分类号 O212/27　　密级 公开
U D C　　编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

论文题目 部分线性可加 Cox 模型的变量选择研究

研究生姓名：雷馨钰
指导教师姓名、职称：郭精军　教授
学科、专业名称：统计学　数理统计学
研究方向：复杂数据分析

提交日期：2023 年 5 月 30 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：雷馨钰 签字日期：2023.5.30

导师签名：李树茂 签字日期：2023.5.30

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：雷馨钰 签字日期：2023.5.30

导师签名：李树茂 签字日期：2023.5.30

Research on Variable Selection of Partially Linear Additive Cox Model

Candidate: Xinyu Lei

Supervisor: Jingjun Guo

摘 要

Cox 比例风险模型在生存分析中扮演着重要的角色，它能够利用风险函数研究变量与生存函数间的关系，初步处理存在删失情况的生存数据，但实际生活中的数据通常不满足 Cox 比例风险假定。针对这类数据，引入部分线性可加 Cox 模型，实现对时依协变量的变量选择研究。本文主要研究内容分为以下三部分：

（1）通过 B-样条曲线拟合部分线性可加 Cox 模型中的非参数部分，将模型中未知分量函数选择问题转变为处理线性组合中选择系数组的问题，实现对部分线性可加 Cox 模型的样条拟合。

（2）针对部分线性可加 Cox 模型中的删失生存数据引入双层变量选择方法，其中协变量自然分组。与组变量选择相比，实现在选定的组内同时进行组选择和个体变量选择，提高模型估计精度。

（3）通过模拟分析对比组变量选择方法和双层变量选择方法在五类指标下的性能，验证了双层变量选择方法在部分线性可加 Cox 模型中的有效性。分别引入两类不同的癌症数据集，结果表明双层变量选择方法筛选出的变量与存活时间相关度最高，即对攻克癌症疾病有现实意义。

研究表明双层变量选择方法在部分加性 Cox 模型中的预测误差优于组变量选择方法，引入的两个数据集都体现了双层变量选择方法的有效性。

关键词：比例风险模型 B-样条 双层变量选择方法 生存分析

Abstract

Cox proportional risk model plays an important role in survival analysis. It can use risk functions to study the relationship between variables and survival functions, and initially process survival data with censoring. However, data in real life often do not meet the Cox proportional risk assumption. For this type of data, a partially linear additive Cox model is introduced to achieve variable selection research for time-dependent covariates. The main research content of this article is divided into the following three parts:

(1)By fitting the nonparametric part of a partially linear additive Cox model with a B-spline curve, the problem of selecting unknown component functions in the model is transformed into the problem of dealing with the selection of system arrays in linear combinations, and spline fitting of partially linear additive Cox models is realized.

(2)Bi-level variable selection method is introduced for censored survival data in partially linear additive Cox models, where covariates are naturally grouped. Compared with group variable selection, group selection and individual variable selection can be performed simultaneously within the selected group, improving model estimation accuracy.

(3)By comparing the performance of the group variable selection method and the two-level variable selection method under five indicators through simulation analysis, the effectiveness of the two-level variable selection method in a partially linear additive Cox model was verified. Two different types of cancer data sets were introduced, and the results showed that the variables screened by the two-level variable selection method had the highest correlation with survival time, which was of practical significance in tackling cancer diseases.

Study has shown that the prediction error of the two-level variable selection method in the partially additive Cox model is better than that of the group variable selection method. The two data sets introduced reflect the effectiveness of the two-level variable selection method.

Keywords: Proportional risk model; B-Spline; Bi-level variable Selection method; Survival analysis

目 录

1 绪论

1.1 研究背景

随着科技的发展，人工智能等科学研究使数据成指数增长，大数据科学在基因组学、金融学、经济学和社会科学等领域中广泛应用，因此人们获取数据的方式越来越灵活、从数据中获取有用信息也越来越多。但同时，伴随着数据数量增多、维数增加等情况也成为数据处理中的难题。数据处理中预测变量数 p 远大于样本量数 n 的这类数据就称为高维数据。例如：建立一个根据年龄，性别和身体指数（BMI）预测血压的模型，假定试验者为 200 人（$n=200$），预测变量为年龄、性别和身体指数，则预测变量为 3（$p=3$），此时 $p>n$，称为低维数据；而建立根据基因来预测高血压的模型时，假定试验者为 200 人（$n=200$）、基因组数为 50 万组（$p=500000$），显然 $p \gg n$，称为高维数据。在处理这样的高维数据问题时，数据集之间的多重共线性、过拟合等问题会使传统数据分析方法失效，从而损失有用信息。如何在庞大的数据中选择出有用的信息是众多统计学家研究的方向，在研究复杂且数量较大的数据时，人们往往倾向于用更少的变量了解到更有用的信息，这便涉及到了变量选择。

针对统计模拟研究，最便捷的就是通过模型来表达数据资料。在发展初期，为了避免出现模型偏差，学者总是尽可能的将大量数据都引入模型中，但这样做的后果就是模型中包含了许多无用或作用较小的变量，导致参数估计出现偏差、分析结果不具统计意义。自 20 世纪 60 年代以来，众多的统计学家开始研究变量选择。好的变量选择方法应该满足计算简单、保证精度及参数估计的一致性等优点。当一个因变量由多个自变量决定时，研究每个自变量对因变量的影响情况是

预测结果的前提，当自变量数量不清楚或某些自变量对模型作用较小时，盲目的将所有自变量应用到模型中会导致出现偏差。则在实际应用中就需要将有用且影响因子较大的变量筛选出来，这个筛选的过程就称为变量选择。首先被提出的变量选择的方法是最优子集筛选法，它是利用 AIC、BIC 等准则在所有子集中选择一个最优子集的方法。这种最优子集筛选法在低维数据中可以得到较好的结果，但随着高维数据的出现，由于上述方法存在 *NP* 困难，会导致变量选择方法在结果中缺少准确性。自然地，对于如何降低数据的维度成为了新的研究目标。针对高维数据，Lasso 方法被提出。它将模型中系数较小的变量估计为零，大大降低了模型的复杂度，提升估计的精确度。

在生存分析中，研究因变量对生存函数的影响是研究者的重点，但在实际中无法准确刻画生存时间与生存函数之间的关系。盲目的选用参数模型会出现参数假定错误和估计误差偏大等问题；而使用非参数模型虽然可以体现风险函数对生存时间的影响，但无法准确建立两者之间的模型关系，缺乏理论支撑。此时，半参数模型一方面可以规避掉模型假设错误的风险，另一方面也可以准确刻画生存时间对风险函数的影响。Cox 比例风险模型[1]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。它是通过刻画风险函数与协变量之间的关系来解决生存时间分布未知的问题。模型以生存结局和生存时间为因变量，通过风险函数来处理实际问题中的无规律分布、数据删失等。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。

1.2 研究目的及意义

1.2.1 研究目的

生存分析中 Cox 比例风险模型只能在协变量之间不存在严重的多重共线性时得到生存时间和风险函数间的模型关系，但当出现高维度、小样本、协变量强相关这类生存数据时，传统的 Cox 模型就无法准确刻画数据的特征。于是，改进 Cox 比例风险模型成为生存分析中需要研究的问题。

本文主要采用部分线性可加 Cox 模型代替 Cox 比例风险模型，并使用双层变量选择方法来研究生存分析中的变量选择问题，改善了 Cox 模型的局限性。将组变量选择方法拓展到双层变量选择中，实现对模型中可能出现的线性与非线性关系的变量选择研究。基于生存分析，对医学数据中处理删失数据带来更可靠的理论依据，使得医学研究更具权威性。

1.2.2 研究意义

变量选择研究是统计学中的重要课题，在生存分析中，常常会收集多维度的变量来刻画问题事件，大量的无关变量会导致模型计算复杂、参数估计误差增大。人们往往希望用较少的变量得到更准确的预测结果，而如何在众多的响应变量中选择信息度较高的变量是进行准确预测的前提。对数据运用合适的变量选择方法可以提高模型的预测精度。研究生存分析中部分线性可加 Cox 模型中的变量选择方法，能够降低实际计算复杂度及增加模型的预测准确度。

（一）理论意义

基于变量选择的重要性，已有众多学者将岭回归、Lasso、Scad 等变量选择方法应用到 Cox 比例风险模型当中，分别讨论了不同情形下变量选择的效果，

能够一定程度上完善生存分析中的变量选择问题。部分学者针对部分线性可加 Cox 模型，利用具有群组特征的 Group Lasso、Group Scad、Group Mcp 等组变量选择方法进行组间稀疏性的研究，但针对部分线性可加 Cox 模型下的双层变量选择而言，其有关的变量选择研究就较为单一。本文在生存分析的背景下，采用样条拟合模型中的未知分量函数部分，基于此得到拟合后的部分线性可加 Cox 模型。将双层变量选择方法应用到部分线性可加 Cox 模型当中，研究双层变量选择的组内与组间的双层效应以及部分线性可加 Cox 模型中分量函数的拟合等问题，在一定程度上使生存分析中可研究数据范围扩大，为之后部分线性可加 Cox 模型的变量选择研究提供了新思路。

（二）现实意义

本文主要研究了生存分析中部分线性可加Cox模型的变量选择问题，多年来，人们因为癌症导致死亡的病例数不胜数，研究人体体内基因与癌症成因的关系是攻克癌症疾病的关键。而利用双层变量选择方法可以将组内不相关区间剔除，得到与癌症死亡率相关度最高的变量，为后续攻克癌症疾病提供了一定的基础，具有现实意义。

1.3 研究现状

1.3.1 变量选择的研究现状

随着数据维数和数量的增多，变量选择在统计领域被广泛研究。在低维情况时，普通最小二乘法是对模型进行参数估计的一种普遍方法。但在实际问题中，不可避免的会因为变量之间相互影响而导致出现多重共线性，故通常在变量选择中，不使用最小二乘法对数据进行筛选。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。此处将有关变量选择的正则化方法分为单变量选择方法、组变量选择方法以及双层变量选择方法。

单变量选择方法能对单个变量进行选择，主要有岭回归、Lasso、Scad、Mcp 等。Hoerl 在 1962 年首先提出了岭回归[2]（Ridge Regression），后来 Hoerl 和 Kennard[3]于 1970 年给予了详细讨论。它是带有 L_2 惩罚项的最小二乘法，是一种有偏估计，以损失部分信息为代价来获取回归系数。岭回归能将系数压缩到尽可能的小，但不能将系数直接压缩到零，这会导致模型精度降低，也无法在本质上解决多重共线性的问题，存在一定的主观性。由于岭回归方法的局限性，Tibshirani 提出了 Lasso[4]方法，它是一个连续的过程，综合了岭回归和逐步回归两者的优点。本质上是在最小二乘法的基础上施加 L_1 惩罚，该方法可以在对模型进行变量选择的同时完成变量系数的估计。但因 Lasso 方法与岭回归都是有偏估计，会损失部分有用信息，Fan 和 lin 提出了 Scad[5]方法，它是基于 Lasso 提出的一种非凹惩罚方法，用 Scad 惩罚代替 L_1 惩罚，在估计具有 Oracle 性质，能将与模型不相关的变量压缩到零，实现无偏估计。Zhang 提出 Mcp[6]方法，它也是基于 Lasso 惩罚的一种非凸惩罚，用 Mcp 惩罚代替 L_1 惩罚，在单变量选择中应用广泛。但在研究组间变量的相关性时，传统的单变量选择方法就不再适用。

组变量选择方法是在单变量选择方法在组特性上的延伸，可以解决变量中成组的选择问题，单变量选择方法对具有组特性的数据没有处理能力。常见的组变量选择方法有组 Lasso（GL）、组 Scad（GS）、组 Mcp（GM）等，它们都是在单变量选择的基础上对一组系数向量添加约束，从而克服单变量选择方法无法从组的水平进行特征选择的这一缺点。许多学者已经在各种统计建模问题中考虑了组变量选择问题。Yuan 和 Lin[7]研究了组 Lasso 以及相关的群组选择方法及算法。

Zhao 等[8]研究群组选择中的一般复合绝对惩罚，对群组惩罚方法有了理论支撑。Breheny 和 Huang[9]提出了广义线性模型中双水平选择中的一般框架，并规范了局部坐标下降算法。但在选择重要变量时，组变量选择方法[10]会将整组的变量同时选入模型当中，这就可能导致组区间内相关程度不高的一些变量选入模型，从而影响模型的精度，而双层变量选择方法就可以解决上述问题。

双层变量选择方法在保持协变量稀疏性的同时保留群组结构，从而实现在个体和群组层面同时选择重要变量。双层变量选择方法一般有两种类型：一类是基于复合惩罚函数，将组内惩罚与组间惩罚结合，在组内和组间都使用单变量选择方法，称为复合惩罚。基于此构建的双层变量选择方法有 Group Bridge 等。Huang 提出的 Group Bridge[11]方法弥补了组变量选择方法的缺陷，通过将 L_1 凹惩罚应用于组结构中，实现组内和组间的稀疏性。另一种类型是基于可加惩罚函数，利用单变量惩罚和组变量惩罚间的线性组合来实现组内和组间的变量选择，被称为可加惩罚，例如稀疏组 Lasso（SGL）、自适应稀疏组 Lasso（ASGL）等。Friedman[12]提出的稀疏组 Lasso 是在组 Lasso 上添加了 L_1 范数惩罚，并推导了其在线性模型中的渐进性质。Matsui[13]研究了在逻辑回归模型中稀疏组 Lasso 的参数估计，完善了双层变量选择方法的理论基础。而自适应稀疏组 Lasso[14]可以被视为稀疏组 Lasso 的改进，添加数据相关权重以提高选择性能，通过将自适应 Lasso 和自适应组 Lasso 结合以实现双层变量选择。

1.3.2 Cox 模型的研究现状

Cox 比例风险模型是统计学家 Cox 在 1972 年提出的一种半参数生存分析模型，它广泛应用于截尾生存数据，在许多领域都有重要应用。Tibshirani(1997)首

次提出在Cox模型中使用Lasso进行变量选择；Fan[15]提出在Cox模型中使用Scad惩罚进行变量选择和估计；Zhang[16]研究了 Cox 模型在自适应 Lasso 下的变量选择和估计，并证明了方法的 Oracle 性质。

对于部分线性模型的研究较为广泛，Liang[17]中对具有测量误差的部分线性模型进行变量选择。Zhang[18]规范了部分线性模型的自动模型选择步骤，但使用 Lasso 方法对模型进行变量选择会导致模型出现估计偏差。故 Xie[19]利用 Scad 惩罚对线性模型进行变量选择，证明了 Oracle 性质。基于此，Zhao[20]和 Kai[21]将变量选择方法拓展到变系数部分线性模型当中，并推导证明了误差边界。Xia[22]在误差不变时进行了部分时变系数模型的变量选择。Yang[23]在变系数模型中研究其变量选择的分位数回归。Hu 和 Lian[24]研究了具有发散维度的部分线性比例风险模型中的变量选择。而上述学者仅在部分线性模型中进行了理论研究，针对部分线性可加模型的研究较少，综上，生存分析中部分线性可加 Cox 模型的变量选择问题值得去研究。

可加 Cox 模型是对 Cox 模型的可加性拓展。可加 Cox 模型显著增加了 Cox 模型的灵活性，它通过转换协变量中的分量函数来分析预测因变量对响应变量的影响，并且克服了维数诅咒。Meier 等[25]考虑研究在高维中可加模型的特征筛选。Xue[26]使用惩罚多项式样条法实现可加模型中的一致变量选择，即当 p 固定时，能在可加模型中能同时进行模型估计和变量选择。Fan 等[27]研究了在稀疏超高维度下可加模型的非参数独立筛选。Lemler[28]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差。Lv 等[29]研究了基于组 Lasso 的可加 Cox 模型的变量选择，研究未知分量函数的惩罚偏似然方法。Zhang 等[30]研究了基于 G-Scad 的可加 Cox 模型的变量选

择，并扩大了惩罚变量选择适用于删失数据的半参数模型的适用范围。Lin 等[31]提出了可加 Cox 模型的全局偏似然方法，并证明估计量的一致性和渐近正态性。Wu 等[32]考虑部分线性可加 Cox 模型中使用 Bernstein 多项式近似非线性协变量及未知风险函数，完成高维区间截尾数据的变量选择和参数估计。Huang[33]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。Lu[34]使用单调 B-样条逼近部分线性加性 Cox 模型中的未知函数，证明非参数分量的样条估计在光滑条件下达到了最优收敛速度，并且回归参数的估计是渐近正态的和有效的。

1.3.3 文献述评

通过梳理已有的文献发现，学者们对变量选择在实际中的应用进行了大量的研究，对本文的研究提供了借鉴与参考，但深入分析，仍有以下提升空间：

（1）目前对生存分析中变量选择问题的讨论集中在 Cox 模型中。但随时间变化的协变量无法适应 Cox 模型，这就需要对 Cox 模型进行可加性改进。因此，采用部分线性可加 Cox 模型代替 Cox 模型，进一步提升模型的精度。

（2）在生存分析中采用的变量选择方法较为单一，大多只采用单变量选择方法或组变量选择方法来对模型进行变量筛选，很少有学者考虑同时组内和组间的变量选择问题。因此，采用双层变量选择方法对部分线性可加 Cox 模型进行变量选择研究是十分有必要的。

1.4 研究内容及框架

本文研究部分线性可加 Cox 模型的变量选择问题，在该模型基础上主要研究以下内容：

（1）考虑部分线性可加 Cox 模型中非参数部分的 B-样条拟合。利用样条展开法逼近部分线性可加 Cox 模型中未知分量函数部分，使其转化为便于后续估计的相关系数组。

（2）对样条拟合后的模型进行变量选择。为避免过度压缩，使用双层变量选择方法对样条拟合后的部分线性可加 Cox 模型施加群组惩罚，对目标函数中的残差平方和进行范数惩罚得到对数偏似然函数，后采用自适应组 Lasso 的惩罚方法进行变量选择研究。

（3）通过模拟与实证分析来证明方法的可靠性。对比组变量选择方法和双层变量选择方法在五类指标下的性能，基于 Monte Carlo 模拟来说明所提方法的有限样本表现，并通过实证分析结果来解释说明所提方法的有效性。

本文结构如下：

第一章，绪论。介绍 Cox 模型以及变量选择方法的研究现状，分别介绍单变量选择方法、组变量选择方法、双层变量选择方法的研究现状，阐述变量选择方法与生存分析模型结合的意义。

第二章，预备知识。介绍生存分析中的基本概念以及生存函数、风险函数等常用的基本函数；介绍样条回归的原理、常用的样条基函数相关理论知识以及在线性模型中的变量选择方法。

第三章，模型设置。对部分线性可加 Cox 模型中的基函数进行样条拟合，

推导了双层变量选择下的部分线性可加 Cox 模型的参数估计。

第四章，数值模拟与实例分析。对组变量选择方法与双层变量选择方法进行模拟研究，证明双层变量选择方发具有组内组间具有有效性。研究真实数据经变量选择后的效果，证明方法有实际意义。

第五章，研究总结和展望。总结本文中的部分线性可加 Cox 模型的变量选择方法，明确双层变量选择在实际中的意义。

1.5 本文创新点

本文主要研究了在生存分析中，利用部分线性可加 Cox 模型来研究协变量与生存时间的关系，通过结合变量选择方法使得部分线性可加 Cox 模型得到的结果更准确，主要创新点为：

（1）在研究内容方面，大量文献对 Cox 模型进行了理论研究，但对部分线性可加 Cox 模型的研究较少，本文将生存分析中对 Cox 模型的研究拓展到对部分线性可加 Cox 模型的研究中，拓展后的模型能够解决协变量随时间变化时产生的生存分析问题。同时考虑了部分线性可加 Cox 模型的变量选择过程，能够将研究方法应用到实际中，为模型的后续研究提供了思路。

（2）在研究方法方面，通过对模型中的非参数函数部分进行样条拟合，将复杂问题转换为选择系数组的问题，后引入双层变量选择方法对部分线性可加 Cox 模型进行组内和组间的变量选择研究，研究了模型中线性与非线性同时存在的情况下的变量选择问题，将问题一般化，更具有实际意义，并通过模拟与实证分析证明方法的有效性。

2 预备知识

2.1 生存分析

在医学、生物学、保险精算学、经济学以及人口统计等领域，往往会存在对某给定事件发生的时间进行估计和预测的问题。比如：疾病的发生时间、种群的灭亡时间、被保险人的索赔时间、经济危机发生的时间等，这类研究事件发生时间规律性的问题就是生存分析问题。生存分析的核心问题就是确定生存分布的模型以及在这些模型的基础上进行统计推断。

2.1.1 生存分析的基本概念

生存时间分布通常不满足正态分布。因观测对象进入或退出实验时间的差别，生存分析中存在删失和截尾两种类型的数据。产生删失数据的原因是由于在实际生存分析问题中，只有部分个体存活时间能够被准确记录，而剩余个体存活时间只知道其发生在某些特定时间之后，这类数据被称为删失数据。比如在利用动物研究某项致癌物质对存活时间的影响时，开始实验前动物的数目是固定的，出于观察时间或费用等因素的考虑，研究者不会等到所有动物都死亡后再结束实验。在上述实验中，删失发生于研究结束时仍然存活的个体中。

截尾数据是指在实际研究中，研究者并未发现因观测等问题自动剔除的样本。例如某项对退休中心居民的生存研究，研究记录了退休居民的死亡年龄和新个体加入中心的年龄。在本项实验中，由于未加入退休中心就提前死亡的人员不会被考虑到实验中，这类提前死亡人员产生的数据集就是截尾数据。

研究协变量是由于其具有影响响应变量的功能，建立协变量和响应变量之间的函数关系是生存分析的基础。在常见的线性回归中，协变量一般不随时间改变，

但在生存分析中，因其特殊性，有些情况下协变量会随时间改变，时间依协变量就用来刻画上述情况。时间依协变量指随时间变化而变化的解释变量。一般分为四种情况：

（1）自定义的时依协变量，一般用于比较两组不含时间变量的数据集的生存情况，此时将两个变量分别与时间相乘，则可得到两组可比较的时依协变量；

（2）内部时依协变量，它是随着时间变化，自变量本身发生变化的时依协变量。比如对有些受试者在随访过程中自身的高血压状态发生变化，自身产生了随时间变化的协变量，被称为内部时依协变量；

（3）外部的时依协变量，随着时间的变化，模型中自变量本身取值并未发生改变，但其效应却在发生变化。比如工作状态中的裁员情况，并非主动辞职，而受到外部的干预导致工作状态的结束，也称作辅助时依协变量；

（4）同时是内部和外部的时依协变量，内部和外部都随时间变化的协变量，这种情况较为少见。典型的例子为器官移植，需要同时满足存在供体以及匹配成功。匹配成功与否产生“未移植”和“移植”两种情况，这是内部随时间变化的情况，供体存在与否也会产生“未移植”和“移植”两种情况，这是外部随时间变化的情况，两者同时受时间的变化而变化的变量称为内部和外部的时依协变量。

