第二章 数据源

数据是事实或观察的结果，是信息的表现形式和载体，可以是符号、文字、数字、语音、图像、视频等。数据和信息是不可分离的，数据是信息的表达，信息是数据的内涵。数据本身没有意义，数据只有对实体行为产生影响时才成为信息。

数据可分为两类：软数据和硬数据。软数据是指人为产生的数据，尤其是在互联网普及的今天，互联网上人为产生大量的软数据，如身份信息、恐怖袭击信息、上网时间、点击率等信息，其特点是表格形式，形式较为单一，但干扰因素较大。软数据是数据分析、数据挖掘、数据科学等领域的研究对象。硬数据是指来自不同传感器的数据（信号），其主要特点是性质不同、实时性强、有噪声等干扰。硬数据是多传感器数据融合的研究对象，本书的数据以后不加特殊说明的话，就指的是来自传感器的测量数据。

2.1 数据源的特点

数据源是提供某种所需要数据的器件或原始媒体，这里主要是指提供观测数据的传感器。例如，在无人车系统中，常见传感器有摄像头、激光雷达、全球定位系统（GPS）、惯性测量单元（IMU）等。摄像头有单、双、三目，经过图像处理后能识别采集目标, 近几年随着深度学习的处理能力越来越强，已经能够实现较好目标的识别与分类效果。但摄像头对光照、天气等条件很敏感, 在采集条件恶劣的情况下，需要复杂的算法支持, 对处理器的计算能力要求也比较高。

激光雷达分单线和多线激光雷达，可以形成精确的 3D 地图, 抗干扰能力强，其缺点是购买成本较高。GPS 是一个中距离圆型轨道卫星导航系统。它可以为地球表面绝大部分地区（98%）提供准确的定位、测速和高精度的时间标准，可满足位于全球任何地方或近地空间的用户连续精确地确定三维位置、三维运动和时间的需要。最少只需 3 颗卫星，就能迅速确定用户端在地球上所处的位置及海拔高度；所能收联接到的卫星数越多，解码出来的位置就越精确。

IMU 包含一个三轴加速度传感器和一个三轴陀螺仪角速度计，前者可以测量一个三维空间的加速度，后者可以测量围绕三维空间三个坐标轴方向的旋转速度。使用三个加速度值，通过两次积分可以获得位移，以此实现相对位置信息，用角速度值积分可以获取姿态信息，结合在一起就可以获得物体的实际运动轨迹。但是，目前的低成本 IMU 采集数据有很大的漂移误差，这种积分运算给出的相对位置精度很低。

其他传感器还包括检测车内温度的温度传感器、检测废气中的成分用来修正喷油量的氧传感器、监测进气量及进气温度的空气流量计、监测机油压力机的油压力传感器、控制点火时间的霍尔传感器、监测发动机点火时间的爆震传感器等等。这些传感器对监控汽车的行驶状态非常重要，是目前的辅助驾驶技术主要部分，也是目前极具前景的无人驾驶系统最关键的技术之一。

本章来讨论传感器的普适性特征。我们发现，传感器的测量性能包含很多不确定因素，而这些因素对系统的测量性能、进而的融合结果将起到决定性的作用。我们要做的是，首先了解传感器包含哪些种类不确定因素，如何使用数学模型去描述它们，进而才能考虑如何在数据融合中克服这些传感器的不确定性，获得更高性能的融合结果。

