

ODHAD ŘÁDU AUTOREGRESNÍHO MODELU

Ladislav Tomášek

1. Úvod

Při studiu autoregresních modelů důležitou roli hraje stanovení řádu modelu. Z literatury jsou známy některé metody, které určují řád modelu nepřímou, na základě testů hypotéz o nulových regresních koeficientech. Byly však navrženy i přímé metody, které využívají skutečnosti, že odhad rozptylu v modelech s podceněným řádem vede obvykle k hodnotám vyšším než je skutečná hodnota rozptylu, zatímco v modelech s přeceněným řádem jsou hodnoty obou rozptylů srovnatelné.

Autoregresi posloupností řádu p budeme rozumět stacionární posloupnost $\{x_t\}$ náhodných veličin, které splňují vztah

$$x_t + a_1 x_{t-1} + \dots + a_p x_{t-p} = e_t, \quad (1.1)$$

kde $a_p \neq 0$ a kde $\{e_t\}$ je posloupnost nekorelovaných náhodných veličin s nulovými středními hodnotami a stejnými rozptyly σ^2 . Parametry a_1, \dots, a_p jsou vázány podmínkou, že všechny kořeny polynomu

$$z^p + a_1 z^{p-1} + \dots + a_p = 0$$

leží uvnitř jednotkového kruhu.

Při odhadech autoregresních parametrů vycházíme z posloupnosti konečné délky N . Neznámé parametry $a' = (a_1, \dots, a_p)$ se odhadují obvykle metodou nejmenších čtverců. Uvažuje se funkce kvadratických odchylek

$$S(a) = \sum_{t=p+1}^N (x_t + a_1 x_{t-1} + \dots + a_p x_{t-p})^2. \quad (1.2)$$

Hodnoty vektorového parametru a , které minimalizují součet $S(a)$ se považují za odhad metodou nejmenších čtverců. Řešení \hat{a} dostaneme ze soustavy normálních rovnic

$$\sum_{t=p+1}^N x_{t-j} (x_t + a_1 x_{t-1} + \dots + a_p x_{t-p}) = 0, \quad j=1, \dots, p.$$

označíme-li

$$c_{ij} = \sum_{t=p+1}^N x_{t-i} x_{t-j}, \quad i, j = 0, \dots, p,$$

$$c_0 = (c_{01}, \dots, c_{0p})', \quad C = (c_{ij}), \quad i, j = 1, \dots, p,$$

můžeme za předpokladu regularity matice C vyjádřit řešení ve tvaru

$$\hat{a} = -C^{-1}c_0$$

Odhad parametru σ^2 se získá na základě veličiny

$$\hat{\sigma}^2 = S(a)/(N-p) = (c_{00} - c_0' C^{-1} c_0)/(N-p) \quad (1.3)$$

2. Některé metody odhadu řádu autoregresního modelu

Jak již bylo poznamenáno v úvodu, existují metody pro přímý odhad řádu modelu, které jsou založeny na odhadu parametru σ^2 za podmínky, že řád modelu je p . Takový odhad budeme dále označovat $\hat{\sigma}_p^2$.

Podstatou takových metod je určení minima v konečné posloupnosti vhodně transformovaných hodnot $\hat{\sigma}_p^2$. Je zřejmé, že tato transformace musí znevýhodňovat modely

s příliš vysokými hodnotami p a dále musí být do transformace zahrnut i parametr N .

Jedním z prvních, kdo se zabývali otázkou odhadu řádu autoregresního modelu, byl H.Akaike. Při konstrukci nejlepší lineární predikce prvku x_{N+1} dospěl k výrazu pro odhad reziduálního rozptylu

$$E(\hat{x}_{N+1} - x_{N+1})^2 = \sigma^2(1 + \frac{p}{N}), \quad (2.1)$$

kde \hat{x}_{N+1} se získá na základě p předcházejících hodnot a odhady autoregresních koeficientů na základě metody nejmenších čtverců. Neznámý parametr σ^2 ve výrazu (2.1) Akaike nahradil veličinou $s_p^2 = S(\hat{a})/(N-2p)$ a tento odhad rozptylu označil jako FPE (final prediction error).