2.1.2 生存分析的基本函数

生存分析[35]是将事件发生的结果与随访时间两个因素结合，对完全或不完全数据进行分析的方法。在生存分析中，一般不直接研究协变量与生存函数的关系，而是借助风险函数来研究。

2.1.2.1 生存函数

生存概率是指单位时间开始时存活的个体到该时段结束时仍然存活的概率，假定按年计数，则年生存概率 P 可用下式表达：

$$\text{年生存概率P}=\frac{\text{同年内生存满一年人数}}{\text{某年年初人数}}$$

生存函数[36]也被称为累积生存概率，它是指生存时间超过某个时间点的概率，假定协变量为 X，生存函数为 $S(\mathrm{t},\mathrm{X})$，则生存时间可以表示为：

$$S(\mathrm{t},\mathrm{X})=\frac{\text{生存时间>t时刻的观察单位数}}{\text{总观察数}}$$

观察生存概率与生存函数可知，生存概率研究的是某段时间的生存情况，而生存函数是指某个较长时间段内的生存情况，即生存函数是生存概率的累积结果。如研究五年的生存函数就表示为每一年生存概率的乘积：

$$S(\mathrm{t},\mathrm{X})=\mathrm{P}_1\times\mathrm{P}_2\times\mathrm{P}_3\times\mathrm{P}_4\times\mathrm{P}_5$$

2.1.2.2 风险函数

假定 X 是连续的随机变量，风险函数 $h(t,\mathrm{X})$ 可以用密度函数 $f(t,X)$ 与生存函数 $S\left(t,X\right)$ 来表示，它是描述观察个体在某时刻存活的条件下，在下一个的单位时间内死亡的（条件）概率，通常用 $h(t,\mathrm{X})$ 表示：

$$h(t,\mathrm{X})=\lim_{dt\to 0}\frac{P\left(\mathrm{t}\le T\le t+dt\right)}{dt\cdot S\left(t,X\right)}=\frac{f(t,X)}{S\left(t,X\right)}$$

2.1.2.3 Cox 比例风险模型

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，研究生存函数与协变量之间的关系。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的解决上述问题。

在大多数应用中，Cox 比例风险模型能够处理生存分析中的问题，但 Cox 模型不直接考察生存函数与协变量之间的关系，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。模型的基本形式为：

$$\mathrm{h}(t,X)=h_0(t)\exp\left(\beta_1X_1+\beta_2X_2+\ldots\beta_mX_m\right)$$

其中，$\beta_1,\beta_2,\ldots\beta_{\mathrm{m}}$ 为自变量 X 的偏回归系数，$\mathrm{h(t,X)}$ 为生存函数中的瞬时死亡率，即风险函数。Cox 比例风险模型是半参数模型，由参数和非参数两部分组成。其中 $\beta_1X_1+\beta_2X_2+\ldots\beta_mX_m$ 是模型中的参数部分；$h_0(t)$ 为基础风险率，分布未知，也称作未知基准风险函数，它与风险函数 $\mathrm{h(t,X)}$ 成比例，是模型中的非参数部分。

生存分析中的 Cox 比例风险模型满足等比例风险假定：在研究期间，任意变量对生存的影响在任何时间都是相同的，不随时间的变化而变化。这就导致当出现与时间相关的协变量时，Cox 比例风险模型就不再适用，此时部分线性可加 Cox 模型就能在上述假定不存在的情况分析变量与生存时间的关系。

2.2 样条回归

2.2.1 样条回归模型

在正交序列中，需要假设向量间为正交。而在样条回归中不用假设向量间正

交，故在样条回归中可以获得更多基函数选择的机会。

样条回归[37]与虚拟变量、时间计数、干预分析、中断时间序列、逐步线性回归等相关。例如，假设有一个随时间变化的连续变量Y，并且其随时间变化的轨迹是因为某个事件或政策改变而发生的变化，比如Y是一个社区每年领取福利的人数，很多年来Y一直上升，假设某项事件发生使得社区福利下降。若这种下降是突然的，就可以使用中断的时间序列模型，因为时间序列模型可以反映出截距的变化；但若领取福利的人数是缓慢降低而不是突然减少的，那么使用样条回归就更合适，因为样条回归可以反映两条回归线在连接点处平滑的斜率变化，而避免回归线中出现断裂。

假设有n组数据$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$, 其中$x_i\in[a,b]$。在很多情况下，有关$(x_i,y_i)$的分布未知，原有的参数模型在此处不适用，则假设$(x_i,y_i)$满足以下关系：

$$y_i=f(x_i)+\varepsilon_i,i=1,2,\cdots,n$$

式中，$f(x_i)$是关于x_i的未知函数，ε_i是满足独立同分布的正态分布$N(0,\sigma^2)$。并且满足$E(\mathrm{y}_i)=f(x_i)$。样条回归步骤见下：

对于未知的函数$f(x_i)$，采用样条基函数来估计未知函数可以避免假定错误等风险，以线性样条基函数为例来逼近未知函数。对于任意$x\in[a,b]$, 关于x的线性样条基函数定义为：

$$1,x,(x-\kappa_1)_+,(x-\kappa_2)_+,\ldots(x-\kappa_k)_+$$

其中，$\kappa_j\in[a,b]$称为结。则未知函数$f(x_i)$可表示为：

$$f(x_i)\approx\beta_0+\beta_1x_i+\sum_{k=1}^{K}b_k(x_i-\kappa_k)_+$$

在上式中，一般认为在差别足够小的情况下，可以取到等号。

在 Ruppret[38]中定义样条回归模型为：

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{K} b_k \left(x_i - \kappa_k\right)_+ + \varepsilon_i, i = 1,2,\ldots,n$$

其中，观测到的因变量为 $y = \left(y_1, y_2, \cdots y_n\right)^T$，此时有矩阵：

$$X = \begin{pmatrix} 1 & x_1 & \left(x_1 - \kappa_1\right)_+ & \left(x_1 - \kappa_2\right)_+ & \ldots & \left(x_1 - \kappa_K\right)_+ \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \left(x_n - \kappa_1\right)_+ & \left(x_n - \kappa_2\right)_+ & \ldots & \left(x_n - \kappa_K\right)_+ \end{pmatrix}$$

这里的矩阵形式可以看出与多元线性回归类似，则参数 $\left(\beta_0, \beta_1, b_1, b_2, \ldots b_K\right)$ 的估计值为：

$$\hat{\beta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_K\right)^T = \left(X^T X\right)^{-1} X^T \mathrm{y}$$

由上可知，$f\left(x_i\right)$ 的估计值为：

$$\hat{f}\left(x_i\right) = \hat{\beta}_0 + \hat{\beta}_1 x_i + \sum_{k=1}^{K} \hat{b}_K \left(x_i - \kappa_k\right)_+$$

利用线性样条基函数估计未知函数的方法较为简便，可以将复杂的模型转化为简单的基函数线性组合，是解决高维问题合适的方式。

2.2.2 样条回归模型的节点

样条回归是加了约束的分段回归，这个约束就指的要求在节点处连续，即在节点处满足连续性。节点可以被认为是分割点，有关节点的选取也是值得关心的问题。在数据量分布较为清晰时，可以根据数据的散点图来推断，即根据点的疏密程度人为进行选择。基本原则为观察散点图若在区间内分布较为均匀，则可以在区间中取等距的点作为节点；若某段的分布较为密集，则可在密集段多加入几个节点。一般情况下，节点的选取可以通过 AIC、BIC 准则直接得出。

2.2.3 常用的样条基函数

随着拟合模型复杂度的增加，简单的线性样条基函数不能满足模型的需要，一些常用的样条基函数有：

（1）二次样条基函数：表达式为 $1,x,\mathrm{x}^2,(x-\kappa_1)^2,(x-\kappa_2)^2,\ldots(x-\kappa_k)^2$，相较线性样条基函数，其在节点处光滑可导。

（2）p 阶截断样条基函数：表达式为 $1,x,\ldots\mathrm{x}^p,(x-\kappa_1)_+^p,(x-\kappa_2)_+^p,\ldots(x-\kappa_k)_+^p$，当 $p=1$时，截断样条基函数就位线性样条基函数；主有在 $p\geq 2$，截断样条基函数才在 p 阶处有定义，且在节点处是可导的。

（3）B-样条基函数：与上述两种样条基函数不同，B-样条基函数需要借助递推公式来定义。首先对 B-样条基函数的节点做出以下假定。

假设 B 是 m+1 个非递减数列的集合，满足 $B_0\leq B_1\leq B_2\leq\ldots\leq B_{\mathrm{m}}$。其中 B_i 就称作 B-样条基函数的节点，而集合 B 就称为节点向量，其中的第 i 个节点区间表示为$\left[B_i,B_{i+1}\right)$。节点的分布情况对 B-样条基函数的拟合效果有一定的影响，一般的，如果一个节点 B_i 出现 K 次，表现为 $B_i=B_{i+1}=\ldots=B_{\mathrm{i+k-1}}(k>1)$, 那么就称 B_i 是一个重复度为 k 的多重节点，记为 $B_i\left(k\right)$，反之，若只出现一次则可知 B_i 为简单节点，若节点间的距离相等，节点向量称为均匀的。定义 0 阶 B-样条基函数为：

$$B_{i,p}(x)=\mathrm{I}\left(\kappa_j\leq x\leq\kappa_{j+1}\right)$$

式中，$\mathrm{I}(\cdot)$ 为示性函数。

假定基函数的次数为 p，p 阶 B-样条基函数由如下递推公式定义：

$$B_{i,p}(x)=\frac{x-\kappa_1}{\kappa_{i+p-1}-\kappa_1}B_{i,p-1}(x)+\frac{\kappa_{i+p}-x}{\kappa_{i+p-1}-\kappa_{i+}}B_{i+1,p-1}(x)$$

上式称为 de Boor-Cox 递推公式。分析可知，若上式次数为 0，则这些基函

数都是阶梯函数，也就等同于 0 阶 B-样条基函数的表达式。若节点有多个，则直接代入上式计算基函数的表达式。

2.3 变量选择方法

考虑线性回归模型：

$$y = \boldsymbol{X}\beta + \varepsilon$$

其中 $\mathrm{y} = (y_1, \cdots, y_n)^T$ 为被解释变量；$\boldsymbol{X} = \left(X_1, \cdots, X_p\right)^T$ 是 p 维解释变量；ε 为随机误差项，假定相互独立且服从正态分布 $\varepsilon_i \sim N\left(0, \sigma^2\right)$。

最小二乘法的的思想是通过最小化真实值和预测值间的平方误差来预测模型，这个平方误差值被称为残差平方和（RSS），满足：

$$RSS = \sum_{\mathrm{i}=1}^{\mathrm{n}} \left(y_i - \hat{y}_i\right)^2$$

最小化残差平方和：

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i\beta_i\right)^2$$

将其化为矩阵形式，可得：

$$\hat{\beta}_{OLS} = \arg\min_{\beta} \left(y - \boldsymbol{X}\beta\right)^T \left(y - \boldsymbol{X}\beta\right)$$

求解得：

$$\hat{\beta}_{OLS} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{X}^T y$$

要求解上述式子，则需要假设 $\boldsymbol{X}^T\boldsymbol{X}$ 为满秩矩阵。但在实际计算中 $\boldsymbol{X}^T\boldsymbol{X}$ 往往不是满秩矩阵。就像在高维数据模型中变量数远远大于样本数时，矩阵中的列数要多于行数，在 $\boldsymbol{X}^T\boldsymbol{X}$ 不满秩时，$\boldsymbol{X}^T\boldsymbol{X}$ 的行列式接近于0，在此时计算误差就会

很大，可解出多个解，从而影响估计结果。在包含多重共线性问题时，使用最小二乘估计会出现系数不稳定的情况。处理方法是对损失函数加上正则项，将不确定问题转化为确定性问题。

分组结构可能出于多种原因而产生，并导致截然不同的建模目标，组变量选择方法常用于分组结构，常见的例子包括通过一组指标变量表示渐进模型中的多级分类协变量，以及通过一组基函数表示连续变量的影响。分组也可以被引入到模型中，希望利用具有科学意义的先验知识。例如，在遗传关联研究中，来自同一基因的遗传标记可以被视为一组。在分析此类数据时，最好考虑分组结构。在本节中主要介绍后续模拟中使用的三种组变量选择方法以及两种双层变量选择方法。

2.3.1 Group Lasso 方法

当分组结构出现时，Lasso方法不能从组的水平进行变量选择，Group Lasso是Lasso的扩展，它能够将分好组的系数向量的每一组系数视为“单个”变量进行选择。与Lasso方法类似，即如果某组的系数向量全为0，则在变量选择中该组数据对应的特征全部舍弃，反之，若某组的系数向量不全为0，则在变量选择中该组数据对应的特征全部保留，这样提高了组间的稀疏性，适用于分组数据。

假设 Y 为 $n\times1$ 维向量，$\boldsymbol{X}_j$ 是第 j 个因子对应的 $n\times \mathrm{P}_j$ 维的矩阵，$\boldsymbol{\beta}_j$ 为 $\boldsymbol{X}_j$ 的系数向量，则由Yuan得目标函数为：

$$\hat{\beta}_{Group\ lasso}=\arg\min_{\beta}\frac{1}{2}\| Y-\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j \|_2^2+\lambda\sum_{j=1}^{J}\| \boldsymbol{\beta}\|_2$$

由上式可知，在Group Lasso中 p 个预测值被分成了 j 个组，这里的正则化系数 λ 满足 $\lambda>0$，与Lasso方法一致，在变量选择中用于控制数据的压缩，当 λ 越

大，就说明数据的压缩程度越高。而引入的系数向量 $\boldsymbol{\beta}_j$ 可以实现每个组的变量选择。

2.3.2 Group Scad 方法

因为Lasso类变量选择是有偏的，Fan于2001年提出的Scad方法是一种具有 L_1 和 L_2 惩罚项之和的凸惩罚函数，采用单调减少的惩罚项来进行无偏估计，而Group Scad方法是Scad的延伸，它将Scad方法延伸到群组结构中，可以将自变量分组，通过组间的稀疏性进行变量选择。

假设 Y 为 $n\times1$ 维向量，$\boldsymbol{X}_j$ 是第 j 个因子对应的 $n\times P_j$ 维的矩阵，由Zeng（2014）得目标函数为：

$$\hat{\beta}_{Group\ Scad}=\arg\min_{\beta}\| Y-\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j \|_2^2+\sum_{j=1}^{J}p_\lambda\left(\left\|\boldsymbol{\beta}_j\right\|_2\right)$$

其中，p_λ 表示为Scad惩罚，表示为：

$$f_\lambda(\theta)=\begin{cases}\lambda|\theta|, & 0\le|\theta|\le\lambda\\ -\dfrac{\theta^2-2a\lambda|\theta|+\lambda^2}{2(a-1)}, & \lambda\le|\theta|\le a\lambda\\ \dfrac{(a+1)\lambda^2}{2}, & |\theta|\ge a\lambda\end{cases}$$

其中，$a>2$。λ 为调节参数，在Fan和Li(2001)在Scad中参数 a 建议取值为3.7，而Zeng(2014)中建议取值为3。

由上述表达式可知Group Scad是将 L_2 惩罚添加到Scad惩罚中的一类惩罚方法，具有更强的稀疏性和组特征。

2.3.3 Group Mcp 方法

与上述相似，Group Mcp 也是对单变量选择方法施加群组的思想，由 Huang 和 Breheny[39]中首次提出，假设 Y 为 $n\times1$ 维向量，$\boldsymbol{X}_j$ 是第 j 个因子对应的 $n\times P_j$ 维的矩阵，则目标函数为：

$$\hat{\beta}_{Group\ Mcp}=\arg\min_{\beta}\| Y-\sum_{j=1}^{J}\boldsymbol{X}_j\boldsymbol{\beta}_j\|_2^2+\rho_\lambda(\boldsymbol{\beta})$$

其中，$\rho_\lambda(\beta)=\sum_{j=1}^{J}sign(\beta_j)\lambda\cdot\int_0^{|\beta_j|}\left(1-\frac{z}{\lambda b}\right)_+dz$

Group Mcp 方法能够在分组数据中进行变量选择，与单变量选择方法比较能够处理更复杂的群组数据，具有较强的适应性，但无法保障组内系数的稀疏性。

2.3.4 Spare Group Lasso 方法

一般来说，数据的稀疏性有两种类型，一类为“组间稀疏”，指变量中至少具有一个非零系数的组的数量，另一类为“组内稀疏”，指变量中每个非零组内存在非零系数的数量。而在组变量选择方法将整组数据选入模型时，要求整个变量组中的系数为非零，也就无法保证组间系数的稀疏性。双层变量选择方法可以实现变量的组内和组间的稀疏性，对整个区间上的数据都能进行有效分析。其中，Simon[40]提出一种 Spare Group Lasso（稀疏组 Lasso）的惩罚方法，假设 Y 为 $n\times1$ 维向量，$\boldsymbol{X}_j$ 是第 j 个因子对应的 $n\times P_j$ 维的矩阵，则目标函数为：

$$\hat{\boldsymbol{\beta}}_{Spare\ Group\ Lasso}=\arg\min_{\boldsymbol{\beta}}\| Y-\sum_{j=1}^{J}X_j\boldsymbol{\beta}_j\|_2^2+(1-\alpha)\lambda\sum_{j=1}^{J}\sqrt{p_j}\left\|\boldsymbol{\beta}_j\right\|_2+\alpha\lambda\sum_{l=1}^{p}|\beta_l|$$

其中，$\sum_{j=1}^{J}p_j=p$，α 是调整 L_1 范数惩罚 L_2 范数惩罚的参数，并且 $\alpha\in[0,1]$。

可以看到，稀疏组 Lasso 实际上为 Lasso 与 Group Lasso 这两种方法的结合。当 $\alpha=0$ 时，稀疏组 Lasso 的目标函数就转化为 Group Lasso 的目标函数，当 $\alpha=1$，

稀疏组 Lasso 的目标函数就转化为 Lasso 的目标函数，如此，确保了变量组间和组内都具有稀疏性。

2.3.5 Adaptive Spare Group Lasso 方法

自适应稀疏组 Lasso[41](Adaptive Sparse Group Lasso)是对稀疏组 Lasso 的扩展和改进。在传统稀疏组 Lasso 中，每个子集的权重向量是固定不变的，而在自适应稀疏组 Lasso 中，每个子集的权重向量是通过自适应的方式来确定，可以更好地适应数据分布的变化。具体来说，自适应稀疏组 Lasso 将每个特征分成一组，每组中的特征具有相似的属性，例如时间相关性或空间相关性；然后，自适应稀疏组 Lasso 通过学习每个特征组的不同权重，来确定每个子集的权重向量。这种方法使得自适应稀疏组Lasso更加灵活和精确，能够更好地适应不同类型的数据。主要思想是给惩罚项选取合适的权重，以解决 Lasso 方法的不一致性。由 Fang(2014)可知目标函数见下：

$$\hat{\boldsymbol{\beta}}_{Spare\ Group\ Lasso} = \arg\min_{\boldsymbol{\beta}} \| Y - \sum_{j=1}^{J} X_j \boldsymbol{\beta}_j \|_2^2 + (1-\alpha)\lambda \sum_{j=1}^{J} \omega_j \sqrt{p_j} \left\| \boldsymbol{\beta}_j \right\|_2 + \alpha\lambda \sum_{l=1}^{p} \nu_l \left| \beta_l \right|$$

其中 $\omega = (\omega_1, ..., \omega_J)^T$， $\nu = (\nu_1, ..., \nu_p)^T$ 分别为组变量以及单个变量的正数权重因子。

2.4 本章小结

本章首先介绍生存分析的基本概念，表明部分线性可加 Cox 能分析时依协变量与生存时间的关系；其次介绍样条回归及相关理论知识，阐述样条展开法逼近模型未知分量函数部分的合理性；最后介绍三种组变量选择方法以及两种双层变量选择方法，体现组变量选择在变量选择中的局限性。

3 模型设置

3.1 部分线性可加 Cox 模型

由于生存数据的复杂性，一般不把基准风险函数参数化。此时，左截断和右删失观测数据下回归参数的半参数估计、基准风险函数和受试者生存曲线的估计的研究具有科学意义。与Cox比例风险模型相比，可加风险模型指定：与一组随时间变化的协变量相关的风险函数是基准风险函数与协变量回归函数之和，而不是与协变量回归函数的乘积。这种表述描述了协变量与失效时间之间关系的不同方面，在许多应用中可加风险回归模型更合理。

对于一个给定的应用，可加风险模型是更合适的选择，尤其是对于连续的协变量，因为在Cox比例风险模型中，当给定的协变量元素每增加一个单位且其他的协变量保持不变时，风险函数被假定按照指数形式增长，在许多实际应用中，指数级增长过于极端。则当Cox比例风险模型拟合的适合性或者乘法协变量效应存在问题时，可加风险模型能解决上述问题。

在实践中，并非所有协变量都与风险函数呈线性相关，即其中一些协变量对风险函数具有非线性影响。此时考虑参数模型过于严格，而非参数模型则会导致“维数诅咒”。在这种情况下，部分线性模型结合了非参数建模的灵活性和参数建模的简约性和易解释性，避免了纯非参数模型维数的诅咒。为了在线性模型中纳入协变量的非线性效应，考虑与Hang（1999）类似的部分线性可加Cox模型。假设条件风险函数为：

$$\lambda\left(t\,|\,W,\ \boldsymbol{X}\right)=\lambda_0\left(t\right)\exp\left\{f\left(W\right)+\boldsymbol{X}\beta\right\},$$

其中，$f\left(\boldsymbol{W}\right)=\sum_{j=1}^{q}f_j\left(W_j\right),\boldsymbol{W}=\left(W_1,\ldots,W_q\right)^T$ 为 q 维协变量向量，$f_j\left(\cdot\right)\left(j=1,\ldots,q\right)$

为未知的非线性光滑函数。

该模型同时包含非参数分量 $f(W)$ 和参数分量 $\boldsymbol{X}\beta$。在实际应用中，很少所有的协变量对响应变量显著相关，也就是说，$\beta=(\beta_1,\cdots,\beta_p)$ 以及 $\{f_j(\cdot)\}_{j=1}^{q}$ 存在一些分量、函数为零或者非参数分量在区间上局部稀疏性等。在这种情况下，有效的双层变量选择方法不仅可以降低模型的复杂度，还可以增加模型预测精度。

3.2 参数估计

本文针对模型，提出应用 B-样条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 $\mathrm{t}\in[0,\mathrm{T}]$ 在区间 [a,b] 上取值，且 $j=1,2,\cdots p$，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 $I_{Kq}=[\xi_q,\xi_{q+1}],\mathrm{q}=0,1,\cdots K$，其中 K 满足 $K=K(n)=n^{\nu}\quad 0<\nu<1/2$ 并使得 $\max_{1\le q\le K+1}\left|\xi_q-\xi_{q+1}\right|=O\left(n^{-\nu}\right)$ 成立。

