也就是我拿这条长样本劈开来去估计分布的，但它估计出来的是我要的这个随机过程的分布吗？它一条样本能代表这个随机过程包含的所有方面的信息吗？

我们用遍历性（Ergodicity）来描述这个性质，它是建立在平稳性的基础上的，即要保证统计特征不随时间推移变化，否则每个部分性质不一样，有足够长的样本也是白搭。

首先，我们需要举一个“平稳但不遍历”的典型例子来说明这个概念存在的必要性。令随机过程 $x(n)$ 中仅 $x(0)$ 是一个一次性抽取的随机变量 z，之后都有 $x(n)=x(n-1)$。这个例子称为随机常数，这是由于该随机过程的所有实现都是一条常值信号，但不同实现中这个常数值不一样，取决于一开始随机的数值。

分析这个过程，首先它一定是平稳的，因为时间变化显然不影响其分布。也可以具体一些证明其满足 WSS 的条件：考察均值，$m_x=E\{x(0)\}$ 与 n 无关；考察自相关函数 $r_x(k,l)=E\{x(k)x^*(l)\}=E\{x(0)x^*(0)\}=\sigma_x^2$，与什么都无关。

然后考察遍历性，注意到样本总体的均值为 $E\{z\}$，但对于一个足够长的实现，其时间平均为该实现中 $x(0)$ 抽到的值，时间再长也是这个值，而这个值不一定等于 $E\{z\}$。换句话说，我们没有办法在这种情况下仅通过一条实现来估计出 z 的性质，即便它足够长。

总结来说，平稳性保证了我们可以用一条足够长的样本得到有意义的时间平均，而遍历性则保证了这个时间平均等于所有可能样本的总体平均，即这样估出来的值是正确的。

> 物理学和统计力学中的相同概念往往翻译为“各态历经性”，这其实更为形象。它表示随着时间的推移，系统将会经历并表现出其所有可能的宏观状态。只有单次实现中有表现出分布所有方面特征的可能性，我们才有机会用单条样本来估计宏观分布。
>
> 我们也可以从偏估计和无偏估计的角度去看待这个问题。对于平稳信号我们可以去估计出一个有意义的统计量，但并不能保证估计是否无偏，此时满足遍历性就可以保证这一点。

严格的遍历性定义比较抽象和复杂，所以类似 WSS 的定义，我们也仅考虑在均值和相关性上考察遍历性，就够满足我们的需要了。当然，以下讨论建立在至少是 WSS 的基础上。

首先是**均值遍历性**（Mean-Ergodicity），我们定义样本均值，也就是一条样本在时间上的平均：

$$\hat{m}_x^{(N)} = \frac{1}{N}\sum_{n=0}^{N-1} x[n] \tag{26}$$

考虑原始定义，即时间平均（样本均值）等于总体平均：

$$\lim_{N\to\infty} E\left\{\left|\hat{m}_x^{(N)} - m_x\right|^2\right\} = 0 \tag{27}$$

不过我们没法真的从真实分布采样再去验证这个条件，所以我们提供一个均值遍历成立的**充分必要**条件（暂不考虑证明）：

$$\lim_{N\to\infty} \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} c_x(n-m) = 0 \tag{28}$$

即自协方差函数衰减得足够快，一个偏直觉的解读是，这种情况下随机变量之间的相关性不会在时间上滞留太久，长期平均可以将前后的关联稀释掉。

还有一个**充分**条件用于简单情况下的快速判断：

$$c_x(0) < \infty \quad \text{and} \quad \lim_{k\to\infty} c_x(k) = 0 \tag{29}$$

或者更强一些的条件，自方差函数绝对可和也是均值遍历的**充分**条件：

$$\sum_{k=-\infty}^{\infty} |c_x(k)| < \infty \tag{30}$$

而后是**相关遍历性**（Correlation-Ergodicity），同样先定义样本自相关函数。因为 WSS 过程的自相关函数仅与时间差有关，所有样本中所有时间差为 k 的样本对都可以用来估计自相关函数的第 k 项，我们直接都加起来取个平均：

$$\hat{r}_x^{(N)}(k) = \frac{1}{N}\sum_{n=0}^{N-1-k} x[n+k]x^*[n] \tag{31}$$

> 你可能会注意到该样本自相关的定义中，总共用了 $N-k$ 组样本对相关结果的加和，但除以的却是 N 而非 $N-k$。
>
> 这实际上确实会导致其值的估计是有偏的，但如果只考虑此处只用于相关遍历性的定义中，其 N 将趋于无穷，使得 k 的影响可以忽略。也就是说，这里定义中我们只需要它满足渐近无偏即可。
>
> 在 Section 9.1.1 中我们将再次讨论实际用这里的样本自相关函数去估计自相关值，以及除以 N 而不是除以 $N-k$ 的解释。

同样考虑时间平均等于总体平均得到定义：

$$\lim_{N\to\infty} E\left\{\left|\hat{r}_x^{(N)}(k) - r_x(k)\right|^2\right\} = 0 \tag{32}$$

同样这只是最朴素的式子，我们不证明地给出实用的**充分必要**条件：

$$\forall k, \lim_{N\to\infty} \frac{1}{N^2}\sum_{n=0}^{N-1}\sum_{m=0}^{N-1} \left|r_x(n-m) - r_x(n-m+k)\right|^2 = 0 \tag{33}$$

此外，绝对可和条件 Equation 30 也是相关遍历的**充分**条件。

3.3. Power Spectrum

（TODO）补充关于能量信号、功率信号、功率谱密度（PSD）。

由 Wiener–Khinchin Theorem，宽平稳（WSS）过程的功率谱密度是其自相关函数的傅里叶变换。

3.4. Filtering Random Processes

（TODO）

若滤波器系数 $h[n]$ 有限长且在 $[0, N-1]$ 外皆为零，则输出过程的功率可以用输入过程 $x(n)$ 的自相关矩阵和滤波器系数向量来表示：

$$\sigma_y^2 = E\{|y(n)|^2\} = \boldsymbol{h}^H \boldsymbol{R}_x \boldsymbol{h} \tag{34}$$

3.5. Random Process and Digital Signals

我们的随机过程 $x(n)$ 是分布，而我们的数字信号 $x[n]$ 可以认为是从中采样得到的某个实现。我们为随机过程定义了均值、自相关等特征，也可以为数字信号定义均值和自相关函数等运算，我们现在希望考察的就是双方的区别与联系。

澄清这一点之后，我们就可以清晰的了解该使用什么手段借助样本估计分布特征了，注意这是我们自始至终不变的目标。我们选取一些常用角度来展开这个话题。

3.5.1. Auto-correlation Estimation with Multiple Realizations

如果我们有大量对随机过程的采样结果（实现），我们就可以直接通过定义（Equation 18）来估计其分布特征，如自相关函数。

对随机过程 $\{x(n)\}$ 进行 N 次实现，每次长度为 M，第 i 次的结果记为 $x_i[n]$，或者用向量表示（当然 M 也可以是无穷大，这里设为有限长度以便展示）：

$$\boldsymbol{x}_i = \begin{pmatrix} x_i[0] \\ x_i[1] \\ \vdots \\ x_i[M-1] \end{pmatrix} \tag{35}$$

多次采样的数据我们给它拼成矩阵形式：

$$X = [\boldsymbol{x}_0 \quad \boldsymbol{x}_1 \quad ... \quad \boldsymbol{x}_{N-1}] = \begin{bmatrix} x_0[0] & x_1[0] & ... & x_{N-1}[0] \\ x_0[1] & x_1[1] & ... & x_{N-1}[1] \\ \vdots & \vdots & \ddots & \vdots \\ x_0[M-1] & x_1[M-1] & ... & x_{N-1}[M-1] \end{bmatrix} \tag{36}$$

由定义，$r_x(k,l) = E\{x(k)x^*(l)\}$，这里的 E 就是在总体上求期望，所以我们直接用不同实现中的 $x_i[k]$ 和 $x_i[l]$ 来估计它即可：

$$\hat{r}_x(k,l) = \frac{1}{N}\sum_{i=0}^{N-1} x_i[k]x_i^*[l] \tag{37}$$

我们把估计自相关函数的值也写成矩阵，其可以通过 X 运算得来，从而得到简洁矩阵表达：

$$\begin{aligned} \hat{\boldsymbol{R}}_x &:= \begin{bmatrix} \hat{r}_x(0,0) & \hat{r}_x(0,1) & ... & \hat{r}_x(0,M-1) \\ \hat{r}_x(1,0) & \hat{r}_x(1,1) & ... & \hat{r}_x(1,M-1) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_x(M-1,0) & \hat{r}_x(M-1,1) & ... & \hat{r}_x(M-1,M-1) \end{bmatrix} \\ &= \frac{1}{N}\sum_{i=0}^{N-1} \begin{bmatrix} x_i[0]x_i^*[0] & x_i[0]x_i^*[1] & ... & x_i[0]x_i^*[M-1] \\ x_i[1]x_i^*[0] & x_i[1]x_i^*[1] & ... & x_i[1]x_i^*[M-1] \\ \vdots & \vdots & \ddots & \vdots \\ x_i[M-1]x_i^*[0] & x_i[M-1]x_i^*[1] & ... & x_i[M-1]x_i^*[M-1] \end{bmatrix} \\ &= \begin{bmatrix} x_0[0] & x_1[0] & ... & x_{N-1}[0] \\ x_0[1] & x_1[1] & ... & x_{N-1}[1] \\ \vdots & \vdots & \ddots & \vdots \\ x_0[M-1] & x_1[M-1] & ... & x_{N-1}[M-1] \end{bmatrix} \begin{bmatrix} x_0[0] & x_0[1] & ... & x_0[M-1] \\ x_1[0] & x_1[1] & ... & x_1[M-1] \\ \vdots & \vdots & \ddots & \vdots \\ x_{N-1}[0] & x_{N-1}[1] & ... & x_{N-1}[M-1] \end{bmatrix}^* \\ &= \frac{1}{N}\boldsymbol{X}\boldsymbol{X}^H \end{aligned} \tag{38}$$

（TODO）没有那么多独立实现的情况下，把 $\boldsymbol{x}_i$ 定义成 $[x[i], ..., x[i+L-1]]^T$。可参考 Slides Lec11 P6。

3.5.2. Auto-correlation Estimation with Correlation-Ergodicity

由上节可见，样本（实现）的数量 N 越大，估计就越准确。但若我们只有一条样本（$N = 1$），采用这种方法的估计就将极不精确。此时，如果随机过程遍历性成立，我们就允许使用这条样本不同时间上的信息来进行估计。

我们定义有限长数字信号 $x[n]$（长度为 N，注意这里的 N 不是上面的实现数量）的自相关函数为：

$$R_{xx}(k) = \sum_{n=0}^{N-1-k} x[n+k]x^*[n] \tag{39}$$

这从公式定义上可以理解为是对一个信号在不同时间延迟下与自身相似程度的衡量。

实际上，只要能表达这个含义即可称为自相关，故不同的信号自相关定义有很多种，只是存在细节上的差别，例如是否除以样本数量（取了一个平均）、是延迟 k 还是提前 k（结果相当于对称了一下）等，哪个方便用哪个即可。

这里我们选取该定义是因为 MATLAB 的互相关函数 xcorr 是按这种方式定义的，文档原文是这么写的[1]：

$$\hat{R}_{xy}(m) = \begin{cases} \sum_{n=0}^{N-m-1} x_{n+m} y_n^*, & m \geq 0, \\ \hat{R}_{yx}^*(-m), & m < 0. \end{cases} \tag{40}$$

回忆 Equation 31 定义的样本自相关函数 $\hat{r}_x^{(N)}(k)$，刚巧和我们定义的信号自相关函数 $R_{xx}(k)$ 形式几乎一致，只差一个系数。

根据前文的定义，如果过程是相关遍历的，那么这个样本自相关函数 $\hat{r}_x^{(N)}(k)$ 就可以用来正确地估计随机分布的自相关函数 $r_x(k)$。由此，与其形式几乎一致的信号互相关函数 $\hat{R}_{xy}(m)$ 也就具有同样的物理意义，即也可以用于估计 $r_x(k)$（需要有系数上的调整）。

总结来说，相关遍历性的成立令我们可以在实现数量不足的情况下，使用其样本的信号自相关函数来估计随机过程的自相关特征。

3.5.3. Auto-correlation Estimation (Comprehensive)

如果我们有多个的实现（样本）可用，同时遍历性还成立，那么就可以综合两种优势进行估计。

具体地，Section 3.5.1 节中我们估计的是 $r_x(k, l)$，如果遍历性成立（平稳性自然也成立），则自相关函数只与时间差有关，那么我们就可以用所有的 $\hat{r}_x(n, n+k)$（$\hat{\boldsymbol{R}}_x$ 中处在同一条斜线上的值）来估计 $r_x(k)$。

[1] 参考 https://nl.mathworks.com/help/matlab/ref/xcorr.html#mw_01b546db-b642-4f02-8625-16078810d80f

4. Digital Signal Processing

（TODO）主要 DTFT、z 变换、频域特性、稳定性、功率、能量等。

4.1. Spectral Analysis

（TODO）非周期信号的频谱连续，离散信号的频谱周期；迪利克雷核，加窗，DFT；zero-padding 提高密度；zero-crossing 与分辨率；谐波高度与分辨率，其他类型窗；WSS 随机信号直接变换，平均 periodogram，bpsk 例子收敛到迪利克雷核。

对于一个较长的、不满足平稳性假设的信号，我们分析它全时间上的频谱意义不大，因为它在随时间变化。所以我们通常会将信号切成小段来分析，切段的方式是直接截断，假设段外的函数值都为 0。我们定义 Dirichlet 核函数：

$$w_R[n] = \begin{cases} 1, & 0 \le n \le N-1 \\ 0, & \text{otherwise} \end{cases} \tag{41}$$

