時系列解析入門 – モデリング

1. 確率分布と統計的モデル

Y が確率変数（random variable）のとき、すべての実数 $y \in R$ に対して、$Y \le y$ となる確率 $\mathrm{Prob}(Y \le y)$ が定められる。これを y の関数とみなして、

$$G(y) = \mathrm{Prob}(Y \le y)$$

とあらわすとき、$G(y)$ を確率変数 Y の分布関数（probability distribution function）と呼ぶ。

時系列解析で用いられる確率変数は通常連続型と呼ばれるもので、その分布関数は $g(t) \ge 0 (-\infty < t < \infty)$ を満たす関数の積分によって、

$$G(y) = \int_{-\infty}^{y} g(t)dt$$

と表現できる。このとき、$g(x)$ を密度関数（density, probability density function）と呼ぶ。逆に分布関数が与えられると、任意の $a < b$ に対して $a < Y \le b$ となる確率が、

$$G(b) - G(a) = \int_{a}^{b} g(x)dx$$

によって求められる。

代表的な密度関数として以下がある。

正規分布（ガウス分布、normal distribution）

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right), -\infty < x < \infty$$

平均 μ、分散 σ^2 で $N(\mu, \sigma^2)$ と表記され、$N(0,1)$ は標準正規分布と呼ばれる。

コーシー分布（Cauchy distribution）

$$g(x) = \frac{\tau}{\pi\left\{(x-\mu)^2 + \tau^2\right\}}, -\infty < x < \infty$$

ピアソン分布族（Pearson family of distribution）

$$g(x) = \frac{c}{\pi\left\{(x-\mu)^2 + \tau^2\right\}^b}, -\infty < x < \infty$$

ただし、$c = \tau^{2b-1}\Gamma(b) \Big/ \left(\Gamma\left(b - \frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right) \right)$ で、$b = 1$ のときコーシー分布と一致する。また k を正の整数として $b = (k+1)/2$ としたとき自由度 k の

t 分布（t-distribution）と呼ばれる。

指数分布（exponential distribution）

$$g(x)=\begin{cases}\lambda e^{-\lambda x}, x \geq 0 \\ 0, x<0\end{cases}$$

χ^2 分布（chi-square distribution）

$$g(x)=\begin{cases}\dfrac{1}{2^{k/2}\Gamma\left(\dfrac{k}{2}\right)} e^{-\frac{x}{2}} x^{\frac{k}{2}-1}, x \geq 0 \\ 0, x<0\end{cases}$$

k は自由度と呼ばれる。$k=2$ のとき指数分布となる。

2 重指数分布（double exponential disbtiution）

$$g(x)=e^{x-e^x}$$

一様分布（uniform distribution）

$$g(x)=\begin{cases}(b-a)^{-1}, a<x \leq b \\ 0, other\end{cases}$$

ある密度関数から得られるデータを、確率変数の実現値（realization）という。反対に、観測するデータの背後に確率変数を想定し、データはその確率変数の実現値として得られたものと考えるとき、この確率変数を特徴付ける密度関数 $g(y)$ を真のモデル（true model）と呼ぶ。

通常この真のモデルは未知であるから、与えられたデータから確率分布を推定する必要がある。このとき、データから推定された密度関数は統計的モデル（statistical model）と呼ばれ $f(y)$ と表される。

時系列データの場合はさらに、同時分布 $f(y_1,...,y_N)$ を考える必要がある。時系列 $y_1,...,y_N$ を標本平均 $\hat{\mu}$ と標本自己共分散関数 $\hat{C}_k$ によって表現するということは、N 次元ベクトル $y=(y_1,...,y_N)^T$ が平均ベクトル $\hat{\mu}=(\hat{\mu},...,\hat{\mu})^T$、分散共分散行列

$$\hat{C}=\begin{bmatrix}\hat{C}_0 & \hat{C}_1 & \cdots & \hat{C}_{N-1}\\ \hat{C}_1 & \hat{C}_0 & \cdots & \hat{C}_{N-2}\\ \cdots & \cdots & \cdots & \cdots\\ \hat{C}_{N-1} & \hat{C}_{N-2} & \cdots & \hat{C}_0\end{bmatrix}$$