此时定义 S_n 为满足以下条件的多项式样条空间：$a)$ I_{Kq} 为 S_n 的子集，且 $1\le q\le K$；$b)$ 对于 $\ell\ge 2$ 与 $0\le\ell'\le\ell-2$，函数 s 是 ℓ 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 ϕ_k 使得对于任意 $f_{nj}\in S_n$ 都存在：

$$f_{nj}(w)=\sum_{k=1}^{m_n}\gamma_{jk}\phi_{jk}(W_j),\quad 1\le j\le p.$$

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。

在本文中对非参数分量部分采用B-样条逼近处理得：

$$\lambda(t|\boldsymbol{W},\ \boldsymbol{X})=\lambda_0(t)\exp\left\{\sum_{j=1}^{q}\sum_{k=1}^{m_n}\gamma_{jk}\phi_{jk}(W_j)+\boldsymbol{X}\beta\right\}$$

为了书写方便，令 $\boldsymbol{\gamma}_j=(\gamma_{j1},\cdots,\gamma_{jm_n})^T$ $\boldsymbol{\gamma}=(\boldsymbol{\gamma}_1^T,\cdots,\boldsymbol{\gamma}_p^T)^T$，对于任意 $w\in[a,b]$，可知存在对应的 B-样条基函数：

$$\boldsymbol{\Phi}=(\boldsymbol{\Phi}_1,\cdots,\boldsymbol{\Phi}_q)=\begin{bmatrix}\phi_{11}(W_{11}) & \cdots & \phi_{1m_n}(W_{11}) & \cdots & \phi_{q1}(W_{1q}) & \cdots & \phi_{qm_n}(W_{1q})\\ \\ \phi_{11}(W_{n1}) & \cdots & \phi_{1m_n}(W_{n1}) & \cdots & \phi_{q1}(W_{nq}) & \cdots & \phi_{qm_n}(W_{nq})\end{bmatrix}_{n\times(q*m_n)}$$

作为新的一组 $n\times(q*m_n)$ 的系数矩阵，为了书写简洁，此处及本文后面都令 $\boldsymbol{\Phi}=\boldsymbol{\Phi}(w)$，利用上式中定义的基函数去逼近未知分量函数 $f_j^*(\cdot)$。将模型矩阵化表示为：

$$\lambda(t|W,\ \boldsymbol{X})=\lambda_0(t)\left\{\boldsymbol{X}^*\boldsymbol{\beta}^*\right\}$$

其中，$X^*=(\boldsymbol{\Phi},X_1,\ldots X_p)$ 为新模型的设计矩阵，$\boldsymbol{\beta}^*=(\boldsymbol{\gamma}_1^T,\ldots,\boldsymbol{\gamma}_q^T,\beta^T)^T$ 为长度为 m_n*q+p 的带估计参数，前 m_n*q 各元素为 q 组B-样条系数，最后 p 个元素可以看做组长度为1的线性结构回归系数。此时对数偏似然函数为：

$$l_n(\boldsymbol{\beta}^*)=\sum_{i=1}^{n}\Delta_i\left[\boldsymbol{X}^*\boldsymbol{\beta}^*-\log\sum_{l=1}^{n}\left\{Y_l(T_i)\exp\left(\boldsymbol{X}^*\boldsymbol{\beta}^*\right)\right\}\right]$$

为了对组变量和个体变量进行同时选择，令 $\boldsymbol{\beta}_j^*=\boldsymbol{\gamma}_j,j=1,\cdots,q$；$\boldsymbol{\beta}_j^*=\beta_j,j=q+1,\cdots,p+q$，采用自适应稀疏组Lasso惩罚方法，惩罚负对数偏似然函数为：

$$pl_n\left(\boldsymbol{\beta}^*\right)=-l_n\left(\boldsymbol{\beta}^*\right)+(1-\alpha)\lambda\sum_{j=1}^{q+p}\omega_j\sqrt{g_j}\left\|\boldsymbol{\beta}_j^*\right\|_2+\alpha\lambda\sum_{l=1}^{p+q^*m_n}\nu_l\left|\beta_l\right|$$

其中，g_j 为第 j 组变量的个数，$\omega=(\omega_1,...,\omega_J)^T$，$\nu=(\nu_1,...,\nu_p)^T$ 分别为组变量以及单个变量的正数权重因子，当取值都相同时就退化为了稀疏组 Lasso。

3.3 本章小结

本章主要对部分线性可加 Cox 模型中的基函数进行样条拟合，实现未知分量函数的有效逼近；推导双层变量选择下的部分线性可加 Cox 模型的参数估计，说明有效的双层变量选择方法不仅可以降低模型的复杂度，还可以增加模型预测精度。

4 数值模拟与实例分析

4.1 块坐标下降法

对于高维问题，最小二乘近似中的伪响应向量有时变得难以或不可能计算。Meier等[42]使用块坐标梯度下降算法解决了逻辑回归中的类似问题。在本文中通过用负对数偏似然代替损失函数，将其算法扩展到部分线性可加Cox模型中。

坐标下降法是每次沿着单个维度方向进行搜索，当得到一个当前维度最小值之后再循环使用不同的维度方向，最终收敛得到最优解。而块坐标下降法是坐标下降法的更一般形式，块坐标下降法的关键思想是将对数似然的二次近似和额外的线搜索相结合来进行迭代，它通过对变量的子集进行同时优化，把原问题分解为多个子问题。

在进行 K+1 次迭代时，利用 $\boldsymbol{\beta}^{*(k)}$ 的二阶泰勒展开式，并且使用正定的海瑟矩阵 $H^{(k)}$ 来代替对数偏似然函数 $l_n\left(\boldsymbol{\beta}^*\right)$，则上文中的 $pl_n\left(\boldsymbol{\beta}^{*(k)}+\boldsymbol{d},\lambda\right)$ 可以表示为：

$$S_\lambda^{(k)}(d)=\left\{L\left(\boldsymbol{\beta}^{*(k)}\right)+\boldsymbol{d}^T\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)+\frac{1}{2}\boldsymbol{d}^T H^{(k)}\boldsymbol{d}\right\}$$
$$+(1-\alpha)\lambda\sum_{j=1}^{p+q}\omega_j\sqrt{g_j}\left\|\boldsymbol{\beta}_j^{*(k)}+\boldsymbol{d}_j\right\|_2+\alpha\lambda\sum_{l=1}^{p+q*m_n}v_l\left|\beta_l^{*(k)}+d_l\right|$$

其中，$\mathbf{d}\in\mathbb{R}^{p+q*m_n}$ 称作步长，用来进行迭代，矩阵 $H^{(k)}$ 是块对角矩阵，并且 $m_n\times m_n$ 维的 $H_{jj}^{(k)}$ 为 $h_j^{(k)}$ 的子集，并且满足 $H_{jj}^{(k)}=h_j^{(k)}I_{m_n}$，$h_j^{(k)}$ 是海瑟矩阵 $(\ddot{L}\left(\boldsymbol{\beta}^*\right))_{jj}$ 的最大对角线值。因为 $H^{(k)}$ 是块对角矩阵，所以惩罚项是块可分的，则对于 $S_\lambda^{(k)}(d)$ 的最小化问题可以理解为并行求解的 P 个子问题。对于第 j 组变量，步长为：

$$\hat{\boldsymbol{d}}_j=\arg\min_{\boldsymbol{d}_j}\left\{\boldsymbol{d}_j^T\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j+\frac{1}{2}\boldsymbol{d}_j^T H_{jj}^{(k)}\boldsymbol{d}_j+(1-\alpha)\lambda\omega_j\sqrt{g_j}\left\|\boldsymbol{\beta}_j^{*(k)}+\boldsymbol{d}_j\right\|_2+\alpha\lambda\sum_{l\in G_j}v_l\left|\beta_l^{*(k)}+d_l\right|\right\}$$

若$\left\|\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j-h_j^{(k)}\boldsymbol{\beta}^{*(k)}\right\|_2\le\lambda$，则上式中的最小值为：

$$\hat{\boldsymbol{d}}_j=-\boldsymbol{\beta}_j^{*(k)}$$

若$\left\|\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j-h_j^{(k)}\boldsymbol{\beta}^{*(k)}\right\|_2>\lambda$，则上式中的最小值为：

$$\hat{\boldsymbol{d}}_j=-\frac{1}{h_j^{(k)}}\left[\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j-\lambda\frac{\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j-h_j^{(k)}\boldsymbol{\beta}_j^{*(k)}}{\left\|\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j-h_j^{(k)}\boldsymbol{\beta}_j^{*(k)}\right\|_2}\right]$$

然而用上述近似代替Hessian矩阵时，步长$\hat{\boldsymbol{d}}_j$会导致当接近静止点时，牛顿方向的收敛速度与牛顿方向的速度不同。沿所获得的方向找到合适的步长$\hat{\boldsymbol{d}}_j$，使用Armijo规则执行不精确的线搜索。建立$\alpha^{(k)}$为几何序列

$\left\{\tau^v\right\}(0<\tau<1,\tau=0,1,2,\cdots)$的最大值，则满足以下:

$$pl_n\left(\boldsymbol{\beta}^{*(k)}+\alpha^{(\mathrm{k})}\hat{\boldsymbol{d}},\lambda\right)-pl_n\left(\boldsymbol{\beta}^{*(k)},\lambda\right)\le\alpha^{(\mathrm{k})}\sigma\Delta^{(k)}$$

其中，$0<\sigma<1$，并且$\Delta^{(k)}$是对目标函数$pl_n(\cdot)$的改进，则有

$$\Delta^{(k)}=-\hat{d}_j^T\left(\dot{L}\left(\boldsymbol{\beta}^{*(k)}\right)\right)_j+\lambda\left(\left\|\boldsymbol{\beta}_j^{(k)}+\hat{d}_j\right\|_2-\left\|\beta_{nj}^{(k)}\right\|_2\right)$$

最后，迭代方程为：

$$\boldsymbol{\beta}_j^{*(k+1)}=\boldsymbol{\beta}_j^{*(k)}+\alpha^{(k)}\hat{\boldsymbol{d}}_j$$

4.2 数值模拟

为证明双层变量选择方法在部分线性可加 Cox 模型的有限样本性能，在本节中利用蒙特卡洛模拟分别将 Group Lasso、Group Scad、Group Mcp、Sparse Group Lasso、Adaptive Sparse Group Lasso 这五种变量选择方法应用在模型中，具体分类见表 4.1。下文将上述变量选择方法简称为 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso。为了与真实模型进行对比，假设已知显著的线性结构变量以及非线性结构变量，并只采用这些变量建立失效时间的部分线性可加 Cox 模型，得到模型的 Orcale 方法估计值，分别将变量选择方法估计值与 Oracle 估计值对比分析。其中，通过改变协变量的分布将模拟分为三种情况：

第一种情况：协变量 $\boldsymbol{X}$ 满足正态分布，协变量 $\boldsymbol{W}$ 满足均匀分布；

第二种情况：协变量 $\boldsymbol{X}$ 满足正态分布且相关系数值设置为 0.5（即 $\rho(X_i, X_j) = 0.5^{|i-j|}; i \neq j; i, j = 1, \cdots, p$ ），协变量 $\boldsymbol{W}$ 满足均匀分布；

第三种情况：协变量 $\boldsymbol{X}$ 与第二种情况相同，协变量 $\boldsymbol{W}$ 满足均匀分布并设置零区间。

表 4.1 模拟惩罚项选择方案

类型	变量选择方法
组变量选择方法	G-Lasso
	G-Scad
	G-Mcp
双层变量选择方法	SG-Lasso
	ASG-Lasso

本文采用三次 B-样条曲线[43]拟合非参数部分中的未知函数，经 AIC 准则验

证取 8 个节点效果最好。通过 R 语言进行模拟分析，主要使用 grpreg、grpreGoverlap、ncvreg、SGL、groCox、splines 和 glmnet 包中的函数进行计算。

将 Oracle 估计值作为基准，使用五种性能指标来对不同的变量选择方法进行评价，五种性能指标分别为选择的真实组数（TG）、选择的零组数（FG）、选择为非零的真实非零变量数（TP）、选择非零的真零变量数（FP）以及预测误差（PE）。PE 定义为：

$$\left\|\left\{\hat{\beta}^T X+\hat{\phi}_1(W_1)+\hat{\phi}_2(W_2)\right\}\right\|-\left\|\left\{\beta^T X+\phi_1(W_1)+\phi_2(W_2)\right\}\right\|.$$

4.2.1 第一种情况

数据生成仿照 Arfan(2020)[44]中的方法设置线性结构变量以及非参数函数分别为 20 个，真实模型由 4 个线性结构协变量以及 2 个非参数函数组成，其中线性结构变量的系数为$(2,2,-2,-2,0,...,0)$，即前四个变量为显著的重要变量；非参数函数为$\phi_1(W_1)=Sin(2\pi w)$、$\phi_2(W_2)=Cox(2\pi w)$、$\phi_l(W_l)\equiv 0, l=3,...,20$。各协变量满足独立同分布，即$X \overset{iid}{\sim} N(0,I)$，$W \overset{iid}{\sim} U(0,I)$，实际例子中非参数函数的协变量的区间不满足[0,1]区间时，采用归一化的方式进行处理以简化计算。风险函数由模型$\lambda(t|\boldsymbol{W}, \boldsymbol{X})=\lambda_0(t)\exp\{f(\boldsymbol{W})+\boldsymbol{X}\beta\}$生成，设置删失变量为[0,C]的均匀分布，并控制 C 使得删失比率为 10%、20%、40%的情况下，假定风险函数$h_0(t)=1.0$，对样本量$n=100$、200、400 时分别进行 500 次模拟研究。

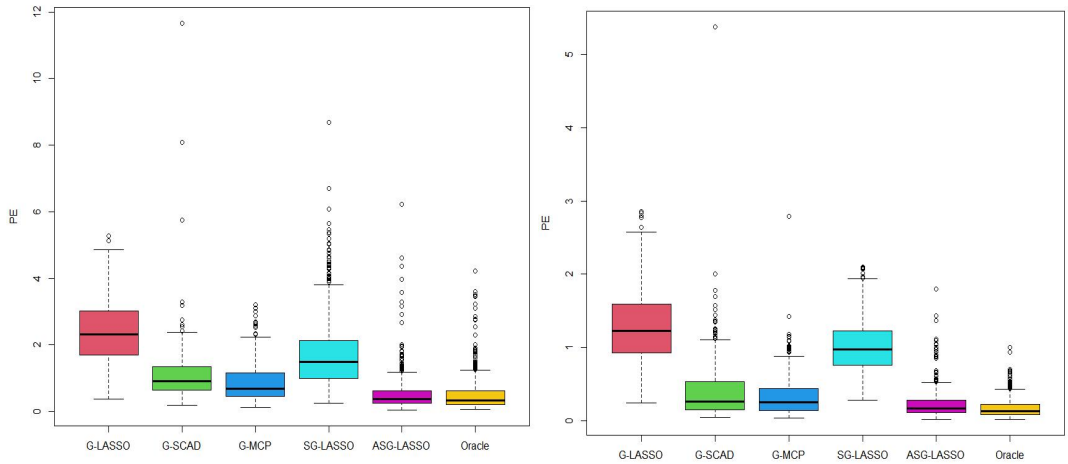
4.2.1.1 删失比为 10%

设置删失比 10%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本$n=100$、200、400 时，评价指标的均值和标准差见表 4.2：

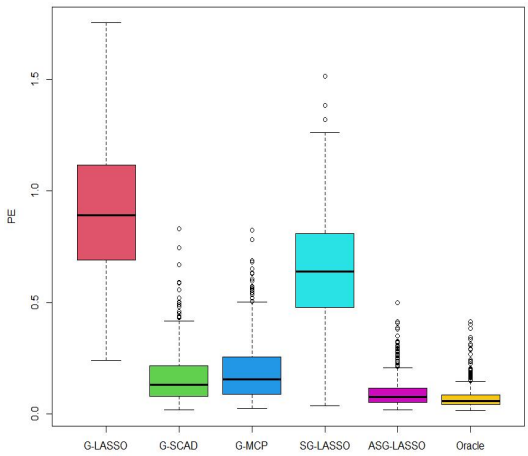
表 4.2 $n=100$、200、400 时不同变量选择方法的均值（标准差）

样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	2（0）	6.13（2.37）	4（0）	4.78（1.95）	2.38（0.94）
	G-Scad	1.90（0.44）	0.89（1.09）	4（0）	1.49（1.25）	1.06（0.78）
	G-Mcp	1.80（0.61）	0.39（0.69）	4（0）	0.85（0.93）	0.88（0.57）
	SG-Lasso	1.60（0.72）	1.36（1.49）	4（0）	4.12（2.65）	1.77（1.17）
	ASG-Lasso	2（0.16）	0.50（0.98）	4（0）	1.19（1.19）	0.55（0.59）
	Oracle	2（0）	NA	4（0）	NA	0.53（0.57）
$n=200$	G-Lasso	2（0）	8.03（2.83）	4（0）	6（2）	1.29（0.51）
	G-Scad	2（0）	1.56（1.77）	4（0）	2.10（1.60）	0.40（0.40）
	G-Mcp	2（0.09）	0.49（0.88）	4（0）	1.10（1.10）	0.33（0.27）
	SG-Lasso	2（0）	1.16（1.14）	4（0）	5.70（2）	1.02（0.35）
	ASG-Lasso	2（0）	0.57（1.13）	4（0）	1.40（1.80）	0.23（0.21）
	Oracle	2（0）	NA	4（0）	NA	0.18（0.14）
$n=400$	G-Lasso	2（0）	10.48（2.85）	4（0）	7.20（2.30）	0.91（0.30）
	G-Scad	2（0）	1（1.41）	4（0）	1.70（1.50）	0.16（0.12）
	G-Mcp	2（0）	0.48（0.91）	4（0）	1（1.20）	0.19（0.14）
	SG-Lasso	2（0）	0.14（0.38）	4（0）	3.20（20）	0.65（0.24）
	ASG-Lasso	2（0）	0.35（0.72）	4（0）	1.60（1.35）	0.10（0.07）
	Oracle	2（0）	NA	4（0）	NA	0.08（0.06）

从表 4.2 中可知，对于评价指标选择的真实组数（TG）以及选择为非零的真实非零变量数（TP），五种变量选择方法的均值和方差与 Oracle 估计值几乎一致，说明模型有效。随着样本量的增大，当 $n=400$ 时，与 G-Lasso、G-Scad、G-Mcp 相比，SG-Lasso 和 ASG-Lasso 选择了更少的零组数并且方差较小，其中，SG-Lasso 选择的零组数（FG）的均值为最小值 0.14。在选择非零的真零变量数（FP）中，G-MCP 表现最好，错误选出的非零变量个数最小。预测误差（PE）箱线图见图 4.1：



（a） $n=100$ （b） $n=200$



（c） $n=400$

图 4.1 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.2 和图 4.1 可知，由于 Lasso 的有偏性，G-Scad、G-Mcp 与 G-Lasso、SG-Lasso 相比具有更小的预测误差，但在 ASG-Lasso 中已得到改善，预测误差（PE）基本与 Oracle 估计值相同。

4.2.1.2 删失比为 20%

设置删失比 20%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.3：

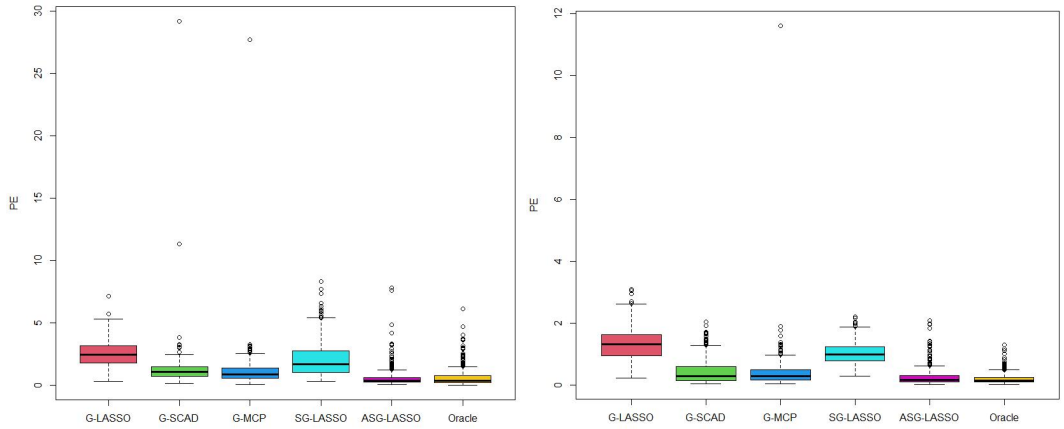
表 4.3 $n=100$、200、400 时不同变量选择方法的均值（标准差）

样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	2（0）	5.82（2.39）	4（0）	4.70（2）	2.25（1.02）
	G-Scad	1.80（0.54）	0.73（1.18）	4（0）	1.28（1.25）	1.24（1.45）
	G-Mcp	1.60（0.73）	0.28（0.61）	4（0）	0.71（0.92）	1.10（1.35）
	SG-Lasso	1.40（0.84）	1.09（1.36）	4（0）	9.63（2.91）	2.08（1.41）
	ASG-Lasso	1.90（0.26）	0.48（0.83）	4（0）	1.11（1.11）	0.60（1.72）
	Oracle	2（0）	NA	4（0）	NA	0.63（0.68）
$n=200$	G-Lasso	2（0）	8.10（2.90）	4（0）	5.83（2.10）	1.34（0.52）
	G-Scad	2（0）	1.25（1.50）	4（0）	1.84（1.40）	0.44（0.38）
	G-Mcp	2（0）	0.45（1.20）	4（0）	0.96（1.30）	0.39（0.58）
	SG-Lasso	2（0）	1.20（1.20）	4（0）	5.63（1.90）	1.04（0.36）
	ASG-Lasso	2（0）	0.63（1.30）	4（0）	1.46（1.70）	0.26（0.27）
	Oracle	2（0）	NA	4（0）	NA	0.21（0.17）

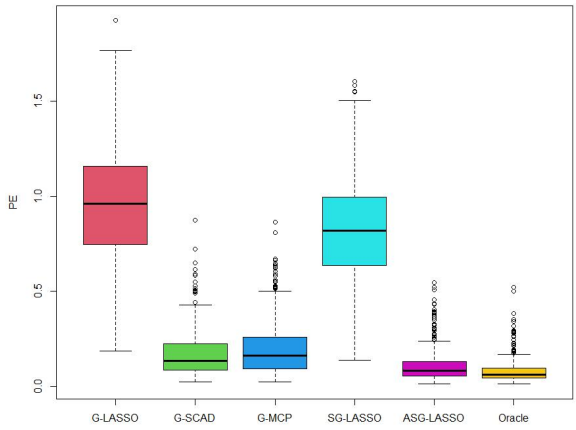
续表 4.3

样本量	方法	TG	FG	TP	FP	PE
$n=400$	G-Lasso	2（0）	10.13（2.90）	4（0）	6.94（2.20）	0.96（0.30）
	G-Scad	2（0）	0.77（1.20）	4（0）	1.48（1.40）	0.17（0.12）
	G-Mcp	2（0）	0.38（0.89）	4（0）	0.97（1.20）	0.20（0.14）
	SG-Lasso	2（0）	0.14（1.39）	4（0）	6.15（1.90）	0.83（0.25）
	ASG-Lasso	2（0）	0.40（0.82）	4（0）	1.19（1.15）	0.11（0.08）
	Oracle	2（0）	NA	4（0）	NA	0.08（0.06）

