联合分布和条件分布的贝叶斯建模

作者：Andriy Norets, Justinas Pelenis

摘要

本篇文章中，我们研究了一种对条件分布灵活建模的贝叶斯方法。这个方法先对独立变量和非独立变量的联合分布使用了一种灵活的建模，然后将我们想要求的的条件分布从预测的联合分布中提取出来。我们用有限个多元正态分布的混合（FMMN）来估计联合分布。条件分布可以通过分析或者模拟得到。离散变量通过隐变量处理。预测过程需要用到马尔科夫链蒙特卡洛算法（MCMC）。我们提供了对 FMMN 中 Kullback-Leibler 闭包的描述并证明了由 FMMN 模型得到的联合和条件预测密度是一大类数据的相合估计量，这些数据通常由连续的和离散的观测量产生。这个方法可以作为一种对各种类型变量（离散，连续，独立和非独立变量）的稳定的回归模型，也可以作为一种对半参数模型和非参数模型（比如分位数回归和核回归）的贝叶斯替代。在实验中，该方法特意和经典的非参数方法和可替代的贝叶斯方法做了比较。

1. 简介

本篇文章中，我们研究了一种对条件分布灵活建模的贝叶斯方法。这个方法对独立变量和非独立变量的联合分布使用了一种灵活的建模，然后将我们想要求得的条件分布从预测的联合分布中提取出来。我们用有限个多元正态分布的混合（FMMN）来估计联合分布。条件分布可以通过分析或者模拟得到。离散变量通过隐变量处理。预测过程需要用到马尔科夫链蒙特卡洛算法（MCMC）。我们提供了对 FMMN 中 Kullback-Leibler 闭包的描述并证明了由 FMMN 模型得到的联合和条件预测密度是一大类数据的相合估计量，这些数据通常由连续的和离散的观测量产生。这个方法可以作为一种对各种类型变量（离散，连续，独立和非独立变量）的稳定的回归模型，也可以作为一种对半参数模型和非参数模型（比如分位数回归和核回归）的贝叶斯替代。

在应用经济学中，条件分布的估计变得越来越重要，然而事实表明，有大量研究仍然使用分位数回归的方法，见 Koenker 和 Hallock (2001)。这个领域在整个贝叶斯框架中似乎有些被忽视了。而且，在贝叶斯领域中，似乎没有一个普遍被人们接受的回归方法可以满足对正态线性模型（比如在经典框架下有着稳定标准误差的普通最小二乘法 OLS）中各种假设的违反足够稳健。回归误差分布的形状可以用正态分布的混合来灵活逼近，见 Geweke(2005)。这个模型的异方差性可以通过乘以由于一个因素引起的误差项来调节，这个误差项取决于协变量，见 Leslie 等人.(2007)。然而，如果有其他的假定的违背，比如取决于协变量的误差分布是不对称的，那

么这个方法之后的细化就会变得很麻烦。一个有关条件分布的灵活方便的模型似乎是在贝叶斯框架下解决这些问题的一个很有吸引力的方法。

如果研究者们只对条件分布感兴趣，那么对协变量的分布的建模就显得是一件不必要的麻烦。一个对我们方法的的贝叶斯替代，即一个光滑的混合回归（SMR）同时也因多个计算机科学领域的专家而出名（见 Jacobs 等人.(1991), Jordan 和 Xu(1995), Peng 等人.(1996), Wood 等人.(2002), Geweke & Keane(2007), Villani 等人.(2009)），通过一些回归的混合直接对我们想要的条件分布建模，这些回归的混合概率通过一个多项选择模型进行建模，因此依赖于协变量。Norets(2010)和 Norets & Pelenis(2011)证实了条件密度的非参数分类可以用几个不同规格的 SMR 和有关的独立狄利克雷过程来近似并相合估计。作为和 SMR 可行结果的对比，FMMN 的结果不需要协变量分布的紧密的支持。这个基于 FMMN 的方法的一个好处。FMMN 相比于 SMR 和其他直接条件方法另外一个好处是用 MCMC 方法估计更简单。对条件密度估计的直接方法的一个好处是，在评估阶段它可以和相关协变量的筛选过程联系在一起。这个可以通过类似于 Villani 等人(2009)用到的方法来完成。我们不考虑以 FMMN 为基础的模型中的协变量选择问题。