の多次元正規分布に従うとするモデルを想定していることに相当する。このようなモデルは正規分布に従う定常時系列を柔軟に表現できるが、データ数Nに対して$N+1$個の未知数$\hat{\mu}$、$\hat{C}_0,...,\hat{C}_{N-1}$を推定することになり、データの情報を効率よく縮約することにはならない。

2. KL 情報量とエントロピー最大化原理

現実のデータを生成する真のモデルを$g(y)$、それを近似した統計的モデルを$f(y)$と表すことにする。統計モデリングでは$g(y)$になるべく「近い」$f(y)$を求めることが主要な目的になる。そのためにはモデル$f(y)$のよさを客観的に評価する基準が必要になる。ここではその基準としてカルバック・ライブラー情報量（Kullback-Leibler information、以下 KL 情報量）を用いる。（2 番目の等式は、モデルが連続型の確率分布の場合）

$$I(g;f)=\mathrm{E}_Y\log\left\{\frac{g(Y)}{f(Y)}\right\}=\int_{-\infty}^{\infty}\log\left\{\frac{g(y)}{f(y)}\right\}g(y)dy$$

この KL 情報量は、以下の性質を持っている。

$$I(g;f)\geq 0$$

$$I(g;f)=0\Leftrightarrow g(y)=f(y)$$

また、KL 情報量の符号を反転した量$B(g;f)=-I(g;f)$は一般化されたエントロピー（entropy）とも呼ばれ、想定した分布$f(y)$からn個の実現値をとった時に、その相対度数分布が真の分布$g(y)$と等しくなる確率の$\frac{1}{n}$を近似的に与える。従って、KL 情報量が小さいほど確率分布fはgに近いと考えることができる。

統計モデルはデータ$y_1,...,y_N$に基づいて真の分布$g(y)$を近似したもので、そのよさは KL 情報量$I(g;f)$で評価できる。統計的モデリングにおいて$B(g;f)=-I(g;f)$を最大とするようにモデルを構築しようとするのがエントロピー最大化原理（entropy maximization principle）である。

例えば、真のモデル$g(y)$およびそれを近似したモデル$f(y)$がともに正規分布である場合を考える。

$$g\left(y\middle|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\}$$

$$f\left(y\middle|\zeta,\tau^2\right)=\frac{1}{\sqrt{2\pi\tau^2}}\exp\left\{-\frac{(y-\zeta)^2}{2\tau^2}\right\}$$

この場合、

$$\log\left\{\frac{g(y)}{f(y)}\right\}=\frac{1}{2}\left\{\log\frac{\tau^2}{\sigma^2}-\frac{(y-\mu)^2}{\sigma^2}+\frac{(y-\zeta)^2}{\tau^2}\right\}$$

となるので、KL 情報量は、

$$I(g;f)=\mathrm{E}_\mathrm{Y}\log\frac{g(Y)}{f(Y)}$$
$$=\frac{1}{2}\left\{\log\frac{\tau^2}{\sigma^2}-\frac{\mathrm{E}_\mathrm{Y}(Y-\mu)^2}{\sigma^2}+\frac{\mathrm{E}_\mathrm{Y}(Y-\zeta)^2}{\tau^2}\right\}$$
$$=\frac{1}{2}\left\{\log\frac{\tau^2}{\sigma^2}-1+\frac{\sigma^2+(\mu-\zeta)^2}{\tau^2}\right\}$$

で与えられる。g と f が正規分布の場合の KL 情報量の計算は簡単だが、そうでない場合は数値計算によって求められる。例えば以下の台形公式などが用いられる。

$$\hat{I}(g;f)=\frac{\Delta x}{2}\sum_{i=1}^{k}\{h(x_i)+h(x_{i-1})\}$$

ただし、

$$x_0=-x_k$$
$$x_i=x_0+(x_k-x_0)\frac{i}{k}$$
$$h(x)=g(x)\log\frac{g(x)}{f(x)}$$

3.　KL 情報量の推定と対数尤度

実際の統計解析の場面では真の分布は未知であるため、KL 情報量は実際の統計モデルの評価に用いられることはほとんどない。真の分布 $g(y)$ の代わりに、$g(y)$ から独立に観測されたデータ $y_1,...,y_N$ が与えられている場合、モデル $f(y)$ の KL 情報量を以下の方法で推定する。

