主成分分析（PCA）详解

——以及其他机器学习经典模型

目录

1 导言

本文的主要目的在于陈述PCA的理论基础，并尽量从基础的理解层面联系到其他一系列的机器学习模型，也即数据分析算法。即便在机器学习和数据科学蓬勃发展的现在，机器学习模型也仍然面临着缺乏宏观的系统性，导致各个模型间缺少联系、单打独斗的困境。看似无关的诸多模型间的内在统一性是这篇说明背后的主题。

第二部分描述了PCA的动机。

第三部分相对比较浅显，基于最大方差理论的基本假设，虽然对其本质的诠释比较模糊，但是只需要基本的线性代数知识就可以理解。

第四部分揭示了PCA的本质，从正态分布的参数估计导出了PCA的全部内容，主要来自于PRML一书中习题2.34的注解。需要矩阵理论知识（包括矩阵性质和矩阵微积分）以及基础的机器学习理论。

第五部分开始往后是关于PCA的一系列扩展，在讨论中，我们将能够看到PCA背后的哲学如何联系到贝叶斯方法、人工神经网络、隐含元模型、Logistic回归、聚类问题乃至隐含马尔可夫模型（HMM）。

必要的数学基础知识基本上都在推导过程以及附录A中给出，为了控制篇幅，少数的遗漏部分尽量给出了参考文献。这篇文章总体上可以被认为self-cotained（不知道怎么翻译比较合适）的。只希望对于PCA的算法有所了解的人推荐详细阅读第二、三部分，理解性地阅读第四部分。想要了解PCA在更多场合中的应用的读者，推荐再阅读第六部分。希望通过阅读这篇文章，对于整个机器学习理论的框架有所把握的读者，推荐着重阅读第七第八部分以及附录BC。

2 动机

图 1: 一个UFO就是一个在三个维度上投影相差很大的形体，通过将其方差最小的维度压缩，我们近似认为UFO是一个平面的“圆盘”

主成分分析是数据降维的方法之一，旨在将输入的n维数据降至k维（$k < n$），而且新的k个维度的基是n维空间中的正交基（这种正交性来源于这样的假设：特征空间中的各个特征之间应该是独立的）。这里面k的大小一方面取决于需要，一方面取决于降维造成的信息损失的大小，基组的正交性会在过程中自然得到。

可以用这样的方法通俗地解释一下主成分分析的效果：假设我们拥有一个分布在三维空间中的数据集合，每个数据项是三维空间中的一个坐标，如果这个数据集合服从这样的特殊分布：在X-Y平面中投影为一个椭圆，而在Z方向基本都靠近于0。那么我们就可以认为，该组数据可以从三维降维至X-Y平面，从而舍弃数据集合在Z方面的震荡。在实际的情况中，整个数据分布的中心可能偏移空间参考点，可降的维度和分布的维度也往往不正交，但是这并不影响PCA的效果。

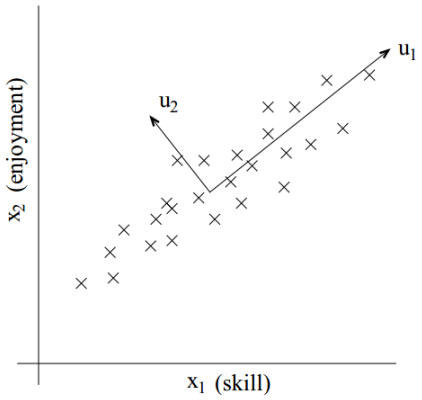

图 2: 如图所示，在正交的两个方向上，一个方向方差很大，另一个方差很小，可以考虑舍弃方差小的维度，将所有数据投影到方差大的方向上去

3 推导方法1：PCA作为最大方差理论下的一种降噪

从上文的例子中看出，在被消去的维度上，数据集合的震荡，即方差比较小。这来源于信号处理上的最大方差理论：即和信号相比，噪音的方差更小，所以通过消去方差小的维度，可以起到消除噪音的效果。换言之，在新的k个维度上，数据集合的方差较大。为了简单地计算方差，数据集合首先经线性移动，矫正至其质心和空间参考点重合，此时新的k维空间的某个基记为u，数据项记为x，我们考察x在u方向的偏移，这个值就是x在u上的投影大小，设投影的向量为ku，则由$p = x - ku$垂直于u，可得$\langle ku + p, u\rangle = \langle ku, u\rangle + \langle p, u\rangle = k = \langle x, u\rangle$，转化为内积运算。我们记数据集合为以数据项为列向量的矩阵X，其列数即数据项数目，记为N。设目标基为u_1（即在u_1方向方差最大）。则要选择u_1来最大化

$$\sum_{i=1}^{N} \langle x_i, u_1\rangle$$

也即最大化

$$\sum_{i=1}^{N} \langle x_i, u_1\rangle^2$$

将内积展开

$$\sum_{i=1}^{N}(x_i^T u_1)^2$$

等价于

$$\sum_{i=1}^{N}(u_1^T x_i)(x_i^T u_1)$$

等价于

$$u_1^T(\sum_{i=1}^{N} x_i x_i^T)u_1$$

上式中的合式即XX^T 换言之，最大化的目标函数为

$$u_1^T XX^T u_1$$

矩阵XX^T是一个对称矩阵因为$(XX^T)^T = (X^T)^T X^T = XX^T$，还是一个半正定矩阵，因为$v^T XX^T v = (X^T v)^T(X^T v) = \langle X^T v, X^T v\rangle \geq 0$，当数据集中没有线性相关项时正定。

可将XX^T分解为$Q^T \Lambda Q$，其中Q为标准正交矩阵，Λ为对角阵，对角元素为XX^T的特征值。因为u_1为单位向量，故目标式中$(Qu_1)^T\Lambda(Qu_1)$的Qu_1也是单位向量，因为$(Qu_1)^T(Qu_1) = u_1^T Q^T Qu_1 = \langle u_1, u_1\rangle = 1$。

不妨设$Qu_1 = (v_1, v_2, ...v_n)$，而且$\sum_{i=1}^{n} v_i^2 = 1$

则目标函数值为 $\sum_{i=1}^{n} v_i^2\lambda_i$

易知其最大值为取对应λ_i最大时的v_i为1，其他均为0。此时因为$Qu_1 = (1, 0....0)$，（将特征值从大到小排列，对应整理Q），则u_1为Q的元素，即对应最大特征值的特征向量。以此类推地，u_2是对应第二大特征值的特征向量..... 换句话说，可以把提取U的过程当做对XX^T进行SVD分解的过程，当对象矩阵为对称矩阵时，SVD退化为$Q^T\Lambda Q$分解。部分资料和相关实例可详见

http://mp.weixin.qq.com/s/lcEGCVOSE4$_T$pQpZmXm3WA

与上链接相比，本文直接用SVD解释了最大化目标函数的过程。

4 推导方法2：PCA作为高维正态分布的最大似然估计和参数压缩

推导方法1真正的核心和基本假设是最大方差理论，而其蕴含的二次型合式有和最小二乘法相近的形式，事实上它们确实同出一辙：二者都是高斯噪声环境下的参数估计方法。为解释协方差矩阵XX^T的意义，以下的推导旨在揭示这样一个事实：PCA的基本假设只包含了一个：数据的分布服从n维正态分布。

在已经拥有分布形式的假设后，任务就是确定该分布的决定参数，n维正态分布一般而言有共$n(n+1)$个参数，而一个好的分布只有$n+n(n+1)/2$个参数（可以将Σ分为一个对称矩阵和一个反对称矩阵的和，后者在二次型中自然消失，所以我们假设Σ有对称性），而且还有额外限制条件（一个良好定义的n维正态分布的协方差矩阵应为一个（半）正定矩阵）。下面演示从X中推断μ和Σ的过程。

μ的推断是直接的，因为高斯分布的一阶统计量可以直接用统计均值作为无偏估计，从广泛的意义上来说，贝叶斯估计不会对一阶统计量产生误差，这一点也蕴含了之前的推导方法中经过线性移动使得数据集合质心与坐标原点重合的方法之有效性。与之相比，Σ的推断比较困难。

首先写出一个参数待定的高斯函数生成X的似然函数的对数函数：

$$\ln(p(X|N(\mu,\Sigma))) = -\frac{Nn}{2}\ln(\frac{\pi}{2}) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

可以看到的是$\sum x_i$和$\sum x_i^T x_i$提供了充分信息(suffcient statistics)，换言之：

观测量 $\rightarrow$ ($\sum x_i$和$\sum x_i^T x_i$) $\rightarrow \mu, \Sigma$是一个马尔可夫过程。

对μ求导数，使之为0得到：

$$\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1} = 0$$

故

$$\mu = \frac{1}{N}\sum_{i=1}^{N}x_i$$

对Σ求偏导，有

$$-\frac{N}{2}\frac{\partial}{\partial\Sigma}\ln|\Sigma|-\frac{1}{2}\frac{\partial}{\partial\Sigma}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

对一个矩阵的行列式求导数，得到其伴随矩阵的转置，所以第一项

$$\frac{\partial}{\partial\Sigma}\ln|\Sigma|=\frac{adj(\Sigma)^T}{|\Sigma|}=\Sigma^{-1}$$

记

$$S=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

$$\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)=NTr[\Sigma^{-1}S]$$

下面分析对于Σ的每一项的导数：

$$\frac{\partial}{\partial\Sigma_{ij}}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)=N\frac{\partial}{\partial\Sigma_{ij}}Tr[\Sigma^{-1}S]=NTr[\frac{\partial}{\partial\Sigma_{ij}}\Sigma^{-1}S]$$

运用

$$\frac{\partial}{\partial x}A^{-1}=-A^{-1}\frac{\partial A}{\partial x}A^{-1}$$

，

上式可以通过

$$\frac{\partial}{\partial x}(AB)=\frac{\partial A}{\partial x}B+A\frac{\partial B}{\partial x}$$

