統計モデル可視化

石黒真木夫
統計思考院@統計数理研究所
2016.5.14

1. 条件付き確率関数・密度関数の可視化

目的変数 y の挙動を説明変数 x との関係で説明する統計モデルを見つけるということは、yを測定したデータの分布に次のような形の表現を与えることであると考えることができる。

[説明変数 x と目的変数 y がともに離散変数の場合]

$$P_Y(y) = \sum_{x=0}^{C} P_{Y|X}(y|x)P_X(x)$$

[説明変数 x が離散、目的変数 y が連続変数の場合]

$$f_Y(y) = \sum_{x=0}^{C} f_{Y|X}(y|x)P_X(x)$$

[説明変数 x が連続、目的変数 y が離散変数の場合]

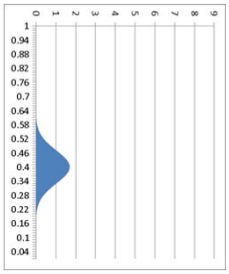
$$P_Y(y) = \int_0^1 P_{Y|X}(y|x)f_X(x)dx$$

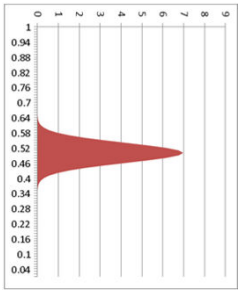
[説明変数 x と 目的変数y がともに連続変数の場合]

