时间序列分析及应用

4.6 AR 模型的预测

刘洋

四川师范大学数学科学学院

2022 年 4 月 2 日

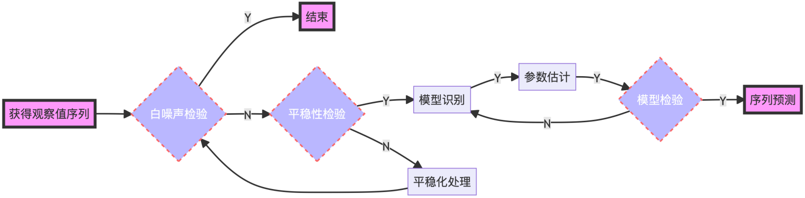

图 1: 时间序列建模流程图

给定时间序列 $X_1, \cdots, X_t$, 如何预测下一时刻序列值 X_{t+1} 或 k 步以后的序列值 X_{t+k} 是时间序列分析的主要目的之一.

人们希望预测得越"准确" 越好. 这里的准确性是在某种准则意义下的, 常见的准则有均方误差准则. 由于平稳序列可以中心化, 故本节仅考虑零均值且方差有限的平稳序列.

本节讨论平稳序列的最佳预测和最佳线性预测, 并将其应用于 AR 模型的预测问题.

1. 4.6.1 最佳预测
2. 4.6.2 最佳线性预测
3. 4.6.3 AR 模型的预测
4. 4.6.4 修正预测
5. 参考资料

最佳预测

设 $\hat{X}_t(k)$ 是由时间序列 $X_1, \cdots, X_t$ 对于 X_{t+k} 的预测值. 为了衡量预测值 $\hat{X}_t(k)$ 的精确程度, 常用的衡量准则为二次损失函数

$$E\left(X_{t+k} - \hat{X}_t(k)\right)^2. \tag{4.42}$$

因此, 我们希望预测值 $\hat{X}_t(k)$ 能使得二次损失函数 (4.42) 达到最小. 若能找到某一 $\hat{X}_t(k)$ 使得 (4.42) 达到最小, 则称 $\hat{X}_t(k)$ 为 X_{t+k} 的最佳预测. 记 $\boldsymbol{X}_t = (X_1, \cdots, X_t)^{\mathrm{T}}$, 我们有下面的结论.

最佳预测

定理 (4.4)

给定时间序列 $X_1, \cdots, X_t$, 则使得二次损失函数 *(4.42)* 达到最小的预测值为条件期望

$$\hat{X}_t(k) = E\left(X_{t+k} \mid \boldsymbol{X}_t\right)$$

最佳预测

证 对于某一预测函数 $f(\boldsymbol{X}_t)$, 其二次损失函数

$$
\begin{aligned}
& E\left(X_{t+k}-f\left(\boldsymbol{X}_t\right)\right)^2 \\
= & E\left(X_{t+k}-E\left(X_{t+k} \mid \boldsymbol{X}_t\right)+E\left(X_{t+k} \mid \boldsymbol{X}_t\right)-f\left(\boldsymbol{X}_t\right)\right)^2 \\
= & E\left(X_{t+k}-E\left(X_{t+k} \mid \boldsymbol{X}_t\right)\right)^2+2 E\left(\xi_{t+k}\right) \\
& +E\left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right)-f\left(\boldsymbol{X}_t\right)\right)^2,
\end{aligned} \tag{4.43}
$$

式中 $\xi_{t+k}=\left(X_{t+k}-E\left(X_{t+k} \mid \boldsymbol{X}_t\right)\right)\left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right)-f\left(\boldsymbol{X}_t\right)\right)$. 易知, $E\left(X_{t+k} \mid \boldsymbol{X}_t\right)$ 和 $f\left(\boldsymbol{X}_t\right)$ 是 $\boldsymbol{X}_t$ 的函数.

最佳预测

因此, 给定数据 $\boldsymbol{X}_t = (X_1, \cdots, X_t)^{\mathrm{T}}$, 我们有

$$
\begin{aligned}
& E\left(\xi_{t+k} \mid \boldsymbol{X}_t\right) \\
=& \left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right) - f\left(\boldsymbol{X}_t\right)\right) \times E\left(X_{t+k} - E\left(X_{t+k} \mid \boldsymbol{X}_t\right)\right) \\
=& \left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right) - f\left(\boldsymbol{X}_t\right)\right) \times \left(E X_{t+k} - E\left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right)\right)\right) \\
=& \left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right) - f\left(\boldsymbol{X}_t\right)\right) \times 0 \\
=& 0
\end{aligned}
\tag{4.44}
$$

上面倒数第二个等式成立的原因是条件期望有一个特殊的性质, 即条件期望的期望等于无条件期望.

最佳预测

从而, 由 (4.44) 式可得

$$E\left(\xi_{t+k}\right)=E\left(E\left(\xi_{t+k} \mid \boldsymbol{X}_t\right)\right)=0 \tag{4.45}$$

把 (4.45) 式带入 (4.43) 可得,

$$\begin{aligned} & E\left(X_{t+k}-f\left(\boldsymbol{X}_t\right)\right)^2 \\ = & E\left(X_{t+k}-E\left(X_{t+k} \mid \boldsymbol{X}_t\right)\right)^2+E\left(E\left(X_{t+k} \mid \boldsymbol{X}_t\right)-f\left(\boldsymbol{X}_t\right)\right)^2 \end{aligned} \tag{4.46}$$

显然, (4.46) 式中等式右边的第一项与预测函数 $f\left(\boldsymbol{X}_t\right)$ 无关, 而第二项是非负的.

最佳预测

欲使 (4.46) 式达到最小, 只需让预测函数

$$f\left(\boldsymbol{X}_t\right)=E\left(X_{t+k} \mid \boldsymbol{X}_t\right)$$

即可. 从而完成了定理 4.4 的证明.

由定理 4.4 可知, 给定数据 $\boldsymbol{X}_t$, 其 k 步最佳预测为其条件期望. 然而, 在实际应用中, 条件期望的具体表达式往往未知, 故难以具体实现. 因此, 人们往往考虑简单的线性预测模型.

最佳线性预测

给定某一中心化的平稳序列 $\boldsymbol{X}_t = (X_1, \cdots, X_t)^{\mathrm{T}}$, 考虑用 $X_1, \cdots, X_t$ 的某一线性组合

$$\mathbf{a}^{\mathrm{T}} \boldsymbol{X}_t = \sum_{i=1}^{t} a_i X_i$$

来预测 k 步之后的序列值 X_{t+k}, 其中 $\mathbf{a} = (a_1, \cdots, a_t)^{\mathrm{T}}$. 人们希望寻找某一向量 $\mathbf{a}$ 使得 $\mathbf{a}^{\mathrm{T}} \boldsymbol{X}_t$ 尽可能逼近 X_{t+k}.

一般地, 其逼近程度用二次损失函数 (4.42) 来衡量.

最佳线性预测

因此, 人们寻找某一向量 a 使得

$$E\left(X_{t+k}-\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t\right)^2 \tag{4.47}$$

达到最小. 设 $\mathbf{a}^*$ 使得 (4.47) 式达到最小, 我们称 $\left(\mathbf{a}^*\right)^{\mathrm{T}}\boldsymbol{X}_t$ 为 X_{t+k} 的最佳线性预测, 并记为 $\hat{X}_t(k)$ 或 $L\left(X_{t+k} \mid \boldsymbol{X}_t\right)$. 由于 $X_{t+k}-\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 为预测误差, 则最佳线性预测是所有线性预测中, 使得均方误差达到最小的预测.

最佳线性预测

比较 (4.42) 式和 (4.47) 式可知，最佳线性预测的预测效果往往比定理 4.4 中给出的最佳预测的预测效果差，因为最佳线性预测只是在线性类里面寻找最佳的向量，而最佳预测在均方误差意义下是最好的预测.

