ARMAX 模型的预测确认 *

陈钧

(东南大学数学系，南京， 210096)

摘要： 本文将围绕 Pena（ 2005 ）提出的关于时间序列中预测确认的新方法进行改进，这种改进的预测确认方法是建立在一种所谓的“ hv- 过滤残差”基础上。这是一种样本内的预测误差，它的估计跟 Pena 提出的“过滤残差”的估计类似。相比于经典交叉确认方法中用删除观察值来进行预测确认，这种改进的方法只需要删除观察值的估计新息。而且可以证明这种“ hv- 过滤残差”直到很小阶都是不相关的。

关键词： 预测误差；交叉确认；预测确认；过滤残差； hv- 过滤残差.

Predictive Validation in ARMAX Time Series Models

Jun Chen

(Department of Mathematics, Southeast University, Nanjing 210096, China)

Abstract: This Paper considers the impact of Pena's(2005) new procedure for multifold predictive validations in time series.The modified procedure is based on the so called "hv-filtered residuals",in-sample prediction errors evaluated in such a way that they are similar to "filtered residuals"presented by Pena. Thus ,instead of using the deletion of observations to validate the predictions,as in classical cross-validation, the modified precedure is based on deletion of the estimated innovations.It is proved that the "hv-filtered residuals" are uncorrelated,up to small order.

Keywords: prediction error; Cross-Validation; predictive validation; filtered residuals; hv-filtered residuals.

一、引言

评估一个模型预测能力的最好方法之一是考察其样本外预测误差的均值。这种样本外误差已经被广泛运用，其中包括判别准则的选取，均方预测误差的估计，时间序列模型的选择等等。对于相互独立的数据，用样本外预测方法进行预测确认的经典方法是交叉确认（ Stone1974,Allen1974 ）。交叉确认通常是把数据分割成任意两部分，一部分用于拟合模型，另一部分用于模型确认。随着不同分割的选取，预测确认不断重复进行，然后我们运用一些估计准则，例如最小均方预测误差（ MSPE ），来估计给定模型的预测误差（ Burman1989,Zhang1993 ）。

*: 作者：陈钧， (1982-), 男，东南大学数学系． E-mail:buyixingzhe@163.com.

交叉确认中一个重要方面就是观察值必须是可以“交叉”的，即每个观察值必须既可以用来拟合模型，又可用于预测确认。然而对于时间序列来说，我们只能用过去已知的观察值来预测未来观察值，从而导致数据的使用受到一定的限制。基于这个原因，时间序列中的预测确认，往往采用分裂样本方法来进行，即将时间序列分成两部分：前一部分用于拟合模型，后一部分用来预测确认。分裂样本预测确认方法中样本的分裂有多种方法，现今最常用的一种叫“滚动预测”。在滚动预测中。每做一次样本分裂，用于模型拟合的样本观察值将得到一定数目的增加；另一方面，用于预测确认的那部分观察值将相应减少。然而，由分裂样本方法得到的样本外误差存在着较大缺陷：第一，模型参数和预测误差方差的估计仅仅用到了部分样本数据，这将会导致预测误差的方差估计偏大，习惯上我们称这一偏大部分的方差为“数据分裂方差”；第二，分裂样本预测确认的结果跟样本初始值的选取有关，而样本初始值的选取是随意的；第三，滚动预测中不同 h 步预测误差的获得所用的样本大小不同，因此各个 h 步预测误差之间很难进行比较。同样的，样本内预测误差也存在着一定缺陷，其中一个重大的缺陷就是在进行模型拟合和计算预测误差时重复使用了相同数据，从而导致预测量估计偏大而预测误差的方差偏小。这个问题已引起很多学者的注意（如 Efron(1986)），因此也有必要对样本内方法进行改进。