$$f_Y(y) = \int_0^1 f_{Y|X}(y|x)f_X(x)dx$$

ここで、P は確率関数、f は確率密度関数を表すものとする。

1.1 説明変数 x が離散、目的変数 y が連続変数の場合、
たとえば正規分布モデルが当てはまれば、、、

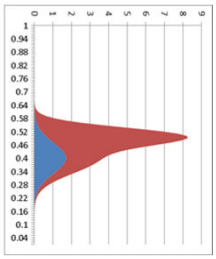
$f_{Y|X}(y|0)P_X(0)$ = 

$f_{Y|X}(y|1)P_X(1)$ = 

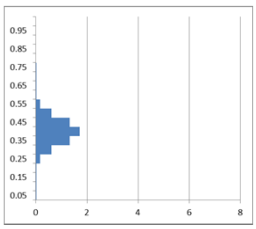
であれば、モデル

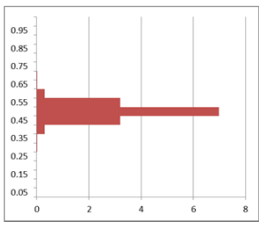
$$\sum_{x=0}^{1} f_{Y|X}(y|x)P_X(x)$$

は右のように「可視化」される。



あるいは、ヒストグラムモデルを使って、

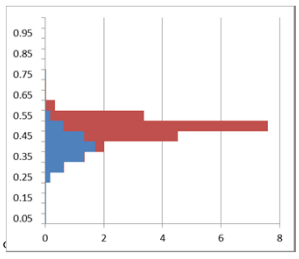
$f_{Y|X}(y|0)P_X(0)$ = 

$f_{Y|X}(y|1)P_X(1)$ = 

であれば、モデル

$$\sum_{x=0}^{1} f_{Y|X}(y|x)P_X(x)$$

は右のように「可視化」される。



説明変数x 説明変数 y がともに離散変数の場合の可視化もこのような表現となる。

1.2 説明変数 x 目的変数 y ともに連続変数の場合

2次元平面上の領域

$$\{(s,y) \mid 0 \le s \le f_Y(y)\}$$

の各点において、関数 $g(s,y)$ を方程式

$$s = \int_0^{g(s,y)} f_{Y|X}(y|x) f_X(x) dx$$

の解として定義すると、$g(s,y)$ は 0 と1 の間の値をとる。

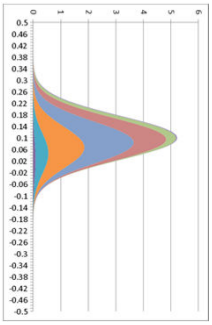
$$0 = g_0 < g_1 < \cdots < g_C = 1$$

として、$g_i < g(s,y) \le g_{i+1}$で定義される領域に色 i をつけることにすると、

モデル $f_Y(y) = \int_0^1 f_{Y|X}(y|x) f_X(x) dx$ が、たとえば

このように →

可視化される。



統計モデルの「地形」を示す $g(s,y)$ という 曲面を「標高」で区分して色付けすることによって「色分け地図」にしているのである。

$g < g(s,y) < g + dg$ で定義される領域の面積は

$$\int f_{Y|X}(y|g) f_X(g) dg dy = f_X(g) dg$$

となる。

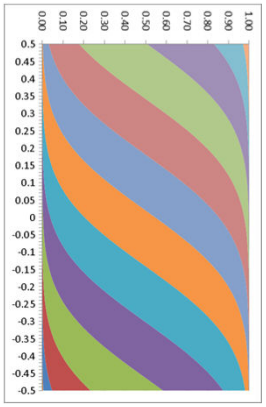
2. ベイズ事後分布関数の色分け地図

$$G(s,y) = g(s f_Y(y), y)$$

を定義すると、G は領域 $\{(s,y) \mid 0 \le s \le 1\}$ で定義される関数となり、

$$0 = g_0 < g_1 < \cdots < g_C = 1$$

として、$g_i < G(s,y) \le g_{i+1}$で定義される領域に色 i をつけると、たとえば、このような図（↓）が得られる。



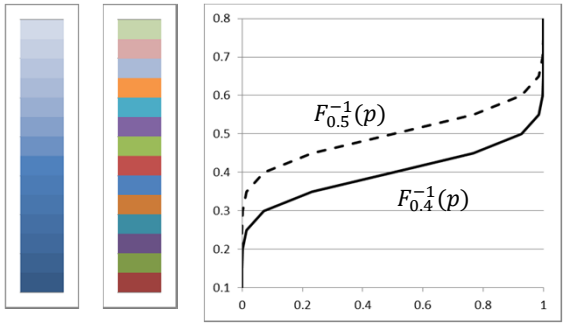
$$s = \frac{s f_Y(y)}{f_Y(y)} = \int_0^{g(s f_Y(y),y)} \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} dx$$

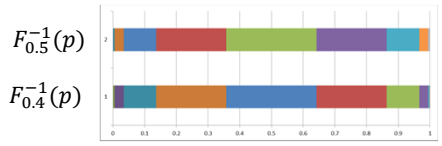
$$= \int_0^{G(s,y)} f_{X|Y}(x|y)\, dx \quad = F_{X|Y}(G(s,y)|y)$$

が成り立つので $G(s,y) = F^{-1}_{X|Y}(s|y)$ となる。これは、左の図が、ベイズ公式で計算される事後分布の分布関数の可視化に他ならないことを意味する。

$f_{Y|X}$ が Xに依存しない場合には、$G(s,y)$ が y に依存せず、左の図は「縦縞模様」になる。この命題の対偶をとって、左図が縦縞でないことから$f_{Y|X}$ が Xに依存すると結論することが出来る。

余談：逆分布関数の色分け地図表示





左図の実線は、平均0.4 標準偏差 0.07 の正規分布の逆分布関数

$$y = F_{0.4}^{-1}(p)$$

のプロット、破線は平均0.5 標準偏差 0.07 の正規分布の逆分布関数$F_{0.5}^{-1}(p)$である。縦軸 y の値を色で表すと、下のような表現が得られる。

左端に示したような単色の濃淡による色付けも可能であるが、曲線の違いを際立たせるには、境界が際立つ色使いが適している。

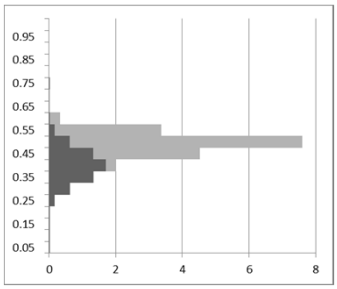
3. 統計モデル可視化ソフト

統計数理研究所機関リポジトリで公開されている「情報量統計学的データ可視化ツール」が簡単な統計モデル可視化機能を備えている。

情報量統計学的データ可視化ツール An AIC-based Tool for Data Visualization http://hdl.handle.net/10787/3614

このソフトは、条件付き分布関数 $f_{Y|X}$ として、ヒストグラムモデルを採用することによって、多様なデータに柔軟に対応できるものとなっており、連続値離散値混在多次元データの各成分相互の間の関係が簡単に網羅的に可視化される 。