源数据通常使用通信的模式进行汇集，进而再进行融合。数据在通信网络上是以数据

包为单位传输的，每个数据包中有表示数据信息和提供给数据路由的“帧”。这就是说，不管网络情况有多好，数据都不是以线性连续传输，中间总是有空洞的，这是一个离散的传输过程。打个比喻，就像我们的快递包裹一样，是一包一包的送过来的。由于物理线路故障、设备故障、病毒攻击、路由信息错误等原因，数据包的传输不可能百分之百的能够完成，总会有一定的损失，也就是所谓的丢包现象。就好像我们很倒霉，快递包裹被快递公司弄丢啦。除此之外，数据在传输过程中会有延时现象，由于传输经过的距离远或者一些故障，或者网络繁忙，导致传输并没有准时达到目的的。这种情况就好比双十一的时候包裹太多，快递公司在运送我们的包裹的时候就不那么及时，耗时要更长一些。如何保证我们传输的数据包快速、准确、无误的到达是通信领域的研究问题，我们在这里不做过多的描述。我们关心的是，当数据到达之后我们如何来处理这些数据，如何在考虑传感器本身的测量误差情况下进行融合，以获得更高质量、更高层次的信息。

多传感器信息融合技术中，传感器的测量特性直接决定了所要采用的融合方法、甚至直接决定了融合结果。很明显，如果传感器的测量性能很好，获得的测量数据和实际的真实情况完全一致的话，融合方法只需考虑传感器测量数据之间的相互关系，如性质上面是否互补、是否覆盖同一空间、传感器的采集时间是什么关系等问题，将不同传感器的测量数据融合在一起即可。

但是如果传感器测量存有误差，我们称其为传感器测量的不确定性。那么，我们就必须要考虑由于传感器的测量引入的误差，才能到使用这些含有误差的数据进行融合后获得反应实际真实情况的信息。我们必须知道由于传感器的测量性能导致的测量数据和实际真实状态有什么不同，然后克服这些不同再进行数据融合。下面我们就对传感器的测量不确定性种类进行分析，并初步探讨如何来克服这些不确定性。

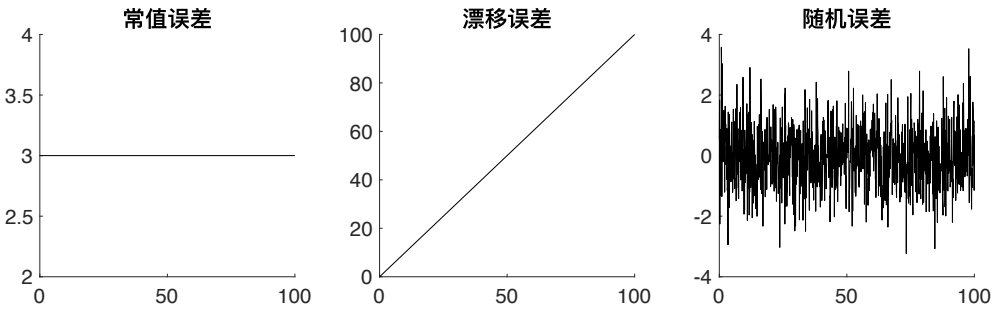
传感器的测量特性容易受到外界的干扰，有时还可能产生较大的测量误差，因此没有一种传感器可以保证在任何时候都能提供完全准确的信息。传感器的测量误差主要有三种：常值误差、漂移误差和随机噪声误差，如图 2.1 所示。如果被测对象的真实数据如图 2.2 所示，那么因为这三种误差，会导致测量的结果如图 2.3 所示。

其中，常值误差指的是传感器的测量数据和真实状态相比，始终存在一个数据上的差别，这个差别是一个常数，有可能是正值也有可能是负值。比如我们在使用家用温度计的时候，由于我们的读数习惯，可能会导致我们的读数始终存在一个这样的偏差。检测设备和电路等安装、布置、调整不当会引起常值误差，如仪表盘刻度不准确就会造成这样的常值误差。

漂移误差指的是随着时间的推移误差会变得越来越大。这种情况是由于我们目前使用的传感器都是电子设备，传感器的运行时间对其电子特性影响非常大。随着使用时间的增加，传感器会发热进一步导致其电磁特性发生改变，甚至连阻值都会发生持续的变化，而导致测量的数据与真实状态相比偏差越来越大。还有一种情况是由于使用环境的改变导致偏差发生不确定的变化，如 IMU 的偏差会随着运动的速度而改变，也会随着使用时间的加长而增加。

