Primer on Probability Theory

一、Probability Density Functions 概率密度函数

1.1 Definitions

1.1.1 PDF Definition

我们缩写概率密度函数为PDF

我们设一个随机变量x，令p(x)为x在[a,b]上的PDF。

那么，我们用Pr代表概率，则：

$$Pr(c \leq x \leq d) = \int_c^d p(x)dx$$

1.1.2 条件概率密度函数

我们还可以类似得到条件概率，令p(x|y)是x的概率密度函数(x∈[a,b])，而条件y∈[r,s]，我们有：

$$\text{任取}\, y, \qquad \int_a^b p(x|y)dx = 1$$

（我们要明确一个思想，条件概率情况下，我们对随机变量进行积分，那么在随机变量的所有取值范围之内积分，其值必为1，不需要管条件是什么）

1.1.3 联合概率密度函数

我们还可以类似得到联合概率密度函数(joint probability densities)，N维情况下：我们有

$$\int_a^b p(x)dx = \int_{a_N}^{b_N} \dots \int_{a_2}^{b_2} \int_{a_1}^{b_1} p(x_1, x_2, \dots, x_N)dx_1 dx_2 \dots dx_N = 1$$

其中，a=(a1,a2,...,aN),b=(b1,b2,...,bN)

1.2 贝叶斯公式

首先，我们易得联合概率密度：

$$p(x,y) = p(x|y)p(y) = p(y|x)p(x)$$

我们重新排列上述公式即可得到贝叶斯公式：

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$$

若我们有先验概率密度p(x)以及传感器模型p(y|x)，我们可以用其来推断给定测量值的状态的后验p(x|y)。

我们对上式的分母进行扩展：

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)dx}$$

我们进行分母的推导：如下

$$p(\mathbf{y}) = p(\mathbf{y}) \underbrace{\int p(\mathbf{x}|\mathbf{y})\, d\mathbf{x}}_{1} = \int p(\mathbf{x}|\mathbf{y}) p(\mathbf{y})\, d\mathbf{x}$$

$$= \int p(\mathbf{x}, \mathbf{y})\, d\mathbf{x} = \int p(\mathbf{y}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x},$$

1.3 Moments 矩

a.k.a also known as 亦称为

当处理许多动态分布(a.k.a 密度函数)，我们通常关注较少的特性，称作moments of mass(e.g. mass,center of mass质心,inertia matrix惯性矩阵).

第零个概率矩总是1，一阶矩就是期望μ：

$$\mu = E[x] = \int xp(x)dx$$

E[·]为期望算子，我们还有：

$$E[F(x)] = \int F(x)p(x)dx$$

但是我们必须这样解释上式：

$$E[F(x)] = [E[f_{ij}(x)]] = [\int f_{ij}p(x)dx]$$

我们如此定义二阶矩：（是我们熟知的协方差矩阵）Σ：

$$\Sigma = E[(x-\mu)(x-\mu)^T]$$

三阶矩和四阶矩分别称作偏度(skewness)和峰度(kurtosis)。

1.4 Sample Mean and Covariance

假设我们有一个随机变量x，以及它的概率密度函数p(x)，我们可以从密度函数进行抽样，记作：

$$x_{meas} = p(x)$$

样本，也可以称作随机变量的一次**实现(realization)**，我们可以直观地将其理解为一次测量值。若我们抽取的样本中有N个值，来估计随机变量x的**均值(mean)和方差(covariance)**，如下：

$$\mu_{meas} = \frac{1}{N}\sum_{i=1}^{N} x_{i,meas}$$

$$\Sigma_{meas} = \frac{1}{N-1}\sum_{i=1}^{N}(x_{i,meas} - \mu_{meas})(x_{i,meas} - \mu_{meas})^T$$