微分$A^{-1}A=I$得到。

$$NTr[\frac{\partial}{\partial\Sigma_{ij}}\Sigma^{-1}S]=-NTr[\Sigma^{-1}\frac{\partial\Sigma}{\partial\Sigma_{ij}}\Sigma^{-1}S]=-NTr[\Sigma^{-1}S\Sigma^{-1}\frac{\partial\Sigma}{\partial\Sigma_{ij}}]$$

考察

$$\frac{\partial\Sigma}{\partial\Sigma_{ij}}$$

这是一个除ij位置为1以外，其他位置均为0的矩阵（这里暂时抛弃了Σ作为对称矩阵的假设）

所以

$$\Sigma^{-1}S\Sigma^{-1}\frac{\partial\Sigma}{\partial\Sigma_{ij}}$$

是一个第j列为$\Sigma^{-1}S\Sigma^{-1}$第i列的矩阵，其迹为该列的第j位，也即$\Sigma^{-1}S\Sigma^{-1}$的ji位，也即ij位。

所以

$$\frac{\partial}{\partial\Sigma_{ij}}Tr[\Sigma^{-1}S\Sigma^{-1}\frac{\partial\Sigma}{\partial\Sigma_{ij}}]=(\Sigma^{-1}S\Sigma^{-1})_{ij}$$

将此结论拓展至整个Σ，得到

$$\frac{\partial}{\partial\Sigma}\sum_{i=1}^{N}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)=-N\Sigma^{-1}S\Sigma^{-1}$$

令先前的偏导数为零，即：

$$-\frac{N}{2}\Sigma^{-1}+\frac{1}{2}N\Sigma^{-1}S\Sigma^{-1}=0$$

得到

$$\Sigma=S$$

可以看出，如果不是直接使得数据中心化的话，协方差矩阵将会处理这一项，这也符合正态分布的性质（先估计均值，再估计方差）。

上述的过程是从实验采集的数据还原正态分布参数的过程，这个模型的参数数量瓶颈在于协方差矩阵，将SVD应用于协方差矩阵，就能压缩协方差矩阵的秩，而这从形式上来说和推导方法1中的后半部分是一样的。

进入N维空间观察，此时正态分布呈现一个各个轴正交（因为Σ是对称的）的椭球，而其中最长轴的方向就是协方差矩阵Σ的对应最大特征值的特征向量的方向（可以将该向量代入分布函数，可以发现在该方向可以最大化概率，因为对Σ取逆使得幂中的二次型对应最大特征值的倒数，反置符号后在对数形式中最大化），次长轴对应次大特征值，以此类推。将比较短的轴压缩至零，也就等价于在Σ的奇异值分解形式里，把对角阵中较小的元素直接置零，从这里能看出，在应用SVD上，两种推导方法是一致的。

将原始数据认为是由一个正态分布所生成的假设是最一般性的假设，这是因为正态分布是限定均值和方差下的最大熵分布（仅限定均值下的最大熵分布是指数分布）。许多模型都是这一基本假设和模型任务的简单结合，除了维度规约以外，回归、分类、聚类都是如此。

4.1 回归任务中的正态分布假设

在回归任务上，对于输入$\mathbf{x}$，我们试图寻找一个函数f，使得$f(\mathbf{x})=$

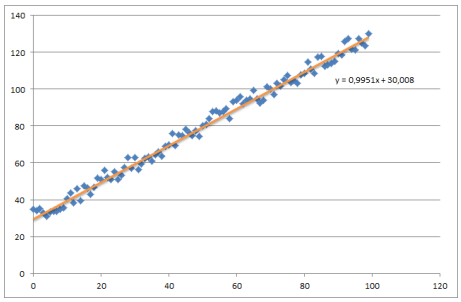

图 3: 线性回归，$\mathbf{x}$=x

y尽可能接近真实值。线性回归假设函数的形式为：

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$$

给定的材料为序对$(\mathbf{x}_n, t_n)$的集合。

我们不失一般性地假设f的输出和真实的输出之间的偏差由一个加性高斯噪声引起，此时：

$$p(\mathbf{T}|\mathbf{X}, \theta) = \prod_{n=1}^{N} N(t_n|\mathbf{w}^T\mathbf{x}_n, \sigma^2))$$

$$\ln p(\mathbf{T}|\mathbf{X}, \theta) = C_1 - C_2 \sum_{n=1}^{N}(t_n - \mathbf{w}^T\mathbf{x}_n)^2$$

平衡条件为：

$$\mathbf{w} = (\sum_{n=1}^{N} \mathbf{x}_n t_n)(\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T)^{-1}$$

通过调节θ来最大化似然概率的过程就是一个标准的回归过程，如果f是关于$\mathbf{x}, \theta$的线性函数，通过取对数后外加$\ln p(\theta)$一项，我们就得到标准的贝叶斯线性回归算法。因为正态分布的解析形式，线性回归的损失函数拥有最小二乘的形式。

4.2 分类任务中的正态分布假设

在分类任务上，我们假设：不同类型的数据元素由不同均值、同协方差的正态分布产生，考虑一个二类分类问题，我们试图计算一个给定条目$\mathbf{x}$属

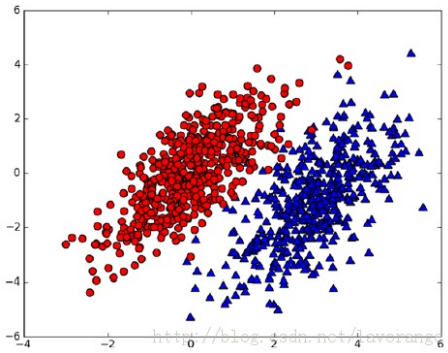

图 4: 分类模型

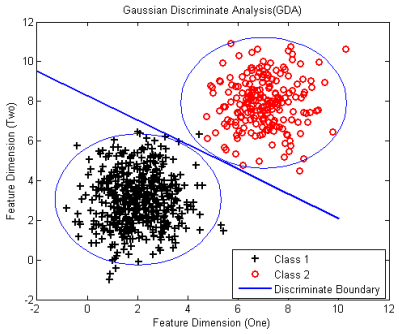

图 5: 分类边界

于某个类型的概率，反复利用贝叶斯公式：

$$p(Class_1|\mathbf{x}) = \frac{p(Class_1,\mathbf{x})}{p(\mathbf{x})} = \frac{p(Class_1,\mathbf{x})}{p(Class_1,\mathbf{x}) + p(Class_2,\mathbf{x})}$$

$$= \frac{1}{1 + \frac{p(Class_2)}{p(Class_1)}\frac{p(\mathbf{x}|Class_2)}{p(\mathbf{x}|Class_1)}}$$

记：

$$\frac{p(Class_2)}{p(Class_1)} = \exp\{z\}$$

$$\frac{p(\mathbf{x}|Class_2)}{p(\mathbf{x}|Class_1)} = \exp\left\{\mathbf{x}^T\Sigma^{-1}(\mu_1-\mu_2)\right\}\exp\left\{-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2^T\Sigma^{-1}\mu_2\right\}$$

总结一下：

$$p(Class_1|\mathbf{x}) = \frac{1}{1+\exp\{\mathbf{x}^T\mathbf{w}+y\}} = \sigma(\mathbf{x}^T\mathbf{w}+y)$$

给出Logistic Regression的基本形式，sigmoid函数其实只是对同协方差、不同均值正态分布参数的压缩表示。我们可以直接对于$\mathbf{w}, y$进行优化来给出分类器，不过如果我们不去求出μ_1, μ_2, Σ，我们的模型就没有生成实验数据的能力。

同协方差的分类模型就是线性分类模型，因为其在特征空间给出的分类边界是线性的。这也是因为关于$\mathbf{x}$的二次项被消去了，只留下一次项。如果我们假设不同的协方差，则一般而言会给出一个二次函数的决策边界。

4.3 聚类任务中的正态分布假设

在聚类任务上，我们可以认为聚类标签是分类标签的一种软近似，但是鉴于一开始没有关于标签的认识，我们需要引入隐含变量（即未被直接观测的变量）来进行建模。整个聚类过程可以被看做是对于一个混合高斯分布的最大似然估计：

$$p(\mathbf{x}) = \sum_{m=1}^{M} \pi_m N(\mathbf{x}|\mu_m, \Sigma_m)$$

$$\pi_m \geq 0$$

$$\sum_{m=1}^{M} \pi_m = 1$$

在这估计中采取一些贪婪近似（准确地说，在EM算法的E步骤将计算期望近似为直接取最大概率的情况为1，其他为0），就会得出最常见的K-means聚类算法。不过混合分布（或任何含隐含变量）的似然估计一般需要EM（Expectation Maximum）算法，EM算法本身是一个很庞大的主题，这里不做展开。

5 贝叶斯主成分分析（Bayesian PCA）

PCA本质上是一个正态分布的最大似然估计，因为略去了先验参数分布，所以将其直接作为后验估计也是可取的.自然而然的贝叶斯扩充方法是选择一个先验参数分布，协方差矩阵先验分布是$Wishart$分布，或者利用$gaussian-wishart$来同时赋予μ和Σ先验参数，无论如何，这都不过是在参数的贝叶斯推断过程中的改良，而PCA无非是在似然估计以后衔接上一个参数数量压缩工具而已。

6 人工神经网络中的PCA

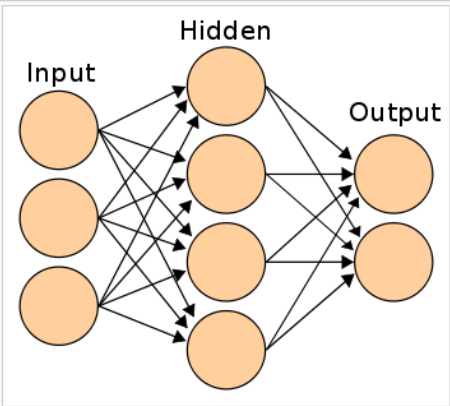

图 6: 一个基础的神经网络结构：输入层+隐含层+输出层

一个或许令人有些失望的事实是，虽然我们经过了上述的推导，达到了现有PCA的精致结果，不过事实上PCA可以在人工神经网络中自然而然地出现。关于人工神经网络（ANN）的知识这里不做补充说明。

