

『AI・データサイエンスのための数学入門』学習補助資料

何時点ずらすと、似た動きになるか

自己相関・コレログラム・白色雑音・差分

1 自己相関は「時間をずらした比較」

ラグ 1 では $(y_2, y_1), (y_3, y_2), \dots, (y_n, y_{n-1})$ を比較します。ラグ 2 では $(y_3, y_1), \dots, (y_n, y_{n-2})$ を比較します。ラグが大きくなると、使える組は $n - k$ 個に減ります。

時点 t	1	2	3	4	5
現在の値 y_t	1	3	2	4	3
1 時点前 y_{t-1}	—	1	3	2	4
2 時点前 y_{t-2}	—	—	1	3	2

何を「似ている」と判断するのか

全体の平均からの偏差 $z_t = y_t - \bar{y}$ を考えます。本文の式は

$$r_k = \frac{\sum_{t=k+1}^n z_t z_{t-k}}{\sum_{t=1}^n z_t^2}$$

です。分子の各項は、偏差の符号が同じなら正、異なれば負になります。

現在の偏差	過去の偏差	積の符号
正	正	正
負	負	正
正	負	負
負	正	負

「前日より上がるか」とは違う

正の自己相関は、平均より高い値の後に平均より高い値が続くやすい、という関係です。前日から増加するかどうかを直接測るものではありません。また、相関だけから因果関係は決まりません。

ラグの単位は観測間隔に依存します。日次データのラグ 7 は 7 日前、月次データのラグ 7 は 7 か月前です。欠測日を詰めて並べると、行の差と日数の差が一致しなくなることに注意します。

2 定義の性質を確かめる

ここでは $0 \leq k < n$ とし、少なくとも 2 つの観測値が異なる、すなわち $D = \sum z_t^2 > 0$ とします。すべて同じ値なら分母が 0 となり、この標本 ACF は定義できません。

ラグ 0 は必ず 1

$$r_0 = \frac{\sum_{t=1}^n z_t^2}{\sum_{t=1}^n z_t^2} = 1.$$

同じ時点どうしの比較なので、情報の中心はラグ 1 以降にあります。

範囲は「おおむね」ではなく、厳密に $[-1, 1]$

コーシー・シュワルツの不等式より、

$$\left| \sum_{t=k+1}^n z_t z_{t-k} \right| \leq \sqrt{\sum_{t=k+1}^n z_t^2} \sqrt{\sum_{t=k+1}^n z_{t-k}^2} \leq D.$$

したがって、本文の定義では $-1 \leq r_k \leq 1$ です。長いラグでは分子の項が減るため、完全な周期データでもピークが 1 になるとは限りません。

普通のピアソン相関とは、計算が少し違う

2 つの短い列 $(y_{k+1}, \dots, y_n)$ と $(y_1, \dots, y_{n-k})$ のピアソン相関では、それぞれの列の平均と偏差平方和を使います。本文の ACF は全系列の平均と全系列の偏差平方和を使うため、両者は一般に異なります。

例：交互に上下するデータ

$y = (1, -1, 1, -1)$ では $\bar{y} = 0, D = 4$ なので、

$$r_1 = \frac{-1 - 1 - 1}{4} = -\frac{3}{4}, \quad r_2 = \frac{1 + 1}{4} = \frac{1}{2}.$$

ラグ 1 で逆向き、ラグ 2 で同じ向きになります。しかし r_2 は 1 ではありません。一方、ラグ 1 の 2 つの短い列のピアソン相関は -1 です。

標本 ACF と理論 ACF を区別する

弱定常な確率過程では、理論 ACF を $\rho_k = \text{Cov}(Y_t, Y_{t-k}) / \text{Var}(Y_t)$ と定めます。 r_k は有限の観測から計算する量であり、標本の偶然によって ρ_k からずれます。