二、样本内误差，样本外误差

记一般时间序列的 ARMAX(p,d,q) 过程如下：

$$\Psi(B)z_t = \Phi(B)x_t + \theta(B)e_t \qquad e_t \sim N(0,\sigma^2) \qquad t = 1,2,\cdots T \tag{1}$$

$$\Psi(B) = 1 - \Psi_1 B - \cdots - \Psi_p B^p$$

$$\Phi(B) = \Phi_1 B + \cdots + \Phi_d B^d$$

$$\theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$$

这里 T 是序列的观测数据，$\Psi(B),\Phi(B),\theta(B)$ 是滞后算子 B 的多项式，且 $\Psi(B),\theta(B)$ 的所有根都在单位圆外。记 $\lambda = (\Psi_1,\cdots,\Psi_p,\Phi_1,\cdots,\Phi_d,\vartheta_1,\cdots,\theta_q)^T$ 。若已知观察时间序列 $Z_n = (z_1,z_2,\cdots,z_n)^T$ 满足模型（ 1 ），记由历史数据 $Z_t = (z_1,z_2,\cdots,z_t)^T$ 对未来 z_{t+h} 进行的 h 步预测为 $\hat{z}_{t+h} = \hat{z}_{t+h}(Z_t,\lambda) = E(z_{t+h}|Z_t,\lambda)$ ，这里假设向量参数 λ 是已知的。其中 $r \le t \le n-h$ ， $z_1,z_2,\cdots,z_r$ 为初始观察值。从而我们得到 h 步预测误差 $e_{t+h} = z_{t+h} - \hat{z}_{t+h}$ 。当 λ 未知时，预测误差可以有多种方法定义。若 $\hat{\lambda}_n = F(Z_n)$ 表示由全部观察值得到的向量参数估计，则称 h 步预测 $\hat{z}^{in}_{t+h} = z_{t+h}(Z_t,\hat{\lambda}_n) = E(z_{t+h}|Z_t,\hat{\lambda}_n)$ 为样本内 h 步预测。从而样本内 h 步预测误差为 $\hat{e}^{in}_{t+h} = z_{t+h} - \hat{z}^{in}_{t+h}, r \le t \le n-h$ 。特别地，当 h=1 时，有 $\hat{e}^{in}_{t+1} = z_{t+1} - \hat{z}^{in}_{t+1} = \hat{a}_{t+1}, r \le t \le n-1$ ，这里 $\hat{a}_{t+1}$ 为模型的残差或新息。再对所有的样本内

预测误差相加取平均值，得到样本内均方预测误差（ MSPE ）的估计，即

$$\hat{V}^{in}(h)=\frac{\sum_{t=r}^{n-h}(\hat{e}_{t+h}^{in})^2}{n-h-r+1} \qquad (2)$$

当 $\hat{\lambda}_m = F(Z_m), m \le t$, 即 $\hat{\lambda}_m$ 是由 t 之前的观察值所得的向量参数估计。就称 h 步预测 $\hat{z}_{t+h}^{out} = z_{t+h}(Z_t, \hat{\lambda}_m) = E(z_{t+h}|Z_t, \hat{\lambda}_m)$ 为样本外预测。从而样本外 h 步预测误差为 $\hat{e}_{t+h}^{out} = z_{t+h} - \hat{z}_{t+h}^{out}, r \le t \le n-h$ ，本文我们假设 $m=t$ 。同样的，将所有的样本外预测误差相加取平均值就得到样本外均方预测预测误差（ MSPE ）的估计，也即

$$\hat{V}^{out}(h)=\frac{\sum_{t=n_h}^{n-h}(\hat{e}_{t+h}^{out})^2}{n-h-n_h+1} \qquad (3)$$

从上可知， $\hat{e}_{t+h}^{out}$ 与要预测的观察值所包含的新息无关，而 $\hat{e}_{t+h}^{in}$ 包含有预测观察值中的新息。