注意到，协方差中用N-1而不是用N，这与**贝塞尔矫正(Bessel's correnction)**有关。

1.5 Statistically Independent, Uncorrelated

我们有两个随机变量x，y，我们说它们**统计学独立(statistically independent)**，若二者联合概率密度函数满足：

$$p(x, y) = p(x)p(y)$$

我们称变量**不相关(uncorrelated)**，若有：

$$E[xy^T] = E[x]E[y]^T$$

若随机变量统计学独立，那么它们不相关，反之则不一定成立。但出于简化计算的目的，我们经常会直接认为(假设)不相关的随机变量是统计独立的。

1.6 Normalized Product

我们翻译为：归一化积

如果p1(x)和p2(x)是随机变量x的两个不同的概率密度函数，那么它们的归一化积p(x)定义为：

$$p(x) = \eta p_1(x)p_2(x)$$
$$\eta = (\int p_1(x)p_2(x)dx)^{-1}$$

其中η是一个常值的归一化因子，用于确保p(x)满足全概率公理。

在贝叶斯理论中，我们可以用归一化积来融合随机变量x的多次独立估计：假设y1，y2为两次独立测量，x为待估计随机变量，则有

$$p(x|y_1,y_2) = \eta p(x|y_1)p(x|y_2)$$

我们讲上式左侧利用贝叶斯公式展开：

$$p(x|y_1,y_2) = \frac{p(y_1,y_2|x)p(x)}{p(y_1,y_2)}$$

假定y1，y2独立，那么有：

$$p(y_1,y_2|x) = p(y_1|x)p(y_2|x) = \frac{p(x|y_1)p(y_1)}{p(x)}\frac{p(x|y_2)p(y_2)}{p(x)}$$

代入上上式我们得到：

$$\eta = \frac{p(y_1)p(y_2)}{p(y_1,y_2)p(x)}$$

若令先验p(x)为均匀分布(常数)，那么η也是一个常量。

1.7 Shannon and Mutual Information

我们翻译为：香农和互信息

当我们在估计某一随机变量的概率密度函数时，我们也希望知道对某个量(比如均值)有多么不确定。那么，描述不确定性的方法有：**计算事件的负熵(negative entropy)或香浓信息量(Shannon information)**，记作H，如下：

$$H(x) = -E[lnp(x)] = -\int p(x)lnp(x)dx$$

另一个重要量则是**互信息(mutual information)，I(x,y)：**

$$I(x,y) = E[ln(\frac{p(x,y)}{p(x)p(y)})] = \int\int p(x,y)ln(\frac{p(x,y)}{p(x)p(y)})dxdy$$

互信息衡量的是已知一个随机变量的信息之后，另一个随机变量不确定性的减少了多少。当x和y统计学独立时，我们可以得到：

$$I(\mathbf{x},\mathbf{y}) = \iint p(\mathbf{x})\,p(\mathbf{y}) \ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right)\,d\mathbf{x}\,d\mathbf{y}$$
$$= \iint p(\mathbf{x})\,p(\mathbf{y}) \underbrace{\ln(1)}_{0}\,d\mathbf{x}\,d\mathbf{y} = 0.$$

而当x和y不独立时，我们有I(x,y)≥0，且：

$$I(x,y) = H(x) + H(y) - H(x,y)$$

上式关联了香农信息和互信息。

1.8 Cramer-Rao Lower Bound and Fisher Information

我们翻译为：克拉美罗下界和费歇尔信息量

假定我们有一个确定的参数θ，其可以影响随机变量x的概率密度，我们用条件概率来描述：

$$p(x|\theta)$$

之后，我们抽样：

$$x_{meas} < -p(x|\theta)$$

这里，xmeas有时可以被称作随机变量x的一个实现(realization)，我们可以看作为测量值。

接下来，**克拉美罗下界(Cramer-Rao lower bound(CRLB))**指出：参数真实值θ的任意**无偏估计θhat**(基于观测值xmeas)的协方差，可以由**费歇尔信息矩阵(Fisher information matrix) I(x|θ)**来定义边界：

$$cov(\hat{\theta}|x_{meas}) = E[(\hat{\theta}-\theta)(\hat{\theta}-\theta)^T] \geq I^{-1}(x|\theta)$$

