【经济学学术前沿书系】

DIMENSION REDUCTION, VARIABLE SELECTION AND THEIR APPLICATIONS

降维、变量选择技术及其应用

戴鹏杰◎著

中国人民大学
RENMIN UNIVERSITY OF CHINA
商学院
SCHOOL OF BUSINESS

大数据分析时代，如何能快速利用大量传统的统计模型？
分析维度的增长，是否还会成为数据分析的“维数诅咒”？
新的数据决策时代，现有方法能否满足分析精度与速度的需求？

经济日报出版社

【经济学学术前沿书系】

降维、变量选择技术及其应用

戴鹏杰◎著

经济日报出版社

图书在版编目（CIP）数据

降维、变量选择技术及其应用／戴鹏杰著. -- 北京：经济日报出版社，2016.7

ISBN 978-7-80257-976-7

Ⅰ. ①降… Ⅱ. ①戴… Ⅲ. ①统计数据-统计分析 Ⅳ. ①O212.1

中国版本图书馆 CIP 数据核字（2016）第 149211 号

降维、变量选择技术及其应用

作　　者	戴鹏杰
责任编辑	陈礼艳
出版发行	经济日报出版社
地　　址	北京市西城区白纸坊东街 2 号（邮政编码：100054）
电　　话	010-63567683（编辑部）
	010-63588446　63567692（发行部）
网　　址	www.edpbook.com.cn
E-mail	edpbook@126.com
经　　销	全国新华书店
印　　刷	北京九州迅驰传媒文化有限公司
开　　本	710×1000 毫米　1/16
印　　张	8.5
字　　数	160 千字
版　　次	2016 年 7 月第一版
印　　次	2016 年 7 月第一次印刷
书　　号	ISBN 978-7-80257-976-7
定　　价	42.00 元

大数据分析时代，如何能快速利用大量传统的统计模型？

分析维度的增长，是否还会成为数据分析的“维数诅咒”？

新的数据决策时代，现有方法能否满足分析精度与速度的需求？

序　言

Hadoop 在 2006 年提出了“大数据”（Big - Data）的概念后，热度不断上升，在经历了几年的狂热与炒作之后，慢慢归于沉寂、慢慢归于实用。大数据在商业是属于后台的内容，前端只需要看到简单的结论与结果，那么只要不是噱头与空壳子，就需要有模型与算法将庞大的数据变成神奇的结果。而在这个变化的过程中，不但要求正确性、科学性，也要求速度，但传统的统计模型与方法并不能在这样的数据容量与计算速度下胜任，因此很多优秀的统计方法并不能在大数据的环境得到应用。虽然“摩尔定律”告诉我们，每 18 个月，机器的性能提升一倍且成本下降一半，但大数据的魔咒在于，可能数据的增长速度会超过摩尔定律的速度，所以我们不能把希望寄托在机器性能提升上，应该更多地关注于算法与模型的改进上。而算法模型的改进过程中，主要困难在于大数据的数据存储量与单个数据记录的巨大维度，前者主要由数值计算学科来完成，而后者需要统计学的相关工具来解决。而降维技术与变量选择技术是解决数据维度过大的主要方法。

降维的思想是希望通过较少的新变量来代替原有较多的变量，并且尽可能多地保持信息的完整性。关于降维的研究历史是较为漫长的，但按其研究的对象不同，可以分为两个大的阶段。第一阶段是专注于将高维度的变量空间投影到低维度，同时保持信息不失（或大部分信息保留），代表方法是 PCA（Principal Component Analysis），主成分分析法；第二阶段是跳出变量空间本身，讨论在回归分析中协变量空间对于响应变量有效的信息保留，代表方法是 SIR（Sliced Inverse Regression）切片逆回归法。就降维效果而言，由于后者的目标是尽可能保留协变量空间对响应变量的有效信息而不是所有的信息，所以后者更好。目前来说，后者的应用范围也较广，而且国际相关研究领域的主流也是讨论后者。

变量选择技术是选择模型中显著重要的变量，从而减少变量维度的一种方法。其研究的历史也比较早，从早期研究线性回归模型，常用的 AIC、BIC 等方法，到近二十年的成熟方法 LASSO、SCAD、自适应 LASSO 等方法，变量选择的

