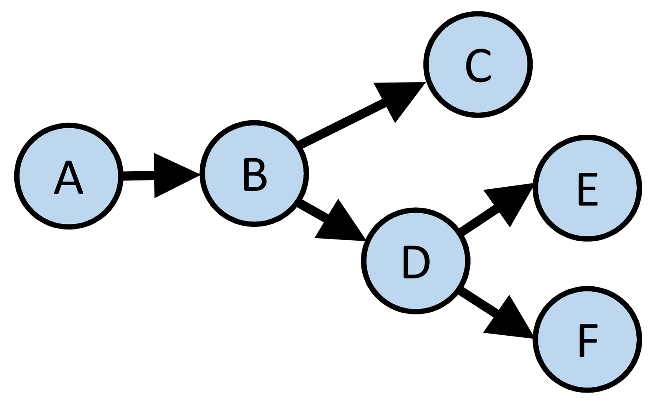
実数値と角度の観測が混在するデータのためのベイジアンネットワーク

加藤　昇吾　数理・推論研究系　准教授

はじめに

ベイジアンネットワーク

ベイジアンネットワーク $\cdots$ グラフィカルモデルの１つ．有向非巡回グラフを通して変数間の依存構造を表す．



図．ベイジアンネットワーク．

それぞれの変数は，実数値や離散値など様々な種類の観測がありうる．

研究の目的

変数として，角度 $[-\pi,\pi)$ の観測が得られることがある．
（例．風向，鳥の移動方向，銃犯罪が起こる時刻，など.）

本研究では，実数値と角度の観測が混在するデータのための木構造のベイジアンネットワークを提案する．

なお本研究は，Ignacio Leguey 氏，Concha Bielza 教授，Pedro Larrañaga 教授 (Technical University of Madrid, Spain) との共同研究である．

ベイジアンネットワーク

モデル

$X_1,\ldots,X_p$ を実数値の確率変数，$\Theta_1,\ldots,\Theta_q$ を角度 $[-\pi,\pi)$ の確率変数，

とする．親ノードと子ノードの同時分布に以下を仮定する：

$(X_j,X_k)\sim$ 2変量正規分布 $N(\mu_j,\mu_k,\sigma_j^2,\sigma_k^2,\rho_{jk})$,

$(\Theta_j,\Theta_k)\sim$ Kato & Pewsey (2015) の分布,

$(X_j,\Theta_k)\sim$ Johnson & Wehrly (1978) の分布の特別な場合.

ただし，Kato & Pewsey (2015) の分布は密度関数

$$f(\theta_j,\theta_k) = C\Big\{c_0 + c_1\cos(\theta_j-\eta_j) + c_2\cos(\theta_k-\eta_k) + c_3\cos(\theta_j-\eta_j)\cos(\theta_k-\eta_k) + c_4\sin(\theta_j-\eta_j)\sin(\theta_k-\eta_k)\Big\}^{-1},$$
$$-\pi\le\theta_j,\theta_k,\eta_j,\eta_k<\pi,\quad 0\le\xi_j,\xi_k<1,\quad -1<\gamma_{jk}<1,$$

で与えられる．$C,c_0,\ldots,c_4$ は，ξ_j,ξ_k,γ_{jk} の関数として陽に表される．

また，Johnson & Wehrly (1978) の分布の特別な場合は，

$$f(x_j,\theta_k) = 2\pi f_j(x_j)g_k(\theta_k)h[2\pi\{F_j(x_j)-G_k(\theta_k)\}],\quad x_j\in\mathbb{R},\ -\pi\le\theta_k<\pi,$$

