带隐变量的回归模型EM算法

韩忠明 吕 涛 张 慧 姜同强
(北京工商大学计算机与信息工程学院 北京 100048)

摘 要 带有隐变量的回归模型具有非常广泛的应用场合,隐回归模型的参数求解问题依赖于自变量的分布假设。基于自变量的beta分布的假设条件,给出了隐回归模型的EM算法,详细地推导了模型中的参数求解过程,给出了使用牛顿法求解beta分布参数的算法,并提出一个合适的初值选择算法。在模拟数据和真实数据的基础上进行了详细的比较性试验,结果表明,对具有不同分布特征的因变量观察值,EM算法能够有效地求解隐回归模型的参数。

关键词 隐回归模型,最大期望算法,回归模型
中图法分类号 TP391.4 **文献标识码** A

EM Algorithm for Latent Regression Model

HAN Zhong-ming LV Tao ZHANG Hui JIANG Tong-qiang
(School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract There have a very wide range of applications for latent variable regression model. Estimation of the parameters of latent variable regression models depends on the assumptions of the distribution of the independent variables. Based on the Beta distribution of the independent variables, an EM algorithm for parameters estimation of latent regression model was proposed in this paper. The detailed solution process in the model was derived. Newton method for solving parameter of Beta distribution was given. Furthermore, an initial value selection algorithm was proposed. Comprehensive experiments were conducted based on simulation datasets and real dataset. The experimental results show that the EM algorithm can efficiently estimate parameters with different distribution shapes of latent regression models.

Keywords Latent regression model, EM algorithm, Regression model

近年来,图模型作为一种非常有效的统计学习模型受到了广泛的关注,应用领域越来越广泛。图模型涉及的变量较多,而受试验条件所限,变量中常存在一些无法观察到的变量,亦即隐变量。如果带隐变量的图模型联合概率分布可以获取,就可以采用EM算法求解,但是在实际情况中,带隐变量的联合概率分布难以确定,如果可以假定隐变量和观察变量之间具有特定的作用形式,例如线性关系或非线性关系,那么隐变量的求解可以转化为带有隐变量的回归问题。如图1所示,假设 $x_1, x_2, \cdots, x_n$ 是隐变量,这些隐变量作用于可以观察到的变量 y 上。假设隐变量与观察变量之间具有的作用关系为 $y=f(x_1, \cdots, x_n)$,则可以通过给定的观察变量的值 $y_1, \cdots, y_M$ 来估计隐变量的分布参数以及对应的值。非线性关系可以通过对数化等转化为线性关系,所以我们只讨论带隐变量的线性回归问题。

假设一个线性回归模型为 $y=\beta_0+\beta_1 x+\varepsilon$,如果自变量 x 为隐变量,随机误差 $\varepsilon \sim N(0, \sigma^2)$,而参数 β_0, β_1 以及 σ^2 为未知参数,则称之为隐变量回归模型。我们需要构造有效的算法来估计参数的值以及隐变量 x 的分布参数。

本质上,带隐变量的回归模型和其他学习模型间存在一些对应关系。对于带隐变量的回归模型 $y=\beta_0+\beta_1 x+\varepsilon$,如果自变量 x 为0-1分类变量,则模型转化为混合模型。如果自变量 x 服从正态分布,则模型为高斯混合模型。求解带隐变量的回归问题就是求解混合模型中的参数。

隐变量回归模型具有很多实际应用领域,现实问题中很多现象存在隐因素的作用。例如在基因网络中,不同的基因之间存在调控关系,但是很多基因的表达无法直接通过试验观察得到,所以要建模不同基因间的作用形式就会形成隐变量模型。在社会网络中,节点之间存在关系的强弱,而关系的强弱可以通过交互的强弱来体现。关系的强弱无法在试验中获取,因此关系强弱可以建模为交互强弱上的回归模型。此外,在分类问题中,如果特征无法准确度量,则也可以采用隐回归模型求解。

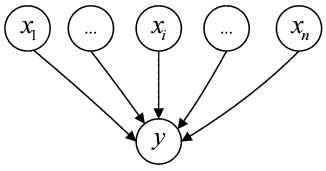

图1 带隐变量的图模型示例

到稿日期:2013-05-20 返修日期:2013-07-16 本文受国家自然科学基金(61170112),北京市属高等学校科学技术与研究生教育创新工程建设项目(PXM2012_014213_000037)资助。