三、过滤残差， hv- 过滤残差

这部分我们将着重介绍过滤残差，并由此引出改进的 hv- 过滤残差。设时间序列 $Z_n = (z_1, z_2, \cdots, z_n)^T$ 满足模型（ 1 ），为叙述方便，我们以 AR（ 1 ）模型为例：

$$z_t = \phi z_{t-1} + a_t, \qquad |\phi| < 1,$$

其中 a_t 为白噪声。

我们的目的是要用从已知样本 Z_n 中得到的预测误差来评估其预测未来 $z_t, t > n$ 的能力。记 $z_{t+h}, 1 \le (t+h) \le n$ 为样本内的预测点，我们将用其来估计预测误差 e_{t+h} 。如果用样本外的方法来估计 e_{t+h} ，我们首先得到参数 ϕ 的估计为 $\hat{\phi}_T = \Sigma_2^T z_t z_{t-1} / \sum_2^T z_{t-1}^2$ ，然后计算出预测量 $\hat{z}_{T+h}^{out} = z_{T+h}(Z_T, \hat{\phi}_T) = \hat{\phi}_T^h z_T$, 最后得到估计误差 $\hat{e}_{T+h}^{out} = z_{T+h} - \hat{\phi}_T^h z_T = e_{T+h} + (\phi^h - \hat{\phi}_T^h) z_T$ ，其中 $e_{T+h} = a_{T+h} + \phi a_{T+h-1} + \cdots + \phi^{h-1} a_{T+1}$ 。从中我们发现新息 $a_{T+1}, \cdots a_{T+h}$ 没有包含在估计量 $\hat{\phi}_T^h$ 中。特别的，若 $h=1$ ，我们可以证明 $E(\hat{e}_{T+1}^{out})^2 \approx \sigma^2(1+(T-1)^{-1})$ 。又因为 $E(\hat{e}_{n+1})^2 \approx \sigma^2(1+(n-1)^{-1})$ 。由此我们得出结论：用样本外方法对 MSPE 进行估计将导致估计偏大。为了更好的估计 $E(\hat{e}_{n+1})^2$, 我们用 $\hat{V}^{out}(1)$ 作为其估计量。于是有

$$E[\hat{V}^{out}(1)] \approx \sigma^2(1+\frac{\sum_{t=n_1}^{n-1} 1/(t-1)}{n-n_1}) > \sigma^2(1+\frac{1}{n-2})$$

从上式可知， $\hat{V}^{out}(1)$ 的渐进偏差为正。对于 $n-n_1$ 足够大时，我们可以认为

$$E[\hat{V}^{out}(1)] \approx \sigma^2(1+\frac{log(n)-log(n_1)}{n-n_1})$$

若我们记 $n_1 = \alpha n, 0 < \alpha < 1$, 则关于 $\hat{V}^{out}(1)$ 的大样本偏差可记为

$$Bias[\hat{V}^{out}(1)] \approx \frac{\sigma^2}{n}(\frac{-log\alpha}{1-\alpha} - 1)$$

从上述偏差中可看出：当减小 α 时， $\hat{V}^{out}(1)$ 的偏差以指数级递增；另一方面，当增大 α 时，用于预测的样本大小 $n(1-\alpha)$ 将相应减少，从而增加 $\hat{V}^{out}(1)$ 的波动。这是一个两难选择！

作为比较,如果用样本内的方法来估计 e_{t+h} ,首先计算 $\hat{\phi}_n$, 然后求得 $\hat{z}^{in}_{T+h} = z_{T+h}(Z_T, \hat{\phi}_n) = \hat{\phi}^h_n z_T$, 最后得到估计误差 $\hat{e}^{in}_{T+h} = z_{T+h} - \hat{\phi}^h_n z_T = e_{T+h} + (\phi^h - \hat{\phi}^h_n)z_T$. 我们发现 e_{T+h} 和 $(\phi^h - \hat{\phi}^h_n)z_T$ 是相关的. 这是因为 $\hat{\phi}_n$ 中已经包含了 $a_{T+1}, \cdots, a_{T+h}$. 同样，我们可以证明 $E(\hat{e}^{in}_{T+1})^2 \approx \sigma^2(1-(n-1)^{-1})$. 从而得到 $Bias[\hat{V}^{out}(1)] \approx -2\sigma^2(1-(n-1)^{-1})$. 从偏差为负可知：样本内方法将导致 MSPE 的估计偏低.