常值误差和漂移误差通常可以通过传感器的标定进行消除，其中，漂移误差的消除会更难一些。但是如果我们通过细致的传感器标定工作，对漂移误差进行较为准确的建模，漂移误差也可以大幅度的削减。

相比之下，随机噪声误差对系统的影响会更大。随机噪声误差可能由传感器电子电路中的不确定因素、环境的随机变化等原因导致。其特点是没有确定的误差值，因此很难进行消除。所以我们在研究多传感器融合的时候，在很多时候都要考虑这类随机噪声误差。幸运的是，随机过程这类数学基础可以描述这些噪声的特性。通常我们假设传感器的随机噪声误差满足高斯白噪声的特性，同时具有已知的均值和方差。



(a) (b) (c)

图 2.1 传感器的测量偏差的类型 (a)常值误差；(b)漂移误差；(c)随机噪声误差

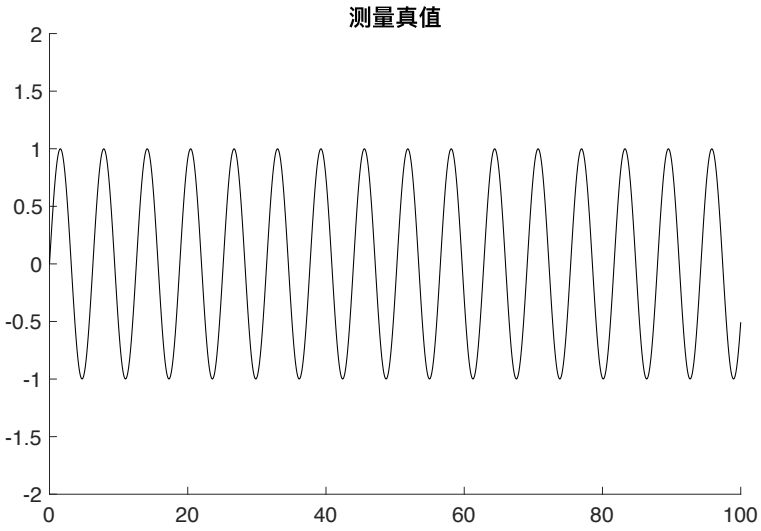


图 2.2 实际状态的真实值

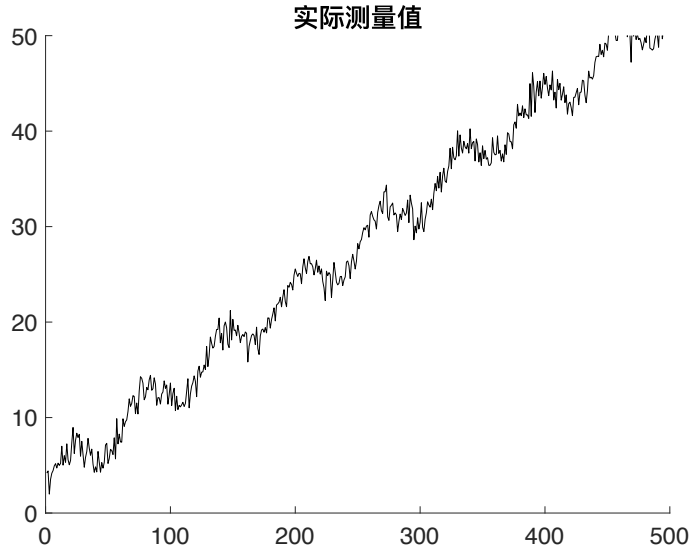


图 2.3 实际获得的传感器测量值

2.2 数据的预处理方法

对于软数据来说，数据预处理也可以称为数据清洗。对于分布在不同平台的相同或不同类型传感器，在对其观测数据进行数据融合前，由于其所在位置各不相同，所选的观测坐标系不一样，加上传感器的采样频率也有很大差别， 因此即使是对同一个目标进行观测，各传感器得到的目标观测数据也会有很大的差别，所以，在进行多传感器数据融合时，首先要做的工作就是统一来自不同平台的多传感器，我们成其为对（硬）数据的预处理 [1]。