韩忠明(1972-),男,博士后,副教授,主要研究方向为大数据分析与挖掘、信息检索等,E-mail:hanzm@th.btbu.edu.cn;**吕 涛**(1987-),男,硕士生,主要研究方向为大数据挖掘;**张 慧**(1989-),女,硕士生,主要研究方向为数据挖掘;**姜同强**(1966-),男,教授,主要研究方向为统计学习。

如果响应变量 y 的分布已知，可以推测隐变量 x 的分布。例如文献[2]针对响应变量 y 和隐变量 x 都满足正态分布的问题，提出了求解方法。但如果响应变量的观察值不服从正态分布，则难以假设隐变量为正态分布。文献[1]提出了一个更为普适的框架，作者假设隐变量服从 beta 分布，因为 beta 分布在不同的参数下可以呈现不同的偏态，所以它能够较好地适应响应变量的分布。文献[2]提出了求解隐回归模型的 EM 算法，但没有给出 EM 算法的完整计算过程。另外 EM 算法对参数初始值比较敏感，文献没有提出合理的初值选择方法。

本文根据文献[1]的启示以及隐回归模型的特点，详细地推导了隐变量为 beta 分布假设条件下的参数求解 EM 算法，提出了一个初值选择算法，并采用模拟数据和真实数据进行了大量试验。

1 相关工作

回归模型的求解以及相关应用研究已经非常成熟，但是对带有隐变量的回归模型的研究及其应用还相对较少。Moustaki 等[2]基于广义潜在特质模型提出了隐回归模型问题，在隐回归模型中，自变量的分布假设为标准正态分布，因变量为一元，所以自变量和因变量的联合概率分布和边缘概率分布都是正态分布。基于这个假设，文献给出了模型参数估计方法。Thaddeus 等[1]提出了基于 beta 分布的隐回归模型，给出了求解模型参数的 EM 算法，并通过很多实例说明了模型的有效性。

Aitchison 等[3]探索了不同回归模型的性质，尤其是逻辑斯蒂回归。由于逻辑斯蒂分布在不同的参数调节下，也可以呈现出不同形态，因此在隐回归模型中也可以采用逻辑斯蒂分布。如果因变量的取值不是连续值，而是离散值，例如计数值等，那么因变量的分布可以采用 beta 二项分布[5]作为隐变量的假设分布。文献[6]提出了一种用来建模分类数据的隐高斯模型，并给出了一种新的似然函数模型。

一般而言，回归模型的误差可假设为高斯白噪声，但在具有长尾效应的因变量观察数据下，随机误差可能不是标准正态分布。Bartolucci 等[4]提出了一个非常灵活的误差模型，它将误差建模为正态分布的混合模型，以便解决不同误差在不同数据点上的差异。

文献[7]提出一个广义线性隐变量模型，模型可以非常灵活地应对非对称、多模态、重尾或轻尾光滑密度的情况，通过有限参数个数的半非参数规范方法计算出广义线性隐变量模型具体应用过程中的适应度要求，其中的参数使用极大似然估计方法进行估算。文献[8]提出一个非参数的隐变量模型的非参数估计方法，并指出非参数的隐变量模型不应该指定具体的基本分布。他们首先估计一个共同的因子分析模型，并假定该模型的基本矩阵结构，然后再使用非参数回归的方法去分析隐变量之间的关系。文献[9]提出了一个更广义的隐分类回归模型，以在因素分析过程中刻画不同隐因子所产生的不同结果。