截断后的信号为：

$$x_N[n] = x[n] w_R[n] \tag{42}$$

时域上的乘积与频域上的卷积相关，具体地：

$$X_N(\omega) = \frac{1}{2\pi} \{X * W_R\}(\omega) \tag{43}$$

从图表和形象的角度考察这一操作：

（TODO）W_R 的频谱，频谱搬移，将冲激函数换为带宽度的峰的直观。

5. Optimization

（TODO）主要关于最小二乘法、拉格朗日乘子法等。

（TODO）要写一下复变函数变量分成自己和其共轭、求解时对共轭求偏导的原因。

6. Signal Modelling

实际上，建模（Modelling）可以视为对信号的压缩（Compressing）过程。一个参数化的模型可以使用比信号样本数更少的参数数量来表示复杂的信号，实现更高效的存储和传输。

压缩得到的参数也可以视为对信号本质特征的描述，例如蕴含其背后的物理规律等，这就允许我们利用模型对信号的未知部分进行预测（Prediction），或称外推（Extrapolation）。

现在，我们的目标是对给定的数字信号 $x[n]$ 进行建模，即找到一个模型 $H(z)$ 使其输出信号 $\hat{x}[n]$ 能够尽可能接近目标信号 $x[n]$。

6.1. Autoregressive and Moving Average (ARMA) Model

我们可以根据实际情况使用不同种类的模型来对信号建模，这里以时间序列分析常用的自回归滑动平均模型（Autoregressive and Moving Average Model）即 $\mathrm{ARMA}(p,q)$ 为例。其传递函数定义如下：

$$\frac{Y(z)}{X(z)}=H(z)=\frac{\sum_{k=0}^{q}b[k]z^{-k}}{1+\sum_{k=1}^{p}a[k]z^{-k}}=\frac{B(z)}{A(z)} \tag{44}$$

设输入和输出信号分别为 $x[n]$ 和 $y[n]$，$X(z)$ 和 $Y(z)$ 为其对应的拉普拉斯变换函数。这里形式上虽然 $a[n]$ 的索引从 1 开始，但实际上可以取 $a[0]=1$ 以得到更统一的形式。于是由定义我们有 $Y(z)A(z)=X(z)B(z)$，变换到时域即：

$$a[n]*y[n]=b[n]*x[n] \tag{45}$$

展开得到经典的线性常系数差分方程（Linear Constant Coefficient Difference Equation，LCCDE）的形式：

$$y[n]+\sum_{k=1}^{p}a[k]y[n-k]=\sum_{k=0}^{q}b[k]x[n-k] \tag{46}$$

写详细一点为：

$$\begin{aligned}&y[n]+a[1]y[n-1]+...+a[p]y[n-p]\\=\ &b[0]x[n]+b[1]x[n-1]+...+b[q]x[n-q]\end{aligned} \tag{47}$$

显然，这个系统是典型的线性移不变（Linear Shift-Invariant，LSI）系统，具有良好的性质。

6.1.1. Autoregressive (AR) Model

若只有自回归的部分，即 $\mathrm{AR}(p)=\mathrm{ARMA}(p,0)$：

$$H(z)=\frac{Y(z)}{X(z)}=\frac{b[0]}{1+\sum_{k=1}^{p}a[k]z^{-k}} \tag{48}$$

在时域即：

$$y[n]+\sum_{k=1}^{p}a[k]y[n-k]=b[0]x[n] \tag{49}$$

写详细一点为：

$$a[0]y[n] + a[1]y[n-1] + ... + a[p]y[n-p] = b[0]x[n] \tag{50}$$

该模型认为当前时刻的输出 $y[n]$ 是前 p 个时刻的输出 $y[n-1], ..., y[n-p]$ 以及当前时刻的输入 $x[n]$ 的线性组合，所以称为自回归模型。由于该模型无零点，故也被称为全极点模型（All-Pole Model）。

6.1.2. Moving Average (MA) Model

若只有滑动平均的部分，即 $\mathrm{MA}(q) = \mathrm{ARMA}(0, q)$：

$$H(z) = \frac{Y(z)}{X(z)} = \sum_{k=0}^{q} b[k] z^{-k} \tag{51}$$

在时域即：

$$y[n] = \sum_{k=0}^{q} b[k] x[n-k] \tag{52}$$

写详细一点为：

$$y[n] = b[0]x[n] + b[1]x[n-1] + ... + b[q]x[n-q] \tag{53}$$

该模型认为当前时刻的输出 $y[n]$ 是前 q 个时刻的输入 $x[n], x[n-1], ..., x[n-q]$ 的线性组合，所以称为滑动平均模型。

6.2. Signal Models

作为一个离散时间系统，$H(z)$ 并不能直接表示信号，而是需要接受一个输入以获得输出。我们将输出信号作为模型对目标信号的估计 $\hat{x}[n]$，将输入信号 $x[n]$ 封装为模型的一部分。由于该模型无极点，故也被称为全零点模型（All-Zero Model）。

6.2.1. Deterministic Modelling

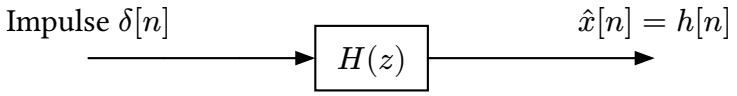

Figure 2: Signal model with deterministic input

我们可以选择一个已知的、确定的信号，将其固定为系统的输入，稳定得到我们想要的输出信号 $\hat{x}[n]$，使其值尽可能接近目标信号 $x[n]$，用于对确定的信号进行建模，称为 Deterministic Modelling。

这个输入信号可以根据实际情况进行选择，符合目标信号的特征的输入信号有时可以减轻模型拟合的负担。在这里我们可以选择使用最简单的单位脉冲信号 $\delta[n]$ 作为输入信号，这使得系统的输出信号 $\hat{x}[n]$ 即为系统的单位冲激响应 $h[n]$。

6.2.2. Stochastic Modelling

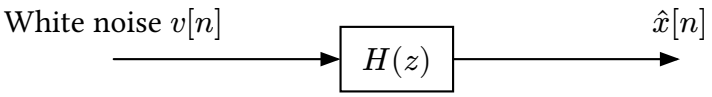

Figure 3: Signal model with stochastic input

我们还可以选择使用一个已知分布的随机噪声作为输入，得到输出信号 $\hat{x}[n]$ 使其统计特征（例如均值、自相关函数）与目标信号 $x[n]$ 一致，从而对随机过程进行建模，称为 Stochastic Modelling。

我们可以选择使用均值为 0、方差为 σ_v^2 的白噪声 $v[n]$ 作为输入信号。这样做的依据是其自相关函数为 $r_v[k] = \sigma_v^2 \delta[k]$，对其进行傅里叶变换得其功率谱密度为常数 $P_v(\omega) = \sigma_v^2$，即在所有频率上均有相同的能量分布。

这样的特性保证了我们可以通过对其进行滤波得到任意频率特性的输出信号 $\hat{x}[n]$，同时任意频率成分能量均匀，使得输出信号的统计特征与输入信号无关。

7. Deterministic Modelling Identification

模型建立后我们还需要对其进行参数辨识，即确定参数序列 $a[k]$ 和 $b[k]$ 的值。选取参数的目标是使模型输出 $\hat{x}[n]$ 尽可能接近目标信号 $x[n]$。

7.1. Least Squares (LS) Method

首先讨论 Deterministic Modelling，我们希望模型输出 $\hat{x}[n]$ 能够精确地重现目标信号 $x[n]$，即每个采样点的值都要尽可能接近。定义误差信号 $e'[n]=x[n]-\hat{x}[n]$，可以通过最小化均方误差 $\mathcal{E}_{\mathrm{LS}}=\sum_{n=0}^{\infty}|e'[n]|^2$ 来确定模型参数，如 Figure 4 所示。

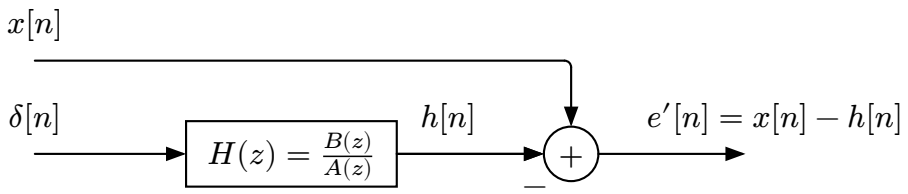

Figure 4: Diagram for deterministic model identification (intractable)

该优化问题可通过求解如下方程组解决（对变量的共轭求偏导的原因详见 Section 5）：

$$\begin{cases}\frac{\partial\mathcal{E}_{\mathrm{LS}}}{\partial a^*[k]}=0, & k=1,2,...,p\\ \frac{\partial\mathcal{E}_{\mathrm{LS}}}{\partial b^*[k]}=0, & k=0,1,...,q\end{cases}\tag{54}$$

但这个方程组是非线性的，求解起来非常复杂。

7.2. Padé Approximation

注意到我们使用的 ARMA 模型具有 $p+q+1$ 个参数，也就是说它拥有 $p+q+1$ 个自由度，所以理论上我们是可以用其完美拟合信号的前 $p+q+1$ 个样本的，我们先考虑这个任务。

我们对传递函数的形式作一个转化：

$$H(z)=\frac{B(z)}{A(z)}\quad\Rightarrow\quad H(z)A(z)=B(z)\quad\Rightarrow\quad h[n]*a[n]=b[n]\tag{55}$$

将卷积展开得到：

$$h[n]+\sum_{k=1}^{p}a[k]h[n-k]=b[n]\tag{56}$$

对于单位脉冲输入 $\delta[n]$，系统的输出 $h[n]$ 就是我们估计的结果 $\hat{x}[n]$。若要完全拟合前 $p+q+1$ 个样本，则直接代入 $h[n]=x[n],0\le n\le p+q$，得到：

$$x[n]+\sum_{k=1}^{p}a[k]x[n-k]=\begin{cases}b[n], & n=0,1,...,q\\ 0, & n=q+1,q+2,...,q+p\end{cases}\tag{57}$$

这就是我们想要的线性方程组了，包含了 $a[\cdot]$、$b[\cdot]$ 和已知的常数 $x[n]$。

上述转化过程体现在系统框图上就是在两路上同乘 $H(z)$ 的分母 $A(z)$。即令新的目标误差 $E(z)=A(z)E'(z)=A(z)X(z)-B(z)$，如 Figure 5 所示。

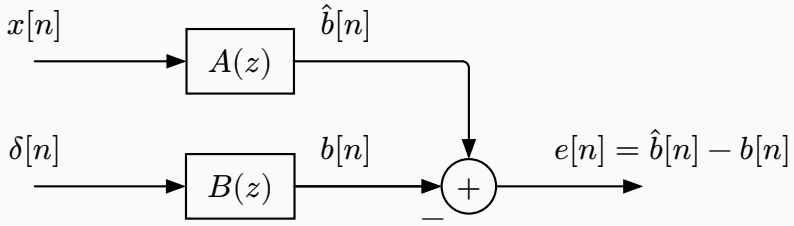

Figure 5: Diagram for deterministic model identification

单位脉冲信号经过 $B(z)$ 得到的输出就是 $b[n]$，将其与用目标信号 $x[n]$ 经过 $A(z)$ 得到的 $\hat{b}[n]$（此时可视为是对系数序列 $b[n]$ 的估计）作差，得到新的误差。

经过这样的操作后的列出的方程就是线性的了。

接下来是对 Equation 57 的求解。该方程组包含相同个数的方程和未知数，非奇异的话恰好可以求解出唯一解。清晰一点，我们把矩阵画出来：

$$\begin{bmatrix} \overbrace{\begin{matrix} x[0] & 0 & 0 & \dots & 0 \\ x[1] & x[0] & 0 & \dots & 0 \\ x[2] & x[1] & x[0] & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[p] & x[p-1] & x[p-2] & \dots & x[0] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[q] & x[q-1] & x[q-2] & \dots & x[q-p] \end{matrix}}^{\boldsymbol{X}_0} \\ \underbrace{\begin{matrix} x[q+1] \\ x[q+2] \\ \vdots \\ x[q+p] \end{matrix}}_{\boldsymbol{x}_{q+1}} \quad \underbrace{\begin{matrix} x[q] & x[q-1] & \dots & x[q-p+1] \\ x[q+1] & x[q] & \dots & x[q-p+2] \\ \vdots & \vdots & \ddots & \vdots \\ x[q+p-1] & x[q+p-2] & \dots & x[q] \end{matrix}}_{\boldsymbol{X}_q} \end{bmatrix} \underbrace{\begin{bmatrix} 1 \\ a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix}}_{\bar{\boldsymbol{a}}} = \begin{bmatrix} \overbrace{b[0]}^{b} \\ b[1] \\ b[2] \\ \vdots \\ b[p] \\ \vdots \\ b[q] \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{58}$$

在这里，我们先用下半部分（后 p 行）求解 $\bar{\boldsymbol{a}}$（即 $a[\cdot]$）：

$$\begin{aligned} [\boldsymbol{x}_{q+1} \ \boldsymbol{X}_q]\boldsymbol{a}=\boldsymbol{0} \quad &\Leftrightarrow \quad [\boldsymbol{x}_{q+1} \ \boldsymbol{X}_q]\begin{bmatrix}1\\ \bar{\boldsymbol{a}}\end{bmatrix}=\boldsymbol{0} \\ \Rightarrow \quad \boldsymbol{X}_q\bar{\boldsymbol{a}}=-\boldsymbol{x}_{q+1} \quad &\Rightarrow \quad \bar{\boldsymbol{a}}=-\boldsymbol{X}_q^{-1}\boldsymbol{x}_{q+1} \end{aligned} \tag{59}$$

注意到 $\boldsymbol{X}_q$ 是一个非对称 Toeplitz 矩阵，存在一些更高效的专用方法用于求解其逆矩阵，如 Trench 算法。接下来，代入上半部分（前 $q+1$ 行）即可得 $b[\cdot]$：

$$\boldsymbol{b} = \boldsymbol{X}_0 \begin{bmatrix} 1 \\ \bar{\boldsymbol{a}} \end{bmatrix} \tag{60}$$

Padé 法很直接，但显然也存在一系列问题：

1. 不保证所得系统是稳定的；
2. 只约束了模型输出 $\hat{x}[n]$ 和目标信号 $x[n]$ 的前 $p+q+1$ 个样本相同，此后的匹配效果可能不佳；
3. $\boldsymbol{X}_q$ 可能奇异且无解。