最佳线性预测

对于均值 μ 不为零的平稳序列 $Y_1, \cdots, Y_t$, 其中 $Y_i = X_i + \mu$, 则其 k 步最佳线性预测可表示为

$$\hat{Y}_t(k) = L\left(X_{t+k} \mid \boldsymbol{X}_t\right) + \mu,$$

即先对序列 $Y_1, \cdots, Y_t$ 都减去均值 μ 得到 $X_1, \cdots, X_t$, 再由其最佳线性预测加上均值 μ 即可. 显然, $E\hat{Y}_t(k) = \mu = E\left(Y_{t+k}\right)$.

最佳线性预测的求解

对于任意 t 维向量 $\boldsymbol{b}$, 我们有

$$
\begin{aligned}
& E\left(X_{t+k}-\boldsymbol{b}^{\mathrm{T}} \boldsymbol{X}_t\right)^2 \\
= & E\left(X_{t+k}-\mathbf{a}^{\mathrm{T}} \boldsymbol{X}_t+\mathbf{a}^{\mathrm{T}} \boldsymbol{X}_t-\boldsymbol{b}^{\mathrm{T}} \boldsymbol{X}_t\right)^2 \\
= & E\left(X_{t+k}-\mathbf{a}^{\mathrm{T}} \boldsymbol{X}_t\right)^2+E\left((\boldsymbol{a}-\boldsymbol{b})^{\mathrm{T}} \boldsymbol{X}_t\right)^2+ \\
& 2(\boldsymbol{a}-\boldsymbol{b})^{\mathrm{T}}\left(E\left(\boldsymbol{X}_t X_{t+k}\right)-E\left(\boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}}\right) \boldsymbol{a}\right)
\end{aligned}
\tag{4.48}
$$

最佳线性预测的求解

显然, 假如如下的关系成立

$$E\left(\boldsymbol{X}_t X_{t+k}\right)-E\left(\boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}}\right) \boldsymbol{a}=\mathbf{0}_t \tag{4.49}$$

其中 $\mathbf{0}_t$ 为 t 维零向量, 则由 (4.48) 式易知,

$$E\left(X_{t+k}-\mathbf{b}^{\mathrm{T}} \boldsymbol{X}_t\right)^2 \geqslant E\left(X_{t+k}-\mathbf{a}^{\mathrm{T}} \boldsymbol{X}_t\right)^2$$

最佳线性预测的求解

此时, $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}_t$ 为最佳线性预测, 且预测均方误差

$$\begin{aligned} E\left(X_{t+k}-\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}_t\right)^2 &= EX_{t+k}^2+\boldsymbol{a}^{\mathrm{T}}E\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\right)\boldsymbol{a}-2\boldsymbol{a}^{\mathrm{T}}E\left(\boldsymbol{X}_tX_{t+k}\right) \\ &= EX_{t+k}^2+\boldsymbol{a}^{\mathrm{T}}E\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\right)\boldsymbol{a}-2\boldsymbol{a}^{\mathrm{T}}E\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\right)\boldsymbol{a} \\ &= EX_{t+k}^2-\boldsymbol{a}^{\mathrm{T}}E\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\right)\boldsymbol{a} \end{aligned}$$

由上面的讨论可知, 假如向量 $\boldsymbol{a}$ 满足 (4.49) 式, 则相应的线性组合 $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}_t$ 即为 X_{t+k} 的最佳线性预测.

因此, 我们称 (4.49) 式为预测方程, 即可以通过解方程 (4.49) 来获得最佳线性预测的向量 a.

最佳线性预测的求解

记 $\mathbf{\Gamma} = E\left(\boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}}\right), \boldsymbol{\eta}_k = E\left(\boldsymbol{X}_t X_{t+k}\right)$, 则预测方程 (4.49) 即为一个非齐次线性方程组

$$\mathbf{\Gamma}\mathbf{a} = \boldsymbol{\eta}_k, \tag{4.50}$$

其中 $\mathbf{\Gamma} \in \mathbb{R}^{t\times t}, \boldsymbol{\eta}_k \in \mathbb{R}^t$. 易知 $\mathbf{\Gamma}$ 为半正定矩阵, 因为对于任意的 t 维向量 $\boldsymbol{c}$, $\boldsymbol{c}^{\mathrm{T}}\mathbf{\Gamma}\boldsymbol{c} = E\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{X}_t\right)^2 \geqslant 0$. 假设矩阵 $\mathbf{\Gamma}$ 可逆, 则由预测方程 (4.50) 式, 易知解向量为

$$\hat{\mathbf{a}} = \mathbf{\Gamma}^{-1}\boldsymbol{\eta}_k \tag{4.51}$$

假如 $\mathbf{\Gamma}$ 不可逆, 即其行列式 $\det(\mathbf{\Gamma}) = 0$, 则由线性代数的知识可知预测方程 (4.50) 的解向量仍然存在. 下面给出其一个解向量.

最佳线性预测的求解

对于半正定矩阵 $\mathbf{\Gamma}$, 总存在一个正交矩阵 $\boldsymbol{Q}$ 使 $\mathbf{\Gamma}$ 对角化, 即

$$\boldsymbol{Q}\mathbf{\Gamma}\boldsymbol{Q}^{\mathrm{T}} = E\left(\boldsymbol{Q}\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\boldsymbol{Q}^{\mathrm{T}}\right) = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0\right)$$

其中 $\lambda_1 \geqslant \cdots \geqslant \lambda_r > 0$ 为 $\mathbf{\Gamma}$ 的 r 个大于零的特征值. 记 $\boldsymbol{Y} = \boldsymbol{Q}\boldsymbol{X}_t = \left(Y_1, \cdots, Y_t\right)^{\mathrm{T}}$, 并设 $\boldsymbol{\xi}$ 为 $\boldsymbol{Y}$ 的前 r 个随机变量, 即 $\boldsymbol{\xi} = \left(Y_1, \cdots, Y_r\right)^{\mathrm{T}}, \mathbf{0}_{t-r}$ 为 $t-r$ 维零向量, 则

$$E\left(\boldsymbol{\xi}\boldsymbol{\xi}^{\mathrm{T}}\right) = \mathrm{diag}\left(\lambda_1, \cdots, \lambda_r\right), \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{\xi} \\ \mathbf{0}_{t-r} \end{bmatrix}, \text{ a.s..}$$

最佳线性预测的求解

若记 $\zeta = \left[E\left(\xi\xi^{\mathrm{T}}\right)\right]^{-1} E\left(\boldsymbol{\xi} X_{t+k}\right)$, 则可以验证

$$\hat{\mathbf{a}} = \boldsymbol{Q}^{\mathrm{T}} \left[\begin{array}{c} \zeta \\ \mathbf{0}_{t-r} \end{array} \right] \tag{4.52}$$

满足预测方程 (4.50),

最佳线性预测的求解

因为

$$
\begin{aligned}
\boldsymbol{Q}\boldsymbol{\Gamma}\hat{\mathbf{a}} &= \boldsymbol{Q}\boldsymbol{\Gamma}\boldsymbol{Q}^{\mathrm{T}}\begin{bmatrix} \boldsymbol{\zeta} \\ \mathbf{0}_{t-r} \end{bmatrix} \\
&= \operatorname{diag}\left(\lambda_1, \cdots, \lambda_r, 0, \cdots, 0\right)\begin{bmatrix} \boldsymbol{\zeta} \\ \mathbf{0}_{t-r} \end{bmatrix} \\
&= \begin{bmatrix} E\left(\boldsymbol{\xi} X_{t+k}\right) \\ \mathbf{0}_{t-r} \end{bmatrix} = E\left[\begin{pmatrix} \boldsymbol{\xi} \\ \mathbf{0}_{t-r} \end{pmatrix} X_{t+k}\right] \\
&= E\left[\boldsymbol{Y} X_{t+k}\right] = E\left[\boldsymbol{Q}\boldsymbol{X}_t X_{t+k}\right] \\
&= \boldsymbol{Q}\eta_k
\end{aligned}
$$

上面等式分别左乘矩阵 $\boldsymbol{Q}^{\mathrm{T}}$, 即得 (4.50) 式. 因此 (4.52) 式是预测方程 (4.50) 式的一个解.