3 時系列とコレログラムを対にして読む

下図はすべて人工データで、右側は左側の 140 個の値から本文の式で計算した標本 ACF です。ラグ 0 は省略しています。

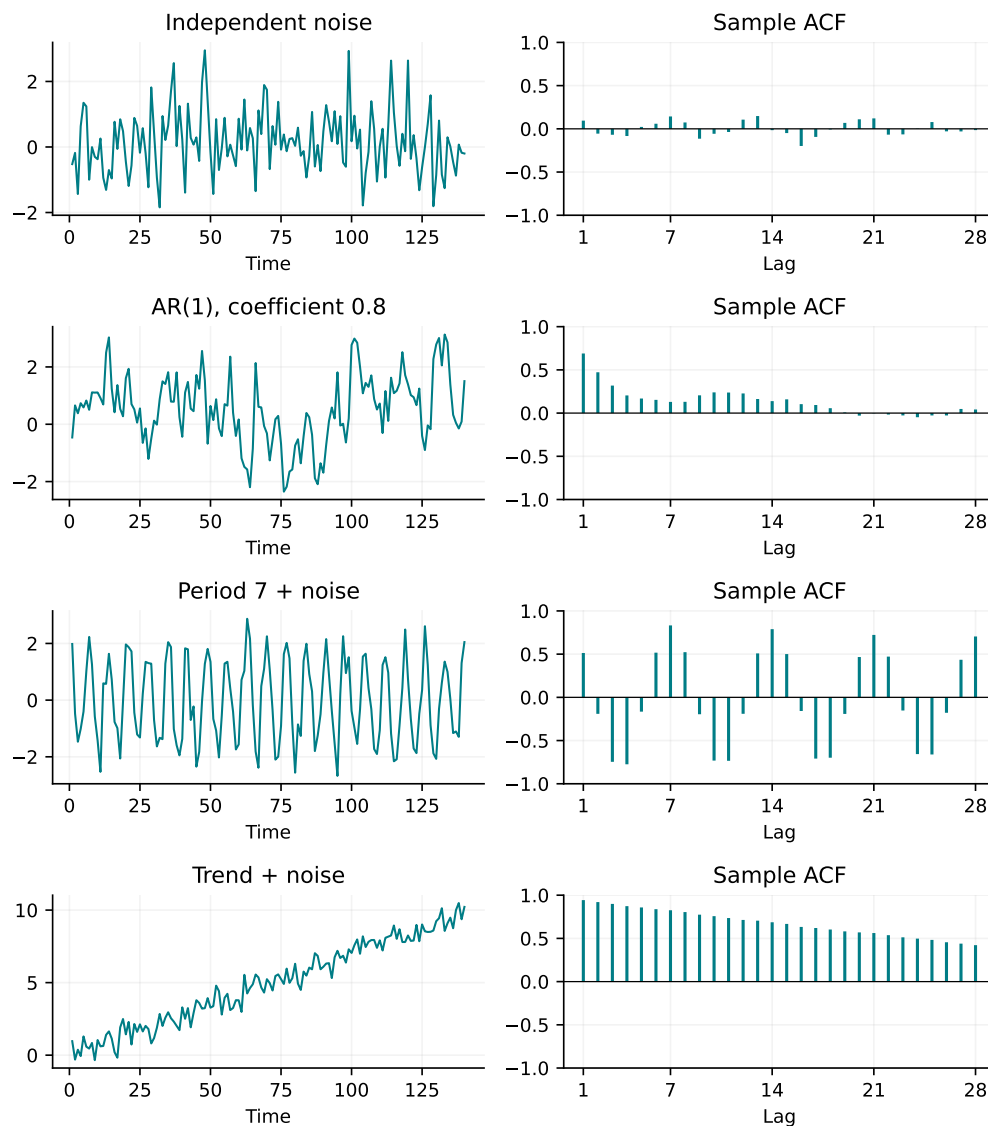


図 1：上から、独立なノイズ、係数 0.8 の自己回帰過程、周期 7 の波とノイズ、上昇トレンドとノイズ。自己回帰の例は初期の 200 時点を除いています。

正の自己相関が大きいだけでは、周期は決まらない

周期性では、特定の間隔でピークが繰り返されることに注目します。トレンドだけでも複数のラグで大きな正の ACF が現れます。左の時系列グラフと併せて読むことが大切です。

4 白色雑音と「有意性の目安」

本文の白色雑音は、平均 0、一定の有限な分散、異なる時点間の共分散 0 を持つ過程です。分散が正なら理論 ACF は、

$$\rho_0 = 1, \quad \rho_k = 0 \quad (k \geq 1)$$

です。それでも、有限標本の r_k がすべて 0 になるわけではありません。

無相関は、独立より弱い条件

独立で平均 0・同じ分散の系列は白色雑音ですが、本文の定義だけでは独立性までは要求しません。したがって、線形相関がないことから、あらゆる非線形な予測も不可能とはいえません。

