MLAPP-C4

4.高斯模型

4.1 引言

4.1.1 符号表达

4.1.2 基础知识

4.1.3 MVN的极大似然解MLE

4.1.4 高斯分布的最大熵性质推导

4.2 高斯判别分析

4.2.1 二次判别分析

4.2.2 线性判别分析

4.2.3 二分类LDA

4.2.4 判别分析的MLE

4.2.5 克服过拟合的策略

4.2.6 含正则项的LDA\*

4.2.7 对角LDA

4.2.8 最近收缩质心分类器\*

4.3 联合高斯分布中的推理问题

4.3.1 相关结论

4.3.2 例子

4.3.3 MVN的信息表达形式

4.3.4 结果证明\*

4.4 线性高斯系统

4.4.1 结果陈述

4.4.2 例子

4.4.3 结果的证明\*

4.5 拓展：威舍特分布**\***

4.5.1 逆威舍特分布

4.5.2 威舍特分布的可视化\*

4.6 MVN的参数推理

4.6.1 参数的后验分布

4.6.2 参数的后验分布\*

4.6.3 参数和的后验分布\*

4.6.4 未知精度下的传感器融合\*

4.高斯模型

4.1 引言

本章我们将介绍**多变量高斯**（multivariate Gaussian）或者**多变量正态**（multivariate normal，MVN）分布，在连续随机变量领域中，它经常被用来建立联合概率密度函数。该分布是我们在后文中介绍的很多模型的基础。

不幸的是，本章所需要的数学知识水平要比很多其他章节的高。特别是在线性代数和矩阵计算方面，本章将严重依赖这些知识。为了处理高维数据，这些知识都是必须的。初学者可以选择跳过那些加\*的章节。除此之外，因为本章有很多公式，我们会将那些特别重要的公式用方框进行标注。

4.1.1 符号表达

首先让我们简单介绍一些符号书写方面的规范。我们使用小写的加粗字母表示向量，比如**x**。使用大写的加粗字母表示矩阵，比如**X**。我们使用大写非加粗字母表示矩阵中的元素，比如*Xij*。

所有的向量在没有特殊说明的情况下都是**列**向量。我们使用表示由*D*个标量组成的**列**向量。类似地，符号的等号左边表示一个特别高的列向量，由每个向量延纵向叠加而成。这种方式还有一种替代写法，但这种写法比较丑陋。符号的等式左边代表一个矩阵，由每个向量延横向叠加而成。

4.1.2 基础知识

根据章节2.5.2的内容，我们知道在*D*维空间中MVN的概率密度函数定义为：



（4.1）

上式中指数项中表达式为向量**x**和期望向量之间的马氏距离。通过将矩阵进行**特征分解**（eigendecomposition），我们可以对马氏距离有更深的理解。的分解式可以写成，其中为特征向量的标准正交矩阵，满足，为对应特征值的构成的对角矩阵。

使用的分解式，我们有：

 （4.2）

其中表示矩阵**U**的第*i*列，表示第*i*个特征向量。所以我们可以重写马氏距离为：

 （4.3）

 （4.4）

其中。回忆一下在2维空间中椭圆的曲线：

 （4.5）

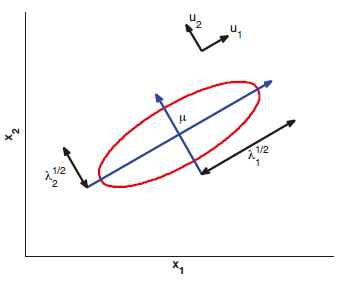


图4.1 2维高斯密度的可视化。椭圆的长短轴分别由协方差矩阵的前两个特征向量（对应的特征值分别为最大和次最大）决定，分别为**u**1和**u**2。

所以，我们发现对于一个多维高斯分布而言，其概率密度等高线的轮廓线是一个椭圆，如图4.1所示。特征向量决定了这个椭圆的方向，特征值决定了对应方向曲线延长的程度。

一般情况下，我们将马氏距离等价为一个新的坐标系中的欧式距离，这个新的坐标系可以通过将现有坐标系（即图4.1中的*x*-*y*坐标系）按方向平移，按**U**旋转得到**。**

4.1.3 MVN的极大似然解MLE

接下来，我们讨论如何使用MLE对MVN中的参数进行估计。在后面的章节中，我们将会讨论使用贝叶斯方法对参数进行推断，这种方法可以缓解过拟合，并且可以用来量化参数估计值的不确定度。

**定理4.1.1**（高斯分布的MLE）。如果我们有*N*个独立同分布的样本，则参数的MLE等于：

 （4.6）

 （4.7）

不难发现，MLE只是单纯的经验期望和经验方差。在单变量高斯分布中，我们有相似的结果：

 （4.8）

 （4.9）

4.1.3.1 证明**\***

为了证明4.1.3的结论，我们需要一些线性代数中的结论，我们将在下面列出。在下式中，**a**和**b**为向量，**A**和**B**为矩阵。符号tr(**A**)表示矩阵的**迹**（trace），定义为矩阵对角线元素的和：。



（4.10）

最后一个等式被称为迹运算的**循环置换性质**（cyclic permutation property）。使用这个性质，我们可以推导出广泛使用的**迹技巧**（trace trick），这种技巧将标量内积进行重新排序。

 （4.11）

基于上述基本定理，接下来我们就可以展开证明。已知对数似然函数为：

 （4.12）

其中为精度矩阵。

令，使用微积分的链式法则，得到：

 （4.13）

 （4.14）

所以：

 （4.15）

 （4.16）

所以的MLE就是经验期望值。

现在我们使用迹技巧重写对数似然函数，以方便求解的MLE:

 （4.17）

 （4.18

其中：

 （4.19）

为以为中心的**散布矩阵**（scatter matrix）。对式（4.18）求关于的导数，得到：

 （4.20）

 （4.21）

所以：

 （4.22）

上式为以为中心的经验协方差矩阵。如果我们使用的极大似然估计（因为两个参数肯定是同时被优化的），我们将得到关于协方差矩阵极大似然解的标准形式。

4.1.4 高斯分布的最大熵性质推导

本节，我们将展示在给定期望和协方差矩阵的前提下，多维高斯分布是所有分布中信息熵最大的分布（9.2.6节同样介绍）。这是高斯分布被广泛使用的理由之一：一个分布的一阶矩（）和二阶矩（）通常是我们能从现有数据中可靠估计出来的全部统计量，所以我们希望概率分布能够充分服从这两个统计量，而尽量不再去做其他额外的假设。

为了简化符号的书写，我们假设期望为0，相应的概率密度函数的形式为：

 （4.23）

如果我们定义和，其中，此时上式的形式与式9.74一致。上述分布的（微分）熵（使用*e*作为对数的底）为：

 （4.24）

我们现在展示，MVN是在给定协方差矩阵的前提下熵最大的分布。

**定理4.1.2.** 令*q*(**x**)为满足的任意概率密度函数。令。我们有.

**证明：**我们有：

 （4.25）

 （4.26）

 （4.27）

 （4.28）

关键步骤4.27式（标\*号）成立的原因是因为我们要求*p*和*q*的协方差矩阵一样。**附：关于式4.27的证明（注意：我们在下面采用ln表示以*e*为底的对数，读者在下面的证明过程中，也可以思考为什么高斯分布的微分熵与期望****无关，即式4.24右边中为什么没有期望项）：**





于是我们有：





4.2 高斯判别分析

MVN的一个重要应用在于：在生成式分类器中定义类条件概率密度，比如：

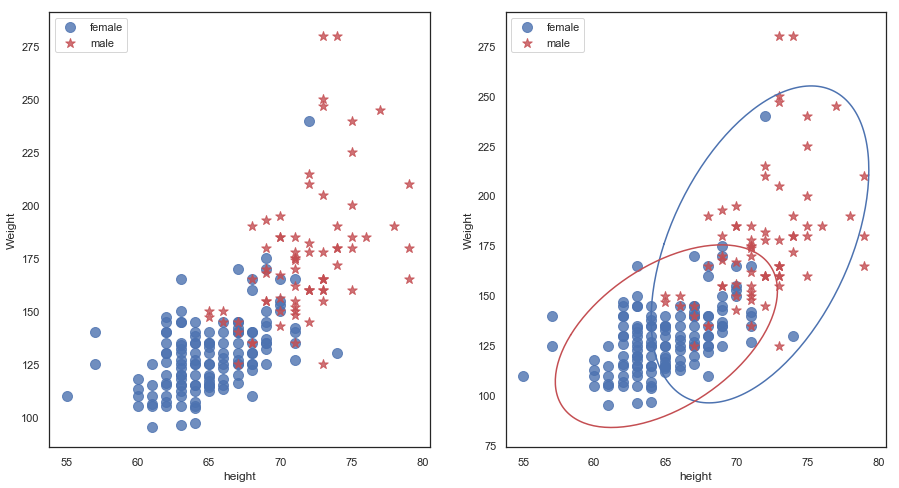
 （4.29）

运用上式的直接结果是引出一个技术叫（高斯）**判别分析**（discriminant analysis）或者叫GDA（尽管这是生成式而非判别式分类器——章节8.6将会进行更多讨论）。如果是对角矩阵，该技术等价于朴素贝叶斯（**读者可以思考为什么？**）。

我们可以使用下式来实现对一个特征向量的分类，由式2.11推理得到：

 （4.30）

当我们基于每个类条件概率密度函数计算**x**的概率密度时，必须测量**x**到每个类中心的马氏距离。所以这种方法又被认为是**最近质心分类器**（nearest centroids classifier）。



（a） （b）

图4.2. （a）身高/体重数据；（b）针对每个类别训练的2维高斯分布的可视化，在椭圆中包含95%的概率质量。图形由程序**gaussHeightWeight**生成。

举例来说，图4.2展示了2个在2维空间中的高斯类条件概率密度函数，分别代表男人和女人的身高和体重。我们可以发现，两个特征是相关的，这与我们的直觉是一致的（高的人胖的可能性更高）。图中每个类的高斯分布边际线（图中椭圆）包含了95%的概率质量。如果我们拥有一个均匀的类先验概率分布，我们可以使用下式对一个新的样本**x**进行分类：

 （4.31）

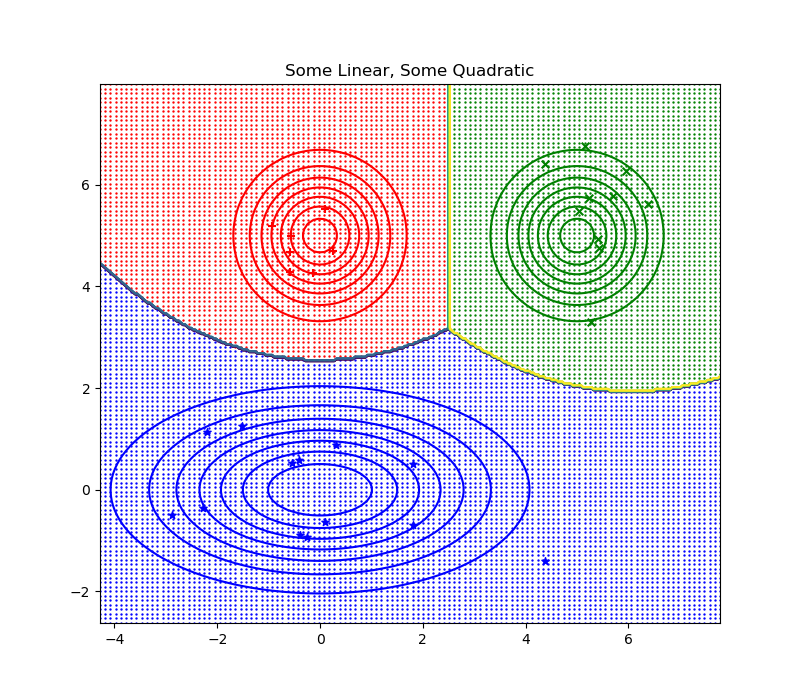
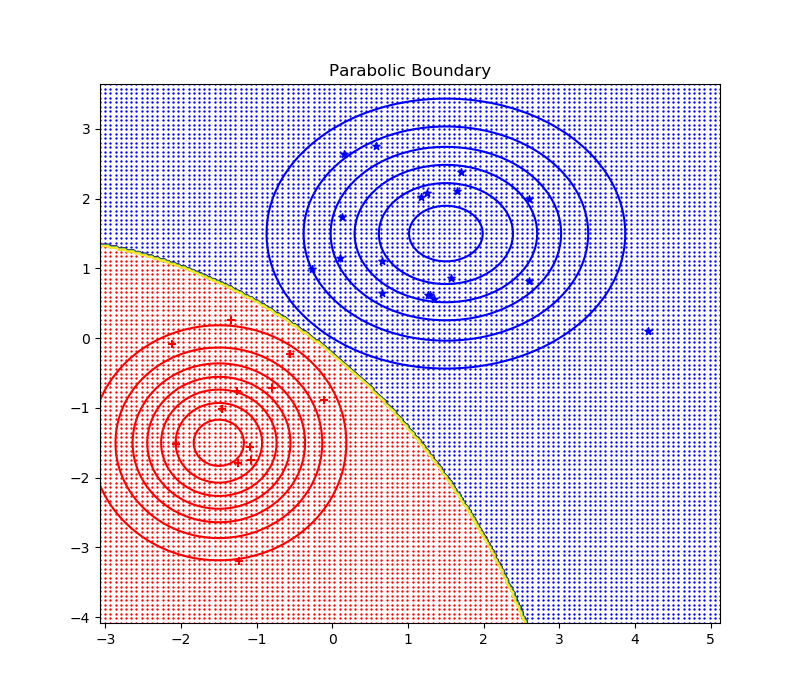


图4.3 2分类和3分类情况下的二次判别模型可视化。图形由程序**discrimAnalysisDboundariesDemo**生成。

4.2.1 二次判别分析

关于类标签*y*的后验分布由式2.11给出，通过引入高斯分布的定义，我们可以对这个模型进行更加深入的探讨，如下所示：

 （4.32）

对上式化简后的函数是关于**x**的二次函数。这个结果被称为**二次判别分析**（quadratic discriminant analysis，QDA）。图4.3给出了一些在2维空间中类与类之间决策边界的形状。

