GS妯″瀷瀛︿範

Leiming Dong

2018-01-28

学习的参考文献是(Lorenz et al. 2011; Campos et al. 2013)。

假定标记之间的连锁相或标记单倍型块与因果多态性在家系间是一致的，因此全体标记效应的估计将是有意义的(Meuwissen2001)。 # large p，small n；预测子间相关或共线性

最小二乘法不能估计全部的效应值，因为自由度不够（ **矩阵奇异** ）。即便自由度足够用，标记间共线性会 **过拟合** 模型。过拟合模型能夸大小幅波动，导致预测能力差。或只能在训练群体中表现很好。

于是提出各种统计模型，大致可分为4类： **收缩、变量选择、核和降维方法**。

3类：收缩、贝叶斯（变量选择）和半参或非参（支持向量、随机森林回归）。这些方法在遗传架构的假设方面不同。(Gamal El-Dien2015)

模型的基本形式一样：

是基因型和表型的关联函数。模型都设法最小化一个特定的成本函数。最小二乘的成本函数是残差的平方和：。最小二乘的回归系数是：

## 岭回归（又叫随机回归BLUP）

遗传值是个标记效应值之和。

与最小二乘的区别是多了一个项，引入的目的是 **避免矩阵奇异，并且降低预测子间的共线性**。 选择的标准是最小化模型误差；还有一个方法是假定标记效应从一个均值为0的正态分布中随机抽取的，然后解混合线性模型方程，是残差方差与标记效应方差的比值var(*e*)/var()。如果var()比var(*e*)小得多的话，会造成标记效应收缩到0，很大。 岭回归的成本函数是

第二项称为约束（constraint）、惩罚（penalty）或正则化（regularizer）。越大，大的收缩得越厉害。假设是标记效应同分布，指的是同样收缩到0，不是所有效应都相等。

## LASSO Least absolute shrinkage and selection operator

LASSO的成本函数：

与岭回归的共同点是都惩罚大的标记效应，二者的一般形式是：

*q*=1时是LASSO，*q*=2时是RR-BLUP，时，称为弹性网回归（elastic net）。因为正则化矩阵公式不同， **LASSO的系数收缩比RR-BLUP的要强，更有效，因为当很小时，比要大得多(Wu et al. 2009)**，而且一些系数收缩到0，产生一个稀疏模型，因此LASSO也能进行变量选择（所以有selection operator的意思）。LASSO的一个问题也是的取值问题。 Park and Casella (2008)提出贝叶斯LASSO。

### Lasso的局限

在非贝叶斯LASSO中，解决方案允许有n - 1个非零回归系数。 这是一个问题，因为在密集的标记数据的情况下，没有理由为什么训练集中的个体数量应该限制具有非零效应的标记的数量。 最多允许n个非零估计的回归系数。在复杂性状的WGR中，没有理由限制具有非零效应的标记的数量被n（观测的数量）限制。 当预测变量相关时，当LD跨越很长的区域时会出现某种情况，执行变量选择的方法如LASSO通常会被岭回归超过。

zou提出了综合RR和LASSO二者优点的方法（弹性网）：建议 **将L1和L2范数的加权平均值用作惩罚** ，即，对于，。模型涉及2个正则参数()和

## Reproducing kernel Hilbert spaces (RKHS) and support vector machine regression

RKHS：经典加性遗传模型结合核函数。核函数将预测变量转换为观测值之间的一组距离，以产生一个正定矩阵，用于线性模型。 **输入空间和特征空间** 是这个文献中经常遇到的术语。 在GS的情况下， **输入数据是标记得分数据**，因此，输入空间是多维空间，其中每个个体的位置由其标记分数确定。 通过 **对输入数据应用内核函数将输入空间转换为特征空间**。 RKHS可以用矩阵符号表示：