其中，无偏估计指：

$$E[\hat{\theta}-\theta] = 0$$

费歇尔信息矩阵：

$$\mathbf{I}(\mathbf{x}|\boldsymbol{\theta}) = E\left[\left(\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^T \left(\frac{\partial \ln p(\mathbf{x}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\right].$$

2、Gaussian Probability Density Functions

我们先明确一下协方差矩阵：定义：

设 $X=(X_1,X_2,\ldots,X_N)^T$ 为n维随机变量，称矩阵

$$C=(c_{ij})_{n\times n}=\begin{pmatrix} c_{11} & c_{12} & \ldots & c_{1n} \\ c_{21} & c_{22} & \ldots & c_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ c_{n1} & c_{n2} & \ldots & c_{nn} \end{pmatrix}$$

为n维随机变量 X 的协方差矩阵（covariance matrix），也记为 $D(X)$ ，其中

$$c_{ij}=Cov(X_i,X_j),i,j=1,2,\ldots,n$$

为 X 的分量 X_i 和 X_j 的协方差（设它们都存在）。

那么，X为向量表示，那么我们还可以如此表示协方差：

$$Cov(x)=E[(x-x_m)(x-x_m)^T]$$

我们举个例子：

例如，二维随机变量 (X_1,X_2) 的协方差矩阵为

$$C=\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

其中 $c_{11}=E[X_1-E(X_1)]^2, c_{12}=E[X_1-E(X_1)][X_2-E(X_2)]$

$c_{21}=E[X_2-E(X_2)][X_1-E(X_1)], c_{22}=E[X_2-E(X_2)]^2$

我们写成矩阵形式：

$$Cov((x_1,x_2)^T)=E[\begin{bmatrix} x_1-x_{1m} \\ x_2-x_{2m} \end{bmatrix}\begin{bmatrix} x_1-x_{1m} & x_2-x_{2m} \end{bmatrix}]$$

结果与展开式相同。

2.1 Definitions

高斯概率密度函数，一维：

$$p(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

高维表示：x∈R^N

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})=\frac{1}{\sqrt{(2\pi)^N\det\boldsymbol{\Sigma}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

其中，μ∈R^N，为均值；Σ∈R^(N×N)是协方差矩阵(正定对称矩阵)。

因此我们有：

$$\boldsymbol{\mu} = E[\mathbf{x}] = \int_{-\infty}^{\infty} \mathbf{x} \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x}, \tag{2.37}$$

and

$$\begin{aligned} \boldsymbol{\Sigma} &= E\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T\right] \\ &= \int_{-\infty}^{\infty} (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}}} \\ &\qquad \times \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) d\mathbf{x}. \end{aligned} \tag{2.38}$$

高维情况下，我们写x~N(0，1)，表示x∈R^N服从高维高斯分布，其中，1为N×N阶单位矩阵。

2.2 Isserlis' Theorem

多维高斯分布的高阶矩比较难计算，但是它们在后续的学习中也很重要。因此我们利用Isserlis理论莱计算高斯分布的高阶矩。

假设x=（x1,x2,...,x2M）∈R^(2M)，通常来讲：

$$E[x_1 x_2 x_3 \ldots x_{2M}] = \sum \prod E[x_1 x_j]$$

这表明：计算2M个变量乘积的期望，可以首先计算所有两两不同的变量的乘积的期望，然后把计算出来这些期望做乘积。这样的组合有：

$$\frac{2M!}{2^M M!}$$

种，最后将这些乘积的值求和即可。例如，M=2时

$$E[x_i x_j x_k x_l] = E[x_i x_j]E[x_k x_l] + E[x_i x_k]E[x_j x_l] + E[x_i x_l]E[x_j x_k]$$