在比较了样本内和样本外方法各自的优缺点，并结合 Stone(1974) 提出的适合独立数据的交叉确认的思想之后， Pena 提出了一种基于过滤残差的预测确认方法. 其对预测误差 e_{T+h} 的估计采用如下方法： (1) 使 e_{T+h} 的估计中不含有新息 $a_{T+1}, \cdots, a_{T+h}$, 以使预测量与 $a_{T+1}, \cdots, a_{T+h}$ 不相关； (2) 但另一方面又必须包含 $a_{T+h+1}, \cdots, a_n$ 以提高估计的精度. 因此，相比经典交叉确认中删除观察值的方法，这种基于过滤残差的预测确认方法只需要删除新息. 这两种方法对于独立数据而言是一致的，因为对独立数据来说，删除新息等价于删除观察值. 而对于相关数据而言，这两种方法是不等价的，因为，在相关数据中，新的观察值中所包含的新息无法从历史数据中预测. 因此，为了避免在估计预测误差时出现预测量所含新息， Pena 构造了一个新的过滤时间序列 y_t 如下：

当 $t = 1, 2, \cdots, T,$ $\quad y_t = z_t;$

当 $t = T+1, \cdots, T+h,$ $\quad y_{T+j} = \hat{\phi}^j_n z_T$ $\quad$ 此时 $a_{T+1} = \cdots = a_{T+h} = 0;$

当 $t = T+h+1, \cdots, n,$ $\quad y_t = \hat{\phi}_n y_{t-1} + \hat{a}_t.$ $\quad$ 这里 $\hat{a}_{t+1} = \hat{e}^{in}_{t+1} = z_{t+1} - \hat{\phi}_n z_t.$

接下来对时间序列 y_t 用样本内预测确认方法得到 $\hat{\phi}^{fil}$, 然后得到预测量 $\hat{z}^{fil}_{T+h} = (\hat{\phi}^{fil})^h z_T$, 最后得到预测误差 $\hat{e}^{fil}_{T+h} = z_{T+h} - (\hat{\phi}^{fil})^h z_T$. 将所有的过滤残差预测误差相加取平均值就得到过滤残差均方预测预测误差（ MSPE ）的估计，也即

$$\hat{V}^{fil}(h) = \frac{\sum_{t=r}^{n-h}(\hat{e}^{fil}_{t+h})^2}{n-h-r+1} \qquad (4)$$

用样本内方法进行预测确认时，由于在估计参数中新息 $a_{T+l}, l = 1, \cdots, h$ 的存在，使得 $E(\hat{z}^{in}_{T+h} a_{T+l}) \neq 0$, 而用样本外方法虽然有 $E(\hat{z}^{out}_{T+h} a_{T+l}) = 0$, 但往往预测精度不高. 与前两种方法相比， Pena 提出的基于过滤残差的预测方法很好的结合了样本内和样本外方法的优点，既保证了预测的精度，又由于显著减小了新息 $a_{T+l}, l = 1, \cdots, h$ 对预测量 $\hat{z}^{fil}_{T+h}$ 的影响. 从下面的定理中我们将得知 $E(\hat{z}^{fil}_{T+h} a_{T+l}), l = 1, \cdots, h$ 比 $E(\hat{z}^{in}_{T+h} a_{T+l})$ 具有更低阶.