结果越来越精确，效果越来越稳定。同时，近几年的变量选择更注重于无模型假设或非参数模型，更加适应在数据的原模型未知情况下计算与分析。

这两种技术在大数据计算中，起到越来越重要的作用。当大数据的维度从我们常见的几维到十几维，快速增长到几百维甚至数万、数十万维（例如医学基因研究）时，降维技术与变量选择技术将被广泛应用到各种传统统计模型、经济模型，并在建立模型的初期与中期起到举足轻重的效果。

近几年中，这两大领域的研究成果不断涌现，最新的论文成果逐年增加，一直是统计学科与经济学科较为热门的话题。但由于学科发展迅速，目前尚没有多少对于这两项技术进行系统性阐述与前沿介绍的书籍，而作者也是看到了现状，希望能填补部分空白。由于作者水平有限，书中不免有诸多错误，万望读者能不吝指正，作者感激不尽。

目　录

第一章　降维、变量选择技术概述

“降维”（Dimension Reduction）与“变量选择”（Variable Selection）是近几年统计学界比较热门的话题，非常多的统计工作者投入其中。这些话题受到追捧的原因是，这些问题有着比较深厚的实际需求背景。

高维度的数据分析已经变得越来越频繁，同时在许多科学领域中的地位也变得越来越重要，比如工程学、医学、财经学、机器学习等。但其也引起了目前统计方面的许多问题。例如，在疾病分类中，我们使用微阵列（microarray）数据，则有数以万计的分子表现成为可能的协变量，即便是一个简单的金融贷款信贷资质问题研究，也会包含数百个协变量。高维数据的种类也比较多，通篇论文中，我们主要处理样本量高于协变量维度情况下的高维问题。

我们通常遇到的问题，一般为研究一个标量的响应变量以及众多与之相关的协变量的回归关系，但协变量的个数并不是越多越好，过多的变量使得最终的模型变得不稳定，或者很难解释与应用。同时，由大量的实践经验表明，对于某一被研究对象，主要的影响因素不会很多，因此，我们作一个假设，我们研究的回归关系只依赖于少数的维度或变量，由此则我们高维问题便有了解决的可能，而如何有效而快速地找到这些少数的显著影响因素就变得十分有意义，“降维”与“变量选择”是这个领域的两大方向。

1.1 降维

降维的历史其实比较久远，我们所熟悉的主成分分析，便是其中一种。主成分分析是将多个变量通过线性变换以选出较少个数的重要变量，又称主分量分析。在很多情形，变量之间是有一定的相关关系的，当两个变量之间有一定相关关系时，可以解释为这两个变量反映此协变量的信息有一定的重叠。主成分分析是对于原先提出的所有变量，建立尽可能少的新变量，使得这些新变量是两两不相关的，而且这些新变量在反映原协变量的信息方面尽可能保持

原有的信息。用空间的思想来看，主成分分析就是希望找到协变量 X 张成空间中的一组基，而基的个数希望低于协变量本身的个数，以此达到降低协变量维度的目的。

主成分分析的降维程度是远远不够的，我们的研究中，经常遇到的形式是一个标量的响应变量 Y 与多个变量组成的协变量 X（设维度为 p）。我们关心的是，两者之间的关系，或者说想要到 Y 关于 X 的模型结构。

那么对于协变量，我需要的是 X 能够给 Y 提供信息的部分，而不是全部的 X。而主成分分析中，只考虑 X 空间本身的冗余，并不考虑 X 对响应变量 Y 不提供信息的那部分冗余。因此，我们可以根据这个思想，对降维问题作进一步的思考与讨论。

1.1.1 充分降维中心子空间

考虑回归关系

$$Y = f(X_1, \cdots, X_p, \varepsilon)$$

其中，误差项 ε 与 Y 并不一定是加和关系。

定义 1.1：如果存在协变量的一个子空间 S，使得下式成立

$$Y \perp X \mid P_S X$$

则我们称 S 为 Y 关于 X 的一个充分降维子空间。其中 P_S 为普通的投影算子（内积计算法则），“ $\perp$ ”表示两者相互独立。

由上述定义，我们可以知道，$P_S X$ 提供了 X 所有可能提供给响应变量 Y 的信息，因此，我们称这种为“充分降维”（Sufficient Dimension Reduction）。事实上，这样的子空间是很多的，而 X 本身也是其中之一，而我们需要的是其中维度最小的一个，于是我们有了如下定义。