从表 4.3 中可知，评价指标选择的真实组数（TG）和选择为非零的真实非零变量数（TP）在五种变量选择方法的均值和方差与 Oracle 估计值几乎一致，说明在删失比为 20%的情形下模型有效。整体趋势与删失比 10%的数值趋势相同，SG-Lasso 选择的零组数（FG）的均值为最小值 0.14。选择非零的真零变量数（FP）中，SG-Lasso 不如 G-MCP，但同样在 ASG-Lasso 中逐渐改善。预测误差（PE）箱线图见图 4.2：



（a） $n=100$ （b） $n=200$



（c） $n=400$

图 4.2 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.3 和图 4.2 可知，当删失比为 20%且样本量较小时，五种变量选择方法的异常值增多，在图中显示为箱体图上下边缘接近坐标轴，但预测误差预测误差（PE）基本与 Oracle 估计值相同。随着样本量的增大，异常值对数据预测误差的影响较小。

4.2.1.3 删失比为 40%

设置删失比 40%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.4：

表 4.4 $n=100$、200、400 时不同变量选择方法的均值（标准差）

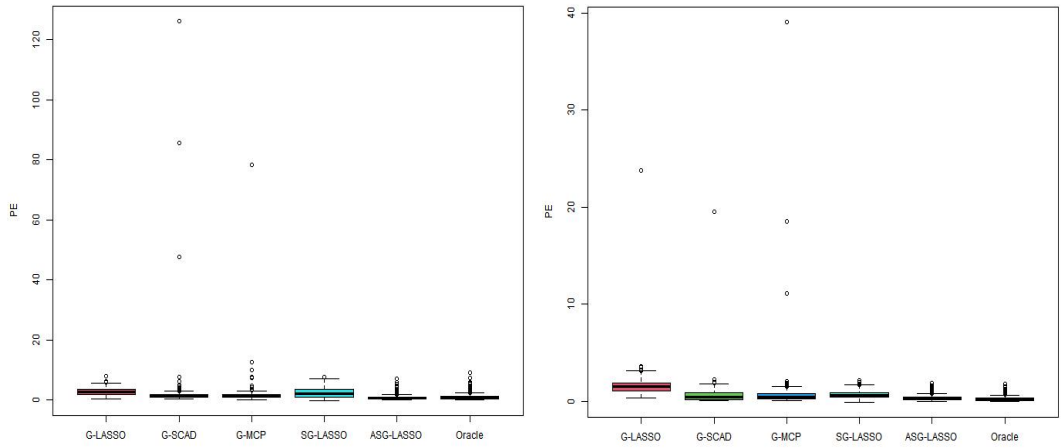
样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	1.97（0.2）	5.58（2.4）	4（0）	4.65（1.92）	2.81（1.15）
	G-Scad	1.36（0.83）	0.45（1.08）	4（0）	0.90（1.19）	1.98（7.06）
	G-Mcp	0.98（0.90）	0.10（0.46）	4（0）	0.41（0.91）	1.70（3.58）
	SG-Lasso	0.86（0.91）	0.59（1.09）	4（0）	8.32（3.19）	2.44（1.60）

续表 4.4

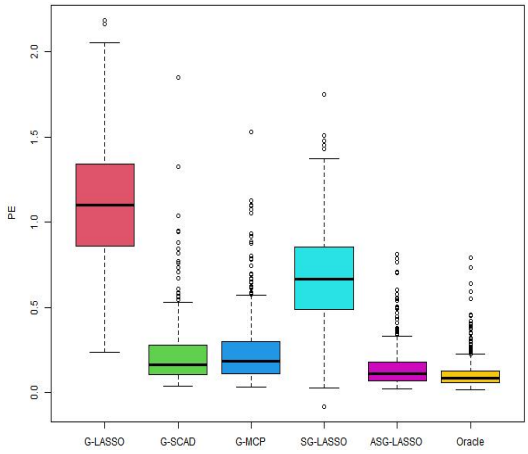
样本量	方法	TG	FG	TP	FP	PE
$n=100$	ASG-Lasso	1.82(0.44)	0.70（1）	4（0）	1.26（1.15）	0.83（0.78）
	Oracle	2（0）	NA	4（0）	NA	0.97（1.06）
$n=200$	G-Lasso	2（0）	7.65（2.88）	4（0）	5.70（2.17）	1.60（1.17）
	G-Scad	1.99（0.11）	1（1.50）	4（0）	1.40（1.45）	0.63（0.97）
	G-Mcp	1.98（0.18）	0.33（1.46）	4（0）	0.71（1.31）	0.66（1.99）
	SG-Lasso	2（0.06）	1.32（1.27）	4（0）	10.73（1.98）	0.72（0.99）
	ASG-Lasso	2（0）	0.58（1.20）	4（0）	1.29（1.55）	0.34（0.29）
	Oracle	2（0）	NA	4（0）	NA	0.29（0.24）
$n=400$	G-Lasso	2（0）	9.75（2.87）	4（0）	6.71（2.20）	1.11（0.35）
	G-Scad	2（0）	0.52（0.97）	4（0）	1.12（1.13）	0.22（0.18）
	G-Mcp	2（0）	0.23（0.62）	4（0）	0.62（0.83）	0.24（0.19）
	SG-Lasso	2（0）	0.19（0.44）	4（0）	9.25（1.99）	0.68（0.27）
	ASG-Lasso	2（0）	0.58（1.07）	4（0）	1.69（1.93）	0.15（0.12）
	Oracle	2（0）	NA	4（0）	NA	0.11（0.09）

从表 4.4 可知，评价指标选择的真实组数（TG）在 $n=100$ 时五种变量选择方法时表现较差；在 $n=200$ 时 SG-Lasso、ASG-Lasso 优于 G-Scad、G-Mcp；在 $n=400$ 时五种变量选择方法的均值和方差与 Oracle 估计值保持一致，也说明了在删失比为 40%的情形下模型仍旧有效。指标选择的零组数（FG）可知 G-Lasso 总会选择出更多不重要的变量，模拟结果较差。G-Mcp 选择非零的真零变量数

（FP）中仍旧表现最好。预测误差（PE）箱线图见图 4.3：



（a） $n=100$ （b） $n=200$



（c） $n=400$

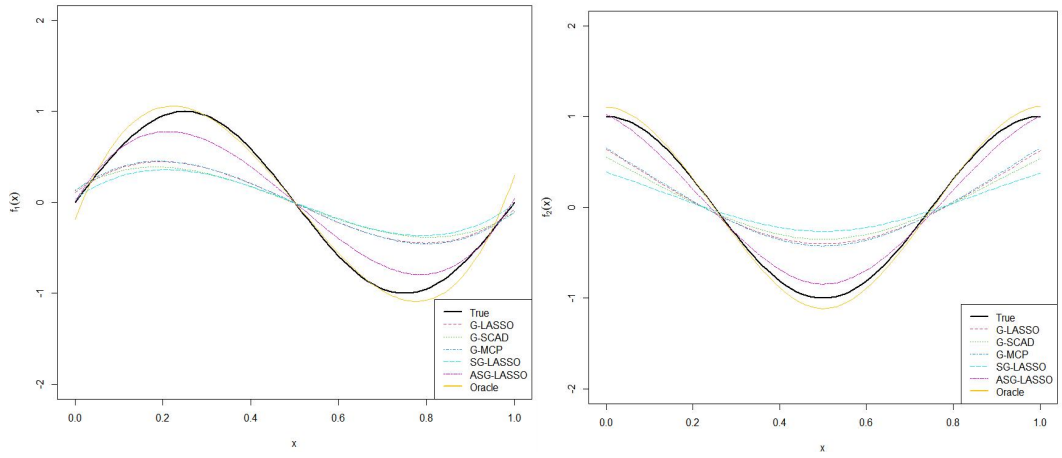
图 4.3 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.4 和图 4.3 可知，当删失比为 40%且样本量较小时，G-Lasso、G-Scad、G-Mcp 存在极端异常值。随着样本量的增大 ASG-Lasso 的预测误差与 Oracle 估计值最为接近，当 $n=400$ 时，均值差为 0.04，方差差为 0.03。

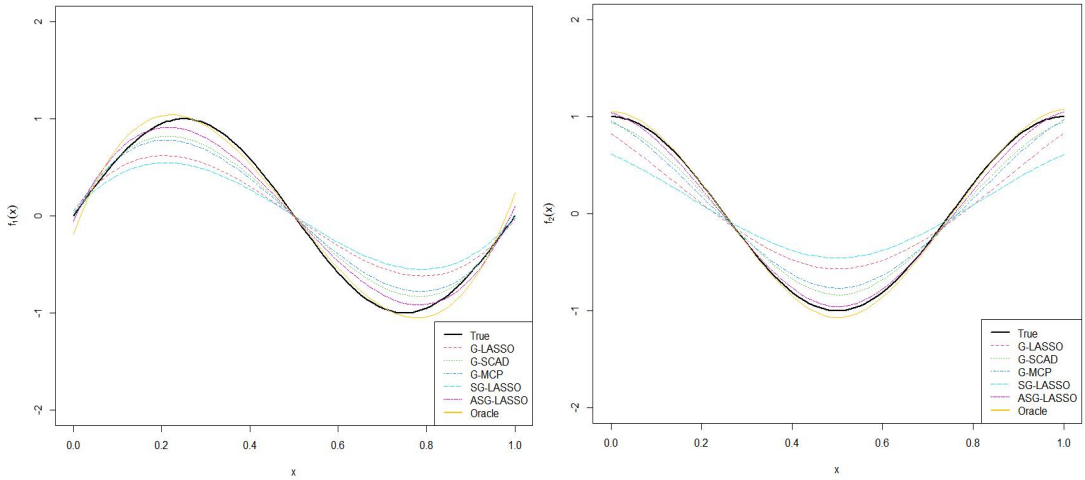
4.2.1.4 不同删失比下的函数拟合图

本小节给出函数样条拟合估计值，黑色实线代表真实值，其余六条曲线分别代表 G-Lasso、G-Scad、G-Mcp、SG-Lasso、ASG-Lasso 以及 Oracle 估计的平均曲线。当协变量满足正态分布，不同删失比下，样本量 $n=100$、200、400 时函数拟合见图 4.4-4.6。

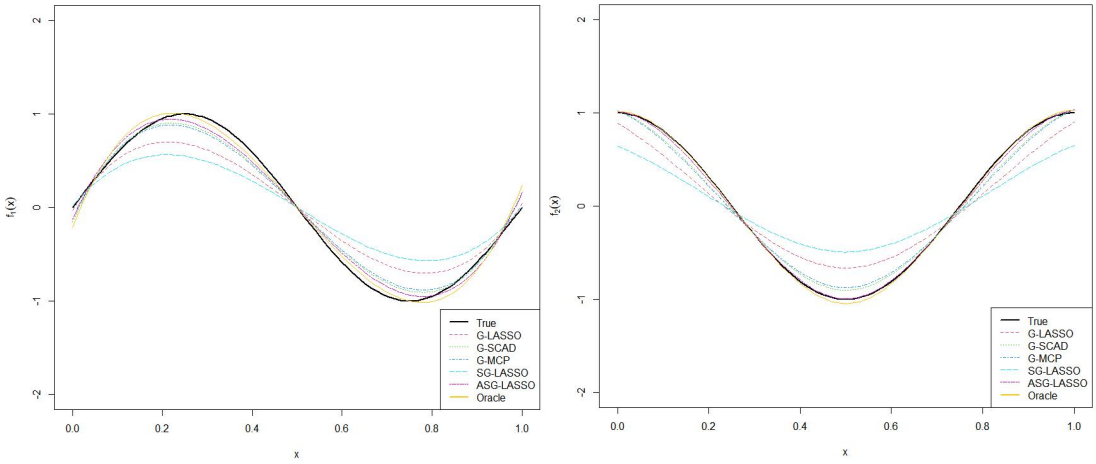
删失比为 10%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



(a) $n=100$



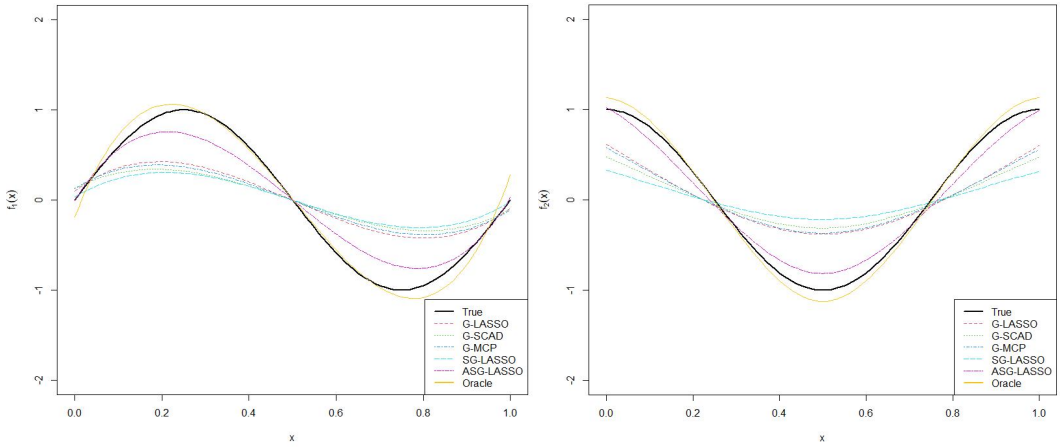
(b) $n=200$



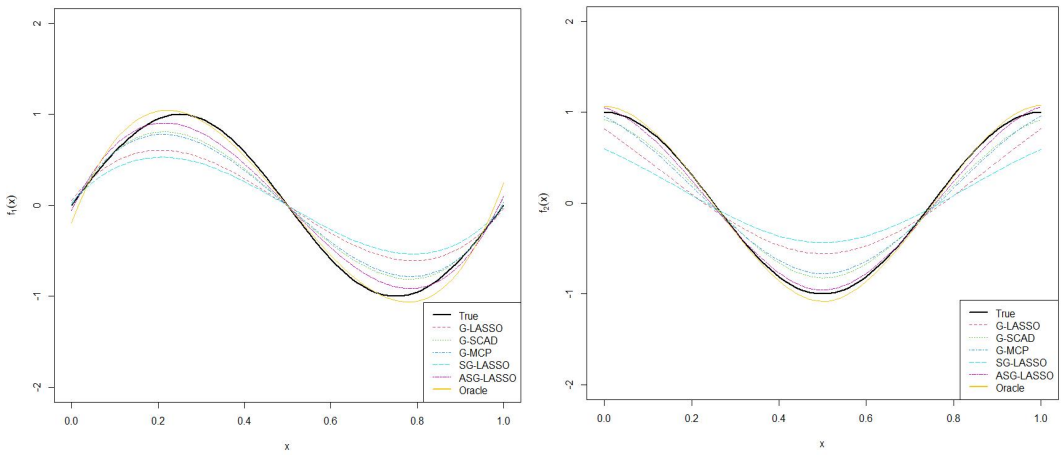
（c） $n=400$

图 4.4 $n=100$、200、400 时 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图

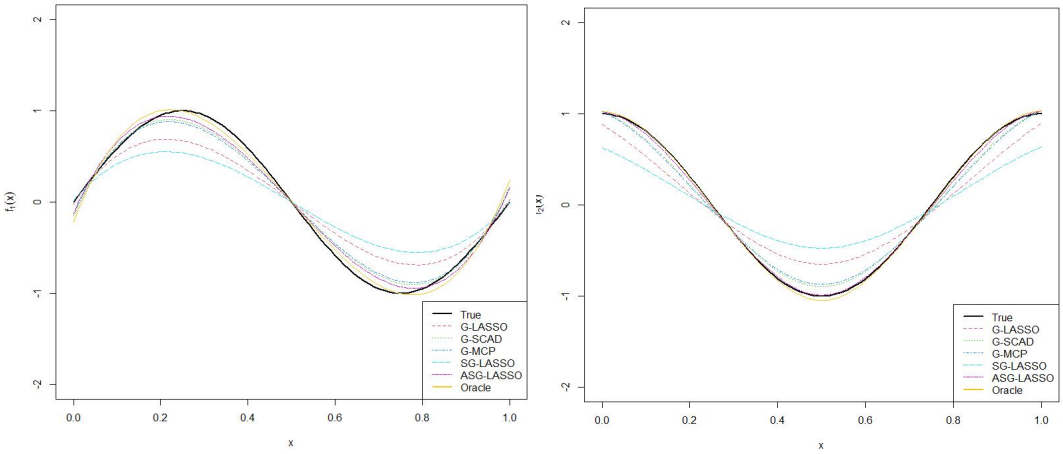
删失比为 20%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



（a） $n=100$



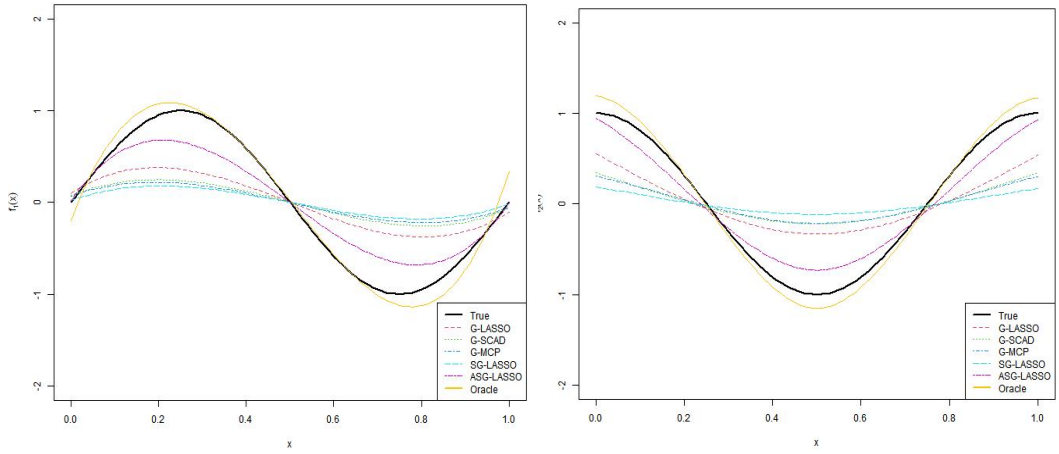
（b） $n=200$



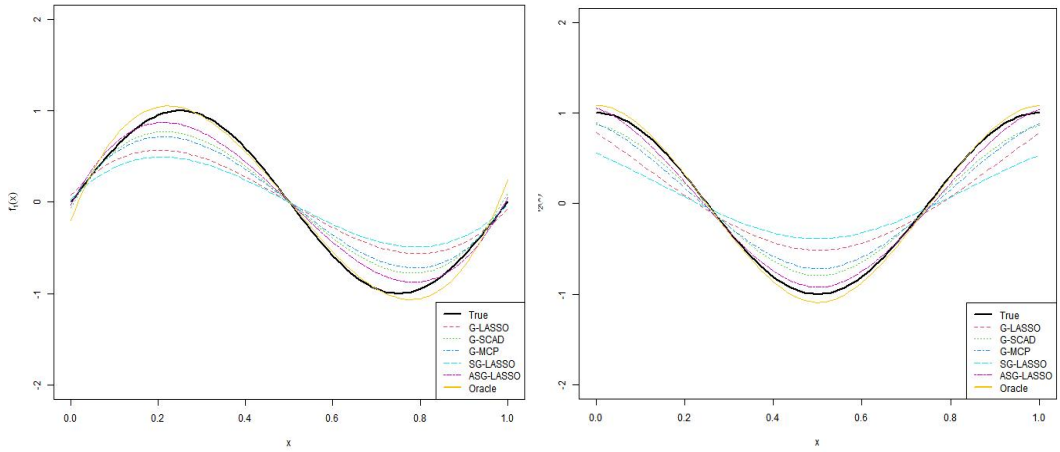
（c） $n=400$

图 4.5　$n=100$、200、400 时 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图

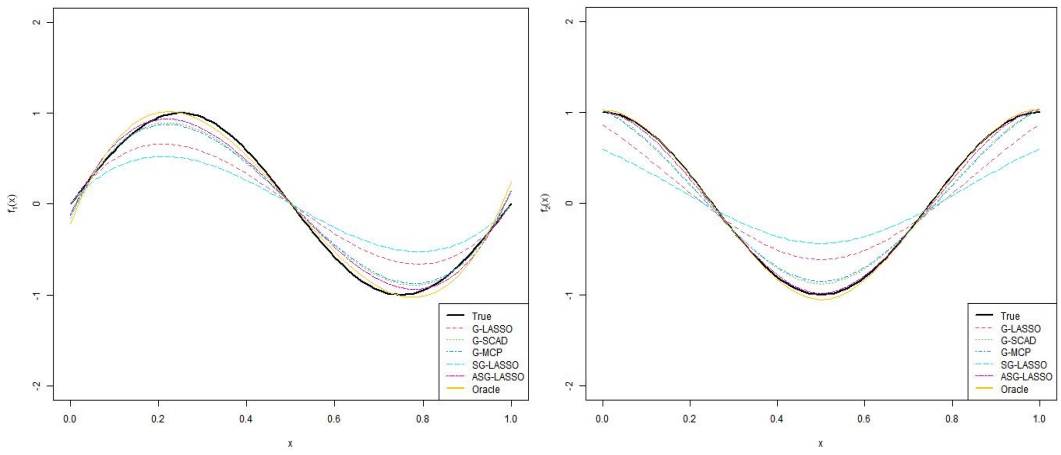
删失比为 40%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



(a) $n=100$



(b) $n=200$



(c) $n=400$

图 4.6 $n=100$、200、400 时 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图

4.2.2 第二种情况

数据生成与第一种情况类似，各协变量满足独立同分布，$X \overset{iid}{\sim} N(0,1)$，$W \overset{iid}{\sim} U(0,1)$。协变量 $\boldsymbol{X}$ 满足正态分布且相关系数值设置为 0.5，协变量 $\boldsymbol{W}$ 满足均匀分布。风险函数由模型 $\lambda(t|\boldsymbol{W},\boldsymbol{X})=\lambda_0(t)\exp\{f(\boldsymbol{W})+\boldsymbol{X}\beta\}$ 生成，设置删失变量为[0,C]的均匀分布，并控制 C 使得删失比率为 10%、20%、40%的情况下，假定风险函数 $h_0(t)=1.0$，对样本量 $n=100$、200、400 时分别进行 500 次模拟研究。