关于 $\boldsymbol{X}_q$ 奇异且无解的情况，可以认为是模型中 $a[0]=1$ 的默认假设存在问题。如果修改模型，令 $a[0]=0$，则方程虽奇异但非无解，而是解不唯一。

如果不是全极点模型，则如此得到的新传递函数在分子和分母中同时具有因子 z，可以约去。这本质上是在零处发生了零极点对消，结果上看等效于模型的阶数下降了，即模型阶数存在冗余。

7.3. Prony's Method

Padé 法将所有的自由度都用在了序列的前 $p+q+1$ 项上，而 Prony 法的想法很简单，就是降低前面这段序列的拟合要求，从而获得一个从信号整体上看更好的拟合。

7.3.1. Prony Normal Equations

我们先从比较形式化的角度来推导 Prony 法。具体地，还是借 Padé 法的想法将问题转化为线性的，如 Figure 5 所示。写出整个信号误差的表达式，而不只是前 $p+q+1$ 项：

$$e[n] = \begin{cases} x[n] + \sum_{k=1}^{p} a[k]x[n-k] - b[n], & n = 0, 1, ..., q \\ x[n] + \sum_{k=1}^{p} a[k]x[n-k], & n > q \end{cases} \tag{61}$$

Prony 法中，我们先通过最小化均方误差的方式求解 $a[\cdot]$：

$$\varepsilon_{p,q} = \sum_{n=q+1}^{\infty} |e[n]|^2 = \sum_{n=q+1}^{\infty} \left| x[n] + \sum_{k=1}^{p} a[k]x[n-k] \right|^2 \tag{62}$$

这里只考虑 $n > q$ 部分的误差是为了令该部分只与 $a[\cdot]$ 有关，这是考虑到分步求解 $a[\cdot]$ 和 $b[\cdot]$ 的需要，确实可能会牺牲一部分定义的准确性，但相对无限长的 $x[n]$ 影响并不大。接下来我们公式化求偏导令其为零计算最优值：

$$\frac{\partial \varepsilon_{p,q}}{\partial a^*[k]} = \sum_{n=q+1}^{\infty} \frac{\partial [e[n]e^*[n]]}{\partial a^*[k]} = \sum_{n=q+1}^{\infty} e[n] \frac{\partial e^*[n]}{\partial a^*[k]} = 0, \quad k = 1, 2, ..., p \tag{63}$$

由定义 Equation 61 我们知道 $\frac{\partial e^*[n]}{\partial a^*[k]} = x^*[n-k]$，代入得：

$$\sum_{n=q+1}^{\infty} e[n]x^*[n-k] = 0, \quad k = 1, 2, ..., p \tag{64}$$

这个式子表达了最小误差和信号间的正交关系，称为 Orthogonality principle。我们继续代入定义 Equation 61 （注意字母 k 用掉了，换用 l，不要混淆）得：

$$\sum_{n=q+1}^{\infty}\left(x[n]+\sum_{l=1}^{p}a[l]x[n-l]\right)x^*[n-k]=0,\quad k=1,2,...,p \tag{65}$$

移项并重排求和符号顺序可以得到：

$$\sum_{l=1}^{p}a[l]\left(\sum_{n=q+1}^{\infty}x^*[n-k]x[n-l]\right)=-\sum_{n=q+1}^{\infty}x^*[n-k]x[n],\quad k=1,2,...,p \tag{66}$$

为了简化表达，我们记：

$$r_x(k,l)\coloneqq\sum_{n=q+1}^{\infty}x^*[n-k]x[n-l] \tag{67}$$

我们可以顺便观察到 $r_x(k,l)=r_x^*(l,k)$。将其代入原式得到：

$$\sum_{l=1}^{p}a[l]r_x(k,l)=-r_x(k,0),\quad k=1,2,...,p \tag{68}$$

这被称为 **Prony normal equations**，写成矩阵形式是：

$$\begin{bmatrix}r_x(1,1) & r_x(1,2) & ... & r_x(1,p)\\ r_x(2,1) & r_x(2,2) & ... & r_x(2,p)\\ \vdots & \vdots & \ddots & \vdots\\ r_x(p,1) & r_x(p,2) & ... & r_x(p,p)\end{bmatrix}\begin{bmatrix}a[1]\\ a[2]\\ \vdots\\ a[p]\end{bmatrix}=-\begin{bmatrix}r_x(1,0)\\ r_x(2,0)\\ \vdots\\ r_x(p,0)\end{bmatrix} \tag{69}$$

记为：

$$\boldsymbol{R}_x\bar{\boldsymbol{a}}=-\boldsymbol{r}_x \tag{70}$$

可以发现 $\boldsymbol{R}_x$ 是一个 Hermitian 矩阵。求得 $a[\cdot]$ 后我们就可以代回 Equation 61，令 $n=1,2,...,q$ 时误差为 0，得到 $b[k]$。

但我们需要注意一件事，原本最小化 $e[n]$ 的问题仍然是一个联合非线性最小二乘问题，$a[\cdot]$ 和 $b[\cdot]$ 是耦合的。所以我们分两步依此求解 $a[\cdot]$ 和 $b[\cdot]$ 的方法实际上是一种简化，并**不能保证全局最小化原始误差**。

7.3.2. An Equivalent Perspective from Pseudoinverse

上节的推导中我们自然地应用最小二乘法将问题视为优化问题处理，实际上我们也可以直接令所有的 $\hat{x}[n]=x[n]$，得到一个超定方程组：

$$
\begin{bmatrix}
\overbrace{\begin{matrix} x[0] & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[q] & x[q-1] & x[q-2] & \dots & x[q-p] \end{matrix}}^{\boldsymbol{X}_0} \\
\underbrace{\begin{matrix} x[q+1] \\ \vdots \\ x[q+p] \\ \vdots \end{matrix}}_{\boldsymbol{x}_{q+1}}
\underbrace{\begin{matrix} x[q] & x[q-1] & \dots & x[q-p+1] \\ \vdots & \vdots & \ddots & \vdots \\ x[q+p-1] & x[q+p-2] & \dots & x[q] \\ \vdots & \vdots & \ddots & \vdots \end{matrix}}_{\boldsymbol{X}_q}
\end{bmatrix}
\underbrace{\begin{bmatrix} 1 \\ a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix}}_{\bar{\boldsymbol{a}}}
=
\begin{bmatrix} b[0] \\ \vdots \\ b[q] \\ 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix}
\tag{71}
$$

接下来，我们可以将原本最小化均方误差的过程“内化”到使用伪逆求该方程组最小二乘解的过程当中。两种思路本质上是等效的，前者逻辑更顺畅，而后者有助于从线性空间的角度去理解问题。使用伪逆求解得到：

$$
\bar{\boldsymbol{a}} = -\boldsymbol{X}_q^{+}\boldsymbol{x}_{q+1} = -\left(\boldsymbol{X}_q^H\boldsymbol{X}_q\right)^{-1}\boldsymbol{X}_q^H\boldsymbol{x}_{q+1} \tag{72}
$$

即最优的系数 $\bar{\boldsymbol{a}}$ 将是如下方程组的解：

$$
\left(\boldsymbol{X}_q^H\boldsymbol{X}_q\right)\bar{\boldsymbol{a}} = -\boldsymbol{X}_q^H\boldsymbol{x}_{q+1} \tag{73}
$$

作如下代换后我们再次得到 Equation 70 的 Prony normal equations：

$$
\boldsymbol{R}_x = \boldsymbol{X}_q^H\boldsymbol{X}_q, \quad \boldsymbol{r}_x = \boldsymbol{X}_q^H\boldsymbol{x}_{q+1} \tag{74}
$$

可以计算验证 $\boldsymbol{R}_x$ 与上节中的定义是一致的：

$$
\begin{aligned}
\boldsymbol{R}_x &= \boldsymbol{X}_q^H\boldsymbol{X}_q \\
&= \begin{bmatrix} x^*[q] & x^*[q+1] & x^*[q+2] & \dots \\ x^*[q-1] & x^*[q] & x^*[q+1] & \dots \\ \vdots & \ddots & \vdots & \ddots \\ x^*[q-p+1] & x^*[q-p+2] & x^*[q-p+3] & \dots \end{bmatrix}
\begin{bmatrix} x[q] & x[q-1] & \dots & x[q-p+1] \\ x[q+1] & x[q] & \dots & x[q-p+2] \\ x[q+2] & x[q+1] & \dots & x[q-p+3] \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \\
&= \sum_{n=q+1}^{\infty}
\begin{bmatrix} x^*[n-1]x[n-1] & x^*[n-1]x[n-2] & \dots & x^*[n-1]x[n-p] \\ x^*[n-2]x[n-1] & x^*[n-2]x[n-2] & \dots & x^*[n-2]x[n-p] \\ \vdots & \vdots & \ddots & \vdots \\ x^*[n-p]x[n-1] & x^*[n-p]x[n-2] & \dots & x^*[n-p]x[n-p] \end{bmatrix} \\
&= \begin{bmatrix} r_x(1,1) & r_x(1,2) & \dots & r_x(1,p) \\ r_x(2,1) & r_x(2,2) & \dots & r_x(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p,1) & r_x(p,2) & \dots & r_x(p,p) \end{bmatrix}
\end{aligned}
\tag{75}
$$

该角度还提供了其他有效信息：关于 A^HA 这种形式的矩阵，对任意向量 $\boldsymbol{a}$ 有 $\boldsymbol{a}^H(A^HA)\boldsymbol{a} = (\boldsymbol{A}\boldsymbol{a})^H(\boldsymbol{A}\boldsymbol{a}) = \|\boldsymbol{A}\boldsymbol{a}\|^2 \geq 0$，这说明 A^HA 是半正定矩阵。

由此，前述 Hermitian 矩阵 $\boldsymbol{R}_x = \boldsymbol{X}_q^H\boldsymbol{X}_q$ 同时也是（半）正定矩阵，而该性质将决定 $A(z)$ 是（临界）稳定的，由此弥补了 Padé 法的一项缺陷（TODO，是吗？是否还需要是 Toeplitz?）。

此外，若 $\boldsymbol{R}_x$ 为正定矩阵，则其特征值都是正数，即行列式不为零，矩阵可逆，解存在；若 $\boldsymbol{R}_x$ 为包含零特征值的半正定矩阵，则奇异，但这实际上说明模型的阶数冗余，可以降低一些再尝试。

7.3.3. The Minimum Error and Augmented Normal Equations

由于我们求的是最小二乘解，所以最终拟合信号和真实信号间还是会存在一个最小误差，由 $e[n]$ 定义（Equation 61）和 $\varepsilon_{p,q}$ 定义（Equation 62）继续推导：

$$\begin{aligned}\varepsilon_{p,q} = \sum_{n=q+1}^{\infty} |e[n]|^2 &= \sum_{n=q+1}^{\infty} e[n]\left(x[n] + \sum_{k=1}^{p} a[k]x[n-k]\right)^* \\ &= \sum_{n=q+1}^{\infty} e[n]x^*[n] + \sum_{n=q+1}^{\infty} e[n]\left(\sum_{k=1}^{p} a[k]x[n-k]\right)^* \\ &= \sum_{n=q+1}^{\infty} e[n]x^*[n] + \sum_{k=1}^{p} a^*[k]\left(\sum_{n=q+1}^{\infty} e[n]x^*[n-k]\right)\end{aligned} \tag{76}$$

代入解最优时成立的 Orthogonality principle（Equation 64）和 $e[n]$ 定义（Equation 61）得：

$$\begin{aligned}\varepsilon_{p,q} = \sum_{n=q+1}^{\infty} e[n]x^*[n] &= \sum_{n=q+1}^{\infty}\left(x[n] + \sum_{k=1}^{p} a[k]x[n-k]\right)x^*[n] \\ &= \left(\sum_{n=q+1}^{\infty} x[n]x^*[n]\right) + \sum_{k=1}^{p} a[k]\left(\sum_{n=q+1}^{\infty} x[n-k]x^*[n]\right)\end{aligned} \tag{77}$$

用自相关序列 $r_x(k,l)$（Equation 67）可简化为：

$$\varepsilon_{p,q} = r_x(0,0) + \sum_{k=1}^{p} a[k]r_x(0,k) \tag{78}$$

化成这种形式后我们可以将 $\varepsilon_{p,q}$ 统一到方程 $\boldsymbol{R}_x\bar{\boldsymbol{a}} = -\boldsymbol{r}_x$ 中去（把常量移到矩阵最左侧一列了）：

$$\left[\begin{array}{c|cccc} r_x(0,0) & r_x(0,1) & r_x(0,2) & \dots & r_x(0,p) \\ \hline r_x(1,0) & r_x(1,1) & r_x(1,2) & \dots & r_x(1,p) \\ r_x(2,0) & r_x(2,1) & r_x(2,2) & \dots & r_x(2,p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_x(p,0) & r_x(p,1) & r_x(p,2) & \dots & r_x(p,p) \end{array}\right]\left[\begin{array}{c} 1 \\ \hline a[1] \\ a[2] \\ \vdots \\ a[p] \end{array}\right] = \left[\begin{array}{c} \varepsilon_{p,q} \\ \hline 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{79}$$

或（$\boldsymbol{u}_1$ 为首元素为 1 其他为 0 的单位向量）：

$$\bar{\boldsymbol{R}}_x\boldsymbol{a} = \varepsilon_{p,q}\boldsymbol{u}_1 \tag{80}$$

这样的形式称为 **Augmented normal equations**。（TODO，书上这里用了和前面一样的 R_x，这里暂时加个上划线区分）

我们也可以用矩阵形式推导，更简洁一些。误差序列构成的向量为：

$$\boldsymbol{e} = \boldsymbol{X}_q \bar{\boldsymbol{a}} + \boldsymbol{x}_{q+1} \tag{81}$$

由最小二乘解的性质，我们取最优解即这个误差被最小化时，它一定是与 $\boldsymbol{X}_q$ 的所有列正交的，即有：

$$\boldsymbol{X}_q^H \boldsymbol{e} = 0 \quad \Leftrightarrow \quad \boldsymbol{X}_q^H \left(\boldsymbol{X}_q \bar{\boldsymbol{a}} + \boldsymbol{x}_{q+1} \right) = 0 \quad \Leftrightarrow \quad \boldsymbol{R}_x \bar{\boldsymbol{a}} = -\boldsymbol{r}_x \tag{82}$$