遵从之前的符号，考虑这样一个神经网络：输入和输出各有n个单元，仅有一个隐含层包含k个单元，所有的激活函数是线性的，以$\{\mathbf{x}\}$为训练集合，以最小二乘误差为损失函数，我们试图在这个网络上训练“相等”这一语义。

对于线性的神经网络写出：

$$\mathbf{y} = \mathbf{xAB}$$

其中$\mathbf{y}$是网络的输出，而$\mathbf{A}, \mathbf{B}$分别为$n * k$和$k * n$的矩阵，显然的：

$$rank(\mathbf{AB}) \leq min(n, k) = k$$

所以可以这样认为：这个神经网络对于输入进行了一次线性变换，该变换的矩阵是秩不大于k的矩阵中，使得变换结果和原始输入相差最小的矩阵。因为整个过程都是线性的，所以梯度下降能够终止于一个全局的最优解。

从神经网络的角度可以自然地得到非线性PCA，只需要把线性单元全部替换为非线性单元即可。从网络构架的角度理解，非线性PCA是一种瓶

颈信道，从训练集中学习到的模式有助于调节瓶颈的功能性，使得经过瓶颈后相对原始输入的损失最少。

瓶颈模型对于信息压缩的一般过程可以参考一个它在马尔可夫过程中的应用（《信息论基础》习题2.14），我们假设一个过程从n种状态中的一个开始，迁移到k个状态中的一个，最后迁移到m个状态中的一个($k \leq n, m \geq n$)。瓶颈对于通信的整体干预可以表达为$I(N;M) \leq \log k$：

链式展开：

$$I(N;M,K) = I(N;M) + I(N;K|M) = I(N;K) + I(N;M|K)$$

因为$N \to K \to M$是一个马尔可夫过程：

$$I(N;M|K) = 0$$

因为互信息的非负性得证：

$$I(N;M) \leq I(N;K) = H(K) - H(K|N) \leq \log k$$

从直观上理解，神经网络的中间层即是被传递信息的上限，而这样的一个三层网络结构就是一个有损编码、解码的过程，所以我们可以在PCA网络被充分训练以后，移除输出层，转而将隐含层作为输出，连接之后的其他可调参网络结构。

在刚才的讨论中，我们一直默认网络的输入和输出为浮点型参数，不过也有一些模型以离散空间作为参数空间，譬如受限玻尔兹曼机（Restricted Boltzman Machine）仅以0和1作为可能的输入、输出。但是RBM的内在原理和PCA还是一致的，本来玻尔兹曼机的模型就来源于Ising模型，而Ising模型除了用来解释磁场，也能描述公民投票（此处投票选出的代表就是大众民意的一种有损编码）以及进行图片降噪等等。和神经网络的不同之处在于，RBM的训练一般采取散度估计而不是梯度下降，因为它严格来说是一个马尔可夫场而不是一个适应性函数。马尔可夫场是一种一般性的概率图模型（Probabilistic Graphic Model），我们会继续提到概率图模型的表示方法来泛化模型，但是受限于篇幅，不就其上的学习过程深入展开。

7 PCA的推广——Probabilistic Principal Component Analysis

7.1 动机

在先前的假设中，我们认为数据由一个高维正态分布生成，并通过对协方差矩阵做SVD分解来进行减秩。另一种类似的思路是：有一个原本是k维的元数据，它被映射到了n维上并成为了现有数据。根据这种思路提出的模型叫做Probabilistic Principal Component Analysis.

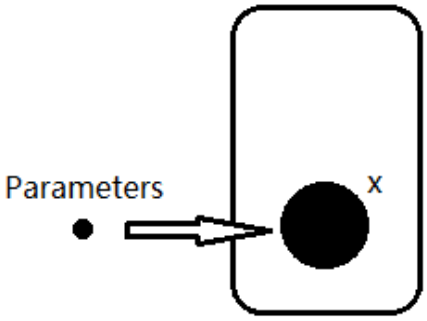

图 7: PCA的概率图模型

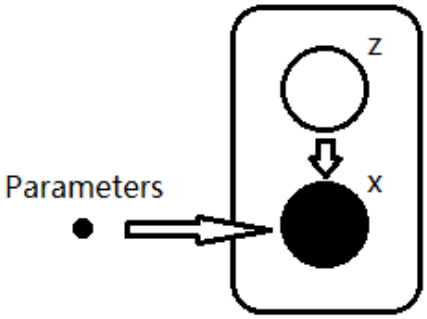

图 8: PPCA的概率图模型

7.2 模型架构

PPCA假设一个k维的隐含变量$\mathbf{z}$服从：

$$p(\mathbf{z}) = N(\mathbf{z}|\mathbf{0}, \mathbf{I})$$

而可观测变量$\mathbf{x}$服从：

$$p(\mathbf{x}|\mathbf{z}) = N(\mathbf{x}|\mathbf{A}\mathbf{z} + \mu, \sigma^2\mathbf{I})$$

此处的$\mathbf{A}, \mu, \sigma^2$为可训练参数。注意到我们并没有对$\mathbf{z}$的分布做出任何非一般性的假设，这是因为在$\mathbf{z}$空间做的任何变动都可以在$\mathbf{x}$空间中通过调节参数被等价地实现。

通过附录A中的结论：

$$p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z})d\mathbf{z} = N(\mathbf{x}|\mu, \mathbf{A}\mathbf{A}^T + \sigma^2\mathbf{I})$$

$$p(\mathbf{z}|\mathbf{x}) = N(\mathbf{z}|(\mathbf{A}^T\mathbf{A} + \sigma^2\mathbf{I})^{-1}\mathbf{A}^T(\mathbf{x} - \mu), \sigma^{-2}(\mathbf{A}^T\mathbf{A} + \sigma^2\mathbf{I}))$$

7.3 参数推断

我们首先尝试使用最大似然估计对可变参数求值，似然概率的对数形式为：

$$\ln p(\mathbf{X}|\mathbf{A}, \mu, \sigma^2) = \sum_{n=1}^{N} \ln p(\mathbf{x}_n|\mathbf{A}, \mu, \sigma^2)$$

将左侧和式的每一项展开，对于μ的求解很简单，因为关于它的导数为零条件是线性的。优化$\mathbf{A}$和σ^2的方法有两种，一种是和提出PPCA的原文里一致的直接求导并分情况讨论，其难度并不高于本文第三部分的推理过程；另一种是使用含隐含元模型的通用方法——EM算法。

EM算法的总体框架如下：

首先，估计隐含变量的分布$p(\mathbf{Z}|\mathbf{X}, \theta_{old})$

其次，将$\{\mathbf{X}, \mathbf{Z}\}$视作“完整的数据集合”，并在$\mathbf{z}$服从上一步的分布的情况下，优化参数：

$$\theta_{new} = max_\theta \{E_\mathbf{z}[p(\{\mathbf{X}, \mathbf{Z}\}|\theta)]\}$$

这一步骤中，常常利用：

$$E_\mathbf{z}[\ln p(\mathbf{X}, \mathbf{Z}|\theta)] = E_\mathbf{z}[\sum_{n=1}^{N} \ln p(\mathbf{z}_n|\theta) + \ln p(\mathbf{x}_n|\mathbf{z}_n, \theta)]$$

并通过内移期望算子来简化运算，使得联合分布的期望最大化转化为先求出$\mathbf{z}$相关变量的期望，再最优化的过程。

譬如在PPCA的例子中，我们可以继续写出：

$$E_{\mathbf{z}}[\ln p(\mathbf{X},\mathbf{Z}|\theta)] = \sum_{n=1}^{N} -\frac{D\ln 2\pi\sigma^2}{2} - \frac{1}{2\sigma^2}(\mathbf{x}_n-\mu)^T(\mathbf{x}_n-\mu)$$

$$-\frac{1}{2\sigma^2}E_{\mathbf{z}}[\mathbf{z}_n^T\mathbf{A}^T\mathbf{A}\mathbf{z}_n] + \frac{1}{\sigma^2}E_{\mathbf{z}}[\mathbf{z}_n^T](\mathbf{x}_n-\mu) + \frac{1}{2}E_{\mathbf{z}}[\mathbf{z}_n^T\mathbf{z}_n]$$

第二行中的三个期望值可由$p(\mathbf{Z}|\mathbf{X})$直接得到，之后的参数更新过程如下给出：

$$\mathbf{A}_{new} = (\sum_{n=1}^{N}(\mathbf{x}_n-\mu)E[\mathbf{z}_n]^T)(\sum_{n=1}^{N}E[\mathbf{z}_n\mathbf{z}_n^T])^{-1}$$

$$\sigma^2_{new} = \frac{1}{ND}\sum_{n=1}^{N}(\mathbf{x}_n-\mu)^T(\mathbf{x}_n-\mu) - 2E[\mathbf{z}_n]^TA^T(\mathbf{x}_n-\mu) + Tr\left\{E[\mathbf{z}_n\mathbf{z}_n^T]A^TA\right\}$$

详细的过程可以参考PRML的习题12.15。

最后提及一点，PPCA中$p(\mathbf{x}|\mathbf{z})$分布的协方差矩阵是对角元素全等的矩阵，如果保持对角阵的性质，但是放宽对角元素全等的假设，我们可以试图从给定训练集上学习不同维度上的不同方差，这一算法叫做独立成分分析（Independent Component Analysis）。这种将同质对角矩阵泛化为一般对角矩阵的思想也在其他的模型中体现得出来，譬如将线性回归模型中的外加白噪声替换成一个有一般对角协方差矩阵的零均值分布，我们就得到了关联向量机（Relevance Vector Machine，虽然名称上和SVM很像，但其背后的设计原理天差地远。RVM的参数估计方法虽然只是贝叶斯公式的反复套用，但是形式比较复杂，这里不再展开讨论）。

8 PPCA的推广——Linear Dynamic System

8.1 动机

在本文正文最后，我们讨论现有的框架在时域上的推广。数据的时变特性是iid模型所欠缺的，iid分布数据集的联合分布可以写做：

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n)$$

而在时变模型中，我们需要考虑数据之间的相互联系：

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{X} - \mathbf{x}_n)$$

有诸多的假设可能帮助我们将过于复杂的条件分布$p(\mathbf{x}_n|\mathbf{X} - \mathbf{x}_n)$简化，譬如：

与未来的无关性：

$$p(\mathbf{x}_n|\mathbf{X} - \mathbf{x}_n) = p(\mathbf{x}_n|\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_{n-1})$$