例如图 2.4 所示，横轴为 N 个传感器，纵轴为 N 个传感器对同一个目标的测量值。已知被测量的真值为恒值，如果存在一个与真值相差很大的值（奇异值），我们称其为测量数据“不一致”的情况。需要对这个值进行预处理——如直接将其去除，再用其他值进行融合就可以得到更准确地结果。如何衡量多个测量值与真实值的差别，然后将奇异值去掉，是多传感器融合数据预处理的主要问题。

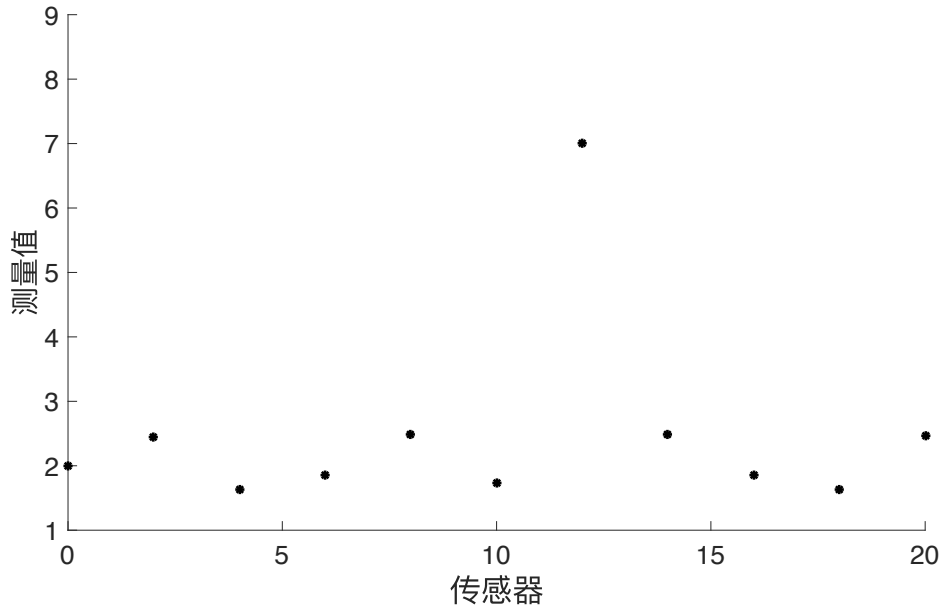


图 2.4 N 个传感器对同一目标的测量值

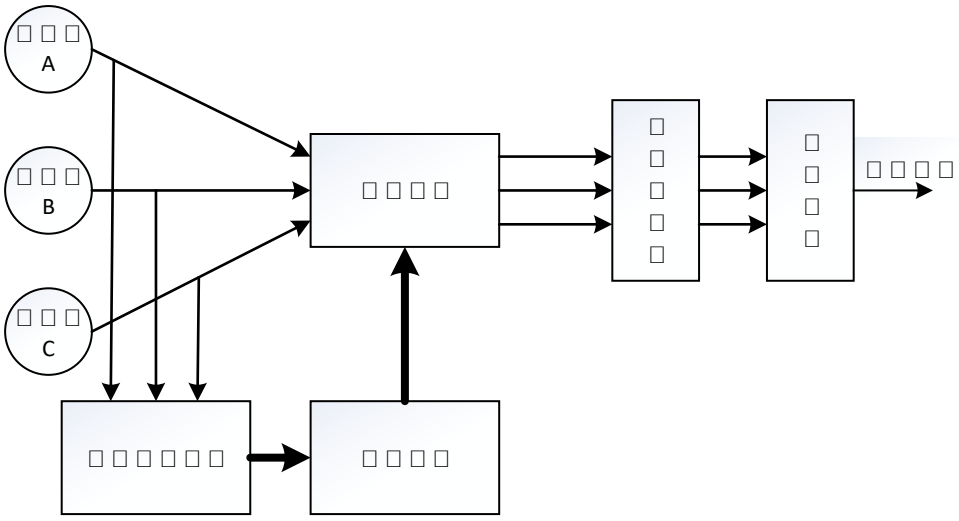


图 2.5 带有传感器数据预处理的融合方法思路

很多研究涉及到如何对传感器数据科学预处理的问题。如肖湘宁[2]提出了对多传感器测量数据进行一致性检验的方法，该方法通过比较置信概率距离和阈值，分析传感器之间的支持程度，建立传感器关系矩阵，剔除处于较大误差状态或失效的传感器信息，但是文献并未给出最大一致传感器组的求取方法。陈文杰[3]对改置信距离进行了改进，采用迭代法，将两个置信距离最近的传感器合并在一组，计算新组的置信距离，按照以上步骤继续合并，直到找出最大一致传感器组。谢运祥[4]采用 2σ 信任度函数来表示信任度，定义置信距离。计算信任度矩阵，根据阈值，画出传感器组的双向连接图，取双向连接节点对应的传感器组为最大一致传感器组。周林[5]在[4]的基础上，得到一致性关系矩阵，按照节点个数由多到少的次序，依次判定子图中的传感器是否满足一致性。

虽然以上文献都具有很好的可行性，但是也有一些不足之处，[1-4]中最大一致传感器组判断的运算量大，时间复杂度高，执行效率低，适用于传感器个数少的情况。盛碧琦等人[6]采用对称的传感器置信距离，将有向图转换成无向图，将关系矩阵转换成对称矩阵，将无向图转化为邻接矩阵，再通过分析矩阵元素判断最大一致传感器组。该方法内存开销小，检测精度高，运算时间短。

带有传感器数据预处理的融合方法思路如图 2.5 所示。我们可以使用置信距离和置信距离矩阵来判断传感器测量数据的一致性。在一个由 N 个传感器构成的多传感器系统中，用高斯概率密度函数 $p_i\left(x\right)$ 描述其测量模型，使用置信距离表示两个传感器间的相互支持性，用于比较测量数据的一致性。具体过程如下：