是核心条目的矩阵，量化个体彼此之间的距离，就像量化系谱关系的加性关系矩阵，即个体之间的遗传距离。这里的区别在于，核函数包括一个或多个平滑参数（ **h**）以影响特征空间中的距离和输入空间中的距离之间的关系。向量可以被解释为可以使用标准混合模型方程估计的特征空间内的单独效应。CV或bootstrap方法可以用来确定h的最优值，例如在群体和性状之间的模型中允许很大的灵活性。RKHS方法可以在标准的定量遗传模型框架内解释。 由于内核方法包含了很大的灵活性，而且没有线性假设，所以它们可以 **同时捕获所有类型的非加性效应**。 **支持向量机也是核方法。其成本函数可称为 -intensive**

是一个用来衡量误差对成本函数的贡献的常数；是 **-intensive损失函数**，意思是不惩罚误差，除非它们大于：

正常情况下，残差的总和不管是大还是小，都会对回归模型进行惩罚。在这里，只有最小化“大”（> ）残差的模型才是有利的，而所有其他残差是可以接受的。 **这个步骤的目的是开发牺牲许多小错误为了排除大错误的模型。** 这个问题的解决方案涉及 **拉格朗日（Lagrange）优化**，其产生一组系数，乘以训练样本与正在预测的观察之间的所有可能的内积的和。训练样本 **i** 和 **j** 的内积定义为，是观测标记得分向量。为了允许非线性，内积项可以被核函数替换，有效地将SVR优化从输入空间传送到特征空间。

## Partial least squares regression and principle component regression降维

从变量中提取 **潜变量** 为预测变量的线性组合，并用于响应预测：

是从原始预测变量中提取的潜在变量，是与该变量相关联的效应。即，是一个的向量，用于在线性组合中给潜变量加权。通常是潜变量的个数，比要小得多，是正交的，目的是消除共线性的问题。这个程序涵盖了各种技术，取决于哪个变异来源被认为是最重要的。在PC回归中，选择潜在变量尽可能多地解释原始预测变量。这种方法产生的预测空间信息向量可能不会与响应的变化相关联。在PLS中，选择潜变量以使潜变量和响应之间的关系尽可能强。 **为了避免由于太少或太多潜在变量而导致模型准确性降低（Solberg等人，2009），可以通过CV或相关技术来确定潜在变量的数量。** 在PCR和PLS中，都可以重新写成预测变量的直接函数，不用潜变量：

## 贝叶斯

贝叶斯方法也可用于估计的变量选择和收缩。 **大多数惩罚估计等价于某些类型的贝叶斯模型的后验模式**。

RR-BLUP所做的遗传效应在基因组上均匀分布的假设并不令人满意，Meuwissen等 （2001）试图用贝叶斯分析来放松它。

### BayesA

每个标记效应是从具有其自身方差的正态分布中抽取的：，这使得每个标记都可以向不同的程度缩小到零。反过来，方差参数从经过缩放的反向分布中采样。这个模型也被称为贝叶斯收缩回归。

### BayesB

没有效应的标记也给了概率。 如果遗传变异存在于少数基因座而在许多基因座缺失，这个 **模型将更好地反映潜在的遗传架构**。

是指示预测模型中标记 \_k\_的存在的指示变量。服从均值为0、有限方差的正态分布。方差的先验分布服从一个混合分布：

时，BayesB和BayesA是一样的。然而，在处理真实的生物体和性状时，这个参数（）的合理值是未知的。为了克服这个问题，一个称为 **BayesC** 的分析本身估计，其中的先验分布在0和1之间是均匀的。另外， **BayesC** 假设所有效应的标记的先验方差相等。即时，时。反过来，该方法一起估计所有非零的。在估计时，以这样的方式对标记进行分组会对所有先验进行加权。 **随机搜索变量选择（SSVS）**，早于Meuwissen的（2001）BayesB，但是与BayesB有大致相同的先验假设。

**所有这些贝叶斯模型共同的问题是参数估计不能以分析方式获得。** 然而，有可能开发所谓的马尔可夫链蒙特卡罗（MCMC）算法，其从后验分布中对参数值进行采样。 然后可以通过重复抽样和计算适当的汇总统计数据（例如分布的均值或中值）来近似这些分布。 不幸的是，这个解决方案在计算上非常昂贵。 因此， **非MCMC方法** 已经被用于估计这些汇总统计。