エントロピー最大化原理に従って最も良いモデルを求めるためには、$B(g;f)=-I(g;f)$ を

最大、$I(g;f)$を最小とするモデルを求めればよい。KL 情報量は、

$$I(g;f)=\mathrm{E}_{\mathrm{Y}}\log g(Y)-\mathrm{E}_{\mathrm{Y}}\log f(Y)$$

と二つの項に分解できる。右辺第 1 項は$g(y)$が与えられないと計算できないが、モデル$f(y)$には依存しない一定の値を取るので無視できる。右辺第 2 項は平均対数尤度（expected log-likelihood）と呼ばれる量で、密度関数を持つ連続型のモデルの場合は、

$$\mathrm{E}_{\mathrm{Y}}\log f(Y)=\int \log f(y)g(y)dy$$

と表現できる。この平均対数尤度も$g(y)$が未知の場合には直接計算できないが、データy_nが密度関数$g(y)$に従って生成されることから、大数の法則によりデータ数が$N\to\infty$の時、

$$\frac{1}{N}\sum_{\mathrm{n=1}}^{N}\log f(y_n)\to\mathrm{E}_{\mathrm{Y}}\log f(Y)$$

が成り立つ。従って KL 情報量$I(g;f)$を最小とするモデルの代わりに、対数尤度（log-likelihood）$l=\sum_{\mathrm{n=1}}^{N}\log f(y_n)$を最大とするようなモデルを選べば、近似的にエントロピーを最大にすることができる。また、その指数をとった、$L=\prod_{\mathrm{n=1}}^{\mathrm{N}} f(y_n)$は尤度（likelihood）と呼ばれる。

時系列解析のモデルでは、通常観測値が独立に得られるという仮定は成り立たない。このような一般の場合には、尤度は$y_1,...,y_N$の同時分布を用いて、

$$L=f(y_1,...,y_N)$$

と定義される。この場合対数尤度は、

$$l=\log L=\log f(y_1,...,y_N)$$

となる。

4. 最尤法によるパラメータの推定

モデルがθをパラメータとするパラメトリックモデルで$f(y)=f(y|\theta)$の形をしている場合には、対数尤度lはパラメータθの関数と考えることができる。したがって、θを明示的に表し、

$$l(\theta)=\begin{Bmatrix}\sum_{n=1}^{N}\log f(y_n|\theta)(\text{独立の場合})\\ \log f(y_1,...,y_N|\theta)(\text{一般の場合})\end{Bmatrix}$$

をθの対数尤度関数と呼ぶ。

対数尤度関数$l(\theta)$はθで定まるモデルの良さを評価した量なので、$l(\theta)$を最大とするθを選

ぶことによって、パラメトリックモデル $f(y|\theta)$ のパラメータの最適な値を定めることができる。このように対数尤度あるいは尤度を最大化することによりパラメータを推定する方法は最尤法（maximization likelihood method）と呼ばれる。また最尤法で推定されたパラメータを $\hat{\theta}$ と表し、最尤推定値（maximum likelihood estimate）と呼ぶ。
例として、平均 μ 分散 1 の正規分布モデル、

$$f(y|\mu)=\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{(y-\mu)^2}{2}\right\}$$

のパラメータ μ を最尤法で推定する。この場合対数尤度関数は、

$$l(\mu)=-\frac{N}{2}\log 2\pi-\frac{1}{2}\sum_{n=1}^{N}(y_n-\mu)^2$$

したがって $l(\mu)$ を最大とするためには、

$$S(\mu)=\sum_{n=1}^{N}(y_n-\mu)^2$$

を最小とする μ を求めればよいので $S(\mu)$ の一階導関数を 0 とおくことによって、

$$\hat{\mu}=\frac{1}{N}\sum_{n=1}^{N}y_n$$

が得られる。 $S(\mu)=\sum_{n=1}^{N}(y_n-\mu)^2$ のように二乗和を最小にすることによりパラメータを推定する方法は、最小二乗法（least squares method）と呼ばれる。
一般に時系列モデルのパラメータ θ の最尤推定値を求めるためには、擬似ニュートン法による数値的最適化（numerical optimization）が用いられる。パラメータ θ の初期値 θ_0 の値を定めたときの対数尤度の値 $l(\theta)$ と一階微分 $\frac{\partial l}{\partial \theta}$ が与えられると、