4.2.2 线性判别分析

现在我们考虑一个特殊的情况，在这种情况中，协方差矩阵在不同的类之间**共享**（shared），即。此时，我们可以对式4.32进行化简：

 （4.33）

 （4.34）

如果我们令

 （4.35）

 （4.36）

我们有：

 （4.37）

其中，为softmax函数，定义为：

 （4.38）

Softmax函数之所以叫“softmax”是因为它的性能与max函数十分相似（**译者注**：从字面上理解就是比较光滑（soft）的max函数）。为了说明这一点，将中的每个分量除以常量*T*，该常量被称为**温度**（temperature）。当时，我们发现：

 （4.39）

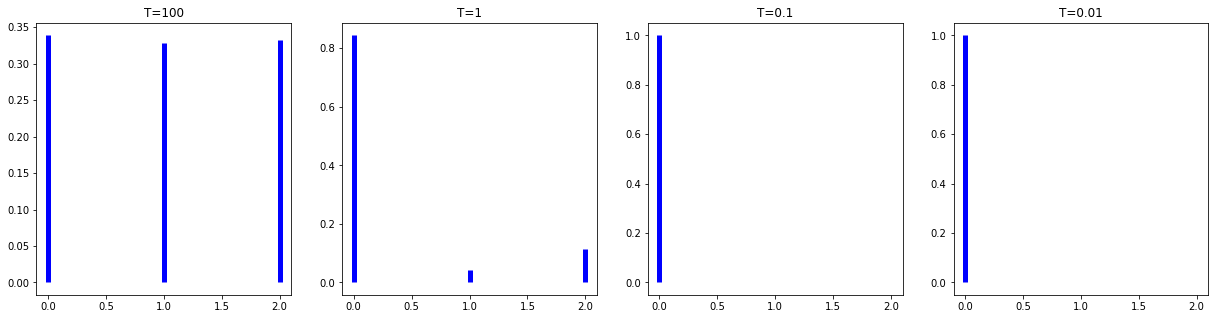


图4.4 不同温度下的Softmax分布，其中。当温度比较高时（左图），分布是比较均匀的，然而当温度比较低时（右图），分布会变得比较“尖”，它将所有的概率质量都集中在了最有可能的那个状态上。图形由程序**softmaxDemo2**生成。

换句话说，当温度比较低时，分布在所有时刻都处在最有可能的状态，然而，当温度变高时，分布所处的状态会比较均匀。图4.4说明了这个原理。值得注意的是，softmax这个术语来自于统计物理学领域，在该领域，一般使用**玻尔兹曼分布**（Boltzmann distribution），它与softmax函数的形式一样。

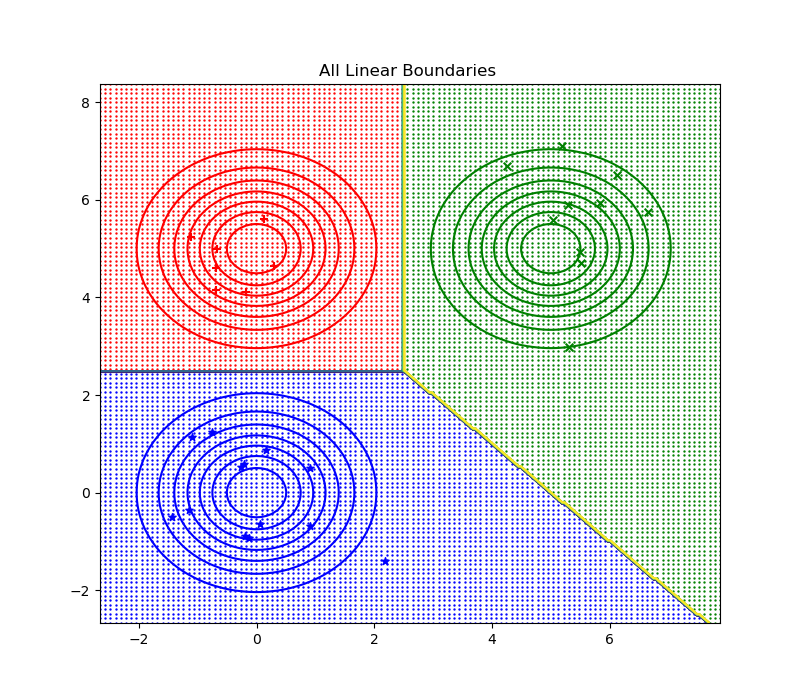
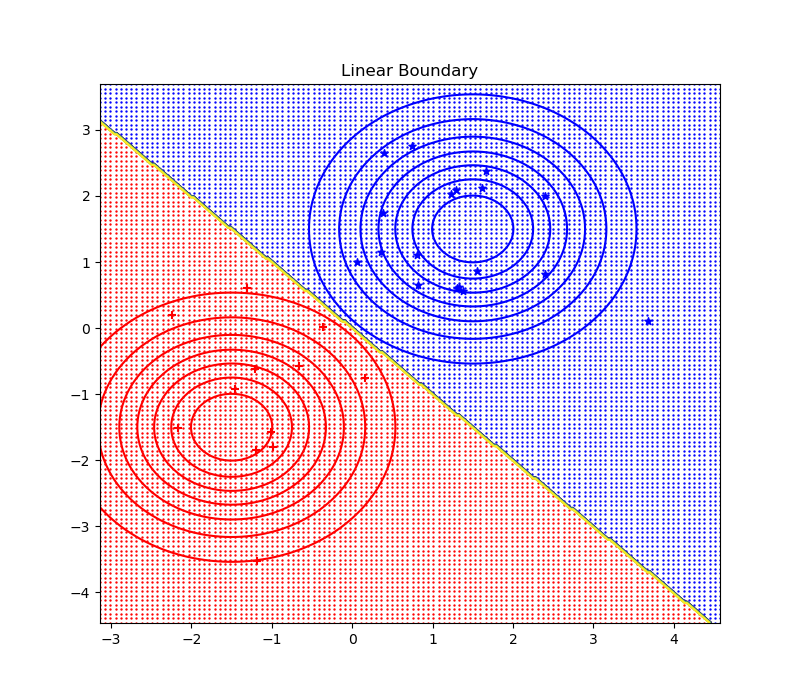


图4.5 2分类和3分类情况下的线性判别模型可视化。图形由程序**discrimAnalysisDboundariesDemo**生成。

式4.37的一个有趣性质是，如果对该式取对数，我们将得到一个关于**x**的线性函数。所以在类*c*和之间的决策边界是一条直线。因此这个技术被称为**线性判别分析**（linear discriminant analysis，LDA）。这条决策边界的形式可由下式导出：

 （4.40）

 （4.41）

 （4.42）

图4.5展示了一些案例。

一种训练LDA模型并且利用这个模型推理类标签后验分布的替代方案是直接训练中的权重矩阵**W**，该矩阵的尺寸大小为*C*×*D*。这被称为**多类别逻辑回归**（multi-class logistic regression），或者**多项式逻辑回归**（multinomial logistic regression）。我们将会在8.2节讨论这个模型的细节，同时在8.6节解释两种方法的区别。

4.2.3 二分类LDA

为了更加深入地理解这些公式的含义，让我们从简单的二分类问题开始分析。在这种情况下，后验分布为：

 （4.43）

 （4.44）

其中为sigmoid函数（式1.10）。

式中：

 （4.45）

 （4.46）

所以，我们定义：

 （4.47）

 （4.48）

然后，我们有，因此，我们有：

 （4.49）

（上式与逻辑回归十分相关，我们将在8.2节对逻辑回归进行讨论。）所以最终的决策规则采用如下的方式得到：考虑向量（**x- x**0），将该向量在方向**w**上进行投影（即内积），观察内积的正负结果。

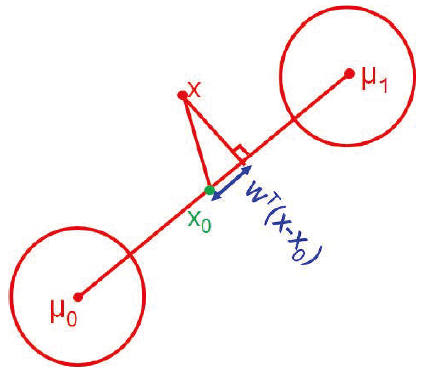


图4.6 2分类问题的LDA，其中

如果，则**w**的方向为。所以我们对某个点的分类只要考虑向量（**x- x**0）的投影是靠近还是。这个过程如图4.6所示。进一步，如果，则，即向量和连线的中点。如果我们令，则**x**0靠近（**译者注：**可以观察式4.48的变化），所以大部分的线段属于类1。相反，如果，则边界将右移。因此我们发现类的类先验分布只是改变决策的阈值，但不会改变决策边界整体的几何形状，正如我们在前面所声明的。（一个相似的讨论同样应用在多分类情况中。）

**w**的幅值决定了逻辑斯特函数sigm的陡峭程度，取决于期望相较于方差的可分离程度有多高。在心理学和信号检测理论中，通常使用物理量d-prime来定义一个信号在有噪音的背景中的**可判别性**（discriminability）：

 （4.50）

其中为信号的期望，为噪声的期望，表示噪音的标准差。如果很大，那么信号可以更容易的从噪声中判别出来。

4.2.4 判别分析的MLE

现在我们讨论如何训练一个判别分析模型。最简单的方式是使用极大似然法。判别分析的对数似然函数为：

 （4.51）

不难发现上式分解成一个包含参数的项和*C*个分别包含参数的项。因此我们对这些参数进行单独的优化。对于类先验分布，我们有，这一点与朴素贝叶斯分类器是一致的。对于类条件概率密度，我们仅仅根据特征向量所属的类别将它们归类，然后对每个类高斯分布参数进行极大似然估计：

 （4.52）

我们在程序**discrimAnalysisFit**中对上述训练过程中给出了python实现。

4.2.5 克服过拟合的策略

MLE方法的速度和简单性是它最大的吸引力之一。然而，MLE对于高维数据会出现严重的过拟合。特别的，当，完全协方差矩阵的MLE是一个奇异矩阵（**译者注：作为一个思考题，读者可以试着证明之**）。就算，完全协方差矩阵也是病态的，意味着它接近奇异矩阵。有几种可能的方法用于解决这个问题：

* 对每个类使用对角协方差矩阵，即假设每个特征之间是条件独立的。这种情况等价于一个朴素贝叶斯分类器（在3.5节中介绍）；
* 使用一个完全协方差矩阵，但是强制每个协方差矩阵都相同，即。这是**参数绑定**（parameter tying）或者**参数共享**（parameter sharing）的应用之一，最终的模型等价于LDA（4.2.2节）。
* 使用一个对角协方差矩阵，并且强制它共享。这被称为对角协方差LDA，将在4.2.7节进行讨论。
* 使用一个完全协方差矩阵，但是为其添加一个先验分布，并最终进行积分。如果我们使用一个共轭先验，这个结果将会是一个闭合的形式，其结果将在4.6.3讨论；这种方法与“贝叶斯方法下的朴素贝叶斯”方法（3.5.1.2节中介绍）类似。
* 通过MAP估计方法训练一个完全或者对角协方差矩阵。我们将会在下文讨论两种不同的先验分布。
* 将数据投影到一个低维空间，再进行高斯分布的训练。章节8.6.3.3将介绍一种方式，从而找到最佳（最具判别性的）的线性投影。

我们将在后文讨论这些方案。

4.2.6 含正则项的LDA\*

假设我们对协方差矩阵实行参数共享，即，这就是在LDA中使用的方法。更进一步的，我们使用一个逆威舍特（Wishart）先验分布（详见4.5.1节）。然后得到关于的MAP估计：

 （4.53）

其中参数控制着正则化的强弱程度，它与先验分布的强度有关（见4.6.2.1介绍更多的细节）。这个技术被称为**正则化判别分析**（regularized discriminant analysis，RDA）。

当我们需要计算一个样本的类条件概率密度时，我们需要计算，也就是，当*D*>*N*时，根本无法计算。然而我们可以使用设计矩阵**X**（**译者注**：由所有样本特征组成的设计矩阵，大小为*N*×*D*）的SVD（奇异值分解，见12.2.3节）来解决这个问题，这一点我们在后面介绍（注意这种方式不能用在QDA，因为它不是关于**x**的线性函数）。

令为设计矩阵的SVD，其中**V**的大小为*D*×*N*，**U**为*N*×*N*的正交矩阵，**D**为尺寸为*N*的对角矩阵。更进一步，定义尺寸为*N*×*N*的矩阵**Z**=**UD**。这类似于一个在低维空间中的设计矩阵（因为我们假设*N*<*D*）。同样，我们定义作为在低维空间中的数据的期望，我们可以使用实现对原来高维空间中的期望的还原，因为（译者注：原文为）。基于这些定义，我们可以重写的MLE：

 （4.54）

 （4.55）

 （4.56）

 （4.57）

 （4.58）

其中为矩阵**Z**的经验协方差矩阵。因此我们可以将MAP估计写成：

 （4.59）

 （4.60）

值得注意的是，实际上我们并不需要计算*D*×*D*的矩阵。因为从式4.38中我们发现，使用LDA进行分类时，我们需要做的只是计算，其中

 （4.61）

对于RDA方法而言，我们在计算重要的项时不需要计算*D*×*D*的矩阵：

 （4.62）

其中为属于类*c*的矩阵**Z**的期望。可参考程序**rdaFit**。

4.2.7 对角LDA

一种替代RDA的简单方法是共享协方差矩阵，正如在LDA中的那样，令。然后对于每个类使用对角协方差矩阵。这被称为**对角LDA**（diagonal LDA）模型，等价于时的RDA。对应的判别式函数为（对比式4.32）：