4.2.2.1 删失比为 10%

设置删失比 10%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.5：

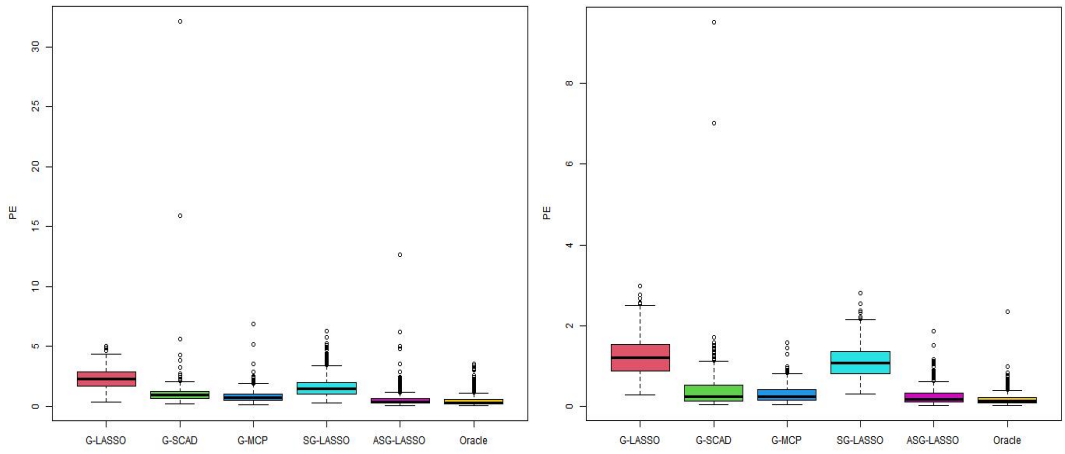
表 4.5 $n=100$、200、400 时不同变量选择方法的均值（标准差）

样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	1.90（0.08）	6.10（2.44）	4（0）	4（1.67）	2.20（0.87）
	G-Scad	1.93（0.30）	1.10（1.48）	4（0）	1.30（1.28）	1.07（1.64）
	G-Mcp	1.80（0.48）	0.40（0.92）	4（0）	0.80（0.99）	0.82（0.60）
	SG-Lasso	1.70（0.63）	1.40（1.54）	4（0）	9.40（2.47）	1.64（0.98）
	ASG-Lasso	1.90（0.14）	0.60（1.19）	4（0）	0.9（1.04）	0.55（0.78）
	Oracle	2（0）	NA	4（0）	NA	0.51（0.57）
$n=200$	G-Lasso	2（0）	8.58（2.75）	4（0）	5.24（2.03）	1.22（0.48）
	G-Scad	2（0）	1.52（1.88）	4（0）	1.82（1.72）	0.39（0.59）
	G-Mcp	2（0）	0.56（1.01）	4（0）	0.99（1.13）	0.31（0.22）

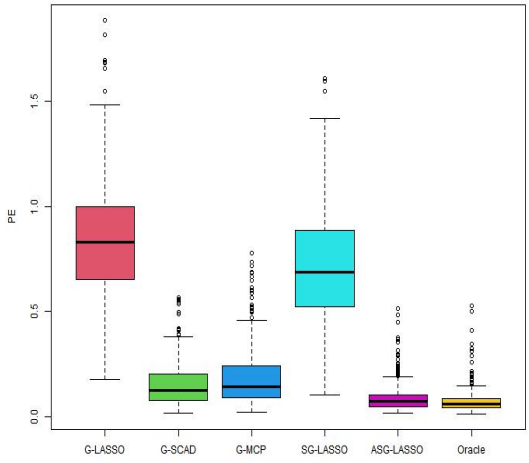
续表 4.5

样本量	方法	TG	FG	TP	FP	PE
$n=200$	SG-Lasso	2（0）	0.86（1.05）	4（0）	9.08（1.89）	1.12（0.39）
	ASG-Lasso	2（0）	0.61（1.20）	4（0）	1.27（1.60）	0.25（0.20）
	Oracle	2（0）	NA	4（0）	NA	0.18（0.17）
$n=400$	G-Lasso	2（0）	10.90（2.75）	4（0）	6.26（2.15）	0.84（0.27）
	G-Scad	2（0）	1.02（1.46）	4（0）	1.56（1.36）	0.15（0.10）
	G-Mcp	2（0）	0.52（0.98）	4（0）	1.04（1.19）	0.18（0.13）
	SG-Lasso	2（0）	0.06（0.25）	4（0）	7.35（1.97）	0.72（0.26）
	ASG-Lasso	2（0）	0.32（0.71）	4（0）	1.45（1.84）	0.09（0.06）
	Oracle	2（0）	NA	4（0）	NA	0.07（0.05）

从表 4.5 可知，当协变量间存在相关系数时，五种变量选择方法在指标选择的真实组数（TG）和选择为非零的真实非零变量数（TP）在删失比为 10%的情形下表现较好，说明了模型的有效性。当协变量存在相关系数时，由指标选择的零组数（FG）可知，随着样本量的增长，SG-Lasso、ASG-Lasso 比 G-Scad、G-MCP 选择了更少的不重要变量，模拟结果较好。G-MCP 选择非零的真零变量数（FP）中表现最好。预测误差（PE）箱线图见图 4.7：



（a） $n=100$ （b） $n=200$



（c） $n=400$

图 4.7 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.5 和图 4.7 可知，当 $n=100$、200 时，预测误差均值多数大于 1，但当 $n=400$ 时，预测误差均值均小于 1，也就说明五种变量选择方法在样本量越大时，产生的预测误差越小。

4.2.2.2 删失比为 20%

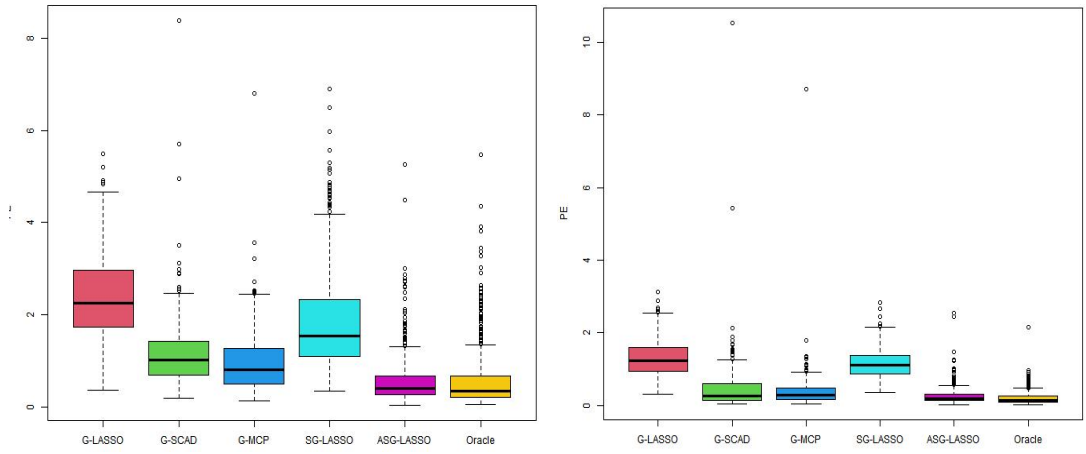
设置删失比 20%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.6：

表 4.6 $n=100$、200、400 时不同变量选择方法的均值（标准差）

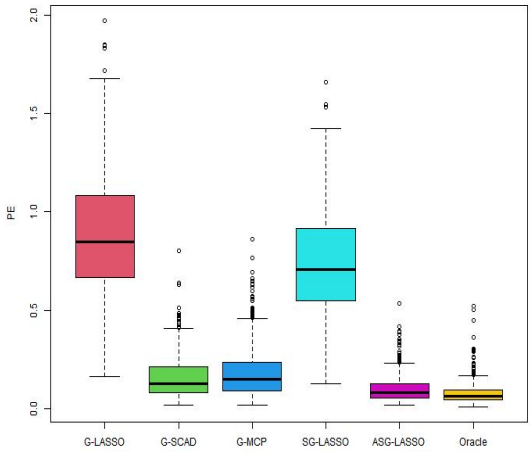
样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	1.99（0.99）	6.11（2.35）	4（0）	4.01（1.65）	2.35（0.89）
	G-Scad	1.83（0.49）	0.83（1.21）	4（0）	1.12（1.08）	1.13（0.68）
	G-Mcp	1.68（0.65）	0.32（0.69）	4（0）	0.65（0.90）	0.96（0.63）
	SG-Lasso	1.52（0.78）	1.23（1.52）	4（0）	8.80（2.67）	1.88（1.14）
	ASG-Lasso	1.93（0.27）	0.63（1.10）	4（0）	0.98（1.06）	0.58（0.55）
	Oracle	2（0）	NA	4（0）	NA	0.60（0.67）
$n=200$	G-Lasso	2（0）	8.32（2.90）	4（0）	5.17（2.01）	1.28（0.50）
	G-Scad	2（0）	1.40（1.85）	4（0）	1.65（1.54）	0.43（0.62）
	G-Mcp	2（0）	0.43（1.13）	4（0）	0.80（1.16）	0.37（0.45）
	SG-Lasso	2（0）	0.89（1.05）	4（0）	9.13（1.93）	1.14（0.39）
	ASG-Lasso	2（0）	0.59（1.25）	4（0）	1.16（1.47）	0.26（0.24）
	Oracle	2（0）	NA	4（0）	NA	0.20（0.18）
$n=400$	G-Lasso	2（0）	10.7（2.93）	4（0）	6.17（2.09）	0.88（0.30）
	G-Scad	2（0）	0.92（1.35）	4（0）	1.46（1.33）	0.16（0.11）
	G-Mcp	2（0）	0.46（0.90）	4（0）	0.90（1.09）	0.19（0.13）
	SG-Lasso	2（0）	0.08（0.29）	4（0）	7.41（1.85）	0.74（0.27）
	ASG-Lasso	2（0）	0.38（0.78）	4（0）	1.44（1.82）	0.10（0.07）
	Oracle	2（0）	NA	4（0）	NA	0.08（0.06）

从表 4.6 可知，评价指标选择的真实组数（TG）和选择为非零的真实非零变

量数（TP）在五种变量选择方法的均值和方差与 Oracle 估计值几乎一致，说明在删失比为 20%的情形下模型有效。SG-Lasso 在指标选择的零组数（FG）的均值为最小值 0.08。G-MCP 选择非零的真零变量数（FP）中仍旧表现最好。预测误差（PE）箱线图见图 4.8：



（a） $n=100$ （b） $n=200$



（c） $n=400$

图 4.8 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.6 和图 4.8 可知，随着样本量的增大，五种变量选择方法的误差逐渐减小，预测误差（PE）最小的 ASG-Lasso，在 $n=400$ 时预测误差与 Oracle 估

计值均值相差 0.02，方差相差 0.01。

4.2.2.3 删失比为 40%

设置删失比 40%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.7：

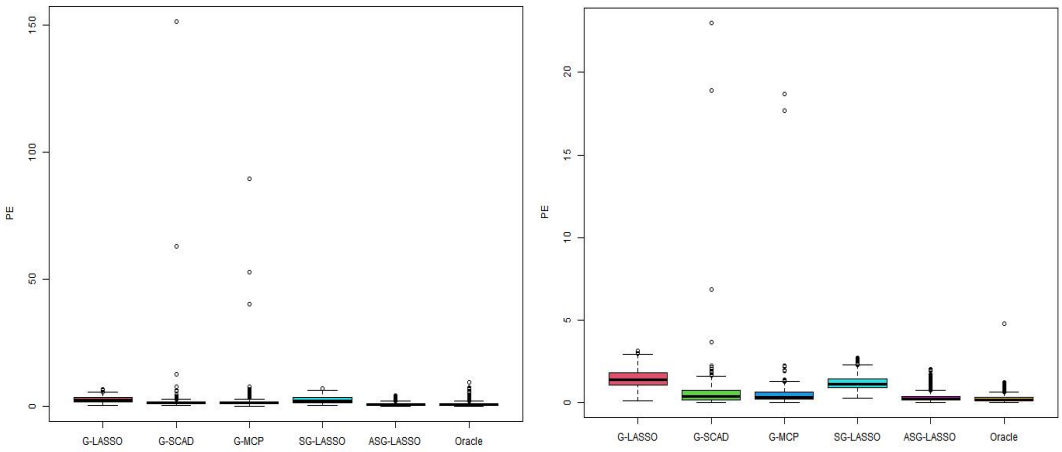
表 4.7 $n=100$、200、400 时不同变量选择方法的均值（标准差）

样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	1.96（0.19）	6.07（2.53）	4（0）	3.87（1.76）	2.66（1.09）
	G-Scad	1.55（0.70）	0.57（1.19）	3.99（0.13）	0.86（1.07）	1.81（7.30）
	G-Mcp	1.21（0.82）	0.20（0.70）	3.98（0.20）	0.40（0.78）	1.78（4.96）
	SG-Lasso	1.04（0.93）	0.91（1.37）	4（0）	7.76（2.98）	2.48（1.40）
	ASG-Lasso	1.79（0.47）	0.66（1）	4（0）	0.89（0.99）	0.78（0.67）
	Oracle	2（0）	NA	4（0）	NA	0.92（1.10）
$n=200$	G-Lasso	2（0）	7.71（2.66）	4（0）	4.85（2.03）	1.48（0.55）
	G-Scad	2（0.06）	0.99（1.69）	4（0）	1.29（1.56）	0.63（1.40）
	G-Mcp	1.98（0.14）	0.31（1.19）	4（0）	0.67（1.34）	0.55（1.17）
	SG-Lasso	1.99（0.10）	0.98（1.13）	4（0）	9.33（1.87）	1.22（0.42）
	ASG-Lasso	2（0）	0.64（1.25）	4（0）	1.13（1.25）	0.34（0.29）
	Oracle	2（0）	NA	4（0）	NA	0.28（0.30）
$n=400$	G-Lasso	2（0）	10.06（2.78）	4（0）	5.96（1.94）	1.02（0.34）
	G-Scad	2（0）	0.70（1.24）	4（0）	1.16（1.32）	0.19（0.13）

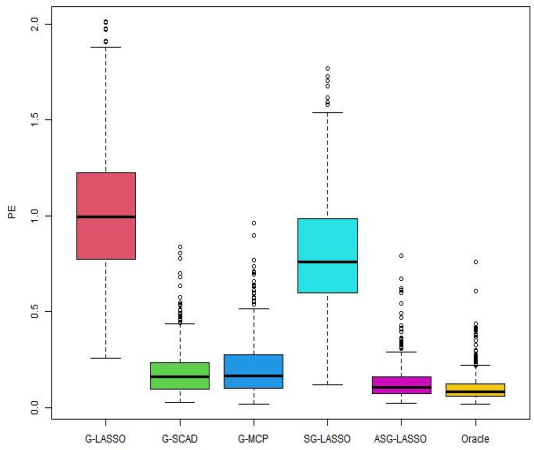
续表 4.7

样本量	方法	TG	FG	TP	FP	PE
$n=400$	G-Mcp	2（0）	0.30（0.69）	4（0）	0.58（0.96）	0.21（0.15）
	SG-Lasso	2（0）	0.11（0.34）	4（0）	7.51（1.85）	0.79（0.29）
	ASG-Lasso	2（0）	0.53（1.07）	4（0）	1.40（1.76）	0.13（0.10）
	Oracle	2（0）	NA	4（0）	NA	0.11（0.08）

从表 4.7 可知，评价指标选择的真实组数（TG）和选择为非零的真实非零变量数（TP）在五种变量选择方法的均值和方差与 Oracle 估计值几乎一致，说明在删失比为 40%的情形下模型有效。由指标选择的零组数（FG）可知，SG-Lasso 选择了更少的不重要变量，均值最小值为 0.11。G-MCP 选择非零的真零变量数（FP）中的均值和方差最小。预测误差（PE）箱线图见图 4.9：



（a） $n=100$ （b） $n=200$



（c） $n=400$

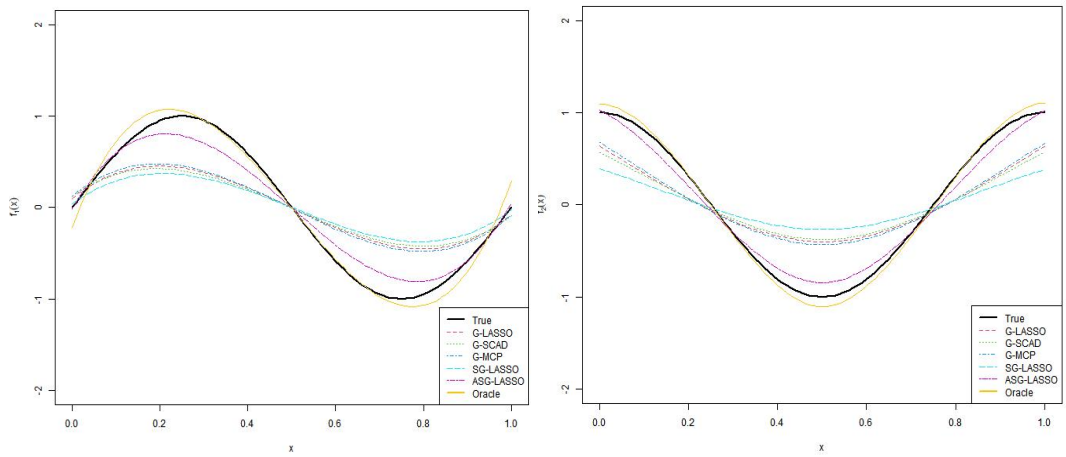
图 4.9 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.7 和图 4.9 可知，随着样本量的增加，五种变量选择方法的预测误差（PE）趋近于 0。其中预测误差（PE）最小的 ASG-Lasso，在 $n=400$ 时预测误差与 Oracle 估计值均值相差 0.02，方差相差 0.02。

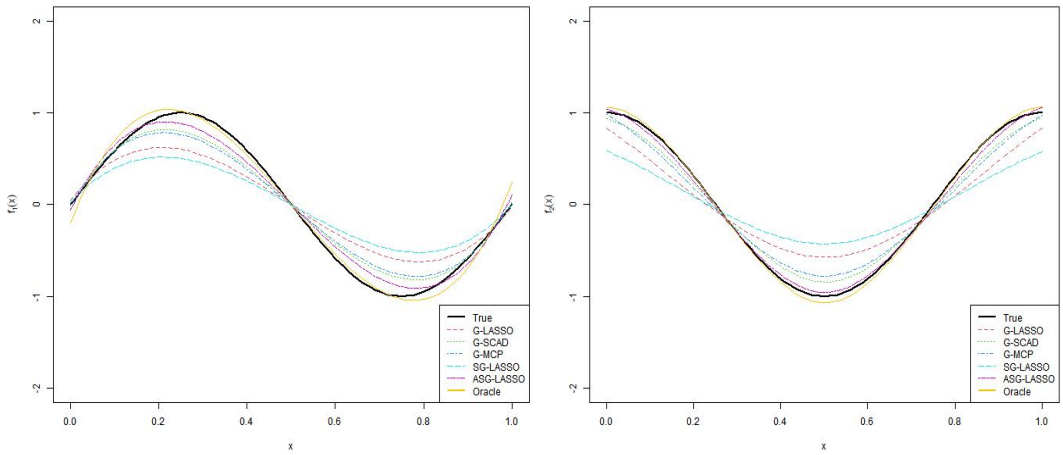
4.2.2.4 不同删失比下的函数拟合图

本小节给出函数样条拟合估计值，黑色实线代表真实值，其余六条曲线分别代表 G-Lasso、G-Scad、G-Mcp、SG-Lasso、ASG-Lasso 以及 Oracle 估计的平均曲线。当协变量满足正态分布且相关系数为 0.5，不同删失比下，样本量 $n=100$、200、400 时函数拟合见图 4.10-4.12。

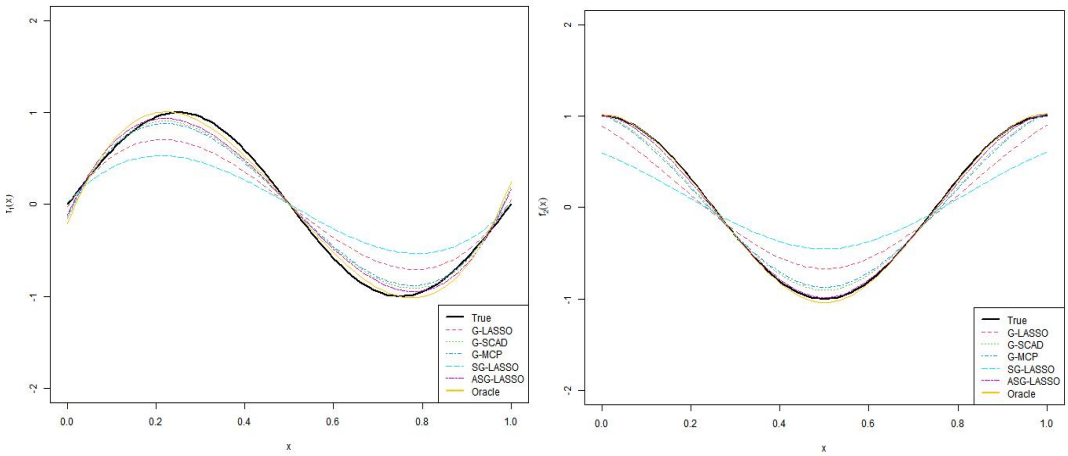
删失比为 10%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