理想上，为了估计条件密度，对直接条件方法和联合密度方法的理论比较应该建立在大量的估计误差或收敛速率上。依靠我们现有的知识，FMMN 和 SMR 的后验收敛速率都还没有求出来（单变量混合模型的后验收敛速率由 Ghosal 和 van der Vaart 在 2007 年得到）。甚至关于条件分布最优收敛速率的经典文献都是很有限的。Efromovich(2007)导出了关于单变量 x 和 y 的条件密度 f(y|x)的最小最大速率。他的结果表明，如果联合和条件密度函数同样光滑，那么它们的最小最大收敛速率就会一样，如果条件密度函数更光滑，那么它就会以更快的速率来被估计。然而，我们现在还不清楚如果联合密度估计量收敛地更慢，从中导出的条件密度估计量是否也会收敛地更慢。因此，关于条件方法和非条件方法的问题的明确的理论方案仍然有待发现，这将是未来研究中的一个重要的方向。

我们的方法是全局的，它不像一些频率学的方法，比如分位数回归中的交叉分位数，存在逻辑不一致性的问题。而且，真实数据的实验表明，FMMN 通过样本得到的预测质量比基于核的方法，DPM 和 SMR 得到的预测质量要好。

一种和我们的方法类似的方法可以通过狄利克雷过程混合（DMP）来实现。Muller 等人.(1996)和 Taddy&Kottas(2010)建议对回归问题和分位数回归问题使用 DPM 模型。然而，这些文章没有研究这些过程的理论性质。基于 DPM 的模型的一个好处是混合成分中的每一个都有正概率，因此不需要去筛选。同时，在有限的样本下，产生数据的混合成分的数量也必须是有限的，而且在估计中出现的成分的数量是由先验分布决定的。FMMN 模型的估计算法也更容易执行并且后验分布也更灵活。因此，我们选择使用 FMMN。

第二节建立了联合分布的模型并且证明了如何提取我们感兴趣的条件分布。探究模型参数后验分布的 Gibbs 取样器和估计模型预测质量的对数评分准则在第三节讨论。预测密度的一致

性在第四节展示。第五节会将我们的方法应用到几个数据集合上，这些集合都是从前用分位数回归和核方法分析过的。附录包含了对各个理论方法的证明。人工数据的实验，检查后验模拟器执行正确性的联合分布测试（Geweke,2004），计算边缘似然的算法和一些额外的估计实验都在一个网页附件里，见 Norets&Pelenis（2009）。

2. 有限正态混合模型

贝叶斯框架下的一个模型详述了观测量，非观测量和我们想要研究的对象的联合分布。首先，我们描述连续观测量的模型。之后，我们证明如何把模型推广到离散观测量的情况。

2.1 连续观测量

这个模型中的观测值用$Y_T = \{y_t, t = 1, \dots, T\}$表示，这里$y_t = (y_{t,1}, \dots, y_{t,d}) \in R^d$。在接下来的回归模型中，$y_{t,1}$是一个独立变量，$y_{t,-1} = (y_{t,2}, \dots, y_{t,d})$是协变量。观测量密度由

$$p(y_t|\theta, \mathcal{M}_m) = \sum_{j=1}^{m} \alpha_j \cdot \phi(y_t; \mu_j, H_j^{-1}) \quad (1)$$

给出，这里$\mathcal{M}$代表有 m 个混合成分的模型，$\phi(y_t; \mu_j, H_j^{-1})$是多元正态分布的密度函数，其均值为$\mu_j$，方差为$H_j^{-1}$（$H_j$被称为精度），$\alpha = (\alpha_1, \dots, \alpha_m)$是混合概率，向量$\theta = (\alpha, \mu_1, H_1, \dots, \mu_m, H_m) \in \Theta_m$收集了模型中的所有参数，$\Theta_m$是参数空间。这里我们用 3.1 节描述的（条件）共轭先验分布$p(\theta|\mathcal{M}_m)$。

预测关于 y 的联合和条件分布是我们想要研究的。预测联合分布是

$$p(y|Y_T, \mathcal{M}_m) = \int p(y|\theta, \mathcal{M}_m)\, p(\theta|Y_T, \mathcal{M}_m) d\theta \quad (2)$$

这里 $p(y|Y_T, \mathcal{M}_m)$由（1）中观测值的分布给出，$p(\theta|Y_T, \mathcal{M}_m)$是参数的后验分布。

预测条件分布是

$$p(y_1|y_{-1}, Y_T, \mathcal{M}_m) = \int p(y_1|y_{-1}, \theta, \mathcal{M}_m) p(\theta|Y_T, \mathcal{M}_m) d\,\theta$$