 （4.63）

一种典型的情况是令和，其中被称为特征*j*的**混合经验方差**（pooled empirical variance），定义为：

 （4.64）

在高维数据中，这个模型的效果比LDA和RDA要好。

4.2.8 最近收缩质心分类器\*

对角LDA的一个缺点在于它依赖于所有的特征。在高维数据的问题中，出于精度和可解释性的原因，我们可能更倾向于只使用全部特征的一个子集。一种方法是使用特征筛选，这种方法可以基于3.5.4节所描述的互信息。我们现在讨论另一种解决这个问题的方法：**最近收缩质心**（nearest shrunken centroids）分类器。

该方法的基本思想是基于一个促进模型稀疏化的先验分布（Laplace先验分布），对对角LDA进行MAP估计（见13.3节）。更加精确的表达为，定义特定类下某个特征的期望为一个与类无关的特征期望和一个与类相关的偏置量的组合：

 （4.65）

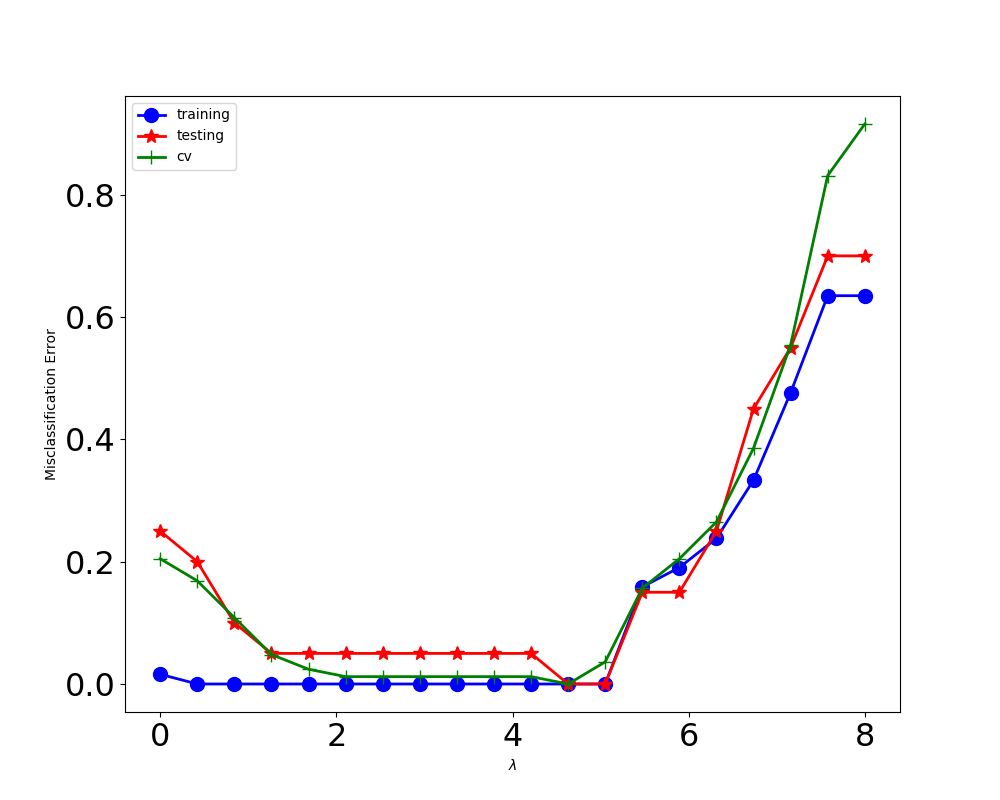
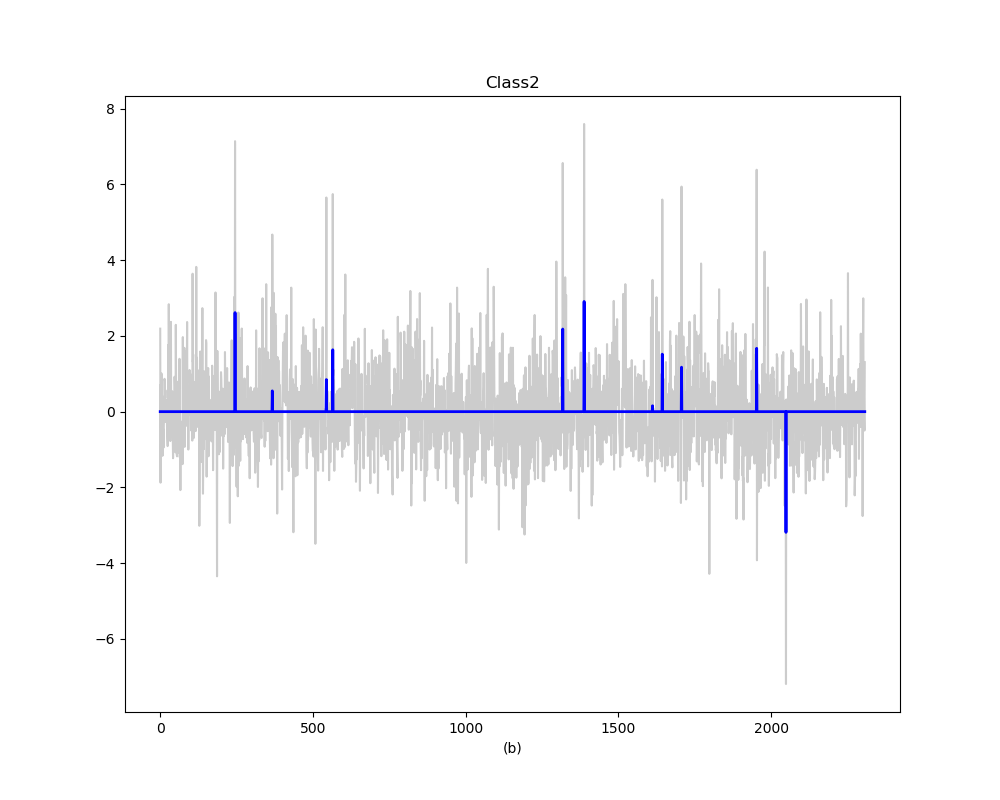
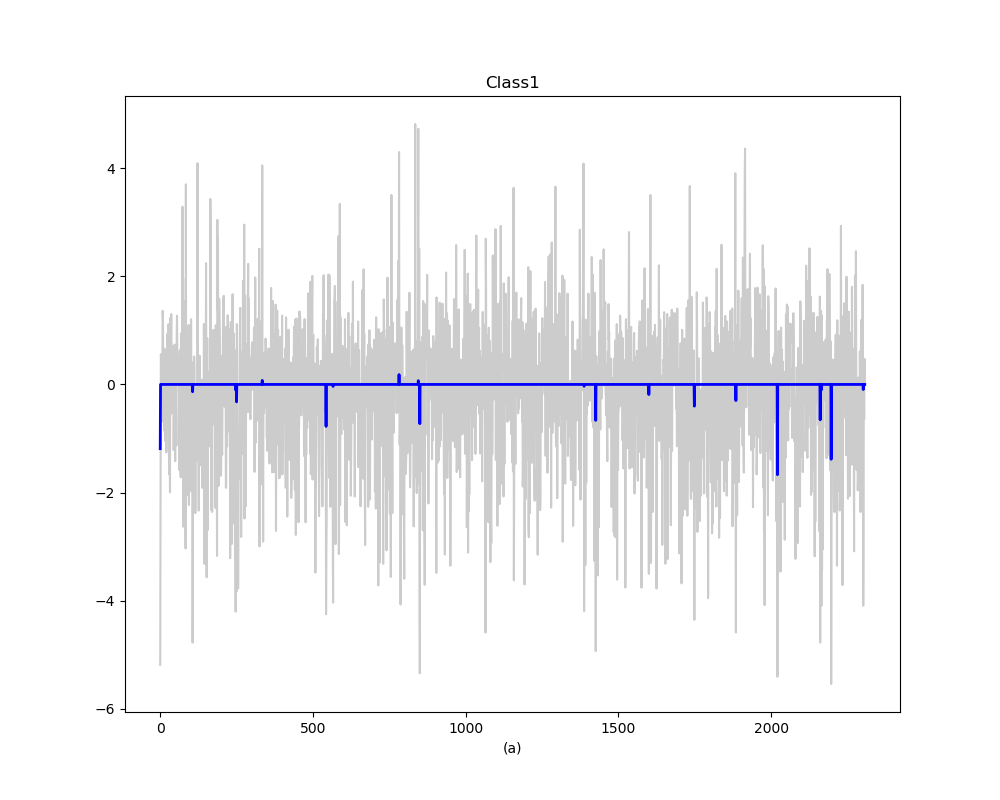


图4.7 在最近收缩质心模型中，误分类率与收缩程度之间的关系，上述模型的数据为SRBCT基因表达数据。图形由程序**shrunkenCentroidsSRBCTdemo**生成。



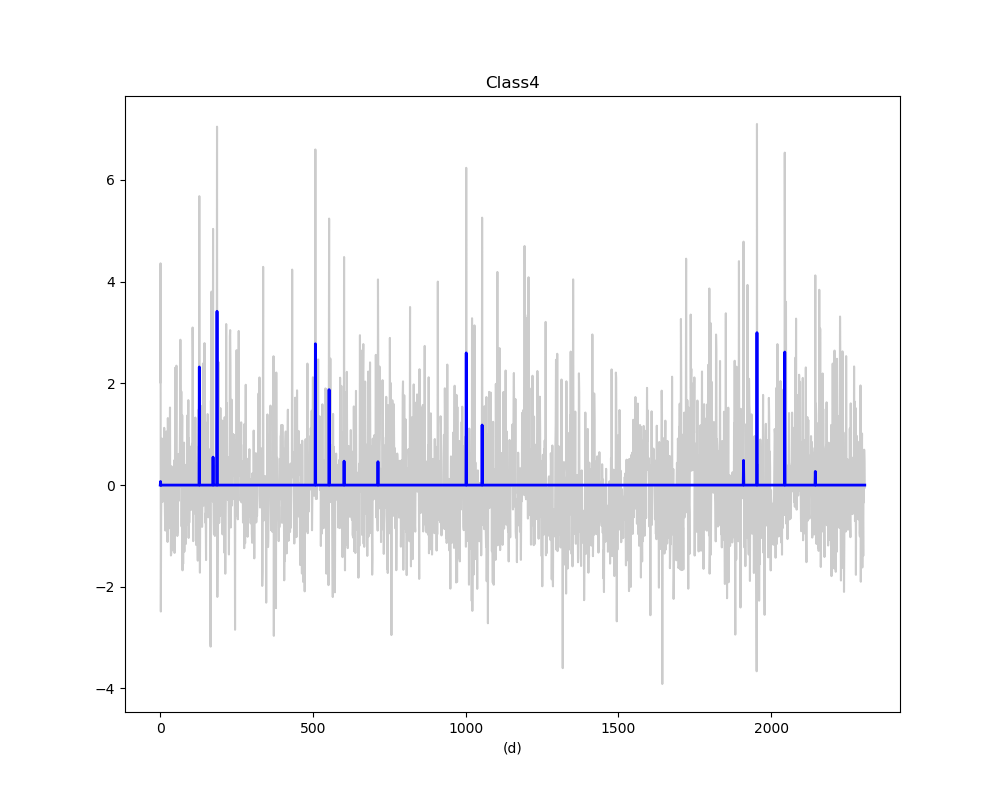
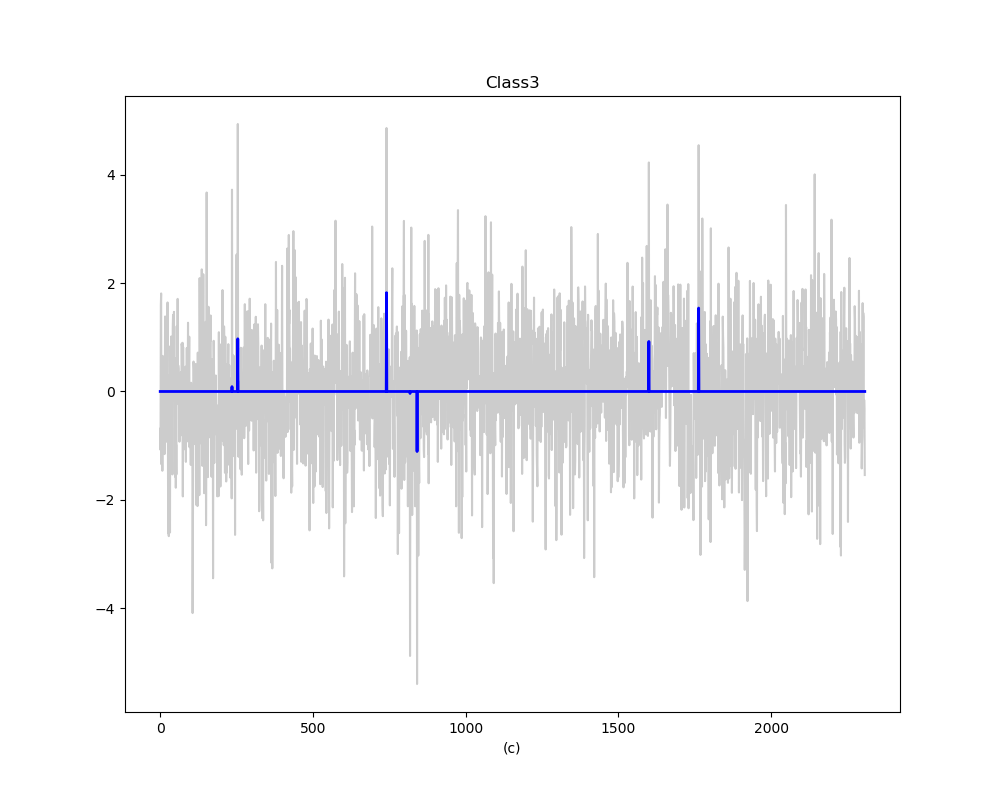


图4.8 通过交叉验证得到的最优参数=4.6（原文为4.4），基于该最优模型绘制的收缩质心。其中与类别相关的特征共50个（原文为39个）。图形由程序**shrunkenCentroidsSRBCTdemo**生成。