定理 1: 设时间序列 z_t 满足 ARMAX 模型 (1), 记 $\hat{z}^{in}_{T+h}$ 为 z_{T+h} 的 h 步样本内预测，其

中 $r<(T+h)\leq n$; 记 $\hat{z}_{T+h}^{fil}$ 为 z_{T+h} 的 h 步过滤预测，则对 $l=1,\cdots h$, 有

$$E(\hat{z}_{T+h}^{in}a_{T+l})=O(n^{-1})$$

$$E(\hat{z}_{T+h}^{fil}a_{T+l})=O(n^{-2})$$

从定理 1 中我们得知 Pena 提出的这种基于过滤残差的预测确认方法确实优于样本内和样本外的预测确认方法，但是我们也注意到在估计参数 $\hat{\phi}^{fil}$ 时并不能完全排除新息 $a_{T+l},l=1,\cdots,h$ 的干扰。因为 $\hat{a}_{T+l}$ 只是预测量中出现的新的新息的估计。这也是导致 $E(\hat{z}_{T+h}^{fil}a_{T+l})$ 不能达到 0 的原因所在。幸运的是，针对这种情况，我们还可以对其进行进一步的改进。具体作法就是在构造新的过滤时间序列 y_t 时在删除新息 $a_{T+l},l=1,\cdots,h$ 之后，我们紧接着再删除 $v,\quad v\geq 1$ 个新息 $a_{T+h+l},l=1,\cdots,v$ 。我们仍以上述 AR（ 1 ）模型为例，构造新的时间序列 x_t 如下：

当 $t=1,2,\cdots,T,$ $x_t=z_t;$

当 $t=T+1,\cdots,T+h+v,$ $x_{T+j}=\hat{\phi}_n^j z_T$ 此时 $a_{T+1}=\cdots=a_{T+h+v}=0;$

当 $t=T+h+v+1,\cdots,n,$ $x_t=\hat{\phi}_n y_{t-1}+\hat{a}_t.$ 这里 $\hat{a}_{t+1}=\hat{e}_{t+1}^{in}=z_{t+1}-\hat{\phi}_n z_t.$

然后还按样本内方法一直往下做．我们称这种改进的方法为基于“ hv- 过滤残差”的预测确认方法．我们首先得到 $\hat{\phi}^{hv-fil}$ ，再求得预测量 $\hat{z}_{T+h}^{hv-fil}=(\hat{\phi}^{hv-fil})^h z_T$, 最后得到预测误差 $\hat{e}_{T+h}^{hv-fil}=z_{T+h}-(\hat{\phi}^{hv-fil})^h z_T$. 再将所有的过滤残差预测误差相加取平均值就得到 hv- 过滤残差均方预测预测误差（ MSPE ）的估计，也即

$$\hat{V}^{hv-fil}(h)=\frac{\sum_{t=r}^{n-h}(\hat{e}_{t+h}^{hv-fil})^2}{n-h-r-v+1} \qquad (5)$$

尽管这种改进的基于“ hv- 过滤残差”的预测确认方法只比 Pena 提出的方法多删除了一些新息，但它却因此进一步减小了新息 $a_{T+l},l=1,\cdots,h$ 对于 $\hat{\phi}^{hv-fil}$ 的干扰，从而使得这种改进的方法较 Pena 提出的新方法在时间序列的预测确认中更有效，这一点我们将从下面的模拟实验中得到证实.

四、模拟举例

这部分我们将对有限样本数据采用蒙特卡罗模拟实验方法，分别讨论样本内方法，样本外方法，基于过滤残差方法以及基于 hv- 过滤残差方法用于预测确认的结果，并作比较以说明我们改进的方法较其它三种方法更有效。我们以 ARMAX（ 3， 0， 0 ）模型为例。

记 ARMAX（ 3， 0， 0 ）模型如下：

$$y_t=1.4y_{t-1}-0.59y_{t-2}-0.07y_{t-3}+e_t \qquad e_t\sim N(0,1)$$

我们将对上述 ARMAX（ 3， 0， 0 ）模型进行 5000 次蒙特卡罗模拟实验。在每个蒙特卡罗模拟中，我们将随机生成 200+n+5 个数据点。为了得到平稳数据，我们将前 200 个数据删