马尔可夫性：

$$p(\mathbf{x}_n|\mathbf{x}_1, \mathbf{x}_2, ...\mathbf{x}_{n-1}) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

一般来说，加长马尔可夫链的长度（即增加条件分布中的条件数量）可以提高模型的效果，不过完全可以通过编码状态空间来不失一般性地将任意有限阶的马尔可夫过程转化为一个一阶马尔可夫过程，所以在理论上我们只讨论一阶马尔可夫过程。

因为PCA的基本假设是正态分布，那么我们很可能会设想一个这样的模型：

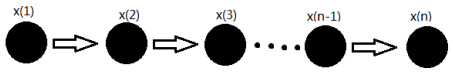

图 9: 线性贝叶斯网络——显式马尔可夫模型

在这一过程中，我们用：

$$p(\mathbf{X}) = p(\mathbf{x}_0) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

来描述联合分布。

这一个模型（显性马尔可夫模型）的缺陷在于：假设中的条件独立性（给定$\mathbf{x}_m$，则$\mathbf{x}_{m-1}$和$\mathbf{x}_{m+1}$相互独立）在现实中不一定成立，但是扩展链的长度则会使得该PGM中的边数以及模型参数数量增长过快。针对这些缺点，隐含马尔可夫模型（HMM）被提出。

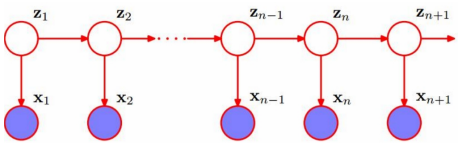

图 10: 隐含马尔可夫模型的概率图模型（z为离散变量）

利用d-separation算法，容易看出除非条件中包含$\mathbf{x}_{m_1}, \mathbf{x}_{m_2}$，否则这两项是不独立的，这就满足了对于时变模型最一般的假设。HMM的另一个优点是其PGM拥有树结构，所以允许一些递归式的算法。HMM本身是需要深入研究和探讨的重要模型，这里不过多赘述。

HMM的隐含变量一般是离散的，如果我们假设一个连续分布的隐含变量，那么我们几乎可以直接地把PPCA的单元扩展形成一个连续性HMM。

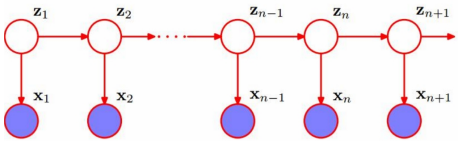

图 11: 线性动态系统的概率图模型（z为连续变量）

在PPCA中，我们已经假设$p(\mathbf{x}_n|\mathbf{z}_n)$是一个线性高斯分布，如果我们假设隐含变量的迁移概率$p(\mathbf{z}_n|\mathbf{z}_{n-1})$也是线性高斯分布的话，我们就得到了一个标准的线性动态系统（Linear Dynamic System），此时的联合分布表达为：

$$p(\mathbf{X}) = p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z}) = (\prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n))(p(\mathbf{z}_1)\prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}))$$

因为整个系统完全由正态分布和线性的条件正态分布组成，所以联合分布也是一个正态分布。

该模型的可训练参数包括$p(\mathbf{z}_1)$和$p(\mathbf{x}|\mathbf{z}), p(\mathbf{z}_n|\mathbf{z}_{n+1})$的参数，具体的训练方法和HMM很类似，可以递归地进行。

不过在实际运用中，LDS的线性假设并不总是成立，反而是HMM用转移矩阵表达的隐含关系更具普遍性，所以推荐大家进一步了解一下HMM的参数训练方法。

在应用层面，LDS相对于HMM也有一个小优势，那就是在对隐含变量进行估计时，只需要取分布$p(\mathbf{Z})$的均值即可，但是同样的任务在HMM上则需要递归地运行Viterbi算法。

9 附录A：正态分布

9.1 多维高斯分布及其几何意义

高斯分布的一维形式是：

$$N(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

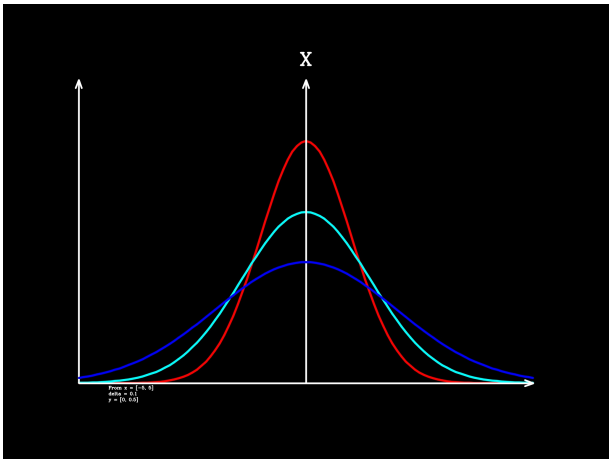

图 12: 1维高斯分布，选取不同的方差σ^2

其归一性的证明为（注意到因为积分上下限为无穷，故μ可以变参消去）：

$$I=\int\exp\left\{-\frac{1}{2\sigma^2}x^2\right\}dx$$

$$I^2=\int\int\exp\left\{-\frac{1}{2\sigma^2}x^2-\frac{1}{2\sigma^2}y^2\right\}dxdy$$

转化为极坐标，积分系数的获取是源于：

$$\|J\|=\left\|\frac{\partial(x,y)}{\partial(r,\theta)}\right\|=\left\|\begin{pmatrix}\frac{\partial x}{\partial r} & \frac{\partial y}{\partial r}\\ \frac{\partial x}{\partial\theta} & \frac{\partial y}{\partial\theta}\end{pmatrix}\right\|=r$$

所以有：

$$dxdy=rdrd\theta$$

$$x^2+y^2=r^2$$

此时积分转化为：

$$
\begin{aligned}
I^2 &= \int_0^{2\pi} \int \exp\left\{-\frac{r^2}{2\sigma^2}\right\} r dr d\theta \\
&= 2\pi \int \exp\left\{-\frac{r^2}{2\sigma^2}\right\} d(\frac{1}{2}r^2) \\
&= (-2\pi\sigma^2) \exp\left\{-\frac{r^2}{2\sigma^2}\right\} |_{r^2=0}^{r^2=\inf} \\
&= 2\pi\sigma^2
\end{aligned}
$$

代入定义式可得归一性。

在高维D，即多变量的情况下，高斯分布的形式为（此处x和μ为向量，Σ为矩阵）：

$$
N(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}
$$

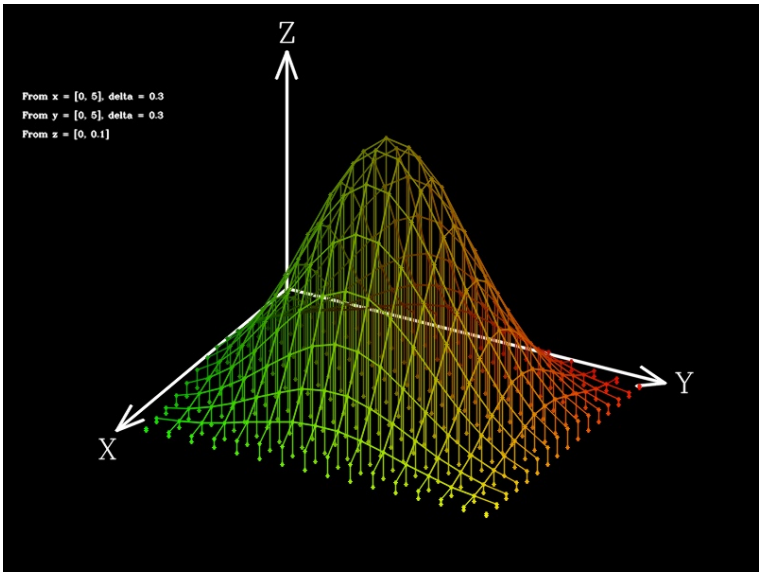

图 13: 2维高斯分布

考察协方差矩阵的逆Σ^{-1}，将其记为一个对称矩阵和一个反对称矩阵之和$\Sigma^{-1} = \Sigma_s + \Sigma_a$：

$$
u^T\Sigma^{-1}u = u^T(\Sigma_s + \Sigma_a)u = u^T\Sigma_s u + u^T\Sigma_a u
$$

对于其中第二项：

$$u^T\Sigma_a u = (u^T\Sigma_a u)^T = u^T\Sigma_a^T u = -u^T\Sigma_a u$$

所以：

$$u^T\Sigma_a u = 0$$

因此我们不妨认为协方差矩阵的逆是一个对称矩阵，所以协方差矩阵本身也是对称矩阵：

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

对称矩阵的特征值均为实数，当且仅当所有的特征值均为正实数时（即为正定矩阵时），一个正态分布是良好定义的，否则$\sqrt{|\Sigma|}$会得到无意义的结果。

从几何上可以找到直观的原因，首先：

$$\Sigma^{-1} = Q^T\Lambda Q$$

$$u^T\Sigma^{-1}u = (Qu)^T\Lambda(Qu)$$

其中Q为正交阵，Λ为对角元素全正的对角阵，该分解是对称矩阵的特征值分解，也是SVD的特例。而幂函数里的二次型可以看做是向量u经过一次正交变换后，以正定矩阵Λ定义的内积空间中的长度。容易知道的是，在该空间中的等值面，也即到原点μ等距离的包络面是一个椭球体，其在各个正交方向上的轴长由Λ的元素，也即Σ的特征值的倒数决定。

经过特征值分解以后，高维高斯分布的归一性可以通过在每个正交方向证明归一性后累乘得到，只要将$(0,0,...1,0,...0)^T$代入u，将$|\Sigma|$记做特征值之积即可。