我们将会为赋予一个先验分布，该先验分布将鼓励中的某些元素严格为零，并基于该先验分布进行最大后验估计。如果对于特征*j*，在所有类*c*中，我们有，则特征*j*对于解决分类问题没有任何帮助（因为将独立于类*c*）。所以那些不具备可判别性的特征将会被自动忽略。程序**shrunkenCentroidsFit**展示了上述过程。

让我们展示一个实际应用案例。考虑一个问题，在这个问题中，我们需要对一个基因表达数据库进行分类，该数据库中有2308个基因，共分为4个类别，63个训练样本和20个测试样本。使用对角LDA分类器在测试集上将产生5（原文中也为5个）个错误。对于一定范围内的值，使用最近收缩质心分类器在测试集上的误分类率为0，如图4.7所示。更为重要的是，该模型是稀疏的，所以更加具备可解释性，图4.8用灰色曲线绘制了不含惩罚项的的估计值，同时用蓝色曲线绘制了的收缩估计值（含惩罚项）。（这些估计值使用通过交叉验证估计出来的最优值进行计算。）我们发现只有50（原文中是39个）个基因被用来做分类任务，而非原来的2308个基因。

现在考虑一个更加困难的问题，我们有16603个基因，训练集包含144个病人，测试集包含54个病人，以及14个不同种类的癌症。Hastie等人指出最近质心收缩分类器只使用6520个基因，在测试集上产生17个误差；RDA（4.2.6节）使用16603个基因，其在测试集上产生12个误差。程序**cancerHighDimClassifDemo**可以用来产生这些数字。

4.3 联合高斯分布中的推理问题

给定一个联合分布*p*(**x**1,**x**2)，如果能够计算边缘概率分布*p*(**x**1)和条件概率分布*p*(**x**1|**x**2)将会是一件十分有用的事。我们将会在下文中进行讨论，并且给出一些具体的应用。这些运算在最坏情况下的时间复杂度为*O*(*D*3)。20.4.3节会给出更快的方法。

4.3.1 相关结论

**定理4.3.1** （MVN的边缘和条件分布）。假设**x**=[**x**1,**x**2]为具备如下参数的联合高斯分布：

 （4.66）

其边缘概率分布为：

 （4.67）

其后验条件分布为：



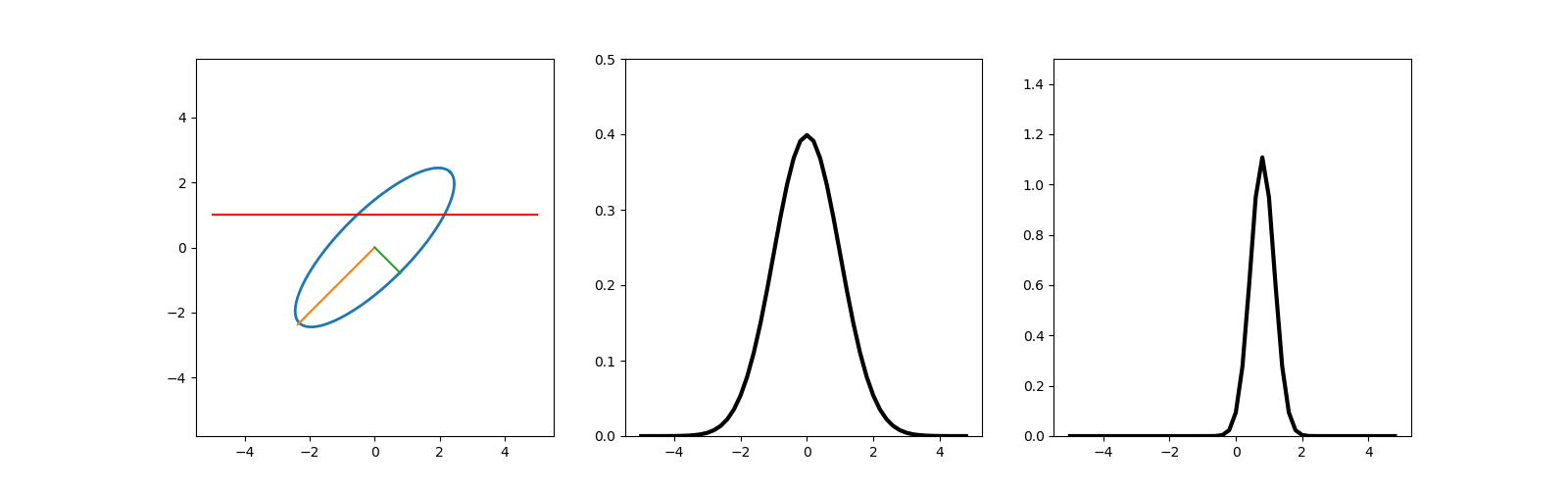
（4.68）

式4.68在本书中是十分重要的结论，所以我们使用方框将其框出，这样你就可以很容易找到它，关于它的证明请参考4.3.4节。

不难发现，无论是边缘概率还是条件概率分布，其本身都还是高斯分布。对于边缘概率分布，我们仅仅只是提取了对应于**x**1或**x**2的参数。对于条件概率分布，我们需要做更多的工作。然而，它并没有想象中的那么复杂：对于条件概率分布的期望，它是关于**x**2的线性函数，条件概率分布的协方差矩阵是一个与**x**2无关的常数矩阵。我们给出关于后验期望的三种（等价的）表达式，以及协方差矩阵的两种（等价的）表达式，每一种表达方式在不同的情况下都有各自的用处。

4.3.2 例子

接下来，我们将给出上式的一些实际应用，从而使得这些结论更加的直观。



（a） （b） （c）

图4.9 （a）联合高斯分布*p*(*x*1,*x*2)，相关系数为0.8。图中我们绘制了95%概率质量的等高线和主轴。（b）非条件边缘分布*p*(x1)。（c）条件概率分布*p*(*x*1|*x*2)=(*x*1|0.8,0.36)。图形由程序**gaussCondition2Demo2**生成。

4.3.2.1 2维高斯分布的边缘分布和条件分布

让我们考虑一个在2维空间中的例子。其协方差矩阵为：

 （4.69）

边缘分布*p*(*x*1)为一个1维高斯分布，其获得的方式是将2维高斯分布投影到*x*1轴上：

 （4.70）

假设我们观察到了*X*2=*x*2；其条件概率分布*p*(*x*1|*x*2)的获取方式是通过直线*X*2=*x*2对联合概率分布进行“切片”（如图4.9所示）：

 （4.71）

如果，我们得到：

 （4.72）

在图4.9中我们展示了当时，*x*2=1例子。我们发现，这与直觉上是一致的，因为，意味着如果*x*2基于它的期望（）增长1，那么*x*1基于它的期望（）的增长为0.8。我们同时发现。这与直觉上也是一致的：通过观察到*x*2值，我们关于*x*1的不确定度会下降，因为我们已经学习到了一些关于*x*1的知识。如果，我们将得到，因为如果*x*2与*x*1不相关（因此互相独立，译者注：此处应该是不存在线性关系），那么*x*2并不会涵盖关于*x*1的任何信息。

4.3.2.2 无噪数据的插值

假设我们需要估计一个1维函数，其定义域为[0,*T*]，该函数在*N*个观察点*t*i处满足。现在假设观察到的数据是无噪的，我们需要对这些数据进行插值（interpolate），即：我们需要拟合一条函数，这个函数精确地通过这些观测点（4.4.2.3将介绍含噪数据的情况）。问题在于：函数在观察到的数据点之间的行为是什么样的呢？通常情况下，“未知函数是光滑的”这一假设是比较合理的。在第15章，我们将看到如何为函数赋予先验分布，并且根据观察到的数据去更新这个先验分布，从而得到关于函数的后验分布。但在本节，我们采用一个更加简单的方式，对于那些定义在1维空间上的函数而言，利用插值法计算出来的函数作为最大后验估计是合适的。

我们首先对这个问题进行形式化描述。我们将该函数的支撑集（定义域）分成*D*的相等的区间。然后我们定义：

 （4.73）

通过假设*xj*为相邻两个点*xj-1*和*xj+1*的均值，外加一些高斯噪音，我们实现了之前“函数是光滑的”这一先验假设：

 （4.74）

其中。精度项控制着我们所认为的这个函数的变化范围：当比较大时，则代表我们相信这个函数非常光滑，当比较小时，则我们认为这个函数相当扭曲。使用向量的表达形式，上式可以被写成：

 （4.75）

其中**L**为尺寸等于（*D*-2）×*D*的二阶**有限差分矩阵**（finite difference matrix）：

 （4.76）

相应的先验分布具备如下形式：

 （4.77）

在后面的内容中，我们假设矩阵**L**已按照比例（译者注：个人觉得应该是）增长，所以在后面的内容中将忽略项，仅仅书写作为精度矩阵。

值得注意的是，尽管**x**为*D*维向量，而精度矩阵的秩仅为*D*-2（读者可以自行验证）。所以这不是一个合适的先验分布，而这种情况对应于一个内在的高斯随机场（19.4.4节给出更多信息）。然而，如果我们观察到*N* ≥ 2个数据点，则相应的后验分布是合理的。

现在我们令**x**2为*N*个已知的无噪音函数观测值，**x**1为*D*-*N*个未知的函数值。不失一般性，假设未知变量被排在前面，其次是观测变量。这样我们就可以将矩阵**L**拆分成：

 （4.78）

同样，我们将联合分布的精度矩阵拆分成：

 （4.79）

使用式4.68，我们可以将条件概率分布写成：

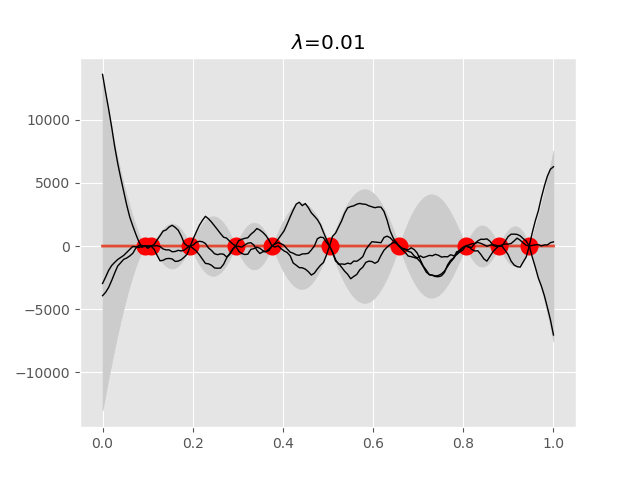
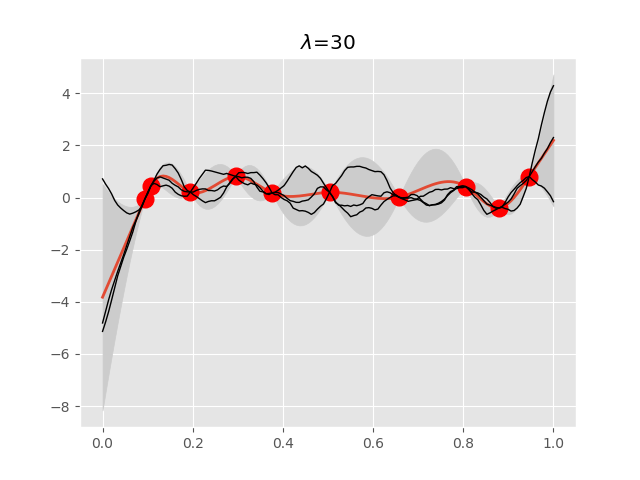
 （4.80）

 （4.81）

 （4.82）

值得注意的是，通过求解下面的线性方程组，我们可以得到期望值：

 （4.83）



（a） （b）

图4.10 对无噪音数据进行插值，其中高斯先验分布的精度为。（a）=30.（b）=0.01。图形由程序**gaussInterDemo**生成。

这种方式是有效的，因为**L**1是一个三对角矩阵。图4.10给出了这些方程的说明。我们发现，后验期望值在特定的点等于观测到的值，并且在观测值之间实现了平滑插值，这一点与我们的目标是一致的。

绘制出每个**点的边缘置信区间**（pointwise marginal credibility intervals）是一件很有趣的事情，这个置信区间用灰色区域表示。不难发现，当我们离开那些观测数据点时，方差变大。同时，当我们降低先验分布的精度时，其方差也会增加。有趣的是，的值并不会对后验分布的期望值有影响，因为它在乘上和时，的影响被消除了。相反，如果我们考虑观测到的是含噪数据点，正如在4.4.2.3中所介绍的，我们将会发现先验分布的精度对后验分布期的望值估计值也有影响。

边缘置信区间并不能捕捉到如下事实：相邻的点是相关的。为了说明这一点，我们可以后验分布中采样出完整的函数（比如向量**x**），并且绘制这个函数曲线。这些曲线在图4.10中以比较细的曲线显示。这些曲线相较于后验分布的期望值来说并没有那么光滑。因为在先验分布中，我们仅仅惩罚了一阶差分。关于这一点，我们会在4.4.2.3中进行进一步的讨论。

