# 高维加性Cox模型的估计

雷馨钰，徐嘉璐

（兰州财经大学，甘肃 兰州 730101）

**摘 要** 半参数模型能够避免完全非参数模型的“维数灾难”及参数模型的模型假定风险，加性Cox模型能处理实际问题中的无规律数据，故通过B样条方法将加性Cox模型中未知的分量函数进行样条基函数展开。将未知的分量函数选择问题变成线性组合中选择系数组的问题，后对加性Cox模型使用组Lasso进行群体惩罚，得到组Lasso估计量，从而实现针对加性Cox模型的变量选择。通过模拟可知，进行样条拟合后的估计量具有较好的性能。

**关键词** 高维数据；B样条；半参数模型

**中图分类号**：O212.1 **文献标识码**：A **文章编号**：1007-0745(2023)01-0018-03

## 1 绪论

在信息爆炸的时代，高维数据的产生便于研究者从多个维度去分析问题，但同时，传统的回归模型就不能满足高维数据的需要，故近年来，半参数回归模型的产生很好地解决了模型构造问题。Cox模型是由英国统计学家D.R.Cox于1972年提出的一种半参数回归模型[1]。该模型以生存结局和生存时间为因变量，引入基线风险函数，对实际问题中的无规律分布、删失数据等问题可以很好地处理。该模型自问世以来，在医学随访研究中得到广泛的应用，也是迄今生存分析中应用最多的多因素分析方法。

然而，在实践中，通常很少或没有先验信息表明协变量的影响呈线性形式或属于任何其他有限维参数族。因此需要通过使用一类更灵活的非参数模型，例如加性Cox模型，加性Cox模型中分量函数的引入显著增加了模型的灵活性，因此，大量学者对加性Cox模型进行了研究。

Tibshirani（1997）首次提出在Cox模型中使用Lasso进行变量选择，Fan和Li（2002a，2002b）提出在Cox模型中使用平滑剪裁绝对偏差（SCAD）惩罚进行变量选择和估计，Huang（1999）利用多项式样条研究了部分线性可加Cox模型下最大偏似然估计的性质，但是，上述作者仅仅研究了加性模型维数p固定时的情况。对于稀疏加性Cox模型，Lemler（2012）考虑了Cox模型中基线风险函数和回归系数的联合估计，但未考虑由分量函数和基线函数的线性组合引起的近似误。基于高维数据与生存分析模型所具有的特殊性，传统的变量选择方法就不再适用，这是由于传统的变量选择方法不满足变量选择应该具有的准确性、可解释性、稳定性等显著特点。因此需要对加性Cox模型在高维情况下的变量选择进行系统分析，以便高效处理高维数据下的变量选择问题。

总体上看，在高维数据中，使用变量选择方法来筛选出数据中的重要信息是未来发展的趋势。大量学者基于惩罚思想对有关模型的变量选择进行不断地改进，常见的变量选择的方法有岭回归、Lasso、SCAD、MCP[2]等。但往往有些变量选择方法的“过度压缩”会导致重要信息的损失，从而损失模型估计的精确度。故如何使模型在变量选择后仍保留更多的有用信息也是研究者们大量关注的问题。

传统Lasso方法对不同系数进行相同程度的加权，造成过度压缩绝对值较大的参数的情况，得到过于稀疏的模型，而且Lasso方法是在单个变量的基础上对模型进行特征选择，不具备处理具有组特性的数据。Yuan(2006)提出了组Lasso(Group Lasso)方法，组Lasso是Lasso的扩展，它的不同之处是对一组系数向量添加约束，因此克服了Lasso方法无法实现从组的水平进行特征选择的这一缺点。组Lasso在各个领域中都被广泛使用：

在医学方面，Ma(2007)将有监督的组Lasso方法用于基因选择和模型预测，并通过组Lasso方法选择集群，从基因簇中找到重要的基因。基于变量选择特征，Kim(2012)将组Lasso方法用于生存数据的分析中，该方法可以有效地结合临床和基因组协变量，并在实际微阵列中进行了实验。