高斯分布被广泛用于噪声和误差的分布，其原因在于：给定均值和方差两个累积量时，高斯分布是连续性分布中信息熵（又称香农熵）最大的，一个分布的信息熵定义为：

$$-\int p(x)\ln p(x)dx$$

给定以下三个约束条件：

$$\int p(x)dx = 1$$

$$\int xp(x)dx = \mu$$

$$\int (x-\mu)^2 p(x)dx = \sigma^2$$

求熵的极值：

$$F = \int \left\{-p(x)\ln p(x) + \lambda_1 p(x) + \lambda_2 xp(x) + \lambda_3(x-\mu)^2 p(x)\right\} dx - \lambda_1 - \lambda_2\mu - \lambda_3\sigma^2$$

该泛函的核函数为：

$$G(x, y, y') = -p(x)\ln p(x) + \lambda_1 p(x) + \lambda_2 xp(x) + \lambda_3(x-\mu)^2 p(x)$$

设置条件$\frac{\partial G}{\partial y} - \frac{d}{dx}(\frac{\partial G}{\partial y'}) = 0$，则：

$$\ln p(x) - \lambda_2 x - \lambda_3(x-\mu)^2 - \lambda_1 + 1 = 0$$

即：

$$p(x) = \exp\left\{-1 + \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2\right\}$$

将上式依次代入三个约束条件中，即可证明标准高斯分布导致的拉格朗日乘子满足条件：

$$\lambda_1 = 1 - \frac{1}{2}\ln(2\pi\sigma^2)$$

$$\lambda_2 = 0$$

$$\lambda_3 = \frac{1}{2\sigma^2}$$

这使得高斯分布往往被作为“可能的最差情况”，纳入通信系统和机器学习任务的分析中。事实上还可以证明，在输出信号和信道加性噪声均为限功率的连续信号的情况下，正态分布是传信率的鞍点平衡条件，即发信者为了发送最多的信息，应该以正态分布进行发信，而干扰者为了最大化地干扰通信，也应该施加一个呈正态分布的噪音。

9.2 多维高斯分布下向量的条件概率和边缘概率

本节中考察的第一个问题是：对于一个高维高斯分布而言，如果已经确定了x其中的一部分，剩下部分服从什么样的分布。不失一般性地，我们如下分割各个参数（这里不显式注明对称性）：

$$x = \begin{pmatrix} x_a \\ x_b \end{pmatrix}$$

$$\mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$\Lambda = \Sigma^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix}$$

考察幂函数里的二次型：

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}\begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}^T \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \begin{pmatrix} x_a-\mu_a \\ x_b-\mu_b \end{pmatrix}$$

$$= -\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a) - (x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b) + constant$$

constant包含了与x_a无关的项。

可以看出，上述的二次型与一般的正态分布二次型相似，一般而言：

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu + constant$$

换言之，比较二次型中x的二次项系数和一次项系数，就可以算出对应的正态分布参数，让我们类似地处理x_a的分布：

$$-\frac{1}{2}(x_a-\mu_a)^T\Lambda_{aa}(x_a-\mu_a) - (x_a-\mu_a)^T\Lambda_{ab}(x_b-\mu_b) + constant$$

$$= -\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T(\Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b))$$

所以$P(x_a|x_b)$也是一个正态分布，而且：

$$\Sigma_{a|b}^{-1} = \Lambda_{aa}$$

$$\Sigma_{a|b}^{-1}\mu_{a|b} = \Lambda_{aa}\mu_a - \Lambda_{ab}(x_b-\mu_b)$$

记$M = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$为$\Sigma^{-1}$的舒尔补的逆，则由分块矩阵求逆公式可得：

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1}+D^{-1}CMBD^{-1} \end{pmatrix}$$

$$\Lambda_{aa} = M = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}$$

$$\Lambda_{ab} = -M\Sigma_{ab}\Sigma_{bb}^{-1} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$$

所以：

$$P(x_a|x_b) = N(x_a|\mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b), \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$$

本节考察的第二个问题是边缘概率，由贝叶斯公式可得：

$$P(x_a) = \int P(x_a, x_b)dx_b$$

我们将证明它也服从高斯分布，并推导其参数。在积分符号中，首先处理关于x_b的二次型：

$$-\frac{1}{2}x_b^T\Lambda_{bb}x_b + x_b^T(\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a))$$

记$m = \Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a)$，试图对上式配平：

$$-\frac{1}{2}(x_b - \Lambda_{bb}^{-1}m)^T\Lambda_{bb}(x_b - \Lambda_{bb}^{-1}m) + \frac{1}{2}m^T\Lambda_{bb}^{-1}m$$

积分的过程仅仅对第一项有效，而且他是一个以Λ_{bb}^{-1}为协方差，$\Lambda_{bb}^{-1}m$为均值的正态分布，所以积分的结果是一个正态分布的归一系数，因此在马上分析x_a二次型的过程中不必加以考虑。现在剩下的与x_a有关的项有$\frac{1}{2}m^T\Lambda_{bb}^{-1}m$和取出$x_b$二次型时没有考虑的部分，它们在幂函数表现为和：

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) + \frac{1}{2}m^T\Lambda_{bb}^{-1}m$$

将m展开：

$$-\frac{1}{2}x_a^T\Lambda_{aa}x_a + x_a^T(\Lambda_{aa}\mu_a + \Lambda_{ab}\mu_b) +$$

$$\frac{1}{2}(\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a))^T\Lambda_{bb}^{-1}(\Lambda_{bb}\mu_b - \Lambda_{ba}(x_a - \mu_a))$$

取出x_a的二次型和一次型的系数：

$$\Sigma_a^{-1} = \Lambda_{aa} - \Lambda_{ba}^T\Lambda_{bb}^{-1}\Lambda_{ba}$$

$$\Sigma_a^{-1}\mu_a = (\Lambda_{aa} - \Lambda_{ab}\Lambda_{bb}^{-1}\Lambda_{ba})\mu_a$$

我们得到了：

$$P(x_a) = N(x_a|\mu_a, \Sigma_{aa})$$

协方差矩阵的逆被叫做Precision Matrix，在协方差矩阵中可以直接读出一对变量间是否相关（观察Σ_{ij}是否为零）而在Precision Matrix中可以直接读出在给定其他所有变量的情况下，一对变量是否相关（观察Λ_{ij}是否为零）。

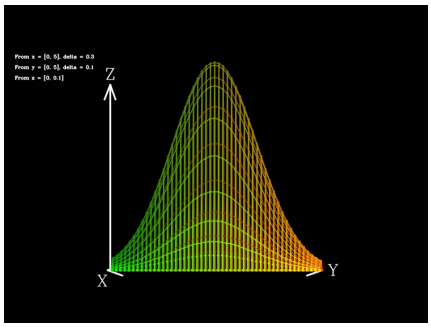

图 14: 2D Gaussian

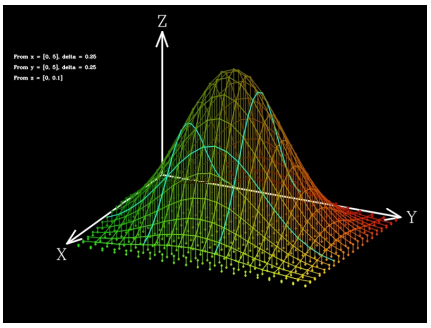

图 15: 2D Gaussian

9.3 多维高斯分布下线性关系的条件概率和边缘概率

上节探讨了，将x_a和x_b作为一个向量中的两部分时，导出的条件概率和边缘概率。这一节中我们讨论更一般的情况，如果y是x的线性函数，在给定$P(x)$和$P(y|x)$时，能否求出$P(y)$和$P(x|y)$的分析形式？

$$P(x) = N(x|\mu, \Lambda^{-1})$$

$$P(y|x) = N(y|Ax + b, Ł^{-1})$$

为了分析x, y的联合分布，我们做向量z：

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

$$P(z = \begin{pmatrix} x \\ y \end{pmatrix}) = P(x)P(y|x)$$

与之前一样，我们的观察重点在于分布的幂函数中，参变量的系数，所以取对数：

$$\ln P(z) = \ln P(x) + \ln P(y|x)$$

$$= -\frac{1}{2}(x-\mu)^T\Lambda(x-\mu) - \frac{1}{2}(y-Ax-b)^TŁ(y-Ax-b) + constant$$

下面利用矩阵，通过提取xy的二次项系数，首先整理出关于z的二次项系数：

$$-\frac{1}{2}x^T(\Lambda + A^TŁA)x - \frac{1}{2}y^TŁy + x^TA^TŁy$$

$$= -\frac{1}{2}\begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda + A^T Ł A & -A^T Ł \\ -Ł A & Ł \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = -\frac{1}{2} z^T R z$$

其中R是z的协方差矩阵的逆，利用分块矩阵求逆公式可得：

$$\Sigma_z = R^{-1} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & \Lambda^{-1} + A\Lambda^{-1}A^T \end{pmatrix}$$

类似地，找出关于x, y的一次项系数，并整理为关于z的形式：

$$x^T \Lambda \mu - x^T A^T Ł b + y^T Ł b = \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} \Lambda\mu - A^T Ł b \\ Ł b \end{pmatrix}$$

所以：

$$\mu_z = \Sigma_z \begin{pmatrix} \Lambda\mu - A^T Ł b \\ Ł b \end{pmatrix} = \begin{pmatrix} \mu \\ A\mu + b \end{pmatrix}$$

至此，我们已经得到了作为x, y联合分布的z的正态分布参数，下面的过程可以直接利用上一小节的成果：

$$P(y) = N(y|\mu_y, \Sigma_{yy}) = N(y|A\mu + b, \Lambda^{-1} + A\Lambda^{-1}A^T)$$

$$P(x|y) = N(x|\mu_x - \Lambda_{xx}^{-1}\Lambda_{xy}(y - \mu_y), \Lambda_{xx}^{-1})$$

$$= N(x|\mu - (\Lambda + A^T Ł A)^{-1}(-A^T Ł)(y - A\mu - b), (\Lambda + A^T Ł A)^{-1})$$