定义 1.2：记所有如定义 1.1 中的降维子空间的交集，为 $S_{Y|X}$，假设其唯一存在，则称为充分降维中心子空间（Central Subspace）。

假设充分降维中心子空间的维度 $\dim(S_{Y|X}) = K$，则我们取 $\vartheta_1, \cdots, \vartheta_K$ 为空间 $S_{Y|X}$ 的一组基。记 $\vartheta = (\vartheta_1, \cdots, \vartheta_K)^T$，则我们将 X 降维到 $\vartheta^T X$，并且没有回归信息损失。关于充分降维中心子空间的存在性，可以参见 Cook（1998）。

定义协变量的标准化形态为 $Z = \sum^{-1/2}(X - E(X))$，其中 $\Sigma = cov(X) > 0$，则充分降维中心子空间 $S_{Y|X} = \Sigma^{-1/2} S_{Y|Z}$。因此，我们将假设协变量 X 已经经过标准化。

我们考虑另一种降维空间。由于在实际研究中，可以出现我们只关心 $E(Y \mid X)$ 的情况，则我们定义这类降维空间，

定义 1.3：如果存在协变量的一个子空间 T，使得下式成立

$$E(Y \mid X) \perp X \mid P_T X$$

则我们称 T 为 Y 关于 X 的一个充分均值降维子空间。其中 $P_{(\cdot)}$ 为普通的投影算子（内积计算法则），“ $\perp$ ”表示两者相互独立。而所有 T 的交集，我们记为 $S_{E(Y\mid X)\mid X}$，称为充分均值降维子空间（Central Mean Subspace）。

很明显，子空间 $S_{E(Y\mid X)\mid X}$ 包含于子空间 $S_{Y\mid X}$，具体的情况可以参见 Cook 与 Li（2002）。

1.1.2 常见的降维方法

估计充分降维中心子空间（CS）的方法很多，最早的是在 Li（1991）中提出的切片逆回归（Sliced Inverse Regression，简称 SIR）。

最初的思想是很自然的。我们正常的研究思路是对 Y 关于 X 的回归，由于协变量的维度比较高，自然出现了高维问题，但如果反过来看，用 X 关于 Y 进行回归，即把响应变量与协变量的位置互换过来，则因为 Y 是一维的，当然不会出现高维问题。同时，我们也可以想象，如果某一维的 X 对 Y 有贡献，则反之，Y 对这一维的 X 也有贡献。但问题出现了，逆回归是避免了高维问题，但与原问题有什么关系呢？

Li（1991）中最重要的一点就是证明了在一定条件下，逆回归曲线 $E(X \mid y)$ 包含于充分降维中心子空间（Li（1991）中的定理 3.1）。由此，我们知道张成的空间 $Span(M_{SIR}) \in S_{Y\mid X}$，其中 $M_{SIR} = cov(E[X \mid Y])$，$Span(\cdot)$ 表示张成的空间。则从 M_{SIR} 的估计 $\widehat{M}_{SIR}$，可以得到其特征向量的的相合估计，如果充分降维子空间的维数为 K，则其中前 K 个非零特征根对应的特征向量即为充分降维子空间的一组基的估计。

而 SIR 方法计算过程中的核心就是 M_{SIR}，作者将响应变量 Y 的取值范围划分为 H 片，分别记为 $I_1,\cdots,I_H$，分别计算每个区域内的 Y 概率与 X 的均值，得结果如下

$$M_{SIR} = \sum_{h=1}^{H} p_h m_h m_h^T$$

其中

$$p_h = P(Y \in I_h)$$

$$m_h = E(X \mid Y \in I_h)$$

过程中的切片数量 H 对于大样本性质影响不大。当每个切片以内只含两个数据点时，HSing 和 Carrol（1992）证明了矩阵 M_{SIR} 是 $\sqrt{n}$ 的速度相合。Zhu 和 Ng