なお、このソフトは可視化にあたってグレーの濃淡で「色付け」するので、このソフトで可視化された統計モデルは右の図のような山水画風の表現となる。



3.1 情報量統計学的データ可視化ツール使用例

テストデータ
条件付き確率関数・密度関数の可視化
ベイズ事後分布関数の可視化
情報量規準 AIC の利用
時系列データの場合

3.1.1 テスト用6次元データ

[確率変数、$x_1(nonGauss)$ と$x_2(yesno)$]

$$x_1 = \begin{cases} w_1 & if\ x_2 = 1 \\ -w_1 & if\ x_2 = 2 \end{cases}$$

とする。ここで、w_1は確率密度関数

$$f(w_1) = w_1 + \mu \qquad (-\mu < w_1 < 1 - \mu)$$

を持つ連続値確率変数。ただし、μ は、w_1の期待値が 0 となるように決められた定数。当然のことながら w_1と $-w_1$ の平均はともに 0であり、分散も等しい。分散分析によって、

$x_1 = w_1$ であるのか、
$x_1 = -w_1$であるのか

を弁別することはできない。x_2 は等確率で値1 あるいは 2 をとる2値確率変数とする。この場合、x_1とx_2の間に関連があるが、その関連を相関解析で見だすことはできない。

[確率変数、x_3 (cos) と$x_4(sin)$]

$$x_3 = w_3 \cos w_2$$
$$x_4 = w_3 \sin w_2$$

とする。ただし、w_3 は期待値 1、標準偏差 $\ll 1$ の正規分布、w_2 は 0 ～ 2π の値をとる一様乱数である。x_3とx_4の間に関連があるが、その関連を相関解析で見だすことはできない。

テストデータ (つづき)

[確率変数、x_5 ($regression$) とx_4(sin)]

$$x_5 = x_4 + w_5$$

ただし w_5 は平均 0 の正規分布。

[確率変数、x_6 ($depend$) とx_2($yesno$)]

$x_2 = 1$ の場合、

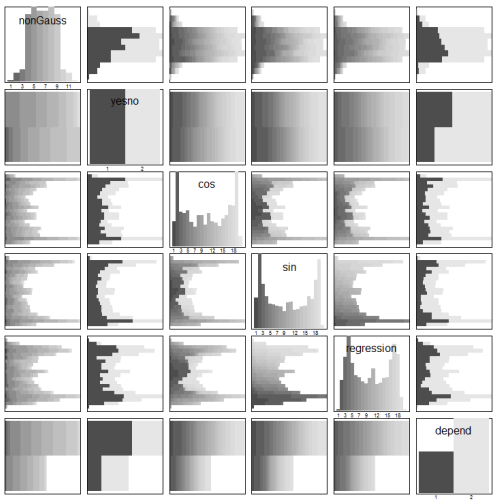
$$x_6 = \begin{cases} 1 & \text{確率 } 1/4\text{で} \\ 2 & \text{確率 } 3/4\text{で} \end{cases}$$

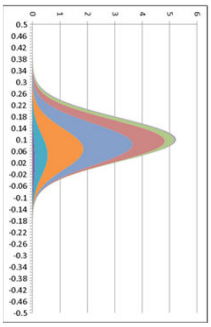
$x_2 = 2$ の場合、

$$x_6 = \begin{cases} 1 & \text{確率 } 1/2\text{で} \\ 2 & \text{確率 } 1/2\text{で} \end{cases}$$

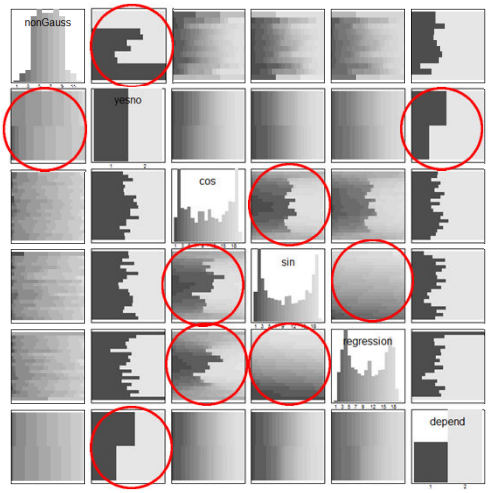
とする。

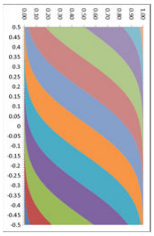
3.1.2 条件付き確率関数・密度関数の可視化





3.1.3 ベイズ事後分布関数の可視化





「縦縞でない」ものに丸印。
2行1列は一見縦縞っぽいがよく見ると縦方向の構造変化がある。

3.1.4 情報量規準 AIC の利用

確率変数 X と Y が関連していることをデータに基いて判断するには、モデル f_Y より $f_{Y|X}$ の方がデータへの当てはまりが良いことを確認すればよい。