带隐变量的混合模型、图模型等是建模数据分析和数据挖掘等问题的有效方法。隐回归模型有很多应用，采用多维隐回归模型来表达异速生长扩展模型的数据具有很好的效果[10,11]，异速生长扩展模型的数据来自于不同的总体，采用隐回归模型可以很好地表达样本间的主成分。文献[12]研究了正态混合的 EM 算法在非混合均匀分布上的情况，指出混合模型需要区分有限混合分布和非混合均匀分布。有限混合分布模型使用范围广泛且常与正态密度混合，因此很难与非混合均匀分布做出有效区分。文献[13]提出了一种灵活的无限混合模型，并将其用于研究值为连续的非确定性安慰剂的治疗效果。由于安慰剂效应的研究非常复杂，产生非确定性治疗效果的因素是潜在的，因此可以用隐回归模型建模非确定性问题。文献[14]提出了一种线性混合效应模型的最优分割方法，以对安慰剂疗效进行研究。文献[15]研究了从样本协方差中导出隐结构的分析方法，样本协方差矩阵导出的隐结构可以用来表示两个观察变量的隐效应。文献[16]在对隐藏变量的分布假设要求的前提下，研究了广义线性隐变量模型，使用一种几何方法，构建了连续半参数估计量，并给出了估计公式。

2 隐回归模型

设一个一般的线性模型为：

$$y=\beta_0+\beta_1 X+\varepsilon$$

其中，$\varepsilon\sim N(0,\sigma^2)$ 为噪声，β_0，β_1 以及 σ^2 为未知参数。如果因变量 X 为一个随机变量，且通过实验无法观察得到 X 的值，而只能观察到响应变量 y 的值，通过响应变量 y 的观测值，估计参数以及 X 的分布参数，则该模型为隐回归模型。

2.1 参数求解的 EM 算法

为了简单起见，我们先讨论变量 X 为一元的随机变量，多元变量可以通过一元变量的计算方法进行推广。由于 beta 分布可以呈现不同的偏态，能够较好地适应响应变量的分布，因此假设随机变量 $x\sim beta(a,b)$，x 的分布密度函数为：

$$g(x;a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

其中，参数 a,b 是未知参数。

根据 2 个随机变量分布的卷积，可以得到随机变量 x,y 的联合概率分布密度为：

$$\begin{aligned}f(x,y)&=f(y|x;\beta_0,\beta_1,\sigma^2)g(x;a,b)\\&=N(y;\beta_0+\beta_1 x,\sigma^2)g(x;a,b)\end{aligned}\tag{1}$$

则响应变量 y 的边缘分布为：

$$\begin{aligned}f(y)&=\int_0^1 N(y;\beta_0+\beta_1 x,\sigma^2)g(x;a,b)\mathrm{d}x\\&=\int_0^1\frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{\frac{-(y-\beta_0-\beta_1 x)^2}{2\sigma^2}}g(x;a,b)\mathrm{d}x\end{aligned}\tag{2}$$

最大期望算法（Expectation-maximization，EM algorithm），在统计学习中常用来寻找依赖于隐性变量的概率模型中的参数最大似然估计。EM 算法经过两个步骤交替进行计算：第一步是期望计算(E)；第二步是最大化(M)，通过最大化 E 步上求得的极大似然值来计算参数的值。

我们采用 EM 算法求解隐回归模型中参数的值，根据响应变量 y 的边缘分布，得到对数似然函数为：

$$\begin{aligned}L(\beta_0,\beta_1,a,b,\sigma^2)=&-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\sum_{i=1}^{n}\\&\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}+\sum_{i=1}^{n}\ln(g(x_i;a,\\&b))\end{aligned}\tag{3}$$

2.1.1 E 步-期望计算

由于对数似然函数中含有隐变量 x，因此无法求得对数

似然函数值，也就无法得到参数的估计值，根据文献[2]，可以求对数似然函数在给定 y 下的条件期望：

$$E[L(\beta_0,\beta_1,a,b,\sigma^2)|y]=$$
$$-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\sum_{i=1}^{n}\frac{y_i^2-2y_i\beta_0+\beta_0^2+\beta_1E[x_i^2|y_i]-2(y_i-\beta_0)\beta_1E[x_i|y_i]}{2\sigma^2}+$$
$$\sum_{i=1}^{n}\ln(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)})+\sum_{i=1}^{n}(a-1)E[\ln(x_i)|y_i]+\sum_{i=1}^{n}(b-1)E[\ln(1-x_i)|y_i] \tag{4}$$