我们利用这个可以推导出一些有用的结果。

假设我们有x~N(0 , Σ)∈R^N。我们计算下式：

$$E[x(x^T x)^p x^T]$$

其中p是一个非负整数。当p=0时，我们有

$$E[xx^T] = \Sigma$$

当p=1时，我们有

$$E[xx_1^T x_1 x^T] = E[[]]$$

2.3 联合高斯概率密度函数，分解与推断

我们有联合随机变量服从高斯分布(x,y)，写为，其PDF写为：

$$p(x,y) = N(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix})$$

我们注意到：Σyx=Σxy^T。这可以将一个联合概率密度拆分为两个概率密度的乘积**(条件概率乘边缘概率)，p(x,y)=p(x|y)p(y)**。特别地，对于高斯分布，我们可以用**Schur complement舒贝尔定理。**

$$\begin{bmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{yx} & \Sigma_{yy}\end{bmatrix}=\begin{bmatrix}1 & \Sigma_{xy}\Sigma_{yy}^{-1}\\ 0 & 1\end{bmatrix}\begin{bmatrix}\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} & 0\\ 0 & \Sigma_{yy}\end{bmatrix}\begin{bmatrix}1 & 0\\ \Sigma_{yy}^{-1}\Sigma_{yx} & 1\end{bmatrix}$$

我们对等式两侧取逆矩阵得到:

$$\begin{aligned}\begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}^{-1} &= \begin{bmatrix}\mathbf{1} & \mathbf{0}\\ -\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} & \mathbf{1}\end{bmatrix}\\ &\times\begin{bmatrix}\left(\boldsymbol{\Sigma}_{xx}-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{yy}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{1} & -\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\\ \mathbf{0} & \mathbf{1}\end{bmatrix}.\end{aligned}$$

之后，我们看联合高斯分布种，指数部分(前面可以认为是常数):

$$\begin{aligned}&\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}-\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}\right)^T\begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy}\\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}^{-1}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}-\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}\right)\\ &=\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}-\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}\right)^T\begin{bmatrix}\mathbf{1} & \mathbf{0}\\ -\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} & \mathbf{1}\end{bmatrix}\begin{bmatrix}\left(\boldsymbol{\Sigma}_{xx}-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^{-1} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Sigma}_{yy}^{-1}\end{bmatrix}\\ &\qquad\times\begin{bmatrix}\mathbf{1} & -\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\\ \mathbf{0} & \mathbf{1}\end{bmatrix}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix}-\begin{bmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{bmatrix}\right)\\ &=\left(\mathbf{x}-\boldsymbol{\mu}_x-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)\right)^T\left(\boldsymbol{\Sigma}_{xx}-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}\right)^{-1}\\ &\quad\times\left(\mathbf{x}-\boldsymbol{\mu}_x-\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\mathbf{y}-\boldsymbol{\mu}_y)\right)+\left(\mathbf{y}-\boldsymbol{\mu}_y\right)^T\boldsymbol{\Sigma}_{yy}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}_y\right),\quad(2.52)\end{aligned}$$

这是两个二次项之和，因此可以进行拆分（指数+ - > ×），因此我们有:

$$\begin{gathered}p(x,y)=p(x|y)p(y)\\ p(x|y)=N(\mu_x+\Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y),\Sigma_{xx}-\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx})\\ p(y)=N(\mu_y,\Sigma_{yy})\end{gathered}$$

我们看到，因子p(x|y),p(y)都是高斯概率密度函数，更进一步：如果我们知道y的值(比如，其观测值)，那么我们可以解算出x的似然值，通过给定y值以及利用上式计算p(x|y)的值。

这也是高斯估计的重要部分：我们从先验概率x~N(μx，Σxx)入手，基于测量值ymeas来缩小(调整、逼近)。上式我们看到，均值改变、方差变小（p(x|y)，在测量值下，出现x状态的可能性，通过测量值ymeas来修正）。