最佳线性预测的求解

当 $\mathbf{\Gamma}$ 不可逆时, 根据矩阵论的知识, 解向量不是唯一的. 设 $\hat{\mathbf{a}}$ 如 (4.52) 所示, 则 $\hat{\mathbf{a}}^{\mathrm{T}}\boldsymbol{X}_t$ 为 X_{t+k} 的最佳线性预测. 对于任意其他向量 $\boldsymbol{b}\in\mathbb{R}^{\mathrm{T}}$, 根据 (4.48) 式, 我们有

$$E\left(X_{t+k}-\mathbf{b}^{\mathrm{T}}\boldsymbol{X}_t\right)^2=E\left(X_{t+k}-\hat{\mathbf{a}}^{\mathrm{T}}\boldsymbol{X}_t\right)^2+E\left((\hat{\mathbf{a}}-\mathbf{b})^{\mathrm{T}}\boldsymbol{X}_t\right)^2$$

因此, 若 $\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}_t$ 也是最佳线性预测, 则有 $E\left((\hat{\mathbf{a}}-\boldsymbol{b})^{\mathrm{T}}\boldsymbol{X}_t\right)^2=0$, 即

$$\hat{\mathbf{a}}^{\mathrm{T}}\boldsymbol{X}_t=\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}_t,\qquad \text{a.s.}$$

换句话说, 最佳线性预测是几乎处处唯一的.

最佳线性预测的求解

综上所述, 我们有如下的结论. 定理 4.5 对于预测方程 (4.49) 式或 (4.50) 式, 当 Γ 可逆时, 最佳线性预测的解向量 $\hat{\mathbf{a}}$ 如 (4.51) 式所示, 当 $\mathbf{\Gamma}$ 不可逆时, 解向量 $\hat{\mathbf{a}}$ 可采用 (4.52) 式.

需要指出的是, 定理 4.5 中讨论的是 $k \geqslant 1$ 的情形. 当 $k \leqslant 0$ 时, 即需预测的时刻只是当前或历史时刻, 则易证其最佳预测和最佳线性预测都是其本身, 即

$$L\left(X_{t+k} \mid \boldsymbol{X}_t\right) = X_{t+k}, \text{当} k \leqslant 0.$$

最佳线性预测的求解

由 (4.49) 式可知,

$$E\left(\left(X_{t+k} - \mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t\right)\boldsymbol{X}_t^{\mathrm{T}}\right) = \mathbf{0}_t^{\mathrm{T}}, \tag{4.53}$$

即预测误差 $X_{t+k} - \mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 和 $\boldsymbol{X}_t$ 线性无关. 我们称预测 $\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 为 X_{t+k} 关于 $\boldsymbol{X}_t$ 的线性投影. 因此, 在线性预测类中, 线性投影是均方误差最小的线性预测.

对于任意的 t 维列向量 $\boldsymbol{b}$,

$$E\left(\left(X_{t+k} - \mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t\right)\left(\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}_t\right)\right) = 0,$$

因此预测误差 $X_{t+k} - \mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 垂直于由 $\boldsymbol{X}_t$ 张成的线性空间 $H = \left\{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}_t, \boldsymbol{b} \in \mathbb{R}^t\right\}$. 因此, $\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 是 X_{t+k} 在空间 H 上的投影.

最佳线性预测的求解

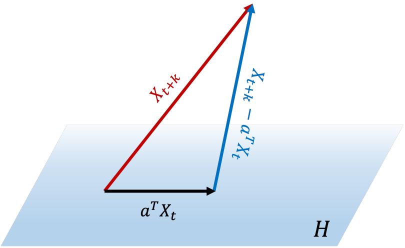

图 2: 投影演示

预测误差 $X_{t+k} - \mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 垂直于由 $\boldsymbol{X}_t$ 张成的线性空间 $H = \{\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}_t, \boldsymbol{b} \in \mathbb{R}^t\}$. 因此, $\mathbf{a}^{\mathrm{T}}\boldsymbol{X}_t$ 是 X_{t+k} 在空间 H 上的投影.

最佳线性预测与最小二乘回归的关系

线性预测是由已有的序列 $\boldsymbol{X}_t$ 的某个线性组合对将来的序列值进行预测, 因此 X_{t+k} 可表示为

$$X_{t+k} = \boldsymbol{a}^{\mathrm{T}} \boldsymbol{X}_t + \varepsilon_{t+k},$$

其中没有常数项的原因是该序列的均值为零. 上式与线性回归的形式非常相似.

最小二乘原理

最小二乘原理

y_i

最小二乘原理

最小二乘原理

$$s(x_i; a_0, \cdots, a_n) \qquad y_i$$

最小二乘原理

最小二乘原理

$$s(x_i; a_0, \cdots, a_n) - y_i$$

最小二乘原理

最小二乘原理

$$| s(x_i; a_0, \cdots, a_n) - y_i |$$

最小二乘原理

最小二乘原理

$$(\ s(x_i; a_0, \cdots, a_n) - y_i\ )^2$$

最小二乘原理

最小二乘原理

$$\sum_{i=1}^{m} \left(s(x_i; a_0, \cdots, a_n) - y_i \right)^2$$

最小二乘原理

$$\mathbf{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} \left(s(x_i;a_0,\cdots,a_n) - y_i \right)^2$$

最小二乘原理

最小二乘原理

$$min \quad \mathbf{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} \left(s(x_i;a_0,\cdots,a_n) - y_i \right)^2$$

最小二乘原理

选择拟合函数类 (模型)

$$min \quad \mathbf{F}(a_0, \cdots, a_n) := \sum_{i=1}^{m} (\, s(x_i; a_0, \cdots, a_n) - y_i \,)^2$$

最小二乘原理

选择拟合函数类 (模型)

$$min \quad \mathbf{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} \left(s(x_i;a_0,\cdots,a_n) - y_i \right)^2$$

最小二乘原理

选择拟合函数类 (模型)

$$\min \quad \mathrm{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} \left(s(x_i; a_0,\cdots,a_n) - y_i \right)^2$$

函数类

函数类$\Phi = \mathrm{Span}\{\varphi_0(x), \varphi_1(x), \cdots, \varphi_n(x)\}$

最小二乘原理

选择拟合函数类 (模型)

$$min \quad \mathbf{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} (\, s(x_i;a_0,\cdots,a_n) - y_i \,)^2$$

函数类

函数类$\Phi = \mathrm{Span}\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$

常用的函数系:幂函数系$\{x^j\}$,
三角函数系$\{sinjx\},\{cosjx\}$,
指数函数系$\{e^{\lambda_j x}\}$,正交函数系等

最小二乘原理

选择拟合函数类 (模型)

$$\min \quad \mathrm{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} \left(s(x_i; a_0,\cdots,a_n) - y_i \right)^2$$

函数类

模型类

函数类$\Phi = \mathrm{Span}\{\varphi_0(x), \varphi_1(x), \cdots, \varphi_n(x)\}$

{线性回归, 决策树, 感知机, 神经网络, 支持向量机}

常用的函数系: 幂函数系$\{x^j\}$,
三角函数系$\{sinjx\}, \{cosjx\}$,
指数函数系$\{e^{\lambda_j x}\}$, 正交函数系等

最小二乘原理

选择拟合函数类（模型)

$$\min \quad \mathbf{F}(a_0,\cdots,a_n) := \sum_{i=1}^{m} \left(s(x_i;a_0,\cdots,a_n) - y_i \right)^2$$

函数类

模型类

函数类$\Phi = \mathrm{Span}\{\varphi_0(x),\varphi_1(x),\cdots,\varphi_n(x)\}$

{线性回归, 决策树, 感知机, 神经网络, 支持向量机}

常用的函数系: 幂函数系$\{x^j\}$,
三角函数系$\{sinjx\}$, $\{cosjx\}$,
指数函数系$\{e^{\lambda_j x}\}$, 正交函数系等

超参数

最小二乘原理

数据拟合是根据测定的数据间的相互关系, 确定曲线 $y = s(x; a_0, a_1, \cdots, a_n)$ 的类型, 然后再根据在给定点上误差的平方和达到最小的原则, 即求解无约束问题

$$\min F(a_0, a_1, \cdots, a_n) = \sum_{i=1}^{m} \left(s(x_i; a_0, a_1, \cdots, a_n) - y_i\right)^2 \tag{1}$$

确定出最优参数 $a_k^*(k = 0, 1, \cdots, n)$, 从而得到拟合曲线 $y = s^*(x)$.