条件分布$p(y_1|y_{-1}, \theta, \mathcal{M}_m)$是（条件）正态分布的混合:

$$p(y_1|y_{-1}, \theta, \mathcal{M}_m) \propto \sum_{j=1}^{m} \alpha_j\, \phi(y_{-1}; \mu_{j,-1}, H_{j,-1}^{-1}) \times \phi(y_1|y_1; \mu, H_j^{-1}) \qquad (3)$$

这里$\phi(y_{-1};\mu_{j,-1},H_{j,-1}^{-1})$是联合分布$\phi(y;\mu_j,H_j^{-1})$隐含的$y_{-1}$的边缘正态分布，$\phi(y_1|y_{-1};\mu_j,H_j^{-1})$是联合分布$\phi(y;\mu_j,H_j^{-1})$隐含的$y_1$的条件正态分布，混合概率为

$$\frac{\alpha_j\phi(y_{-1};\mu_{j,-1},H_{j,-1}^{-1})}{\sum_k\alpha_k\,\phi(y_{-1};\mu_{k,-1},H_{k,-1}^{-1})}$$

我们想要研究的联合和条件密度以及它们的期望可以通过模拟来估算出来：$\theta^k\sim p(\theta|Y_T,\mathcal{M}_m)$(由后验分布得出)，$y^k\sim p(y|\theta^k,\mathcal{M}_m)$，$y_1^k\sim p(y_1|y_{-1},\theta^k,\mathcal{M}_m)$

2.2 离散观测量

由于计算方面的原因，在贝叶斯框架中我们常常用连续的隐变量对离散变量建模，见 Albert&Chib(1993)和 Geweke(2005)的第六章。同时我们用隐变量来处理离散观测量。我们用$y_c\in R^d$代表观测向量$y\in R^{d+K}$中连续的组成部分，离散的用y_{-c}表示，这里下标 c 代表连续，K 是离散变量的个数。假设第 k 个离散变量可以取N_k个不同的值，这里$k\in\{1,\dots,K\}$。我们把每一个离散变量可能的值都映射到 R 上的一个分割。因此，$y_{-c}=\left[a_{l_1}^1,b_{l_1}^1\right]\times\cdots\times\left[a_{l_K}^K,b_{l_K}^K\right]$，$l_k\in\{1,\dots,N_k\}$，对每一个$k\in\{1,\dots,k\}$，$R=\cup_{l_k=1}^{N_k}\left[a_{l_k}^k,b_{l_k}^k\right]$。对每一个离散变量，我们在模型中引入一个相应的隐变量$y_{-c}^*\in y_{-c}$。隐变量以及连续观测量的密度用正态分布的混合构建，

$$p(y_c,y_{-c}^*|\theta,\mathcal{M}_m)=\sum_{j=1}^m\alpha_j\cdot\phi(y_c,y_{-c}^*;\mu_j,H_j^{-1})\quad(4)$$

就计数测度来说，离散观测量的条件密度是一个指示函数

$$p(y_{-c}|y_c,y_{-c}^*,\theta,\mathcal{M}_m)=1_{y_{-c}}(y_{-c}^*)\quad(5)$$

在勒贝格和在参数限定下的计数测度乘积下，观测密度通过对（4）（5）式的乘积关于y_{-c}^*求积分得到。

$$p(y_c,y_{-c}|\theta,\mathcal{M}_m)=\sum_{j=1}^m\alpha_j\,\phi(y_c;\mu_{j,c},H_{j,c}^{-1})\times\int_{y_{-c}}\phi(y_{-c}^*|y_c;\mu_j,H_j^{-1})d(y_{-c}^*)\quad(6)$$

这里$\phi(y_c;\mu_{j,c},H_{j,c}^{-1})$是联合正态分布$\phi(y_c,y_{-c}^*;\mu_j,H_j^{-1})$隐含的$y_c$的边缘正态分布，$\phi(y_{-c}^*|y_c;\mu_j,H_j^{-1})$是在给定$y_c$时，联合正态分布$\phi(y_c,y_{-c}^*;\mu_j,H_j^{-1})$隐含的$y_{-c}^*$的条件正态分布。正如之前 2.1 小节中所提到的，$p(y_1|y_{-1},\theta,\mathcal{M}_m)$得出的结论可以用来求我们感兴趣的预测的条件密度。