Vzhledem k poměru hodnot p a N můžeme jednoduchými úpravami tuto veličinu postupně převést na tvar

$$FPE_p = \frac{N-p}{N-2p} \hat{\sigma}_p^2 (1 + \frac{p}{N}) = \hat{\sigma}_p^2 (1 + 2\frac{p}{N}).$$

Tato veličina bývá uváděna obvykle v logaritmickém tvaru a označována jako AIC

$$\ln \hat{\sigma}_p^2 + 2p/N. \quad (2.2)$$

Později zkoumal otázku odhadu řádu modelu G.Schwarz. Na základě bayesovských úvah o řádu modelu jako veličině s apriorním rozdělením dospěl ke kritériu pro odhad řádu modelu založeném na veličinách

$$\ln L_p(x_1, \dots, x_N) - \frac{1}{2} p \ln N, \quad (2.3)$$

kde L_p je věrohodnostní funkce v modelu o řádu p . Aplikací na autoregresní model lze dospět k transformační funkci

$$\ln \hat{\sigma}_p^2 + p \frac{\ln N}{N}. \quad (2.4)$$

Ke stejnému výsledku došel i J.Rissanen na základě úvah založených na co nejúspornějším záznamu posloupnosti x_1, \dots, x_N .

Přibližně ve stejné době zkoumali uvažovaný problém E.J.Hannan a B.G.Quinn. Jako funkci penalizující počet parametrů navrhl

$$h(n) = 2c \frac{\ln \ln N}{N}, \quad c > 1. \quad (2.5)$$

Odhad řádu modelu založený na takové funkci je konzistentní. Důkaz je založen na odhadu rozptylu pomocí parciálních autokorelací a na zákonu o iterovaném logaritmu.

Později navrhl J.Anděl, M.G.Perez a A.I.Negrão kritérium pro odhad řádu lineárního regresního modelu v obecném tvaru

$$A_p = \hat{\sigma}_p^2 (1 + p w_N), \quad (2.6)$$

kde $w_N \rightarrow 0$ a $N w_N \rightarrow \infty$ při $N \rightarrow \infty$. Za těchto předpokladů je asymptoticky

$$P(A_p < A_k, k \neq p) = 1.$$

Konkrétně byla navržena penalizační funkce w_N ve tvaru

$$w_N = c N^{-\alpha}, \quad c > 0, \quad 0 < \alpha < 1/2. \quad (2.7)$$

Z předcházejících poznámek vyplývá, že kritérium pro odhad řádu autoregresního modelu můžeme obecně založit na veličinách

$$Q(p, N) = \ln \hat{\sigma}_p^2 + \frac{p}{N} h_N, \quad (2.8)$$

kde h_N je neklesající funkce N taková, že $h_N/N \rightarrow 0$ při $N \rightarrow \infty$. Veličinu h_N budeme dále nazývat penalizační faktor.

3. Simulace

K posouzení kvality různých metod odhadu řádu p a k prohloubení znalostí o penalizačním faktoru byly provedeny simulace autoregresních posloupností do třetího řádu. Simulované posloupnosti byly získány na základě vztahu (1.1), přičemž pro $t \leq p$ bylo $x_t = 0$ a pro zlepšení stacionarity posloupností bylo prvních 60 členů posloupnosti vynecháno. Generující autoregresní koeficienty byly zvoleny následovně:

p	a_1	a_2	a_3
1	0.36		
2	0.24	0.36	
3	0.30	0.40	0.36

Autoregresní posloupnosti byly simulovány pro hodnoty $N = 50, 75, 100, \dots, 250$. U každé posloupnosti byly pak vypočteny odhady $\hat{\sigma}_p^2$ pro $p = 0, 1, \dots, 7$. Pro každý řád a každou hodnotu N bylo provedeno 1000 realizací.

Kvalitu metod uvedených v části 2 ilustrují následující tabulky.

N	Akaikeova metoda								Schwarzova metoda							
	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
50	87	22	442	116	92	70	75	96	36	20	89	419	149	96	92	99
75	40	12	552	132	94	55	49	66	1	7	53	530	152	98	81	78
100	13	7	601	131	81	56	49	62	1	1	8	573	154	101	89	73
125	3	3	582	144	98	70	49	51	0	0	4	609	143	91	82	71
150	2	1	626	131	85	50	48	57	0	0	2	606	149	96	74	73
175	2	0	601	129	91	71	56	50	0	0	0	614	137	105	70	74
200	0	0	618	125	95	49	67	46	0	0	0	651	120	95	71	63
225	0	0	638	128	71	65	47	51	0	0	0	603	165	93	58	81
250	0	0	599	144	91	65	50	51	0	0	0	635	147	72	73	73