在机器学习方面，Yeh(2014)将组Lasso多核学习

方法应用于异构特征选择，并证明了在选择紧凑特征子集方面是有效的。在金融风险投资方面，Qi 等 (2021) 利用非负稀疏组 Lasso 方法 [3]，用于成分股的选择和权重系数的估计。

针对现有文献中存在的问题，本文使用了一类正则化方法，通过对对数偏似然函数施加群组惩罚，并基于一些温和的假设条件可以同时对高维 Cox 加性模型进行结构识别，变量选择及其估计。特别地，我们将模型的结构识别和变量选择问题转化为一个对于分量函数的判别问题，通过构建正交 B 样条基可以将这些问题参数化，并通过快坐标最优下降法 lv(2017)[4] 对提出的变量选择方法进行识别。

## 2 稀疏加性 Cox 模型

一般来说，医学中生存分析的研究应用在观察时间与事件发生时间不一致的情况，它将事件发生的结果与观察时间两因素结合起来，研究生存函数与斜变量之间的关系，可以分别对完全、不完全数据进行分析，通常可用生存率、生存曲线等指标来估计生存时间。但当生存时间的分布过于复杂时，简单的计算指标不能满足现实的需要，而 Cox 比例风险模型就可以很好地解决上述问题。

Cox 模型不直接考察生存函数与斜变量之间的关系，而是用风险函数作为因变量，将参数与非参数结合，排除混杂因素影响，筛选出影响生存时间的因素。但在 Cox 模型中，当引进的斜变量对时间的响应较为敏感时，偏似然函数损失的信息较多。故在本文中对带有时间变量的 Cox 模型进行假设。

建立关于时间 $t$ 的 $P$ 维协变量 $X(t)=(X_1(t),X_2(t),\cdots X_p(t))^{\mathrm{T}}$。用 $Q$ 和 $R$ 分别表示在生存过程中时间发生的时间和截尾时间，则观察到的生存时间和截尾时间满足 $W=\min\{Q,R,\Delta=1|Q\leq R\}$。此时，建立一个 $n$ 维独立同分布的随机样本 $\{(X_i(t),W_i,\Delta_1),0\leq t\leq T\}_{i=1}^{n}$，其中 $T$ 表示为观察结束时间。

考虑一个 $n$ 维计数过程 $N^{(n)}(t)=(N_1(t),N_2(t)\cdots N_n(t)),t\geq 0$，其中任意第 $i$ 项满足 $N_i(t)=1\{W_i\leq,\Delta_i=1\}$，对于 $t\geq 0$，建立包含截止到时间 $t$ 中所有有用信息构成的 $\sigma$ 族：$\Im_t=\sigma\{N_i(s),Y_i(s),X_i(s);s\leq t,i=1,\cdots,n\}$，此时，假定对于 $\{\Im_t,t\geq 0\}$，$n$ 维随机过程是可预测的，并且存在 $\Lambda^{(n)}=(\Lambda_1,\Lambda_2,\cdots\Lambda_n)$。

由 lv(2017) 知稀疏加性 Cox 模型如下：

$$d\Lambda_i(t)=Y_i(t)\exp\left\{f^*\left(X_i(t)\right)\right\}d\Lambda_0(t) \qquad (1)$$

其中，$Y_i(t)$ 为关于 i 的主观时变风险过程，为参数部分，$f^*(X_i(t))$ 为具有 P 维斜变量的真实分量函数，为非参数部分。$\Lambda_0(t)$ 为未知的基线累积函数。并且针对稀疏加性 cox 模型要满足 $p\geq$n。但在实际中关于 $f(x)$ 的重要协变量相对较少，所以，针对式（1）中的分量函数可以表示为：

$$f^*\left(X(t)\right)=\sum_{j\in\vartheta}f_j^*\left(X_j(t)\right) \qquad (2)$$

其中，$f_j^*$ 中的元素都为单变量，并且 $\vartheta\subseteq\{1,2,...,p\}$ 是基底 $|\vartheta|=d_0$ 的子集，满足 $d_0\ll p$。