$$\theta_k=\theta_{k-1}+\lambda_k H_{k-1}\frac{\partial l}{\partial \theta}$$

を繰り返して $l(\theta)$ の極大点を自動的に求める。ステップ幅 λ_k とヘッセ行列の逆行列 H_{k-1} は自動的に決定される。

5. AIC（赤池情報量基準）

最大対数尤度はそのままでは異なるモデル間の比較には用いることができない。最尤推定値 $\hat{\theta}$ で規定されるモデルは、$N^{-1}l(\hat{\theta})$ が $\mathrm{E}_Y \log f(Y|\hat{\theta})$ の推定量として正の偏りを持つためである。この偏差は、パラメータの推定とモデルの評価のための平均対数尤度の推定に同じ

データを 2 度用いたことによって生じる。

$\mathrm{E_Y} \log f\left(Y\middle|\hat{\theta}\right)$を$N^{-1}l\left(\hat{\theta}\right) \equiv N^{-1}\sum_{n=1}^{N} \log f\left(y_n\middle|\hat{\theta}\right)$によって推定したときに生じる平均的な偏りを、

$$C \equiv \mathrm{E_x}\left\{\mathrm{E_Y} \log f\left(Y\middle|\hat{\theta}\right) - N^{-1}\sum_{n=1}^{N} \log f\left(y_n\middle|\hat{\theta}\right)\right\}$$

とおく。このとき、$N^{-1}l\left(\hat{\theta}\right)$を$C$だけ補正し、$N^{-1}l\left(\hat{\theta}\right)+C$とすることにより$\mathrm{E_Y} \log f\left(Y\middle|\hat{\theta}\right)$の偏りのない推定量を求めることができる。ここで$C=-N^{-1}k$となることから、赤池情報量基準（AIC: Akaike Information Criterion）が得られる。

$$AIC = -2l\left(\hat{\theta}\right)+2k$$
$$= -2(\text{最大対数尤度}) + 2(\text{パラメータ数})$$

6.　データ変換

正規分布しなかったり分散が一定でない時系列のデータも、対数変換すれば変動が小さくなったり正規分布に近づいたりする。
対数変換を含む一般的のデータ変換として Box-Cox 変換

$$z_n = \begin{Bmatrix} \lambda^{-1}\left(y_n^{\lambda}-1\right), \lambda \neq 0 \\ \log y_n, \lambda = 0 \end{Bmatrix}$$

がある。Box-Cox 変換は定数を無視すると、$\lambda=0$のとき対数、$\lambda=-1$のとき逆数、$\lambda=0.5$のとき平方根、$\lambda=1$のとき原データをとる変換となる。
AIC を用いると、データに適した変換を定めるパラメータλを選択することができる。
Box-Cox 変換によって変換されたデータ$z_n=h\left(y_n\right)$が密度関数$f\left(z\right)$に従う場合、元データy_nの密度関数は、

$$g(y) = \left|\frac{dh}{dy}\right| f(h(y))$$

ただし、$\left|dh/dy\right|$は変換のヤコビアン（Jacobian）と呼ばれる。これは、変換したデータのモデルが変換前のデータに関してもひとつのモデルを定めていることを示している。
例えば、原データy_nおよび変換されたデータz_nに正規分布を当てはめたときの AIC の値をそれぞれAIC_y、AIC_zとする。このとき、

$$\mathrm{AIC}_z' = \mathrm{AIC}_z - 2\sum_{i=1}^{N} \log \left| \frac{dh}{dy} \right|_{y=y^i}$$

の値を AIC_y と比較することにより。原データと変換後データのどちらかが正規分布に近いかを判断することができる。すなわち $\mathrm{AIC}_y < \mathrm{AIC}_z'$ の場合は原データのほうがよいことになる。一方 $\mathrm{AIC}_y > \mathrm{AIC}_z'$ の場合は変換した方がよいことがわかる。さらに AIC_z' が最小になるようにすることによって、Box-Cox 変換の最適な λ の値を選択することもできる。実際の時系列データでは Box-Cox 変換を行った後、色々な時系列モデルを当てはめることが多いので、その場合には時系列モデルの AIC を Box-Cox 変換のヤコビアンを使って補正する必要がある。