数据科学基础 I (Matlab)

东北大学

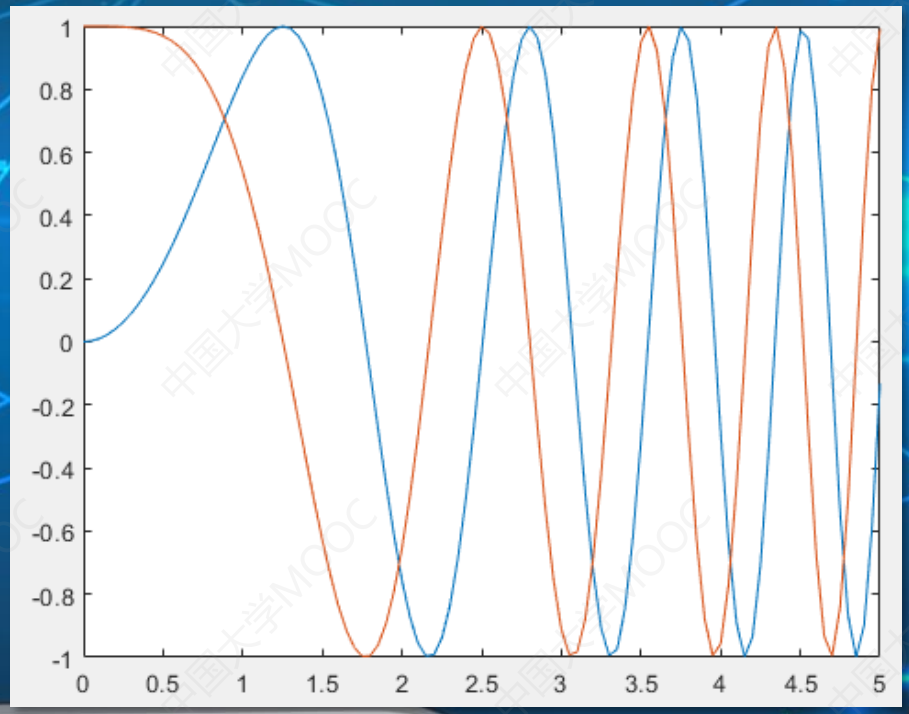
点线类基本命令

命令	功能说明
plot	x轴和y轴均为线性刻度（Linear scale）
loglog	x轴和y轴均为对数刻度（Logarithmic scale）
semilogx	x轴为对数刻度，y轴为线性刻度
semilogy	x轴为线性刻度，y轴为对数刻度
plotyy	双纵坐标绘图(不推荐，建议用yyaxis替换)
scatter	散点图
plotmatrix	散点图矩阵
spy	矩阵的稀疏模式散点图
fplot	根据表达式或函数绘图
polarplot	在极坐标系中绘图

plot基本绘图

可以在一次绘制多条线

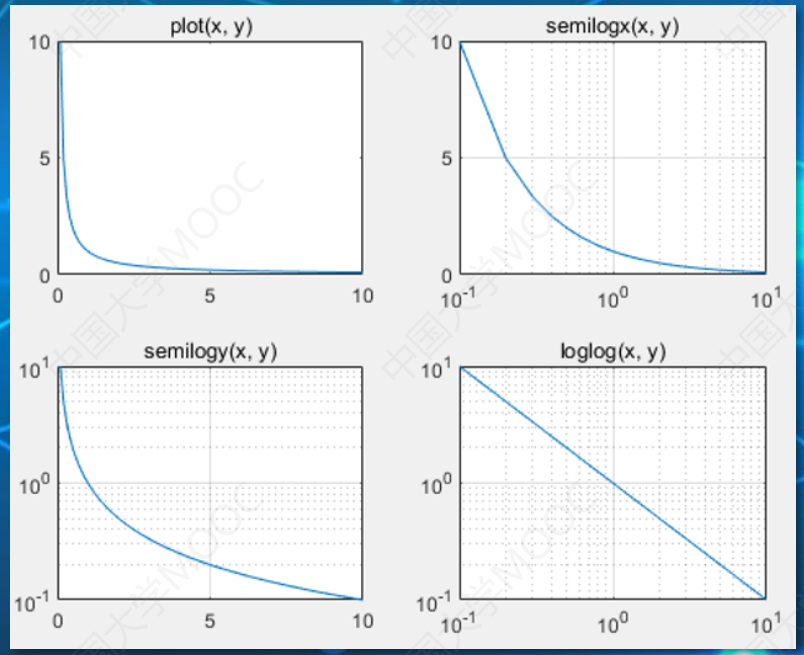
```
>> x = 0:0.05:5;
>> y1 = sin(x.^2);
>> y2 = cos(x.^2);
>> plot(x,y1,x,y2)
```



Loglog/semilogx/semilogy对数刻度

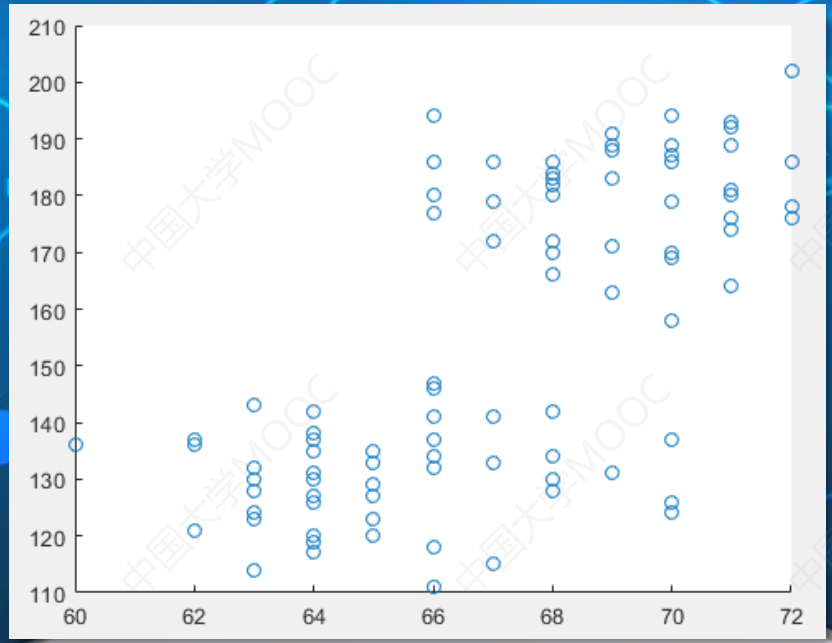
相比于线性刻度，对于变化快数据范围大的信息更有利

```
>> x = 0:0.1:10;
>> y = 1./x;
>> subplot(2, 2, 1);
>> plot(x, y);title('plot(x, y)');
>> subplot(2, 2, 2);
>> semilogx(x, y); grid on;
……
```