2.4 Statisitically Independent , Uncorrelated

我们翻译为：统计学独立、不相关

高斯概率密度函数的情况下，统计学独立和不相关是等价的。如同上一节：我们设p(x,y)，若x，y独立，那么有:

$$\begin{gathered}p(x,y)=p(x)p(y)\\ then,\quad p(x|y)=p(x)=N(\mu_x,\Sigma_{xx})\\ then,\quad \Sigma_{xy}=0\\ then,\quad \Sigma_{xy}=E[(x-\mu_x)(y-\mu_y)^T]=E[xy^T]-E[x]E[y]^T\end{gathered}$$

那么，我们就得到了不相关成立的条件:

$$E[xy^T]=E[x]E[y]^T$$

2.5 Linear Change of Variables

我们翻译为：高斯分布随机变量的线性变换

假设，我们有服从高斯分布的随机变量x：

$$x \in R^N \sim N(\mu_x, \Sigma_{xx})$$

并且，我们还有另一个随机变量y∈R^M：

$$y = Gx$$
$$G \in R^{M\times N}$$

其中，G为常数矩阵。

我们想研究随机变量y的统计特性，那么**最简单的方法就是计算均值和方差：**

$$\mu_y = E[y] = E[Gx] = GE[x] = G\mu_x$$
$$\Sigma_{yy} = E[(y-\mu_y)(y-\mu_y)^T] = GE[(x-\mu_x)(x-\mu_x)^T]G^T = G\Sigma_{xx}G^T$$

因此，我们得到：

$$y \sim N(\mu_y, \Sigma_{yy}) = N(G\mu_x, G\Sigma_{xx}G^T)$$

另外一种方法，我们假设这个映射是**单射**，意思是两个x值不可能和同一个y值对应；我们通过假定一个更严格的条件来简化单射条件，即G是可逆的(因此M=N)，根据全概率公理：

$$\int_{-\infty}^{\infty} p(x)dx = 1$$

一个小区域内的x映射到y上，变为：

$$dy = |detG|dx$$

代入：

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} p(\mathbf{x})\, d\mathbf{x} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}_{xx}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_x)^T \boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{x}-\boldsymbol{\mu}_x)\right) d\mathbf{x} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^N \det \boldsymbol{\Sigma}_{xx}}} \\
&\qquad \times \exp\left(-\frac{1}{2}(\mathbf{G}^{-1}\mathbf{y}-\boldsymbol{\mu}_x)^T \boldsymbol{\Sigma}_{xx}^{-1}(\mathbf{G}^{-1}\mathbf{y}-\boldsymbol{\mu}_x)\right) |\det \mathbf{G}|^{-1}\, d\mathbf{y} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^N \det \mathbf{G} \det \boldsymbol{\Sigma}_{xx} \det \mathbf{G}^T}} \\
&\qquad \times \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{G}\boldsymbol{\mu}_x)^T \mathbf{G}^{-T} \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{G}^{-1}(\mathbf{y}-\mathbf{G}\boldsymbol{\mu}_x)\right) d\mathbf{y} \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{(2\pi)^N \det(\mathbf{G}\boldsymbol{\Sigma}_{xx}\mathbf{G}^T)}} \\
&\qquad \times \exp\left(-\frac{1}{2}(\mathbf{y}-\mathbf{G}\boldsymbol{\mu}_x)^T (\mathbf{G}\boldsymbol{\Sigma}_{xx}\mathbf{G}^T)^{-1}(\mathbf{y}-\mathbf{G}\boldsymbol{\mu}_x)\right) d\mathbf{y},
\end{aligned}$$