(a) $n=100$



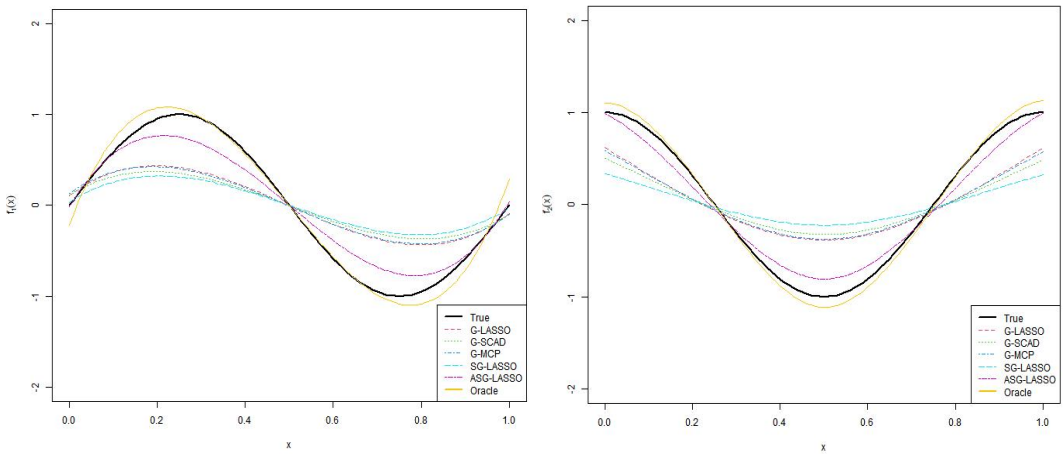
(b) $n=200$



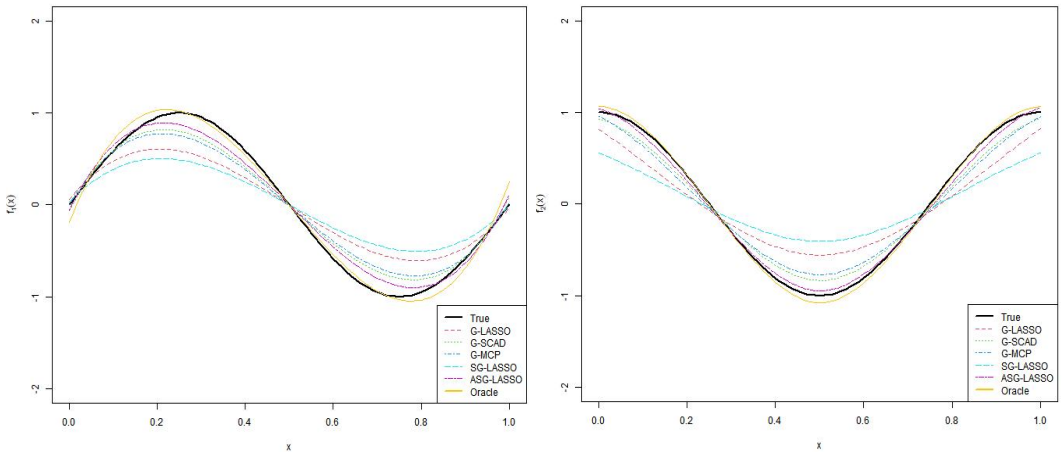
(c) $n=400$

图 4.10 $n=100$、200、400 时 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图

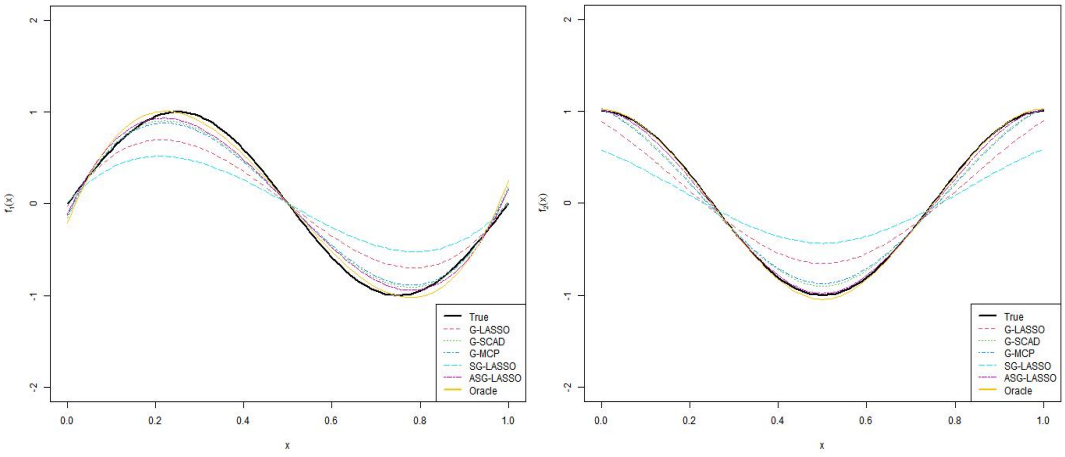
删失比为 20%时，非参数分量$\phi_1(W_1)$与$\phi_2(W_2)$的函数拟合图为：



（a） $n=100$



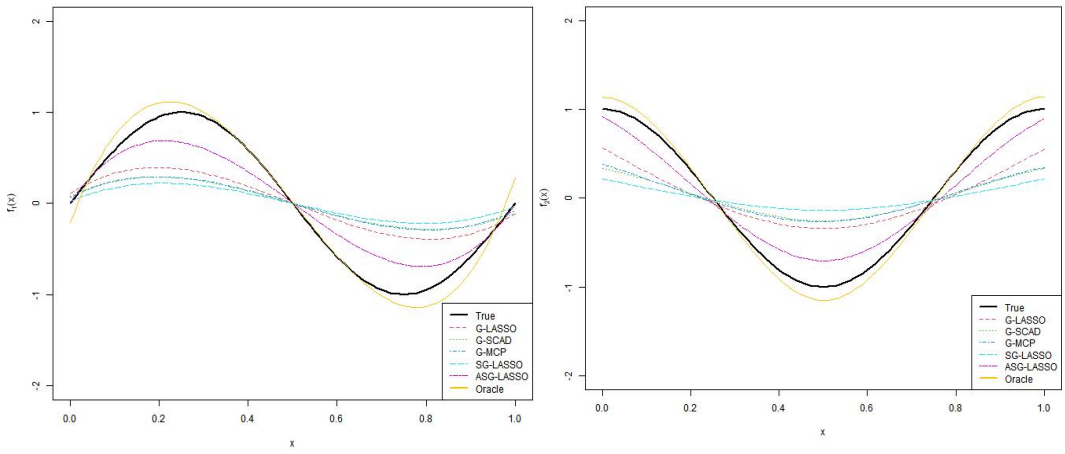
（b） $n=200$



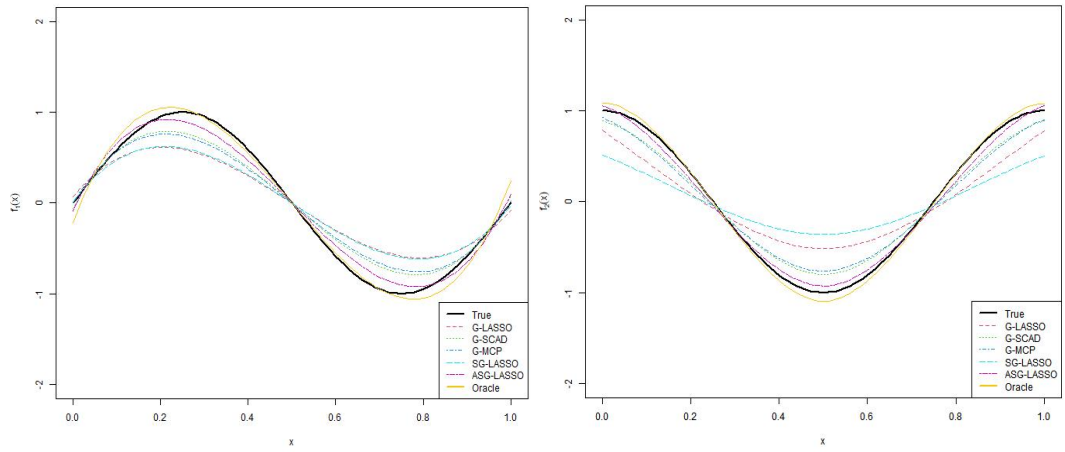
（c） $n=400$

图 4.11　$n=100$、200、400 时$\phi_1(W_1)$与$\phi_2(W_2)$的函数拟合图

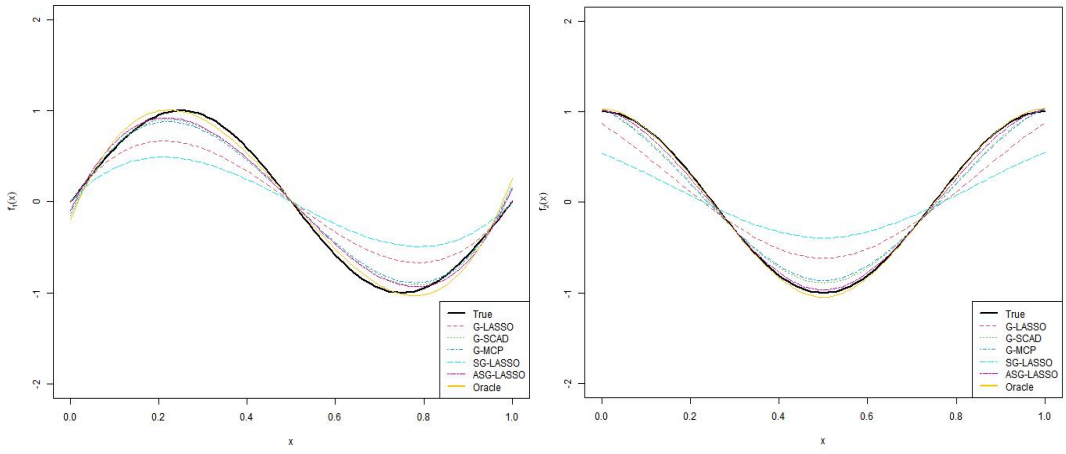
删失比为 40%时，非参数分量$\phi_1(W_1)$与$\phi_2(W_2)$的函数拟合图为：



(a) $n=100$



(b) $n=200$



(c) $n=400$

图 4.12 $n=100$、200、400 时$\phi_1(W_1)$与$\phi_2(W_2)$的函数拟合图

4.2.3 第三种情况

数据生成与第一种情况类似，各协变量满足独立同分布，$X \overset{iid}{\sim} N(0,1)$，$W \overset{iid}{\sim} U(0,1)$。协变量 $\boldsymbol{X}$ 满足正态分布且相关系数值设置为 0.5，协变量 $\boldsymbol{W}$ 满足均匀分布且存在零区间，非参数函数为：

$$\phi_1(W_1)=\begin{cases}2Sin(2\pi w), \mathrm{w}<0.5\\ 0 \quad\quad , \mathrm{w}\geq 0.5\end{cases}$$

$$\phi_2(W_2)=\begin{cases}0 \quad\quad , w\leq 0.25\\ 2Cox(2\pi w), 0.25<w<0.75\\ 0 \quad\quad , w\geq 0.75\end{cases}$$

$$\phi_l(W_l)\equiv 0, l=3,...,20$$

风险函数由模型 $\lambda(t\,|\,\boldsymbol{W},\ \boldsymbol{X})=\lambda_0(t)\exp\{f(\boldsymbol{W})+\boldsymbol{X}\beta\}$ 生成，设置删失变量为[0,C]的均匀分布，并控制 C 使得删失比率为 10%、20%、40%的情况下，假定风险函数 $h_0(t)=1.0$，对样本量 $n=100$、200、400 时分别进行 500 次模拟研究。

4.2.3.1 删失比为 10%

设置删失比 10%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.8：

表 4.8 $n=100$、200、400 时不同变量选择方法的均值（标准差）

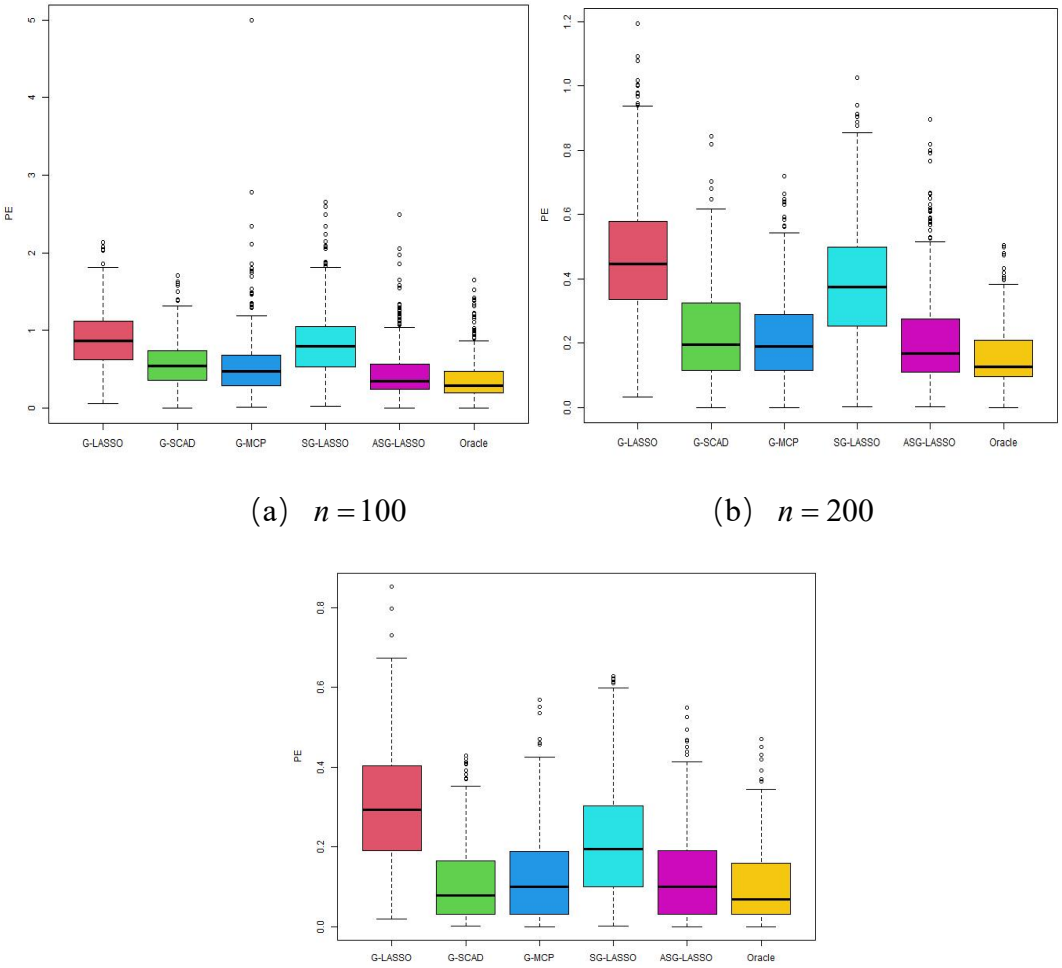
样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	2（0.06）	4.36（2.24）	4（0）	3.97（1.84）	0.87（0.34）
	G-Scad	1.98（0.17）	1.05（1.23）	4（0）	1.80（1.26）	0.57（0.25）
	G-Mcp	1.95（0.26）	0.53（1.14）	4（0）	1.21（1.12）	0.54（0.67）

续表 4.8

样本量	方法	TG	FG	TP	FP	PE
$n=100$	SG-Lasso	1.94（0.31）	2.26（1.91）	4（0）	3.10（2.18）	0.82（0.38）
	ASG-Lasso	1.98（0.16）	0.42（0.93）	4（0）	0.93（1.08）	0.40（0.32）
	Oracle	2（0）	NA	4（0）	NA	0.33（0.23）
$n=200$	G-Lasso	2（0）	6.12（2.76）	4（0）	4.78（2.05）	0.46（0.19）
	G-Scad	2（0）	1.67（1.66）	4（0）	2.14（1.50）	0.23（0.15）
	G-Mcp	2（0）	0.53（0.88）	4（0）	1.10（1.10）	0.21（0.13）
	SG-Lasso	2（0）	3.83（2.50）	4（0）	5.03（2.10）	0.38（0.18）
	ASG-Lasso	2（0）	0.74（1.36）	4（0）	1.23（1.53）	0.20（0.15）
	Oracle	2（0）	NA	4（0）	NA	0.15（0.10）
$n=400$	G-Lasso	2（0）	8.49（2.94）	4（0）	6.05（2.30）	0.29（0.13）
	G-Scad	2（0）	1.22（1.52）	4（0）	1.87（1.46）	0.09（0.08）
	G-Mcp	2（0）	0.69（1.17）	4（0）	1.27（1.27）	0.10（0.08）
	SG-Lasso	2（0）	5.22（2.77）	4（0）	5.79（2.15）	0.20（0.12）
	ASG-Lasso	2（0）	0.70（1.23）	4（0）	1.68（2.04）	0.11（0.09）
	Oracle	2（0）	NA	4（0）	NA	0.08（0.07）

从表 4.8 可知，评价指标选择的真实组数（TG）在 $n=100$ 时，G-Scad、G-Mcp、SG-Lasso、ASG-Lasso 的均值和方差略有波动，这种误差随着样本量的增加而消失，说明在协变量 X 中存在零区间时，模型仍旧有效。由指标选择的零组数（FG）可知，ASG-Lasso 选择了更少的不重要变量，优于 SG-Lasso。G-MCP、ASG-Lasso

选择非零的真零变量数（FP）的均值和方差较小。预测误差（PE）箱线图见图 4.13：



（a） $n=100$ （b） $n=200$

（c） $n=400$

图 4.13 $n=100$、200、400 时预测误差（PE）箱线图

结合表 4.8 和图 4.13 可知，当协变量 X 存在零区间时，五种变量选择方法的预测误差都小于 1，预测效果优于上述两种情况。随着样本量的增加，预测误差（PE）值趋近于零。

4.2.3.2 删失比为 20%

设置删失比 20%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.9：

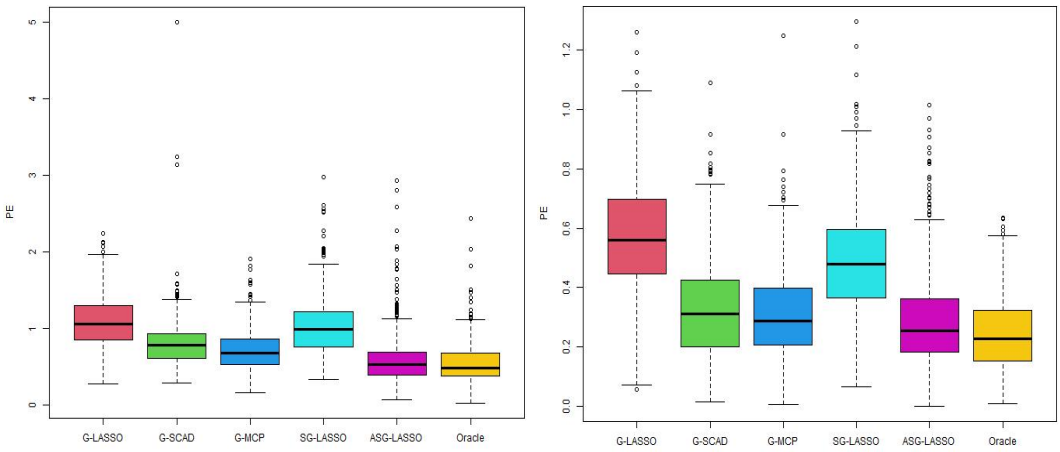
表 4.9 $n=100$、200、400 时不同变量选择方法的均值（标准差）

样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	2（0.06）	4.36（2.24）	4（0）	3.97（1.84）	0.98（0.33）
	G-Scad	1.98（0.17）	1.05（1.23）	4（0）	1.80（1.26）	0.71（0.35）
	G-Mcp	1.95（0.26）	0.53（1.14）	4（0）	1.21（1.12）	0.62（0.38）
	SG-Lasso	1.94（0.31）	2.26（1.91）	4（0）	3.10（2.18）	0.93（0.38）
	ASG-Lasso	1.98（0.16）	0.42（0.93）	4（0）	0.93（1.08）	0.50（0.35）
	Oracle	2（0）	NA	4（0）	NA	0.43（0.25）
$n=200$	G-Lasso	2（0）	6.12（2.76）	4（0）	4.78（2.05）	0.57（0.20）
	G-Scad	2（0）	1.67（1.66）	4（0）	2.14（1.50）	0.33（0.18）
	G-Mcp	2（0）	0.53（0.88）	4（0）	1.10（1.10）	0.32（0.15）
	SG-Lasso	2（0）	3.83（2.50）	4（0）	5.03（2.10）	0.49（0.18）
	ASG-Lasso	2（0）	0.74（1.36）	4（0）	1.23（1.53）	0.29（0.17）
	Oracle	2（0）	NA	4（0）	NA	0.23（0.12）
$n=400$	G-Lasso	2（0）	8.49（2.94）	4（0）	6.05（2.30）	0.39（0.13）
	G-Scad	2（0）	1.22（1.52）	4（0）	1.87（1.46）	0.17（0.10）
	G-Mcp	2（0）	0.69（1.17）	4（0）	1.27（1.27）	0.19（0.10）
	SG-Lasso	2（0）	5.22（2.77）	4（0）	5.79（2.15）	0.31（0.12）

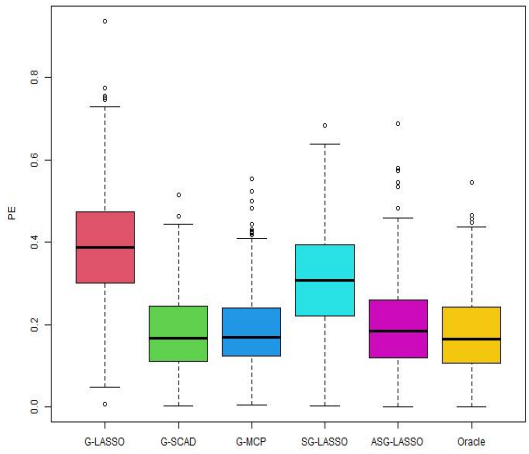
续表 4.9

样本量	方法	TG	FG	TP	FP	PE
$n=400$	ASG-Lasso	2（0）	0.70（1.23）	4（0）	1.68（2.04）	0.20（0.11）
	Oracle	2（0）	NA	4（0）	NA	0.16（0.09）

从表 4.9 可知，评价指标选择的真实组数（TG）的波动于删失比为 10%时类似，五种变量选择方法在指标选择为非零的真实非零变量数（TP）的表现与 Oracle 估计值一致，说明模型有效。在协变量 X 存在零区间时，SG-Lasso 在指标选择的零组数（FG）中表现较差，但 ASG-Lasso 选择了更少的不重要变量。由指标选择非零的真零变量数（FP）可知，ASG-Lasso 选择了最少的错误变量。预测误差（PE）箱线图见图 4.14：



（a） $n=100$ （b） $n=200$



（c） $n=400$

图 4.14　$n=100$、200、400 时预测误差（PE）箱线图

结合表 4.9 和图 4.14 可知，在删失比为 20%时预测误差（PE）分布较为均匀，预测结果均小于 1，说明了删失比适当增加时五种变量选择方法对模型预测精度影响较小。

4.2.3.3 删失比为 40%

设置删失比 40%，当 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 方法分别在样本 $n=100$、200、400 时，评价指标的均值和标准差见表 4.10：

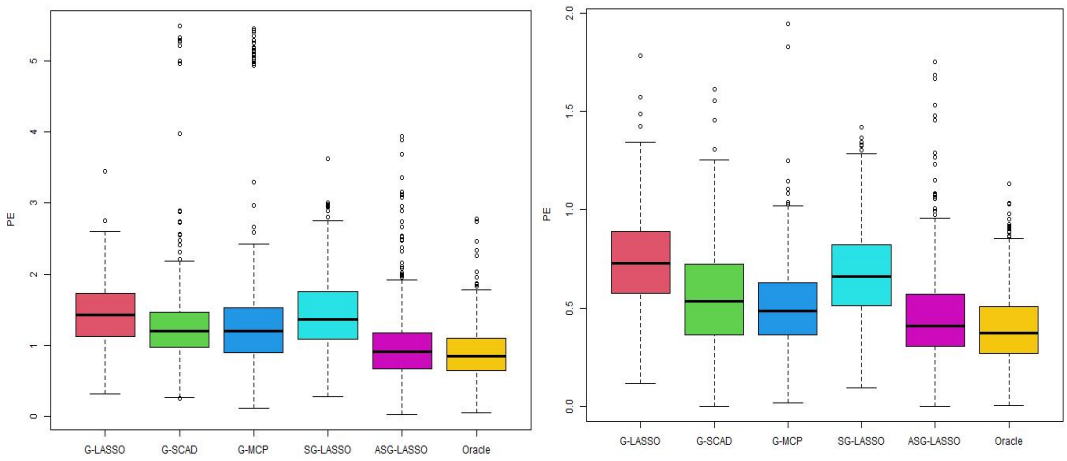
表 4.10　$n=100$、200、400 时不同变量选择方法的均值（标准差）