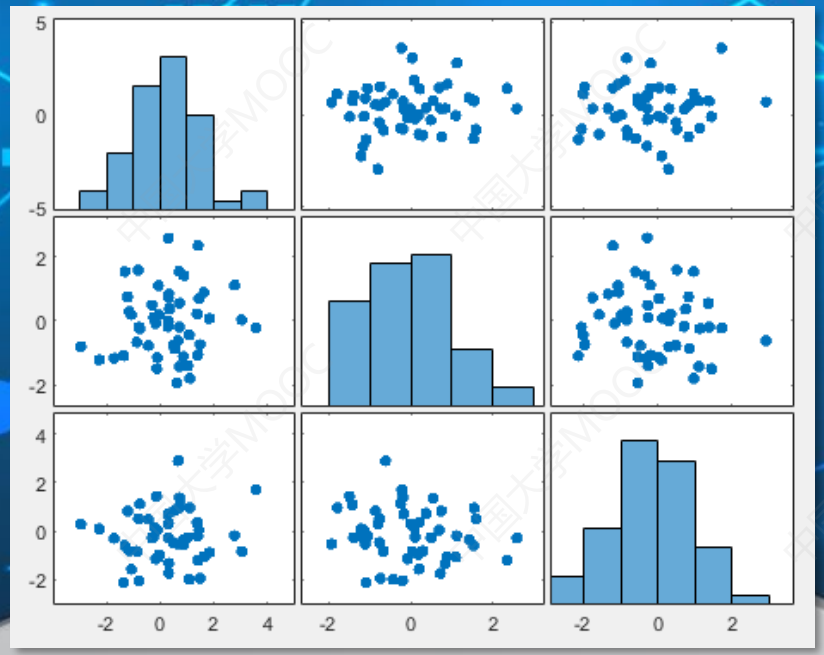
scatter

```
>> load patients Height Weight Systolic ;
>> scatter(Height,Weight) ;
```



plotmatrix矩阵散点图

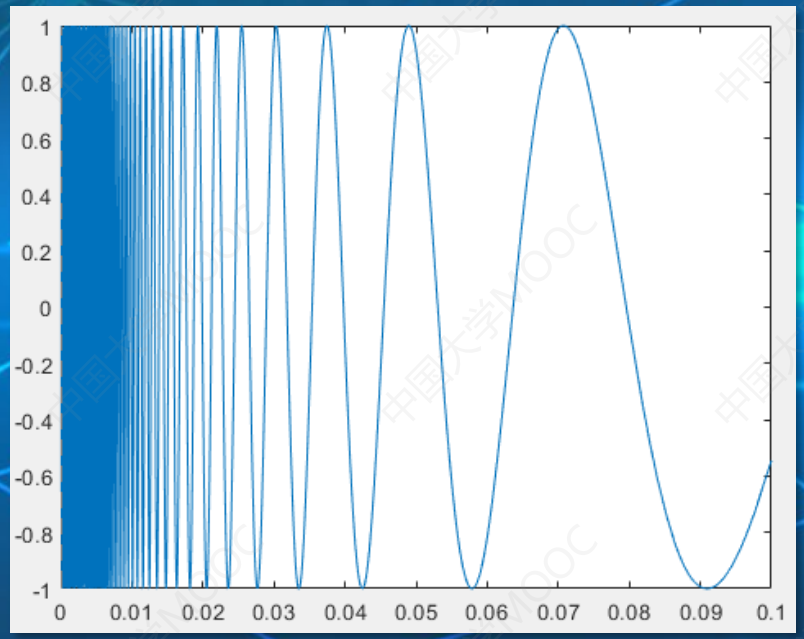
```
>> X = randn(50,3);
>> plotmatrix(X);
```



fplot

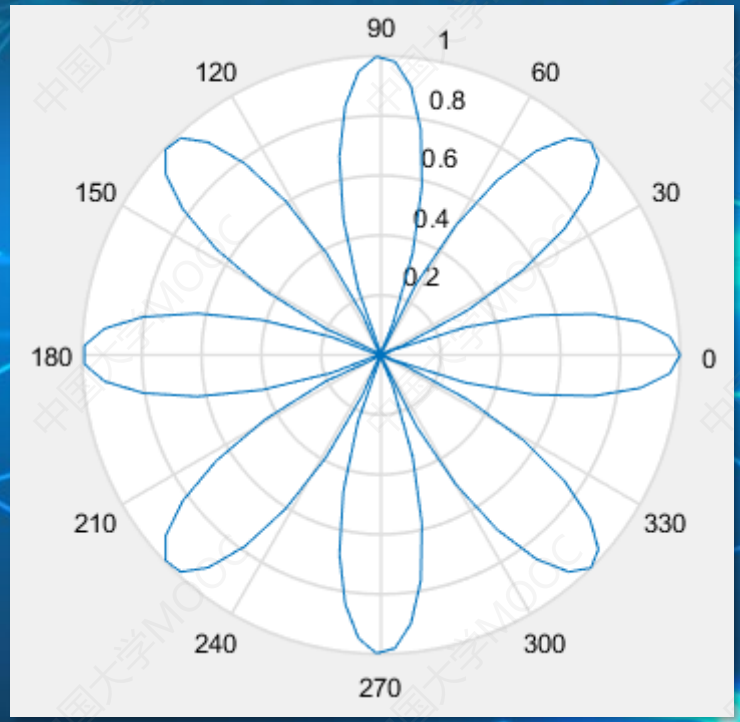
对于变化剧烈的函数，可用fplot来进行较精确的绘图，会对剧烈变化处进行较密集的取样。

```
>> fplot(@(x)sin(1./x),[0 0.1]);
```



polarplot极坐标绘图

```
>> theta = linspace(0,2*pi);
>> r = cos(4*theta);
>> polarplot(theta,r);
```