对于阶数为 p, q 的滤波器我们关心的最小误差为：

$$\varepsilon_{p,q} = \|\boldsymbol{e}\|^2 = \boldsymbol{e}^H \boldsymbol{e} = \left(\boldsymbol{X}_q \bar{\boldsymbol{a}} + \boldsymbol{x}_{q+1} \right)^H \boldsymbol{e} = \boldsymbol{x}_{q+1}^H \boldsymbol{e} \tag{83}$$

最后一步是由于 $\boldsymbol{X}_q^H \boldsymbol{e} = 0$。继续代入 $\boldsymbol{e}$ 最终得到同样的式子：

$$\begin{aligned} \varepsilon_{p,q} &= \boldsymbol{x}_{q+1}^H \boldsymbol{e} = \boldsymbol{x}_{q+1}^H \left(\boldsymbol{X}_q \bar{\boldsymbol{a}} + \boldsymbol{x}_{q+1} \right) \\ &= \boldsymbol{x}_{q+1}^H \boldsymbol{x}_{q+1} + \left(\boldsymbol{x}_{q+1}^H \boldsymbol{X}_q \right) \bar{\boldsymbol{a}} \\ &= r_x(0,0) + [r_x(0,1), r_x(0,2), ..., r_x(0,p)] \bar{\boldsymbol{a}} \\ &= r_x(0,0) + \sum_{k=1}^{p} a[k] r_x(0,k) \end{aligned} \tag{84}$$

7.4. Special Case: All-pole Modelling

我们来研究全极点模型（Equation 48），它在一些物理过程中很常见。

7.4.1. All-pole Normal Equations

首先，我们照例将求解 $a[\cdot]$ 的过程视作优化问题。参考 Equation 62 的误差定义，$q = 0$ 时我们有：

$$\varepsilon_{p,0} = \sum_{n=1}^{\infty} |e[n]|^2 \tag{85}$$

其中 $e[n]$ 定义仍来自 Equation 61，但由于 $n = 0$ 时 $n - k < 0$，故 $a[\cdot]$ 的系数 $x[n-k]$ 值为 0，只剩 $x[0] - b[0]$：

$$e[n] = \begin{cases} x[0] - b[0], & n = 0 \\ x[n] + \sum_{k=1}^{p} a[k] x[n-k], & n > 0 \end{cases} \tag{86}$$

这时候我们要作个妖，注意 $e[0] = x[0] - b[0]$ 对于 $a[\cdot]$ 可视作常数，故求解 $a[\cdot]$ 时最小化 $\varepsilon_{p,0}$ 和最小化我们定义的一个新误差 ε_p 是等价的：

$$\varepsilon_p = \sum_{n=0}^{\infty} |e[n]|^2 \tag{87}$$

区别就是把 $e[0]$ 也放进去了。幸运的是，这样更换误差函数之后，我们依然会导出 Prony normal equations 的形式（见 Equation 68），**但其中 $r_x(k,l)$ 的定义变化为**：

$$r_x(k,l) := \sum_{n=0}^{\infty} x^*[n-k]x[n-l] \tag{88}$$

区别是求和下限从 $q+1=1$ 换成了 0，**这是我们更换误差函数的影响**，具体的原因需要从 Equation 63 处开始用新的误差定义重新推导。

好吧其实很简单，我们换 ε_p 后令其对 $a[\cdot]$ 偏导为零：

$$\frac{\partial \varepsilon_p}{\partial a^*[k]} = \sum_{n=0}^{\infty} \frac{\partial [e[n]e^*[n]]}{\partial a^*[k]} = \sum_{n=0}^{\infty} e[n]\frac{\partial e^*[n]}{\partial a^*[k]} = 0, \quad k = 1,2,...,p \tag{89}$$

由上面的 $e[n]$ 定义，我们还是有 $\frac{\partial e^*[n]}{\partial a^*[k]} = x^*[n-k]$，因为 $n=0$ 时 $x^*[n-k]=0$，恰好和 $\frac{\partial (x[0]-b[0])}{\partial a^*[k]} = 0$ 一致，所以可以统一代入得：

$$\sum_{n=0}^{\infty} e[n]x^*[n-k] = 0, \quad k = 1,2,...,p \tag{90}$$

继续代入 $e[n]$ 定义，依旧由于 $x^*[n-k]$ 项的存在，$n=0$ 的情况可以合并进来，直接得：

$$\sum_{n=0}^{\infty} \left(x[n] + \sum_{l=1}^{p} a[l]x[n-l] \right) x^*[n-k] = 0, \quad k = 1,2,...,p \tag{91}$$

移项并重排求和符号顺序可以得到：

$$\sum_{l=1}^{p} a[l] \left(\sum_{n=0}^{\infty} x^*[n-k]x[n-l] \right) = -\sum_{n=q+1}^{\infty} x^*[n-k]x[n], \quad k = ... \tag{92}$$

该式同 Prony normal equations 的形式一致，区别只是 $r_x(k,l)$ 的定义需要如 Equation 88 中所示，变为从 0 开始求和。

到这里我们已经得到了全极点模型情况下求解 $a[\cdot]$ 的 “normal equations”，但观察一下还可以发现，由于 $x[n]$ 在 $n<0$ 时值皆为 0，代入 Equation 88 有：

$$
\begin{aligned}
r_x(k+1,l+1) &= \sum_{n=0}^{\infty} x^*[n-(k+1)]x[n-(l+1)] \\
&= \sum_{n=0}^{\infty} x^*[n-1-k]x[n-1-l] \\
&= \sum_{n=-1}^{\infty} x^*[n-k]x[n-l] \\
&= x^*[-1-k]x[-1-l] + \sum_{n=0}^{\infty} x^*[n-k]x[n-l] \\
&= 0 + \sum_{n=0}^{\infty} x^*[n-k]x[n-l] \\
&= r_x(k,l), \quad (\forall k,l \geq 0)
\end{aligned} \tag{93}
$$

由此，我们可以令：

$$
r_x(k-l) \coloneqq r_x(k,l) = \sum_{n=0}^{\infty} x^*[n-k]x[n-l] \tag{94}
$$

即：

$$
r_x(k) = \sum_{n=0}^{\infty} x^*[n-k]x[n] \tag{95}
$$

观察可知 $r_x(k)$ 是共轭对称的，即 $r_x(k) = r_x^*(-k)$。代入可以得到适用于 All-pole 模型的更简洁的方程组：

$$
\sum_{l=1}^{p} a[l] r_x(k-l) = -r_x(k), \quad k = 1,2,...,p \tag{96}
$$

或：

$$
\begin{bmatrix}
r_x(0) & r_x^*(1) & ... & r_x^*(p-1) \\
r_x(1) & r_x(0) & ... & r_x^*(p-2) \\
\vdots & \vdots & \ddots & \vdots \\
r_x(p-1) & r_x(p-2) & ... & r_x(0)
\end{bmatrix}
\begin{bmatrix}
a[1] \\ a[2] \\ \vdots \\ a[p]
\end{bmatrix}
= -
\begin{bmatrix}
r_x(1) \\ r_x(2) \\ \vdots \\ r_x(p)
\end{bmatrix} \tag{97}
$$

这被称为 **All-pole normal equations**。由于矩阵 $\boldsymbol{R}_x$ 是共轭对称而且 Toeplitz 的，这使得我们可以使用 Levinson-Durbin 算法对它进行高效的求解。

对于最小误差值的计算，类似地我们有：

$$
\varepsilon_p = r_x(0) + \sum_{k=1}^{p} a[k] r_x^*(k) \tag{98}
$$

也可以和方程组一起写成类似 Augmented normal equations 的形式，此处不再赘述。

7.4.2. Issues on the Numerator Selection

按常规方法，求得 $a[\cdot]$ 后我们会通过 Equation 60 得 $b[0]=x[0]$。但在 Section 7.3.1 的最后我们提到 Prony 法中分步求解的方法并不保证全局最优。我们在这里修改 $b[0]$ 的取值并不一定会破坏结果的最优性，因为结果本来就不是最优的；相反，我们甚至有可能通过换一种 $b[\cdot]$ 的取值方式达到更好的效果。

而所谓更好的效果也不一定是误差均方值的降低，也可能综合其他因素的考量。这里的全极点模型就是一个例子，如果原信号由于噪声或其他干扰导致 $x[0]$ 的值实际上不是那么可信，为了防止整个模型参数都被这一个 $b[0]=x[0]$ 带偏，我们更倾向于令拟合信号 $\hat{x}[n]$（在我们的模型中就等于单位脉冲响应 $h[n]$）与目标信号 $x[n]$ 的能量相等：

$$r_{\hat{x}}(0)=r_h(0)=r_x(0) \tag{99}$$

推导可以得到应取 $b[0]=\sqrt{\varepsilon_p}$。

（TODO）关于如何推导得出该 $b[0]$ 的取值还没搞懂，参考书中称在其 5.2.3 节会讲。

7.5. Finite Data Records for All-pole Cases

前面对 Prony 法的分析都基于一个假设，即 $x[n]$ 定义在整个正时间域上，从 0 到 ∞。而现在我们需要考虑当我们只拥有 $[0,N]$ 上的 **$N+1$ 个样本**的情况。至于为什么不是 N 个样本我不知道，可能作者觉得后面的索引简洁一点吧，防止出错也这样写了，但挺难受的。

下面要介绍的两种思路通常用于全极点模型，所以我们也只默认讨论全极点模型。

7.5.1. Auto-correlation Method

第一种方法称为自相关法，我们考虑对 $x[n]$ 加矩形窗，或者说视 $x[n]$ 在 $[0,N]$ 以外的部分值为 0：

$$x_N[n]=\begin{cases}x[n], & 0\le n\le N\\ 0, & \text{otherwise}\end{cases} \tag{100}$$

然后我们直接应用 Prony 法求解。注意，用加窗后的 $x_N[n]$ 去估计的 $r_x(k)$ 会变为：

$$r_x(k)=\sum_{n=0}^{\infty}x_N^*[n-k]x_N[n]=\sum_{n=k}^{N}x^*[n-k]x[n],\quad k=0,1,...,p \tag{101}$$

方便展示，我们将超定方程组写出：

$$\boldsymbol{X}_p \bar{\boldsymbol{a}}_p = -\boldsymbol{x}_1$$

$$\Rightarrow \left[\begin{array}{ccccc} x[0] & 0 & 0 & \dots & 0 \\ x[1] & x[0] & 0 & \dots & 0 \\ x[2] & x[1] & x[0] & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline x[p-1] & x[p-2] & x[p-3] & \dots & x[0] \\ x[p] & x[p-1] & x[p-2] & \dots & x[1] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[N-2] & x[N-3] & x[N-4] & \dots & x[N-p-1] \\ x[N-1] & x[N-2] & x[N-3] & \dots & x[N-p] \\ \hline x[N] & x[N-1] & x[N-2] & \dots & x[N-p+1] \\ 0 & x[N] & x[N-1] & \dots & x[N-p+2] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & x[N] \end{array}\right] \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \left[\begin{array}{c} x[1] \\ x[2] \\ x[3] \\ \vdots \\ \hline x[p] \\ x[p+1] \\ \vdots \\ x[N-1] \\ x[N] \\ \hline 0 \\ 0 \\ \vdots \\ 0 \end{array}\right] \tag{102}$$

而 normal equations 除了 $r_x(k)$ 的定义如 Equation 101 所述修改，形式上则同前 All-pole normal equations（见 Equation 96）：

$$\begin{bmatrix} r_x(0) & r_x^*(1) & \dots & r_x^*(p-1) \\ r_x(1) & r_x(0) & \dots & r_x^*(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p-1) & r_x(p-2) & \dots & r_x(0) \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(p) \end{bmatrix} \tag{103}$$

该法的最小误差值形式也同 All-pole 分析中的 Equation 98 一致，需修改自相关函数。

自相关法将信号直接截断，即使信号在区间外值不为零，所以给出的结果是可能同实际解是**有偏差**的。

（TODO）但该方法有一条重要的性质，可以保证所得模型是**稳定的**，对一些需要大量外推或分析的情况来说非常有用。书中称证明在第 5 章提及，此处暂略。

7.5.2. Covariance Method

Auto-correlation Method 把定义域以外的部分设为 0 本质上是改变了 $x[n]$ 的形态，因为 0 也是正常的信号值。而在一些情况下，这样做并不能达到最好的效果。

第二种协方差法（Covariance Method）的结果通常更加准确，其不对信号本身作假设，而是在优化过程中不考虑定义域以外的样本。

从定义误差、求解优化问题的正规流程来说，我们如果不能考虑定义域外的样本，则误差只能定义在有效区间上。由之前的误差定义，计算 $e[n]$ 需要用到 $x[n], x[n-1], \dots, x[n-p]$，所以我们只能将误差定义在 $[p, N]$ 上：

$$\mathcal{E}_p^C = \sum_{n=p}^{N} |e[n]|^2 \tag{104}$$

然后我们可以用这个误差对系数求偏导，再走一遍推导过程得到 normal equations，但这里就不展开了。

换超定方程组的角度来说更简单，其实就是把自相关法中涉及到定义域以外样本（如 $x[N+1]$ 等）的式子给删了。参考自相关法的超定方程组 Equation 102，只把虚线中间的部分抠出来，就是协方差法的超定方程组：

$$\begin{bmatrix} x[p-1] & x[p-2] & x[p-3] & \dots & x[0] \\ x[p] & x[p-1] & x[p-2] & \dots & x[1] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x[N-2] & x[N-3] & x[N-4] & \dots & x[N-p-1] \\ x[N-1] & x[N-2] & x[N-3] & \dots & x[N-p] \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \begin{bmatrix} x[p] \\ x[p+1] \\ \vdots \\ x[N-1] \\ x[N] \end{bmatrix} \tag{105}$$

而其 normal equations 则同最原始的 Prony normal equations（见 Equation 68）：

$$\begin{bmatrix} r_x(1,1) & r_x(1,2) & \dots & r_x(1,p) \\ r_x(2,1) & r_x(2,2) & \dots & r_x(2,p) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(p,1) & r_x(p,2) & \dots & r_x(p,p) \end{bmatrix} \begin{bmatrix} a[1] \\ a[2] \\ \vdots \\ a[p] \end{bmatrix} = - \begin{bmatrix} r_x(1,0) \\ r_x(2,0) \\ \vdots \\ r_x(p,0) \end{bmatrix} \tag{106}$$