通常情况下，利用多个传感器测量某参数的过程中包含两个随机变量，被测参数 μ 和每个传感器的输出 $X_i, i=1,2,\cdots,m$，一般认为它们服从正态分别，用 x_i 表示第 i 个测量值的一次测量输出，它是随机变量 X_i 的一次取样。设

$$\begin{aligned} \mu &\sim N\left(\mu_0, \sigma_0^2\right) \\ X_k &\sim N\left(\mu, \sigma_k^2\right) \end{aligned} \tag{2-1}$$

为对传感器输出数据进行选择，必须对其可靠性进行估计，下面我们定义各数据间的置信距离。用 X_i、X_j 表示第 i 个和第 j 个传感器的输出，则其一次读数 x_i 和 x_j 之间的置信距离为：

$$\begin{aligned} d_{ij} &= 2\int_{x_i}^{x_j} p_i\left(x \mid x_i\right)dx \\ d_{ji} &= 2\int_{x_j}^{x_i} p_j\left(x \mid x_j\right)dx \end{aligned} \tag{2-2}$$

置信距离表示两个传感器 i 和 j 之间的相互支持性。注意到我们发现这两个距离的值可能会不一致，这是这种度量方法的一个缺陷。

已知 x_i, x_j 服从正态分布，则上式中：

$$p_i\left(x \mid x_i\right)=\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left\{-\frac{1}{2}\left(\frac{x-x_i}{\sigma_i}\right)^2\right\}$$
$$p_j\left(x \mid x_j\right)=\frac{1}{\sqrt{2\pi}\sigma_j}\exp\left\{-\frac{1}{2}\left(\frac{x-x_j}{\sigma_j}\right)^2\right\} \quad (2\text{-}3)$$