(1995) 推广了这个结果，证明了当片内数据点可以为 2 ~ $n/2$ 的任意值，M_{SIR} 总是 $\sqrt{n}$ 相合的。

需要注意的是，SIR 方法并不能保证把所有的降维方向（也就是充分降维子空间的一组基）都找到，计算时也存在着一些问题。而许多后来的统计学者也对 SIR 进行了改进，例如 Hsing 与 Carroll (1992)，Zhu 与 Ng (1995)，Zhu 与 Fang (1996) 对矩阵 M_{SIR} 的计算作了更深入的研究。

自从 Li (1991) 的文章在 JASA 发表之后，降维的思想开始兴起，并不断有新的方法问世。同年，Cook 与 Weisberg (1991)，提出了切片平均方差估计 (Sliced Average VarianceEstimation，简称 SAVE)，它与 SIR 方法是比较相像的，只是矩阵的定义有点区别

$$M_{SAVE} = E[I - cov(X \mid Y)^2]$$

其中 I 是 p 维的单位阵。对该矩阵的估计进行谱分解，前 K 个非零特征根对应的特征向量即为降维方向。

它是对 SIR 的一种改进与补充，能找到一些 SIR 方法找不到的降维方向，但自身也有一些缺陷，也不能保证能够找到所有的降维方向。

之后，Li (1992) 在 JASA 又发表了一篇，提出了基本 Hessian 方向 (PrincipalHessianDirections，简称 PHD)。理论工作基本同出一辙，但主要利用 Hessian 阵，考量非线性部分的降维方向，矩阵的定义也变化为

$$M_{PHD} = \Sigma_{yxx}$$

其中 $\Sigma_{yxx} = E[(Y - E(Y))XX^T]$。对该矩阵的估计进行谱分解，前 K 个非零特征根对应的特征向量即为降维方向。此项工作经由 Cook (1998) 进一步发展完善。

十年之后，Cook 与 Li (2002)，提出了迭代 Hessian 变换法 (Iterative Hessian Transformation，简称 IHT)，理论思想一脉相承，核心的矩阵为

$$M_{IHT} = \widehat{\Sigma}_{yxx}$$

其中 $\widehat{\Sigma}_{yxx} = (\beta_{yx}, \cdots, \sum_{yxx}^{p-1} \beta_{yx}), \beta_{yx} = E[YX]$。

此处，还有许多其它的降维中心子空间基的估计方法。Cook (1994) 与 Cook (1998) 均提出了图形回归法 (graphical regression)，Bura&Cook (2001) 提出了参数逆回归方法 (parametric inverse regression)，Chiaromonte，Cook&Li (2002) 提出了在协变量为分类变量时的偏 SIR 方法 (partial SIR)，Yin&Cook (2002) 提出了用协方差估计 K 阶中心子空间的方法，Fung etal.，(2002) 提出了经典相关系数估计 (Canonical Correlation Estimation)，Li，Zha&Chiaromonte (2005) 提出了 contour 回归法 (contour regression)，Li&Wang (2007) 中提出了方向性回归

(directional regression), Cook&Forzani (2009) 提出了似然方向方法 (likelihood acquired directions), 等等。

"降维"第一个问题自然是如何找到降维方向, 而第二个问题是如何估计降维空间的维度。在实际应用中, 我们不可能事先知道充分降维中心子空间的维数究竟是多少, 因此, 估计这一维数, 也是重要工作之一。几乎每一篇关于降维的文章都会涉及一些维数的估计问题, 但目前来看, 主要分为两大类, 一类是序列检验方法 (sequential test), 该方法是由 Li (1991) 首先提出的, 其中令 $\bar{\lambda}_{p-K}$ 为最小 $p-K$ 特征根平均值的渐近分布, 文章中证明, 在协变量 X 是正态分布时,

$$n(p-K)\,\bar{\lambda}_{p-K} \sim \chi^2_{(p-K)(H-K-1)}$$

其中 $\chi^2_{(p-K)(H-K-1)}$ 指自由度为 $(p-K)(H-K-1)$ 的卡方分布, H 为 SIR 中的切片数量。

由此, 依次对特征根作检验, 最终找到维数的估计。后经由 Schott (1994), Velilla (1998), Bura 与 Cook (2001) 以及 Ferr'e (1998) 发展完善。