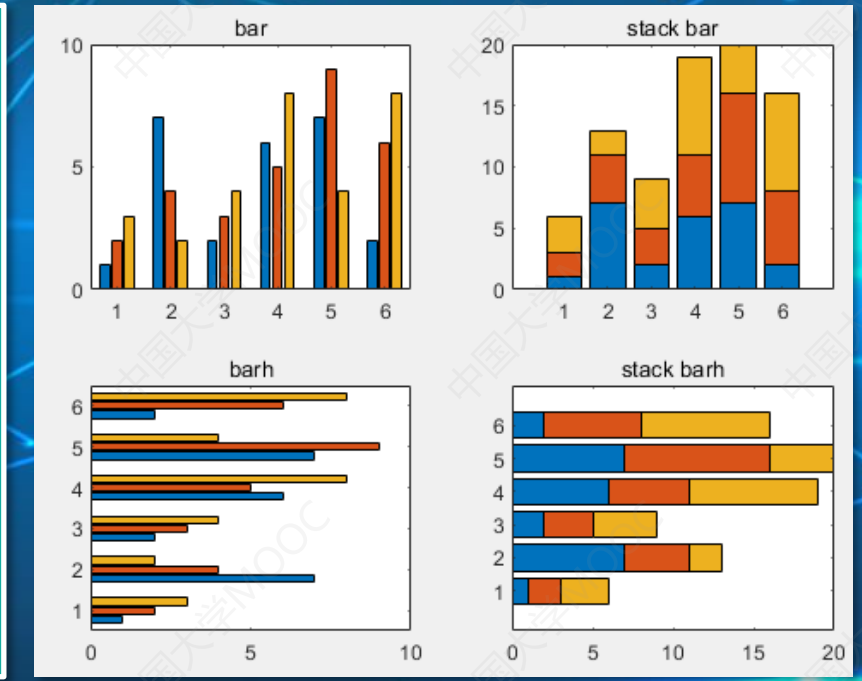


点线类基本命令——填充类

命令	功能说明
bar	方块图（垂直柱）
barh	方块图（水平柱）
fill	填充图
area	面积图

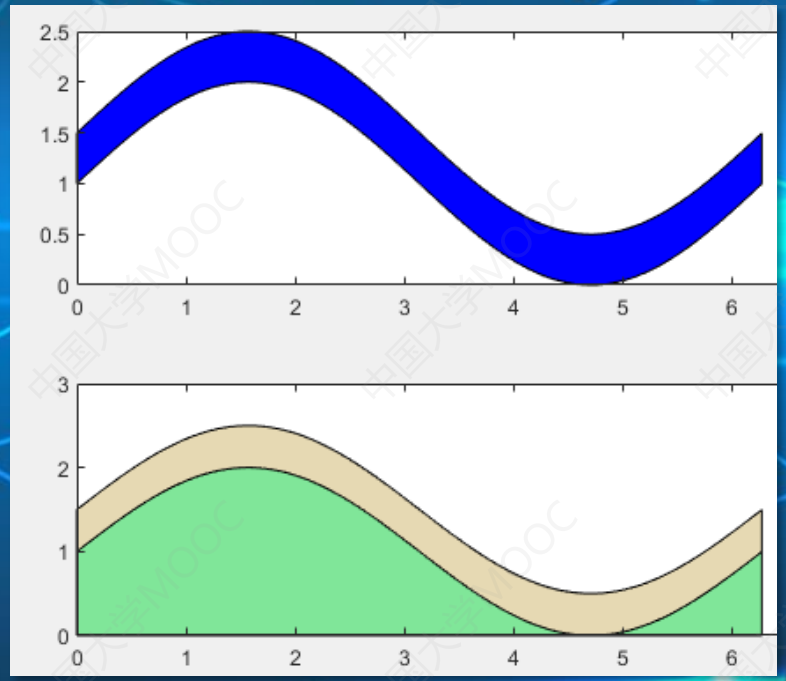
bar/barh 方块图

```
>> Y = [1 2 3; 7 4 2; 2 3 4; 6 5 8; 7 9 4; 2 6 8];
>> subplot(2,2,1)
>> bar(Y)
>> title('bar')
>> subplot(2,2,2)
>> bar(Y,'stack')
>> title('stack bar')
>>……
```

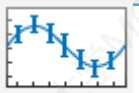
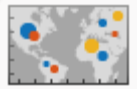
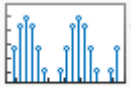
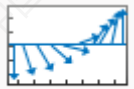
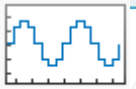
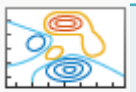
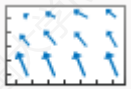


fill /area填充

```
>>x=0:0.01:2*pi;
>> y1=sin(x);
>> y2=sin(x)+.5;
>> X=[x,fliplr(x)];
>> Y=[y1,fliplr(y2)];
>> subplot(2,1,1);
>> fill(X,Y,'b');
>> subplot(2,1,1);
>> area(x,y2,'facecolor',[0.9 0.85 0.7]);
>> hold on;
>> area(x,y1,'facecolor',[0.5 0.9 0.6]);
```