拟合算法步骤:

最小二乘原理

拟合算法步骤:

- Step1. 获取数据集 $Data = \{(x_i, y_i) | i = 1, 2, \cdots, N\}$

最小二乘原理

拟合算法步骤:

- Step1. 获取数据集 $Data = \{(x_i, y_i) | i = 1, 2, \cdots, N\}$
- Step2. 分析数据集分布情况确定拟合函数类 Φ 并确定其超参数.
- Step3. 选择某种度量方式刻画“拟合结果数据”与“真实数据”之间的分布差距 (误差).

最小二乘原理

拟合算法步骤:

- Step1. 获取数据集 $Data = \{(x_i, y_i) | i = 1, 2, \cdots, N\}$
- Step2. 分析数据集分布情况确定拟合函数类 Φ 并确定其超参数.
- Step3. 选择某种度量方式刻画“拟合结果数据”与“真实数据”之间的分布差距 (误差).
- Step4. 通过使用某种优化算法最小化损失 (误差) 函数得到拟合函数模型.

最小二乘原理

拟合算法步骤:

- Step1. 获取数据集 $Data = \{(x_i, y_i) | i = 1, 2, \cdots, N\}$
- Step2. 分析数据集分布情况确定拟合函数类 Φ 并确定其超参数.
- Step3. 选择某种度量方式刻画“拟合结果数据”与“真实数据”之间的分布差距 (误差).
- Step4. 通过使用某种优化算法最小化损失 (误差) 函数得到拟合函数模型.

最佳线性预测与最小二乘回归的关系

假设观测值序列 $\boldsymbol{x}_t = (x_{t-n+1}, x_{t-n+2}, \cdots, x_t)^{\mathrm{T}}$, 给定 T 个数据对 $\{x_{t+k}, \boldsymbol{x}_t\}$, $t = 1, \cdots, T$, 则 x_{t+k} 的线性预测为

$$x_{t+k} = b^T x_t + \varepsilon_{t+k}.$$

最佳线性预测与最小二乘回归的关系

则使残差平方和

$$\sum_{t=1}^{T}\left(x_{t+k}-\boldsymbol{b}^{\mathrm{T}}\boldsymbol{x}_t\right)^2$$

达到最小的参数 $\boldsymbol{b}$ 的最小二乘解为

$$\begin{aligned}\boldsymbol{b} &= \left[\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t^{\mathrm{T}}\right]^{-1}\left[\sum_{t=1}^{T}\boldsymbol{x}_t x_{t+k}\right] \\ &= \left[\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t\boldsymbol{x}_t^{T}\right]^{-1}\left[\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{x}_t x_{t+k}\right]\end{aligned} \tag{4.54}$$

最佳线性预测与最小二乘回归的关系

比较 (4.54) 式和最佳线性预测的参数估计 (4.51) 式可知, (4.54) 式中的参数 $\boldsymbol{b}$ 是由样本矩 $\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t^{\mathrm{T}}$ 和 $\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t x_{t+k}$ 估计的, 而 (4.51) 式中的参数 $\boldsymbol{a}$ 是由总体矩 $E\left(\boldsymbol{X}_t\boldsymbol{X}_t^{\mathrm{T}}\right)$ 和 $E\left(\boldsymbol{X}_t X_{t+k}\right)$ 估计的.

因此, 线性回归是对观测序列的处理, 而最佳线性预测是对于总体信息的处理. 不过对于平稳序列 $\boldsymbol{X}_t$, 当该序列具有各态历经 (遍历) 性时, 可以证明, 随着数据量 T 的增大, 样本矩将依概率收玫于总体矩. 因此, $\boldsymbol{b}$ 也依概率收敛于 $\boldsymbol{a}$. 所以, 线性回归得到的系数是最佳线性预测的参数的一个一致估计.

最佳线性预测和最佳预测相等的情形

一般而言，在均方误差意义下，最佳预测比最佳线性预测效果好. 然而，在某些情形下，这两者是相等的. 下面给出最佳预测和最佳线性预测相等的两种特殊情形. 需要指出的是，这里的相等都是在几乎处处 (a.s.) 意义下相等的.

定理 (4.6)

对于零均值的时间序列 X_t, 若 $(X_1, \cdots, X_t, X_{t+k})^{\mathrm{T}}$ 服从多元正态分布 $N(\mathbf{0}, \Sigma)$, 则最佳预测和最佳线性预测相等, 即

$$E\left(X_{t+k} \mid \boldsymbol{X}_t\right) = L\left(X_{t+k} \mid \boldsymbol{X}_t\right)$$

最佳线性预测和最佳预测相等的情形

证设 $\hat{X}_t(k)=L\left(X_{t+k}\mid \boldsymbol{X}_t\right)$, 则根据 (4.53) 式可知，预测误差 $X_{t+k}-\hat{X}_t(k)$ 和 $\boldsymbol{X}_t$ 线性无关, 根据多元正态分布的性质可知, $X_{t+k}-\hat{X}_t(k)$ 和 $\boldsymbol{X}_t$ 相互独立. 从而 $X_{t+k}-\hat{X}_t(k)$ 与任意可测函数 $g(\cdot)$ 作用于 $\boldsymbol{X}_t$ 而得到的随机变量 $g\left(\boldsymbol{X}_t\right)$ 都相互独立.

最佳线性预测和最佳预测相等的情形

从而对于任意的预测函数 $g(\boldsymbol{X}_t)$, 我们有

$$
\begin{aligned}
E\left(X_{t+k}-g\left(\boldsymbol{X}_t\right)\right)^2 =& E\left(X_{t+k}-\hat{X}_t(k)+\hat{X}_t(k)-g\left(\boldsymbol{X}_t\right)\right)^2 \\
=& E\left(X_{t+k}-\hat{X}_t(k)\right)^2+E\left(\hat{X}_t(k)-g\left(\boldsymbol{X}_t\right)\right)^2+ \\
& 2E\left(\left(X_{t+k}-\hat{X}_t(k)\right)\left(\hat{X}_t(k)-g\left(\boldsymbol{X}_t\right)\right)\right) \\
=& E\left(X_{t+k}-\hat{X}_t(k)\right)^2+E\left(\hat{X}_t(k)-g\left(\boldsymbol{X}_t\right)\right)^2 \\
\geqslant& E\left(X_{t+k}-\hat{X}_t(k)\right)^2
\end{aligned}
\tag{4.55}
$$

最佳线性预测和最佳预测相等的情形

上面倒数第一个等式成立的原因是在于, $\hat{X}_t(k)$ 和 $g(\boldsymbol{X}_t)$ 都是由 $\boldsymbol{X}_t$ 变化而来的, 从而 $X_{t+k}-\hat{X}_t(k)$ 与 $\hat{X}_t(k)-g(\boldsymbol{X}_t)$ 相互独立, 故根据序列的零均值可知,

$$E\left[\left(X_{t+k}-\hat{X}_t(k)\right)\left(\hat{X}_t(k)-g(\boldsymbol{X}_t)\right)\right]=0.$$

则由最佳预测的定义及 (4.55) 式可知, $\hat{X}_t(k)$ 即为最佳预测.