如果记$\Sigma = (\Lambda + A^T Ł A)^{-1}$，那么：

$$\Sigma_{x|y} = \Sigma$$

$$\mu_{x|y} = \Sigma\left\{(\Lambda + A^T Ł A)\mu + A^T Ł(y - A\mu - b)\right\}$$

$$= \Sigma\left\{\Lambda\mu + A^T Ł(y - b)\right\}$$

9.4 似然参数估计

这一节中介绍正态分布的似然估计方法，贝叶斯估计方法放在最后一节介绍。

首先考虑一维的正态分布，分别考虑在已知方差的情况下估计均值和在已知均值的情况下估计方差，以及同时估计二者。

因为：

$$P(\mu, \sigma^2|x) \propto P(x|\mu, \sigma^2)$$

对于一个iid的数据集合D：

$$P(\mu,\sigma^2|D)\propto\prod_{i=1}^{N}P(x_i|\mu,\sigma^2)$$

$$\propto\ln\prod_{i=1}^{N}P(x_i|\mu,\sigma^2)=\sum_{i=1}^{N}-\frac{1}{2}\ln 2\pi\sigma^2-\frac{(x_i-\mu)^2}{2\sigma^2}$$

$$=-\frac{N}{2}\ln 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

定义损失函数，则最大化后验概率即最大化似然概率即最小化损失函数：

$$E(\mu,\sigma^2)=\frac{N}{2}\ln 2\pi\sigma^2+\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

首先固定方差：

$$E(\mu|\sigma^2)\propto\sum_{i=1}^{N}(x_i-\mu)^2$$

$$\frac{\partial}{\partial\mu}E(\mu|\sigma^2)=C\sum_{i=1}^{N}(x_i-\mu)$$

对导数置零，就得到了固定方差下的均值估计：

$$\mu=\frac{1}{N}\sum_{i=1}^{N}x_i$$

注意到均值的似然估计中不存在和方差有关的项，所以即便没有关于方差的信息，也可以直接地先进行均值的似然估计。

对方差的导数置零得到：

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

似然估计在二阶累积量上是有差的，一个直接的结论是用因子$\frac{N-1}{N}$来校正方差的估计，不过在数据量足够大的时候，这是不重要的误差。可以证明，在平均意义上，贝叶斯学习方法中，参数的方差总是随着学习资料的增加而减小，即趋向于确定数值的。

对于一阶累积量，我们将证明：

$$E_\theta[\theta]=E_D[E_\theta[\theta|D]]$$

因为：

$$E_\theta[\theta] = \int P(\theta)\theta d\theta$$

$$E_D[E_\theta[\theta|D]] = \int \left\{ \int P(\theta|D)\theta d\theta \right\} P(D)dD$$

$$= \int \left\{ \int P(\theta|D)P(D)dD \right\} \theta d\theta$$

$$= \int P(\theta)\theta d\theta$$

对于二阶统计量，我们将证明：

$$var_\theta[\theta] = E_D[var_\theta[\theta|D]] + var_D[E_\theta[\theta|D]]$$

其中：

$$var_\theta[\theta] = E_\theta[\theta^2] - (E_\theta[\theta])^2$$

$$= \int \theta^2 P(\theta)d\theta - (\int \theta P(\theta)d\theta)^2$$

$$E_D[var_\theta[\theta|D]] = E_D[E_\theta[\theta^2|D] - (E_\theta[\theta|D])^2]$$

$$= E_D[E_\theta[\theta^2|D]] - E_D[(E_\theta[\theta|D])^2]$$

$$= \int \left\{ \int \theta^2 P(\theta|D)d\theta \right\} dD - \int \left\{ \int \theta P(\theta|D)d\theta \right\}^2 dD$$

$$= \int \theta^2 P(\theta)d\theta - \int \left\{ \int \theta P(\theta|D)d\theta \right\}^2 dD$$

$$var_D[E_\theta[\theta|D]] = E_D[E_\theta^2[\theta|D]] - (E_D[E_\theta[\theta|D]])^2$$

$$= \int \left\{ \int \theta P(\theta|D)d\theta \right\}^2 dD - (\int \left\{ \int \theta P(\theta|D)d\theta \right\} dD)^2$$

$$= \int \left\{ \int \theta P(\theta|D)d\theta \right\}^2 dD - (\int \theta P(\theta)d\theta)^2$$

将后两项相加可得证明。

明确上述等式的物理意义是很有价值的：等式左侧是参数的先验方差，也可以看做参数的先验不确定性。右侧的第一项表示平均意义上的后验方

差，因为右侧第二项非负，所以后验方差在平均意义上总是小于先验方差，也即“观测总是不增加不确定性”。如果参数均值的后验方差更大，也就意味着观测带来了更多的信息，从而后验方差将愈发减小，也即观测者对于参数的把握通过观测而增大。

9.5 贝叶斯参数估计

贝叶斯参数估计利用公式：

$$P(w|D) \propto P(D|w)P(w)$$

这是一个看似比较直接的结论，但是实际的证明并不明显，因为这个等式蕴含着，对于所有的D，它们的生成概率$P(D)$相等，这蕴含生成的数据集属于典型集内，需要利用弱大数定律来证明渐进均分性（AEP），这部分知识仍旧属于信息论的范畴。渐进均分性意味着：如果采样的次数足够多，几乎所有生成的数据集合D（不一定要iid）都是等可能的。而在典型集中进行分析允许我们进行一系列的平均近似，上方的等比例也是这一近似的结果。

通过给定$P(w)$来防止过拟合。考虑此时的损失函数：

$$E(w) = -\ln P(w|D) = C - \sum_{i=1}^{N} \ln P(x_i|w) - \ln P(w)$$

略去常数项，对于一维正态分布：

$$E(\mu, \sigma^2) = \frac{N}{2} \ln 2\pi\sigma^2 + \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 - \ln P(\mu, \sigma^2)$$

首先考虑固定方差的情况，我们希望：增加$P(w)$后，不改变损失函数的代数形式，换言之，我们希望有：

$$E(\mu|\sigma^2) = \frac{1}{2x} \sum_{i=1}^{N'} (x_i - \mu)^2$$

一个合适的方法是选择μ的分布为另一个正态分布：

$$P(\mu) = N(\mu|\mu_0, \sigma_0^2)$$

$$-\ln P(\mu) = C + \frac{1}{2\sigma_0^2}(\mu_0 - \mu)^2$$

$$E(w) = \frac{1}{2\sigma^2}\left\{\sum_{i=1}^{N}(x_i-\mu)^2 + \frac{\sigma^2}{\sigma_0^2}(\mu_0-\mu)^2\right\}$$

这等价于在对于观测集最小二乘误差的累积中增加了一项，而：

$$\mu_{post} = \frac{1}{N+\frac{\sigma^2}{\sigma_0^2}}\left\{\sum_{i=1}^{N}x_i + \frac{\sigma^2}{\sigma_0^2}\mu_0\right\}$$

$$= \mu_0\frac{\sigma^2}{N\sigma_0^2+\sigma^2} + \mu_{ML}\frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}$$

这起到的效果是：后验均值将是先验均值和最大似然均值的凸组合，组合的偏向性由先验方差决定。如果先验不确定很大，也就是说先验分布很扁平，则后验分布将偏向似然估计的结果，反之，若先验分布很肯定，先验方差小，则结果将偏向先验均值。

对于方差的后验估计：

$$E(\sigma^2|\mu) = \frac{N}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \ln P(\sigma^2)$$

$-\ln P(\sigma^2)$将有一个对数和一个倒数的形式，以下为讨论方便，取$\lambda = \sigma^{-2}$：

$$E(\lambda|\mu) = -\frac{N}{2}\ln\lambda + \frac{\lambda}{2}\sum(x-\mu)^2 - \ln P(\lambda)$$

则：

$$P(\lambda|a,b) = f(a,b)\exp\{a\lambda + b\ln\lambda\}$$

$$= f(a,b)\lambda^b\exp\{a\lambda\}$$

重整参数至符合规范的命名，并归一化，我们得到了伽马分布：

$$P(\lambda|a,b) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda) = Gam(\lambda|a,b)$$

将其带入损失函数并写出σ_{post}^2的过程是一个形式化的过程，这里不展开探讨。

我们也可以同时选择均值和方差的先验分布，因为其表征为损失函数中和式的两项，而且因为向左结合任一项都不影响函数的代数形式，所以二者的联合分布就是：

$$P(\mu,\sigma^2) = P(\mu|\sigma^2)P(\sigma^2)$$

不失一般性，我们假设$P(\mu|\sigma^2)$的方差经过了一个线性变换：

$$P(\mu,\lambda|\mu_0,\sigma_0^2,a,b)=N(\mu|\mu_0,(\beta\lambda)^{-1})Gam(\lambda|a,b)$$

对于高维高斯分布，其均值的先验分布的最优候选是另一个同维高斯分布，$Wishart$分布是协方差矩阵的先验分布：

$$W(\Lambda|w,v)=B(w,v)|\Lambda|^{(v-D-1)/2}\exp(-\frac{1}{2}Tr(w^{-1}\Lambda))$$

联合先验分布仍旧由两个先验分布的乘积给出。

9.6 指数分布族

高斯分布是更广义定义的指数分布族中的一个，而幂指数布族中成员的先验分布有一个类似的查找过程，而且通过将二项分布、多项分布等等归纳为幂分布族，我们将在分类问题中获得更一般化的分析结果。