Tab. 3.1 Akaikeova metoda pro AR(2) a AR(3)

N	Akaikeova metoda								Schwarzova metoda							
	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
50	347	38	511	54	24	12	8	6	204	42	168	459	78	26	9	14
75	216	31	666	61	17	6	1	2	53	39	127	697	53	18	9	4
100	121	17	784	66	9	2	0	1	21	9	60	819	71	15	3	2
125	39	13	859	65	17	5	2	0	10	4	27	887	48	20	3	1
150	24	10	925	33	6	1	1	0	1	0	11	931	42	12	1	2
175	16	1	926	39	15	3	0	0	1	1	2	944	39	11	1	1
200	3	0	943	36	13	3	2	0	0	0	4	941	41	12	2	0
225	1	0	956	37	5	1	0	0	0	0	0	955	36	7	1	1
250	0	1	952	41	5	0	1	0	0	0	0	971	27	1	1	0

Tab. 3.2 Schwarzova metoda pro AR(2) a AR(3)

N	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
50	374	34	501	48	20	10	7	6	226	43	171	447	73	22	7	11
75	223	32	665	56	15	6	1	2	55	40	133	691	52	17	9	3
100	120	17	783	66	10	2	0	2	21	9	60	817	72	16	3	2
125	33	13	862	65	18	6	2	1	10	4	27	885	49	21	3	1
150	21	9	926	35	7	1	1	0	1	0	8	926	49	13	1	2
175	13	1	927	41	15	3	0	0	0	1	1	936	45	15	1	1
200	3	0	840	39	13	3	2	0	0	0	1	937	46	14	2	0
225	1	0	941	48	9	1	0	0	0	0	0	942	44	11	2	1
250	0	0	942	50	6	1	1	0	0	0	0	962	35	1	2	0

Tab. 3.3 Hannan-Quinnova metoda pro AR(2) a AR(3), $c=1.5$

N	0	1	2	3	4	5	6	7	0	1	2	3	4	5	6	7
50	236	39	536	78	46	21	20	24	112	34	139	497	101	55	33	29
75	179	32	674	73	25	8	7	2	42	37	123	691	61	24	14	8
100	130	17	783	58	9	2	0	1	27	11	63	816	66	13	3	1
125	71	11	855	49	11	2	1	0	14	6	37	892	36	14	0	1
150	53	13	908	24	2	0	0	0	4	4	18	938	32	4	0	0
175	37	1	936	20	6	0	0	0	4	0	10	965	16	5	0	0
200	13	2	970	10	4	1	0	0	2	0	6	972	17	3	0	0
225	10	1	980	9	0	0	0	0	0	1	4	987	8	0	0	0
250	5	1	987	7	0	0	0	0	0	0	1	994	4	1	0	0

Tab. 3.4 Andělova metoda pro AR(2) a AR(3), $c=0.3, \alpha=0.4$

Z četností uvedených tabulek je patrné, že Akaikeova metoda obecně přeceňuje řád modelu, což je výrazné zejména u dlouhých posloupností. Tato skutečnost je ostatně známa z literatury. O něco lépe fungují kritéria Schwarz a Hannana, kde podíl přeceněných modelů je u delších posloupností relativně nižší, u krátkých posloupností je naopak vyšší podíl nedocenených modelů (vzhledem k Akaikeově kritériu). Relativně nejlepší výsledky dává metoda Andělova.

4. Empirické odhady penalizačního faktoru

Na základě výsledků simulací byla dále studována otázka, jaké mezní hodnoty penalizačního faktoru lze připustit při předepsané chybě přecenění, resp. podcenění řádu modelu.

Uvažujme autoregresní posloupnost řádu p . Jestliže má kritérium pro volbu řádu tvar

$$Q(j, N) = \ln \hat{\sigma}_j^2 + h_N \frac{j}{N}, \quad j=0, \dots, K, \quad (4.1)$$

je minimální hranice penalizačního faktoru pro odhad, který nepřecení řád p , dána výrazem

$$h_N^- = -N \min_{j > p} \frac{\ln \hat{\sigma}_j^2 - \ln \hat{\sigma}_p^2}{j - p}. \quad (4.2)$$

Analogicky maximální hranicí penalizačního faktoru proti podcenění řádu p lze

vyjádřit ve tvaru

$$h_N^+ = -N \max_{j < p} \frac{\ln \hat{\sigma}_j^2 - \ln \hat{\sigma}_p^2}{j - p} \quad (4.3)$$

Na základě hodnot h^- a h^+ vypočtených v jednotlivých simulovaných posloupnostech lze si utvořit představu o hodnotách penalizačního faktoru pro různé hodnoty N . Nejvhodnější charakteristiky jsou v tomto případě empirické kvantily, které jsou uvedeny v následujících tabulkách.