我们有当 $x_i = x_j$时， $d_{ij} = d_{ji} = 0$，当 $x_i >> x_j$或 $x_j >> x_i$时， $d_{ij} = d_{ji} = 1$。

那么置信距离矩阵表示为：对 m 个传感器的一次测量数据，利用上述方法可以分别计算任意两个传感器数据之间的置信距离 d_{ij} 和 d_{ji}，其中 $i = 1,2,\cdots,m$， $j = 1,2,\cdots,m$。利用置信距离 d_{ij} 和 d_{ji} 得到如下 $m \times m$ 矩阵

$$D_m=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m1} & d_{m2} & \cdots & d_{mm} \end{bmatrix} \quad (2\text{-}4)$$

根据具体问题选择合适的阈值 β_{ij}对数据的可靠性进行判定，使用下式将上式中的数据置换成 1 和 0。

$$r_{ij}=\begin{cases} 1 & d_{ij} \le \beta_{ij} \\ 0 & d_{ij} > \beta_{ij} \end{cases} \quad (2\text{-}5)$$

将所有的 r_{ij} 组合在一起，构成如下关系矩阵

$$R_m=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mm} \end{bmatrix} \quad (2\text{-}6)$$

根据关系矩阵 R_n，可用有向图表示传感器间的支持关系。对于两传感器共存在两种支持关系: 当 r_{12}=r_{21}=0时，表明传感器 1 与传感器 2 无支持关系；当 r_{12}=r_{21}=1时，表明传感器 1 与传感器 2 相互强支持。找出传感器系统中所有相互强支持的传感器，即为最大一致传感器组，从而获得具有一致性的传感器测量数据。

在获得置信距离与置信距离矩阵，选择 l 个数据作为最佳融合数之后，可以得到融合结果 $\hat{\mu}$ 为：

$$\hat{\mu}=\frac{\sum_{k=1}^{l}\frac{x_k}{\sigma_k^2}+\frac{\mu_0}{\sigma_0^2}}{\sum_{k=1}^{l}\frac{1}{\sigma_k^2}+\frac{1}{\sigma_0^2}} \tag{2-7}$$

其中被测参数满足 $\mu\sim N\left(\mu_0,\sigma_0^2\right)$，而第 k 个传感器的测量数据为 $X_k\sim N\left(\mu,\sigma_k^2\right)$。

基于一致性估计的数据融合算法可以总结为：

（1） 计算 m 个传感器数据的置信距离矩阵，为简化计算，当测试数据服从正态分布时可利用误差函数计算置信距离。

$$d_{ij}=erf\left(\frac{x_j-x_i}{\sqrt{2}\sigma_i}\right)$$

$$erf\left(\theta\right)=\frac{2}{\pi}\int_0^{\theta}e^{-u^2}du$$

（2） 选择合适的距离临界值，由置信距离矩阵产生关系矩阵，表示第 j 个传感器对第 i 个传感器的支持。

$$\mathrm{r}=\begin{cases}1 & d_{ij}\le\beta_{ij}\\ 0 & d_{ij}>\beta_{ij}\end{cases}$$

（3） 构造关系矩阵（2-6）对多传感器数据进行选择，产生最佳融合数。原则是如果一个传感器被一组传感器(所谓的一组传感器也需要设定一个阈值来描述，我们称其为支持传感器个数 K）支持，则它的读数是有效的。否则它的读数无效，在融合中不考虑。

