Об оптимальном прогнозировании авторегрессионных временных рядов с интервальным цензурированием

Харин Ю.С.
НИИ прикладных проблем
математики и информатики БГУ
пр. Независимости, 4, г. Минск
Беларусь
kharin@bsu.by

Бодягин И.А.
НИИ прикладных проблем
математики и информатики БГУ
пр. Независимости, 4, г. Минск
Беларусь
bodiagin@cosmostv.by

Аннотация

Рассматривается модель авторегрессии 1-го порядка в случае интервального цензурирования. Построена оптимальная прогнозирующая статистика в случае, когда последнее наблюдение принадлежит заданному интервалу цензурирования. Вычислен риск прогнозирования. Проведено сравнение оптимальной прогнозирующей статистики с прогнозирующими статистиками, часто используемыми на практике. Представлены численные результаты.

1 Введение

Модель авторегрессии широко используется для описания временных рядов с зависимыми наблюдениями в технике, экономике, промышленности, транспорте, астрономии, химии, метеорологии (Бокс и Дженкинс, 1974). Для данной модели случай полных данных, когда наблюдается весь временной ряд, хорошо изучен (Андерсон, 1976). Также хорошо изучен и случай, когда часть наблюдений временного ряда пропущена (Литтл и Рубин, 1990; Харин 2008). Особый интерес вызывает пока малоизученная ситуация, когда часть наблюдений временного ряда известна точно, а остальные элементы ряда подвержены интервальному цензурированию (Park, Genton и Ghosh, 2007). Такая ситуация может возникать из-за наличия у приборов пределов измерения, высокой стоимости проведения точных измерений, разладки оборудования и возникает в технике, экономике, медицине, метеорологии, физике и других областях (Sen, 1995; Gomez, Espinal и Lagakos, 2003).

2 Оптимальная прогнозирующая статистика

Пусть временной ряд описывается моделью авторегрессии первого порядка AP(1):

$$x_t = \theta x_{t-1} + u_t, \tag{1}$$

где θ — коэффициент авторегрессии, такой что $0 \le |\theta| < 1$; $u_t,\ t \in \mathbb{Z}$ — независимые одинаково распределенные случайные величины, имеющие нормальный закон распределения вероятностей: $u_t \sim \mathcal{N}(0, \sigma^2)$. Наблюдаются случайные события:

$$A_1^* = \{x_1 \in A_1\}, \ldots, A_T^* = \{x_T \in A_T\}, \tag{2}$$

где T — длительность наблюдения, $A_1, \ldots, A_T \in \mathcal{B}(\mathbb{R}^1)$ — борелевские множества. Если $A_i = \{x_i\}$ — одноэлементное множество, то наблюдение в момент времени $t = i$ известно точно. Если $A_i = (a_i, b_i)$ — числовой интервал $(b_i > a_i)$, то мы имеем случай интервального цензурирования; интервал (a_i, b_i) называется интервалом цензурирования, длину его будем обозначать $\tau_i = b_i - a_i$.

Прогнозирующая статистика для величины x_{T+1} в будущий момент времени $t = T + 1$ является числовой функцией наблюдаемых событий и имеет общий вид:

$$\hat{x}_{T+1} = f(A_T^*, A_{T-1}^*, \ldots, A_1^*).$$

Условный среднеквадратический риск:

$$r(T) = \mathrm{E}\left\{(\hat{x}_{T+1} - x_{T+1})^2 | A_T^*, A_{T-1}^*, \ldots, A_1^*\right\}. \tag{3}$$

Пусть для последних q событий ($T > q \geq 1$) среди наблюдаемых событий (2) выполнено $A_T = (a_T, b_T), \ldots, A_{T-q+1} = (a_{T-q+1}, b_{T-q+1})$, т.е. точные значения последних q наблюдений не известны; известно лишь, что они цензурированы и принадлежат заданным интервалам, а для остальных событий выполнено $A_{T-q} = \{x_{T-q}\}, \ldots, A_1 = \{x_1\}$.