从上式可以看出，计算条件对数似然期望需要分别计算 x_i，x_i^2，$\ln(x_i)$ 以及 $\ln(1-x_i)$4 个随机变量函数在给定 y_i 下的条件期望值。计算随机变量函数的条件期望的一般形式为：

$$E(X|Y=y)=\int_{-\infty}^{+\infty}xf_{X|Y}(x|y)\mathrm{d}x$$

其中

$$f_{X|Y}(x|y)=\frac{f(x,y)}{f_Y(y)}$$

由于对数似然期望函数中难以直接求得变量 σ^2 以及 $\beta=(\beta_0,\beta_1)'$的极值，因此需要采用不同的方法估计参数在给定 y 下的值。

2.1.2 M 步-参数估计

对于 $\beta=(\beta_0,\beta_1)'$，我们可以单独考虑回归方程：

$$Y=\beta X+\Psi$$

其中，X 和 Y 分别表示由 x_i 和 y_i 组成的向量，Ψ 是随机噪声 ε 的协方差矩阵。这样原来估计参数 $\beta=(\beta_0,\beta_1)'$的值转化为多元回归问题的参数估计，可使用最小二乘法估计参数的值：

$$\hat{\beta}=[X'X]^{-1}X'Y$$
$$\hat{\sigma^2}=(Y'Y-2\hat{\beta}'X'Y+\hat{\beta}(X'X)\hat{\beta})/n$$

由于未知变量 X 的存在，因此我们采用条件期望代替 X，设$\widetilde{X}=E[X|Y]$，则变量估计为：

$$\hat{\beta}=[E[X'X|Y]]^{-1}\widetilde{X}'Y \tag{5}$$
$$\hat{\sigma^2}=(Y'Y-2\hat{\beta}'\widetilde{X}'Y+\hat{\beta}(E[X'X|Y])\hat{\beta})/n \tag{6}$$

其中需要计算矩阵条件期望 $E[X'X|Y]$，$X'X$ 的期望可以转化为和 X 的 X^2 期望。

接下来，构造参数 a 和 b 的估计计算方法，根据条件期望(4)，难以直接求导计算极值点，因此参数 a 和 b 的极值也无法直接得到，设

$$q(a,b)=\sum_{i=1}^{n}\ln(\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)})+\sum_{i=1}^{n}(a-1)E[\ln(x_i)|y_i]+\sum_{i=1}^{n}(b-1)E[\ln(1-x_i)|y_i] \tag{7}$$

式(7)中参数 a 和 b 的极值点也就是条件期望(4)中参数 a 和 b 的极值点，可采用牛顿法得到估计参数 a 和 b 的极值，采用牛顿法需要求出函数 $q(a,b)$对 a 和 b 的一阶和二阶偏导数，对一阶偏导有：

$$\frac{\partial q}{\partial a}=n\frac{\mathrm{d}(\ln(\Gamma(a+b)))}{\mathrm{d}a}-n\frac{\mathrm{d}(\ln(\Gamma(a)))}{\mathrm{d}a}+\sum_{i=1}^{n}E[\ln(x_i|y_i)] \tag{8}$$
$$\frac{\partial q}{\partial b}=n\frac{\mathrm{d}(\ln(\Gamma(a+b)))}{\mathrm{d}b}-n\frac{\mathrm{d}(\ln(\Gamma(b)))}{\mathrm{d}b}+\sum_{i=1}^{n}E[\ln(1-x_i|y_i)] \tag{9}$$

同理可以求出二阶偏导数：

$$\frac{\partial^2 q}{\partial a\partial b}=\frac{\partial^2 q}{\partial b\partial a}=n\frac{\mathrm{d}(\frac{\mathrm{d}(\ln(\Gamma(a+b)))}{\mathrm{d}a})}{\mathrm{d}b} \tag{10}$$
$$\frac{\partial^2 q}{\partial a^2}=n\frac{\mathrm{d}^2(\ln(\Gamma(a+b)))}{\mathrm{d}a^2}-n\frac{\mathrm{d}^2\ln(\Gamma(a))}{\mathrm{d}a^2} \tag{11}$$
$$\frac{\partial^2 q}{\partial b^2}=n\frac{\mathrm{d}^2(\ln(\Gamma(a+b))}{\mathrm{d}b^2}-n\frac{\mathrm{d}^2\ln(\Gamma(b))}{\mathrm{d}b^2} \tag{12}$$