発展例：値は無相関でも、二乗は関連する

Z_t が独立な標準正規変数で、 $\varepsilon_t = Z_t Z_{t-1}$ とします。平均は 0、分散は 1 です。隣同士の積の期待値も

$$E[\varepsilon_t \varepsilon_{t-1}] = E[Z_t]E[Z_{t-1}^2]E[Z_{t-2}] = 0$$

で、離れた時点も無相関です。しかし $E[Z_t^4] = 3$ より

$$\text{Cov}(\varepsilon_t^2, \varepsilon_{t-1}^2) = 3 - 1 = 2.$$

白色雑音の定義を満たしても、独立ではない例です。

破線 $\pm 2/\sqrt{n}$ の読み方

独立同分布のノイズなど適切な条件の下で、標本数が十分大きくラグを固定すれば、標本 ACF の標準偏差を約 $1/\sqrt{n}$ とみなせます。 $\pm 2/\sqrt{n}$ は各ラグごとの、およそ 95% の参考範囲です。

$$n = 100 : \quad \pm 0.20, \quad n = 400 : \quad \pm 0.10.$$

すべてのラグを一度に保証する帯ではない

多数のラグを見れば、無相関でも偶然範囲を超える棒が出やすくなります。例えば各ラグの超過確率を約 5% とすると、20 本中の超過本数の期待値は約 1 本です。これは必ず 1 本出るという意味ではありません。

本文の無相関という白色雑音の定義だけでは、この帯の近似は一般に保証されません。学習後の残差や大きなラグにも機械的に適用せず、参考線として扱います。

5 定常性と差分：何が変わり、何が残るか

弱定常性では、平均が一定で、共分散が時刻そのものではなく時間差だけに依存します。つまり、

$$E[Y_t] = \mu, \quad \text{Cov}(Y_t, Y_{t-k}) = \gamma_k$$

です。分散は γ_0 で一定です。計算できることと、1 つの ACF で全期間を代表できることは別です。

直線トレンドに差分を取る

$Y_t = a + bt + \varepsilon_t$ なら、

$$\Delta Y_t = b + \varepsilon_t - \varepsilon_{t-1}.$$

時間とともに増える $a + bt$ は定数 b になります。ただし、ノイズの構造も変わります。独立なノイズの分散が $\sigma^2 > 0$ なら、

$$V[\Delta Y_t] = 2\sigma^2, \quad \text{Cov}(\Delta Y_t, \Delta Y_{t-1}) = -\sigma^2, \quad \boxed{\rho_1 = -\frac{1}{2}}.$$