Рассмотрим задачу построения оптимальной прогнозирующей статистики для x_{T+1} по имеющимся наблюдениям при известных параметрах модели (1). Обозначим: $n(\cdot)$ - гауссовскую плотность,

$$I_1(l,m) = \int_{a_T}^{b_T} \ldots \int_{a_{T-m+1}}^{b_{T-m-1}} x_T^l p(x_T, \ldots, x_{T-m+1}|x_{T-m}) dx_{T-m+1} \ldots dx_T,\ l, m \in \mathbb{N}_0,$$

где $p(x_T, \ldots, x_{T-q+1}|x_{T-q}) = n(x_T, \ldots, x_{T-q+1}|\bar{\mu}, \bar{\Sigma})$, $\bar{\mu} = \Sigma_{12}\Sigma_{22}^{-1}x_{T-q}, \bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$;

$$\sigma_{ij} = Cov\left\{x_{T-i+1}, x_{T-j+q}\right\}, \Sigma = (\sigma_{ij}) = \begin{matrix} \\ q \\ 1 \end{matrix} \begin{matrix} \begin{matrix} q & 1 \end{matrix} \\ \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix} \end{matrix},\ i, j = 1, 2, \ldots, q+1.$$

Теорема 1. *Пусть наблюдаются значения $x_1, \ldots, x_{T-q}$ и события $A_{T-q+1}^* = \{x_{T-q+1} \in (a_{T-q+1}, b_{T-q+1})\}$, ..., $A_T^* = \{x_T \in (a_T, b_T)\}$. Тогда оптимальной в смысле минимума риска (3) является прогнозирующая статистика:*

$$\hat{x}_{T+1} = \theta \mathrm{E}\left\{x_T|A_T^*, \ldots, A_{T-q+1}^*, x_{T-q}\right\} = \theta\frac{I_1(1,q)}{I_1(0,q)},\ r_0(T) = \sigma^2 + \theta^2\left(\frac{I_1(2,q)}{I_1(0,q)} - \left(\frac{I_1(1,q)}{I_1(0,q)}\right)^2\right). \quad (4)$$

Теорема 2. *Пусть наблюдаются значения $x_1, \ldots, x_{T-q}$, события $A_{T-q+1}^* = \{x_{T-q+1} \in (a_{T-q+1}, b_{T-q+1})\}$, ..., $A_T^* = \{x_T \in (a_T, b_T)\}$ и для некоторого k ($k \in \{0, \ldots, q-1\}$) $\tau_{T-k} \to 0$. Тогда для оптимальной прогнозирующей статистики выполнено следующее соотношение:*

$$\hat{x}_{T+1} = \theta \mathrm{E}\left\{x_T|A_T^*, \ldots, A_{T-q+1}^*, x_{T-q}\right\} \to \theta \mathrm{E}\left\{x_T|A_T^*, \ldots, A_{T-k+1}^*, x_{T-k}\right\}.$$

Из теоремы 2 следует, что для построения оптимального прогноза в случае модели AP(1), достаточно знать только последнее наблюдаемое значение x_{T-k} и все интервалы цензурирования $A_{T-k+1}, \ldots, A_T$ после него.

Следствие 1. *Пусть выполнены условия теоремы 1 и $a_T \to -\infty, \ldots, a_{T-q+1} \to -\infty$, $b_T \to +\infty, \ldots, b_{T-q+1} \to +\infty$, тогда оптимальная прогнозирующая статистика и ее риск имеют вид:* $\hat{x}_{T+1} = \theta^{q+1}x_{T-q},\ r_0(T) = \sigma^2\sum\limits_{i=0}^{q}\theta^{2i}$.

Условия следствия 1 означают, что в моменты времени $T, \ldots, T-q+1$ наблюдения $x_T, \ldots, x_{T-q+1}$ пропущены. Такая ситуация является изученной ранее и для нее получены результаты (Харин, 2008), которые совпадают со следствием 1.

Следствие 2. *Пусть выполнены условия теоремы 1 и $a_T \to b_T, \ldots, a_{T-q+1} \to b_{T-q+1}$. Тогда оптимальная прогнозирующая статистика и ее риск примут вид: $\hat{x}_{T+1} = \theta x_T,\ r_0(T) = \sigma^2$.*

Условия следствия 2 означают, что в пределе известны значения $x_T = a_T = b_T, \ldots, x_{T-q+1} = a_{T-q+1} = b_{T-q+1}$, т.е. мы имеем случай полных данных. Для этого случая также ранее получены результаты, например Андерсоном (1976), которые совпадают со следствием 2.