（4） 将 μ_0、σ_0^2和最佳融合数对应的 x_k、σ_k^2代入融合估计公式（2-7）求的参数估计值。

下面我们给出一道例题，来说明如何实现上述方法。我们使用 Matlab 语言对这道例题进行了编程，具体的程序见本章的附录。

例 2.1 利用 8 个传感器对一个恒温槽的温度进行测量，已知恒温槽温度满足正态分布，其中 $\mu_0=850.50℃$，$\sigma_0^2=4.5025$，8 个传感器的测量结果及测量方差如下：

传感器编号	1	2	3	4	5	6	7	8
测量方差	25.73	23.81	24.95	25.75	35.65	21.33	23.94	22.96
测量值	848.1	850.5	851.9	849.9	854.6	849.3	848.0	848.3

注意：这是每次测量的实际真实值的概率分布，它和测量的真实值是不同的。

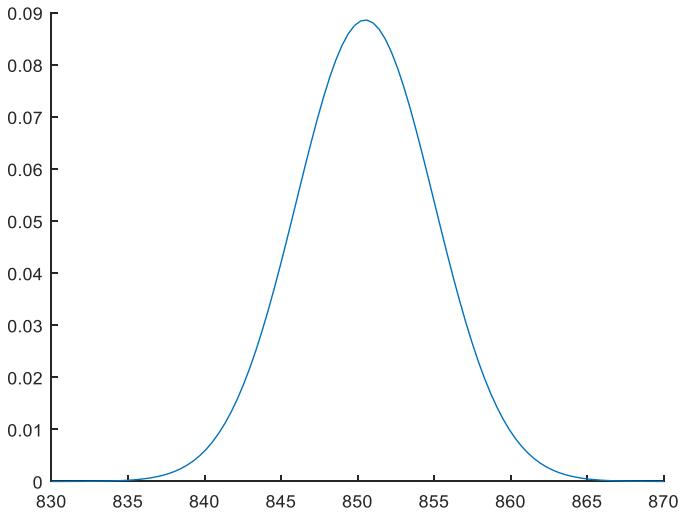


图 2.6 恒温槽温度的分布

实际温度的概率分布如图 2.6 所示，可以看出温度的变化服从正态分布，大多数时间的温度在 850.50℃左右，也有少数情况会低到 840℃或高到 860℃。

按照本章前述的方法我们编写了 Matlab 程序见本章附录。当置信距离的阈值取为 0.4、关系矩阵支持传感器个数 K 取为 5 时，得到的融合估计结果为 849.7129℃。

下面研究一下置信距离的阈值和融合结果的关系。将阈值取为 0.01、0.02、0.03、……、1 共 100 个值，分别考察融合结果的变化，如图 2.7 所示。我们看到，阈值的变化对融合结果有很大的影响，融合结果的值随着阈值的变化而变化。当阈值比较小的时候，如小于 0.22 时，所有的传感器都没有入选参与融合机制，融合的结果只是恒温槽温度的期望值，测量数据没有对融合结果产生任何价值，所有的测量值都被抛弃了。

另外当阈值大于 0.72 时，所有的测量数据都被认为很彼此接近，也就是说所有传感器都参与了融合，因此融合结果保持在了一个不变的值。而当阈值在 0.4-0.5 时，由于最终根据阈值所选择的传感器数量不同，所以融合结果也就发生了改变。