である．ここで，f_j,F_j はそれぞれ正規分布 $N(\mu_j,\sigma_j^2)$ の密度関数と分布関数，g_k,G_k はそれぞれ Wrapped Cauchy 分布 $WC(\eta_k,\xi_k)$ の密度関数と分布関数，h は $WC(0,\delta_{jk})$ の密度関数，$WC(\eta,\xi)$ の密度関数は，

$$f(\phi)=\frac{1}{2\pi}\frac{1-\xi^2}{1+\xi^2-2\xi\cos(\phi-\eta)},\quad -\pi\le\phi,\eta<\pi,\ 0\le\xi<1.$$

相互情報量

確率変数 $Z_j(Z_k)$ が $\mathcal{M}_j(\mathcal{M}_k)$ に値をとるとする．このとき，2つの確率変数 (Z_j,Z_k) 間の相互情報量を以下で定義する．

$$\mathrm{MI}(Z_j,Z_k)=\int_{\mathcal{M}_j}\int_{\mathcal{M}_k} f(z_j,z_k)\log\left\{\frac{f(z_j,z_k)}{f_j(z_j)f_k(z_k)}\right\}dz_k dz_j.$$

ただし，$f(z_j,z_k)$ は (Z_j,Z_k) の同時密度関数，$f_j(f_k)$ は $Z_j(Z_k)$ の周辺密度関数をあらわす．

このとき，提案したモデルの相互情報量は以下で与えられる：

$$\mathrm{MI}(X_j,X_k)=-\frac{1}{2}\log(1-\rho_{jk}^2),\quad \mathrm{MI}(\Theta_j,\Theta_k)=-\log(1-\gamma_{jk}^2),$$
$$\mathrm{MI}(X_j,\Theta_k)=-\log(1-\delta_{jk}^2).$$

周辺分布と条件付分布

$$X_j\sim N(\mu_j,\sigma_j^2),\quad \Theta_j\sim WC(\eta_j,\xi_j),$$
$$X_j|X_k=x_k\sim N\left(\mu_j+\sigma_j\sigma_k^{-1}\rho_{jk}(x_k-\mu_k),(1-\rho_{jk}^2)\sigma_j^2\right),$$
$$2\pi F_j(X_j)|\Theta_k=\theta_k\sim WC(2\pi G_k(\theta_k),\delta_{jk}),$$
$$\Theta_j|\Theta_k=\theta_k\sim WC\left(\eta(\theta_k),\xi(\theta_k)\right),\quad \Theta_k|X_j=x_j\sim WC\left(\eta'(x_j),\xi'(x_j)\right).$$

ここで，$\eta(\theta_k),\xi(\theta_k)$ は θ_k の関数として，$\eta'(x_j),\xi'(x_j)$ は x_j の関数として，それぞれ陽に表すことができる．

モデルの推定

よく知られているように，2変量正規分布のパラメータの最尤推定量は，陽な形で表現できる．

Kato & Pewsey (2015) の分布と Johnson & Wehrly (1978) の分布のパラメータについては，最尤推定値は数値的に求めることになるが，モーメント推定量は陽な形で表すことができる．

これらの結果を利用し，提案したベイジアンネットワークを下記の Chow Liu アルゴリズムにより推定する．

アルゴリズム

1. 確率変数 $X_1,\ldots,X_p,\Theta_1,\ldots,\Theta_q$ に対して，すべての周辺分布のパラメータと，すべての２変量同時分布の相関パラメータを推定する．
2. 推定されたすべての２変量同時分布に対して，相互情報量をエッジの重みとして計算する．
3. 最も大きい重みのエッジを木に含める．
4. 次に大きい重みを持つエッジを，ループが起こらない場合は木に含める．
5. ステップ 4 を $p+q-1$ のエッジが選ばれるまで繰り返す．
6. 根ノードを選択し，それをもとに有向グラフを構成する．

References

[1] Johnson, R.A. & Wehrly, T.E. (1978). Some angular-linear distributions and related regression models. *Journal of the American Statistical Association* **73**(363), 602–606.

[2] Kato, S. & Pewsey, A. (2015). A Möbius transformation-induced distribution on the torus. *Biometrika* **102**(2), 359–370.