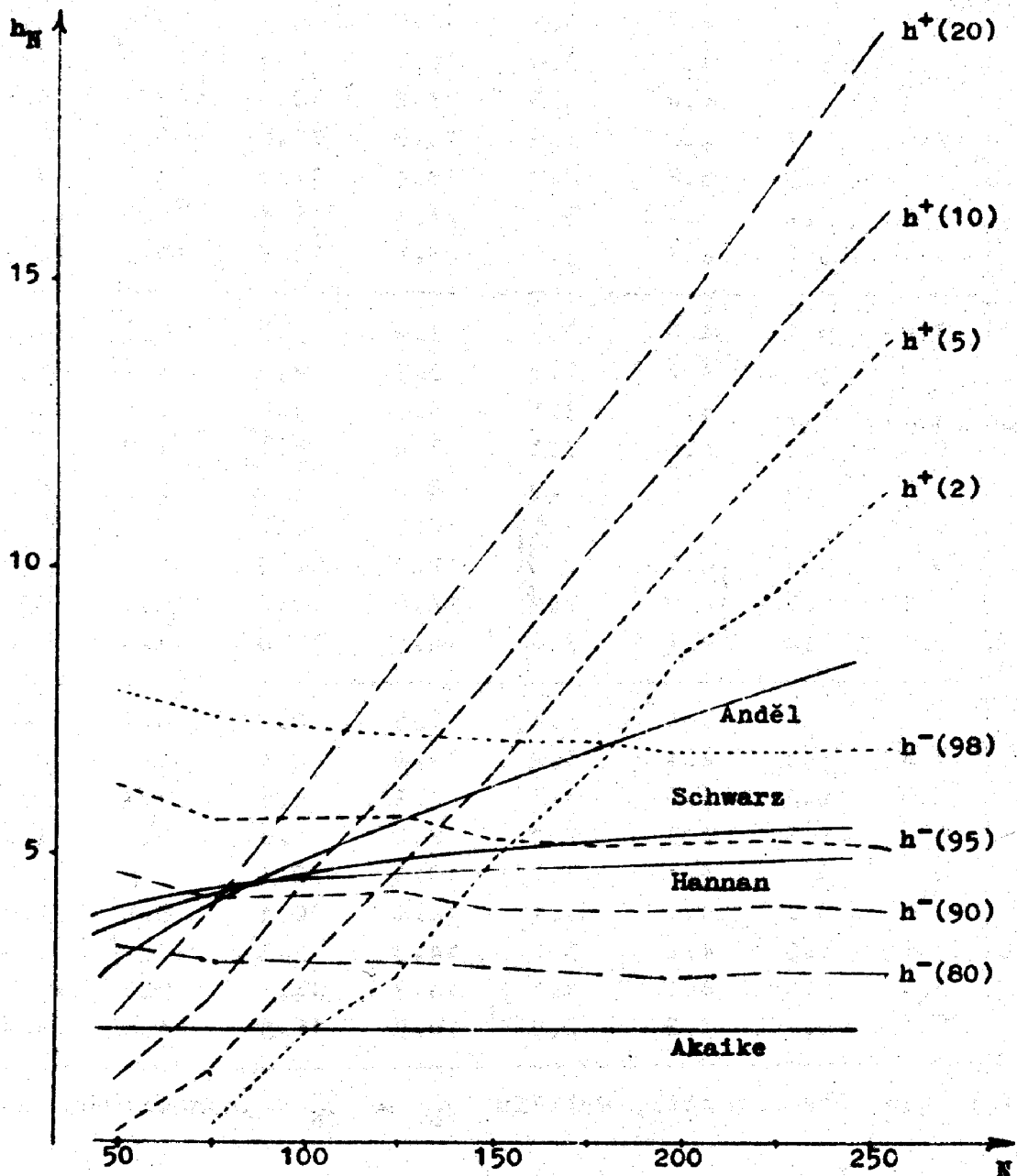
p	N	$h_N^-(98)$	$h_N^-(95)$	$h_N^-(90)$	$h_N^-(80)$	$h_N^+(20)$	$h_N^+(10)$	$h_N^+(5)$	$h_N^+(2)$
1	50	7.6	5.8	4.5	3.4	2.5	1.1	0.1	-0.7
	75	7.2	5.5	4.1	3.0	4.9	2.9	1.3	0.2
	100	7.1	5.1	4.2	3.1	7.6	5.5	3.5	2.2
	125	7.3	5.6	4.1	3.0	10.7	7.1	4.5	2.6
	150	7.4	5.8	4.4	3.3	13.2	10.3	7.7	5.1
	175	7.0	5.1	4.0	2.9	15.8	12.6	9.5	6.6
	200	6.3	4.9	3.8	2.8	18.4	14.5	11.8	8.8
	225	6.5	5.2	4.2	3.1	22.1	18.4	15.6	11.4
	250	7.2	5.4	4.1	3.0	23.6	19.7	16.3	12.9
2	50	7.8	6.1	4.6	3.3	2.3	1.3	0.6	-0.1
	75	7.8	5.6	4.4	3.2	3.7	2.5	1.7	1.0
	100	7.1	5.4	4.3	3.1	5.5	3.8	2.8	1.6
	125	7.4	5.8	4.6	3.3	7.6	5.7	4.8	2.9
	150	6.6	4.6	3.8	2.9	8.8	7.0	5.5	4.2
	175	7.4	5.4	4.1	3.0	10.9	9.0	7.2	5.8
	200	6.7	5.4	4.1	2.9	12.9	10.9	9.4	7.4
	225	6.6	5.2	4.2	2.9	14.8	12.2	10.7	8.6
	250	7.0	5.4	4.1	3.0	16.3	13.8	11.4	10.0
3	50	8.0	6.6	5.0	3.6	1.9	0.8	0.1	-0.4
	75	7.4	5.6	4.4	3.2	3.8	2.4	1.1	0.1
	100	7.5	6.2	4.6	3.3	6.1	4.6	3.1	1.9
	125	6.6	5.6	4.3	3.2	8.2	6.6	5.2	3.1
	150	7.2	5.3	4.1	3.0	10.4	8.4	7.3	5.7
	175	6.9	5.2	4.1	3.1	12.4	10.4	8.9	7.0
	200	7.1	5.5	4.2	3.1	14.3	12.5	10.5	8.9
	225	7.0	5.2	4.4	3.1	16.9	14.4	12.6	10.4
	250	6.1	4.7	3.9	2.9	18.9	16.5	14.8	12.4