3 Случай $q = 1$

Чтобы более детально исследовать зависимость риска прогнозирования от длины интервала цензурирования, рассмотрим частный случай $q = 1$ модели (1), т.е. последнее значение временного ряда x_T цензурировано интервалом (a_T, b_T), а предпоследнее значение x_{T-1} известно точно. Для упрощения обозначений вместо a_T и b_T будем писать a и b. Обозначим: $\varphi(\cdot)$, $\Phi(\cdot)$ — плотность и функция стандартного нормального распределения $\mathcal{N}(0,1)$; $(a-\theta x_{T-1})/\sigma = A_1$, $(b-\theta x_{T-1})/\sigma = B_1$, $(a\sqrt{1-\theta^2})/\sigma = A_2$, $(b\sqrt{1-\theta^2})/\sigma = B_2$, $(a^k b^l\varphi(a) - a^l b^k\varphi(b))/(\Phi(a) - \Phi(b)) = \psi_{k,l}(a,b), k, l \in \{0,1\}$, $\tau = b - a$.

Теорема 3. *Пусть наблюдаются значение x_{T-1} и событие $A_T^* = \{x_T \in (a,b)\}$. Тогда оптимальная прогнозирующая статистика и ее минимальный риск примут вид:*

$$\hat{x}_{T+1} = \theta \mathrm{E}\{x_T | A_T^*, x_{T-1}\} = \theta^2 x_{T-1} + \theta\sigma\psi_{0,0}(A_1, B_1), \tag{5}$$
$$r_0(T) = \sigma^2 \left(1 + \theta^2 \left(1 - \psi_{0,0}^2(A_1, B_1) + \psi_{1,0}(A_1, B_1)\right)\right).$$

При $\tau \to 0$ справедливо асимптотическое разложение риска прогнозирования:

$$r_0(T) = \sigma^2 + \frac{\theta^2}{4}\tau^2 + o(\tau^2).$$

Известно (Андерсон, 1976), что в случае полных данных риск прогнозирования для оптимального прогноза равен $r_0 = \sigma^2$. Для оценки чувствительности риска к длине τ интервала цензурирования (a,b) воспользуемся коэффициентом неустойчивости риска (Харин, 2008): $æ = (r - r_0)/r_0$.

Следствие 3. *В условиях теоремы 3 для коэффициента неустойчивости риска справедливо приближение: $æ_0 \approx \theta^2\tau^2/4\sigma^2$.*

ε-допустимой ($\varepsilon > 0$) длиной интервала цензурирования называется (Харин, 2008) такая наибольшая длина $\tau(\varepsilon)$, при которой $æ \le \varepsilon$.

Следствие 4. *В условиях теоремы 3 для ε-допустимой длины интервала цензурирования справедливо приближение: $\tau_0(\varepsilon) \approx 2\sigma\sqrt{\varepsilon}/\theta$.*

Рассмотрим другие возможные прогнозирующие статистики. Одной из часто используемых на практике статистик является прогнозирующая статистика:

$$\hat{x}_{T+1} = \theta \mathbf{E}\left\{x_T | x_T \in (a,b)\right\}. \tag{6}$$

Теорема 4. *Пусть наблюдаются значения $x_1, \ldots, x_{T-1}$, событие $A_T = \{x_T \in (a,b)\}$ и используется прогнозирующая статистика (6). Тогда:*

- *прогнозирующая статистика (6) может быть вычислена по следующей формуле:*

$$\hat{x}_{T+1} = \frac{\theta\sigma}{\sqrt{(1-\theta^2)}}\psi_{0,0}(A_2, B_2);$$

- *риск прогнозирования равен:*

$$r_1(T) = \frac{\sigma^2}{1-\theta^2}\left(1 + \theta^2\left(\psi_{1,0}(A_2, B_2) - \psi_{0,0}^2(A_2, B_2)\right)\right);$$

- *справедливо следующее асимптотическое разложение риска прогнозирования при $\tau \to 0$:*

$$r_1(T) = \sigma^2 + \frac{\theta^2}{4}\tau^2 + o(\tau^2);$$

- *для коэффициента неустойчивости риска и ε-допустимой длины интервала цензурирования справедливы приближения:*

$$æ_1 \approx \theta^2\tau^2/4\sigma^2, \quad \tau_1(\varepsilon) \approx 2\sigma\sqrt{\varepsilon}/\theta.$$

Рассмотрим еще одну применяемую на практике прогнозирующую статистику:

$$\hat{x}_{T+1} = \theta\frac{a+b}{2}, \tag{7}$$

которая основывается на предположении, что все значения x_T в (a,b) равновозможны.