样本量	方法	TG	FG	TP	FP	PE
$n=100$	G-Lasso	2（0.06）	4.36（2.24）	4（0）	3.97（1.84）	1.24（0.42）
	G-Scad	1.98（0.17）	1.05（1.23）	4（0）	1.80（1.26）	1.17（1.29）
	G-Mcp	1.95（0.26）	0.53（1.14）	4（0）	1.21（1.12）	1.23（1.67）
	SG-Lasso	1.94（0.31）	2.26（1.91）	4（0）	3.10（2.18）	1.24（0.48）

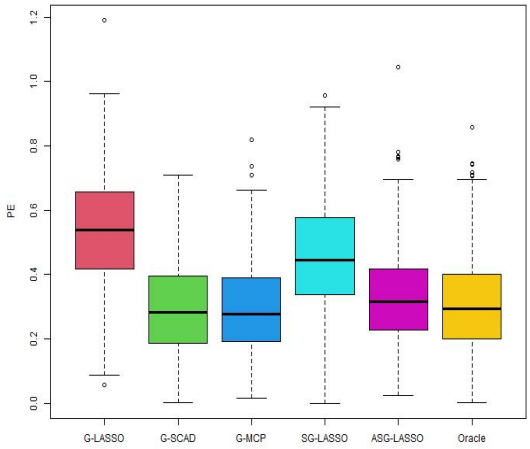
续表 4.10

样本量	方法	TG	FG	TP	FP	PE
$n=100$	ASG-Lasso	1.98（0.16）	0.42（0.93）	4（0）	0.93（1.08）	0.79（0.52）
	Oracle	2（0）	NA	4（0）	NA	0.67（0.35）
$n=200$	G-Lasso	2（0）	6.12（2.76）	4（0）	4.78（2.05）	0.75（0.24）
	G-Scad	2（0）	1.67（1.66）	4（0）	2.14（1.50）	0.55（0.27）
	G-Mcp	2（0）	0.53（0.88）	4（0）	1.10（1.10）	0.52（0.23）
	SG-Lasso	2（0）	3.83（2.50）	4（0）	5.03（2.10）	0.68（0.23）
	ASG-Lasso	2（0）	0.74（1.36）	4（0）	1.23（1.53）	0.46（0.25）
	Oracle	2（0）	NA	4（0）	NA	0.38（0.18）
$n=400$	G-Lasso	2（0）	8.49（2.94）	4（0）	6.05（2.30）	0.54（0.18）
	G-Scad	2（0）	1.22（1.52）	4（0）	1.87（1.46）	0.29（0.15）
	G-Mcp	2（0）	0.69（1.17）	4（0）	1.27（1.27）	0.31（0.15）
	SG-Lasso	2（0）	5.22（2.77）	4（0）	5.79（2.15）	0.45（0.17）
	ASG-Lasso	2（0）	0.70（1.23）	4（0）	1.68（2.04）	0.30（0.17）
	Oracle	2（0）	NA	4（0）	NA	0.28（0.14）

从表 4.10 可知，评价指标选择的真实组数（TG）在 $n=100$ 时，变量选择方法的均值和方差都与 Oracle 估计值有偏差，但随着样本量的增长这种偏差就消失了，并且由指标选择为非零的真实非零变量数（TP）可知，五种变量选择方法的均值和方差都与 Oracle 估计值一致，说明在删失比为 40%的情况下模型仍旧有效。G-Lasso、SG-Lasso 在指标选择的零组数（FG）中表现较差，选择了更多的不重要变量。预测误差（PE）箱线图见图 4.15：



（a） $n=100$ （b） $n=200$



（c） $n=400$

图 4.15 $n=100$、200、400 时预测误差（PE）箱线图

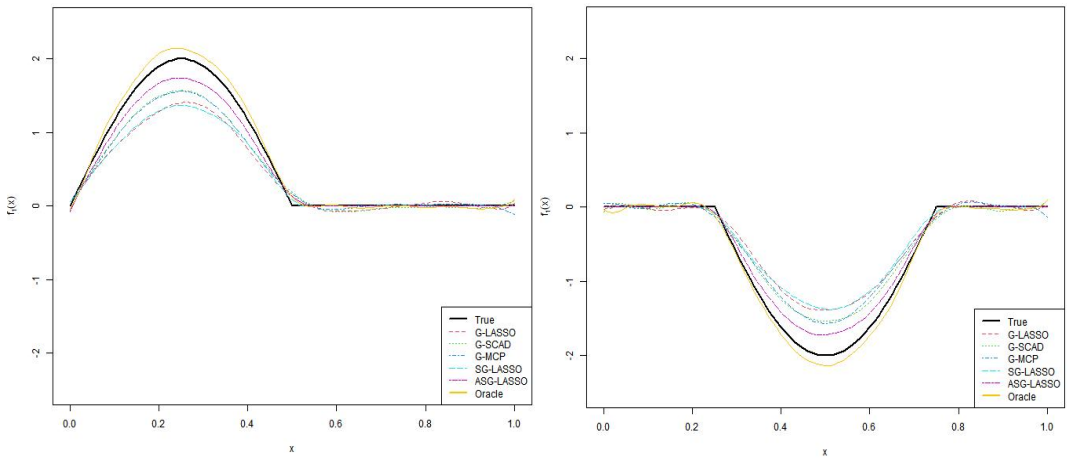
结合表 4.10 和图 4.15 可知，当 $n=100$ 时，G-Lasso、G-Scad、G-Mcp、SG-Lasso 预测误差的均值和方差均大于 1，ASG-Lasso 表现良好。表明当删失比增大为 40% 时，会出现极端值以影响模型的精度，但随着样本量的增大，预测误差（PE）逐渐减小。

4.2.3.4 不同删失比下的函数拟合图

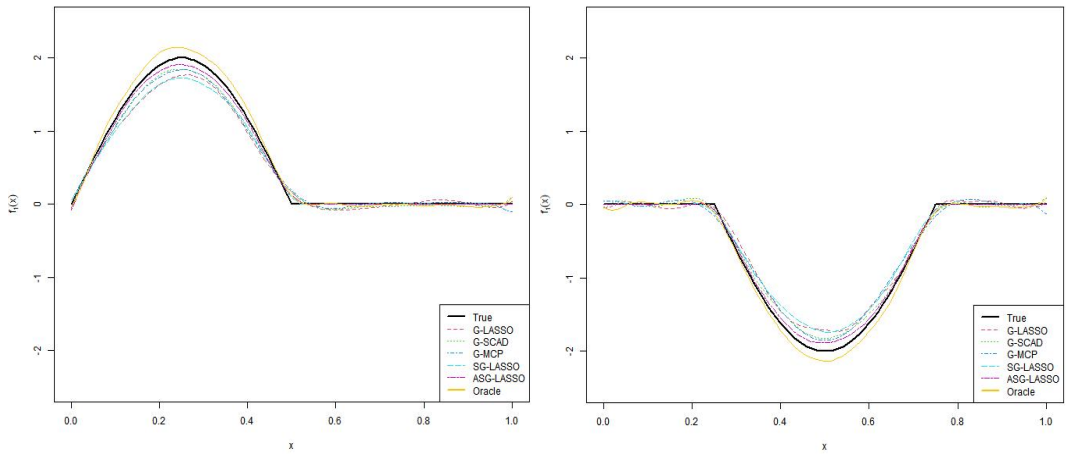
本小节给出函数样条拟合估计值，黑色实线代表真实值，其余六条曲线分别代表 G-Lasso、G-Scad、G-Mcp、SG-Lasso、ASG-Lasso 以及 Oracle 估计的平均

曲线。当协变量满足正态分布且存在零区间时，不同删失比下，样本量 $n=100$、200、400 时函数拟合见图 4.16-4.18。

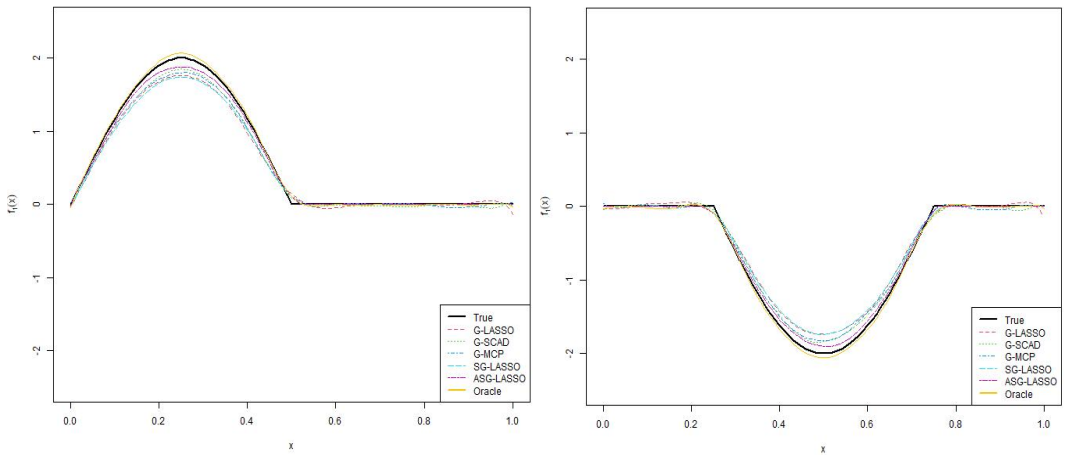
删失比为 10%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



（a） $n=100$



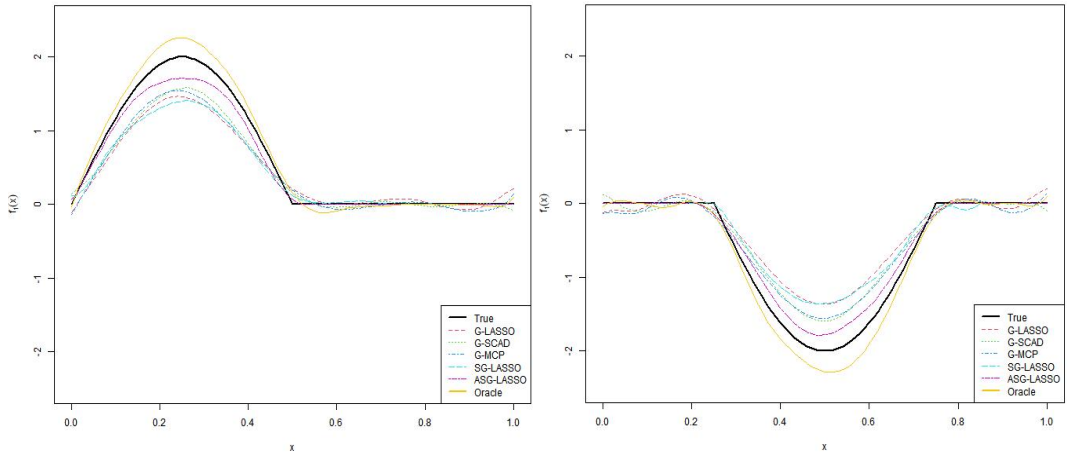
（b） $n=200$



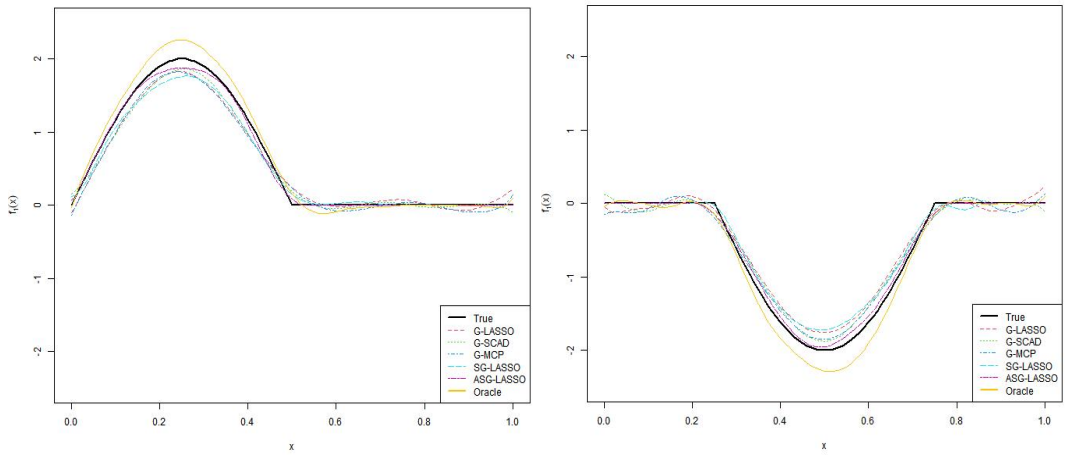
（c） $n=400$

图 4.16 $n=100$、200、400 时 $\phi_1(W_1)$、$\phi_2(W_2)$ **的函数拟合图**

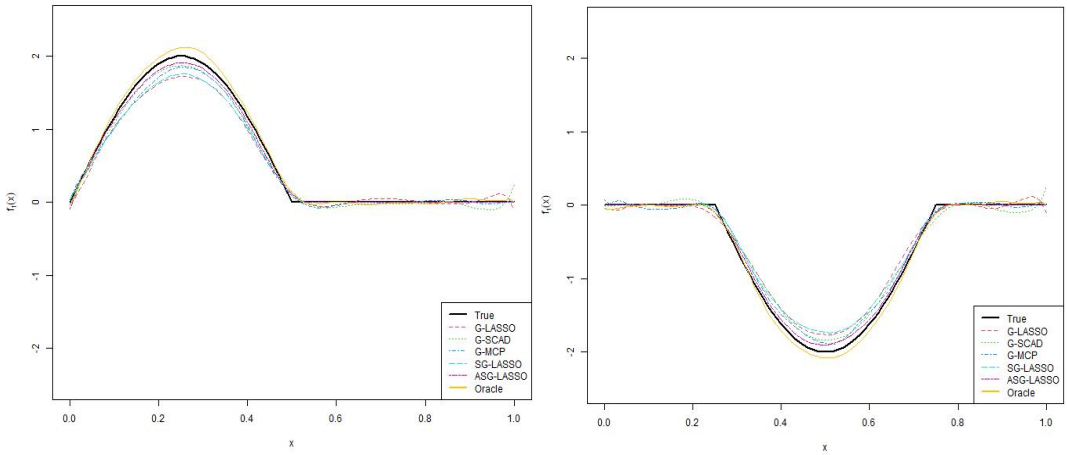
删失比为 20%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



（a） $n=100$



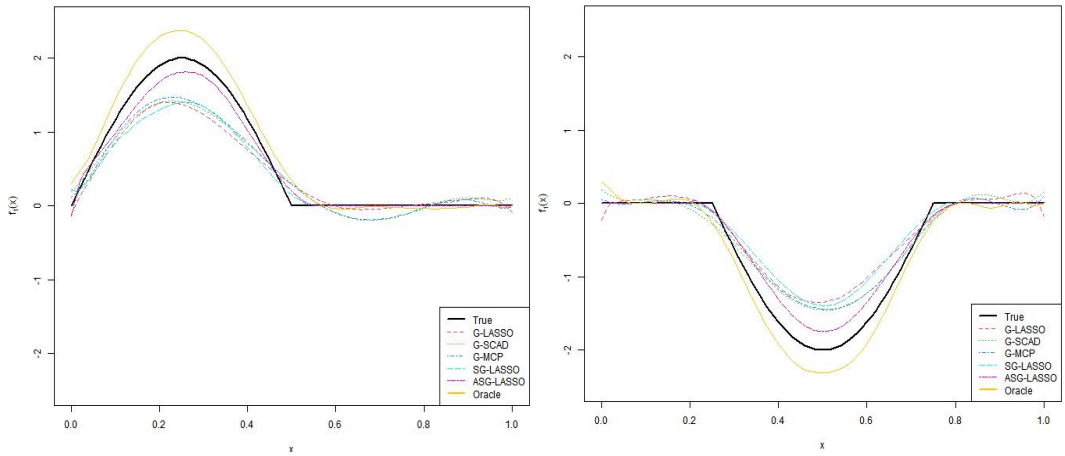
（b） $n=200$



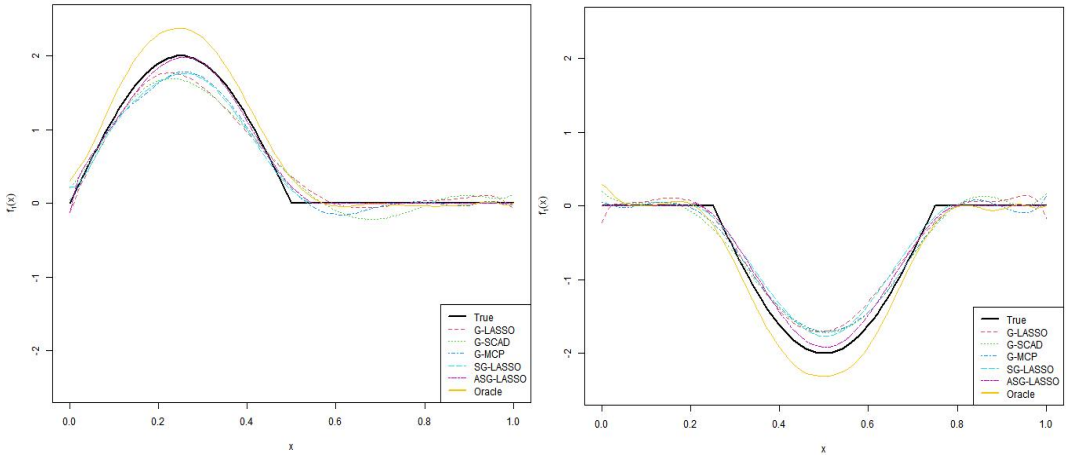
（c） $n=400$

图 4.17 $n=100$、200、400 时 $\phi_1(W_1)$、$\phi_2(W_2)$ **的函数拟合图**

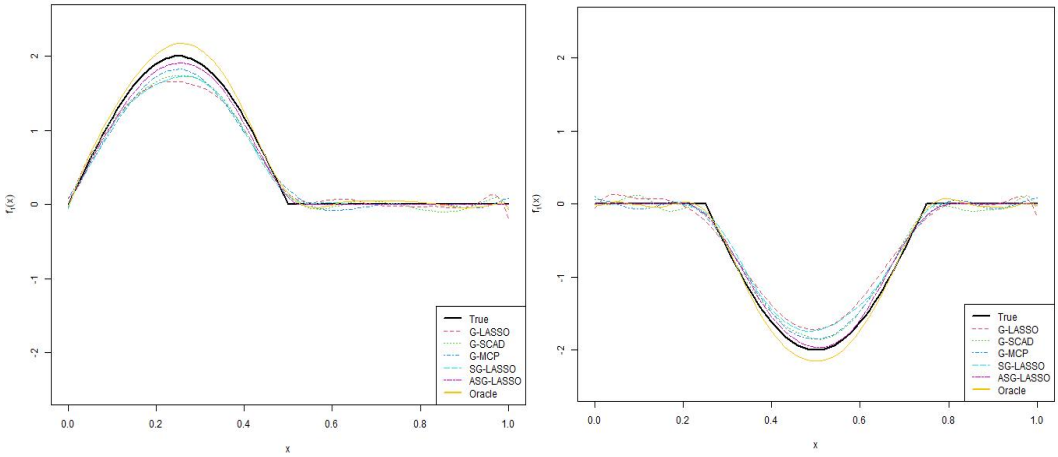
删失比为 40%时，非参数分量 $\phi_1(W_1)$ 与 $\phi_2(W_2)$ 的函数拟合图为：



(a) $n=100$



(b) $n=200$



(c) $n=400$

图 4.18 $n=100$、200、400 时 $\phi_1(W_1)$、$\phi_2(W_2)$ 的函数拟合图

4.2.4 总结

结合上述三种情况，可以得到：五种变量选择方法都能在不同样本量的情形下选择出正确的选择的真实组数（TG）与非零的真实非零变量数（TP），可以说明模型中变量选择的有效性。尽管 SG-Lasso 方法在某些情况下选择了更多的组和更不重要的变量，但这种性能在 ASG-Lasso 中得到了显著改善，在选择更多重要组和选择不重要变量的数量显著降低。因此，如果已知协变量之间存在已知的分组结构，双层变量选择方法优于单变量选择方法，并且双层变量选择方法在样本量增加的情形下明显优于组变量选择方法。综上所述，当样本量与删失比例合适时，我们的 SG-Lasso 方法在变量的误选率上比 G-SCAD，G-MCP 差一些，ASG-Lasso 方法具有 Oracle 性质，变量的正确选择以及错误选择的表现性能和 G-SCAD，G-MCP 相当，整体 PE 值要比这两种方法表现的好一些。并且在样本量少以及删失比例高时，ASG-Lasso 方法表现出更好的模型稳健性。

在样条曲线拟合方面，可以发现 SG-Lasso 要比 G-Lasso 好一些，说明双层变量选择的方法是有效的，由于 SG-Lasso 不具备 Oracle 性质，在曲线拟合的表现上要比 G-Scad 以及 G-Mcp 差一些，这与 Tian[45]的结论类似。同时在五种变量选择方法中，ASG-Lasso 表现最好，与 Afzal(2020)的结论一致，尤其协变量取值存在零区间时表现稳定。随着样本量的增大，估计效果有明显的增强，说明本文提出的估计方法具有良好的相合性。

4.3 实例分析

4.3.1 乳腺癌数据集

一直以来，癌症是人类死亡的高危因素，而乳腺癌一度成为女性患癌种类的榜首，但乳腺癌的成因并不清楚，所以研究乳腺癌患者体内特征成为预防乳腺癌疾病的重要手段。在本节中选用 gbsg 数据集。它包含了 1984-1989 年间德国乳腺癌研究小组对 720 名淋巴结阳性乳腺癌患者进行的患者记录，它包含了 686 名患者的完整数据作为预后变量。数据集包括 686 个观测值和 8 个可能影响生存结局的变量。变量介绍见表 4.11：

表 4.11 乳腺癌数据变量解释

编号	变量名	变量解释
变量1	年龄（age）	患者患病年龄
变量2	更年期（meno）	患者更年期状态
变量3	肿瘤大小（size）	乳腺肿瘤大小
变量4	肿瘤恶性等级（grade）	乳腺肿瘤恶性程度
变量5	阳性淋巴结数（nodes）	患者体内阳性淋巴结数
变量6	黄体酮（pgr）	孕酮受体的个数
变量7	雌激素（er）	患者雌激素的个数
变量8	荷尔蒙（hormon）	是否进行激素治疗