根据函数 $q(a,b)$对 a 和 b 的一阶和二阶偏导数，可以得到牛顿法求解参数的过程，如算法 1 所示。算法 1 中涉及到对函数 $\ln(\Gamma(x))$求一阶和二阶导数。在 R[1] 系统中，函数 digamma 和 trigamma 可以直接求解一阶和二阶导数。采用矩估计法计算参数 a 和 b 的初值。

算法 1 求参数 a 和 b 的估计

输入：误差阈值 e，迭代次数 N

1. ex←$\overline{E[x_i|y_i]}$
2. ex2←$\overline{E[x_i^2|y_i]}$
3. $a_0\leftarrow\frac{ex*(1-ex)^2}{(ex2-ex^2)}+ex-1$
4. $b_0\leftarrow a_0*\frac{ex}{(1-ex)}$
5. for i=1;N do
6. $(a^i,b^i)\leftarrow(a^{i-1},b^{i-1})+(H^k)^{-1}g^k$
7. if($|(a^i,b^i)'\cdot(a^i,b^i)-(a^{i-1},b^{i-1})'\cdot(a^{i-1},b^{i-1})|<e$) then break;
8. end

算法 1 中的 H^k 和 g^k 分别为根据式(8)－式(11)计算而得的函数 $q(a,b)$的 Hessian 矩阵和梯度。算法采用 2 个停止条件，第一个是预定义的参数估计误差阈值。

2.2 参数初值选择算法

EM 算法求解隐回归模型的参数时，首先需要给定一组参数的初值。初值可以随机选择，即在参数定义域范围内随机选择一个值作为初值，但由于随机性强，参数最优值和收敛速度都难以保证。如果因变量的观察值分布呈现不同成分组合，则可以采用 k-means 聚类对每个初值进行估计。由于 k-means 算法需要预先确定好 k 的个数，且聚类的时间复杂度较高，因此我们给出了一个简单而有效的初值估计方法，如算法 2 所示。

算法 2 参数 $\beta=(\beta_0,\beta_1)'$、σ^2、a、b 的初值选择

输入：因变量 y 的观察值序列 $y_i,\cdots,y_n$

1. $\hat{\beta}_0\leftarrow\min(y_i)$
2. $\hat{\beta}_1\leftarrow\max(y_i)-\hat{\beta}_0$
3. $\hat{x}_i\leftarrow\frac{y_i-\hat{\beta}_0}{\hat{\beta}_i}$
4. $\hat{\sigma}\leftarrow\frac{\sum_{i=1}^{n}(y_i-\hat{x}_i)^2}{n-2}$
5. $\hat{a}\leftarrow\frac{0.5}{\log(\overline{\hat{x}_i})-\overline{\log(\hat{x}_i)}}$
6. $\hat{b}\leftarrow\frac{\log(\hat{x}_i)}{\hat{a}}$

[1] http://www.r-project.org

算法 2 的基本思想是观察值归一化，归一化的参数作为回归参数，归一化的结果作为隐变量的近似值，采用极大似然估计来计算隐变量的分布参数。$\hat{\beta}_0$，$\hat{\beta}_1$ 采用了归一化方法估计，而在估计参数 a 和 b 初值时采用了极大似然估计方法。由于对 Gamma 分布无法计算参数 a 的极大似然估计值，故采用拟牛顿法求解或者采用近似解。

3 试验与分析

为了衡量 EM 算法对隐回归模型参数求解的效果，我们采用 3 个模拟数据集进行参数学习。另外采用一个真实的数据集进行建模分析。本文的实验均在同一平台之下进行，采用 R 语言实现了模型。

3.1 模拟数据试验

为了衡量 EM 算法对隐回归模型参数求解的效果，我们采用模拟数据集进行参数学习，首先给定一个 beta 分布，基于这个分布生成一组随机数（每组随机数的个数为 500 个），然后采用给定的线性模型生成因变量的观察值。根据 beta 分布的不同特性，我们分别生成了 3 组模拟数据，每组数据对应的参数值如表 1 所列。