定理 4.6 说明当时间序列为零均值的正态过程时, 最佳预测与最佳线性预测是相等的. 需要注意的是, 该结论并没有要求序列必须满足平稳性. 进一步的, 若取消零均值的限制时, 该结论仍然成立. 具体证明过程留给读者.

最佳线性预测和最佳预测相等的情形

定理 (4.7)

若平稳序列 X_t 可以表示为

$$X_t = \sum_{j=0}^{\infty} a_j \varepsilon_{t-j} \tag{4.56}$$

其中 ε_t 为白噪声序列, 则最佳预测和最佳线性预测相等的充要条件为

$$E\left(\varepsilon_{t+1} \mid \varepsilon_t, \varepsilon_{t-1}, \cdots\right) = 0 \tag{4.57}$$

最佳线性预测和最佳预测相等的情形

定理 4.7 的证明过程可参考何书元 (2003). 特别地, 当 $\{\varepsilon_t\}$ 为高斯白噪声时, (4.57) 式显然成立. 由于 $\mathrm{AR}(p)$ 序列 X_t 可以表示为 (4.56) 式, 因此, 由定理 4.7 可知, 当 ε_t 为独立白噪声时, 最佳预测与最佳线性预测是相等的.

特别地, ε_t 为高斯白噪声时, 任意两个时刻的噪声都相互独立, 故 (4.57) 式成立, 此时最佳预测与最佳线性预测是相等的.

AR 模型的预测

假设平稳序列 X_t 服从 $\mathrm{AR}(p)$ 模型

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + \varepsilon_t$$

其中 $\{\varepsilon_t\}$ 为零均值白噪声. 考虑用 $\boldsymbol{X}_t = (X_1, \cdots, X_t)^{\mathrm{T}}$ 预测 k 步之后的序列值 X_{t+k}.

$\mathbf{AR}(p)$ 序列的最佳预测

考虑 $\{\varepsilon_t\}$ 为零均值高斯白噪声情形下的 $\mathrm{AR}(p)$ 序列 X_t 的最佳预测. 根据定理 4.4, X_{t+k} 的最佳预测 $\hat{X}_t(k)$ 为条件期望 $E\left(X_{t+k} \mid \boldsymbol{X}_t\right)$. 当 $k=1$ 时,

$$
\begin{aligned}
\hat{X}_t(1) &= E\left(X_{t+1} \mid \boldsymbol{X}_t\right) \\
&= E\left(\left(a_1 X_t + a_2 X_{t-1} + \cdots + a_p X_{t+1-p} + \varepsilon_{t+1}\right) \mid \boldsymbol{X}_t\right) \\
&= E\left(\left(a_1 X_t + a_2 X_{t-1} + \cdots + a_p X_{t+1-p}\right) \mid \boldsymbol{X}_t\right) + E\left(\varepsilon_{t+1} \mid \boldsymbol{X}_t\right) \\
&= \sum_{j=1}^{p} a_j X_{t+1-j},
\end{aligned}
$$

$\mathbf{AR}(p)$ 序列的最佳预测

上面最后一个等式成立的原因如下:

已知序列 $\boldsymbol{X}_t = (X_1, \cdots, X_t)^{\mathrm{T}}$, 则条件期望 $E\left(X_{t+1-j} \mid \boldsymbol{X}_t\right) = X_{t+1-j}, j = 1, \cdots, p$; 且 ε_{t+1} 为将来时刻的噪声, 与已有数据 $\boldsymbol{X}_t$ 独立, 故 $E\left(\varepsilon_{t+1} \mid \boldsymbol{X}_t\right) = E\left(\varepsilon_{t+1}\right) = 0.$

$\mathbf{AR}(p)$ 序列的最佳预测

当 $k=2$ 时, 最佳预测 $\hat{X}_t(2)$ 为条件期望 $E\left(X_{t+2} \mid \boldsymbol{X}_t\right)$, 因此,

$$
\begin{aligned}
\hat{X}_t(2)= & E\left(X_{t+2} \mid \boldsymbol{X}_t\right) \\
= & E\left(\left(a_1 X_{t+1}+a_2 X_t+\cdots+a_p X_{t+2-p}+\varepsilon_{t+2}\right) \mid \boldsymbol{X}_t\right) \\
= & a_1 E\left(X_{t+1} \mid \boldsymbol{X}_t\right)+E\left(\left(a_2 X_t+a_3 X_{t-1}+\cdots+a_p X_{t+2-p}\right) \mid \boldsymbol{X}_t\right) \\
& +E\left(\varepsilon_{t+2} \mid \boldsymbol{X}_t\right) \\
= & a_1 \hat{X}_t(1)+a_2 X_t+a_3 X_{t-1}+\cdots+a_p X_{t+2-p}
\end{aligned}
$$

AR(p) 序列的最佳预测

当 $k=3$ 时，最佳预测 $\hat{X}_t(3)$ 可表示为，

$$
\begin{aligned}
\hat{X}_t(3) =& E\left(X_{t+3} \mid \boldsymbol{X}_t\right) \\
=& E\left(\left(a_1 X_{t+2}+a_2 X_{t+1}+a_3 X_t+\cdots+a_p X_{t+3-p}+\varepsilon_{t+3}\right) \mid \boldsymbol{X}_t\right) \\
=& a_1 E\left(X_{t+2} \mid \boldsymbol{X}_t\right)+a_2 E\left(X_{t+1} \mid \boldsymbol{X}_t\right) \\
& +E\left(\left(a_3 X_t+a_4 X_{t-1}+\cdots+a_p X_{t+3-p}\right) \mid \boldsymbol{X}_t\right)+E\left(\varepsilon_{t+3} \mid \boldsymbol{X}_t\right) \\
=& a_1 \hat{X}_t(2)+a_2 \hat{X}_t(1)+a_3 X_t+a_4 X_{t-1}+\cdots+a_p X_{t+3-p}
\end{aligned}
$$

$\mathbf{AR}(p)$ 序列的最佳预测

由数学归纳法易知，当 $k>1$ 时，$\hat{X}_t(k)$ 的最佳预测为

$$\hat{X}_t(k)=\begin{cases}\sum\limits_{j=1}^{k-1} a_j\hat{X}_t(k-j)+\sum\limits_{j=k}^{p} a_jX_{t+k-j}, & k\leqslant p\\ \sum\limits_{j=1}^{p} a_j\hat{X}_t(k-j), & k>p\end{cases}$$

$\mathbf{AR}(p)$ 序列的最佳预测

综上所述, 当 $\{\varepsilon_t\}$ 为零均值高斯白噪声时, $\mathrm{AR}(p)$ 序列 X_t 的 k 步最佳预测 $\hat{X}_t(k)$ 可表示为

$$\hat{X}_t(k)=\begin{cases}\sum_{j=1}^{p}a_jX_{t+1-j}, & k=1\\ \sum_{j=1}^{k-1}a_j\hat{X}_t(k-j)+\sum_{j=k}^{p}a_jX_{t+k-j}, & 1<k\leqslant p\\ \sum_{j=1}^{p}a_j\hat{X}_t(k-j), & k>p\end{cases}\tag{4.58}$$

从 (4.58) 式可知, 若已知 $\mathrm{AR}(p)$ 序列的系数 $a_1,\cdots,a_p$, 我们只用数据 $\{X_t,X_{t-1},\cdots,X_{t+1-p}\}$ 即可预测 X_{t+k}, 而不需要其他数据. 因此, 该预测式子非常简单, 这也是实践中常用 AR 模型的原因之一. 然而, 当 ε_t 不是独立序列时, (4.58) 式往往不再成立, 即此时最佳预测难以有显式表达式.