除。然后对余下的 n+5 个数据，取前 n 个数据用于模型拟合，将最后 5 个数据看成是未来要预测的观察数据。我们将分别取 n=25 和 n=100 两个样本进行模拟，并对每个样本分别考察 h=1 和 h=3 两种情形。在具体模拟实验中，我们先利用最后 5 个数据得到总体样本外 MSPE 的估计，记作 V_h^{pop} 。然后选取一定的 v, 分别计算 $\hat{V}^{out}(h)$, $\hat{V}^{in}(h)$, $\hat{V}^{fil}(h)$, $\hat{V}^{hv-fil}(h)$, 并逐个与 V_h^{pop} 作差得到偏差 $Bias[\hat{V}(h)-V_h^{pop}]$ 和均方误差 $MSE[\hat{V}(h)-V_h^{pop}]^2$ 。模拟结果如下表所示：

表 1: $Bias[\hat{V}(h)-V_h^{pop}]$ **和** $MSE[\hat{V}(h)-V_h^{pop}]^2$

		Bias				MSE			
h	$\hat{V}^{pop}$	$\hat{V}^{in}$	$\hat{V}^{out}$	$\hat{V}^{fil}$	$\hat{V}^{hv-fil}$	$\hat{V}^{in}$	$\hat{V}^{out}$	$\hat{V}^{fil}$	$\hat{V}^{hv-fil}$
n=25									
1	2.0127	-1.0962	2.5846	-0.8912	-0.2714	1.2016	6.6800	0.7943	0.0737
3	5.7401	-2.0436	5.0378	-1.7359	0.5949	4.1763	25.3798	3.0133	0.3539
n=100									
1	1.3070	-0.3367	1.8958	-0.2792	-0.2106	0.1134	3.5939	0.0780	0.0443
3	5.0364	-0.7036	6.9380	-0.5170	-0.1985	0.4951	48.1350	0.2673	0.0393

五、结论和进一步的问题

本文改进的基于“ hv- 过滤残差”的预测确认方法计算简单，并且容易发现其模型拟合过程与观察时间序列中出现异常点时模型参数估计类似，故可将这一方法推广运用到时间序列异常点的探测问题中。同时，这种改进的预测确认方法中合适的 v 的选取仍然有待于进一步的讨论 。

参考文献

[1] Burman,P.(1989) *A Comparative Study of Ordinary Cross-Validation,v-Fold Cross-Validation,and the Repeated Learning-Testing Methods[J].* Biometrika.76:503-514.

[2] Burman,P,Chow E,and Nolan,D.(1994)*A Cross-Validatory Method for Dependent Data[J].* Biometrika.81,351-358.

[3] Pena,D,and Sanchez,I.(2005)*Multifold Predictive Validation in ARMAX Time Series Models[J].* Journals of the American Statistical Association ,March 2005,Vol,100,No.489,Theory and Method.

[4] Stone,M.(1974)*Cross-Validation Choice and Assessment for Statistical Predictions[J].* Journals of the Royal Statistical Society,Ser,B,36,111-147.

[5] West,K,D.(1996).*Asymptotic Inference About Predictive Ability[J].* Econometrica,64,1067-1084.

[6] White,H.(2000)*A Reality Check for Data Snooping[J].* Econometrica,68,1097-1126

[7] Zhang,P.(1993).*Model Selection via Multifold Cross-Validation[J].* The Annals of Statistics,21,299-313.

[8] 陈平，达庆利 (2001). 运用 *SAS* 软件系统对我国农作物受灾及成灾面积的预测分析 *[J]* 系统工程理论和实践． 21(4): 141-144.

[9] 韦博成，鲁国斌，史建清 (1991). 统计诊断引论 *[M].* 南京，东南大学出版社.