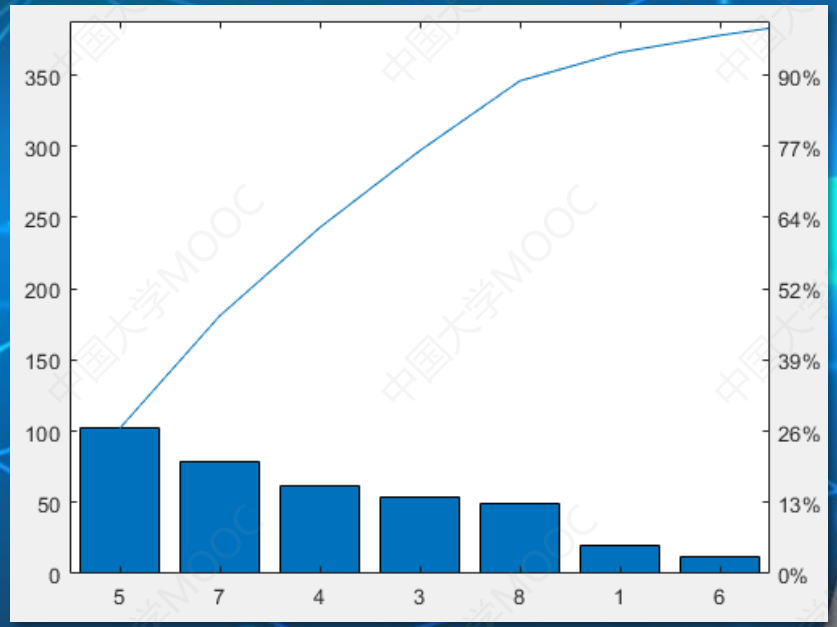
特殊二维图形

命令	功能说明	样例	命令	功能说明	样例
pareto	帕累托图		wordcloud	文字云图	
errorbar	误差图		geobubble	气泡地理图	
stem	针状图		feather	羽毛图	
stairs	阶梯图		compass	罗盘图	
contour	等高线图		quiver	向量场图	

pareto帕累托图

用于寻找问题或主要因素

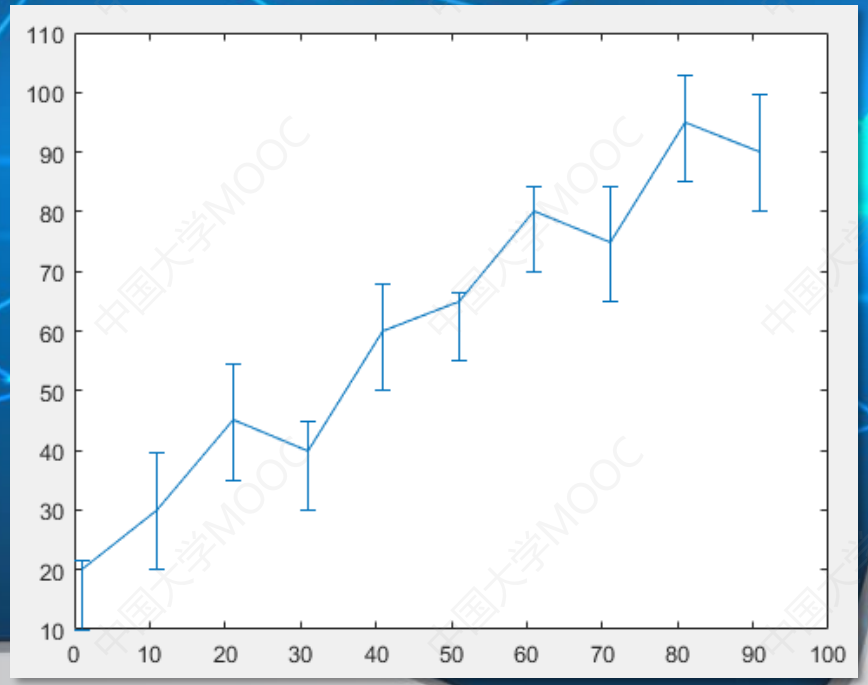
```
>> y = [20,10,54,62,102,12,79,49];
>> pareto(y)
```



errorbar误差图

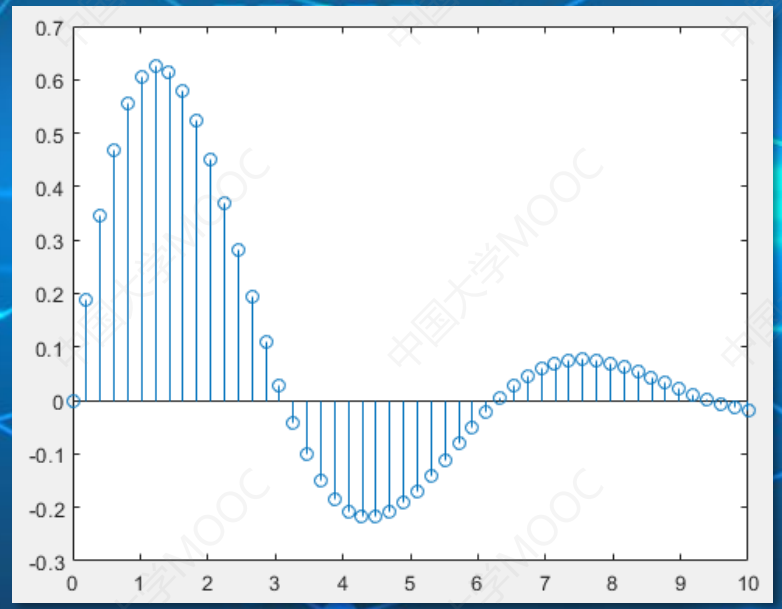
误差棒是数据可变性的图形表示，并用于图表以指示所报告的测量中的误差或不确定性。

```
>> x = 1:10:100;
>> y = [20 30 45 40 60 65 80 75 95 90];
>> err1 = 10*ones(size(y));
>> err2 = 10*rand(size(y));
>> errorbar(x,y,err1,err2)
```