AR(p) 序列的最佳线性预测

由前一小节的讨论可知，当 $\{\varepsilon_t\}$ 为高斯白噪声时，最佳线性预测与最佳预测是一致的. 然而，当 $\{\varepsilon_t\}$ 不是独立白噪声时，最佳线性预测与最佳预测一般不相等. 然而当历史数据足够多时，最佳线性预测往往近似于最佳预测. 并且由于最佳线性预测有显式表达式，易于操作，人们往往考虑 AR(p) 序列的最佳线性预测.

$\mathbf{AR}(p)$ 序列的最佳线性预测

假设 $\{\varepsilon_t\}$ 为白噪声. 当 $1 \leqslant t \leqslant p-1$ 时, 自协方差矩阵 $\boldsymbol{\Gamma} = E\left(\boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}}\right)$ 往往可逆. 此时, 根据 (4.51) 式, X_{t+k} 的最佳线性预测的预测向量 $\boldsymbol{a} = \boldsymbol{\Gamma}^{-1}\boldsymbol{\eta}_k$, 其中 $\boldsymbol{\eta}_k = E\left(\boldsymbol{X}_t X_{t+k}\right)$. 则最佳线性预测为

$$\hat{X}_t(k) = L\left(X_{t+k} \mid \boldsymbol{X}_t\right) = \boldsymbol{\eta}_k^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} \boldsymbol{X}_t$$

预测的均方误差为

$$E\left(X_{t+k} - \hat{X}_t(k)\right)^2 = EX_{t+k}^2 - \boldsymbol{a}^{\mathrm{T}} E\left(\boldsymbol{X}_t \boldsymbol{X}_t^{\mathrm{T}}\right)\boldsymbol{a} = \gamma_0 - \boldsymbol{\eta}_k^{\mathrm{T}} \boldsymbol{\Gamma}^{-1} \boldsymbol{\eta}_k \tag{4.59}$$

$\mathbf{AR}(p)$ 序列的最佳线性预测

由于 $\mathrm{AR}(p)$ 序列 X_t 可表示为

$$X_t = \sum_{j=0}^{\infty} g_j \varepsilon_{t-j}$$

则根据白噪声的性质可知, X_t 与 ε_{t+k} 线性无关. 从而当 $t \geqslant p$ 时,

$$\begin{aligned}\hat{X}_t(k) &= L\left(\sum_{j=1}^{p} a_j X_{t+k-j} + \varepsilon_{t+k} \mid \boldsymbol{X}_t\right) = L\left(\sum_{j=1}^{p} a_j X_{t+k-j} \mid \boldsymbol{X}_t\right) \\ &= \sum_{j=1}^{p} a_j L\left(X_{t+k-j} \mid \boldsymbol{X}_t\right).\end{aligned}$$

$\mathbf{AR}(p)$ 序列的最佳线性预测

当 $j \geqslant k$ 时, $L\left(X_{t+k-j} \mid \boldsymbol{X}_t\right) = X_{t+k-j}$. 因此, X_{t+k} 的最佳线性预测可表示为

$$
\begin{aligned}
\hat{X}_t(k) &= L\left(X_{t+k} \mid \boldsymbol{X}_t\right) \\
&= \begin{cases} \sum_{j=1}^{p} a_j X_{t+1-j}, & k = 1, \\ \sum_{j=1}^{k-1} a_j \hat{X}_t(k-j) + \sum_{j=k}^{p} a_j X_{t+k-j}, & 1 < k \leqslant p, \\ \sum_{j=1}^{p} a_j \hat{X}_t(k-j), & k > p \end{cases}
\end{aligned}
\tag{4.60}
$$

上式和 (4.58) 式非常类似, (4.60) 式只不过把 (4.58) 式中最佳预测替换为最佳线性预测而已. 然而这两个式子的条件是有所不同的. (4.58) 式中要求噪声为高斯白噪声, 而 (4.60) 式中仅要求噪声为白噪声即可.

$\mathbf{AR}(p)$ 序列的最佳线性预测

例 (4.12)

$\mathrm{AR}(1)$ 模型的预测. 设

$$X_t = a_1 X_{t-1} + \varepsilon_t.$$

其中 ε_t 为白噪声.

记 $\boldsymbol{X}_t = (X_1, \cdots, X_t)^{\mathrm{T}}$, 则根据 (4.60) 式, X_{t+k} 的最佳线性预测为

$$\hat{X}_t(k) = L\left(X_{t+k} \mid \boldsymbol{X}_t\right) = a_1 \hat{X}_t(k-1) = a_1^2 \hat{X}_t(k-2) = \cdots = a_1^k X_t \tag{4.61}$$

则预测均方误差为 $E\left(X_{t+k} - \hat{X}_t(k)\right)^2 = E\left(X_{t+k} - a_1^k X_t\right)^2$.

$\mathbf{AR}(p)$ 序列的最佳线性预测

根据 AR(1) 序列的平稳性, 可知 $|a_1| < 1$. 则当 $k \to \infty$ 时, $a_1^k \to 0$. 故预测均方误差

$$E\left(X_{t+k} - \hat{X}_t(k)\right)^2 \to EX_{t+k}^2 = \gamma_0, \quad k \to \infty$$

上式说明, 当预测的步长越大, 预测的精度越低. 当 $\{\varepsilon_t\}$ 为高斯白噪声时, (4.61) 式也是 AR(1) 模型的最佳预测.

$\mathbf{AR}(p)$ 序列的最佳线性预测

例 (4.13)

AR(2) 模型的预测. 设

$$X_t = a_1 X_{t-1} + a_2 X_{t-2} + \varepsilon_t$$

其中 $\{\varepsilon_t\}$ 为白噪声.

$\mathbf{AR}(p)$ 序列的最佳线性预测

根据 (4.60) 式, X_{t+k} 的最佳线性预测为

$$
\begin{aligned}
\hat{X}_t(1) &= L\left(X_{t+1} \mid \boldsymbol{X}_t\right) = a_1 X_t + a_2 X_{t-1} \\
\hat{X}_t(2) &= L\left(X_{t+2} \mid \boldsymbol{X}_t\right) = a_1 \hat{X}_t(1) + a_2 X_t \\
\hat{X}_t(3) &= L\left(X_{t+3} \mid \boldsymbol{X}_t\right) = a_1 \hat{X}_t(2) + a_2 \hat{X}_t(1) \\
&\cdots \\
\hat{X}_t(k) &= L\left(X_{t+k} \mid \boldsymbol{X}_t\right) = a_1 \hat{X}_t(k-1) + a_2 \hat{X}_t(k-2)
\end{aligned}
$$

即 AR(2) 模型的预测可以采用递推的形式得到. 从中可见，最佳线性预测 $\hat{X}_t(k)$ 最终可以由 X_t 和 X_{t-1} 的线性组合表出.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

前面讨论的是有限历史数据的情形. 若历史数据为无穷多, 即已知数据 $\{X_t, X_{t-1}, \cdots\}$, t 为当前时刻, 则 $\mathrm{AR}(p)$ 序列的最佳预测有另外一种表达形式, 而且最佳预测的预测误差有简单表达式. 此时, 根据定理 4.7 可知, $\mathrm{AR}(p)$ 序列的最佳预测与最佳线性预测相等.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

根据 Green 公式, AR(p) 序列 X_t 可表示为

$$X_t = \sum_{j=0}^{\infty} g_j \varepsilon_{t-j} \tag{4.62}$$

其中 g_j 为 Green 函数的系数. 则

$$X_{t+k} = (\varepsilon_{t+k} + g_1\varepsilon_{t+k-1} + \cdots + g_{k-1}\varepsilon_{t+1}) + (g_k\varepsilon_t + g_{k+1}\varepsilon_{t-1} + \cdots).$$