共有する ε_{t-1} の係数が一方では -1 、他方では $+1$ だからです。ラグ 2 以降の理論 ACF は 0 です。

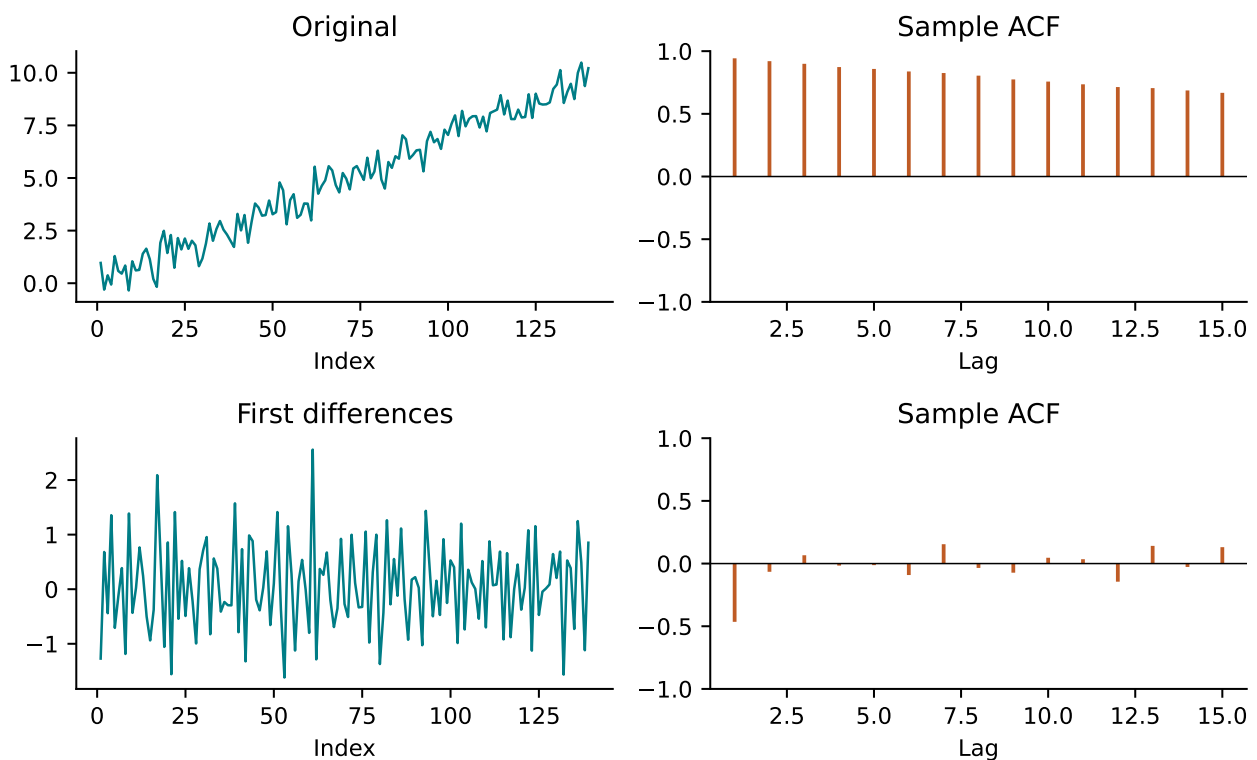


図 2：図 1 最下段の系列と、その 1 階差分。差分で上昇傾向は弱まりますが、負のラグ 1 自己相関が現れます。差分後の左図の横軸は差分系列内の番号です。

差分を取れば必ず定常になるわけではない

分散の変化や構造の急変は、差分だけでは解消されないことがあります。また、既に定常な系列への不要な差分は、新しい相関を生みます。変換後のグラフと ACF を確認します。

6 本文・問 1 を、偏差の積から計算する

データ $y = (1, 3, 2, 4, 3)$ の平均は

$$\bar{y} = \frac{13}{5} = 2.6.$$

t	1	2	3	4	5
y_t	1	3	2	4	3
$z_t = y_t - \bar{y}$	-1.6	0.4	-0.6	1.4	0.4
z_t^2	2.56	0.16	0.36	1.96	0.16

分母はすべてのラグで共通で、

$$D = 2.56 + 0.16 + 0.36 + 1.96 + 0.16 = 5.20.$$

ラグ 1：隣り合う偏差を掛ける

$$\begin{aligned} \sum_{t=2}^5 z_t z_{t-1} &= (0.4)(-1.6) + (-0.6)(0.4) \\ &\quad + (1.4)(-0.6) + (0.4)(1.4) \\ &= -0.64 - 0.24 - 0.84 + 0.56 = -1.16. \end{aligned}$$

よって、

$$r_1 = \frac{-1.16}{5.20} = -\frac{29}{130} \approx -0.2231.$$

ラグ 2：1 つ飛ばした偏差を掛ける

$$\begin{aligned} \sum_{t=3}^5 z_t z_{t-2} &= (-0.6)(-1.6) + (1.4)(0.4) + (0.4)(-0.6) \\ &= 0.96 + 0.56 - 0.24 = 1.28. \end{aligned}$$

したがって、

$$r_2 = \frac{1.28}{5.20} = \frac{16}{65} \approx 0.2462.$$

この 5 個のデータから言える範囲

ラグ 1 では逆方向、ラグ 2 では同方向の偏差の積が優勢です。ただし、組数はそれぞれ 4 組と 3 組しかありません。安定した周期 2 の構造が存在すると断定できるほどの情報ではありません。

