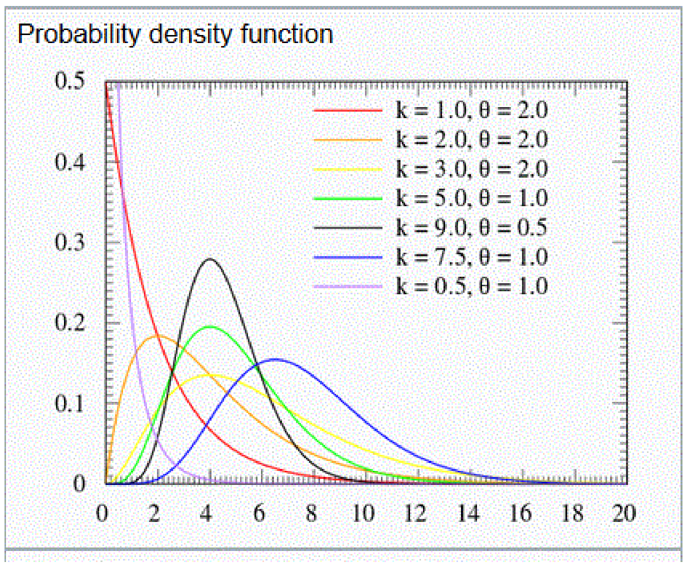
Odhad směsi s gama komponentami

Gama rozdělení je velice blízké exponenciálnímu. Jeho hustota pravděpodobnosti je

$$f(y) = \frac{1}{\Gamma(k)\, d^k} y^{k-1} \exp\left\{-\frac{y}{k}\right\} \tag{1}$$

kde d a k jsou parametry rozdělení. Parametr d má obdobný význam jako u exponenciálního rozdělení, parametr k určuje tvar distribuce. Pro $k < 1$ má maximum v nule (jako exponenciální), pro $k > 1$ maximum není v nule a hustota pravděpodobnosti má konkávní část (kopeček) s vrcholem poblíž nuly. Obrázek ukazuje různé průběhy této distribuce



Odtud plyne, že gama rozdělení může modelovat děle, které jsou stabilně nulové a mají kladné odchylky od nuly, nebo i děje, které jsou stabilně blízké nule a mohou se odchylovat, přičemž výsledná hodnota je vždy nezáporná.

Příklad modelování

Abychom výroky "stabilně nulové" a "stabilně blízké nule" lépe osvětlili, budeme uvažovat dva modely jízdy automobilem. Jedna je klidná (šetřivá) a druhá sportovní (divoká). Jako měřenou veličinu vezmeme absolutní hodnotu příčného zrychlení. U klidné jízdy bude tato hodnota ponejvíce nulová, občas (v zatáčce) se objeví odchylka. Při divoké jízdě bude průměrná hodnota nenulová a občas se objeví odchylka k nule (při dlouhé rovné silnici) nebo k větším hodnotám, při zvláště ostrých zatáčkách.

Střední hodnota je $E[y] = kd$.

Rozptyl $D[y] = kd^2$.

Modus (maximum) $\operatorname{mod}(y) = (k-1)\,d$.

Věrohodnostní funkce

$$L_N(\alpha, \beta) = \left(\frac{\beta^\alpha}{\Gamma(\alpha)\, d^k}\right)^N P_N^{k-1} \exp\left\{-\frac{S_N}{d}\right\}$$

kde $P_N = \prod_{i=1}^{N} x_i$ a $S_N = \sum_{i-1}^{N} x_i$.

Přepočet statistik

$$S_t = S_{t-1} + x_t, \quad P_t = P_{t-1} x_t, \quad \kappa_t = \kappa_{t-1} + 1 \tag{2}$$

Bodové odhady

.. zavedeme veličinu s

$$s = \ln\left(\frac{S_N}{N}\right) - \frac{1}{N}\ln\left(P_N\right)$$

a potom

$$\hat{k} \doteq \frac{3 - s + \sqrt{(s-3)^2 + 24 \cdot s}}{12 \cdot s}$$

a

$$\hat{d} = \frac{S_N}{\hat{\alpha} N} = \frac{\bar{y}}{\hat{k}} \tag{3}$$

kde $\bar{y} = \frac{S_N}{N}$ je průměrná hodnota výstupu.

Odhad směsi v každém čase probíhá podle následujícího schematu

1. Výpočet vah w pro jednotlivé komponenty, kde hlavním prvkem jsou tzv. proximity - tj. hodnoty hustoty pravděpodobnosti příslušné komponenty s dosazenými existujícími bodovými odhady parametrů a aktuálně změřenými daty, tj. $f_c\left(y_t|\hat{\Theta}_{c;t-1}\right)$, kde $c = 1, 2, \cdots, nc$ (pro všechny komponenty).
2. Přepočet statistik, kde data se přidávají s příslušnou váhou.
3. Výpočet bodových odhadů.

Pro tyto kroky potřebujeme

1. Model 1.
2. Vzorec pro přepočet statistik 2.
3. Vzorec pro výpočet bodového odhadu parametrů 3

Všechny tyto kroky jsou specifické pro určité rozdělení (zde gama).