上式中, $\varepsilon_t, \varepsilon_{t-1}, \cdots$ 为当前及过去的噪声序列.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

记

$$\begin{aligned} e_t(k) &= \varepsilon_{t+k} + g_1\varepsilon_{t+k-1} + \cdots + g_{k-1}\varepsilon_{t+1} \\ \hat{X}_t(k) &= g_k\varepsilon_t + g_{k+1}\varepsilon_{t-1} + \cdots \end{aligned} \tag{4.63}$$

则 $X_{t+k} = e_t(k) + \hat{X}_t(k)$. 由 AR 模型的传递形式 (4.10) 可知, $\hat{X}_t(k)$ 可表示为 $\{X_t, X_{t-1}, \cdots\}$ 的线性组合, 即

$$\hat{X}_t(k) = \sum_{j=0}^{\infty} d_j X_{t-j}$$

若 ε_t 是均值为 0，方差为 σ^2 的高斯白噪声, 则 $\{\varepsilon_{t+k}, \cdots, \varepsilon_{t+1}\}$ 与 $\{\varepsilon_t, \varepsilon_{t-1}, \cdots\}$ 相互独立.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

故由无限多的历史数据 $\boldsymbol{X}_t=\{X_t, X_{t-1}, \cdots\}$ 可得
$E\left(\hat{X}_t(k) \mid \boldsymbol{X}_t\right)=\hat{X}_t(k), \operatorname{Var}\left(\hat{X}_t(k) \mid \boldsymbol{X}_t\right)=0.$

根据定理 4.4, X_{t+k} 的最佳预测为

$$
\begin{aligned}
E\left(X_{t+k} \mid \boldsymbol{X}_t\right) & =E\left(\left(e_t(k)+\hat{X}_t(k)\right) \mid \boldsymbol{X}_t\right) \\
& =E\left(e_t(k) \mid \boldsymbol{X}_t\right)+E\left(\hat{X}_t(k) \mid \boldsymbol{X}_t\right) \\
& =E\left(e_t(k)\right)+\hat{X}_t(k)=\hat{X}_t(k).
\end{aligned} \tag{4.64}
$$

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

即最佳预测为 $\hat{X}_t(k)$, 且预测的方差为

$$
\begin{aligned}
\operatorname{Var}\left(X_{t+k} \mid \boldsymbol{X}_t\right) &= \operatorname{Var}\left(e_t(k) \mid \boldsymbol{X}_t\right) + \operatorname{Var}\left(\hat{X}_t(k) \mid \boldsymbol{X}_t\right) \\
&= \operatorname{Var}\left(e_t(k)\right) \\
&= \left(1 + g_1^2 + \cdots + g_{k-1}^2\right) \sigma^2
\end{aligned} \tag{4.65}
$$

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

由于噪声服从正态分布, 故 $\hat{X}_t(k)$ 也服从正态分布. 由 (4.64) - (4.65) 式可知，当噪声为高斯白噪声且知无限多的历史样本数据 $\boldsymbol{x}_t=\{x_t, x_{t-1}, \cdots\}$ 时, 对于 X_{t+k} 的最佳预测将服从正态分布 $N\left(\hat{x}_t(k), \operatorname{Var}\left(e_t(k)\right)\right)$, 从而得到预测的 $(1-\alpha)$ 的置信区间为

$$\left(\hat{x}_t(k) \pm z_{\frac{\alpha}{2}}\left(1+g_1^2+\cdots+g_{k-1}^2\right)^{\frac{1}{2}} \sigma\right), \tag{4.66}$$

式中 $z_{\frac{\alpha}{2}}$ 表示标准正态分布的上 $\frac{\alpha}{2}$ 分位数.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

例 (4.14)

设某城市每年的降雨量 (单位: mm) 近似服从 AR(2) 模型

$$X_t = 200 + 0.5X_{t-1} + 0.3X_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 100)$$

已知 2009 – 2011 这三年的降雨量分别为: 1100 1000 1120. 请预测该市今后四年的降雨量及其 95% 的置信区间.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

例 (4.14)

设某城市每年的降雨量 (单位: mm) 近似服从 AR(2) 模型

$$X_t = 200 + 0.5X_{t-1} + 0.3X_{t-2} + \varepsilon_t, \quad \varepsilon_t \sim N(0, 100)$$

已知 2009 – 2011 这三年的降雨量分别为: 1100 1000 1120. 请预测该市今后四年的降雨量及其 95% 的置信区间.

由于噪声为高斯白噪声, 而且模型给定, 因此该模型可看为根据无限多的历史数据而得到的, 故最佳预测与最佳线性预测相等. 因此我们只需考虑最佳线性预测.

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

(1) 预测值的计算. 根据 (4.60) 式, 今后四年降雨量的预测值如下所示:

2012 年: $\hat{x}_{2011}(1) = 200 + 0.5 \times x_{2011} + 0.3 \times x_{2010} = 1060,$

2013 年: $\hat{x}_{2011}(2) = 200 + 0.5 \times \hat{x}_{2011}(1) + 0.3 \times x_{2011} = 1066,$

2014 年: $\hat{x}_{2011}(3) = 200 + 0.5 \times \hat{x}_{2011}(2) + 0.3 \times \hat{x}_{2011}(1) = 1051,$

2015 年: $\hat{x}_{2011}(4) = 200 + 0.5 \times \hat{x}_{2011}(3) + 0.3 \times \hat{x}_{2011}(2) = 1062.1.$

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

(2) 预测方差的计算. 根据 Green 函数的递推公式, 可得该 AR(2) 模型的 Green 系数为

$$
\begin{aligned}
g_0 &= 1 \\
g_1 &= a_1 g_0 = 0.5 \\
g_2 &= a_1 g_1 + a_2 g_0 = 0.5 \times 0.5 + 0.3 \times 1 = 0.55 \\
g_3 &= a_1 g_2 + a_2 g_1 = 0.5 \times 0.55 + 0.3 \times 0.5 = 0.425
\end{aligned}
$$

从而可得预测方差为

$$
\begin{aligned}
\operatorname{Var}\left(\hat{x}_{2011}(1)\right) &= g_0^2 \sigma^2 = 100 \\
\operatorname{Var}\left(\hat{x}_{2011}(2)\right) &= \left(g_0^2 + g_1^2\right) \sigma^2 = 125 \\
\operatorname{Var}\left(\hat{x}_{2011}(3)\right) &= \left(g_0^2 + g_1^2 + g_2^2\right) \sigma^2 = 155.25 \\
\operatorname{Var}\left(\hat{x}_{2011}(4)\right) &= \left(g_0^2 + g_1^2 + g_2^2 + g_3^2\right) \sigma^2 = 173.3125
\end{aligned}
$$

基于无限数据的 $\mathbf{AR}(p)$ 序列预测

(3) 置信区间的计算. 由于标准正态分布的 0.975 分位点 $z_{0.975} = 1.96$, 则根据 (4.66) 式, 我们可得 $2012 - 2015$ 年预测的置信区间如下所示.

2012 年: $(1040.40, 1079.60)$,

2013 年: $(1044.09, 1087.91)$,

2014 年: $(1026.58, 1075.42)$,

2015 年: $(1036.30, 1087.90)$.

从中可知, 预测的置信区间随着预测步长的增加而增大.

1. 4.6.1 最佳预测
2. 4.6.2 最佳线性预测
3. 4.6.3 AR 模型的预测
4. 4.6.4 修正预测
5. 参考资料

修正预测

预测的步长越长, 未知信息就越多, 从而估计的精度就越差. 然而, 随着时间的发展，我们在原有的观测值 $\{\cdots, x_{t-1}, x_t\}$ 的基础上，不断获得新的观测值 $\{x_{t+1}, x_{t+2}, \cdots\}$. 这些新观测值带来更多的信息, 从而预测未来时刻的未知信息逐渐减少.