3. 估计方法

3.1 Gibbs 取样器

参数的后验分布通过 Gibbs 取样器进行探究。对有限混合模型，Gibbs 取样器的有利的参数化包含了对隐含的状态变量的引入（Diebolt&Robert,1994）：$s_t \in \{1,\dots,m\}$，$p(y_t|s_t,\theta,\mathcal{M}_m)=\phi\left(\cdot;\mu_{s_t},H_{s_t}^{-1}\right)$以及$P(s_t=j|\theta,\mathcal{M}_m)=\alpha_j$。后验分布与观测量和非观测量的联合分布成比例

$$p\left(\{y_t,s_t\}_{t=1}^{T};\{\alpha_j,\mu_j,H_j\}_{j=1}^{m}\middle|\mathcal{M}_m\right)$$

$$\propto\prod_{t=1}^{T}\alpha_{s_t}\left|H_{s_t}\right|^{0.5}exp\{-0.5\left(y_t-\mu_{s_t}\right)'H_{s_t}\left(y_t-\mu_{s_t}\right)\}\times\alpha_1^{\underline{a}-1}\cdots\alpha_m^{\underline{a}-1}$$

$$\times\prod_j\left|H_j\right|^{0.5}exp\{-0.5\left(\mu_j-\underline{\mu}\right)'\underline{\lambda}H_j(\mu_j-\underline{\mu})\}$$

$$\times\prod_j|H|^{(\underline{\nu}-d-1)/2}exp\{-0.5tr\underline{S}H_j\}\qquad(7)$$

我们用条件共轭先验分布：Wishart 正态表示（μ_j,H_j）,狄利克雷函数表示α。超参数（$\underline{\nu}$，$\underline{S}$，$\underline{\mu}$，$\underline{\lambda}$，$\underline{a}$）必须由研究者在每一个特定应用中明确指出。我们会在第五节中提供一些建议。

Gibbs 取样器每部分的密度都和（7）中的联合分布成比例。其关于隐性状态量的部分为一个多元分布，$p(s_t=j|\dots)\propto\alpha_j\left|H_j\right|^{0.5}exp\{-0.5\left(y_t-\mu_j\right)'H_j(y_t-\mu_j)\}$。混合概率部分是狄利克雷函数，

$$p(\alpha|\dots)\propto\alpha_1^{\sum_t\mathbf{1}\{s_t=1\}+\underline{a}-1}\qquad(8)$$

混合组成部分的均值和精度由下列式子给出

$$p\left(\mu_j,H_j\middle|\dots\right)\propto\prod_{t;s_t=j}\left|H_j\right|^{(0.5)}exp\{-0.5\left(y_t-\mu_j\right)'H_j\left(y_t-\mu_j\right)\}$$

$$\times\left|H_j\right|^{0.5}exp\{-0.5\left(\mu_j-\underline{\mu}\right)'\underline{\lambda}H_j\left(\mu_j-\underline{\mu}\right)\}$$

$$\times\left|H_j\right|^{(\underline{\nu}-d-1)/2}exp\{-0.5tr\underline{S}H_j\}$$

$$\propto |H_j|^{(T_j+\underline{\nu}-d)/2} \times exp\{-0.5tr\left(H_j\left[\sum_{t:s_t=j}(y_t-\mu_j)(y_t-\mu_j)' + \underline{\lambda}(\mu_j-\mu)(\mu_j-\mu)' + \underline{S}\right]\right)\}$$

$$\propto |H_j|^{(T_j+\underline{\nu}-d)/2} \times exp\{-0.5tr(H_j[\sum_{t:s_t=j}\left(y_t-\overline{y}_j\right)\left(y_t-\overline{y}_j\right)'$$

$$+T_j\left(\overline{y}_j-\mu_j\right)\left(\overline{y}_j-\mu_j\right)' + \underline{\lambda}(\mu_j-\mu)(\mu_j-\mu)' + \underline{S}\,])\}$$

这里$T_j=\sum_j 1\{s_t=j\}$，$\overline{y}_j=T_j^{-1}\sum_{t:s_t=j}y_t$ 。因此，$p(\mu_j|H_j,\ldots)p(H_j|..)$是一个 Wishart 正态分布：

$$H_j \sim Wishart(T_j+\underline{\nu},$$

$$\left[\sum_{t:s_t=j}\left(y_t-\overline{y}_j\right)\left(y_t-\overline{y}_j\right)' + \frac{T_j\underline{\lambda}}{T_j+\underline{\lambda}}\left(\overline{y}_j-\underline{\mu}\right)\left(\overline{y}_j-\underline{\mu}\right)' + \underline{S}\right]^{-1})$$

$$\mu_j \sim N\left(\frac{T_j\overline{y}_j+\underline{\lambda}\underline{\mu}}{T_j+\underline{\lambda}}, \left[(T_j+\underline{\lambda})H_j\right]^{-1}\right)$$