其中，变量 2、变量 4 和变量 8 为二分类变量，当做模型的参数部分，其余五个变量为连续型变量，当做模型中的非参数部分处理，所以要分别计算出参数

部分的线性估计值以及非参数部分的分量估计图。

4.3.1.1 数据预处理

首先对 8 个可能影响变量进行描述性统计，由图 4.19 初步可知：乳腺癌患者的年龄主要集中在 45-65 岁；患病时间与更年期前后关系不大；肿瘤大小集中在 20mm；肿瘤等级严重程度分为三个等级（1＜2＜3）；阳性淋巴结数集中在 0-20 个。

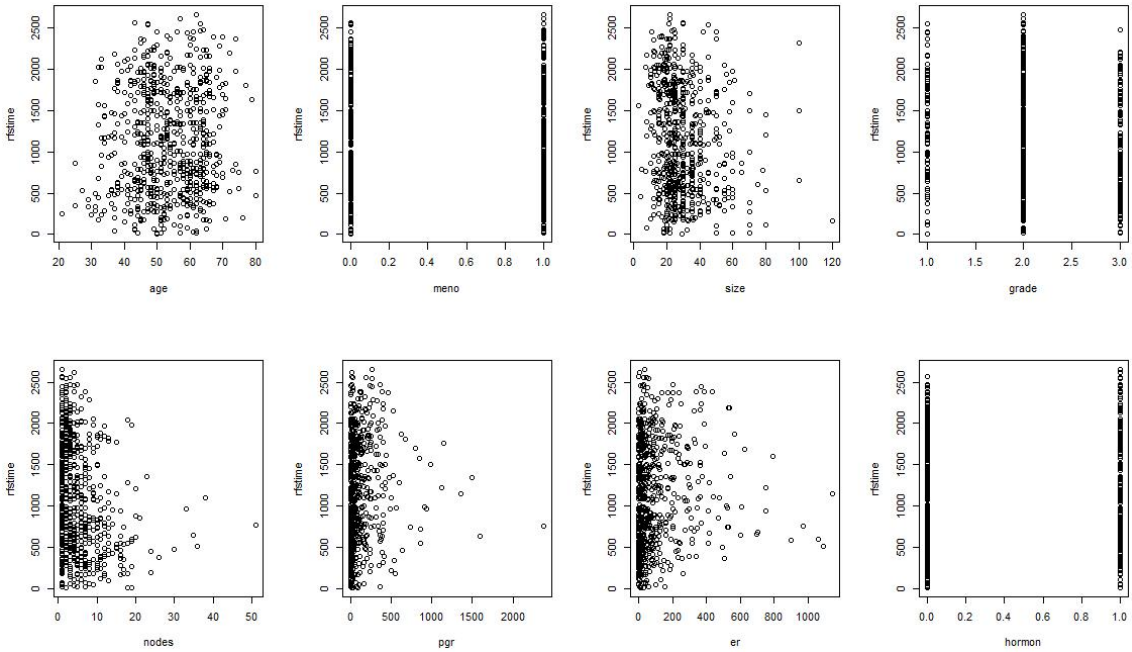


图 4.19 乳腺癌患者 8 个变量的分布散点图

4.3.1.2 模型求解

经数据处理后，研究共纳入 686 名患者，8 个可能影响乳腺癌死亡时间的因素进入模型。其中包含 3 个线性分量以及 5 个非参数分量，首先对 8 个变量进行斯皮尔曼（spearman）相关系数分析，热力图见与图 4.20：

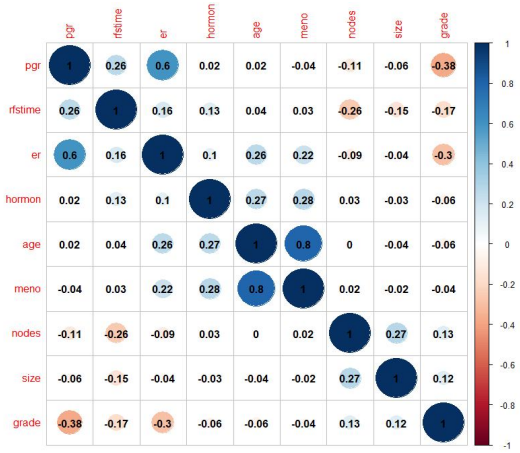


图 4.20　斯皮尔曼相关系数热力图

变量 2、变量 4、变量 8 的估计值见表 4.12，由表可知，变量 2 更年期（meno）与乳腺癌风险函数无关；变量 4 肿瘤恶性等级（grade）与风险函数呈正相关；变量 8 荷尔蒙（hormon）与风险函数呈负相关。

表 4.12　变量 2、4、8 的线性估计值

方法 / 变量	G-Lasso	G-Scad	G-Mcp	SG-Lasso	ASG-Lasso
meno	0	0	0	0	0
grade	0.19	0.25	0.26	0.24	0.25
hormon	-0.26	-0.31	-0.31	-0.29	-0.3

对变量 1、变量 3、变量 5、变量 6、变量 7 分别进行 G-Lasso、G-Scad、G-Mcp、SG-Lasso 以及 ASG-Lasso 变量选择，得到的估计图见图 4.21：

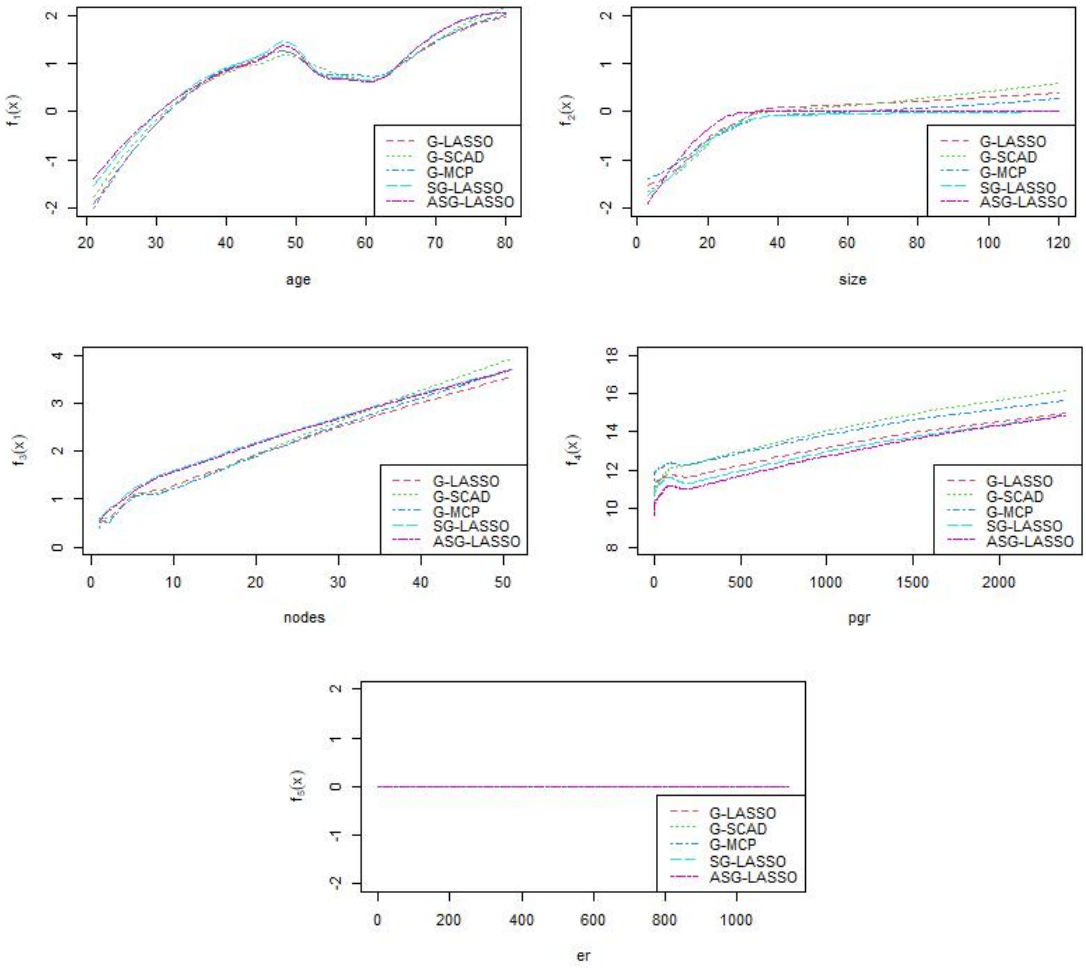


图 4.21　估计图

由图 4.21 可知，五种变量选择方法对各个变量的估计趋势一致，可以得到结论：变量 1 年龄（age）显示随着年龄的增大，患乳腺癌的风险在加速增加，和实际医学现象一致；变量 3 肿瘤大小（size）主要表现为 30mm 以下，随着肿瘤的增长，患乳腺的风险再增加，肿瘤大小在 30mm 左右时，双层变量选择方法将函数压缩为了 0 区间，认为肿瘤在 30-120mm 之间患乳腺癌风险函数是一样的，这与其他三种组变量选择方法一些细微差距；变量 5 阳性淋巴结数（nodes）与变量 6 黄体酮（pgr）与风险函数呈加速正相关关系；变量 7 雌激素（hormon）不影响患病的风险函数。

4.3.2 癌症基因数据集

在本节中选用来自“tcga”数据库的癌症病人体内基因与其存活时间的数据关系，该数据集包含了 559 名癌症病人的 396 种不同基因的数据，剔除无效变量，分析 373 个自变量与生存结局与时间的关系。分别对这些基因进行五种变量选择方法，得到相关度较大的基因。

表 4.13 不同变量选择方法筛选的基因名称

变量选择方法	变量名称
G-Lasso	EFS_2008、OS_2008、EFS_MO_2008、OS_MO_2008、EFS_censor、OS_censor、fustat.y、fustatE.y、AK2、CBX7、CIPC、CSE1L、DDX39A、DOCK7、EIF3B、ARF5_g、CCT6A_g、COX6C_g、CRIP1_g、DRAM1_g、EIF3B_g、EZH2_g、FABP5_g、SAMD8、STIL、TBCB、TMEM178A、TUBA1C（共 28 个）
G-Scad	EFS_2008、OS_2008、EFS_MO_2008、OS_MO_2008、EFS_censor、OS_censor、fustat.y、fustatE.y（共 8 个）
G-Mcp	EFS_2008、OS_2008、EFS_MO_2008、OS_MO_2008、EFS_censor、OS_censor、fustat.、fustatE.y、SAMD8（共 9 个）
SG-Lasso	LDH、EFS_2008、EFS_TIME_2008、OS_2008、OS_TIME_2008、EFS_MO_2008、OS_MO_2008、EFS_censor、OS_censor、EFS_month、OS_month、fustat.y、futime.y、fustatE.y、futimeE.y、COX6C_g、FABP5_g、TAGLN2（共 18 个）

续表 4.13

变量选择方法	变量名称
ASG-Lasso	EFS_2008、EFS_TIME_2008、OS_2008、EFS_MO_2008、OS_MO_2008、EFS_censor、OS_censor、EFS_month、fustat.y、futime.y、fustatE.y、futimeE.y*（共 12 个）

由表 4.13 可知，五种变量选择方法都能在 373 个变量中筛选出与存活时间相关度最高的变量，并且存在重复基因。结合实际，说明在生存分析中引入变量选择方法是十分有必要的，通过变量选择，在众多变量中筛选出与生存时间相关度最高的变量，既降低模型复杂度，也为癌症疾病的攻克提供了条件。

4.4 本章小结

本章分数值模拟与实例分析两部分。在数值模拟中，通过改变协变量的分布，基于 Monte Carlo 模拟对比组变量选择方法和双层变量选择方法在五类指标下的性能，验证了双层变量选择方法在部分线性可加 Cox 模型中的有效性；在实例分析中，引入乳腺癌数据集及癌症基因数据集，结果都表明双层变量选择方法筛选出的变量与存活时间相关度最高，对攻克癌症疾病有现实意义。

5 研究总结和展望

5.1 研究结论

在生存分析中，Cox 模型的产生为医学数据的处理奠定了基础，针对 Cox 模型的变量选择研究是处理删失数据的重要手段。而部分线性可加 Cox 模型在保留半参数模型优点的同时，将生存分析中随时间变化的协变量纳入模型，提升了模型的使用范围。

对部分线性可加 Cox 模型的变量选择方法的研究是十分有意义的。组变量选择方法与双层变量选择方法均能对部分线性 Cox 模型进行变量选择，但双层变量选择方法能够实现在组内和组间的变量选择。针对模型，通过 B-样条曲线拟合，实现非参数部分的样条基函数展开，后将双层变量选择方法引入模型，建立了更完善的部分线性可加 Cox 模型的变量选择过程。

通过模拟比较五种变量选择方法在模型筛选中的表现，证明了双层变量变量选择方法在部分线性可加 Cox 模型中效果最好，并通过实例分析表明研究具有现实意义。其中自适应稀疏组 Lasso（ASG-Lasso）在协变量满足正态分布、满足正态分布且存在相关系数以及满足协变量存在零区间这三种情形下对模型变量筛选都有最小的均值和方差，并且模型中的非参数部分的拟合与真实曲线趋势相同，证明了模型有效性以及双层变量选择方法的优势。在实证方面，用两个医学数据集说明模型和变量选择方法的有效性，首先通过乳腺癌数据集进行五种变量选择方法的对比，说明部分线性可加 Cox 模型能够适应实际数据。其次通过变量选择方法对癌症基因分别进行筛选，得到与生存结局相关度最高的基因，为后续癌症的攻克起到一定的现实意义。

5.2 研究展望

首先，本文对模型中的非参数部分使用 B-样条曲线进行拟合，样条回归中可用于拟合的曲线有很多，后续可以近一步考虑样条回归的拟合效果。

其次，本文主要研究了双层变量选择方法中的基于加性惩罚函数下的变量选择方法，另一种复合惩罚函数也可以完成双层变量选择，后续可考虑这类变量选择方法在模型中的应用。

最后，在本文的基础上，可以探究更多生存分析中的变量选择问题，可以考虑不同种类生存数据最适合的变量选择方法，为医学研究提供条件。

参考文献

[1] Cox DR. Regression models and life tables (with Discussion)[J]. Journal of the Royal Statistical Society:Series B, 1972, 34:187-220.

[2] Hoerl A E. Application of ridge analysis to regression problems[J]. Chemical Engineering Progress, 1962, 58:54–59.

[3] Hoerl A E, Kennard R W. Ridge Regression[J]. Applications to Non-orthogonal Problems Technometrics, 1970, 12:34-47.

[4] Tibshirani R. Regression Shrinkage and Selection via the Lasso[J]. Journal of the Royal Statistical Society:Series B, 1996, 58(1):267-288.

[5] Fan J , Li R . Variable selection via nonconcave penalized likelihood andits oracle properties[J]. Journal of the American Statistical Association, 2001, 96(456):1348-1360.

[6] Zhang. Nearly Unbiased Variable Selection under Minimax Concave Penalty[J]. The Annals of Statistics, 2010, 38(2):894-942.

[7] Yuan M, Lin Y. Model Selection and Estimation in Regression with Grouped Variables[J]. Journal of the Royal Statistical Society, 2006, 68(1):49-67.

[8] Zhao Y. Grouped and Hierarchical Model Selection through Composite Absolute Penalties[J]. The Annals of Statistics, 2009, 37:3468-3497.

[9] Blazere M, Loubes J M, Gamboa F. Oracle Inequalities for a Group Lasso Procedure Applied to Generalized Linear Models in High Dimension[J]. IEEE Transactions on Information Theory, 2014, 60(4):2303-2318.

[10]王小燕, 谢邦昌, 马双鸽, 等. 高维数据下群组变量选择的惩罚方法综述[J].

数理统计与管理, 2015, 34(6):978-988.

[11] Huang J, Ma S, Xie H, Zhang C, et al. A group bridge approach for variable selection. Biometrika, 2009, 96(2):339-355.

[12] Friedman J, Hastie T, Tibshirani R. A note on the group lasso and a sparse group lasso; 2010, 78(2):736-754.

[13] Matsui H. Sparse group lasso for multiclass functional logistic regression models[J]. Communications in Statistics -Simulation and Computation, 2018,0:1-14.

[14] Fang K G,Wang X,Zhang S, et al. Bi-level Variable Selection Via Adaptive Sparse Group Lasso, Journal of Statistical Computation and Simulation, 2014, 34(3):28-45.

[15] Fan J,Li R. Varible Selection for Cox's Proportion Hazards Model and Frailty Model[J]. Annals of Statistics, 2002, 30(1):74-99.

[16] Zhang,H H,Lu W. Adaptive lasso for cox ' s proportion hazards model[J]. Biometrika, 2007, 94:691-703.

[17] Liang H,Li R.Variable Selection for Partially Linear Models with Measurement Errors[J].Journal of the American Statistical Association,2009,104(485):234-248.

[18] Ni X, Zhang H H,Zhang, D.Automatic Model Selection for Partially Linear Models[J]. Journal of Multivariate Analysis, 2009,100(9):2100-2111.

[19] Xia X,Yang H.Variable Selection for Partially Time-varying Coefficient Error-in-variables Models[J]. Statistics, 2016,50(2):278-297.

[20] Zhao P, Xue L. Variable Selection for Semiparametric Varying Coefficient

Partially Linear Errors-in-variables Models[J]. Journal of Multivariate Analysis,2010,101(8):1872-1883.

[21].Kai B, Li R, Zou H. New Efficient Estimation and Variable Selection Methods for Semiparametric Varying-coefficient Partially Linear Models[J]. The Annals of Statistics,2011,39(1):305-332.

[22]Xie H,Huang J. Scad-penalized Regression in High-dimensional Partially Linear Models[J]. The Annals of Statistic,2009,37(2),673-696.

[23]Yang J, Lu F,Yang H. Quantile Regression for Robust Estimation and Variable Selection in Partially Linear varying-coefficient Models[M].2017.

[24]Hu Y, Lian H. Variable Selection in a Partially Linear Proportional Hazards Model With a Diverging Dimensionality[J].Statistics Probability Letters,2013,83(1):61-69.

[25]Meier L, Van de Geer S,Bühlmann, P. High-dimensional Additive Modeling. Annals of Statistics, 2009, 37:3779-3821.

[26]Xue L. Consistent variable selection in additive models[J]. Stat Sin. 2009,19:1281-1296.

[27]Fan J,Feng Y,Song R. Nonparametric Independence Screening in Sparse Ultrahigh Dimensional Additive Models[J].J Amer Stat Assoc, 2011,116:544-557.

[28]Lemler S. Oracle inequalities for the Lasso for the conditional hazard rate in a high-dimensional setting[J]. arXiv preprint arXiv, 2012,1206.5628.

[29]Lv S H, Jiang J K, Zhou F Y, et al. Estimating high-dimensional additive Cox model with time-dependent covariate processes[J]. Scandinavian Journal of

Statistics, 2018:1-23.

[30] Zhang S,Wang L,Lian H. Estimation by Polynomial Spline with Variable Selection in Additive Cox Models[J]. Statistics, 2014 ,48:67-80.

[31] Lin H,He Y,and Huang J. A global partial likelihood estimation in the additive Cox proportional hazards model[J]. J Stat Plan Inference, 2016,169:71-87.

[32] Wu Q, Zhao H, Zhu L, et al. Variable selection for high-dimensional partly linear additive Cox model with application to Alzheimer's disease. Statistics in Medicine, 2020:3120-3134.

[33] Huang, J. Efficient estimation of the partly linear additive Cox model[J]. Annals of Statistics, 1999, 27:1536-1563.

[34] Lu M, Lu T, Li C. Efficient estimation of partially linear additive Cox model under monotonicity constraint[J]. Journal of Statistical Planning and Inference,2018,192:18-34.

[35] 彭非,王伟.生存分析[M].北京:中国人民大学出版社:2004:25-35.

[36] Kleinbaum D G. Survival Analysis, a Self-Learning Text[J]. Biometrical Journal: Journalof Mathematical Methods in Biosciences, 1998, 40(1):107-108.

[37] 王星.非参数统计[M].北京:电子工业出版社:2020.11,257-260.

[38] Ruppret D.Semiparametric Regression[M]. Cambridge University Press,Cambridge,2003.

[39] Huang J, Breheny P, Ma S. A Selective Review of Group Selection in High-Dimensional Models[J]. Statistical Science, 2012, 27(4):481-499.

[40] Simon N, Friedman J, HastieT, et al.A Sparse-Group Lasso[J]. Journal of

Computational and Graphical Statistics, 2013,22(2):231-245.

[41] Li J T,Liang K,Song X. Logistic regression with adaptive sparse group lasso penalty and its application in acute leukemia diagnosis[J]. Computers in Biology and Medicine, 2022, 141:105154.

[42] Meier L, van de Geer S,Bühlmann P.The Group Lasso for Logistic Regression. Journal of the Royal Statistical Society[J].Statistical Methodology,2008,70:53-71.

[43] Aydin D,Yilmaz E.Modified Spline Regression Based on Randomly Right-censored Data[J].Journal of Statistical Computation and Simulation, 2018, 88(8):2587-2611.

[44] Afzal A R, Lu X W.Variable Selection in Partially Linear Proportional Hazards Model with Grouped Covariates and a Diverging Number of Parameters[J].Biomedical Research,2020:411-448.

[45] Tian T,Sun J. Variable selection for nonparametric additive Cox model with interval-censored data[J].Biomedical Journal,2023,65:2011301.

致谢

转眼间三年的研究生生涯就要结束了，回首过往，初来学校的点滴浮上心头。初次离开家乡来到 600 公里以外的陌生之地求学，心中郁闷难以自愈，但好在这种情感随着时间慢慢减弱。再回首，心中尽是不舍。

首先，感谢我的老师，他严谨的教学态度始终鼓舞着我，从论文选题到形成终稿离不开老师对我的指导。师者，传道授业解惑也，三年来每周的讨论班他从不缺席，为我们营造良好的学习习惯，认真务实的形象我将铭记于心。

其次，感谢我的好友，漫漫人生路离不开知己相伴。虽隔千里但仿若近在咫尺，无数次难熬的日子多亏了她的陪伴。在此衷心的愿你万事顺遂、健康长乐，也愿我们友谊长存。感谢师门的师兄师姐师弟师妹们，对我的学习生活给予了莫大的帮助；感谢统计学院相识的伙伴们，为我留下美好的回忆；感谢我的妹妹，她是最可爱的存在，每次看见她仿佛世间的阴霾都可消散，感谢有你。

最后，最要感谢的是我的父母，感谢他们的生养之恩；感谢他们给了我无限的包容与宠爱；感谢他们成为我最坚强的后盾。父母在，不远游，游必有方。女儿求学之路即将结束，归期已至，再无分别之日。愿你们生活吉祥如意，幸福相伴到老！

行文至此，心中颇有感慨。唯愿山水有相逢，未来皆可期，我们来日再见。