值得注意的是：若M<N，那么线性映射就不是单射了，我们无法通过定积分变量代换的方法，求得y的分布。

但是，若rank(G)=M，我们还是可以求。

2.6 Normalized Product of Gaussians

翻译为：高斯分布的归一化积

高斯概率密度函数一个很有用的性质：K个高斯概率密度函数的归一化积还是高斯概率密度函数。

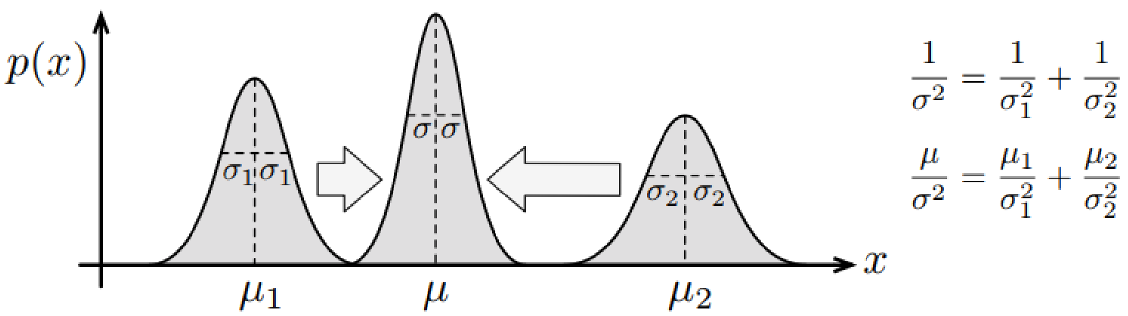
$$\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$
$$\equiv \eta\prod_{k=1}^{K}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right), \quad (2.68)$$

其中：

$$\boldsymbol{\Sigma}^{-1} = \sum_{k=1}^{K}\boldsymbol{\Sigma}_k^{-1}, \quad (2.69a)$$

$$\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \sum_{k=1}^{N}\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k, \quad (2.69b)$$

η是归一化常数，保证其满足概率公理。高斯概率密度函数在融合过程中会起到作用，如下图所示：



2.7 Sherman-Morrison-Woodbury Identity

Identity 译作 等式

Sherman-Morrison-Woodbury 恒等式有时合称为矩阵求逆引理。这个等式是从一个恒等式衍生出来的四个不同的等式。

对于可逆矩阵，我们可以将它分解为一个下三角——对角——上三角（LDU）形式，或上三角——对角——下三角（UDL）形式，如下所示：

$$
\begin{aligned}
&\begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{CA} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}+\mathbf{CAB} \end{bmatrix} \begin{bmatrix} \mathbf{1} & -\mathbf{AB} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \quad \text{(LDU)} \\
&= \begin{bmatrix} \mathbf{1} & -\mathbf{BD}^{-1} \\ \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{A}^{-1}+\mathbf{BD}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{D}^{-1}\mathbf{C} & \mathbf{1} \end{bmatrix}. \quad \text{(UDL)}
\end{aligned} \tag{2.72}
$$

我们对两侧取逆并且取等号，得到如下四个恒等式：

$$
\begin{aligned}
(\mathbf{A}^{-1}+\mathbf{BD}^{-1}\mathbf{C})^{-1} &\equiv \mathbf{A}-\mathbf{AB}(\mathbf{D}+\mathbf{CAB})^{-1}\mathbf{CA}, && \text{(2.75a)} \\
(\mathbf{D}+\mathbf{CAB})^{-1} &\equiv \mathbf{D}^{-1}-\mathbf{D}^{-1}\mathbf{C}(\mathbf{A}^{-1}+\mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1}, && \text{(2.75b)} \\
\mathbf{AB}(\mathbf{D}+\mathbf{CAB})^{-1} &\equiv (\mathbf{A}^{-1}+\mathbf{BD}^{-1}\mathbf{C})^{-1}\mathbf{BD}^{-1}, && \text{(2.75c)} \\
(\mathbf{D}+\mathbf{CAB})^{-1}\mathbf{CA} &\equiv \mathbf{D}^{-1}\mathbf{C}(\mathbf{A}^{-1}+\mathbf{BD}^{-1}\mathbf{C})^{-1}. && \text{(2.75d)}
\end{aligned}
$$