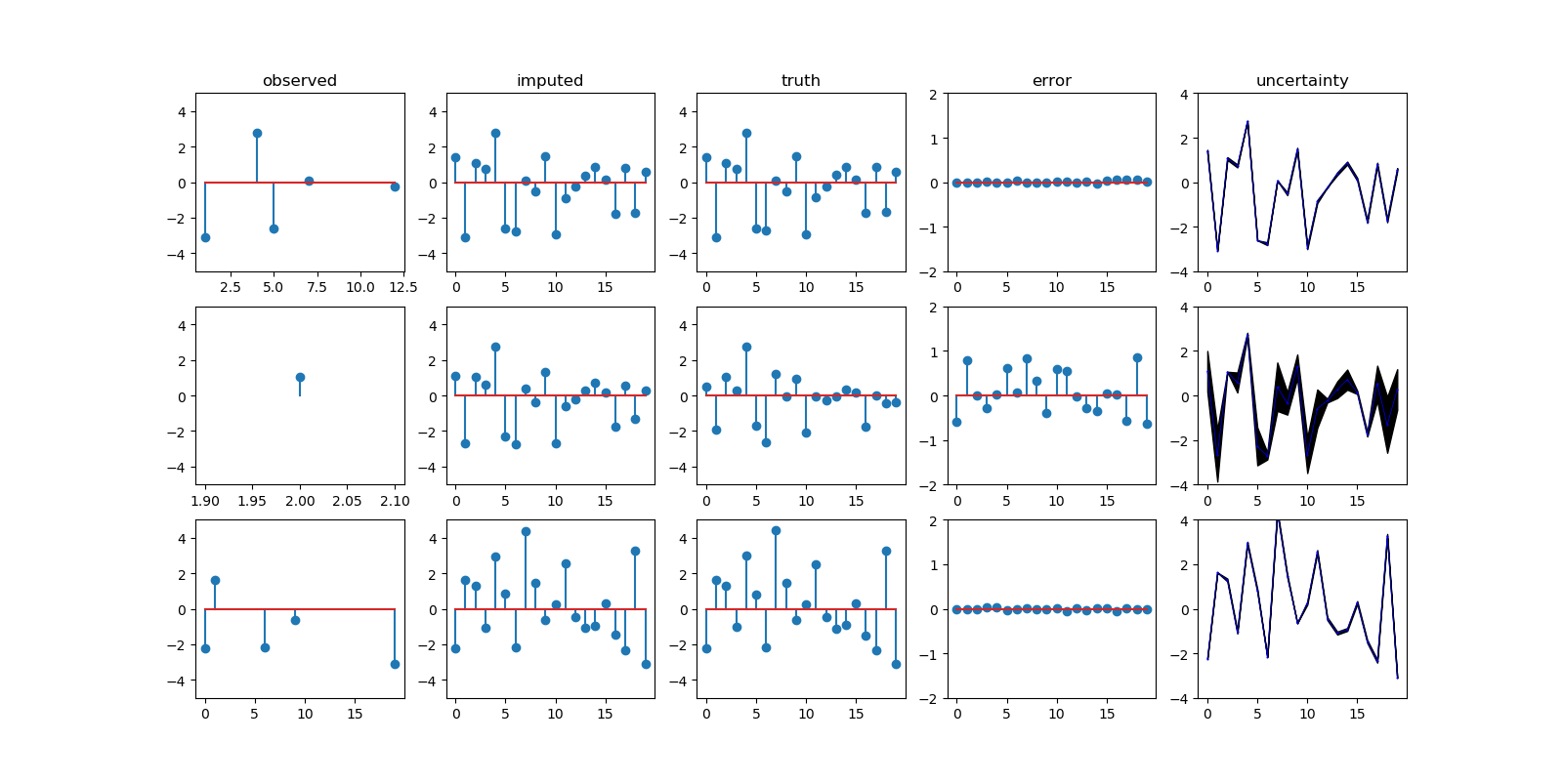


图4.11 数据重构的示意图。左侧：相关行中观测到的数据点；第2列：数据重构后的后验预测期望；第3列：真实的数据；第4列：真实数据与重构数据的偏差；第5列：后验预测的不确定度。（图形由程序**gaussImputationDemo**生成）

4.3.2.3 数据重构

假设我们在一个设计矩阵中缺失了一些数据项。如果列与列（即变量与变量）之间是相关的，那么我们可以使用看到的数据去预测缺失的数据。图4.11展示了一个简单的例子。我们从一个20维的高斯分布中采样得到一些数据，然后在每一行故意“隐藏”50%的数据。然后根据观察到的数据，使用真实的（生成）模型，去推理那些缺失的数据。更加精确的表达是，对于每一行*i*，我们计算，其中和为在样例*i*中观测值和隐藏值的索引值。基于，我们计算每个缺失变量的边缘分布。然后绘制出该分布的期望；从最小化我们的期望方差角度而言，该值表示我们对这个缺失项的真实值的最佳猜测（5.7节给出更多细节）。图4.11说明预测的值与真实值十分接近。（当然，如果，则期望值等于观测值）

我们可以使用作为对我们猜测值的不确定度的衡量，尽管在图中这一点并没有展示（译者注：我们在图4.11的第5列展示了这一点）。另一个可选的方案是，我们可以从分布采集多个样本，这被称为**多重重构**（multiple imputation）。

除了对那些缺失值进行重构，我们可能对计算每一行中那些观测值的似然函数感兴趣，可以使用式4.67进行计算。这对检测异常值是有用的（非典型应用）。

4.3.3 MVN的信息表达形式

假设。我们可以知道为期望向量，为协方差矩阵。这些被称为该分布的**矩参数**（moment parameters）。然而，有些时候使用**典范参数**（canonical parameters）或者**自然参数**（natural parameters）是有用的，定义为：

 （4.84）

我们可以使用下式实现矩参数的转换：

 （4.85）

（关于上述的两个典范参数我们将在9.2节讨论更多的细节，那个时候我们将讨论指数族分布。）

使用典范参数，我们可以将MVN写成**信息形式**（information form）（又被称为**典范形式**canonical form）：

 （4.86）

其中我们使用符号以区别于矩参数形式。

同样，我们也可以推导出信息形式的边缘和条件概率分布。我们发现：

 （4.87）

 （4.88）

所以我们发现矩形式下的边缘分布更加容易，而信息形式下的条件分布更加容易。

当我们采用信息形式时，有一个操作也变得非常的容易，即：两个高斯分布相乘。我们有：

 （4.89）

然而，在矩参数形式下，事情会变得比较麻烦：

 （4.90）

4.3.4 结果证明\*

现在我们讨论定理4.3.1。对那些繁重的线性代数感到反感的读者可以安全地跳过这个章节。我们首先会推导一些我们会在本节或者书中其他地方使用到的结果。最后我们会回到证明本身。

4.3.4.1 利用舒尔补实现分块矩阵的逆

我们需要的关键工具是对一个分块矩阵求逆。利用下面的结果我们可以完成这一点。

定理4.3.2 （分块矩阵的逆矩阵）。考虑一个一般形式的分块矩阵：

 （4.91）

其中我们假设**E**和**H**是可逆的。我们有

 （4.92）

 （4.93）

其中：

 （4.94）

 （4.95）

我们称为矩阵**M**关于**H**的**舒尔逆**（Schur complement）。式4.92被称为**分块求逆公式**（partitioned inverse formula）。

证明：如果我们可以对矩阵**M**进行分块对角化，那么对矩阵**M**的求逆会变得更加容易。为了使矩阵**M**的右上角元素变成0，我们进行如下的矩阵欲乘：

 （4.96）

相似的，为了使其左下角元素变成0，我们进行如下的矩阵后乘：

 （4.97）

将上式放在一起，我们得到：

 （4.98）

对上式两边求逆：

 （4.99）

所以：

 （4.100）

将上式用相应的矩阵定义替换，得到：

 （4.101）

 （4.102）

 （4.103）

另一种可选方案是，将矩阵**M**分解为**E**和项，得到：

 （4.104）

4.3.4.2 矩阵求逆引理

我们现在基于上述结果推理一些有用的引理。

**引理4.3.1** （矩阵求逆引理）。考虑一个一般的分块矩阵，我们假设**E**和**H**式可逆的。我们有：

 （4.105）

 （4.106）

 （4.107）

前两个等式被称为**矩阵求逆引理**（matrix inversion lemma）或者**谢尔曼 - 莫里森 - 伍德伯里公式**（Sherman-Morrison-Woodbury formula）。第三个等式被称为**矩阵行列式引理**（matrix determinant lemma）。在机器学习或者统计学中的一个典型应用如下。令为*N*×*N*的对角矩阵，大小为*N*×*D*，其中，令。然后我们有：

 （4.108）

等式的左边计算的时间复杂度为*O*(*N*3)，等式右边的时间复杂度为*O*(*D*3)。

另一个相关的应用是计算一个逆矩阵的**秩一矫正**（rank one update）。令*H* = -1（一个标量），**F** = **u**（一个列向量），**G** = **v**T（一个行向量）。然后我们有：

 （4.109）

 （4.110）

当我们不断将一个数据向量添加到一个设计矩阵上，并且想更新我们的充分统计量时，上述的结论十分有用。

证明：为了证明等式4.105，我们只需要简单的令式4.92和4.93的左上方分块相等即可.为了证明等式4.106，我们只需要简单的令式4.92和4.93的右上方分块相等即可。读者可以试着证明式4.107。

4.3.4.3 高斯条件分布公式的证明

现在回到我们最开始的目的，即证明等式4.68。将联合概率分布*p*(x1,x2)分解成*p*(x1)和*p*(x1|x2)：

 （4.111）

使用式4.101，上式指数可以写成：

 （4.112）

 （4.113）

 （4.114）

 （4.115）

上式具备如下的形式：

exp(关于x1和x2的二次项)×exp(关于x2的二次项) （4.116）

所以，我们成功地将联合分布因式分解为：

 （4.117）

 （4.118）

条件分布的参数可以从上面的等式中解析出来，为：

 （4.119）

 （4.120）

我们同样可以使用来验证归一化常数是正确的：

 （4.121）

 （4.122）

其中*d*1=dim(**x**1)，*d*2=dim(**x**2)。

4.4 线性高斯系统

假设我们有两个变量**x**和**y**。令为隐变量，为变量**x**的含噪观测值。假设我们有如下的先验分布和似然函数：



（4.123）

其中矩阵**A**的大小为。上式是一个**线性高斯系统**（linear Gaussian system）的例子。用图示的方式表达为，意思是说**x**生成**y**。本节我们将讨论如何逆转这个箭头，也就是说如何从**y**推测出**x**。我们将在下文展示相关结论，并且给出一些案例，最后再给出这个结论的推导过程。在后面的章节中我们会看到这些结论的更多应用。

4.4.1 结果陈述

定理4.4.1（线性高斯系统的贝叶斯法则）。给定一个线性高斯系统，如式4.123所述，其后验概率分布*p*(**x**|**y**)由下式给出：



（4.124）

除此之外，归一化常数*p*(**y**)为：



（4.125）

证明过程见章节4.4.3。

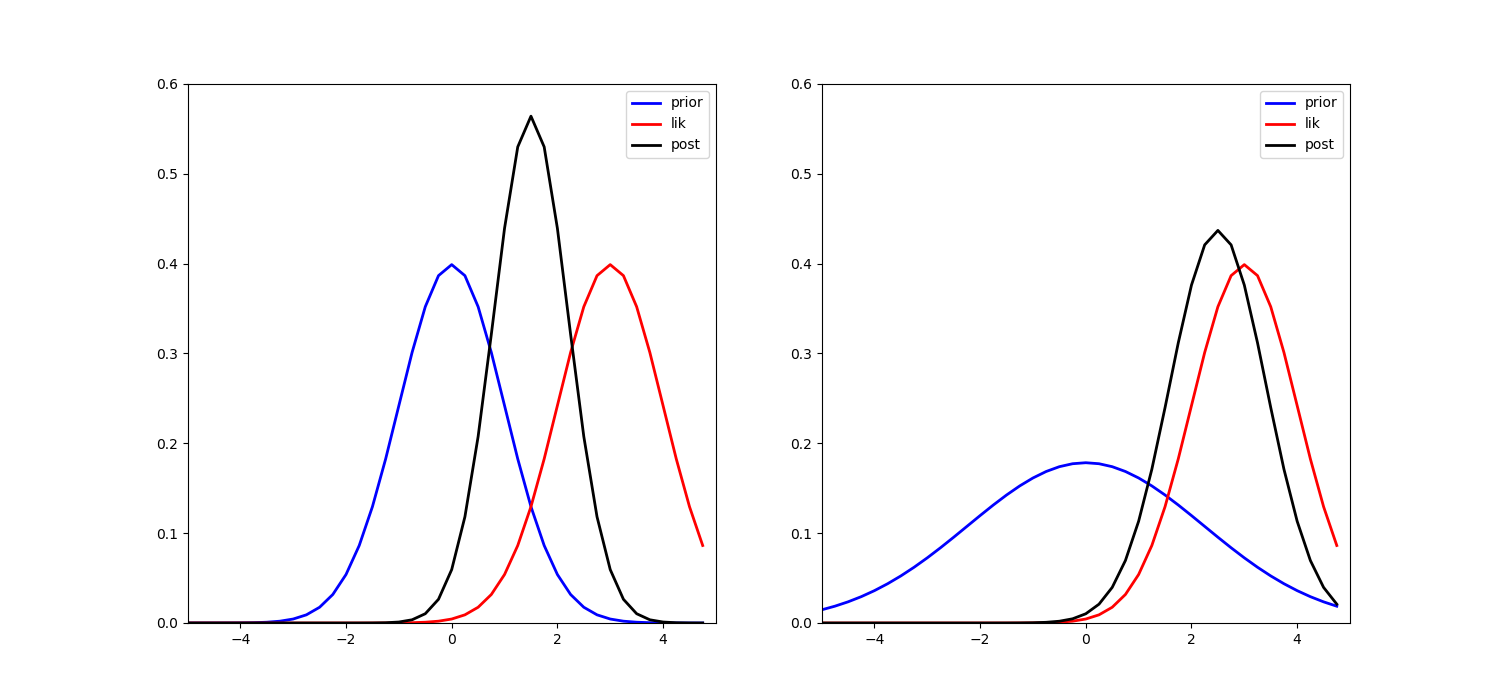


图4.12 基于含噪观察值y=3的基础上进行的推理。（a）强先验，后验分布的期望值向先验分布的期望值0靠近。（b）弱先验分布。后验分布的期望与MLE相似。图形由程序**gaussInferParamsMean1d**生成。