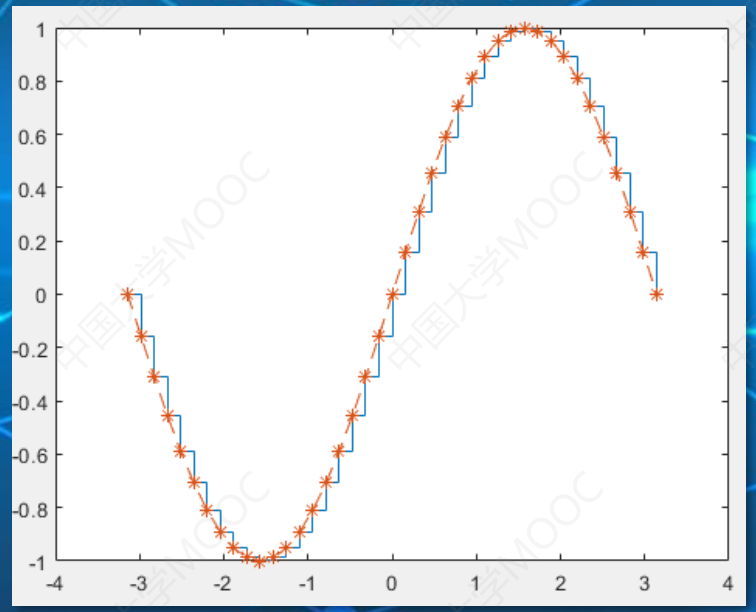
stem针状图

```
>> x = linspace(0,10,50);
>> y = sin(x).*exp(-x/3);
>> stem(x,y);
```



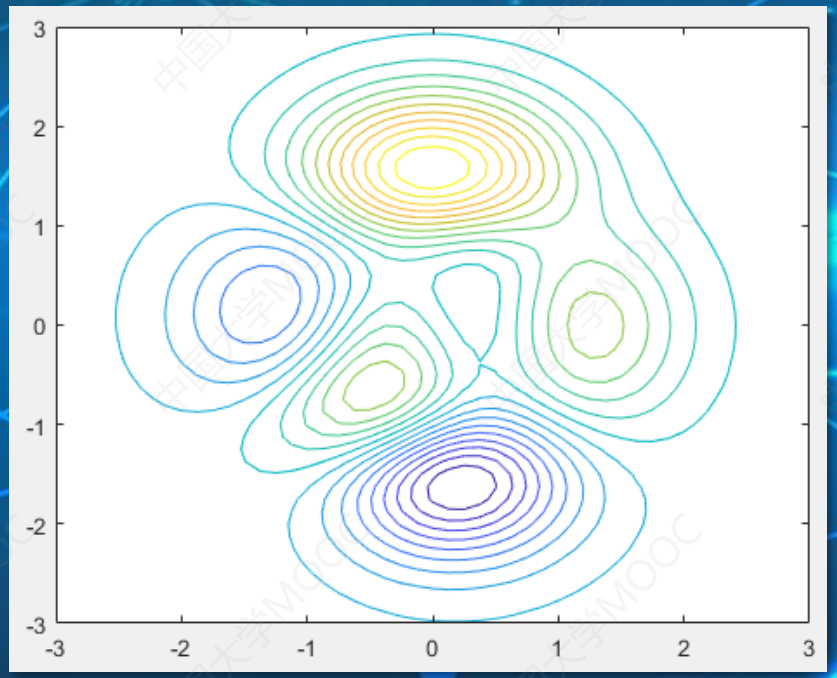
stairs绘制阶梯图

```
>> x = -pi:pi/20:pi;
>> y = sin(x);
>> stairs(x,y)
>>hold on
>>plot(x,y,'--*')
```



contour等高线图

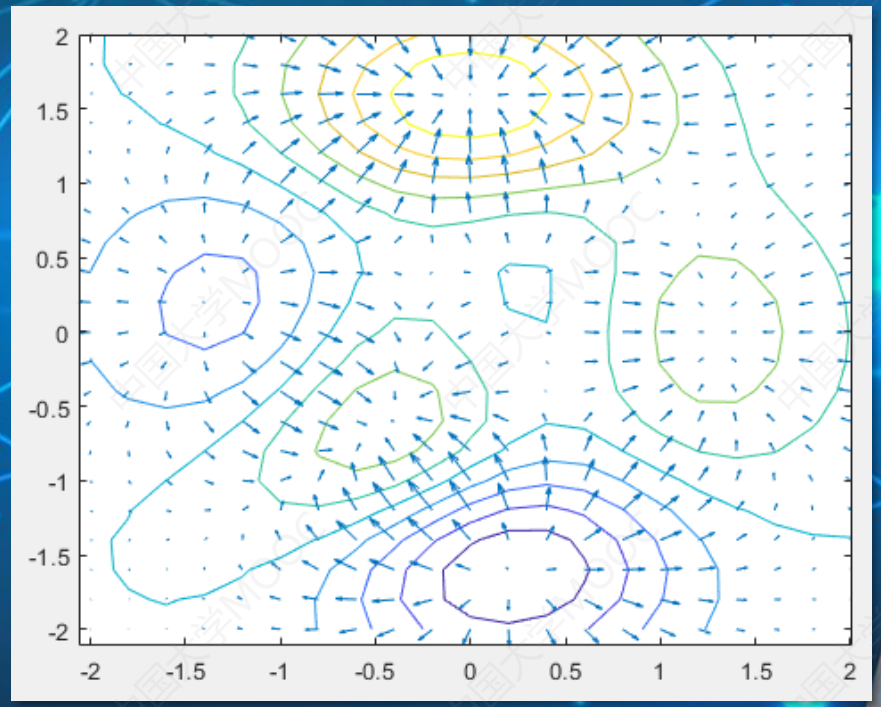
```
>> [X,Y,Z] = peaks;
>> contour(X,Y,Z,20)
```