这样做其实舍弃了 All-pole 分析中矩阵 Toeplitz 的性质。同样地，其中自相关函数需要改为有限数据的版本：

$$r_x(k,l) := \sum_{n=p}^{N} x^*[n-k]x[n-l] \tag{107}$$

最小误差值形式上也同 Equation 78，只需修改自相关函数。

7.6. Example: Channel Inversion

（TODO）

8. Stochastic Modelling Identification

对于随机过程的建模同对确定信号的建模主要存在两方面差别。第一，确定建模中由于已知 $x[n]$ 的具体样本值，故误差依赖于确定样本值定义，而随机建模中我们只掌握 $x[n]$ 的统计特征，不再适合使用先前的 $e[n]$ 定义。第二就是输入信号的差别，由于对随机过程进行建模，所以输入也不再适合使用单位脉冲信号，而是使用单位方差的白噪声（White Noise），见 Figure 3。

在考虑这些区别的前提下，我们还要对要建模的随机过程作平稳性假设，即假定随机过程是 WSS 的。

还是类似地，对于随机过程我们可以将 Section 7.1 中的最小二乘误差换成均方误差 $\mathcal{E}_{\text{MS}} = E\{|x[n] - \hat{x}[n]|^2\}$ 来进行优化，但也会遇到同样的非线性问题，难以处理，需要寻找别的方案。

> 注意这里开始的 $x[n]$、$v[n]$ 等不再是具体的离散信号，而是代表一个随机过程。实际应用中我们可能并不直接知道其统计特征，此时才需要从具体的信号（实现、样本）中估计出统计特征。如果只有一条实现用于估算，则应满足相应的平稳性、遍历性假设才有意义。
>
> 方便起见就不将后面所有方括号改成圆括号了。

8.1. Autoregressive Moving Average (ARMA) Processes

我们先定义一种叫 ARMA 过程的随机过程。考虑使用 ARMA 模型的传递函数（Equation 44）对方差为 σ_v^2 的白噪声 $v[n]$ 进行滤波，得到输出 $x[n]$。

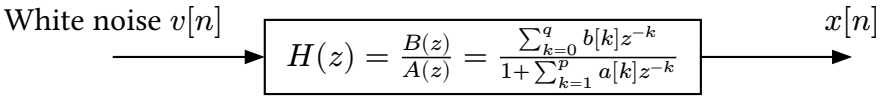

Figure 6: Diagram of ARMA process generation

这里假设 $H(z)$ 是稳定（Stable）的，那么该模型输出的随机过程 $x[n]$ 将会是 WSS 的（证明略）。由于白噪声的功率谱为 $P_v(z) = \sigma_v^2$，得到 $x[n]$ 的功率谱：

$$P_x(z) = \sigma_v^2 \frac{B(z)B^*(1/z^*)}{A(z)A^*(1/z^*)} \tag{108}$$

在频域即为：

$$P_x(e^{j\omega}) = \sigma_v^2 \frac{\left|B(e^{j\omega})\right|^2}{\left|A(e^{j\omega})\right|^2} \tag{109}$$

我们定义用于这种形式的功率谱的过程为 $\text{ARMA}(p, q)$ 过程。可以注意到由于对称性，其功率谱有 $2p$ 和极点和 $2q$ 个零点。

再次澄清，Section 6.1 中提及 ARMA 模型，此处是在说 ARMA 过程。后者是指将白噪声放入 ARMA 模型后输出信号满足的随机过程。

8.1.1. Yule-Walker Equations

在随机建模问题上，我们希望构建的模型输出具有与目标过程相同的统计特征，如考虑输出的自相关 $r_{x(k)}$ 同目标过程一致。所以接下来我们需要**建立起模型输出的自相关 $r_x(k)$ 与系统参数 $a[\cdot]$、$b[\cdot]$ 乃至单位冲激响应 $h[n]$ 之间的统计关系**。

由定义，对于来自 $v[n]$ 的 ARMA 过程 $x[n]$，满足如下方程：

$$x[n]+\sum_{l=1}^{p}a[l]x[n-l]=\sum_{l=0}^{q}b[l]v[n-l] \tag{110}$$

我们可以由此式推得 $x[n]$ 的自相关与 $x[n]$ 同 $v[n]$ 的互相关之间满足同样形式的关系，具体操作是在等式两侧同乘以 $x^*[n-k]$ 并取期望：

$$E\left\{x[n]x^*[n-k]+\sum_{l=1}^{p}a[l]x[n-l]x^*[n-k]\right\}=E\left\{\sum_{l=0}^{q}b[l]v[n-l]x^*[n-k]\right\} \tag{111}$$

即：

$$E\{x[n]x^*[n-k]\}+\sum_{l=1}^{p}a[l]E\{x[n-l]x^*[n-k]\}=\sum_{l=0}^{q}b[l]E\{v[n-l]x^*[n-k]\} \tag{112}$$

在平稳性假设成立的前提下，代入自相关和互相关的定义（见 Section 3.1）得：

$$r_x(k)+\sum_{l=1}^{p}a[l]r_x(k-l)=\sum_{l=0}^{q}b[l]r_{vx}(k-l) \tag{113}$$

互相关项 $r_{vx}(k-l)$ 的存在让式子仍然包含 v，我们可以用表示系统属性的单位冲激响应 $h[n]$ 来换掉它，方法是继续代入 $x[n]=v[n]*h[n]=\sum_{m=-\infty}^{\infty}v[m]h[n-m]$：

$$\begin{aligned}r_{vx}(k-l)&=E\{v[k]x^*[l]\}\\&=E\left\{v[k]\left(\sum_{m=-\infty}^{\infty}v[m]h[l-m]\right)^*\right\}\\&=\sum_{m=-\infty}^{\infty}E\{v[k]v^*[m]\}h^*[l-m]\end{aligned} \tag{114}$$

这里使用的 $r_{vx}(k-l)=E\{v[k]x^*[l]\}$ 由于平稳性假设同前面定义的 $r_{vx}(k-l)=E\{v[n-l]x^*[n-k]\}$ 是一致的，因为 $(n-l)-(n-k)=k-l$。这样换一下推导更简洁。

由于 $v[n]$ 是独立同分布、方差为 σ_v^2 的白噪声，故有：

$$E\{v[k]v^*[m]\} = \begin{cases} \sigma_v^2, & m = k \\ 0, & \text{otherwise} \end{cases} \tag{115}$$

即该项在 $m \neq k$ 时都为 0，代入前式得到：

$$r_{vx}(k-l) = \sigma_v^2 h^*[l-k] \tag{116}$$

于是我们得到了不包含 v 的表达式：

$$r_x(k) + \sum_{l=1}^{p} a[l] r_x(k-l) = \sigma_v^2 \sum_{l=0}^{q} b[l] h^*[l-k] \tag{117}$$

最后考虑现实情况修饰一下，**假设系统是因果（Causal）的**，即 $h[n]$ 在 $n < 0$ 时取值皆为 0，那么 $h^*[l-k]$ 在 $l < k$ 时就为 0，可以修改等式右侧项所含求和的上下限，并记为 $c[k]$：

$$\begin{aligned} c[k] &:= \sum_{l=0}^{q} b[l] h^*[l-k] = \sum_{l=k}^{q} b[l] h^*[l-k] = \sum_{l=0}^{q-k} b[l+k] h^*[l] \\ &= b[k] * h^*[-k] \end{aligned} \tag{118}$$

顺便，$k > q$ 时该项为 0，最后我们得到 **Yule-Walker Equations**：

$$r_x(k) + \sum_{l=1}^{p} a[l] r_x(k-l) = \begin{cases} \sigma_v^2 c[k], & 0 \leq k \leq q \\ 0, & k > q \end{cases} \tag{119}$$

注意，之后我们将默认应用单位方差假设，即取 $\sigma_v^2 = 1$。

这就是本节的目标，即模型输出的自相关 $r_x(k)$ 与系统参数 $a[\cdot]$、$b[\cdot]$ 和单位冲激响应 $h[n]$ 之间的统计关系。

顺带一提，$k > q$ 时：

$$r_x(k) = -\sum_{l=1}^{p} a[l] r_x(k-l) \tag{120}$$

可以用滤波器参数和已知的自相关函数值外推自相关函数之后的值。

清晰一点，我们也把它的矩阵形式画出来：

$$\left[\begin{array}{cccc} r_x(0) & r_x(-1) & \dots & r_x(-p) \\ r_x(1) & r_x(0) & \dots & r_x(-p+1) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(q) & r_x(q-1) & \dots & r_x(q-p) \\ \hline r_x(q+1) & r_x(q) & \dots & r_x(q-p+1) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(q+p) & r_x(q+p-1) & \dots & r_x(q) \\ \vdots & \vdots & \ddots & \vdots \end{array}\right] \left[\begin{array}{c} 1 \\ a[1] \\ a[2] \\ \vdots \\ a[p] \end{array}\right] = \sigma_v^2 \left[\begin{array}{c} c[0] \\ c[1] \\ \vdots \\ c[q] \\ \hline 0 \\ \vdots \\ 0 \\ \vdots \end{array}\right] = \left[\begin{array}{c} c[0] \\ c[1] \\ \vdots \\ c[q] \\ \hline 0 \\ \vdots \\ 0 \\ \vdots \end{array}\right] \tag{121}$$

8.1.2. Modified Yule-Walker Equation (MYWE) Method

Yule-Walker 方程可以用于从自相关函数求解滤波器参数，但由于 $h^*[l]$ 的存在，它仍然是一个较难处理的非线性问题。

再次澄清，在该问题中我们对随机过程建模而不是对具体的信号建模，故视目标过程的统计特征（如自相关函数）是已知的，如果我们不知其自相关函数 $r_v(k)$ 的值，才需要通过统计方法从一些实现（样本）中估算得到 $\hat{r}_v(k)$。

回到参数辨识的问题上来，我们可以先仿照 Pade 法的思路，通过分步求解来**近似**最优结果。先用 $q < k \leq q+p$ 的部分估计 $a[\cdot]$，对应的式子为：

$$\left[\begin{array}{cccc} r_x(q) & r_x(q-1) & \dots & r_x(q-p+1) \\ r_x(q+1) & r_x(q) & \dots & r_x(q-p+2) \\ \vdots & \vdots & \ddots & \vdots \\ r_x(q+p-1) & r_x(q+p-2) & \dots & r_x(q) \end{array}\right] \left[\begin{array}{c} a[1] \\ a[2] \\ \vdots \\ a[p] \end{array}\right] = -\left[\begin{array}{c} r_x(q+1) \\ r_x(q+2) \\ \vdots \\ r_x(q+p) \end{array}\right] \tag{122}$$

该方程组称为 **Modified Yule-Walker equations**（MYWE），于是该方法称为 MYWE 法。值得注意的是该方程组形式与 Pade 法中 Equation 59 的形式完全一样，只是将自相关函数换成了 $x[n]$ 的值。该矩阵也是 Toeplitz 的，所以可以使用 Trench 算法等加速求解。

得到 $a[\cdot]$ 后，**第二步需要求解 $b[\cdot]$**。若将 $a[\cdot]$ 代回 Yule-Walker 方程我们可以得到 $c[\cdot]$ 的值。但 $c[k] := b[k] * h^*[-k]$，$h[k]$ 甚至还依赖 $b[k]$，想求出 $b[\cdot]$ 非常困难。课件称 “We skip this”，似乎不打算管这部分。参考书的相关内容大致从 190 页开始，主要提到几种方法。

第一，我们已知 $a[\cdot]$，用其构造一个 AR 滤波器 $A(z)$ 对 $x[n]$ 进行滤波可以得到新过程 $y[n]$：

$$P_x(z) = \frac{B(z)B^*(1/z^*)}{A(z)A^*(1/z^*)} \quad \overset{A(z)}{\longrightarrow} \quad P_y(z) = B(z)B^*(1/z^*) \tag{123}$$

这个过程是一个 MA 过程，我们再用之后 Section 8.3 中的方法处理，估计 $b[\cdot]$。

第二，不显式地进行滤波，不过本质应该和第一种一样。通过 Yule-Walker 方程上半部分求出 $c[\cdot]$ 后，求正半轴的拉普拉斯变换得到（因为通过 Yule-Walker 方程只能求出其正半轴的值）：

$$[C(z)]_+ = \sum_{k=0}^{\infty} c[k] z^{-k} \tag{124}$$

相应地，虽然我们不知道，但其负半轴的拉普拉斯变换为：

$$[C(z)]_- = \sum_{k=-\infty}^{-1} c[k]z^{-k} = \sum_{k=1}^{\infty} c[-k]z^k \tag{125}$$

由定义 $c[k] := b[k] * h^*[-k]$ 又得到 MA 过程的功率谱：

$$\begin{aligned} C(z) &= B(z)H^*(1/z^*) = B(z)\frac{B^*(1/z^*)}{A^*(1/z^*)} \\ \Rightarrow \quad P_y(z) &\equiv C(z)A^*(1/z^*) = B(z)B^*(1/z^*) \end{aligned} \tag{126}$$

我们将其拆开写：

$$P_y(z) = C(z)A^*(1/z^*) = [C(z)]_+ A^*(1/z^*) + [C(z)]_- A^*(1/z^*) \tag{127}$$

由于 $a[k]$ 负半轴值为 0，则 $A^*(1/z^*)$ 只包含 z 的正功率，同时 $[C(z)]_+$ 也是如此（TODO，这里书上写的减号？），故 $P_y(z)$ 的 causal part 即：

$$\left[P_y(z)\right]_+ = [C(z)]_+ A^*(1/z^*)]_+ \tag{128}$$

于是虽然我们不知道 $c[k]$ 的负半轴部分的值，但通过该式，我们可以用已知的 $c[\cdot]$ 正半轴的值伙同 $a[\cdot]$ 求出 $\left[P_y(z)\right]_+$，再由共轭对称性得到完整的 $P_y(z)$。最后对其进行谱分解（Spectral Factorization）得到系数 $b[\cdot]$：

$$P_y(z) = B(z)B^*(1/z^*) \tag{129}$$

（TODO，要不抄过来）参考书 192 页还有一个清晰的例子。

8.1.3. Extended Yule-Walker Equation Method

相应地，在第一步中我们也可以使用类似 Prony 法的方式，将 $k > q$ 的所有式子都纳入考虑，得到的超定方程组称为 **Extended Yule-Walker equations**。