这四个恒等式我们可能在后续经常用到。

2.8 Passing a Gaussian through a Nonlinearity

译作：高斯分布随机变量的非线性变换

高斯分布经过一个随机非线性变换之后的情况：

$$
p(y)=\int_{-\infty}^{\infty} p(y|x)p(x)dx
$$

其中：

$$
\begin{aligned}
p(y|x) &= N(g(x),R) \\
p(x) &= N(\mu_x,\Sigma_{xx})
\end{aligned}
$$

这里：g(·)表示一个非线性映射：g:->y，被N(0,R)高斯分布干扰。我们经常需要这类随机非线性变换来对传感器建模。

2.8.1 Scalar Deterministic Case via Change of Variables

译作：标量情况下的非线性映射

一个简单的情况：x是一个标量，g(·)是一个非线性函数，且：

$$
x \sim N(0,\sigma^2)
$$

对于高斯概率密度函数：

$$
p(x)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2}\frac{x^2}{\sigma^2}}
$$

特殊例子：

$$
y=e^x
$$

反函数：

$$
x=ln(y)
$$

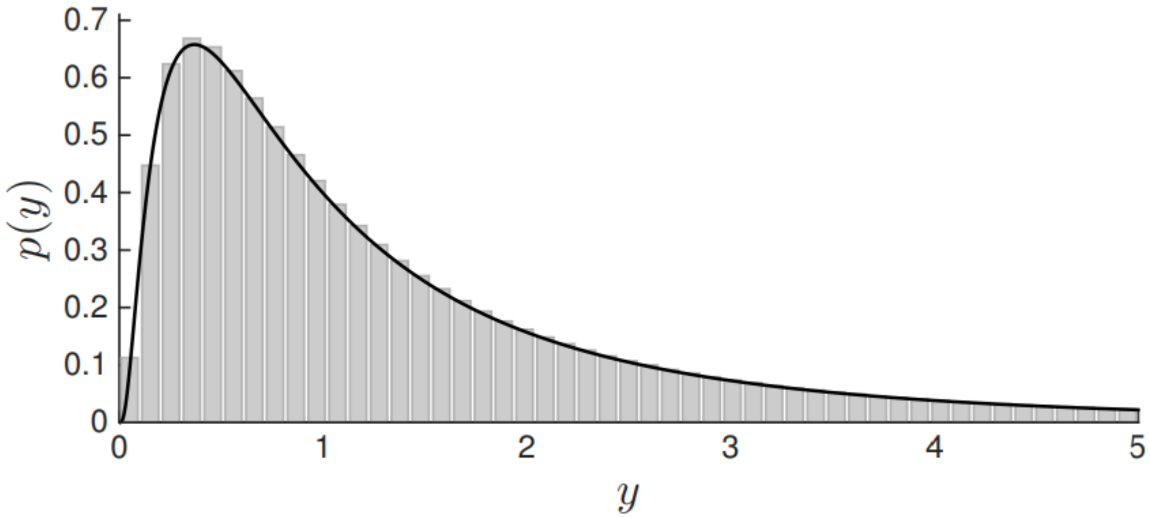
在很小区间我们有：

$$
\begin{gathered}
dy = e^x dx \\
or \\
dx = \frac{1}{y} dy
\end{gathered}
$$

根据上述推导：

$$
\begin{gathered}
1 = \int_{-\infty}^{\infty} p(x)dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{x^2}{\sigma^2}} dx \\
= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(ln(y))^2}{\sigma^2}} \frac{1}{y} dx = \int_{0}^{\infty} p(y)dy
\end{gathered}
$$