表 1 模拟试验数据集

试验编号	Beta0	Beta1	a	b	Sigma
1	0.3	1.5	0.5	1.5	0.1
学习参数	0.29	1.52	0.47	1.48	0.09
2	1.5	2.5	1.5	1.5	0.1
学习参数	1.49	2.51	1.48	1.52	0.08
3	1.5	1.8	0.4	0.5	0.1
学习参数	1.50	1.79	0.44	0.51	0.09

参数学习过程中，我们设定迭代次数为 100 次，参数迭代误差为 0.01。3 组拟合试验学习到的参数值列在表 1 中，图 2—图 4 分别给出了每组数据拟合的效果。

表 1 中的学习参数表示对应试验编号 EM 算法学习到的参数结果。从表 1 中可以看出：

(1)在有限迭代次数下，EM 算法可以学习到参数的近似结果。我们的试验迭代了 100 次，得到的参数结果与模拟数据的真实误差基本一致，其中最大的参数误差为 0.04，最小的参数误差为 0.001。这个结果说明了参数初值选择得较好，使得 EM 算法可以在较少的迭代次数下获得精度较高的参数结果。

(2)在 3 组实验中，第一组和第二组的参数结果优于第三组的参数。由于第三组模拟数据的分布为 U 型分布，因变量的观察值分布呈现两个成分的组合，而我们采用了简单的归一化初值选择，没有采用 k-means 聚类学习参数的初值，这是导致迭代在一定次数下和真实参数之间具有一定误差的潜在原因。

图 2、图 3 与图 4 分别给出了每组数据拟合的效果。每个图由 3 个子图组成，(a)子图为模拟数据的直方图；(b)子图为隐变量在 EM 算法学习到的参数下的 beta 分布密度曲线；(c)子图为因变量观察值和 EM 算法计算得到的自变量的期望值的点对图，其中横坐标为自变量的期望值，而纵坐标为因变量观察值。

从 3 组模拟数据的直方图可以看出，第一组模拟试验数据呈现偏态分布，用来模拟自变量呈指数或幂律分布下降的情况；第二组模拟试验数据呈现正态分布特征，用来模拟自变量呈正态分布的特征；第三组模拟试验数据呈现两个成分组合的特征，用来模拟自变量也呈多成分组合或 U 型分布的特征。综合图 2、图 3 与图 4 的模拟数据的结果，可以得出如下基本结论：

(1)从图 2、图 3 与图 4 的每个(b)子图可以看出，每组模拟数据得到的自变量的 beta 分布密度曲线与因变量的分布一致。对第一组模拟数据，自变量的 beta 分布呈指数分布的特征下降；对第二组模拟数据，自变量的 beta 分布呈峰值双侧缓慢下降，与因变量的观察值分布形态一致；对第三组模拟数据而言，因变量的数据呈现两个成分组合的特征，可以看成 2 个正态分布的混合，但由于采用 beta 分布作为自变量的分布，因此自变量的分布呈现 U 型。

(2)结合图 2、图 3 与图 4 的(c)子图可以看出，第二组数据下 EM 算法得出的自变量期望值和因变量的观测值相关性最高，呈直线相关；而第三组数据下自变量期望值和因变量的观测值相关性最差，尤其是在值的两端，也就是最大值和最小值附近（如图 4(c)所示），自变量期望值和因变量的观测值相关性较低。这个现象与第一组数据的结果一致，如图 2(c)所示。造成这种不相关性的原因在于 beta 分布在偏态下降和 U 性分布时都是简单的下降，没有上升的过程，而在因变量的观测值分布上都有上升过程，用 beta 分布建模这种上升过程只能依赖随机偏差，结果导致了在最大值和最小值附近相关性较低。

(3)综合观察上述结论(1)(2)，可以得出采用 beta 分布能够有效地建模和刻画各种形态下的因变量分布，采用噪声作为自变量和因变量的偏差可以实现因变量观察值中的波动。但是自变量的分布确定了自变量的取值特性，所以对于具有与 beta 分布不同形态特征的因变量观察值，也可采用正态分布、Gamma 分布以及指数分布等进行建模。