## 3 惩罚对数偏似然函数

本文针对模型，提出主要应用 B 样条 [5] 的方法对未知的分量函数进行样条基函数展开，从而进行后续估计。在样条估计中，主要利用样条基函数的线性组合来逼近未知的光滑函数，这种组合可以拟合不同形状或分布的数据，因此，为了使得 B 样条估计方法可以对更复杂的模型进行逼近求解，对于合适的基函数的选取也是我们值得关心的问题。

假定 $X_j(t)$ 在任意 t∈[0,T] 在区间 [a,b] 上取值，且 $j=1,2,\cdots p$，假定多项式空间 $S_n$ 中有 $K$ 个点，满足 $a=\xi_0<\xi_1<\cdots\xi_{K+1}=b$，则 $K$ 个点就为多项式空间 $S_n$ 中的 $K$ 个节点。用 $I_{Kq}$ 表示为区间 [a,b] 上的子集，建立 $I_{Kq}=[\xi_q,\xi_{q+1}],q=0,1,\cdots K$，其中 $K$ 满足 $K=K(n)=n^v\ 0<v<1/2$ 并使得 $\max_{1\leq q\leq K+1}|\xi_q-\xi_{q+1}|=O(n^{-v})$ 成立。

此时定义 $S_n$ 为满足以下条件的多项式样条空间：（1）$I_{Kq}$ 为 $S_n$ 的子集，且 $1\leq q\leq K$；（2）对于 $\ell\geq 2$ 与 $0\leq\ell'\leq\ell-2$，函数 $s$ 是 $\ell$ 次连续可微的。

由上述可知，在空间 $S_n$ 上，当 $1<k<m_n$，$m_n=K(n)+l$ 时存在一个 B 样条基 $\phi_k$ 使得对于任意 $f_{nj}\in S_n$ 都存在：

$$f_{nj}\left(z\right)=\sum_{k=1}^{m_n}\beta_{jk}\phi_k\left(Z\right),1\leq j\leq p. \qquad (3)$$

基于光滑性假定，基函数 $f_{nj}(z)$ 可以逼近 $S_n$，在上述近似下，每个分参数分量都可以表示为样条基函数的线性组合，则通过 B 样条可以将模型中未知的分量选择问题变成了线性组合中选择系数组的问题，便于之后的估计。

建立两组向量：$\beta_{nj}=(\beta_{j1},\cdots,\beta_{jm_n})^T\beta_{nj}=(\beta'_{n1},\cdots,\beta'_{np})^T$，对于任意 $x\in[a,b]$ 可知存在 $\Phi(x)=(\phi_1(x),...,\phi_{m_n}(x))^T$，利用（2,3）中定义的基函数去逼近未知分量函数 $f_j^*(\cdot)$，则可得部分对数似然函数：

$$C(\beta_n)=\sum_{i=1}^{n}\int_0^T \beta_{nj}^{'}\Phi\left(X_{ij}(s)\right)dN_i(s)-\int_0^T \log\left[\sum_{i=1}^{n}Y_i(s)\exp\left\{\sum_{j=1}^{p}\beta_{nj}^{'}\Phi\left(X_{ij}(s)\right)\right\}\right]d\bar{N}(s)$$

其中 $\bar{N}=\Sigma_{i=1}^{n}N_i\ell_1$，已知 Cox 模型在高维环境下选择变量的一种流行方法是最小化 $\ell_1$ 惩罚负对数偏似然准则，在本节中，对稀疏加性模型 Cox 模型使用组 lasso 施加群体惩罚，通过考虑以下惩罚目标函数，得到了用于选择系数组的组 lasso 估计量：

目标函数：

$$\Gamma_n(\beta_n,\lambda)=L(\beta_n)+\lambda\sum_{j=1}^{p}\left\|\beta_{nj}\right\|_2 \quad (4)$$

其中，$L(\beta_n)=-\frac{C(\beta_n)}{n}$，$\lambda$ 是控制稀疏性和模型偏差的参数，定义估计值为：

$$\hat{\beta}_n=\beta_n(\lambda)=\arg\min_{\beta_n}\left\{\Gamma_n(\beta_n,\lambda)\right\} \quad (5)$$