## 统计方法总结

在RR-BLUP模型下，标记效应是来自具有固定方差的正态分布的样本。通过缩小所有的标记效应到相同的程度，并包括模型中的所有标记，RR-BLUP的使用意味着从业者相信这个性状被许多具有小影响的基因座所控制。相反，贝叶斯B假设大多数位点对性状没有影响，因此大多数标记都不在预测模型中。包含在模型中的标记具有从具有不同方差的分布中采样的效果。因此，贝叶斯B有效地假定，性状是由相对较少的效应大小不同的基因座控制的。在BayesC的情况下，包含在模型中的标记的比例是根据数据估计的。包括在模型中的标记效应是从估计方差的相同分布从数据。这使得BayesC更具灵活性来模拟寡源到多基因的性状。 我们强调这些模型的先验和它们的含义，强调这些模型假定不同的基因结构。因此，就遗传结构不同而言，所有性状和种群都没有单一的最佳模型。 **如果遗传结构由多个小效应位点组成（Buckler et al。，2009），RR-BLUP和利用全球遗传关系信息的模型运作良好。如果大效QTL解释了很多遗传变异（Anderson等，2001; Munkvold等，2009），贝叶斯B等模型应该受到青睐。此外，这些示例模型假定可加性，并且如果非加性遗传效应是重要的，则可能不会很好地执行。一些人群中存在流行的非加性效应的证据（Dudley and Johnson，2009）; RKHS和SVR在这些情况下可能占有优势。**

# 非参方法（机器学习）

## RKHS

与BLUP类似 在称为基因组BLUP（G-BLUP）的方法中使用的基因组关系矩阵 **被更一般的核矩阵代替**，所述矩阵在个体之间产生相似性，即使遗传无关。 **核是定义个体之间的协方差结构的任何平滑函数K**。 该方法假定可以使用测量希尔伯特空间中的objects（焦点）之间的距离的核矩阵来表示欧几里德空间中的距离。 希尔伯特空间中，更相似的基因型比那些具有较少的相似性更密切。

# 

均方误差有非常好的几何意义，它对应了常用的欧几里得距离或简称“欧氏距离”（Euclidean distance）。基于均方误差最小化来进行模型求解的方法称为“最小二乘法”（least square method）。在线性回归中，最小二乘法就是试图找到一条直线，使所有样本到直线上的欧氏距离之和最小

估计量的准确度可以用估计的和参数的真实值之间的 **平方欧几里得距离** 来测量。在标量的情况下，这只是平方的偏差：。写成的目的是强调预测子是抽样数据的函数。 均方误差（MSE）是平方欧几里得距离的期望值（超过数据的可能实现）：。这能拆成2项：估计值的方差加上其偏差的平方，。 处罚和贝叶斯方法是两个最常用的收缩估计程序

# 支持向量机

SVM回归的基本思想是通过非线性映射函数将样本从预测空间映射到高维特征空间，并在后一空间进行线性回归(Jannink2010)。

Campos, Gustavo de los, John M Hickey, Ricardo Pong-Wong, Hans D Daetwyler, and Mario P L Calus. 2013. “Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding.” *Genetics* 193 (2): 327–45. doi:[10.1534/genetics.112.143313](https://doi.org/10.1534/genetics.112.143313).

Lorenz, Aaron J., Shiaoman Chao, Franco G. Asoro, Elliot L. Heffner, Takeshi Hayashi, Hiroyoshi Iwata, Kevin P. Smith, Mark E. Sorrells, and Jean-Luc Jannink. 2011. “Genomic Selection in Plant Breeding.” In *Advances in Agronomy*, 110:77–123. C. doi:[10.1016/B978-0-12-385531-2.00002-5](https://doi.org/10.1016/B978-0-12-385531-2.00002-5).

Wu, Tong Tong, Yi Fang Chen, Trevor Hastie, Eric Sobel, and Kenneth Lange. 2009. “Genome-wide association analysis by lasso penalized logistic regression.” *Bioinformatics* 25 (6): 714–21. doi:[10.1093/bioinformatics/btp041](https://doi.org/10.1093/bioinformatics/btp041).