Program (komentáře jsou za programem)

```
// Pokusy s gamm proximity jako likelihood
// - dvourozměrná data
// - více komponent
// - dvourozměrna data
// -----------------------------------------
exec SCIHOME/ScIntro.sce, mode(0), rand('seed',0);
```

```
nc=5;          // number of components
nv=2;          // number of variables
nd=1000;       // number of data
prox=1;        // 1=gamm proximity, 0=gaussian-like prox.

thS=list();
//      a b
thS(1)=[1 2      // x1
       10 2];    // x2        Egam=a/b
thS(2)=[8 1
        1 2];
thS(3)=[8 2
       12 3];
thS(4)=[28 2
       5 3];
thS(5)=[.8 2
        .2 3];
alS=[.2 .1 .3 .2 .2]; alS=fnorm(alS(1:nc));
cS(1)=1;

for t=2:nd     // ============== SIMULATION ============
  cS(1,t)=sum(rand(1,1,'u')>cumsum(alS))+1;
  u=rand(nv,1,'u');
  thP=thS(cS(t));
  x(:,t)=cdfgam("X",thP(:,1),thP(:,2),u(:),1-u(:));
end
//scatt(x)
//return

              // ============ INITIALIZATION ============
ni=50;
ka=ni*(.1*rand(1,nc,'u')+1);
nu=ni*(rand(1,nc,'u')+5);          // zatim se nepoužije (s tím je horší)
al=nu/sum(nu);                     // \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
zz=[];
for i=1:nc
  z=zeros(2,ni);
  for j=1:ni
    u=rand(nv,1,'u');
    P=thS(i); //+2*rand(thS(i),'u');
    P(:,2)=ones(2,1);
    z(:,j)=cdfgam("X",P(:,1),P(:,2),u,1-u);
  end
  Cy.V=[sum(z,2) sum(log(z),2)];
  zz=[zz z];

  s=log(Cy.V(:,1)/ka(i))-Cy.V(:,2)/ka(i);
  Cy.th(:,1)=(3-s+sqrt((3-s)^2+24*s))./(12*s);
  Cy.th(:,2)=Cy.th(:,1)*ka(i)./Cy.V(:,1);
```

```
  Est.Cy(i)=Cy;
end
ct=ones(1,nd);
tht=list();
for i=1:nc, tht(i)=zeros(4,nd); end

printf(' ')  //  ============= TIME LOOP ==============
for t=2:nd
  if t/fix(nd/10)==fix(t/fix(nd/10)), printf('.'); end

  // proximity
  for i=1:nc
    if prox==1
      [xxx,G(i)]=GammaG(x(:,t),Est.Cy(i).th);
    else
      Ex=Est.Cy(i).th(:,1)./Est.Cy(i).th(:,2);
      G(i)=(x(:,t)-Ex)'*(x(:,t)-Ex)/1000;
    end
  end
  Gn=G-max(G);
  g=exp(Gn);
//  g=g.*al';
  w=g/sum(g);
  wt(:,t)=w(:);

  // přepočet statistik pro gamma
  for i=1:nc
    Est.Cy(i).V(:,1)=Est.Cy(i).V(:,1)+w(i)*x(:,t);        // data
    Est.Cy(i).V(:,2)=Est.Cy(i).V(:,2)+w(i)*log(x(:,t));   // log()

    // bodové odhady parametrů komponent pro gamma
    s=log(Est.Cy(i).V(:,1)/ka(i))-Est.Cy(i).V(:,2)/ka(i);
    Est.Cy(i).th(:,1)=(3-s+sqrt((3-s)^2+24*s))./(12*s);
    Est.Cy(i).th(:,2)=(Est.Cy(i).th(:,1)*ka(i))./Est.Cy(i).V(:,1);
    tht(i)(:,t)=Est.Cy(i).th(:);
  end
  ka=ka+w';
  nu=nu+w';    // zatim se nepoužije
  al=nu/sum(nu);
end
[xxx,cE]=max(wt,'r');
printf('\n')

// VÝSLEDKY
[xxx,Ect]=max(wt,'r');
[q,T]=c2c(cS,Ect);
Ecc=q(Ect);
wrong=sum(cS(:)~=Ecc(:))
```

```
from=nd

// sim -----------------------
CS=list();
for i=1:nc
  j=find(cS==i);
  CS(i)=x(:,j);
end

tx=['b.';'r.';'g.';'m.';'k.';'y.'];
scf(1);
for i=1:nc
  plot(CS(i)(1,:),CS(i)(2,:),tx(i),'markersize',3)
end
title 'Simulated clusters'
set(gca(),'data_bounds',[0 45 0 40])

// est ------------------------
C=list();
for i=1:nc
  j=find(Ecc==i);
  C(i)=x(:,j);
end

set(scf(2),'figure_position',[850 200]);
for i=1:nc
  if ~isempty(C(i))
  plot(C(i)(1,:),C(i)(2,:),tx(i),'markersize',3)
  end
end
title 'Estimated clusters'
set(gca(),'data_bounds',[0 45 0 40])

set(scf(3),'position',[10 50 1500 400])
for i=1:nc
subplot(1,nc,i),plot(tht(i)')
end
legend('11','21','12','22');
```

Inicializace - dodělat !!!!!