Tab. 4.1 Empirické kvantily veličin h_N^- a h_N^+ v jednotlivých modelech

N	$h_N^-(98)$	$h_N^-(95)$	$h_N^-(90)$	$h_N^-(80)$	$h_N^+(20)$	$h_N^+(10)$	$h_N^+(5)$	$h_N^+(2)$
50	7.7	6.2	4.7	3.4	2.2	1.1	0.2	-0.4
75	7.4	5.6	4.3	3.2	4.0	2.5	1.3	0.3
100	7.2	5.6	4.3	3.2	6.2	4.5	3.1	1.9
125	7.1	5.7	4.4	3.2	8.4	6.3	4.8	2.9
150	6.9	5.3	4.1	3.1	10.4	8.1	6.3	4.8
175	7.0	5.2	4.1	3.0	12.5	10.2	8.4	6.4
200	6.8	5.3	4.1	2.9	14.5	12.1	10.2	8.5
225	6.7	5.3	4.2	3.1	16.8	14.2	11.9	9.5
250	6.8	5.2	4.0	3.0	18.6	15.9	13.7	11.1

Tab. 4.2 Souhrnné empirické kvantily veličin h_N^- a h_N^+

Vztah získaných kvantilů a penalizačních faktorů známých metod můžeme dobře posoudit z následujícího grafického vyjádření.



Obr. 4.1 Souhrnné empirické kvantily veličin h_N^- a h_N^+ v závislosti na N

Literatura

- [1] Akaike, H.: Fitting autoregressive models for prediction. *Ann.Inst.Stat. Math.* 21, 1969
- [2] Anděl, J., Perez, M.G., Negrao, A.J.: Estimating the dimension of a linear model. *Hybernetika*, Vol.17, 1981
- [3] Hannan, E.J., Quinn, B.G.: The determination of the order of an autoregression. *J.Roy.Stat.Soc., Ser.B* 41, 1979
- [4] Schwarz, G.: Estimating the dimension of a model. *Ann.Stat.* 6, 1978