因为对于任意 $\lambda$，$\Gamma_n(\beta_n,\lambda)$ 为关于 $\beta_n$ 的凸函数，则可知其存在最小值，对于任意 $j$，定义：

$$\hat{f}_j^{*}\left(X_j(t)\right)=\hat{\beta}_{nj}^{'}\Phi\left(X_j(t)\right)$$

$$\bar{\phi}_k\left(X_j(t)\right)-\frac{1}{n}\sum_{i=1}^{n}\int_0^T\phi_k\left(X_{ij}(s)\right)dN_i(s)$$

其中基向量为 $\bar{\Phi}\left(X_j(t)\right)=\left(\bar{\phi}_1\left(X_j(t)\right),\ldots,\bar{\phi}_{m_n}\left(X_j(t)\right)\right)$，则可以得：

$$\hat{f}_j\left(X_j(t)\right)=\hat{\beta}_{nj}^{'}\bar{\Phi}\left(X_j(t)\right)$$

由上文中对数似然函数 $C(\beta_n)$，可知当且仅当 $(\hat{\beta}_n,\hat{f}_1,\ldots,\hat{f}_p)$ 存在极大似然估计时，$(\hat{\beta}_n,\hat{f}_1^{*},\ldots,\hat{f}_p^{*})$ 的极大似然估计值才可以求出。

## 4 模拟研究

本节对整合后的加性 Cox 模型进行蒙特卡洛模拟分析，因高维数据的特殊性，分别考虑当 P=10 和 P=50 时的拟合情况。

$$d\Lambda_i=Y_i(t)\exp\left\{\sum_{j=1}^{p}\left[f_j\left(X_{ij}(t)\right)\right]\right\}d\Lambda_0(t) \quad (6)$$

其中，假定在上式中前三个变量当 j=1,2,3 时定义为 $f_1(x)=\sin x+2$，$f_2(x)=\sin(2x)^2+12$，$f_3(x)=10(x-2)^2$，当 j=4,…$p$ 定义为 $f_j(x)=0$，且协变量和残差都满足均匀分布。

情形 1：当 P=10 时，分别取 $n$=100、200、500。可得表 1：

表 1

| | $mse_1$ | $mse_2$ | $mse_3$ |
|---|---|---|---|
| n=100 | 0.0042 | 0.025 | 0.115 |
| n=200 | 0.0041 | 0.024 | 0.104 |
| n=500 | 0.0039 | 0.025 | 0.102 |

情形 2：当 P=50 时，分别取 $n$=100、200、500。可得表 2：

表 2

| | $mse_1$ | $mse_2$ | $mse_3$ |
|---|---|---|---|
| n=100 | 0.0039 | 0.024 | 0.105 |
| n=200 | 0.0041 | 0.025 | 0.104 |
| n=500 | 0.0040 | 0.023 | 0.095 |

由情形 1 和情形 2 可知，随着维数增加，误差会增大，但数值普遍较小，可知估计量有良好的性能。

## 5 总结展望

从大量数据中选择出重要变量对于模拟研究及探寻事物变化的本质有着重要的意义，因此变量选择方法在高维数据中就显得尤为重要。在本文中，考虑加性 Cox 模型在高维数据中的情况，通过 B 样条曲线拟合模型，将函数中的未知函数用样条基函数展开，结合具有 Oracle 性质的组 Lasso 惩罚方法，建立了更完善的加性 Cox 模型的变量选择过程。后续可考虑在更高维度下的变量选择问题。

## 参考文献：

[1] Cox DR.Regression models and life tables(with Discussion)[J].J.R.Statist.Soc.B,1972(34):187-220.

[2] 白玥，田茂再．几种高维变量选择方法的比较及应用 [J]. 统计与决策，2017(22):11-16.

[3] Qi Kai,Yang Hu.Nonnegative Sparse Group Lasso with an Application in Finanxial Index Tracking[J].Applied Probability and Statistics,2021,37(03):221-240.

[4] Shaogao Lv.Estimating high-dimensional additive Cox model with time-dependent covariate processes[J]. Scandinavian Journal of Statistics,2018(04):900-922.

[5] 马晓跃，武新乾．非参数回归模型基于残差的样条估计 [J]. 河南科技大学学报（自然科学版），2021,42(04):91-96.