7 本文・問 2、問 3 と追加問題

本文・問 2：周期性と有意性の目安

日次データでラグ 7、14、21 にピークがあれば、7 日周期の曜日効果などが疑われます。 $n = 100$ なので参考範囲は

$$\pm \frac{2}{\sqrt{100}} = \pm 0.20.$$

$r_7 = 0.35$ は上側の目安を超えます。複数の 7 の倍数にピークがあることも合わせると、週周期の構造を検討する理由になります。ただし、周期性の確定や原因の特定はできません。

本文・問 3：1 階差分

100, 108, 96, 120 の差分は

$$8, -12, 24.$$

水準そのものから、1 時点ごとの増減量へと見方が変わります。直線的な上昇なら差分は一定に近くなるため、平均の時間変化を弱められる場合があります。この 4 個の観測だけで定常性を判定することはできません。

追加問 A：符号と周期

$y = (2, -2, 2, -2, 2, -2)$ について、本文の定義で r_1, r_2 を求めよ。なぜ r_2 が 1 にならないのか説明せよ。

追加問 B：単位の変更

$w_t = ay_t + b$ ($a \neq 0$) とする。 w_t の標本 ACF が y_t の標本 ACF と同じになることを示せ。ただし元の系列は定数ではないとする。

追加問 C：参考範囲

$n = 400$ のときの目安を求めよ。 $r_3 = 0.14$ だけが範囲を超えたとき、「周期 3 が証明された」と言えるか。

追加問 D：差分で相関が生まれる

独立同分布の ε_t は平均 0、分散 4 とする。 $X_t = \varepsilon_t - \varepsilon_{t-1}$ の分散とラグ 1 の共分散、理論 ACF を求めよ。

追加問 E：平滑化後の自己相関【発展】

同じノイズに対して $M_t = (\varepsilon_t + \varepsilon_{t-1} + \varepsilon_{t-2})/3$ とする。 M_t の理論 ACF ρ_1, ρ_2, ρ_3 を求めよ。

8 追加問題の解答と、読み違いの確認

問 A：周期が完全でも、標本 ACF は 1 とは限らない

平均 0、分母 $D = 6 \cdot 4 = 24$ 。ラグ 1 は 5 個の積がすべて -4 、ラグ 2 は 4 個の積がすべて 4 なので、

$$r_1 = -\frac{20}{24} = -\frac{5}{6}, \quad r_2 = \frac{16}{24} = \frac{2}{3}.$$

分子に使う組数が減る一方で、分母は全 6 個の偏差平方和のままだからです。

問 B：平行移動と一定倍率に不変

$\bar{w} = a\bar{y} + b$ なので $w_t - \bar{w} = a(y_t - \bar{y})$ 。分子・分母がともに a^2 倍され、比は変わりません。単位を円から千円に変えても、系列全体の符号を反転しても ACF は同じです。

問 C：一本のピークで周期を確定しない

目安は ± 0.10 です。0.14 は超えていますが、多数のラグを調べた場合の偶然の超過や、周期以外の構造も考えられます。元のグラフ、他のラグ、データ数を合わせて読みます。

問 D：隣接差分が共有する項に注目

$V[X_t] = 4 + 4 = 8$ 。共分散に残るのは $(-\varepsilon_{t-1})(\varepsilon_{t-1})$ の期待値だけなので、

$$\text{Cov}(X_t, X_{t-1}) = -4, \quad \rho_1 = -\frac{1}{2}.$$

ラグ 2 以降は共通するノイズがなく、理論 ACF は 0 です。

問 E：共有する観測が正の相関を作る

分散は $3 \cdot 4/9 = 4/3$ 。ラグ 1 では 2 項、ラグ 2 では 1 項を共有するので、共分散はそれぞれ $8/9, 4/9$ 。したがって、

$$\rho_1 = \frac{2}{3}, \quad \rho_2 = \frac{1}{3}, \quad \rho_3 = 0.$$

元のデータが独立でも、重なりのある移動平均を取ると自己相関が生まれます。平滑化後の相関を、元の現象の記憶性とそのまま解釈しないことが大切です。

コレログラムを読む順序

まず時間順のグラフでトレンド・季節性・急変を確認し、次に ACF の符号・減衰・ピーク間隔を読みます。参考範囲は最後の補助判断とし、差分や平滑化が相関に与えた影響も考えます。