quiver向量场图

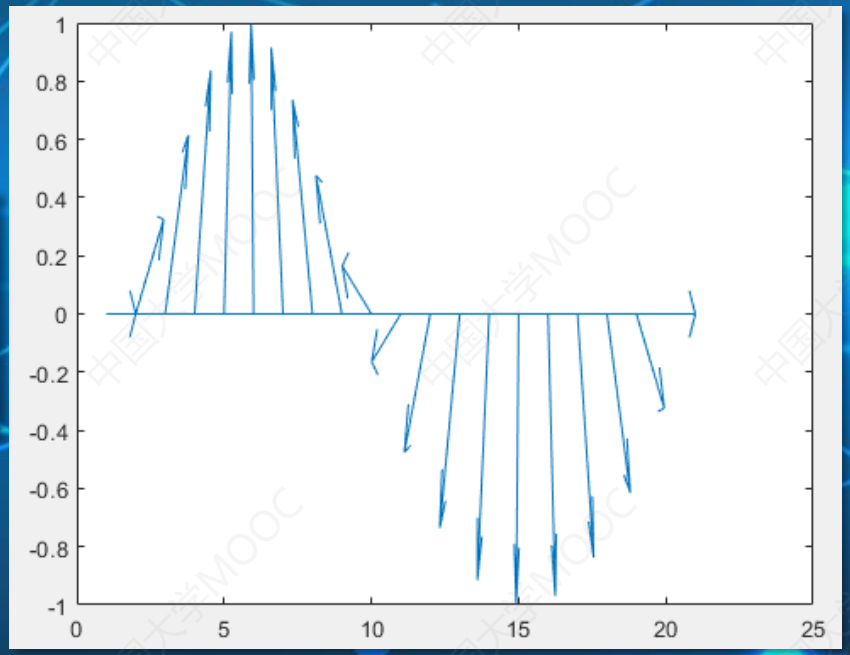
箭头指向的方向为向量方向，箭头长短代表向量大小

```
>> n=-2.0:0.2:2.0;
>> [x,y,z]=peaks(n);
>> contour(x,y,z,10)
>> [u,v]=gradient(z,0.2);
>> hold on
>> quiver(x,y,u,v)
```



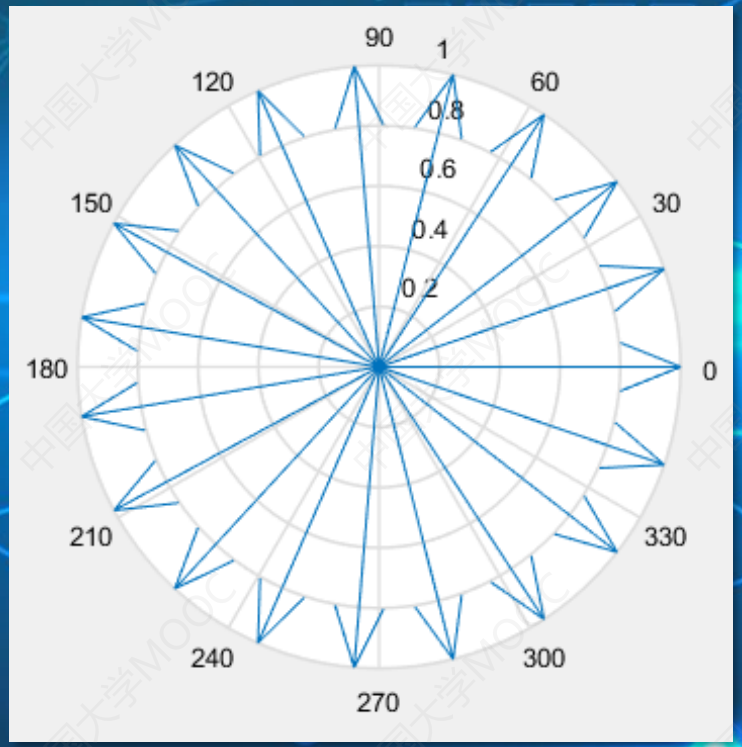
feather羽毛图

```
>> theta = linspace(0,2*pi,20);
>> z = cos(theta)+i*sin(theta);
>> feather(z);
```



compass罗盘图

```
>> theta=linspace(0, 2*pi, 20);
>> z = cos(theta)+i*sin(theta);
>> compass(z);
```

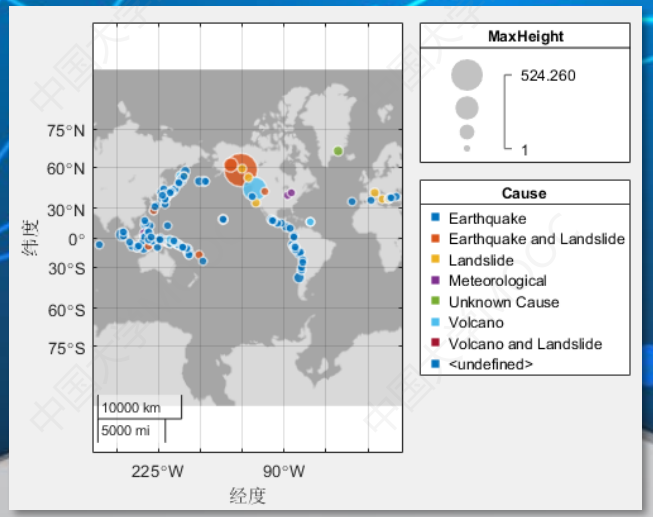


wordcloud文字云图

```
>> load sonnetsTable
>> wordcloud(tbl,'Word','Count');
```

geobubble 气泡地理图

```
>> tsunamis = readtable('tsunamis.xlsx');
>> tsunamis.Cause = categorical(tsunamis.Cause);
>> gb = geobubble(tsunamis,'Latitude','Longitude', ...
        'SizeVariable','MaxHeight','ColorVariable','Cause')
```