然后我们公式化求一个最小二乘解，过程也与 Prony 法相似，具体不再赘述。

8.2. Autoregressive (AR) Processes

我们又来考虑 all-pole 的情况，由于只剩下一个 $b[0]$，方程可以简化很多：

$$r_x(k) + \sum_{l=1}^{p} a[l]r_x(k-l) = |b[0]|^2\delta(k), \quad k \geq 0 \tag{130}$$

由于不存在复杂的 $c[k]$ 的问题，可以直接沿用 Prony 法，先用除了第一个以外的式子求 $a[\cdot]$，可以注意到式子和 Equation 97 完全一样；然后再用第一个式子推得 $b[0]$。这被称为 Yule-Walker 法（什么混乱的起名方式）。

（TODO，这前头的 1/N 是？虽然好像求解 a[·] 时抵消了，b[·] 也因为直接选取，不受影响。还有星号呢？这个式子见书 P194，4.153）再次说明，如果我们不知道目标过程的自相关，就需要从（满足遍历性假设的）样本中去估计，例如：

$$\hat{r}_x(k) = \frac{1}{N}\sum_{n=0}^{N-1} x[n]x^*[n-k] \tag{131}$$

这么一搞其实就和前面的 Auto-correlation 法（Section 7.5.1）等价了，也符合直觉。

8.3. Moving Average Processes

对于 MA 过程，我们代入 Yule-Walker 方程后得到：

$$r_x(k) = \sum_{l=0}^{q} b[l]b^*[l-k] = b[k] * b^*[-k] \tag{132}$$

即有：

$$P_x(z) = B(z)B^*(1/z^*) \tag{133}$$

总结来说，就是将自相关函数 $r_x(k)$ 作 z-变换后得到功率谱 $P_x(z)$，然后进行谱分解即可得到结果，举个例子：

$$r_x(k) = 17\delta(k) + 4[\delta(k-1) + \delta(k+1)] \tag{134}$$

z-变换得到：

$$P_x(z) = 17 + 4z^{-1} + 4z = (4 + z^{-1})(4 + z) \tag{135}$$

于是有：

$$B(z) = 4 + z^{-1} \quad \text{or} \quad B(z) = 1 + 4z^{-1} \tag{136}$$

此外，还有 Durbin's method 等方法，此处不记录。

9. Spectrum Estimation

功率谱的估计显然是个很有用的东西。例如，加性噪声、信号与噪声不相关前提下的非因果 Wiener 平滑滤波器有如下频率响应：

$$H(e^{j\omega}) = \frac{P_{dx}(e^{j\omega})}{P_x(e^{j\omega})} = \frac{P_d(e^{j\omega})}{P_d(e^{j\omega}) + P_v(e^{j\omega})} \tag{137}$$

如果目标信号与噪声的功率谱密度已知则可以直接求出其频率响应，若不知道则可以通过谱估计得到结果；再例如对窄频带信号的检测与追踪等。

总之本节考虑宽平稳随机过程的功率谱密度的估计。具体地，给定一个随机过程产生的随机信号，如何根据它有效地估计这个随机过程的谱？

最直接的想法是，由 Wiener–Khinchin Theorem，我们知道自相关函数就可以通过傅里叶变换得到功率谱：

$$P_x(e^{j\omega}) = \sum_{k=-\infty}^{\infty} r_x(k) e^{jk\omega} \tag{138}$$

而对于一个遍历的随机过程产生的随机信号 $x[n]$，我们可以这样得到该过程的自相关函数：

$$r_x(k) = \lim_{N\to\infty} \left\{ \frac{1}{2N+1} \sum_{n=-N}^{N} x[n+k]x^*[n] \right\} \tag{139}$$

这是由于满足遍历性假设，我们用一条无限长的实现就可以无偏地估计出该过程的统计特征。

但显然这么做存在一些问题：第一，我们拥有的样本长度往往是有限的，例如地震波等本身就很短的信号，以及语音信号等只有很短时间内才近似满足平稳性假设的信号；第二，信号样本往往自己还包含噪声。

谱估计的方法可以分为两类：一类是无参数（Nonparametric）的方法，如从估计序列的自相关入手，变换得到功率谱；另一类是在对随机过程的模型有先验了解的情况下可以使用的有参（Parametric）估计的方法，从估计模型参数入手，再由模型计算功率谱。

9.1. Nonparametric Spectrum Estimation

9.1.1. Periodogram

本章开头提到从样本中估计自相关函数，然后傅里叶变换得到功率谱的方法。即使现在样本长度有限（例如 N 个），我们也就用这点样本来直接估计自相关函数：

$$\hat{r}_x(k) = \frac{1}{N} \sum_{n=0}^{N-1-k} x[n+k]x^*[n], \quad k = 0, 1, ..., N-1 \tag{140}$$

此处实际上是假定了 $x[n]$ 在 $[0, N-1]$ 以外的部分值都为 0，等效于将求和范围设置为只从 0 到 $N-1-k$，即求和共有 $N-k$ 项。但由于前面除以的是 N，最终得到的是一个有偏（Biased）估计：

$$
\begin{aligned}
E\{\hat{r}_x(k)\} &= \frac{1}{N}\sum_{n=0}^{N-1-k} E\{x[n+k]x^*[n]\} \\
&= \frac{1}{N}\sum_{n=0}^{N-1-k} r_x(k) \\
&= \frac{N-k}{N} r_x(k), \quad k = 0, 1, ..., N-1
\end{aligned} \tag{141}
$$

当 N 趋于无穷时期望就和实际值相等了，所以它是渐近无偏（Asympotically Unbiased）的。k 为负的部分我们可以由 WSS 过程自相关函数的共轭对称性质翻转得到，于是可以将其重写为：

$$
E\{\hat{r}_x(k)\} = w_B(k) r_x(k) \tag{142}
$$

其中：

$$
w_B(k) = \begin{cases} \frac{N-|k|}{N}, & |k| \le N \\ 0, & |k| > N \end{cases} \tag{143}
$$

这是一个 Bartlett (triangular) window。

接下来再对它使用傅里叶变换估计功率谱：

$$
\hat{P}(e^{j\omega}) = \sum_{k=-N+1}^{N-1} \hat{r}_x(k) e^{-jk\omega} \tag{144}
$$

以上这个两步走的过程主要是顺着定义来，了解原理后我们可以化简一下直接通过随机信号 $x[n]$ 求得上述频谱估计。首先我们对 $x[n]$ 在 $[0, N-1]$ 以外的部分值都为 0 的假设可以用乘以一个 Dirichlet 核（Equation 41）的过程替代，即 $x_N[n] = x[n]w_R[n]$，其傅里叶变换为：

$$
X_N(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x_N[n] e^{-j\omega n} = \sum_{n=0}^{N-1} x[n] e^{-j\omega n} \tag{145}
$$

由：

$$
\hat{r}_x(k) = \frac{1}{N}\sum_{n=-\infty}^{\infty} x_N[n+k] x_N^*[n] = \frac{1}{N} x_N[k] * x_N^*[-k] \tag{146}
$$

再由 DTFT 的性质，可得功率谱估计为：

$$
\hat{P}_{\text{per}}(e^{j\omega}) = \frac{1}{N} X_N(e^{j\omega}) X_N^*(e^{j\omega}) = \frac{1}{N}\left|X_N(e^{j\omega})\right|^2 \tag{147}
$$

将满足该定义式的功率谱估计称为**周期图**（Periodogram）。

9.1.1.1. An Equivalent Perspective from a Filter Bank

前面是基于 Wiener–Khinchin 定理，从“估计自相关函数 $\rightarrow$ 傅里叶变换得谱估计”的逻辑推导周期图的定义，接下来我们换一个角度解释周期图的含义。

一个直观的想法是，如果我们知道某个信号每个频率分量的功率值，那么我们实际上就**直接可以拼出它的功率谱密度图像**。那么如何知道一个信号某个频率分量的功率？自然地，我们可以先用一个很窄的带通滤波器把这个频率分量滤出来，然后再想办法处理得到它的功率，这里可以用上 Parseval 定理，见后。

那么具体地，我们定义一组长度为 N 的 FIR 滤波器如下：

$$h_i[n] = \frac{1}{N} e^{jn\omega_i} w_R[n] = \begin{cases} \frac{1}{N} e^{jn\omega_i}, & 0 \le n < N \\ 0, & \text{otherwise} \end{cases} \tag{148}$$

这些滤波器的傅里叶变换为：

$$H_i(e^{j\omega}) = \sum_{n=0}^{N-1} h_i[n] e^{-jn\omega} = e^{-j(\omega-\omega_i)(N-1)/2} \frac{\sin(N(\omega-\omega_i)/2)}{N\sin((\omega-\omega_i)/2)} \tag{149}$$

其主瓣的带宽大致为 $\Delta\omega = \frac{2\pi}{N}$。

这样设计滤波器的原因是，滤波器的单位脉冲响应 $h_i[n]$ 在时域上同 $x[n]$ 卷积，对应它们的傅里叶变换 $H_i(e^{j\omega})$ 和 $X(e^{j\omega})$ 在频域上的乘积。而我们希望频域上它们相乘之后只留下 $X(e^{j\omega})$ 在频率 ω_i 处的分量，其他全部为零，相当于采了个样。所以理想情况下的 $H_i(e^{j\omega})$ 应该为 $\delta(\omega-\omega_i)$，它的时域表达式就是 $\frac{1}{2\pi}e^{jn\omega_i}$。

此外，我们的序列长度有限，无法实现理想的 $h_i[n]$，只能在长度为 N 处截断。这里我们设置 $e^{jn\omega_i}$ 前的系数为 $1/N$，是为了使其满足 $\left|H_i(e^{j\omega})\right|_{\omega=\omega_i} = 1$，方便计算。

注意，我们接下来考察的是随机过程的真实功率谱密度 $P_x(e^{j\omega_i})$，所以下文所有的 $x[n]$ 都是指信号背后的**随机过程**，考察其经过滤波后的整体行为，而非一个具体的信号。

于是，$x[n]$ 经过 $h_i[n]$ 滤波后得到的也是一个随机过程 $y_i[n]$：

$$y_i[n] = x[n] * h_i[n] = \sum_{k=n-N+1}^{n} x[k]h_i[n-k] = \frac{1}{N}\sum_{k=n-N+1}^{n} x[k]e^{j(n-k)\omega_i} \tag{150}$$

由于 $\left|H_i(e^{j\omega})\right|_{\omega=\omega_i} = 1$，所以在 ω_i 频率点上，输入 $x[n]$ 和输出 $y_i[n]$ 的功率谱密度值应该是一致的：

$$P_x(e^{j\omega_i}) = P_y(e^{j\omega_i}) \tag{151}$$

接下来我们要想办法得到这个 $P_y(e^{j\omega_i})$，从而得到我们希望知道的 $P_x(e^{j\omega_i})$。方法是考虑 Parseval 定理：

$$E\left\{|y_i[n]|^2\right\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} P_y(e^{j\omega})\left|H_i(e^{j\omega})\right|^2 \mathrm{d}\omega \tag{152}$$

如果 $H_i(e^{j\omega})$ 的带宽足够窄、旁瓣足够小，接近理想带通滤波器，那么可以认为在 passband 内 $P_y(e^{j\omega}) = P_y(e^{j\omega_i})$ 均匀不变，stopband 内均为零，于是：

$$E\left\{|y_i[n]|^2\right\} \approx \frac{1}{2\pi}\left(\Delta\omega \cdot P_y(e^{j\omega_i})\right) = \frac{1}{N}P_y(e^{j\omega_i}) = \frac{1}{N}P_x(e^{j\omega_i}) \tag{153}$$

这个结论说明我们可以用 $y_i[n]$ 的功率来估计 $P_x(e^{j\omega_i})$：

$$P_x(e^{j\omega_i}) \approx NE\left\{|y_i[n]|^2\right\} \tag{154}$$

好了，由于 $y_i[n]$ 在这里代表一个随机过程，我们无法考察其具体值，所以接下来问题变成了怎么从实际的样本中估计这个功率 $\hat{E}\left\{|y_i[n]|^2\right\}$。

到这一步我们**有很多选择**，但这里是为了推导出周期图，所以我们用单点样本平均（One-point average）去估计它，令：

$$\hat{E}\left\{|y_i[n]|^2\right\} = |y_i[N-1]|^2 = \frac{1}{N^2}\left|\sum_{k=0}^{N-1} x[k]e^{-jk\omega_i}\right|^2 \tag{155}$$

这实际上是非常极端、不精确的近似，即使单点样本平均确实是功率的无偏估计，但太容易受到信号实际值波动的影响。**不过可以说这里只是为了凑出周期图的定义，仅提供一种诠释方式**。用这个估计方式，最终我们得到：

$$\hat{P}_x(e^{j\omega_i}) = N|y_i[N-1]|^2 = \frac{1}{N}\left|\sum_{k=0}^{N-1} x[k]e^{-jk\omega_i}\right|^2 \tag{156}$$

这同 Equation 147 中周期图的定义是一致的。

在后面的 Section 9.1.5 中我们会用到和这里滤波器组类似的思路，但根据 Equation 34 用自相关矩阵和滤波器系数得到更好的功率估计。

9.1.1.2. Performance of the Periodogram

我们接下来评估用周期图作为功率谱估计的表现。首先我们一定希望从样本中计算出的周期图可以收敛到实际随机过程的功率谱。由于其随机性质，我们需要从统计意义上考虑收敛性，如均方收敛：

$$\lim_{N\to\infty} E\left\{\left[\hat{P}_{\text{per}}(e^{j\omega}) - P_x(e^{j\omega})\right]^2\right\} = 0 \tag{157}$$