另一类是对于矩阵 $cov(E[X \mid Y])$ 谱分解后特征根的 BIC 类型判断准则来估计, 此类方法是由 Zhu, Miao&Peng (2006) 的文章中首先提出, 在 Zhu et al. (2009) 中完善。具体形式如下:

$$\hat{K} = \arg\max_{k=1,\cdots,p}\left\{\frac{N}{2}\times\frac{\sum_{i=1}^{k}[\log(\hat{\lambda}_i+1)-\hat{\lambda}_i]}{\sum_{i=1}^{p}[\log(\hat{\lambda}_i+1)-\hat{\lambda}_i]} - 2\times C_N\times\frac{k(k+1)}{2p}\right\}$$

其中, C_N 为一与 N 有关的系数, $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ 为估计矩阵 M_{SIR} 谱分解的特征根。

值得一提的是, 充分降维的研究依然还有可以提高之处, 例如充分降维中心子空间基的搜索, 许多方法都不能保证是够穷尽所有降维方向等。因此, 许多的统计工作者依然在不断地探索前进。

1.2 变量选择

变量选择是另一种处理高维问题的方法, 主要通过一定的挑选标准, 从众多的协变量中选取部分变量, 从而达到降低协变量维度的目的。"充分降维"与"变量选择"谁好谁劣? 标准不同, 结论不同, 仁者见仁, 智者见智。"充分降维"是在回归信息不损失的前提下, 用尽量少的协变量线性组合来替代原协变量, 涉及到的协变量个数可能并没有减少, 并且这些新变量很难附于实际含义,

可解释性比较差，主要可以用来处理非参数分析中遇到“维数祸根”问题；而变量选择则是在一定判定准则下，挑选出了部分变量作为子集，认为其已经包含了回归问题中的全部或者大部分信息，可解释性很强，可作为建立模型时的重要参考，但选入变量个数应该大于或等于“充分降维”中的线性组合个数，故不适合用于非参数分析。

变量选择问题的出发点可能有两个：

1. 预测准确性：最小二乘法经常会出现模型的偏差（Bias）比较小而方差比较大的情况，而其模型的预测性，有时可以通过把某些系数改为零而得到改进。这样，我们牺牲一点偏差，而使方差有比较大的改善，从而使模型的整体预测性得到提高。

2. 可解释性：通常我们建立模型，更希望能选用大量预测变量中能起到决定性作用的一小部分来完成建模，这样虽然减低了一些预测精度，但可以使我们把目光放在模型的大局上。

因此，我们需要变量选择来抽取大量协变量中的显著部分。出于理论上的考虑，我们假设，对响应变量产生作用的是少数变量（我们称为显著变量）。

变量选择问题历史悠久，除去贝叶斯类的方法，大体可以分为两个时期，第一个时期被称为子集选择（Subset Selection）：主要为在参数模型下，以最二小乘法（Least Square）的残差最小为标准，并辅以选入变量个数相关的惩罚系数，搜索寻找一定意义下的协变量最优子集；第二时期被称为收缩算法时期（shrinkage method）：主要在线性模型下，对协变量的系数大小加惩罚项，将比较小的系数收缩到0，从而达到挑选变量的目的。近几年也出现了在不设定模型下，用收缩算法达到变量选择的方法。

下面我们按这两大类分别作简单介绍。

1.2.1 子集选择法

如前文中所说，子集选择法是以最二小乘法的残差辅以选入变量个数的惩罚函数为判定准则，而这里有两个方面可以有不同的标准，第一：惩罚函数可以不同，于是有了AIC、BIC、TIC、RIC等准则，一般的统计书中都可以找到各自的惩罚函数，这里不再重复；第二：如何搜索这个协变量的子集。

我们所熟知的，一般有最优子集法（Best Subset Selection）、向前选择法（Forward Stepwise Selection）、向后选择法（Backward Stepwise Selection）、向前阶段回归法（Forward Stagewise Regression）。

