V – SELEÇÃO DE MODELOS

e

ESTIMADORES TENDENCIOSOS

V.1 INTRODUÇÃO

O problema da seleção da estrutura de um modelo não é um problema em aberto. Na verdade existem duas opções aceitas universalmente:

- a seleção de modelos suficientemente simples e complexos;

- a utilização de modelos complexos mas com conjuntos de parâmetros obtidos de uma maneira particular que permite regular a inflação e a falta de ajuste.

Da primeira destas opções temos a famosa seleção de regressores, muito utilizada na década de 50, ainda hoje muito utilizada em disciplinas como a identificação de sistemas. A seleção de regressores trata da seleção de alguns termos de uma função para serem utilizados na representação do processo.

A outra opção, a utilização de estimadores tendeciosos, tratam da seleção de uma parâmetro de tendência, seja or exemplo no caso de um PLS, do número de variáveis latentes a serem utilizadas na construção do modelo, seja do valor da constante de tendência no caso de um estimador de “ridge”.

Em todos os casos, o grande problema é que o ajuste do modelo não é um bom critério de seleção porque o critério de ajuste sempre melhora à medida

que o modelo se torna mais complexo. Apesar que em geral é mais difícil calcular as estimativas dos parâmetros.

Portanto sempre terá que ser usado uma critério de seleção.

V.2 CRITÉRIOS DE SELEÇÃO DE MODELOS: A VALIDAÇÃO CRUZADA

O principal critério a se estudar para decidir se um modelo ou parâmetro de tendência são adequados é a capacidade de predição do modelo. E a maneira mais natural de fazer isto é o que se chama de avalidação cruzada. Esta metodologia é a mais usada atualmente.

A idéia básica da validação cruzada é tomar o conjunto de dados experimentais e dele retirar uma parte dos dados, ajustar o modelo na parte restante e verificar como o modelo prevê o conjunto que não foi utilizado. A este método se dá o nome de validação externa pois os conjunto usados para validação (às vezes chamado de teste) quanto para ajuste (muitas vezes chamado aprendizado, grande sacada de marketing) nunca trocam de papéis.

Na validação cruzada um ponto oras faz o papel de validação e oras é utilizado para o ajuste. Ou seja, o ajuste é realizado um grande número de vezes. O método em que o ajuste é repetido o maior número de vezes é o método de PRESS.

No método PRESS cada ponto é retirado do conjunto vez a vez, e o ajuste realizado com os n-1 pontos restantes. Desta forma o ajuste é repetido n vezes, cada vez sem um dos pontos no conjunto de ajuste. O critério utilizado é calculado pela soma dos resíduos ao quadrado, entre as predições “puras”, isto é, sem que o ponto seja usado para ajustar o modelo e os valores experimentais:

$$\mathrm{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2$$

Felizmente, para o caso de modelos lineares em relação aos parâmetros existe uma expressão analítica que permite avaliar o PRESS sem ter que repetir o procedimento de ajuste n vezes.

Em geral, para evitar ter que repetir o procedimento n vezes, o procedimento é simpificado dividindo o conjunto em um número pequeno de subconjuntos (da ordem de 10). Cada conjunto é deixado por sua vez de fora do procedimento de ajuste e posteriormente utilizado para testar a capacidade de predição do modelo, similarmente ao PRESS mas sem ser tão completo.

Em geral, na literatura utilizam-se duas numenclaturas, a RMSECV, raíz quadrada média de validação cruzada, e a MSECV, média quadrática de validação cruzada. A grande vantagem da primeira é que tem a mesma dimensão da variável à qual ela se refere.

V.2 CRITÉRIOS DE SELEÇÃO DE MODELOS: CRITÉRIOS APROXIMADOS

Existe uma infindade de critérios, cada qual fazendo referência a um autor famoso (ou que se tornou famoso em consequência de tal). Eles são muito utilizados, ainda, em controle de processos por exemplo. Eles começaram a surgir nos anos 60.

Alguns destes critérios estão listados a seguir. Mas antes algumas definições. Vamos supor que temos um modelo que podemos chamar de "o mais complexo". A partir dele queremos saber quais os regressores que podemos extrair. Este modelo "mais complexo" tem p parâmetros, ou regressores.

Chamamos $\hat{\sigma}^2$ à estimativa da variância do erro obtida com o ajuste deste modelo "mais complexo", sendo $\hat{y}_i$ as predições obtidas deste mesmo modelo:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2}{n - p}$$

Quando usarmos um modelo mais simples, com r parâmetros, chamaremos a variâmcia obtida com este modelo mais simples de $\hat{\sigma}_r^2$ e é dada por:

$$\hat{\sigma}_r^2 = \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_{ir}\right)^2}{n - r}$$

Onde $\hat{y}_{ir}$ são as predições obtidas com o modelo reduzido.