要满足这一点，我们需要其均值渐近无偏、方差在样本量足够大时趋于零：

$$\begin{aligned}\lim_{N\to\infty} E\left\{\hat{P}_{\text{per}}(e^{j\omega})\right\} &= P_x(e^{j\omega}) \\ \lim_{N\to\infty} \mathrm{Var}\left\{\hat{P}_{\text{per}}(e^{j\omega})\right\} &= 0\end{aligned} \tag{158}$$

换句话说，我们希望周期图是对功率谱密度的一致估计（Consistent Estimate）。**先说结论**，以上第一个条件成立，但第二个条件不成立。

具体地，首先考虑**第一个条件**，即渐近无偏。从样本自相关开始，其期望如 Equation 142 所示，再由周期图的推导可以得到：

$$\begin{aligned} E\left\{\hat{P}_{\text{per}}(e^{j\omega})\right\} &= E\left\{\sum_{k=-N+1}^{N-1} \hat{r}_x(k)e^{-jk\omega}\right\} = \sum_{k=-N+1}^{N-1} E\{\hat{r}_x(k)\}e^{-jk\omega} \\ &= \sum_{k=-N+1}^{N-1} w_B(k)r_x(k)e^{-jk\omega} \end{aligned} \tag{159}$$

这就是自相关函数和窗函数乘积的傅里叶变换，于是由频域卷积性质得到：

$$E\left\{\hat{P}_{\text{per}}(e^{j\omega})\right\} = \frac{1}{2\pi}P_x(e^{j\omega}) * W_B(e^{j\omega}) \tag{160}$$

其中 Bartlett 窗的频域表达式是：

$$W_B(e^{j\omega}) = \frac{1}{N}\left[\frac{\sin(N\omega/2)}{\sin(\omega/2)}\right]^2 \tag{161}$$

> 注意这里的窗函数是加在自相关序列上的，所以称为滞后窗（Lag window），区别于之后直接加在数据上的数据窗（Data window）。无实际意义，仅作概念上的区分。

注意到 $N \to \infty$ 时，$W_B(e^{j\omega})$ 收敛到一个脉冲函数（周期积分为 2π，面积集中到原点，即 $2\pi\delta(\omega)$，证明略），于是周期图满足渐近无偏的条件。

从频谱图像的角度，本来是用一堆理想的脉冲函数拼出整个谱，现在变成用 Bartlett 窗函数来拼出整个谱。样本越多，窗函数越接近理想脉冲信号，频谱估计越准确，如 Figure 7 所示。

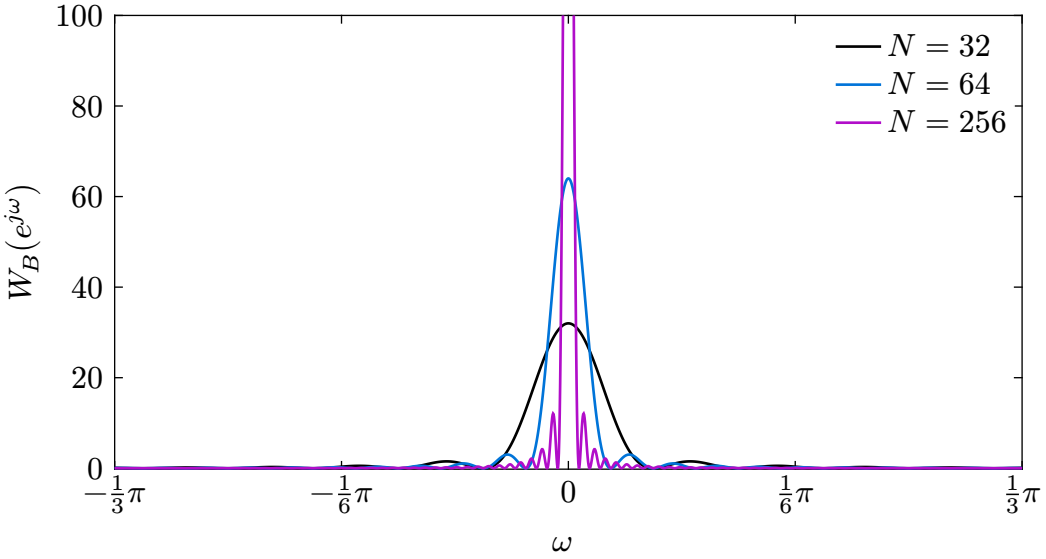

Figure 7: The Fourier transforms of Bartlett windows

该函数的主瓣带宽大约为 $\frac{2\pi}{N}$，太过靠近的两个脉冲信号在同窗函数卷积后可能导致两个峰叠加合并，无法清晰分辨。故定义分辨率为此处窗函数的 6dB 带宽：

$$\operatorname{Res}\left[\hat{P}_{\text{per}}(e^{j\omega})\right] = (\Delta\omega)_{6\text{dB}} = 0.89\frac{2\pi}{N} \tag{162}$$

−6dB 约为 0.5，即大约在这个位置两峰交叠处的值为峰值的一半，考虑 Equation 160 中的卷积过程，这会导致两峰叠加后只剩一个峰而使结果无法分辨。

考虑在实际计算中的一个细节问题，即这里给出的 $\Delta\omega$ 的数值，是在傅里叶变换归一化到 $[-\pi, \pi]$ 区间后的测度意义下的。

而通常我们的指标是在实际频谱单位下的，这时我们需要除以采样率。例如，一个采样率为 10kHz 的信号，我们指出要求分辨率至少达到 10Hz，那么 $\Delta\omega$ 应该是 $2\pi \times \frac{10\text{Hz}}{10\text{kHz}}$。

第二个条件，考虑方差是否趋于零。由于周期图同样本是二阶关系，现在又计算方差，就是对随机过程四阶矩的计算，这太过复杂；但我们可以考虑随机过程是方差为 σ_x^2 的高斯白噪声（Gaussian white noise）的特殊情况。经过一系列计算（详见参考书第 8.2.2 节，404 页）可得该情况下的方差为：

$$\operatorname{Var}\left\{\hat{P}_{\text{per}}(e^{j\omega})\right\} = \sigma_x^4 \tag{163}$$

与 N 无关，不会随其增长收敛到零。实际上，如果考虑普遍情况，我们有如下近似（对于高斯白噪声的情况就是 $\left(\sigma_x^2\right)^2$）：

$$\operatorname{Var}\left\{\hat{P}_{\text{per}}(e^{j\omega})\right\} \approx P_x^2(e^{j\omega}) \tag{164}$$

所以结论就是，第二个条件并不满足，即**周期图不是对功率谱密度的一致估计**。

9.1.2. Modified Periodogram

所以我们当然想着要改进一下，我们先不管前面的各种推导，直接考虑在定义上修一修，然后再验证效果。回过来观察周期图的定义式：

$$\hat{P}_{\text{per}}(e^{j\omega}) = \frac{1}{N}\left|X_N(e^{j\omega})\right|^2 = \frac{1}{N}\left|\sum_{n=-\infty}^{\infty} x[n]w_R[n]e^{-j\omega n}\right|^2 \tag{165}$$

式中体现的是我们对原信号使用 Dirichlet 核（也就是方形窗，Rectangular window）后进行谱估计的过程。那么一个很直观的想法是，如果这里用别的窗函数会不会有什么效果？

（TODO）书 408、409 页在推期望和方差。注意书上好像有关于 $w_B(k)$ 的定义不一致性，我们这里采用归一化的 $w_B(k) = \frac{1}{N}w_R(k) * w_R(-k) = \sum_{n=-\infty}^{\infty} w_R(k)w_R(n-k)$，和前面保持一致，所以和书上略有系数的差异，要改。

我们定义修正周期图（Modified periodogram）为：

$$\hat{P}_M(e^{j\omega}) = \frac{1}{NU}\left|\sum_{n=-\infty}^{\infty} x[n]w[n]e^{-jn\omega}\right|^2 \tag{166}$$

其中 N 为窗函数的长度，常数 U 为窗函数的功率，也就是能量对时间的平均（后续会说明这是为了使修正周期图渐近无偏）：

$$U = \frac{1}{N}\sum_{n=0}^{N-1}|w[n]|^2 \tag{167}$$

9.1.2.1. Performance of Modified Periodogram

类似地，我们评估修正周期图的表现。首先是偏差（Bias），由类似的推导有：

$$E\left\{\hat{P}_M(e^{j\omega})\right\} = \frac{1}{2\pi NU}P_x(e^{j\omega}) * \left|W(e^{j\omega})\right|^2 \tag{168}$$

其中 $W(e^{j\omega})$ 是 $w[n]$ 的傅里叶变换，故由前 U 的设置有：

$$U = \frac{1}{N}\sum_{n=0}^{N-1}|w[n]|^2 = \frac{1}{2\pi N}\int_{-\pi}^{\pi}\left|W(e^{j\omega})\right|^2 \mathrm{d}\omega \tag{169}$$

即：

$$\int_{-\pi}^{\pi}\frac{1}{2\pi NU}\left|W(e^{j\omega})\right|^2 \mathrm{d}\omega = 1 \tag{170}$$

这使得 $E\left\{\hat{P}_M(e^{j\omega})\right\}$ 在 $N \to \infty$ 时趋于功率谱密度，即渐近无偏，也是这样设置 U 的目的。

接下来是方差，加数据窗并不有助于降低方差，故同 Equation 164 一样：

$$\mathrm{Var}\left\{\hat{P}_M(e^{j\omega})\right\} \approx P_x^2(e^{j\omega}) \tag{171}$$

即修正周期图同样不是对功率谱密度的一致估计。

9.1.2.2. Trade-off between Resolution and Confusion

不影响估计的偏差与方差，那么加数据窗的操作到底对什么有影响？

不同数据窗的傅里叶变换形态有差异，主要体现在主瓣（Main lobe）和旁瓣（Sidelobe）上。参考 Section 4.1，前者将影响谱估计的分辨率（Resolution），后者则会引入旁瓣的干扰和混淆。

我们定义分辨率为数据窗主瓣的 3dB 带宽，这个值越大说明越不清晰：

$$\mathrm{Res}\left[\hat{P}_{\mathrm{per}}(e^{j\omega})\right] = (\Delta\omega)_{3\mathrm{dB}} \tag{172}$$

注意到，前文分析 Periodogram 时定义的分辨率是 Bartlett 窗的 6dB 带宽而不是 3dB，但实际上这是一致的。

这是由于前文分析的是加在自相关序列上的窗（即滞后窗），注意 Equation 160 中同功率谱密度卷积的是 $W_B(e^{j\omega})$；而此处分析的是数据窗，注意 Equation 168 中同功率谱密度卷积的是 $\frac{1}{NU}\left|W(e^{j\omega})\right|^2$，二者之间存在平方关系。

所以前者的 -6dB 点同后者的 -3dB 点是一致的，都是相对信号来说的半功率点。

常见窗函数旁瓣抑制和分辨率的大致值总结如 Table 1 所示。

	Sidelobe (dB)	Resolution
Rectangular	-13	$0.89(2\pi/N)$
Bartlett	-27	$1.28(2\pi/N)$
Hanning	-32	$1.44(2\pi/N)$
Hamming	-43	$1.30(2\pi/N)$
Blackman	-58	$1.68(2\pi/N)$

Table 1: Properties of a few commonly used windows with length N

可以观察到往往旁瓣抑制效果越好，分辨率就越差（即3dB带宽越大），这是一个 Trade-off。

9.1.3. Periodogram Averaging

至此，由于方差不收敛，以上方法都无法得到关于功率谱密度的一致估计。而接下来我们通过几种对周期图做平均的方法来得到我们想要的一致估计。

考虑之前对于一个随机变量 x，我们通过收集其大量不相关的测量样本并计算样本均值的方式，得到了对该随机变量均值的一致估计 $E\{x\}$。类比之，理论上我们需要用随机过程 $x(n)$ 的多个不相关的实现（Uncorrelated realizations），分别求周期图后再平均，估计周期图的期望。

具体地，设我们有 K 个不相关的实现 $x_i[n]$，每个长度为 L，有总样本点数为 $N = LK$。计算每个实现的周期图：

$$\hat{P}_{\text{per}}^{(i)}(e^{j\omega}) = \frac{1}{L}\left|\sum_{n=0}^{L-1} x_i[n]e^{-jn\omega}\right|^2, \quad i = 0, 1, ..., K-1 \tag{173}$$

然后求平均得最终的谱估计：

$$\hat{P}_x(e^{j\omega}) = \frac{1}{K}\sum_{i=0}^{K-1}\hat{P}_{\text{per}}^{(i)}(e^{j\omega}) = \frac{1}{N}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1} x_i[n]e^{-jn\omega}\right|^2 \tag{174}$$

下面照例评估其偏差和方差。首先，因为只是再做了一次平均，期望同前面是一样的：

$$E\left\{\hat{P}_x(e^{j\omega})\right\} = E\left\{\hat{P}_{\text{per}}^{(i)}(e^{j\omega})\right\} = \frac{1}{2\pi}P_x(e^{j\omega}) * W_B(e^{j\omega}) \tag{175}$$

故在 $L \to \infty$ 时渐近无偏。然后考虑方差，由于不同实现之间不相关，有：

$$\begin{aligned}\text{Var}\left\{\hat{P}_x(e^{j\omega})\right\} &= \frac{1}{K^2}\text{Var}\left\{\sum_{i=0}^{K-1}\hat{P}_{\text{per}}^{(i)}(e^{j\omega})\right\} \\ &= \frac{1}{K}\text{Var}\left\{\hat{P}_{\text{per}}^{(i)}(e^{j\omega})\right\} \approx \frac{1}{K}P_x^2(e^{j\omega})\end{aligned} \tag{176}$$

综上，使用这种基于平均的方法可以在 L 和 K 都趋于无穷时给出对功率谱密度的**一致估计**。

9.1.3.1. Bartlett's Method

一般实际情况中我们没有那么多独立的实现，但若我们有一条足够长的实现，并且其背后的随机过程满足遍历性假设，那么我们就可以把它切成小段当作不相关的实现来用，求得谱估计，称为 Bartlett 法。

设信号长度为 N，切成**不重叠**（尽量保证不相关）的 K 段，每段长度 L。若我们再令：

$$x_i[n] = x[n + iL], \quad n = 0, 1, ..., L-1; \quad i = 0, 1, ..., K-1 \tag{177}$$

