卡尔曼滤波器和优化的本质理解 - ziliwangmoe的博客 - CSDN博客

非线性最小二乘优化中，最重要的是那个雅克比矩阵。每一行对应一个观察两，里面的值代表要减少这个观察量的误差，其他被优化的量需要做的改变。而每一列就是这个观察量需要怎么变才能减少不同的观察量。对于一个优化量，对有的观察量需要增加值，有的需要减少。那到底是要减少还是增加呢？就把所有观察量对应的变化求个平均，得到这个迭代优化量的跟新值。

如果所有观察对应的更新量都是一致的，说明观察量质量高，反之亦然。

换个角度想，即使只有一行，我们仍然可以算出一个跟新量。所以我们有两个选择，一个是用所有观察量结算出一个更新量，还是每次用一个观察算出一个更新量，跟新变量，再用第二个观察量算出一个更新量，然后。。。这两种方式就是全局优化和滤波器的本质不同。

而滤波器为什么很在意不确定度，因为不确定度给出了两次观察的重要性的比值。就还比求均值，我们是把所有东西加一起了，除以个数。还是每次先把一个值除以一个系数再累加上去。当你用第一种方法的时候，你不关系某个数据对最终的结果的总要性。

==========

假设我们有两个未知量想要知道他们的值。最直接的方法是找到两个和这两个未知量相关的方程，求解方程组就能得到他们的值。

但如果我们只能得到一个方程呢？也许你会说这个问题无解。但是换个角度想，虽然只有一个方程，但也不没有好，至少我们还是多了一些关于这两个位置量的信息。

很多情况下，我们不能一下得到足够的关于所有未知量的方程，但随着时间的推进，我们能不断地得到越来越多的方程（信息）。再往后面，我们不仅能得到足够的信息，还能得到更多的信息。所以我们想要有一种方法能够在信息不足的时候，尽可能的缩小对位置量的不确定性（未知量可能的取值范围），在信息有余的时候，还能不断的优化对未知量的判断。甚至在位置量发生变化的时候，我们还能通过新的信息来跟新未知量。这就是卡尔曼滤波的背景理解了。

一次使用绝对足够，甚至多余的方程来求解的方法就是优化发。而不断的融合新的信息的方法就会滤波法。

假设我们有x，y两个未知量。

有关于他们的两个方程：

$x+2y=0$

$x-y=1$

我们可以得出$x=2/3$，$y=-1/3$

但如果我们在第一个时刻只知道$x+2y=0$这个信息呢？虽然不能求出x，y的具体值，但是我们大概知道x，y的一个关系。画在一个二维的空间中，就是一条直线。如果我们还对这个关系不太确定，这条线就会变成一个带状物。

换个说法，一开始我们对x，y一无所知。所以这个x,y组成的二维空间上，任何一点都是有可能的。当我们得到第一条消息$x+2y=0$后，我们对x,y的不确定度缩小到那个带状区域了。当又收到$x-y=1$这个消息的时候，我们对x,y

的不确定度缩小到一个点状物。如果还有更多的消息来，这个不确定度区域还能不断缩小。

基于这个理解还可以思考下无偏估计的原理：我们希望这些消息对应的不确定区域都是有一个共同重叠区的，这样才能对真值有越来越稳定的估计，随着获得的信息越来越多。

再举一个计算机视觉的例子。假设摄像头只观测到一个空间位置已知的点，我们没法只用这一个点确定摄像机的位置。但是我们对摄像机未知的不确定度，已经从之前的6d空间均匀分布，缩小到一个用高斯分布表示的区域。这个区域由相机pose这个随机变量的均值和方差同时决定。但我们观察到越来越多的点后，pose的不确定区域也会越来越小。但有一点很重要，方差和均值同时确定了pose的可能取值。这也是为什么对于滤波法，我们一般只需要把协方差的初值取得足够大就行了。

这里还有一点需要注意：如果我们想要增量的融合信息，我们必须知道未知量的协方差，不然我们没有办法去融合新的信息。但最终有用的信息只有均值。这也是为什么对于一次性使用所有信息进行的优化法，就不需要关注最终结果的协方差了，虽然这个协方差是存在，也能求出来。

再来想想卡尔曼滤波器的场景：一开始我们对相机pose一无所知，当进行了第一次观察后，pose的不确定区域有所缩小。然后通过一次运动方程变换。pose的总体不确定区域虽然没有变小，但是形状和位置发生了变化。然后再进行一次观察，不确定区域又缩小一些。就这样我们对pose的变化规律越来越确定。但因为每次观察和运动都会带来额外的不确定度（加噪音）。所以最终观察带来的不确定性减少和额外的不确定性增加达到平衡状态。

还可以换一种理解角度。假设不同时刻的pose都是不同的随机变量。但和第一个pose不同，在我们用观察值来跟新这个pose的时候，他就已经有一个先验的概率分布了，而不是一无所知。而这个先验概率正好是由上一个时刻的pose通过运动方程推算过来的。这样就是卡尔曼滤波的问题划分为多个最普通的高斯概率合并的问题。(这里不是边缘化，后面会讲区别)

y=a*x+n中，x和y都是变量，没有不确定度一说，而不是随机变量。而这个式子代表的事y和x的一个联合概率分布。只是一个联合随机变量的两种不同表示。而n其实是p(y|x)，大概是这个意思，不太严格。

obs=a*x中，x和obs都有自己的不确定度。代表随机变量的变换。

b-a*x~n，只有x是变量，这个表示的是X这个随机变量满足的方程

卡尔曼滤波问题中：运动方程其实属于第二种意义，从一个随机变量变换到另外一个随机变量，再加上一个随机变量。

观察方程 其实属于第二种意义，观察值是已知的，我们需要求能让观察值概率最大的那个x的值。

而融合不是讲把一个随机变量变换为另一个随机变量，而是把一个随机变量的两种分布相乘，得到这个随机变量的第三种分布。上面三种表达都和融合无关。

所以他们讲的不同的东西，不用强行把他们联系在一起。运动方程通过原理2得到一个关于x的分布，观察方程通过原理3得到关于x的另外一个分布。然后通过融合x的这两种分布，得到x的第三种分布。为什么要用相乘的方式来融合呢？因为两个分布同时发生的概率分布就是这两个分布的乘积。