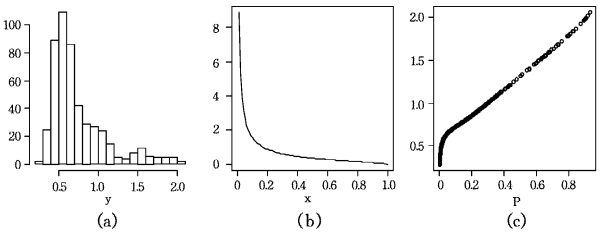

图 2 模拟数据 1 的拟合结果

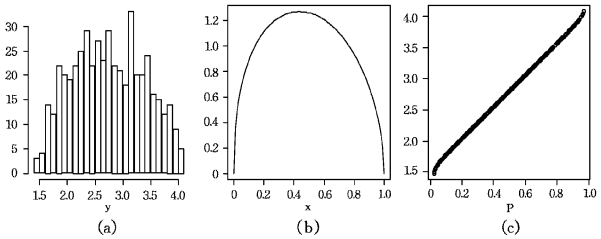

图 3 模拟数据 2 的拟合结果

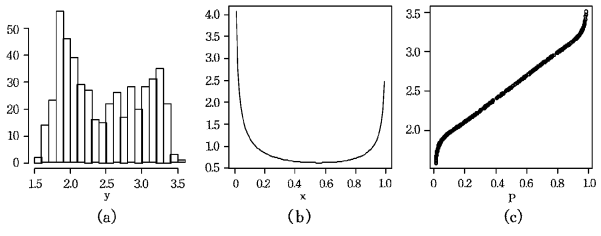

图 4 模拟数据 3 的拟合结果

3.2 真实数据试验

真实数据集来自人人网的443个用户，这些用户来自于人人网中一个用户的所有好友，这些好友基本为初中、高中以及大学同学。我们采集了这些用户的基本信息、新鲜事、日志、相册以及照片，并采集了这些信息的全部评论，形成了完整的测试数据集。

我们试图定量评价真实人类社会网络中的关系强度，但关系强度难以计算。线上的社会网络为计算关系强度提供了可能，线上社会网络中用户间的关系通过交互体现，用户交互的强弱刻画了社会关系网络中用户关系的强弱。据直观分析，两个用户间的交互强度越高，用户的关系强度越高。所以，我们采用用户的交互强度作为因变量，用交互次数实现的计算模型作为观察值[17]；关系强度作为自变量，由于关系强度难以观察，因此它为隐变量，所以构建隐回归模型。

我们采用与模拟数据相同的平台和程序计算隐回归模型的参数值，结果为：$\beta_0=0.13$，$\beta_1=1.88$；beta分布的两个参数$a=0.42$，$b=4.18$；而随机噪声的方差$\sigma^2=0.02$。

试验结果中的变量分布曲线如图5所示，其中(a)子图为通过交互次数计算得到的用户交互强度直方图；(b)子图为隐变量在EM算法学习到的参数下的beta分布密度曲线；(c)子图为交互强度和EM算法计算得到的关系强度期望值的比较图，其中黑色曲线为关系强度期望值，而蓝色曲线为交互强度。

从(a)子图上可以看出交互强度呈现指数下降的趋势，对应的EM算法得出的关系强度期望值的理论分布也呈指数下降，如子图(b)所示。比较(a)(b)两个子图可以看出，两者的趋势基本一致。我们将交互强度和关系强度期望值进行比较，如子图(c)所示。对比分析交互强度和关系强度期望值可以看出：

(1)交互强度和关系强度的趋势一致，尤其是在长尾效应上一致，但是关系强度在长尾部分的值更小，说明偶然交互带来的关系强度可以通过隐回归模型进行弱化；

(2)对于交互强度较高的用户对，对应的关系强度也较高，但在中间值部分，通过隐回归模型计算得到的关系强度期望比交互强度的区分能力更强，这表明关系强度期望值能够较好地区分不同用户间的关系紧密程度。