非线性变换后的概率密度函数如下图所示：



2.8.2 General Case via Linearization

我们引入**线性化**：

$$
\begin{gathered}
g(x) \approx \mu_y + G(x - \mu_x) \\
G = \frac{\partial g(x)}{\partial x}|_{x=\mu_x} \\
\mu_y = g(\mu_x)
\end{gathered}
$$

其中，G为g(·)的雅可比矩阵。

2.9 Shannon Information of a Gaussian

译作：高斯分布的香农信息、

在高斯概率密度情况下，我们有如下Shannon information：

$$
\begin{aligned}
H(\mathbf{x}) &= -\int_{-\infty}^{\infty} p(\mathbf{x}) \ln p(\mathbf{x})\, d\mathbf{x} \\
&= -\int_{-\infty}^{\infty} p(\mathbf{x}) \left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) - \ln \sqrt{(2\pi)^N \det \boldsymbol{\Sigma}} \right) d\mathbf{x} \\
&= \frac{1}{2} \ln \left((2\pi)^N \det \boldsymbol{\Sigma} \right) + \int_{-\infty}^{\infty} \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\, p(\mathbf{x})\, d\mathbf{x} \\
&= \frac{1}{2} \ln \left((2\pi)^N \det \boldsymbol{\Sigma} \right) + \frac{1}{2} E\left[(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right], \qquad (2.91)
\end{aligned}
$$

其中，我们用期望算子来表示第二项。实际上，第二项就是平方**马氏距离(Mahalanobis distance)**的期望值，与欧式距离差一个协方差权重。我们由：

$$x^T A x = tr(A x x^T)$$

可得：

$$(x-\mu)^T \Sigma^{-1} (x-\mu) = tr(\Sigma^{-1} (x-\mu)(x-\mu)^T)$$

2.10 Mutual Information of a Joint Gaussian PDF

译作：联合高斯概率密度函数的互信息

2.11 Cramer-Rao Lower Bound Applied to Gaussian PDFs

译作：高斯概率密度函数的克拉美罗下界

3、Gaussian Processes

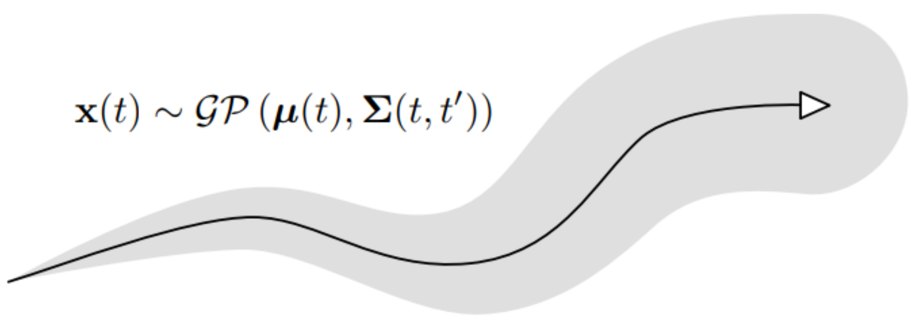
译作：高斯过程

我们将满足高斯分布的变量x∈R^N记为：

$$x \sim N(\mu, \Sigma)$$

我们会大量使用这类随机变量表达离散时间的状态量。接下来我们要着手讨论时间t上的连续的状态量。

首先，引入**高斯过程(Gaussian processes , GPs)**。下图描述了高斯过程表示的轨迹：



其中，其均值函数为黑色的实线，协方差函数为阴影区域。

我们认为整个轨迹是一个单独的随机变量，其属于一个函数集合。一个函数越接近均值函数，轨迹就越相似。协方差函数通过描述两个时刻t，t'的随机变量的相关性来刻画轨迹的平滑程度。我们把这个随机变量函数记为：

$$x(t) \sim GP(\mu(t),\ \Sigma(t,t'))$$

这表明了连续时间轨迹是一个高斯过程。实际上高斯过程不仅限于表达对于时间是一维的情况，但不需要考虑那么多。

4、习题

①