指数分布族的一般形式为：

$$p(x|\eta)=h(x)g(\eta)\exp\left\{\eta^T u(x)\right\}$$

其中h,g,u为三个函数，而g是一个归一化算子，即：

$$g(\eta)=(\int h(x)\exp\left\{\eta^T u(x)\right\}dx)^{-1}$$

伯努利分布是指数分布族的一员：

$$Bern(x|\mu)=\mu^x(1-\mu)^{1-x}$$

$$=\exp\left\{x\ln\mu+(1-x)\ln(1-\mu)\right\}$$

$$=\exp\left\{\ln\frac{\mu}{1-\mu}x\right\}(1-\mu)$$

所以对于伯努利分布：

$$\eta=\ln\frac{\mu}{1-\mu}$$

$$h(x)=1$$

$$u(x)=x$$

$$g(\eta)=\frac{1}{1+\exp(-\eta)}=\sigma(\eta)$$

对于多项分布：

$$P(x|\mu) = \prod_{k=1}^{K} \mu_k^{x_k}$$

可以得到类似的结果，利用$\sum x = 1$的条件，可以发现η_k和μ_k的关系由softmax函数维系。

对于一维正态分布：

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\}$$

直接地：

$$\eta = \begin{pmatrix} -\frac{1}{2\sigma^2} \\ \frac{\mu}{\sigma^2} \end{pmatrix}$$

$$u(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}$$

$$g(\eta) = \exp\left\{-\frac{1}{2}\eta_2\right\} \sqrt{-\frac{\eta_1}{\pi}}$$

$$h(x) = 1$$

对于多维正态分布，$u(x)$为向量x的所有一次、二次型组合的排列。

指数分布族的估计可以通过对η微分归一式的两侧得到，对于一阶矩：

$$g(\eta) \int h(x) \exp\left\{\eta^T u(x)\right\} dx = 1$$

$$-\nabla g(\eta) \int h(x) \exp\left\{\eta^T u(x)\right\} dx = g(\eta) \int h(x) \exp\left\{\eta^T u(x)\right\} u(x) dx$$

$$-\frac{1}{g(\eta)} \nabla g(\eta) = \int P(x|\eta) u(x) dx$$

$$-\nabla \ln g(\eta) = E[u(x)]$$

对于二阶矩，只需要微分两次：

$$-\frac{1}{g(\eta)} \nabla\nabla g(\eta) - \frac{1}{g(\eta)} \nabla g(\eta) E[u(x)] = \frac{1}{g(\eta)} \nabla g(\eta) E[u(x)] + E[u(x)u(x)^T]$$

事实上，指数分布族的任意阶矩都可以通过微分来递归地得到。

而似然估计的形式也很直接，考虑iid产生的观测集X：

$$P(X|\eta) = \prod_{i=1}^{N} P(x_i|\eta)$$

$$E(\eta) = -\sum_{i=1}^{N} \ln h(x_i) - N \ln g(\eta) - \eta^T \sum_{i=1}^{N} u(x_i)$$

$$\nabla E(\eta) = -\frac{N}{g(\eta)} \nabla g(\eta) - \sum_{i=1}^{N} u(x_i)$$

$$\nabla E(\eta) = -N \nabla \ln g(\eta) - \sum_{i=1}^{N} u(x_i)$$

得到关于η的估计为：

$$-\nabla \ln g(\eta) = \frac{1}{N} \sum_{i=1}^{N} u(x_i)$$

指数分布族的贝叶斯先验分布形式为：

$$P(\eta|\chi, v) = f(\chi, v) g(\eta)^v \exp\left\{v\eta^T \chi\right\}$$

其中f为归一项，此时的后验分布为：

$$P(\eta|X, \chi, v) \propto P(X|\eta) P(\eta|\chi, v)$$

$$\propto \left\{\prod_{i=1}^{N} h(x_i) g(\eta) \exp\left\{\eta^T u(x_i)\right\}\right\} \left\{f(\chi, v) g(\eta)^v \exp\left\{v\eta^T \chi\right\}\right\}$$

$$\propto \left\{\prod_{i=1}^{N} g(\eta) \exp\left\{\eta^T u(x_i)\right\}\right\} \left\{g(\eta)^v \exp\left\{v\eta^T \chi\right\}\right\}$$

$$= \prod_{i=1}^{N+v} g(\eta) \exp\left\{\eta^T u(x_i)\right\}$$

定义了当$i \geq N$时：

$$x_i = \chi$$

可见，引入指数分布族的先验分布等价为：给指定观测集预先增加一些先验观测结果，它们的数量为v，均值为χ。

机器学习理论偏爱指数分布族是有原因的：它是基于团位势的概率图模型所能表达的唯一一种分布。

9.7 拉普拉斯近似

非混合高斯分布不能很好地模拟拥有多个模的分布，但是根据中心极限定理，我们可以用高斯分布来拟合一个单模的分布，也即对于一个非高斯的单模分布函数$f(x)$，求得高斯分布$N(x)$：

$$f(x) \to N(x)$$

根据中心极限定理，分布的模保持不变，因而我们首先求得使得f最大的x_m：

$$\frac{d}{dx}f(x)|_{x=x_m} = 0$$

$$\mu = x_m$$

因为高斯分布的对数拥有一个二次项，所以我们尝试对f的对数进行泰勒展开到二次：

$$g(x) = \ln f(x)$$

$$g(x) \approx g(x_m) + g'(x_m)(x - x_m) + \frac{1}{2}g''(x_m)(x - x_m)^2$$

因为：

$$g'(x_m) = \frac{d}{dx}\ln f(x)|_{x=x_m} = \frac{1}{f(x_m)}\frac{d}{dx}f(x)|_{x_m} = 0$$

令：

$$A = -g''(x_m)$$

则：

$$g(x) \approx g(x_m) - \frac{1}{2}A(x - x_m)^2$$

也即：

$$f(x) \approx f(x_m)\exp\left\{-\frac{1}{2}A(x - x_m)^2\right\}$$

得到：

$$N(x) = (\frac{A}{2\pi})^{1/2}\exp\left\{-\frac{1}{2}A(x - \mu)^2\right\} = N(x|x_m, A^{-1})$$

对于高维分布，类似地有：

$$N(x) = \frac{|A|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{1}{2}(x - x_m)^T A(x - x_m)\right\} = N(x|x_m, A^{-1})$$

其中A为原始分布函数对数的海森矩阵的取负：

$$A = -\nabla\nabla\ln f(x)|_{x=x_m}$$

9.8 例：麦克斯韦速率分布

本节通过举例分析麦克斯韦速率分布，来演示一下前两节中知识的应用。

麦克斯韦速率分布揭示了在一定温度条件下，一个封闭气体集合中，各种能量的微观粒子所占的比重，它的定义式是：

$$f(v) = 4\pi(\frac{m}{2\pi kT})^{\frac{3}{2}} \exp\left\{-\frac{mv^2}{2kT}\right\} v^2$$

这是一个单标量参数的分布，记$\frac{m}{kT} = \theta, v = x$：

$$P(x|\theta) = 4\pi(\frac{\theta}{2\pi})^{\frac{3}{2}} \exp\left\{-\frac{\theta x^2}{2}\right\} x^2$$

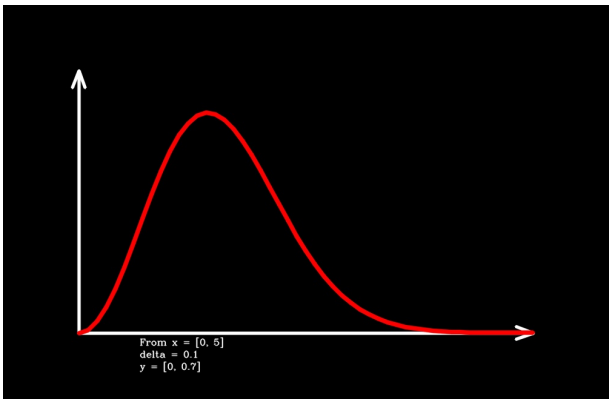

图 16: 麦克斯韦速率分布，$\theta = 1$

可以看出麦克斯韦速率分布是幂分布族的一种，满足：

$$\eta = -\frac{\theta}{2}$$

$$u(x) = x^2$$

$$h(x) = x^2$$

$$g(\eta) = 4\pi(-\frac{\eta}{\pi})^{\frac{3}{2}}$$

为求得麦克斯韦速率分布的近似高斯分布，拉普拉斯近似过程如下。首先，求目标分布的模：

$$\frac{d}{dx}P(x|\theta) = C\left\{2x \exp\left\{-\frac{\theta x^2}{2}\right\} + x^2 \exp\left\{-\frac{\theta x^2}{2}\right\}(-x\theta)\right\} = 0$$

$$x = \sqrt{\frac{2}{\theta}}$$

再求目标分布的对数的二阶导的相反数：

$$\ln P(x|\theta) = C + 2\ln x - \frac{\theta x^2}{2}$$

$$-\nabla\nabla \ln P(x|\theta) = \theta + \frac{2}{x^2} = 2\theta$$

故：

$$\mu = x = \sqrt{\frac{2}{\theta}} = \sqrt{2}$$

$$\sigma^2 = \frac{1}{-\nabla\nabla \ln P(x|\theta)|_{x=\mu}} = \frac{1}{2\theta} = 0.5$$

如图所示的是取$\sigma^2 = 0.3, \sigma^2 = 0.5$的两个同均值正态分布和原分布的对比，正态分布用蓝色画出：

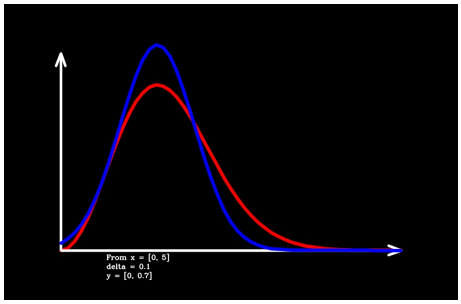

图 17: 拉普拉斯近似，$\sigma^2 = 0.3$

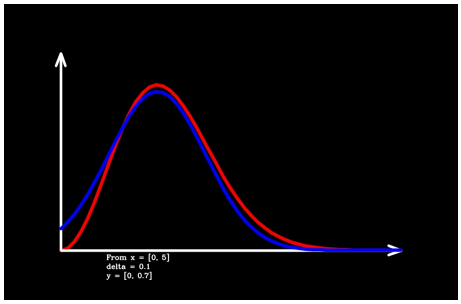

图 18: 拉普拉斯近似，$\sigma^2 = 0.5$

可以看出，拉普拉斯近似算法给出的估计更接近原始分布。

10 附录B：概率图模型

概率图模型是一种用以表达系统中变量间因果联关系的一般性框架，两种典型的概率图模型是：有向图果模型（贝叶斯网络）和无向图联模型（马尔可夫场络）。

10.1 贝叶斯网络

贝叶斯网络基于一个可以被如下拆分的联合分布：

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x}_n | Parents(\mathbf{x}_n))$$