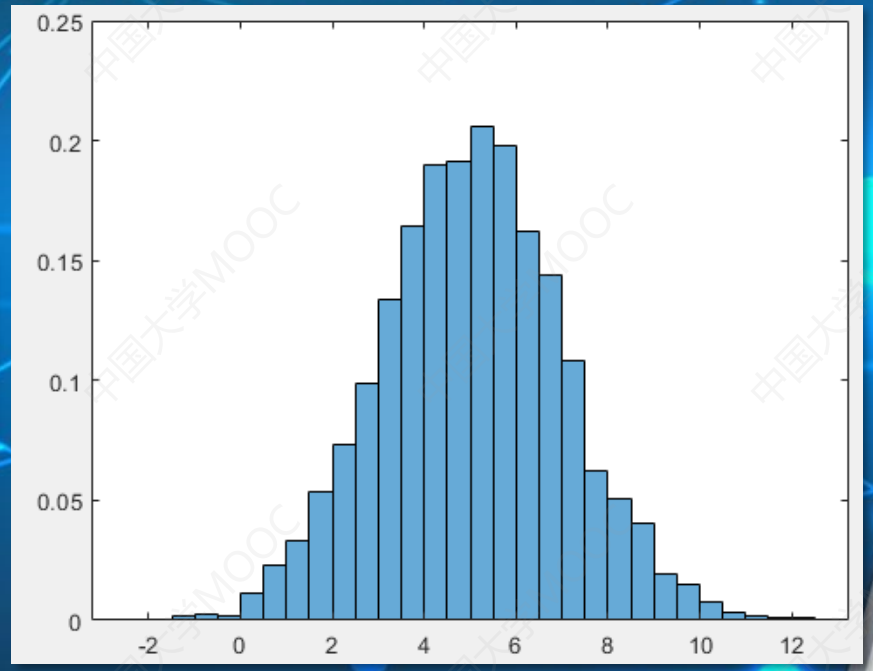


概率/统计类数据可视化

命令	功能说明
histogram	绘制直方图
rose	极坐标累计图
scatterhistogram	绘制带直方图的散点图
boxplot	箱线图
histfit	附加正态密度曲线的直方图
normplot	正态概率图
wblplot	wbl分布概率图
probplot	通用概率图
qqplot	q-q图
cdfplot	经验累积分布图

histogram直方图

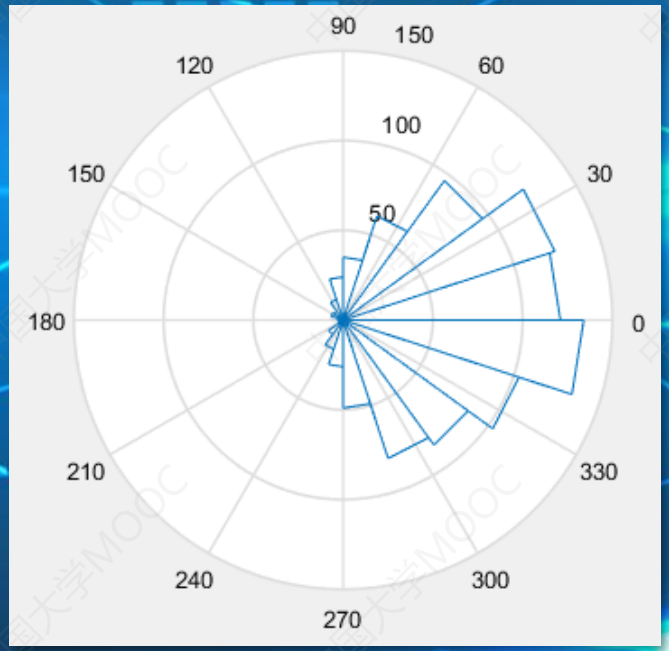
```
>>x = 2*randn(5000,1) + 5;
>> histogram(x,'Normalization','pdf')
```



rose极坐标系中的统计图

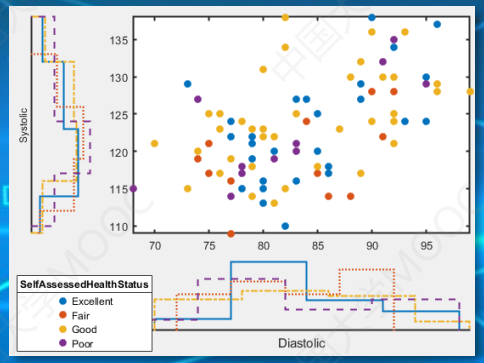
rose和histogram很接近，将信息大小视为角度，信息个数视为距离，并在极坐标系绘制

```
>> x=randn(1000, 1);
>> rose(x);
```



scatterhistogram带直方图的散点图

将scatter和histogram的结果集成显示

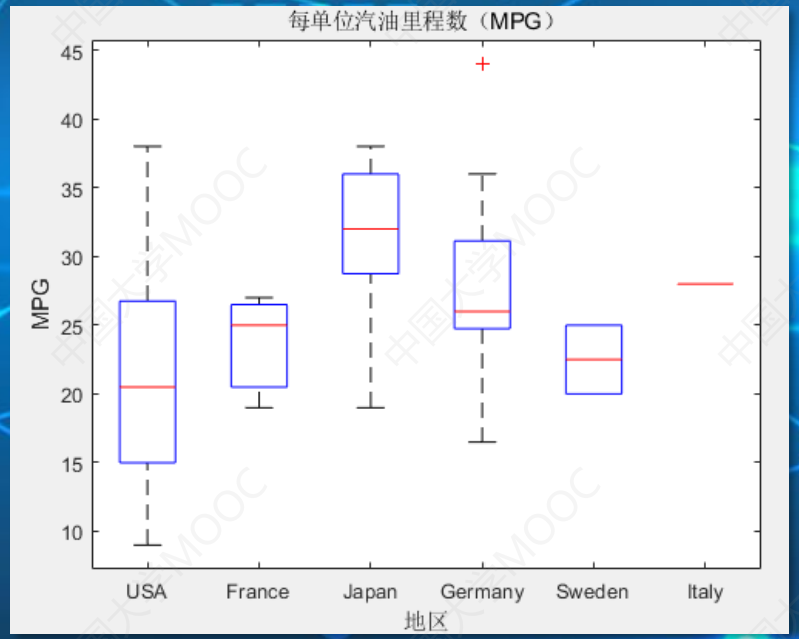



```
>> load patients;
>> tbl = table(LastName,Diastolic,Systolic,SelfAssessedHealthStatus);
>> s =
scatterhistogram(tbl,'Diastolic','Systolic','GroupVariable','SelfAssessedHealthStatus',
...
'NumBins',4,'LineWidth',1.5,'ScatterPlotLocation','NorthEast','LegendVisible','on');
```