Assim, temos:

- C_p de Mallows: $C_p = \frac{\hat{\sigma}_r^2}{\hat{\sigma}^2}(n - r) + 2r$

- J_p de Hocking (também chamado de *FPE, Final Prediction Error*, por Ljung, 1999): $J_p = \hat{\sigma}_r^2(n + r)$

- S_p de Hocking também: $S_p = \hat{\sigma}_r^2 \frac{(n - r)}{(n - r + 1)(n - r - 1)}$

- Akaike: $AIC = \ln\left(\hat{\sigma}_r^2 \frac{(n - r)}{n}\right) + 2\frac{r}{n}$

ou aproximadamente: $AIC = \ln\left(\hat{\sigma}_r^2 \frac{(n - r)}{n}\left(1 + 2\frac{r}{n}\right)\right)$

- Press: $PRESS = \sum_{i=1}^{n} \frac{(y_i - \hat{y}_{ir})^2}{(1 - h_{iir})^2}$ onde h_{iir} são os elementos da diagonal da matriz chapéu do modelo reduzido.

- BIC (*Bayesian Information Criterion)* ou MDL(*Minimum Descriptor Length*) : $BIC = \ln\left(\hat{\sigma}_r^2 \frac{(n - r)}{n}\left(1 + \frac{r\ln(n)}{n}\right)\right)$

V.3 SELEÇÃO DE PARÂMETROS DE TENDÊNCIA

A seguir apresentamos alguns métodos que utilizam parâmetros de tendência para controlar o sobreajuste são:

V.3.1 o estimador de Componentes Principais, chamado de PCR. Nele o parâmtero de tendência é o número de componentes principais a serem retidos

no modelo. Seja a decomposição em auto-valores e auto-vetores, em que **D** contém os auto-valores ordenados do maior ao menor:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{S}\,\mathbf{D}\,\mathbf{S}^{\mathrm{T}}$$

A base do estimador de componentes principais é tomar esta decomposição e truncá-la em um determinado componente r:

$$\mathbf{S} = \left[\mathbf{s}_1\ \mathbf{s}_2 \cdots \mathbf{s}_{\mathrm{r}}\ \mathbf{s}_{\mathrm{r+1}} \ldots \mathbf{s}_{\mathrm{p}}\ \right]$$

$$\mathbf{S}_{\mathrm{r}} = \left[\mathbf{s}_1\ \mathbf{s}_2 \cdots \mathbf{s}_{\mathrm{r}}\right]$$

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & & & & \\ 0 & 0 & & \lambda_{\mathrm{r}} & & & \\ 0 & 0 & & & \lambda_{\mathrm{r+1}} & & \\ \vdots & \vdots & & & & \ddots & \\ 0 & 0 & & & & & \lambda_{\mathrm{p}} \end{bmatrix}$$

$$\mathbf{D}_{\mathrm{r}} = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & \lambda_{\mathrm{r}} \end{bmatrix}$$

E o estimador definido por:

$$\hat{\boldsymbol{\theta}}^{\mathrm{PCR}} = \mathbf{S}_{\mathrm{r}}\,\mathbf{D}_{\mathrm{r}}^{-1}\,\mathbf{S}_{\mathrm{r}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

V.3.2 O estimador de Ridge. Nele utiliza-se na equação normal um parâmetros de tendência k (hiperparâmetro):

$$\left(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \mathrm{k}\,\mathbf{I}\right)\hat{\boldsymbol{\theta}}^{\mathrm{RR}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$$

Pode-se mostrar que isto é equivalente a resolver:

$$\hat{\theta}^{\mathrm{RR}} = \arg \min_{\theta} \sum_{\mathrm{i=1}}^{\mathrm{n}} \left(\mathrm{y_i} - \mathrm{x_i^T}\theta\right)^2 + \mathrm{k}\sum_{\mathrm{i=1}}^{\mathrm{p}} \theta_{\mathrm{i}}^2$$

Esta fórmula também pode ser interpretada como a fórmula do estimador Bayesiano de máximo a posteriori com um a priori gaussiano com média **0** e variância σ^2/k .

V.3.3 Estimador de Lasso

O estimador de Lasso pode ser definido como:

$$\hat{\theta}^{LA} = \arg \min_{\theta} \sum_{i=1}^{n} \left(y_i - x_i^T \theta \right)^2 + k \sum_{i=1}^{p} \left| \theta_i \right|$$

Esta fórmula também pode ser interpretada como a fórmula Bayesiano de máximo a posteriori com um a priori que obedece à distribuição de Laplace com média **0** e variância inversamente proporcional a k. A vantagem do Lasso sobre o estimador de ridge é que a solução apresenta mais elementos zeros.

V.3.4 Estimador *Elastic Net*

O estimador *Elastic Net* pode ser definido como:

$$\hat{\theta}^{EN} = \arg \min_{\theta} \sum_{i=1}^{n} \left(y_i - x_i^T \theta \right)^2 + k_1 \sum_{i=1}^{p} \theta_i^2 + k_2 \sum_{i=1}^{p} \left| \theta_i \right|$$

Este estimador é um combinação entre o estimador de Lasso e o de Ridge.