4.4.2 例子

本节，我们将展示一些上述结果的应用。

4.4.2.1 从含噪的测量数据中预测一个未知标量

假设我们针对一个潜在量*x*作出*N*个含噪的测量*y*i；假设测量值的精度都是一样的，等于。所以，相应的似然函数为：

 （4.126）

假设未知真实值*x*的先验分布为高斯分布：

 （4.127）

我们希望计算。为了在高斯分布中方便使用贝叶斯法则，我们将上述的形式进行改写，定义（一个1×*N*的行向量，元素都为1），。然后我们有：

 （4.128）

 （4.129）

 （4.130）

这些等式十分符合我们的直观感受：后验精度为先验精度加上*N*个测量值的精度。同样，后验分布的期望为最大似然估计和先验期望的凸组合。这就说明了后验分布的期望是在MLE和先验期望之间的一个权衡。如果相较于信号（即观测到的数据）的强度，先验较弱（相较于较小），我们将赋予MLE更大的权重。如果先验相较于信号的强度较强（相较于较大），我们将赋予先验更多的权重。这个过程在图4.12中进行展示，该图与图3.6中展示的beta-binomial分布十分相似。

值得注意的是，在后验分布的期望中包含项，所以拥有*N*个测试精度为的测量值等价于拥有1个值为，精度为的测量值。

我们将之前的结果利用后验分布的方差而非后验分布的精度进行重写：

 （4.131）

 （4.132）

 （4.133）

其中表示先验分布的方差，表示后验分布的方差。

我们可以在每个观察值之后采用序列化的方式对后验分布进行更新。如果*N*=1，我们可以在观察到一个测量值之后重写后验分布（其中我们定义分别为似然函数，先验分布和后验分布的方差）：

 （4.134）

 （4.135）

 （4.136）

我们可以用如下的三种形式重写后验分布的期望：

 （4.137）

 （4.138）

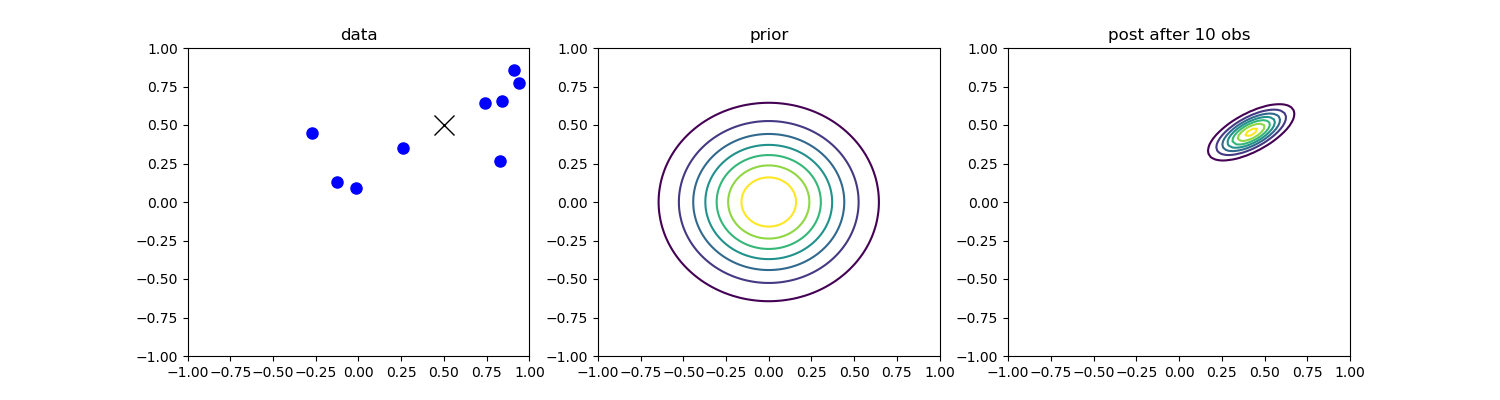
 （4.139）

第一个等式为先验分布期望和观测数据之间的凸组合。第二个等式表示先验期望向数据方向进行调整。第三个等式表示数据向先验期望进行调整，这被称为**收缩**（shrinkage）。这些等式是等价的，它们都说明后验分布在似然函数与先验分布之间的一种权衡。如果相较于小，即对应于一个更强的先验分布，那么收缩的程度将会很大（见图4.12(a)）。然而，如果相较于大，即对应于一个更弱的先验分布，那么收缩的程度很小（见图4.12(b)）。

另一种衡量收缩的程度的方式是使用**信噪比**（signal-to-noise ratio），定义为：

 （4.140）

其中为真实的信号，为观测到的信号，为噪音项。



（a） （b） （c）

图4.13 2维高斯分布情况下的贝叶斯推理。（a）生成的数据服从分布，其中，。我们假设传感器的协方差矩阵已知，但是未知。黑色的×代表。（b）先验分布。（c）在观察到10个数据之后的后验分布。图形由程序**gaussInferParamsMean2d**生成。

4.4.2.2 从含噪的测量数据中预测一个未知矢量

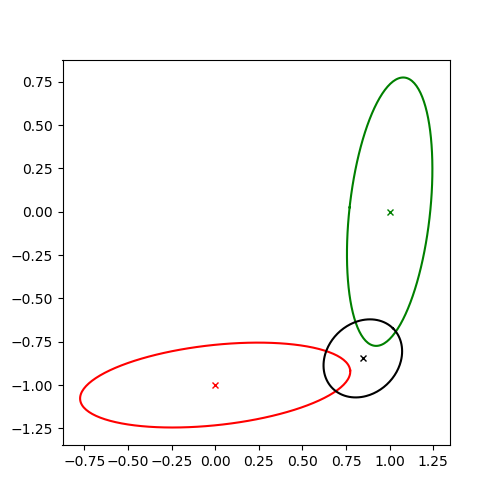
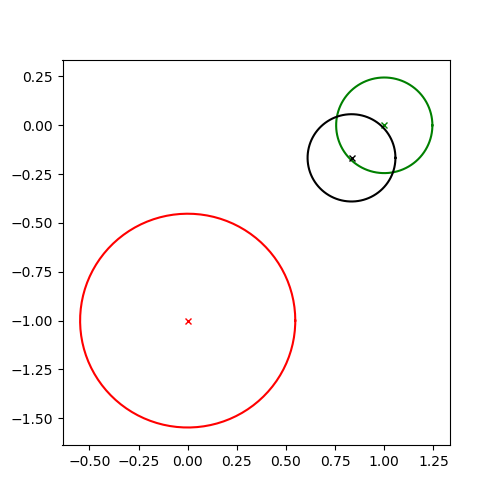
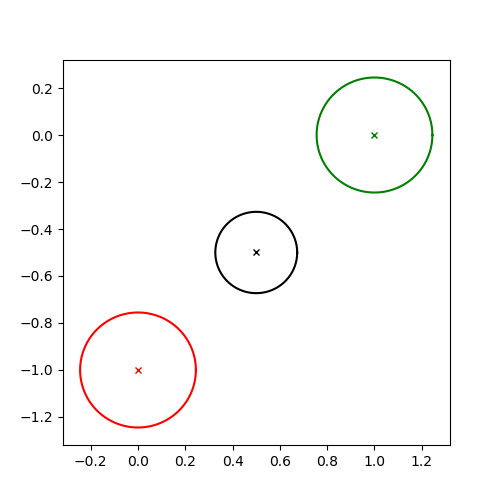
现在考虑*N*个向量观测值和一个高斯先验分布。令，使用表示精度为的有效观测值，我们有：

 （4.141）

 （4.142）

 （4.143）

图4.13展示了一个在2维空间中的情况。我们可以将图中符号x当作一个目标在2维空间中正确但未知的坐标位置，比如一个导弹或者飞机，将当作含噪的观测值，不如说雷达的“杂音”。当我们收到更多的雷达杂音，我们将更好地去定位真实的目标。在18.3.1节，我们将介绍如何扩展这个案例，利用著名的卡尔曼滤波算法实现对移动目标的追踪。



（a） （b） （c）

图4.14 在我们观察到**y**1=(0, -1)（红×）和**y**2=(1, 0)后（绿×）后进行的推理（黑×）。（a）相同精度的传感器，所以后验分布的期望值在两个圆圈之间；（b）传感器2更加可靠，所以估计值向绿色圆圈靠拢。（c）传感器1在垂直分量上更加可靠，传感器2在水平分量上更加可靠。估计值是两个测量值的一个组合。图形由程序**sensorFusion2d**生成。

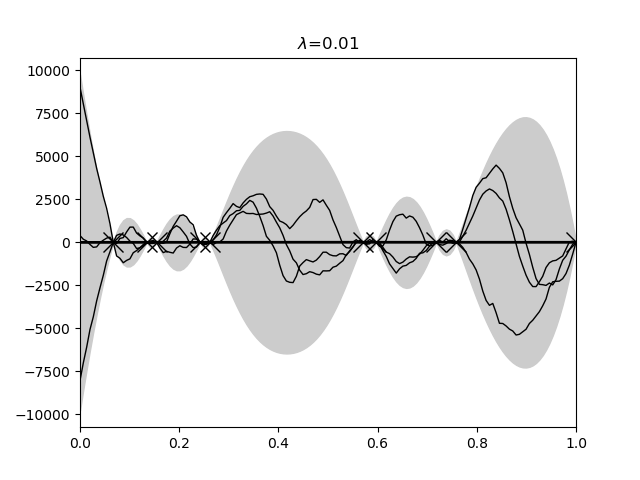
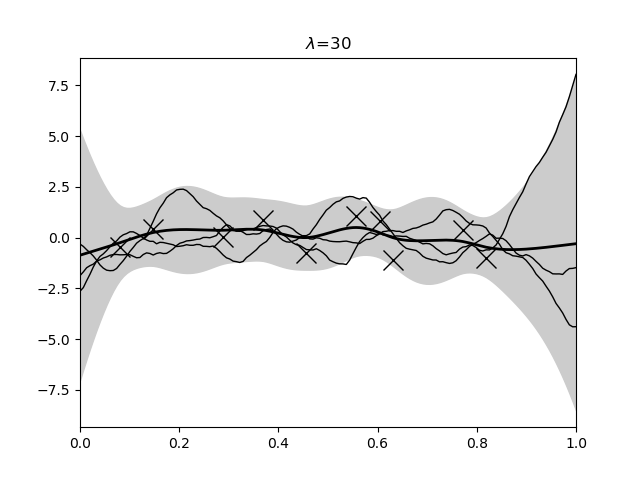
现在假设我们有多个检测设备，我们需要将它们的信号组合起来。这种方式被称为**传感器融合**（sensor fusion）。如果不同的观测值之间具备不同的协方差（即不同设备的可靠度不同），那么后验预测将会是不同数据的加权平均。考虑图4.14中的案例，我们使用一个关于**x**的无信息先验分布，。我们得到2个含噪观测值，和。然后，我们计算。

在图4.14(a)中，我们令，也就是说两个测试设备的可靠性是一致的。在这种情况下，后验期望在观测值**y**1和**y**2的连线中点处。在图4.14(b)中，我们令，即传感器2比传感器1更加可靠。在这种情况下，后验期望将更加接近**y**2。在图4.14(c)中，我们令



所以传感器1在分量*y*2上要比传感器2更加可靠（垂直方向），传感器2在分量*y*1上要比传感器1更加可靠（水平方向）。在这种情况下，后验期望使用**y**1的垂直分量和**y**2的水平分量。

值得注意的是，这个技术关键之处在于它依赖于我们对每个传感器不确定度的建模。如果使用无加权的平均值进行计算将导致错误的结果。然而，我们已经假设每个传感器的精度已知。如果它们的不确定度未知，我们需要对不确定度和进行建模。章节4.6.4将介绍更多细节。



（a） （b）

图4.15 含噪数据的插值（噪音方差等于1）。（a）=30；（b）=0.01.图形由程序**gaussInterpNoisyDemo**生成。

4.4.2.3 含噪数据的插值

我们现在重新考虑章节4.3.2.2中介绍的案例。此刻我们不再假设观测值是无噪声的。取而代之的是，假设我们有*N*个含噪的观测值*y*i；不失一般性，假设观测值对应的真实值分别对应*x*1,…,*xN*。我们使用线性高斯系统对这种情况进行建模：

 （4.145）

其中，为观测噪音，**A**为*N*×*D*的投影矩阵，它用于选择哪些值是观测值。比如说，如果*N*=2，*D*=4，我们有：

 （4.146）

像之前一样，我们使用一个不合适的先验分布，令，我们可以很容易地计算出后验分布的期望和方差。在图4.15中，我们绘制了后验分布的期望，方差和一些从后验分布中采样的样本。现在我们发现先验分布精度对后验分布期望和方差都有影响。特别的，对于一个强先验（较大），估计值将会很平滑，不确定度很低。但对于一个弱先验（较小），估计值很扭曲，不确定度（远离观测值的）会很高。

后验期望也可以通过解决下面的优化问题实现求解：

 （4.147）

其中为了符号上书写便捷，我们定义*x*0=*x*1和*xD*+1=*xD*。上述优化问题是如下问题的一个离散化近似：

 （4.148）

其中为函数*f*的一阶导。第一项衡量了对数据的拟合程度，第二项惩罚那些过于扭曲的函数。这是**吉洪诺夫正则化**（Tikhonov regularization）的一个例子，它是**泛函数据分析**（functional data analysis）中的一个著名方式。第15章将给出更加复杂的方法，它们将实现更高的阶平滑度（所以最终的样本不再那么呈现锯齿状）。

4.4.3 结果的证明\*

现在我们推导公式4.124。其基本思想是推导联合概率分布，然后使用4.3.1节介绍的结果去计算***p***(**x**|**y**)。

更多的细节我们将在下面展示。联合概率分布的对数为（舍去那些无关的常数项）：

 （4.149）

这显然是一个联合高斯分布，因为它是二次项形式的指数。

将二次项中包含**x**和**y**的项展开，忽略那些线性项和常数项，我们有：

 （4.150）

 （4.151）

 （4.152）

其中联合分布的精度矩阵定义为：

 （4.153）

根据式4.68，使用事实，我们有：

 （4.154）

 （4.155）

 （4.156）

 （4.157）

4.5 拓展：威舍特分布**\***

威舍特分布是伽玛分布（译者注：伽玛分布的支撑集为{*x*|*x* > 0}）向正定矩阵的一种泛化。相关书籍指出“在多变量统计领域中，从重要性和实用性来说，威舍特分布仅次于（多变量）正态分布”。我们将使用该分布来对协方差矩阵或者它们的逆矩阵的不确定度进行建模。