情報量統計学的データ可視化ツールにいいては

$$\text{AIC差} = (\text{モデル } f_{Y|X} \text{ の } AIC) - (\text{モデル } f_Y \text{ の } AIC) < 0$$

であればX と Y が関連しているとする。

Y が離散変数の場合は、f_Y を確率関数、連続変数の場合は、f_Y を確率密度関数とする。

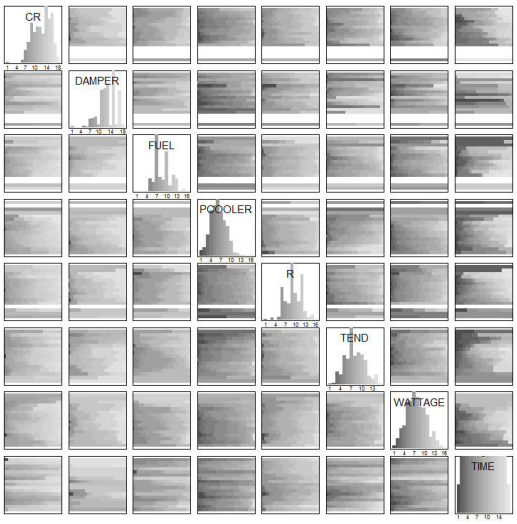
AIC差による判断の性能

この方法を、「テストデータ」に適用した結果は以下の通りである。 nonGauss と yesno, cos と sin, sin と regression、yesno と depend の間に関連があるという結果であるが、これは正しい結果である。

目的変数＼説明変数	nonGauss	yesno	cos	sin	regression	depend
nonGauss	0.00	−0.19	0.00	0.00	0.00	0.00
yesno	−1.18	0.00	0.14	0.13	0.07	−0.06
cos	0.00	0.00	0.00	−2.16	−1.30	0.00
sin	0.00	0.00	−2.16	0.00	−4.12	0.00
regression	0.00	0.00	−1.41	−3.92	0.00	0.00
depend	0.00	−0.06	0.13	0.07	0.11	0.00

目的変数＼説明変数	NG	YN	CO	SI	RG	DP
nonGauss(NG)		*	−	−	−	
yesno(YN)	*					#
cos(CO)	−			*	*	
sin(SI)	−	−	*		*	
regression(RG)	−	−	*	*		
depend(DP)	−	#				

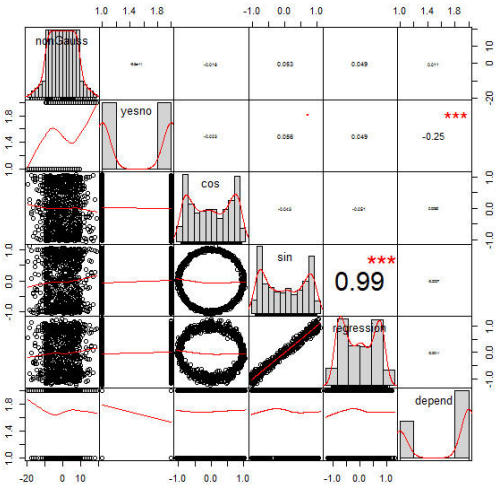
3.1.5 時系列データの可視化



「時間」が有効な説明変数となっていることが分かる。ただしこの方法で捕まるのは低周波変動である。

3.2 蛇足

3.2 1 相関解析が有効な場面は限られている

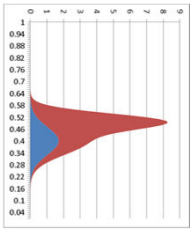


3.2.2 記述統計と推測統計

1.1節で、モデル

$$\sum_{x=0}^{1} f_{Y|X}(y|x)P_X(x)$$

が右のように「可視化」される例を示した。



この「可視化」は、当然のことながらP_X に依存する。P_X としてデータにおける X の周辺分布をつかえば上の「可視化」はデータの特徴を記述する記述統計的性格を持つことになる。このモデルを推測目的で使うときには、推測の場面に適合するP_X を用いるべきである。データをとったときと同じ状況での予測に使うのならデータ周辺分布のP_X を使うのが当然であるが。。

参考文献

- 坂元・石黒・北川（1983）.情報量統計学、共立出版
- Katsura,K. & Sakamoto,Y.(1980). CATDAP, A categorical data analysis program package, Computer Science Monographs, No.14, The Institute of Statistical Mathematics,Tokyo.
- 坂元慶行（1985）.カテゴリカルデータのモデル分析.共立出版.
- 情報量統計学的データ可視化ツール　An AIC-based Tool for Data Visualization　http://hdl.handle.net/10787/3614