Теорема 5. *Пусть наблюдаются значения $x_1, \ldots, x_{T-1}$, событие $A_T = \{x_T \in (a,b)\}$ и используется прогнозирующая статистика (7). Тогда риск прогнозирования равен:*

$$r_2(T) = \frac{\sigma^2}{1-\theta^2} + \frac{\theta^2(a+b)^2}{4} + \frac{\theta^2\sigma^2}{1-\theta^2}\psi_{0,1}(B_2, A_2).$$

Полученные результаты обобщаются для случая модели АР(p) авторегрессии порядка $p > 1$.

4 Численные результаты

В случае когда $q = 1$, для сравнения точности прогнозирующих статистик (5), (6), (7) и иллюстрации результатов теорем 3 – 5, были проведены численные эксперименты. Для оценивания риска прогнозирования при каждом фиксированном τ использовался метод Монте-Карло с числом прогонов равным $N = 10000$. Были взяты следующие значения параметров: $\theta = 0.8$, $\sigma^2 = 1$, $T = 100$, $\tau = b - a \in \{0, 0.5, 1, \ldots, 15\}$. Последнее наблюдение заменялось случайным интервалом цензурирования (a, b) фиксированной длины τ так, что длина интервала (a, x_T) равна $\alpha\tau$, а длина интервала (x_T, b) равна $(1 - \alpha)\tau$, где α — случайная величина, равномерно распределенная на отрезке $[0, 1]$.

На рисунке 1(а) изображены зависимости от τ экспериментальных значений риска прогнозирования для всех трех прогнозирующих статистик. На рисунках 1(б)–1(г) изображены экспериментальные (с 95% доверительными границами) и теоретические значения риска прогнозирования для статистик (5), (6) и (7) в зависимости от τ. Экспериментальные и теоретические результаты находятся в хорошем согласии.

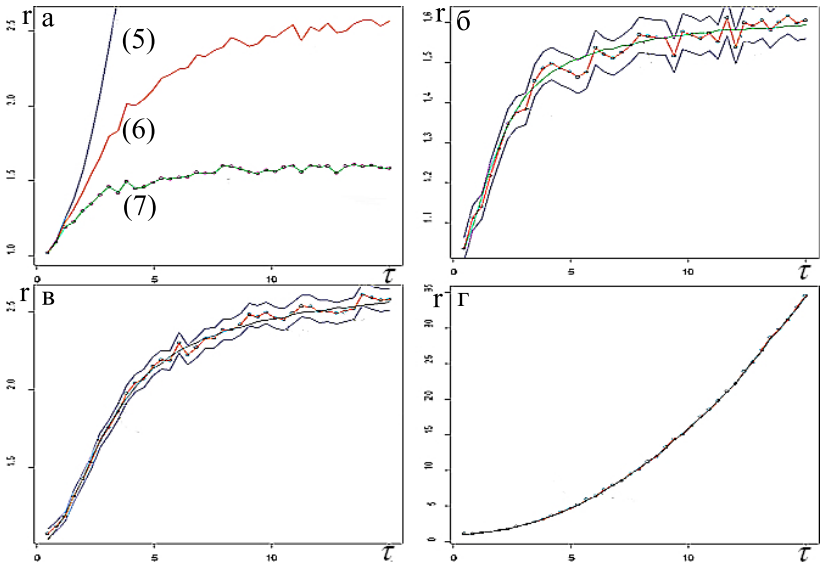

Рис. 1: Результаты численных экспериментов

Список литературы

[1] Андерсон Т. (1976). *Статистический анализ временных рядов.* Москва: Мир.

[2] Бокс Дж., Дженкинс Г. (1974). *Анализ временных рядов. Прогноз и управление* / Пер. с англ. Т.1. Москва: Мир.

[3] Харин Ю.С. (2008). *Оптимальность и робастность в статистическом прогнозировании.* Минск: БГУ.

[4] Литтл Р. Дж. А., Рубин Д.Б. (1990). *Статистический анализ данных с пропусками* / Пер. с англ. Москва: Финансы и статистика.

[5] Park J.W., Genton M.G., Ghosh S.K. (2007). Censored time series analysis with autoregressive moving average models. *The Canadian Journal of Statistics 35*(1), 151–168.

[6] Sen P.K. (1995). Censoring in theory and practice: statistical perspectives and controversies. *IMS Lectures Notes 27.*

[7] Gomez G., Espinal A., Lagakos W. (2003). Inference for a linear model with an interval-censored covariate. *Statistics in medicine 22*, 409 – 425.