威舍特分布的概率密度函数定义为：

 （4.158）

其中表示“自由度”，**S**表示“尺度矩阵”。（我们不久将会对这些参数给出直观的感受。）该分布的归一化常数（需要对所有对称的正定矩阵进行积分）为下面那个令人生畏的表达式：

 （4.159）

其中表示**多变量伽玛函数**（multivariate gamma function）：

 （4.160）

所以，且

 （4.161）

归一化常数只在时存在（所以概率密度函数也只在这种情况下是被良好定义的）。

威舍特分布与高斯分布之间存在一定的关联。特别的，令，则散步矩阵服从威舍特分布：。因此，。更一般的情况，威舍特分布的期望和众数为：

 （4.162）

其中众数只在时存在。

如果*D*=1，则威舍特分布将退化为伽玛分布：

 （4.163）

4.5.1 逆威舍特分布

在第二章我们指出，如果，则。类似的，如果，那么，其中IW被称为**逆威舍特**（inverse Wishart），它是逆伽玛分布向多维空间的泛化。对于和（**译者注：表示S为正定矩阵**），逆威舍特定义为：

 （4.164）

 （4.165）

逆威舍特分布具备如下的性质：

 （4.166）

如果*D* = 1，该分布将退化为逆伽玛分布：

 （4.167）

4.5.2 威舍特分布的可视化\*

因为威舍特分布是定义在矩阵上的密度函数，所以很难绘制出它的概率密度函数。然而，我们可以很容易地从该分布上进行采样，在2维空间的情况

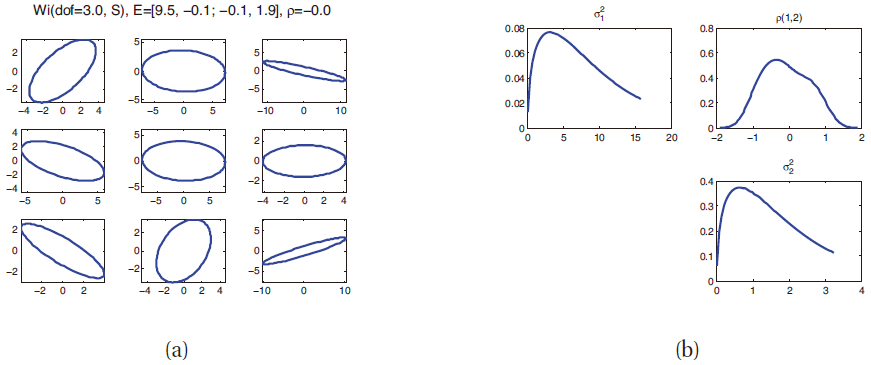


图4.16 威舍特分布的可视化。左图：从威舍特分布中采样得到的一些（二维）矩阵，其中**S**=[3.1653,-0.0262;-0.0262,0.6477]，。右图：边缘分布（服从Gamma分布）的可视化，相关系数的近似（基于采样样本的）边缘分布。如果，则关于相关系数会有许多不确定性（图中大部分质量分布都在[-1,1]之间）。采样得到的矩阵具备很高的不确定性，且有一些是奇异矩阵，当不断增大时，采样得到的矩阵将更多地集中在先验值S上。（**图形来自于原书**）

下，我们可以使用采样得到的矩阵的特征向量去定义一个椭圆，这一点在4.1.2节已做解释。图4.16给出了一些案例。

对于高维矩阵，我们可以绘制出相应的边缘分布。威舍特分布矩阵的对角线服从伽玛分布，所以很容易就可以绘制出来。一般情况下，我们很难计算出非对角元素的分布，但是我们可以从这些分布中对矩阵进行采样，然后计算出经验分布。特别的，我们可以将每个样本矩阵转化为一个相关系数矩阵，然后利用蒙特卡洛近似法（2.7节）去计算期望的相关系数：

 （4.168）

其中，表示将协方差矩阵转换为相关系数矩阵：

 （4.169）

为了绘制图形的目的，然后我们可以使用核密度估计（14.7.2介绍）来产生一个单变量密度的光滑近似，图4.16给出了一些例子。

4.6 MVN的参数推理

截止目前，我们已经讨论了在高斯分布中参数已知的情况下的推理问题。现在我们讨论如何对参数本身进行推断。我们将假设数据服从分布，其中*i* =1 : *N*。每个样本被全部观测到，即不存在缺失值（章节11.6.1将介绍如何在有缺失的情况下对参数进行估计）。为了简化我们的表达，我们分三个部分推导后验分布：首先计算；其次计算；最后计算。

4.6.1 参数的后验分布

我们已经讨论了如何计算参数的MLE，现在我们讨论如何计算它的后验分布，它将有助于对参数取值的不确定度进行建模。

似然函数的形式如下：

 （4.170）

为了简单起见，我们使用一个共轭先验分布，在这种情况下，即选择高斯先验分布。特别的，如果，那么我们可以基于4.4.2.2节得出的结论去计算的后验分布：

 （4.171）

 （4.172）

 （4.173）

这与之前介绍的根据雷达的含噪“杂音”去推理目标位置的过程是一样的，只不过现在我们使用含噪的样本去推理分布的期望值（从贝叶斯角度来看，参数的不确定度与任何其他事物的不确定度没有什么区别。）

我们可以使用一个无信息先验分布，即令。在这种情况下，我们有，此时后验分布的期望等于MLE。同时后验分布方差的下降比例为1/*N*，这是频率统计学派的标准结果。

4.6.2 参数的后验分布\*

我们现在讨论如何计算。似然函数具备如下形式（**译者注：参考式4.18**）：

 （4.174）

相应的共轭先验为前文介绍的逆威舍特分布（见4.5.1节），其概率密度函数为：

 （4.175）

其中表示自由度，为对称正定矩阵。我们发现扮演的角色是先验散布矩阵，控制着先验分布的强度，因此与样本的数量*N*扮演同样的角色。

将似然函数与先验分布相乘，我们发现后验分布也是一个逆威舍特分布：

 （4.176）

 （4.177）

 （4.178）

 （4.179）

 （4.180）

总而言之，后验分布的强度为先验分布强度加上观测值数目*N*，后验的散布矩阵为先验散布矩阵加上数据的散布矩阵。

4.6.2.1 MAP估计

从式4.7中我们可以发现，是一个秩为min(*N*,*D*)的矩阵。如果*N*＜*D*，则该矩阵不是满秩矩阵，即它是不可逆的。就算*N*＞*D*，也是一个病态的矩阵（**则几乎是奇异矩阵**）。

为了解决这个问题，我们可以使用后验分布的众数（或者期望）。结果（使用与推导MLE的类似方法）表明MAP估计为：

 （4.181）

如果我们使用一个不合适的均匀先验分布，对应于，我们将得到MLE。

现在让我们考虑一个合适的包含信息的先验分布，当*D*/*N*很大（比如说大于0.1）时，这种先验分布是很有必要的。令，所以。然后我们可以将MAP估计重新写成先验分布众数和MLE的凸组合。为了看清楚这一点，令为先验分布的众数。然后后验分布的众数可以重写成：

 （4.182）

其中，它控制着向先验分布收缩的程度。

这会导致一个问题：先验分布的参数该从何处取得？通常情况下，通过交叉验证去设置。可选的方案是，我们可以使用一些文献中提供的封闭解，如果我们使用平方损失，这些解是最优的频率学派估计值。对于协方差矩阵而言，这并不是最自然的损失函数（因为它忽略了协方差矩阵是正定矩阵的约束），但它却是一个简单的估计方法。我们在后面讨论如何通过贝叶斯方法去估计。

关于先验分布的协方差矩阵，通常情况下是使用下面的先验：。在这种情况下，MAP估计为：

 （4.183）

所以我们发现那些对角元素与它们的ML估计是一致的，那些非对角元素在某种程度上向0收缩。这种技术所以被称作**收缩估计**（shrinkage estimation）或者叫**正则估计**（regularized estimation）。

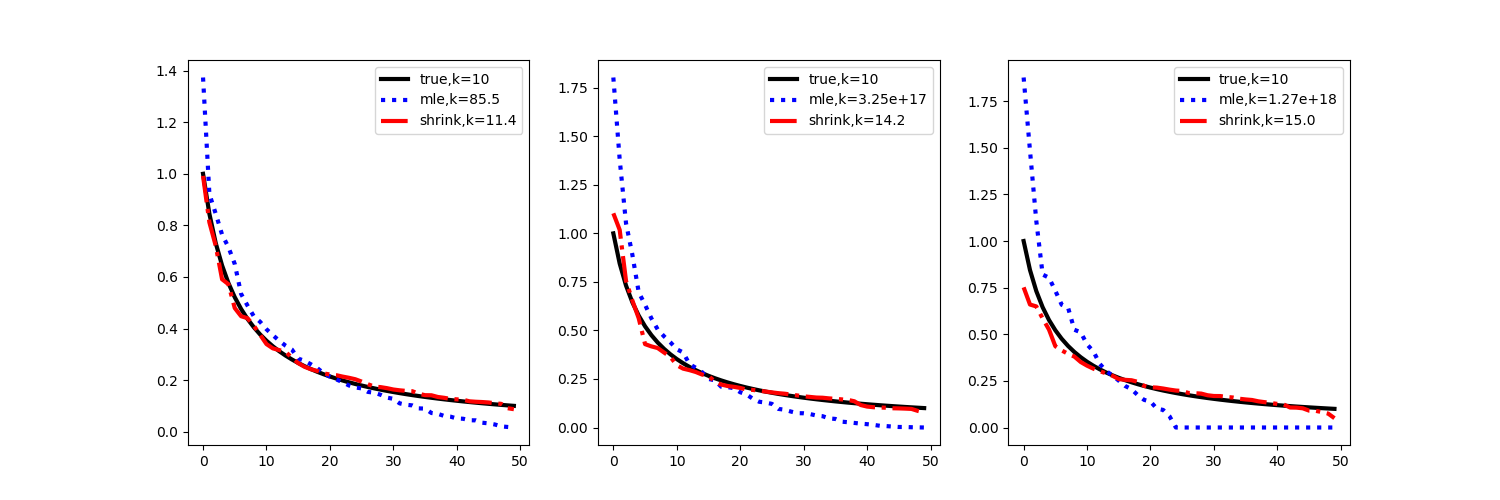


图4.17 使用*N*∈{100,50,25}个样本对一个50维的协方差矩阵进行估计。我们以降序的方式绘制了真实协方差矩阵的特征值（黑色实线），MLE的特征值（蓝色点线），MAP估计的特征值（红色实线），其中MAP估计使用式4.183（=0.9）。我们同时列出了在标签中每个矩阵的条件数。（图形由程序**shrinkcovDemo**生成）

MAP估计的好处在图4.17中作出说明。我们考虑训练一个50维的高斯分布，其中三个图分别对应的样本数量为：*N*=100, *N*=50和*N*=25。我们发现MAP估计往往是条件良好的，这一点与MLE不同。特别的，我们发现相较于MLE，MAP的**特征值谱**（eigenvalue spectrum）与真实矩阵的特征值谱更加接近。然而，特征向量不受影响。

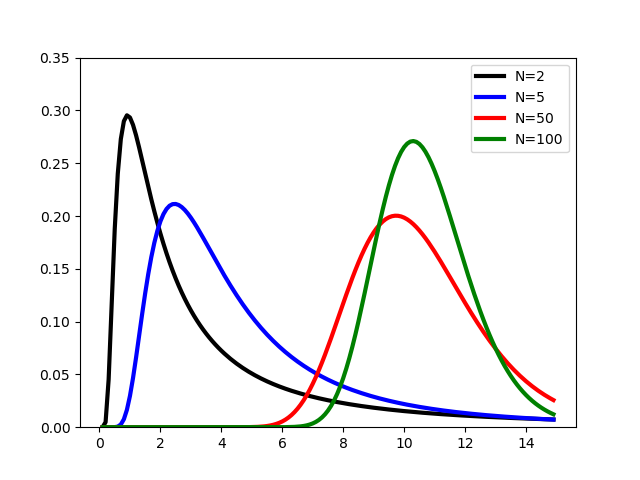


图4.18 关于参数的序列化更新，更新的步骤从无信息先验分布开始。数据从一个高斯分布中采样得到，该高斯分布的已知参数，未知参数。（**图形由程序gaussSeqUpdateSigma1D生成**）

当我们考虑对高维数据训练协方差矩阵时，对估计值进行正则化的重要性将变得很明显，这一点将在后面的章节中体现。

4.6.2.2 单变量后验分布

在1维空间中，似然函数的形式为：

 （4.184）

标准的共轭先验分布为逆伽玛分布，它是逆威舍特分布的标量版本：

 （4.185）

将似然函数与先验分布相乘，我们发现后验分布也是一个IG分布：

 （4.186）

 （4.187）

 （4.188）

图4.18给出了一些说明。

后验分布的形式显然没有在多变量情况下的那么完美，因为存在一个因子1/2。这是因为。另一个使用分布的问题在于先验分布的强度同时由*a*0，*b*0控制。为了避免这些问题，通常情况下（在统计学的相关文献中）使用IG分布的一种备选的参数化，被称作**逆卡方分布**（inverse chi-squared distribution）。定义为：