因此, 利用新观测值的信息, 我们可以更好地预测未来的序列值 x_{t+k}, 预测精度将提高. 这就是所谓的修正预测.

修正预测

修正预测有两种处理方式.

一种处理方法是把新的观测值和原数据合并, 重新拟合模型, 然后再利用拟合后的模型预测 x_{t+k};

另一种处理方法是利用原来的拟合模型, 然后利用新观测值修正原来的拟合模型, 从而得到新的拟合模型. 当新的观测序列很多时或易于操作时, 可采用第一种方法. 然而, 当新的观测并不多时, 第一种方法不是最佳选择.

此时, 第二种方法将更加简便. 下面介绍第二种处理方法.

修正预测

假设已知观测值 $\{\cdots, x_{t-1}, x_t\}$, 则其 k 步最佳预测如 (4.63) 式所示. 若获得时刻 $t+1$ 的观测值 x_{t+1}, 即现有数据为 $\{\cdots, x_{t-1}, x_t, x_{t+1}\}$, 则根据 (4.63) 式, x_{t+k} 的预测值为

$$\begin{aligned}\hat{x}_{t+1}(k-1) &= g_{k-1}\varepsilon_{t+1} + g_k\varepsilon_t + g_{k+1}\varepsilon_{t-1} + \cdots \\ &= g_{k-1}\varepsilon_{t+1} + \hat{x}_t(k)\end{aligned} \tag{4.67}$$

其中 $\varepsilon_{t+1} = x_{t+1} - \hat{x}_t(1)$ 表示 x_{t+1} 的一步预测误差, $\hat{x}_{t+1}(k-1)$ 表示根据数据 $\{\cdots, x_{t-1}, x_t, x_{t+1}\}$ 对 $k-1$ 步后的 x_{t+k} 的预测值, 而 $\hat{x}_t(k)$ 表示根据数据 $\{\cdots, x_{t-1}, x_t\}$ 对 k 步后的 x_{t+k} 的预测值. 实际上, 修正后的预测值 $\hat{x}_{t+1}(k-1)$ 和原来的预测值 $\hat{x}_t(k)$ 都是对序列值 x_{t+k} 的预测.

修正预测

此时, 修正后的预测误差为

$$e_{t+1}(k-1)=x_{t+k}-\hat{x}_t(k)=g_0\varepsilon_{t+k}+\cdots+g_{k-2}\varepsilon_{t+2}.$$

因此, 其预测误差的方差为

$$\operatorname{Var}\left(e_{t+1}(k-1)\right)=\left(1+g_1^2+\cdots+g_{k-2}^2\right)\sigma^2,$$

其中 σ^2 为白噪声的方差. 与原预测的误差方差 (4.65) 相比, 修正后的预测方差减少了 $g_{k-1}^2\sigma^2$, 因此, 修正后的预测精度有所提高.

修正预测

一般地, 假如获得新观测值 $x_{t+1}, \cdots, x_{t+l}(1 \leqslant l < k)$, 则 x_{t+k} 的修正预测值为

$$
\begin{aligned}
\hat{x}_{t+1}(k-1) &= g_{k-l}\varepsilon_{t+l} + \cdots + g_{k-1}\varepsilon_{t+1} + g_k\varepsilon_t + g_{k+1}\varepsilon_{t-1} + \cdots \\
&= g_{k-l}\varepsilon_{t+l} + \cdots + g_{k-1}\varepsilon_{t+1} + \hat{x}_t(k)
\end{aligned}
$$

其中 $\varepsilon_{t+j} = x_{t+j} - \hat{x}_{(t+j-1)+1}(1 \leqslant j \leqslant l)$ 为 x_{t+j} 的一步预测误差. 此时, 修正后的预测方差为

$$
\operatorname{Var}\left(e_{t+l}(k-l)\right) = \left(1 + g_1^2 + \cdots + g_{k-l-1}^2\right)\sigma^2.
$$

从上面的分析可知, 当我们获得新的观测值时, 修正后的预测方差将减少, 从而提高了预测精度. 而且这种修正方式简单, 易于操作.

修正预测

例 (4.15)

(例 4.14 续) 假如已知 2012 年该市的降雨量为 1100 mm, 预测后三年该市的降雨量及其 95% 的置信区间.

修正预测

例 (4.15)

(例 4.14 续) 假如已知 2012 年该市的降雨量为 1100 mm, 预测后三年该市的降雨量及其 95% 的置信区间.

(1) 计算 2012 年的一步预测误差:

$$\varepsilon_{2012} = x_{2012} - \hat{x}_{2011}(1) = 1100 - 1060 = 40.$$

修正预测

(2) 计算修正预测值:

$$\hat{x}_{2012}(1) = g_1\varepsilon_{2012} + \hat{x}_{2011}(2) = 1086$$
$$\hat{x}_{2012}(2) = g_2\varepsilon_{2012} + \hat{x}_{2011}(3) = 1073$$
$$\hat{x}_{2012}(3) = g_3\varepsilon_{2012} + \hat{x}_{2011}(4) = 1079.1.$$

修正预测

(2) 计算修正预测值:

$$
\begin{aligned}
\hat{x}_{2012}(1) &= g_1\varepsilon_{2012} + \hat{x}_{2011}(2) = 1086 \\
\hat{x}_{2012}(2) &= g_2\varepsilon_{2012} + \hat{x}_{2011}(3) = 1073 \\
\hat{x}_{2012}(3) &= g_3\varepsilon_{2012} + \hat{x}_{2011}(4) = 1079.1.
\end{aligned}
$$

(3) 计算修正预测方差:

$$
\begin{aligned}
\operatorname{Var}\left(e_{2012}(1)\right) &= g_0^2\sigma^2 = 100 \\
\operatorname{Var}\left(e_{2012}(2)\right) &= \left(g_0^2 + g_1^2\right)\sigma^2 = 125 \\
\operatorname{Var}\left(e_{2012}(3)\right) &= \left(g_0^2 + g_1^2 + g_2^2\right)\sigma^2 = 155.25
\end{aligned}
$$

修正预测

(4) 该市降雨量 $l(1 \leqslant l \leqslant k-1)$ 步预测值的 95% 的置信区间为

$$\left(\hat{x}_{2012}(l) - 1.96\sqrt{\mathrm{Var}\left(e_{2012}(l)\right)}, \hat{x}_{2012}(l) + 1.96\sqrt{\mathrm{Var}\left(e_{2012}(l)\right)}\right).$$

相应结果见表 4.7, 从中可知, 修正预测的置信区间比原预测的置信区间有所缩小, 即预测精度有所提高.

表 4.7 修正预测与原预测的比较

	原预测		修正预测	
年份	置信区间	长度	置信区间	长度
2012	$(1040.40, 1079.60)$	39.2		
2013	$(1044.09, 1087.91)$	43.82	$(1066.40, 1105.60)$	39.2
2014	$(1026.58, 1075.42)$	48.84	$(1051.09, 1094.91)$	43.82
2015	$(1036.30, 1087.90)$	51.60	$(1054.68, 1103.52)$	48.84

参考资料

[1] 周永道, 王会琦, 吕王勇. 时间序列分析及应用. 北京: 高等教育出版社,2015.

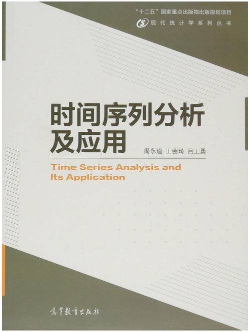

图 3: 时间序列分析及应用

感谢

Thanks

本人非统计专业，若有不妥之处，恳请批评指正.

作者: 图灵的猫 邮箱: turingscat@126.com 主页: turingscat.club