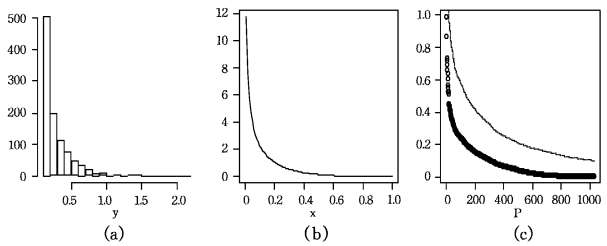

图5 真实数据的拟合结果

结束语 隐回归模型在图模型、社会网络以及生物信息学等领域都具有广泛的应用。由于无法直接得到隐回归模型中自变量的分布，只能依赖因变量的观察值分布进行假设，因此需要选择具有普适性的自变量分布。beta分布在不同的参数条件下，可以呈现出不同的分布特性，本文采用beta分布作为自变量的假设分布，给出了隐回归模型的EM算法，详细地推导了模型中的参数求解过程，给出牛顿法求解beta分布参数的算法，并提出一个初值选择算法。模拟数据和真实数据的试验结果表明，对具有不同分布特征的因变量观察值，利用EM算法都能够有效地求解隐回归模型的参数，同时也说明简单有效的初值选择能够满足实际需要。

由于因变量的观察值分布具有不同的形式，对所有的分布都采用beta分布作为自变量的假设会带来一定的偏差，如何根据因变量观察值的分布特性，选择其它更具有普适性的分布或者根据不同的因变量观察值分布自动选择特定的自变量分布还有待研究。另外，非线性隐回归模型、多变量隐回归模型的求解也是非常有意义的研究课题。

参 考 文 献

[1] Tarpey T, Petkova E. Latent regression analysis[J]. Statistical Modelling, 2010, 10(2): 133-158

[2] Moustaki I, Knott M. Generalized latent trait models[J]. Psychometrika, 2000, 65(3): 391-411

[3] Aitchison J, Shen S M. Logistic-normal distributions: Some properties and uses[J]. Biometrika, 1980, 67(2): 261-72

[4] Barto lucci F, Scaccia L. The use of mixtures for dealing with non-normal regression errors [J]. Computational Statistics & data Analysis, 2005, 48(4): 821-48

[5] Dorazio R M, Andrew Royle J. Mixture models for estimating the size of a closedpopulation when capture rates vary among individuals[J]. Biometrics, 2003, 59(2): 351-64

[6] Khan M E, Mohamed S, Marlin B M, et al. A stick-Breaking Likelihood for categorical data analysis with latent Gaussian models[C]//Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). 2012: 610-618

[7] Irncheeva I, Cantoni E, Genton M G. Generalized linear latent variable models with flexible distribution of latent variables[J]. Scandinavian Journal of Statistics, 2012, 39(4): 663-680

[8] Kelava A, Kohler M, Krzyzak A, et al. Nonparametric estimation of a latent variable model[OL]. http://www3.mathematik.tu-darmstadt.de/hp/stochastik-homepages/kohler-michael/publikationen.html, 2012

[9] Guo J, Wall M, Amemiya Y. Latent class regression on latent factors[J]. Biostatistics, 2006, 7(1): 145-63

[10] Tarpey T, Ivey C T. Allometricex tension for multivariate regression models[J]. Journal of Data Science, 2006, 4(4): 479-95

[11] Bartoletti S, Flury B D, Nel D G. Allometric extension[J]. Biometrics, 1999, 55(4): 1210-1214

[12] Tarpey T, Yun D, Petkova E. Model misspecification finite mixture or homogeneous? [J]. Statistical modeling, 2008, 8(2): 199-218

[13] Tarpey T, Petkova E. Modeling Placebo Response via Infinite Mixtures[J]. Jpn Journal of Biostatistics, 2010, 4(2): 161-179

[14] Tarpey T, Petkova E, Lu Y, et al. Optimal partitioning for linear mixed effects models: Applications to identifying placebo responders[J]. Journal of the American Statistical Association, 2010, 105(491): 968-977

[15] Tsou C M. On the exploration of linear latent effect for multivariate modeling[J]. Applied Mathematical Modelling, 2012, 36(12): 6154-6166

[16] Ma Y, Genton M G. Explicit estimating equations for semiparametric generalized linear latent variable models[J]. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2010, 72(4): 475-495

[17] 韩忠明，苑丽玲，杨伟杰，等. 加权社会网络中重要节点发现算法[J]. 计算机应用，2013，33(6)：1553-1557