boxplot箱线图

在每个box上，中心标记表示中位数，箱子的底边和顶边分别表示第25个和75个百分位数。虚线会延伸到不是离群值的最远端数据点，离群值会以 '+' 符号单独绘制。

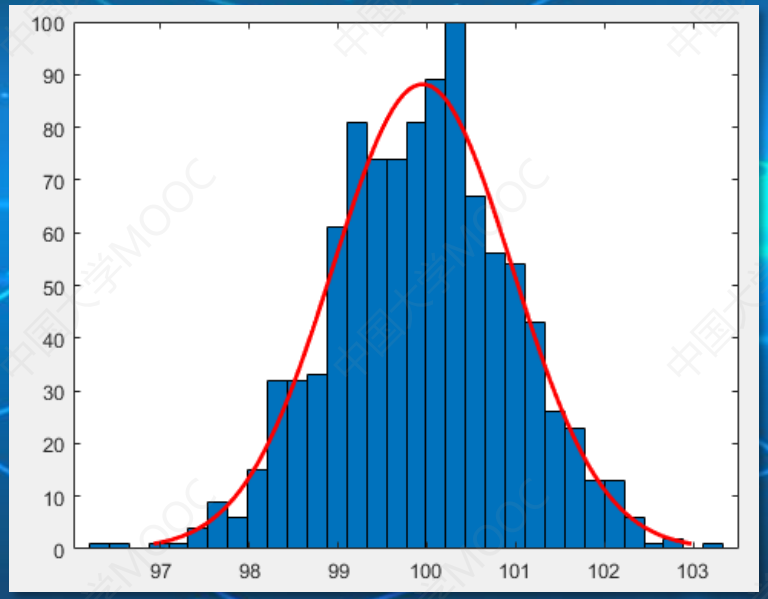
```
>> load carsmall
>> boxplot(MPG,Origin)
>> title('每单位汽油里程数（MPG）')
>> xlabel('地区')
>> ylabel('MPG')
```



histfit在直方图上附加正态密度曲线

用以判断是否符合正态分布

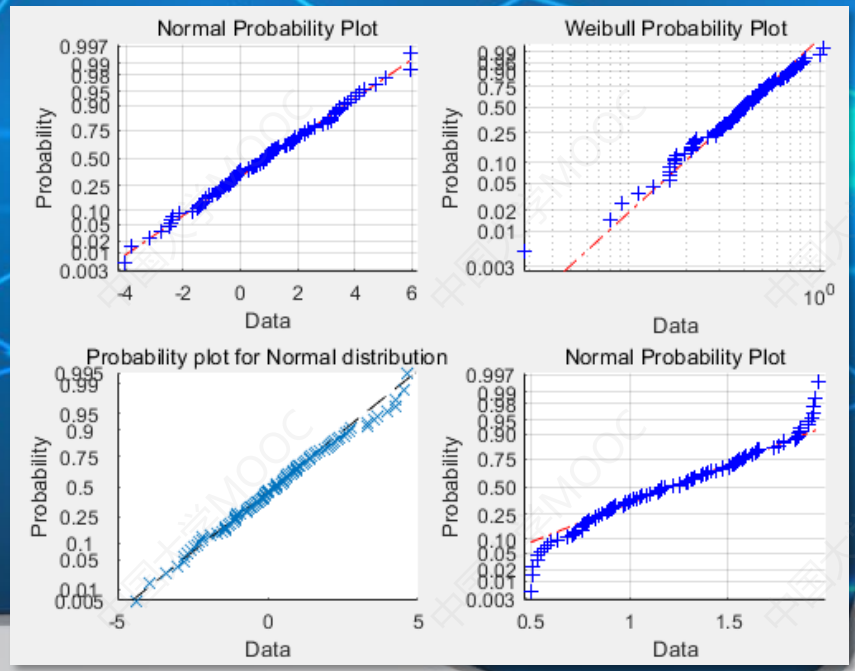
```
>> x = normrnd(100,1,1000,1);
>> histfit(x)
```



normplot/wblplot/probplot/qqplot

检查数据是否满足某种特定的分布

```
>> r1 = normrnd(0.5,2,[100,1]);
>> subplot(2,2,1)
>> normplot(r1)
>> r2 = wblrnd(0.5,2,[100,1]);
>> subplot(2,2,2)
>>wblplot(r2)
>>r3 = normrnd(0.5,2,[100,1]);
>>subplot(2,2,3)
>>probplot('normal',r3)
>>r4=unifrnd(0.5,2,[100,1]);
>>subplot(2,2,4)
>>normplot(r4)
```



cdfplot 经验累积分布图

为了观测随机变量的取值在哪个附近出现的概率比较大

```
>> x1=normrnd (0,1,1000,1);
>> subplot(2,1,1)
>> cdfplot(x1)
>> x2=unifrnd (0,1,1000,1);
>> subplot(2,1,2)
>>cdfplot(x2)
```