 （4.189）

其中控制着先验分布的强度，对先验分布的取值进行编码。基于这个先验分布，后验分布变成：

 （4.190）

 （4.191）

 （4.192）

我们发现后验分布的自由度为先验分布自由度加上*N*，后验分布的平方和为先验分布的平方和加上数据的平方和。

通过令=0，我们可以一个模拟一个无信息先验分布，，这与直觉是一致的（因为它对应于0虚拟样本尺寸）。

4.6.3 参数和的后验分布\*

我们现在讨论如何计算。这些结果会有一些复杂，但是在后面的内容中会很有用。在第一次阅读时可以自由地跳过该章节。

4.6.3.1 似然函数

似然函数定义为：

 （4.193）

同时我们有：

 （4.194）

所以我们可以将似然函数重新写成：

 （4.195）

 （4.196）

我们会在下面的内容中使用这个形式。

4.6.3.2 先验分布

显然我们使用如下形式的先验分布：

 （4.197）

不幸的是，这并不是似然函数的共轭先验。为了说明这一点，我们发现在似然函数中参数和并非出现在分解的因子式中；所以它们也会联合出现在后验分布中。

上述的先验分布有时也被称为**半共轭**（semi-conjugate）或者**条件共轭**（conditionally conjugate），因为两个条件分布和是单独共轭的。为了创建一个全共轭先验，我们需要使用一个参数和彼此相关的先验分布。我们将使用一个具备如下形式的联合分布作为先验分布：

 （4.198）

考察式4.196的似然函数的形式，我们发现一个自然的共轭先验具备**正态逆威舍特**（Normal-inverse-wishart）或者NIW分布的形式，定义为：

 （4.199）

 （4.200）

 （4.201）

 （4.202）

 （4.203）

 （4.204）

 （4.205）

其中为多变量伽玛函数。

NIW的参数可以作出如下解释：为参数的先验期望，反映了我们相信这个先验的强度；为（或正比于）参数的先验期望，反映了我们相信这个先验的强度。

非信息（并不合适）先验分布具备如下的形式：

 （4.206）

 （4.207）

实际上，通常情况下使用一个比较弱的含信息数据依赖先验分布是比较好的。一种常见的选择是使用和，这种设置可以确保，同时设置，并将设置为一些比较小的数字，比如说0.01。

4.6.3.3 后验分布

后验分布是一个参数更新后的NIW分布：

 （4.208）

 （4.209）

 （4.210）

 （4.211）

 （4.212）

 （4.213）

其中我们已经定义为未中心化的平方和矩阵（这比中心化版本更容易实现逐步更新）。

这个结果事实上更符合直觉：后验分布的期望是先验分布期望和MLE的凸组合，其强度为；后验散布矩阵等于先验散布矩阵加上经验散布矩阵，再加上一个因为期望不确定度而导致的额外项（它创建了自己的虚拟散布矩阵）。

4.6.3.4 后验众数

联合分布的众数具备如下的形式：

 （4.214）

如果我们令，上式将退化为：

 （4.215）

上式相应的估计值与式4.182几乎一样，但不同之处在于，1出现在分母上，因为这是联合分布的众数，而非边缘分布的众数。

4.6.3.5 后验边缘分布

参数的后验边缘分布为：

 （4.216）

边缘分布的众数和期望由下式给定：

 （4.217）

关于参数的后验边缘分布服从多变量学生T分布：

 （4.218）

上述遵循一个事实：学生T分布可以由一个尺度化的混合高斯分布表示（见式11.61）。

4.6.3.6 后验预测分布

后验预测分布由下式给定：

 （4.219）

所以它可以很容易地表达为边缘似然函数的比例。

结果表明，这个比例具备一个多变量学生-T分布的形式：

 （4.220）

 （4.221）

学生T分布比高斯分布拥有一个更宽的尾域，这考虑到了未知的事实。然而，这将会迅速地变成一个类高斯分布。

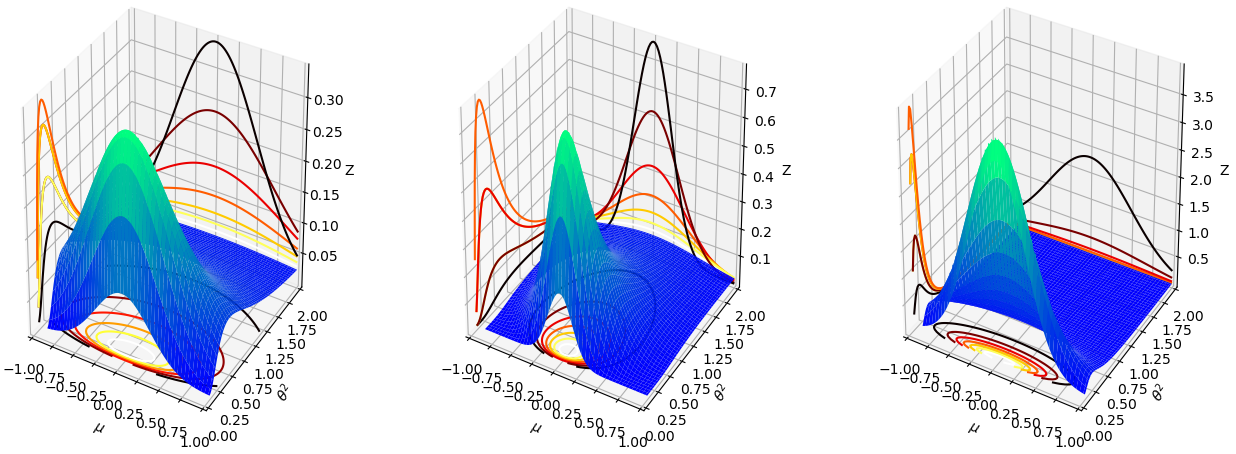


图4.19 正态逆卡方分布。为先验分布的期望，表征对该期望的信念度；表示先验的方差，表征对该方差的信念度。（a）。注意轮廓线（底部投影）形似一个“压缩的鸡蛋”。（b）我们增加了关于期望的信念强度，图形变得更加窄：。（c）我们增加了关于方差的信念强度，图形变得更加窄：。（图形由程序**NIXdemo2**生成）

4.6.3.7 标量数据的后验分布

我们现在讨论当x是一个1维变量（标量）时的情况。这些结果在统计学文献中广泛使用。与章节4.6.2.2一样，按照惯例，我们不使用正态逆威舍特分布，而是使用**正态逆卡方**（normal inverse chi-squared）分布或者NIX分布，定义为：

 （4.222）

 （4.223）

图4.19绘制了一些图形。沿着坐标轴，分布的形状类似于一个高斯分布，沿着坐标轴，分布的形状类似于；联合分布的密度是一个“压扁的鸡蛋”的形状。有趣的是，我们发现当比较小时，的轮廓线峰值更窄，这与我们的直觉是一致的，因为如果数据的方差较小，我们就可以更加可靠地估计它的期望。

其后验分布由下式给定：

 （4.224）

 （4.225）

 （4.226）

 （4.227）

 （4.228）

的后验边缘分布为：

 （4.229）

对应的后验期望为。

关于参数的后验边缘分布服从一个学生分布，它遵循一个事实：学生分布是高斯分布的（不同尺度下的）混合：

 （4.230）

其后验分布的期望为。

如果我们使用下面的非信息先验分布，那么结果看起来又是如何的呢：

 （4.231）

使用这个先验分布，其后验分布具备如下的形式

 （4.232）

其中：

 （4.233）

为**采样标准方差**（sample standard deviation）。（在章节6.4.2，我们将显示这是一个方差的无偏估计。）因此期望的边缘后验分布由下式给定：

 （4.234）

参数的后验分布方差为：

 （4.235）

上式的均方根被称为**期望的标准差**（standard error of the mean）：

 （4.236）

所以期望的95%后验**可靠区间**（credible interval）为：

 （4.237）

（贝叶斯可靠区间将在章节5.2.2进行细节讨论；我们会在章节6.6.1将其与频率学派的置信区间进行对比。）

4.6.3.8 贝叶斯t检测

在给定数据的情况下，假设我们希望检测一个假设 （），其中是某个已知值（通常是0）。这被称为双边，单采样**t-检测**。一种简单的方式是执行这个检测从而判断是否满足。如果不满足，那么我们可以95%的相信。一个常见的情景是我们希望测试两个样本是否有同样的期望。更加精确的表达，假设和。我们希望通过使用数据判断是否成立。我们可以使用下式对这个量进行估计：

 （4.238）

这被称为单边，**成对t-检测**（paired t-test）。（章节5.2.3将介绍在二项式分布数据中，成对t-检测与非成对检测的区别。）

为了计算后验分布，我们必须指定一个先验分布。假设我们使用一个非信息先验分布。正如我们在前面介绍的，我们发现参数的后验边缘分布具备如下形式：

 （4.239）

现在我们定义**t统计量**（t statistic）：

 （4.240）

上式的分母为期望的标准差。我们发现：

 （4.241）

其中为标准学生t分布的概率密度函数。

4.6.3.9 与频率学派统计的关联\*

如果我们使用一个非信息先验，结果将表明上述的贝叶斯分析方法与使用频率学派的方法一致。（我们将在第6章讨论频率学派统计的相关知识。）特别的，从上述结果，我们可以发现：

 （4.242）

这与MLE的采样分布具有相同的形式：

 （4.243）

原因在于学生分布的前两个参数是对称的，所以；因此关于参数的后验分布的表述与采样分布的表述具备一样的形式。因此，由频率学派测试返回的（单边）p值（在6.6.2节定义）与由贝叶斯方法返回的是一样的。程序**bayesTtestDemo**给出了一个案例。

尽管这两种方法表面上具有相似性，这两个结果还是有一个不一样的解释：在贝叶斯方法中，是未知的，是固定的，然而在频率学派的方法中，是未知的，而是固定的。章节7.6.3.3将介绍更多的细节。

4.6.4 未知精度下的传感器融合\*

本节，我们将使用4.6.3节得到的结果，来解决如下的问题：在每一个测量设备精度未知的情况下如何实现传感器融合。这个结果是4.4.2.2节结果的泛化版本，在之前的结果中，测量值被建模为服从精度已知的高斯分布。当精度未知时，将得到本质上不同的结果，正如我们即将看到的，我们将得到一个多模型后验分布。

假设我们想利用由多个信源发出的数据来估计某个量，但每一个信源的可靠度未知。特别地，假设我们有两个不同的测试设备*x*和*y*，这两个设备具备不一样的精度：和。基于每个设备，我们得到两个独立的测试值：

 （4.244）

我们将针对使用一个非信息先验分布：，我们可以使用一个无穷宽的高斯分布进行模拟。如果和项是已知的，那么后验分布将会是一个高斯分布：

 （4.245）

 （4.246）

 （4.247）

其中为设备*x*的测量值数目，为设备*y*的测量值数目。和。这个结果之所以成立，是因为后验分布的精度为测量设备精度的和，后验分布的期望是先验期望（等于0）与数据期望的加权和。

然而，测量精度我们是不知道的。直觉上我们可以通过最大似然去估计它们。对数似然函数由下式给定：

 （4.248）

通过求解下面的联立方程，得到MLE：

 （4.249）

 （4.250）

 （4.251）

我们得到：

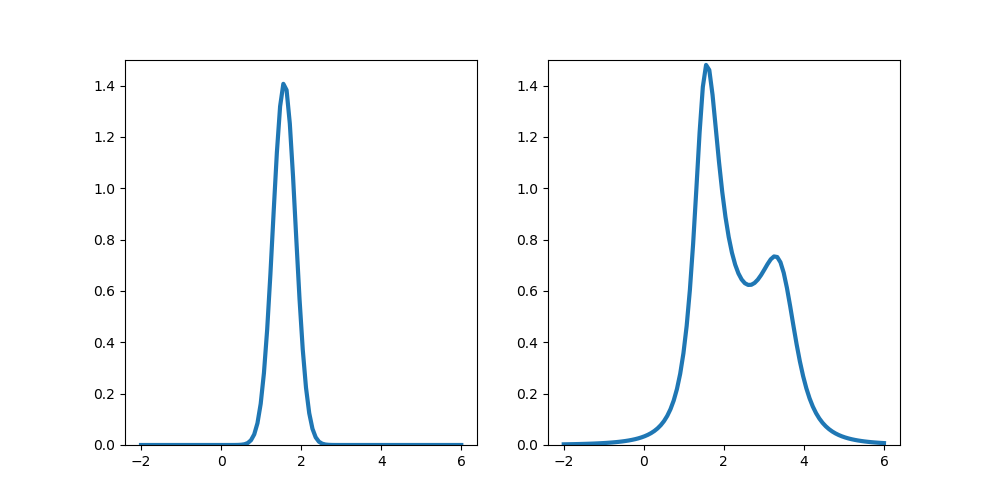
 （4.252）

 （4.253）

 （4.254）

我们发现参数的MLE与后验期望的形式一样。

我们可以通过固定点迭代的方式去解这些方程。我们首先估计和实现参数初始化，其中，。使用这些，我们得到，所以。如果我们现在迭代，我们将收敛于。



（a） （b）

图4.20 参数的后验分布。（a）点估计。（b）精确的后验分布。 （**图形由程序sensorFusionUnknownPrec生成**）

后验分布的点估计在图4.20（a）中展示。这个点估计根据估计得到的精度值对每个传感器实现加权。因为根据估计，传感器*y*的可靠度较传感器*x*低，所以我们有，我们可以有效地忽略传感器*y*。

我们现在使用贝叶斯方法，对未知的参数精度进行积分，而不是尝试估计它们的值。意味着，我们需要计算：

 （4.255）

我们使用非信息杰弗里先验分布和。因为*x*和*y*项是对称的，我们可以仅仅针对其中一个。主要的积分项为：

 （4.256）

 （4.257）

考虑当前场景，上式简化为：

 （4.258）

我们发现上式与非归一化Gamma概率密度函数的积分成比例

 （4.259）

其中*a* = 1和。因此积分项与伽玛分布的归一化常数成比例，所以我们得到：

 （4.260）

后验分布将变成：

 （4.261）

精确的后验分布在图4.20（b）中展示。我们发现它有两个峰值，一个接近，一个接近。这与如下的事实相对应：我们相信传感器*x*比*y*更加可靠。反之亦然。第一个众数的权重更大，因为从传感器*x*中获取的数据更加一致，所以传感器*x*看上去好像更加可靠。（两个传感器的可靠度显然不是一样的，因为它们所输出的数据并不一致。）然而，贝叶斯方法保留了传感器*y*更加可靠的这一可能性；从两次测量中，我们无法分辨并且正好选择传感器*x*，就像点估计所做的那样，导致过度置信（一个太窄的后验分布）。