我们最初选用 Wishart 正态分布作为（μ_j,）的先验分布是因为它可以帮助我们简化边缘似然函数的计算（见网页附录，Norets&Pelenis（2009））。当μ_j，H_j有独立的条件共轭先验分布时，Gibbs 取样器对μ_j就有一个正态块，对H_j就有一个 Wishart 块（此时，块的导数时相似的）。

如果观测量有离散的组成成分，那么在上述的 Gibbs 取样器中，我们可以用（$y_{t,c},y^*_{t,-c}$）来代替y_t，并在隐变量$y^*_{t,-c}$的组成部分添加块。$y^*_{t,-c}$第 k 个组成成分的块是一个被截断的正态分布，

$$p(y^*_{t,-c,k}|\ldots) \propto exp\{-0.5\left(\left(y_{t,c},y^*_{t,-c}\right)-\mu_{s_t}\right)' H_{s_t}\left(\left(y_{t,c},y^*_{t,-c}\right)-\mu_{s_t}\right)\}\cdot 1_{y_{t,-c}}(y^*_{t,-c})$$

在我们所讨论的模型中，参数的后验分布关于其标签交换是对称的。例如，的边缘后验分布对于每一个 j 都是一样的。当 m 很大时，MCMC 算法可能无法产生足够多的标签交换来得到关于（μ_j,H_j,α_j）完全一致的边缘后验分布。然而，正如 Geweke(2007)所证明的，只要我们的目标是标记不变量，那么 MCMC 中标签交换的缺乏在混合模型中通常不是问题，本文讨论的就是这种情况。

我们最初使用边缘似然（ML）来估计模型的表现。一种基于 Chib(1995)和 Marin&Robert(2008)的计算 ML 的算法请见网站附录 Norets&Pelenis(2009)。当变量的数量

很大时，特别是离散变量，那么 ML 的数值计算就会变得很不稳定。因此，我们用对数评分准则。另一个重要的原因是它们可以用于计算非贝叶斯模型，因此可以在经典代替下和 FMMN 进行有效的比较。一个完整的交叉验证的对数准则由 Gelfand 等人在 1992 年给出

$$\sum_{t=1}^{T} log\, p\left(y_t \middle| Y_{T/t}, \mathcal{M}_m\right) \approx \sum_{t=1}^{T} log(\frac{1}{N}\sum_{n=1}^{N} p\left(y_t \middle| Y_{T_t}, \theta^n, \mathcal{M}_m\right) \quad (10)$$

$$\sum_{t=1}^{T} log\, p(y_{t,i}|y_{t,-i}, Y_{T/t}, \mathcal{M}_m) \approx \sum_{t=1}^{T} log\left(\frac{1}{N}\sum_{n=1}^{N} p\left(y_{t,i}\middle|y_{t,i}, Y_{T/t},, \mathcal{M}_m\right)\right) \quad (11)$$

这里$\mathrm{Y_{T/t}}$是去除第 t 个观测量的样本，θ^{n}是从后验分布$\mathrm{p}\left(\theta\middle|\mathrm{Y_{T/t}}, \mathcal{M}_m\right)$推出来的。若联合概率分布是我们需要的，选择等式（10）；若第 i 个元素的条件分布是我们需要的，选择等式（11）（相比于 **ML**，若条件分布是我们想求的，对数评分准则还可以用来评估模型）。一个完整的交叉验证的对数评分准则要求对每个特定的模型都要有 T 个后验模拟器。一个模型化的交叉验证对数评分准则（Geweke&Keane，2007）计算上会更高效。在这个准则下，样本是随机排列的，前$\mathrm{T_1}$个观测量用于估计，其它的用于计算对数评分。重复 K 次这个过程，然后将得到的对数评分的均值或中位数用于模型的比较。下面的这个公式详述了如何计算平均对数分数

$$\frac{1}{K}\sum_{k=1}^{K}\left(\sum_{t=T_1+1}^{K} log\, p\left(y_{t,i}^{k}\middle|y_{t,-i}^{k}, Y_{T_1}^{k}, \mathcal{M}_m\right)\right) \quad (12)$$

这里$\mathrm{Y^k}$代表Y的一个新的随机排序，$p\left(y_{t,i}^{k}\middle|y_{t,-i}^{k}, Y_{T_1}^{k}, \mathcal{M}_m\right)$是式（11）中的结果。