关于这里提到的一个随机变量的不同分布可以这样理解：不同分布其实对应的还是不同的随机变量，这是这两个随机变量的取值一模一样，然后问他们同时取相同值得概率分布是什么。而上面原理2的两个随机变量的取值可以是完全不同的，就算相同，要说他们取相同值的概率也是么有太大意义的。

滤波法和优化法的等价理解

这两个方法都在解决的问题是：如果一个随机变量的概率分布可以用一个自变量x来表示。求x取值为什么的时候，能够让概率最大。这本来就是个

优化问题，优化法直接进行求解。而滤波法是把x和一个正太分布之间用一个变换矩阵联系起来。先求变换矩阵的表示，然后由变换矩阵反推x的分布。最后取出平均值。这样做的前提是x也是正态分布，才能用一个矩阵来表示变换。优化法没有这个限制。当x不是高斯分布的时候，只有对这个变换进行线性展开，这里就出现了误差。所以当x不是正太分布的时候，两者就出现了差异。

另外当目标分布是标准正太分布的时候，优化法中的模就是普通的二次模而不是协方差的模。但一般来说观察的误差值之间都是独立同分布的，其实就是标准正太分布了。

在EKF中，虽然我们总把pose看做一个正态分布。但是旋转或者通过观察得到的pose都不再是正态分布。所以这里每个时刻的pose的正太分布表示都是近似表示。

总结一下就是：让误差最小的那个普通变量x的取值对应让概率函数取最大的那个随机变量x的取值。这里的的概率分布是指随机变量x通过一定变换后（包含了求误差的变换）得到的那个正太分布的概率分布。而这里的误差是指普通变量x通过某个变化后和应该取得的值之间的差异。

先验概率和边缘化的理解

卡尔曼滤波和贝叶斯定律虽然都是讲概率的，而且过程都是之前和之后的关系。但严格的说卡尔曼滤波没有用到任何贝叶斯的东西。

如果硬要分析他们的相关性，贝叶斯是和融合相关的：一个随机变量x本来的分布是先验分布，然后和一个特殊条件下的条件分布融合，最后求两个分布相乘后的分布。相乘就是融合，所以说先验概率就是在做融合。这里不是上面的原理3，所以和优化没关系。

实际情况中，条件概率会有多个：t时刻的pose基于t-1, t-2等时刻的条件概率。这些概率都是通过融合（相乘）在一起，然后最终得到的那个概率分布又要满足某个概率分布，然后问里面各个变量的取值。融合后的概率分布函数通过log后。里面各个相乘的项就变成了相加的项。这也是为先验项

在优化问题里面，就变成了加一个误差项。这也是为啥我们在融合不同数据的时候，能够简单的把这些误差项加起来。这个就是因子。

但是因为是不同的观察量，所以虽然是独立的，但是分布就不一样了，不同的观察量的方差不一样。所以需要在每个因子前面都要乘以一个超参系数。

如果这么看的话他们都是在做融合。但这里我们先抛开贝叶斯，因为卡尔曼滤波的整个推导并没有用到贝叶斯。所以要注意先验概率和卡尔曼滤波里的被更新前的状态没有关系，注意区分他们。

当状态只包含一个pose的时候，事情很简单。直接把前一个时刻的分布变换后给后一个时刻，再在这个基础上更新。但如果硬要使用优化法就会出现一个问题：优化法假设处理的变量是系统中所有的变量，但其实又只优化了当前帧。但明显过去的pose对当前pose肯定是有限制作用的。虽然不管历史pose，情况好也能优化出很好的结果。这就相当于滤波法里面跟新前的状态很差，但是更新非常有精确。滤波法把历史信息用起来是很自然的，但优化法就需要把历史限制作为一个误差项来利用历史信息了。这个也很好理解，卡尔曼滤波其把更新前的概率乘以更新的概率。那么更新前的概率自然就成为一个误差项了。

另外还有种做法是把状态的先验信息当做非线性优化的初值来用。这个其实是说不通的，因为如果优化很完美的话，不管初值是什么，都会达到相同的值。也就是结果和先验无关。但是因为问题的非线性性，初值的好坏对结果影响很大。而用先验来做初值不失为一个好方法。重点的差异是，这种做法并没有把先验信息加入到优化的constrain里面。

马尔科夫假设

对于多个随机变量组成的联合概率密度函数（x1,x2,x3等），这里一般指不同时刻的状态变量。可以用他们之间的条件概率函数以很多不同的组合相乘表示出来。

马尔科夫假设指:

$p(x3,x2,x1)=p(x3|x2x1)p(x2|x1)p(x1)=>p(x3,x2,x1)=p(x3|x2)p(x2|x1)p(x1)$

因为有：$p(x3|x2x1)=p(x3|x2)$

$p(x3,x2,x1)$是我们最终要求得东西，虽然我们一般只用x3,x2,x1的均值。而这个被求的东西可以分解为多个两两耦合的因子。

正是有马尔科夫假设，我们才能像卡尔曼滤波器那样把上个时刻的概率变换后直接当做当前时刻的概率。也就是x3的分布由x2唯一决定。

文章最后发布于: 2018-10-04 17:02:21