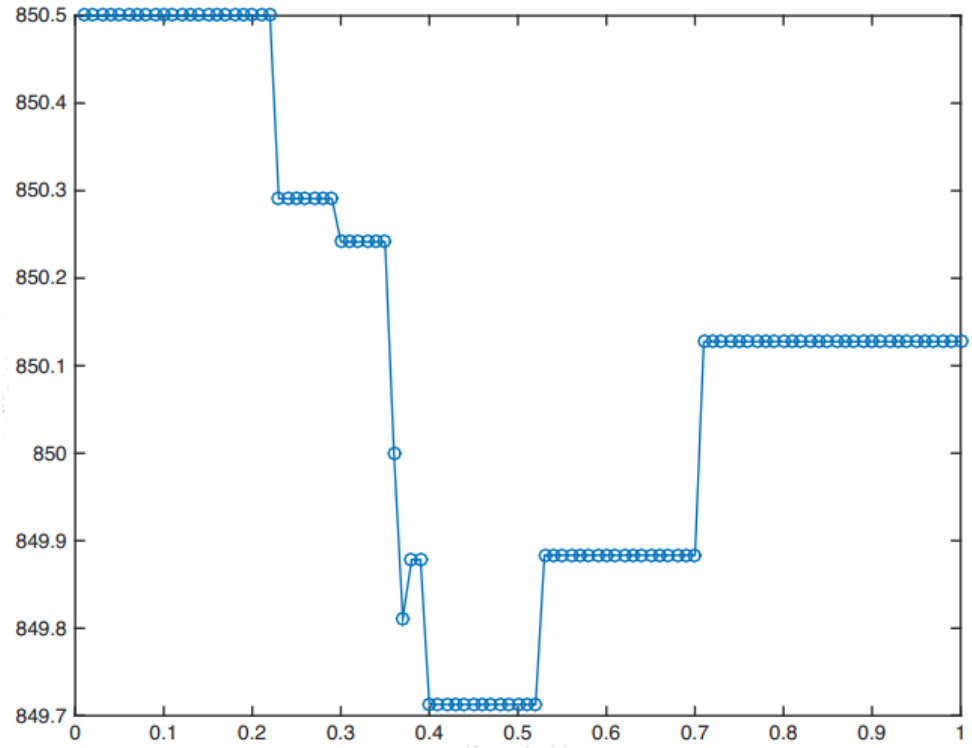


图 2.7 支持传感器个数为 5 时，不同置信距离阈值的融合结果

2.3 小结

本章主要介绍了数据源的特点和数据预处理的方法。在多传感器数据融合的研究中，硬数据是主要的研究对象，指来自不同传感器的数据信号，其主要特点是性质不同、实时性强、有噪声等干扰。因此，我们需要对数据进行预处理，以便在融合中获得更好的效果。数据预处理的方法是置信距离、置信距离矩阵和关系矩阵。通过给定阈值的方法，衡量两个不同传感器数据间的距离以获得有效的传感器组合，并最终用于融合获得最佳结果。

参考文献

[1] 周炳玉, 卢野, 刘珍阳. 多传感器数据融合中的数据预处理技术研究[J]. 红外与激光工程, 2007, 36(z2):246-249.

[2] 肖湘宁. 电能质量分析与控制［M］. 中国电力出版社，2010: 175－180.

[3] 陈文杰. 基于瞬时无功功率理论的三相电路谐波和无功电流检测［J］. 机电信息，2013(3) : 140－142.

[4] 谢运祥. 电力有源滤波器及其应用技术的发展［J］. 电工技术杂志，2000(4) : 1－3.

[5] 周林． 基于小波变换的谐波测量方法综述［J］. 电工技术学报，2006，21 (9) : 67－74.

[6] 盛碧琦, 应忠于, 胡云琴, 等. 基于改进置信距离的多传感器一致性校验[J]. 工业

仪表与自动化装置, 2014(2):102-104.

附录

```
measurements=[848.1,850.5,851.9,849.9,854.6,849.3,848.0,848.3]';
covv=[25.73,23.81,24.95,25.75,35.65,21.33,23.94,22.96]';
RealValue=850.5;
RealCov=4.5025;
for i=1:8
    for j=1:8
    d(i,j)=erf((measurements(j)-measurements(i))/(sqrt(2)*covv(i)));
    end
end
Threshold=0.4;
R=(abs(d)-Threshold*ones(8))<0
SupportNumber=5;
support=(sum(R,2)>SupportNumber)
FusionData=(support'*(measurements./covv)+RealValue/RealCov)/(support'*(1./covv)+1/RealCo
v)
x = linspace(830,870,100);
y = normpdf(x,RealValue,RealCov);
plot(x,y)
```