其中$Parents(\mathbf{x})$表示的是在因果性上作为$\mathbf{x}$的原因的变量，这使得每一个$p(\mathbf{x}|Parents(\mathbf{x}))$都是有严格的因果定义的。

我们把每一个变量$\mathbf{x}_n$当成图中的一个节点，并根据$Parents()$关系连接父与子形成有向边。因为不可以有两个变量以任何形式互为因果，所以贝叶斯网络是一个无环图，通过在这个图上运行深度优先搜索获得拓扑排序以后，我们可以确保联合分布的计算和采样过程都是良好定义的（按照拓扑序采样并依次计算边缘分布并采样即可）。

如果把$p(\mathbf{x}|Parents(\mathbf{x}))$理解为一个确定性的函数，那么我们就得到了一个因果决策树的框架，在一次采样的过程中，我们都会顺着因果网络的关系由因走到果。

不过理论上决策树是允许一个变量多次出现的，它允许同一变量在不同的前提条件下处于不同的因果层次，和概率模型相比，决策树的构建是贪心的（通过计算当前条件下的最大熵特征来递归构建），没有迭代的参数训练过程，而且为了应对过拟合，需要引入一系列的剪枝方法。不过总体而言，无论是贝叶斯网络还是决策树，都提供了一种容易结合人工特征、先决知识的知识模型。

10.2 马尔可夫网络

马尔可夫网络面向的是变量间对称的关系，此时联合分布的计算可以用变量表（离散情况）、或者团势方法定义。和能够在每一个条件概率上进

行归一化的贝叶斯网络不同，马尔可夫网络的联合分布：

$$p(\mathbf{X}) = \frac{1}{Z}\prod_{r=1}^{R}\phi_r(\{\mathbf{x}\}_r)$$

需要在每一个团势（无向图的每一个完全连接的子图）上计算，并在最后一步归一。这使得马尔可夫网络的参数推断和采样形式比较复杂。

10.3 d-separation

概率图模型最大的优点就是变量间的条件独立性可以从图中直接地看出，这一判断算法叫做d-separation，为了判断给出$\mathbf{Z}$时两个变量$\mathbf{x}, \mathbf{y}$的独立性，我们试图在贝叶斯网络中寻找这样一条连接两个变量的迹，其中若满足：

没有v形结构（即一个节点，同时为其左侧节点和右侧节点的孩子），且迹上任意一节点不属于$\mathbf{Z}$；

有v形结构，且v形结构的顶点属于$\mathbf{Z}$或者$\mathbf{Z}$的后代

若存在这样一条迹则$\mathbf{x}$和$\mathbf{y}$不相互独立，读者可以运用这一条件验证HMM可观测变量间的耦合性。

马尔可夫网络上的独立性判断比较简单，只要有一条连通的迹上没有$\mathbf{Z}$的节点即可。为条件独立性付出的代价是联合分布的复杂程度。

11 附录C：层次分析法

AHP是一种复数前提下的决策框架，为了进行一个决策，我们并行地考虑多个因素，并在这些因素的指导下，对于复数个可能的选项进行评估。换言之这一贝叶斯网络中的所有有向路径都拥有“决策目的→决策因素→决策选项”的模式，而同一层次的因素之间则互不直接相连。

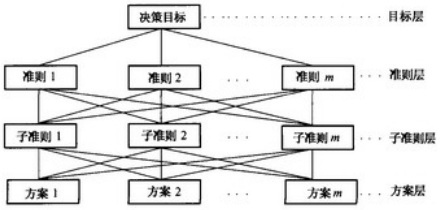

图 19: 一个层次分析法的实例

AHP嵌入式地融合了人工特征来表达一层间各个变量的重要性，对于变量a_i和a_j，它们的相对影响力由一个人工赋值$a_{ij} = \frac{a_i}{a_j}$刻画，所以对于记录了一层间所有变量相互重要性的矩阵A而言，有$a_{ij}a_{ji} = 1$，此时的A被称为成对比较矩阵。

考虑变量间重要性的传递，我们希望$a_{ij} * a_{jk} = a_{ik}$否则的话有可能出现a比b重要，b比c重要，同时c还比a重要的谬误，一个符合上述等式的矩阵被称为一致性矩阵，比方如：

$$A' = \begin{pmatrix} \frac{w_1}{w_1} & \frac{w_1}{w_2} & \cdots & \frac{w_1}{w_N} \\ \frac{w_2}{w_1} & \frac{w_2}{w_2} & \cdots & \frac{w_2}{w_N} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{w_N}{w_1} & \frac{w_N}{w_2} & \cdots & \frac{w_N}{w_N} \end{pmatrix}$$

既满足了一致性，又满足了变量间重要性比较表示为商的性质，记：

$$\mathbf{w} = (w_1, w_2, ...w_N)^T$$

显然的：

$$A'\mathbf{w} = N\mathbf{w}$$

换言之，$\mathbf{w}$是A'的对应于特征值N的特征向量，但是我们只能近似地取$A' = Q\Lambda Q^T$，其中Q是ASVD的结果，Λ是SVD对角阵仅保留最大一项的近似。

我们希望我们人工设计的重要性关联矩阵A和A'尽可能接近，但是我们很容易验证A的最大的特征值不小于n，假定A的特征值为λ对应特征向量为$\mathbf{w}$：

根据定义：

$$\sum_{j=1}^{N} a_{ij} w_j = \lambda w_i$$

对于j求和：

$$n\lambda - n = \sum_{i,j,i\neq j}^{N} a_{ij} w_j w_i^{-1}$$

则：

$$\lambda = \frac{\sum_{i,j,i\neq j}^{N} a_{ij} w_j w_i^{-1} + N}{N}$$

因为合式中的因子两两成对显现出$x + \frac{1}{x}$的形式，所以我们有：

$$\lambda \geq \frac{2 * \frac{N(N-1)}{2} + N}{N} = N$$

取等条件等价于该正互反矩阵的一致条件：

$$a_{ij} = \frac{w_i}{w_j}$$

我们主观提出的矩阵与一致性矩阵的差别记为：

$$CI = \frac{\lambda - N}{N - 1}$$

对于多个正互反矩阵，定义它们的一致性指标为：

$$RI = \frac{\sum_{m=1}^{M} CI_m}{M}$$

当我们选择的A使得$CR = \frac{CI}{RI} \leq 0.1$时，我们可以近似认为我们选择了一个有效的特征矩阵。这种一致性的判断可以在网络的各层间传递，譬如对于$a_1, ...a_N$一层后接$b_1, ...b_N$为决策对象，则这两层之间的一致性指标通过在$\mathbf{a}$上加权获得：

$$CR = \frac{\sum_{n=1}^{N} a_n CI_n}{\sum_{n=1}^{N} a_n RI_n}$$

其中CI_n和RI_n由a_n引导的B矩阵计算得出，我们习惯上以$CR \leq 0.1$为一致判定条件。

AHP的优化基本完全依赖于人工判断，虽然在因果性上有很强的说服力，但是自适应性偏弱。

不过我们可以以一种近似的方式来机械地运行类似AHP的算法，因为AHP本身就是一个类似神经网络的结构，所以我们可以使用类似的思路，将AHP的权重计算转换成一个有诸多限制条件下的最优化问题并使用梯度下降求解。

我们一一个单层结构为例，其中有N个要素，权重分别为$a_1, a_2, ...a_N$，而我们先验进行的要素间的相对重要性记为g_{ij}，则我们想要最小化的损失函数形式为：

$$J(\mathbf{a}) = \lambda_0(1 - \sum_{n=1}^{N} a_n) + \sum_{i=1}^{N}\sum_{j=1}^{N}(\frac{a_i}{a_j} - g_{ij})^2$$

为了最优化$\mathbf{a}$，我们直接求梯度：

$$\frac{\partial}{\partial a_n}J(\mathbf{a}) = -\lambda_0 + \sum_{j=1}^{N} 2(\frac{a_n}{a_j^2} - 2\frac{g_{ij}}{a_j}) + \sum_{i=1}^{N} 2(\frac{-a_i^2}{a_n^3} + \frac{g_{ij}a_i}{a_n^2})$$

选定先验权重λ和学习速率α，那么理论上，理想的$\mathbf{a}$可以用梯度下降法迭代地求解：

$$\mathbf{a}_{new} = \mathbf{a}_{old} - \alpha\nabla_{\mathbf{a}}J(\mathbf{a})$$

这种算法的基本思想和AHP相仿，但是此法中的参数可以机械地求解，而不需要诉诸于人为的尝试。

12 附录D：生成模型

本文中提到的PPCA可以看做是PCA的生成模型，机器学习模型可以分为生成模型（generative mode）和判别模型（discrimative model）两个大类，譬如本文提到的分类任务中，Logistic Regression和正态分布似然推断相比就是一个判别模型。生成模型一般拥有更多的参数，更复杂的训练过程，但是也因此拥有生成原始数据的能力。

生成模型按照是否对于生成分布显式建模、是否对于分布采取可解析形式分成几个大类：全显信念网络（类似贝叶斯网络，显式建模，可解析分布）、玻尔兹曼机（显式建模，采样模拟分布）、生成随机网络（Generative Stochastic Network，隐式建模，采样分布）、生成对抗网络（Generative Adversarial Network，隐式建模，采样分布），在附上的文章里可以笼统地了解到各个生成模型的特点和优缺点。

更一般而言，机器学习是一种用分布拟合材料数据生成分布的概率建模手段，模型框架的建立是选定一个特殊的分布，模型参数的训练是选定这个特殊分布最适合的参数，以使得模型的分布和真实数据的分布尽量地吻合。

13 延伸阅读

PPCA:"Probabilistic Principal Component Analysis, Michael E.Tipping"

AHP:"A Scaling Method for Priorities in Hierarchical Structures, Thomas L.Saaty"

Generative Models:"NIPS 2016 Tutorial: Generative Adversarial Networks, Ian Goodfellow"