符号就同前面的分析一致了，我们直接得到公式：

$$\begin{aligned}\hat{P}_B(e^{j\omega}) &= \frac{1}{K}\sum_{i=0}^{K-1}\left(\frac{1}{L}\left|\sum_{n=0}^{L-1}x[n+iL]e^{-jn\omega}\right|^2\right)\\ &= \frac{1}{N}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1}x[n+iL]e^{-jn\omega}\right|^2\end{aligned} \tag{178}$$

偏差和方差则同前文平均法的分析一致。不过由于即便不重叠，切分出的序列间也一定存在相关性(遍历性只是保证在较长时间下相关性会逐渐消弭)，所以方差可能比前面分析的还小一些。不过我们还是近似地认为片段之间不相关，取这个方差的近似值。

接下来我们再分析一下该法对分辨率的影响，由于原本长度为 N 的序列被我们切成长度为 L 的小段了，所以计算周期图时实际序列长度只有 L，故分辨率为：

$$\mathrm{Res}\left[\hat{P}_B(e^{j\omega})\right] = 0.89\frac{2\pi}{L} = 0.89K\frac{2\pi}{N} \tag{179}$$

可以看到相比用整个长度为 N 的序列计算周期图，分辨率变差了 K 倍，这是代价。

9.1.3.2. Welch's Method

Barlett 法用周期图做平均，接下来我们沿用这个思想，用修正周期图做平均，称为 Welch 法。

修正周期图的想法是对数据加窗函数，而这一操作实际可以减弱相邻两个片段边缘处信号的相关性。所以我们可以考虑放宽要求，允许切分数据时有**重叠**（Overlapping）。每个片段依旧长度为 L，但每段的起点只间隔 D 个样本（$D<L$ 时就有重叠了）：

$$x_i[n] = x[n+iD],\quad n=0,1,...,L-1;\quad i=0,1,...,K-1 \tag{180}$$

于是总样本点数就为 $N=L+D(K-1)$。通常，我们取 50% 重叠，即 $D=L/2$。数据窗加到每个片段上，窗长度同片段长度 L 一致。于是我们得到 Welch 法的谱估计为：

$$\begin{aligned}\hat{P}_W(e^{j\omega}) &= \frac{1}{K}\sum_{i=0}^{K-1}\left(\frac{1}{LU}\left|\sum_{n=0}^{L-1}w[n]x[n+iD]e^{-jn\omega}\right|^2\right)\\ &= \frac{1}{KLU}\sum_{i=0}^{K-1}\left|\sum_{n=0}^{L-1}w[n]x[n+iD]e^{-jn\omega}\right|^2\end{aligned} \tag{181}$$

照例分析指标，首先是偏差，与 Modified periodogram 一致，将 N 换成 L：

$$E\left\{\hat{P}_W(e^{j\omega})\right\} = E\left\{\hat{P}_M^{(i)}\right\} = \frac{1}{2\pi LU}P_x(e^{j\omega}) * \left|W(e^{j\omega})\right|^2 \tag{182}$$

方差与重叠程度有关，难以计算，我们只考虑 50% overlap 的情况：

$$\mathrm{Var}\left\{\hat{P}_W(e^{j\omega})\right\} \approx \frac{9}{8K}P_x^2(e^{j\omega}) \approx \frac{9}{16}\frac{L}{N}P_x^2(e^{j\omega}) \tag{183}$$

系数 $9/8$ 看似表明方差表现变差了一点，但由于片段数量 $K \approx \frac{2N}{L}$ 差不多翻了一倍，方差表现实际上是提升的。

9.1.4. Periodogram-based Methods Summary

参考书中还提到了一个 Blackman-Tukey 法，此处略去不表。前面除此之外的每种方法我们都分析了其均值、方差、分辨率等表现，在此做一个总结，如 Figure 8 所示。

	Definition $\hat{P}_x(e^{j\omega})$	Expectation	Variance (approx.)	Resolution $\Delta\omega$
Periodogram	$\frac{1}{N}\left\|\sum_{n=-\infty}^{\infty} x[n]w_R[n]e^{-j\omega n}\right\|^2$	$\frac{1}{2\pi}P_x(e^{j\omega}) * W_B(e^{j\omega})$	$P_x^2(e^{j\omega})$	$0.89\frac{2\pi}{N}$
Modified periodogram	$\frac{1}{NU}\left\|\sum_{n=-\infty}^{\infty} x[n]w[n]e^{-j\omega n}\right\|^2$	$\frac{1}{2\pi NU}P_x(e^{j\omega}) * \left\|W(e^{j\omega})\right\|^2$	$P_x^2(e^{j\omega})$	See Table 1
Bartlett's $N = KL$	$\frac{1}{N}\sum_{i=0}^{K-1}\left\|\sum_{n=0}^{L-1} x[n+iL]e^{-j\omega n}\right\|^2$	$\frac{1}{2\pi}P_x(e^{j\omega}) * W_B(e^{j\omega})$	$\frac{1}{K}P_x^2(e^{j\omega})$	$0.89K\frac{2\pi}{N}$
Welch's (50% overlap) $N = L + D(K-1)$	$\frac{1}{KLU}\sum_{i=0}^{K-1}\left\|\sum_{n=0}^{L-1} w[n]x[n+iD]e^{-j\omega n}\right\|^2$	$\frac{1}{2\pi LU}P_x(e^{j\omega}) * \left\|W(e^{j\omega})\right\|^2$	$\frac{9}{16}\frac{L}{N}P_x^2(e^{j\omega})$	Window dependent

Figure 8: Properties of a few commonly used windows with length N

注意，其中修正周期图的 $U = \frac{1}{N}\sum_{n=0}^{N-1}|w[n]|^2$，而 Welch 法中由于每个片段长度为 L，其 $U = \frac{1}{L}\sum_{n=0}^{L-1}|w[n]|^2$。

参考书中定义了两个指标来衡量以上方法的表现，其一为变异性（Variablilty）：

$$\mathcal{V} = \frac{\mathrm{Var}\left\{\hat{P}_x(e^{j\omega})\right\}}{E^2\left\{\hat{P}_x(e^{j\omega})\right\}} \tag{184}$$

说白了就是归一化的方差。其二是品质因数（Figure of merit）：

$$\mathcal{M} = \mathcal{V}\Delta\omega \tag{185}$$

是变异性和分辨率的乘积，这个值越小越好。顺带一提，这里品质因数这种定义为两个量相乘的指标，一般都是把 Trade-off 的量乘起来，所以我们会发现前面这些无参估计方法的品质因数都差不太多。

9.1.5. Minimum Variance (MV) Spectrum Estimation

首先，我们围绕 Section 9.1.1.1 中将信号送入滤波器组的思路来展开。在那一节中，我们滤波器是固定好的，和数据 $x[n]$ 无关，称为 data independent 的。这种情况下，如果不巧某些滤波器旁瓣滤到的能量有点多，就会导致明显的干扰。

本节中要介绍的 Minimum Variance (MV) Spectrum Estimation 方法的思路是根据输入信号 $x[n]$ 来为每个频率点设计滤波器，使得每个滤波器在滤波时：1、在目标频率点 ω_i 增益为一，无损通过；2、尽可能阻止旁瓣的能量通过。由此得到更好的估计结果。

我们先定义符号，令 $g_i[n]$ 是 p 阶复值 FIR 带通滤波器，为了满足**第一条要求**，应有：

$$G_i(e^{j\omega_i}) = \sum_{n=0}^{p} g_i[n]e^{-jn\omega_i} = 1 \tag{186}$$

方便起见我们写成向量形式，令 $\boldsymbol{g}_i = [g_i[0], g_i[1], ..., g_i[p]]^T$ 和 $\boldsymbol{e}_i = \left[1, e^{j\omega_i}, ..., e^{jp\omega_i}\right]^T$，则上述约束化为：

$$\boldsymbol{g}_i^H \boldsymbol{e}_i = \boldsymbol{e}_i^H \boldsymbol{g}_i = 1 \tag{187}$$

注意，虽然 $\boldsymbol{g}_i^H \boldsymbol{e}_i = \left(\boldsymbol{e}_i^H \boldsymbol{g}_i\right)^*$ 但这里令其为 1 了，结果已知是实数所以没事，但要记得这个等式只是这里成立。

第二条要求需要我们尽可能减小输出过程的功率，关于这个功率值，由 Equation 34 有：

$$E\left\{|y_i[n]|^2\right\} = \boldsymbol{g}_i^H \boldsymbol{R}_x \boldsymbol{g}_i \tag{188}$$

于是，我们要最小化这个值的同时，满足前面提到的线性约束，即求解：

$$\min_{\boldsymbol{g}_i} \boldsymbol{g}_i^H \boldsymbol{R}_x \boldsymbol{g}_i \quad \text{s.t. } \boldsymbol{e}_i^H \boldsymbol{g}_i = 1 \tag{189}$$

该问题的解为：

$$\begin{aligned} \boldsymbol{g}_i &= \frac{\boldsymbol{R}_x^{-1}\boldsymbol{e}_i}{\boldsymbol{e}_i^H \boldsymbol{R}_x^{-1}\boldsymbol{e}_i} \\ \min_{\boldsymbol{g}_i} \boldsymbol{g}_i^H \boldsymbol{R}_x \boldsymbol{g}_i &= \frac{1}{\boldsymbol{e}_i^H \boldsymbol{R}_x^{-1}\boldsymbol{e}_i} \end{aligned} \tag{190}$$

这是一个典型的可以用拉格朗日乘子法解决的优化问题，求解过程如下。令：

$$L(\boldsymbol{g}_i, \lambda) = \boldsymbol{g}_i^H \boldsymbol{R}_x \boldsymbol{g}_i - \lambda\left(\boldsymbol{e}_i^H \boldsymbol{g}_i - 1\right) \tag{191}$$

令其对两个参数的偏导分别为零得（$\boldsymbol{R}_x$ 是 Hermitian 的）：

$$\begin{cases} 2\boldsymbol{R}_x \boldsymbol{g}_i - \lambda \boldsymbol{e}_i = 0 \\ \boldsymbol{e}_i^H \boldsymbol{g}_i = 1 \end{cases} \tag{192}$$

由第一个式子有 $\boldsymbol{g}_i = \frac{\lambda}{2}\boldsymbol{R}_x^{-1}\boldsymbol{e}_i$，代入第二个式子并整理得：

$$\frac{\lambda}{2} = \frac{1}{\boldsymbol{e}_i^H \boldsymbol{R}_x^{-1}\boldsymbol{e}_i} \tag{193}$$

再代回前式得到解：

$$\boldsymbol{g}_i = \frac{\boldsymbol{R}_x^{-1}\boldsymbol{e}_i}{\boldsymbol{e}_i^H \boldsymbol{R}_x^{-1}\boldsymbol{e}_i} \tag{194}$$

再代回得最小值的解析式。

由于上面的推导对任意 ω_i 都成立，所以我们可以直接将式子中的下标去掉，并写为关于 ω 的函数：

$$\hat{\sigma}_x^2(\omega) = \frac{1}{\boldsymbol{e}^H \boldsymbol{R}_x^{-1} \boldsymbol{e}} \tag{195}$$

其中 $\boldsymbol{e} = \begin{bmatrix} 1 & e^{j\omega} & e^{j2\omega} & \dots & e^{jp\omega} \end{bmatrix}^T$。该过程对应的滤波器参数为：

$$\boldsymbol{g} = \frac{\boldsymbol{R}_x^{-1} \boldsymbol{e}}{\boldsymbol{e}^H \boldsymbol{R}_x^{-1} \boldsymbol{e}} \tag{196}$$

但 $\hat{\sigma}_x^2(\omega)$ 只是输出过程的功率估计，**还不能直接作为功率谱密度估计**，我们还需要除以滤波器的带宽，原因可类比 Equation 153。带宽有多种定义方式，我们可以直接取最简单的一种，令其代入后可使白噪声的例子得到正确的估计结果（见书第 429 页的例子）：

$$\frac{\Delta}{2\pi} = \frac{1}{p+1} \tag{197}$$

于是最终我们得到功率谱密度估计：

$$\hat{P}_{\text{MV}}(e^{j\omega}) = \frac{\hat{\sigma}_x^2(\omega)}{\Delta / 2\pi} = \frac{p+1}{\boldsymbol{e}^H \boldsymbol{R}_x^{-1} \boldsymbol{e}} \tag{198}$$

称为 **minimum variance spectrum estimate**，注意其中使用了随机过程的自相关矩阵 $\boldsymbol{R}_x$，如果我们只有样本数据，那么就需要使用估计的 $\hat{\boldsymbol{R}}_x$：

$$\begin{gathered} \hat{\boldsymbol{R}}_x = \frac{1}{K} \sum_{i=0}^{K-1} \boldsymbol{x}_i \boldsymbol{x}_i^H \\ \boldsymbol{x}_i = \begin{bmatrix} x[i] & x[i+1] & x[i+2] & \dots & x[i+L-1] \end{bmatrix}^T \end{gathered} \tag{199}$$

这里我们是将长度为 N 的信号样本交叠切分为 K 段，每段长度 L，起点间隔只有 $D = 1$，即 $K = N - L + 1$，如此来估计样本自相关矩阵。**为了维度匹配，前面的滤波器阶数同序列长度存在关系 $\boldsymbol{L = p + 1}$**，即有：

$$\hat{P}_{\text{MV}}(e^{j\omega}) = \frac{L}{\boldsymbol{e}^H \boldsymbol{R}_x^{-1} \boldsymbol{e}} \tag{200}$$

(TODO) 累了，毁灭吧。

9.2. Parametric Spectrum Estimation

9.2.1. For Autoregressive (AR) Models

(TODO) the Yule-Walker Method (autocorrelation method) and the covariance method.

9.2.2. Mulitple Signal Clasification (MUSIC)

(TODO)

10. Optimum Filtering

(TODO)

10.1. FIR Wiener Filter

10.1.1. Wiener-Hopf Equations

10.2. Discrete Kalman Filter

11. Adaptive Filtering

(TODO)

11.1. Least Mean Squares (LMS) Algorithm

11.2. Recursive Least Squares (RLS) Algorithm