最优子集法需要对每一个协变量的子集都进行一次搜索计算，再根据一定的

准则挑定最优子集，可以找到全局最优的结果，但主要问题是计算量过大（计算阶为 2^p），对于高维问题并不适用，且选出的子集稳定性较差。向前选择法是从不选入变量的状态开始，先选入一个使回归残差最小的变量，再选入一个，使得与已有子集一起作回归的残差最小，并依次进行，直到在某准则下收敛，相比最优子集法，计算的强度下降并且最终结果子集比较稳定，但其为一种贪婪算法，不一定能找到全局最优解。向后选择法与之相似，只是过程相反，先假设全部变量已经选入，再去除一个变量，使得剩余的子集的回归残差最小。而向前阶段回归法与前两种方法不太一样，虽然也是先假设没有变量选入，再选入一个使得回归残差最小的，到选择第二个变量时，则是依照新变量与前一次回归的残差作回归，得到的新残差最小为标准，选入第二个变量，依次进行，得到最终子集，这同样是一种贪婪算法，不一定能找到全局最优解。

以上方法，一直被广泛应用，虽然存在着许多问题，但胜在简单易用。

1.2.2 收缩算法

子集选择法，均为通过某一种步骤过程来找到最优子集。为了能自动地并且一次性地选择重要变量，许多收缩算法被提出，算法一般都能一次性处理大量的协变量，而不必担心维数过大时计算量过大。我们主要介绍几种比较有名的方法，且不作深入讨论。

桥回归法（bridge regression）在 Frank 与 Friedman（1993）被提出。算法的核心为下式：

$$\widehat{\beta}^{bridge} = arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$

其中 λ 是一个控制系数。上式有一个等价的写法

$$\widehat{\beta}^{bridge} = arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

$$subjectto\sum_{j=1}^{p}\beta_j^2 \leq t$$

其中，λ 与 t 有一一对应关系，但并不相等。这个写法，含义更加清楚，即为在限制条件求目标函数极小值。请注意，此方法对于变量选择起的作用很小，几乎不能剔除不显著的变量。

最小绝对收缩选择算子（least absolute shrinkage and selection operator，**简称**（LASSO））在 Tibshirani（1996，1997）中被提出，其算法核心为下式：

$$\widehat{\beta}^{lasso} = arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} + \lambda\sum_{j=1}^{p}|\beta_j|$$

LASSO 方法并不能保证变量选择的相合性（即为在大样本的情况下，显著的变量都选入，不显著的变量都不选入）。

自适应的 LASSO（Adaptive LASSO）在 Zou（2006）中被提出，其算法核心为下式：

$$\widehat{\beta}^{lasso} = arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} + \lambda\sum_{j=1}^{p}\widehat{\omega}_j|\beta_j|$$

其中 $\widehat{\omega}_j = 1/|\widehat{\beta}_j|$，$\widehat{\beta}_j$ 为 β_j 的某一个 $\sqrt{n}$ 相合估计。

平滑夹取绝对导数惩罚（smoothly clipped absolute deviation penalty，**简称**（SCAD））在 Fan（2001）中被提出，其算法核心为下式：

$$\widehat{\beta}^{lasso} = arg\min_{\beta}\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\} + n\sum_{j=1}^{p} P_\lambda(\beta_j)$$

$$P_\lambda(\beta_j) = \lambda^2 - (|\beta| - \lambda)^2 I(|\beta| < \lambda)$$

其中 λ 为某一常数。

从以上的数个收缩算法来看，我们注意到，核心思想是一致，只是惩罚函数取得不同。那么我们用什么来评价惩罚函数取的优劣呢？在 Fan（2001）中提出了可能的三条标准。

1. 无偏性：系数估计是无偏的。

2. 稀疏性：能有效地减少系数估计中的非零个数，达到变量选择的目的。

3. 连续性：对于数据的微小扰动能保持系数估计的稳定性。

同时，他也提出了，好的惩罚函数能使系数估计有神性（Oracle 性质），意思为，由于在变量选择前，我们是不知道哪些变量能入选的，即不知道真实模型是哪些变量组成的，而某些惩罚函数可以使得线性模型系数估计的渐近方差，与事前知道了真实模型而得到的系数估计的渐近方差一样。而以上四种惩罚函数只有 SCAD 与 adative – lasso 是有这种性质的。

注意，以上所有的收缩算法均为在线性模型的假设下，无模型设定的变量选择问题将在后续章节中仔细介绍。