Advances in Applied Mathematics 应用数学进展, 2017, 6(4), 583-588
Published Online July 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.64068

Hans汉斯

Model for Detecting QTL Controlling for Dynamic Trait Based on B-Spline Function

Xiaojing Zhou[1], Qianyu He[2], Qiaosheng Zhang[1], Ming Fang[3], Li Yan[4], Yangyang Li[2], Qi Li[2]

[1]Department of Mathematics, Heilongjiang Bayi Agriculture University, Daqing Heilongjiang
[2]College of Animal and Veterinary Medicine, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang
[3]College of Information Technology, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang
[4]College of Life Science and Technology, Heilongjiang Bayi Agricultural University, Daqing Heilongjiang
Email: zhouxiaojing7924@126.com

Received: Jul. 1st, 2017; accepted: Jul. 18th, 2017; published: Jul. 24th, 2017

Abstract

Dynamic traits are those phenotypic values change with time and other quantifiable factors such as age, parities, physiological status, performance level and environment etc. Because of the special economic status of the dynamic traits in breeding and production, it is very important to reveal the genetic regularity and improvement of these traits. The choice of body shape is beneficial to the improvement of the whole health and milk production of dairy cows. On the basis of B-spline function, a random regression model (RRM) has been developed to detect the QTLs controlling the dynamic traits. A real dataset for China Holstein cows, which contains the records of body weight from the local dairy farm, was analyzed and the biological conclusions were derived.

Keywords

Dynamic Trait, B-Spline, Detection, Model

检测控制奶牛动态性状的QTL方法研究——基于B样条插值函数

周晓晶[1]，何倩毓[2]，张巧生[1]，方　铭[3]，闫　丽[4]，李洋洋[2]，李　琦[2]

[1]黑龙江八一农垦大学理学院，黑龙江　大庆
[2]黑龙江八一农垦大学动物科技学院，黑龙江　大庆
[3]黑龙江八一农垦大学生命科学技术学院，黑龙江　大庆

文章引用: 周晓晶, 何倩毓, 张巧生, 方铭, 闫丽, 李洋洋, 李琦. 检测控制奶牛动态性状的 QTL 方法研究——基于 B 样条插值函数[J]. 应用数学进展, 2017, 6(4): 583-588. https://doi.org/10.12677/aam.2017.64068

[4]黑龙江八一农垦大学信息技术学院，黑龙江 大庆
Email: zhouxiaojing7924@126.com

收稿日期：2017年7月1日；录用日期：2017年7月18日；发布日期：2017年7月24日

摘 要

随着时间(生命时期、年龄、胎次等)或其他可以量化的因素(生理状态、生产水平、代谢率和环境条件等)变化的性状，称为动态性状，如身高、体重、胸围、产奶量等等。由于动态性状在育种和生产中特殊的经济地位，揭示这类性状遗传及其改良提高的研究工作尤显重要。体型性状的选择有利于奶牛整体健康和产奶性状的提高。本文基于B样条插值函数，建立检测调控动态性状基因位点的随机回归模型，实际数据分析表明模型的合理性和适应性。

关键词

动态性状，B-样条插值，检测，拟合

1. 前言

表型值随着时间(生命时期、年龄、胎次等)或其他可以量化的因素(生理状态、生产水平、代谢率和环境条件等) 变化的性状称为动态性状。这些性状的表型值有无限个，而且其变化可用一些连续的函数(或随机过程)来描述，在指定时间点观测动态性状的有限变化点就获得了实验所需的重复记录[1] [2] [3]。

为检测控制动态类型的数量性状位点(QTL)，学者们采用参数和非参数的模型来描述表型和基因型效应随时间的变化。大多数采用具有生物学意义的数学模型去拟合数量性状位点(QTL)的表型效应。然而，这样的参数方法具有局限性，因为数学函数是非线性的。相反，学者们较青睐于将非参数的 Legendre 多项式嵌入随机回归模型中作为子模型检测动态性状位点。Legendre 多项式的优点除了能够拟合任何性状的生物曲线之外，更重要的一点是它是线性模型，这样的理论和方法论促使它能够广泛地应用于 QTL 定位的线性模型用来估计 QTL 参数。高阶的 Legendre 多项式能够很好地拟合均值和方差的变化，但是，这样的多项式经常在曲线的极值点过高地估计观测值，导致 Runge(龙格)现象，也就是说，由于曲线上极值点处的震荡,曲线的拟合度随着阶数的升高而显著降低。相反，B-spline [4]能够更广泛地应用于非参数的数据分析去推断变量的经验分布。样条插值函数应用于检测 QTL 是杨于 2006 提出的[5]。

目前，奶牛育种的总体趋势是在保持奶牛产奶量以及乳成分等优良遗传性状的同时，兼顾奶牛的躯体结构、趾蹄健康、使用寿命、繁殖性能等综合遗传性能的选育，以获得奶牛养殖的最大经济效益。有研究表明-加强体型性状的选择对奶牛产奶量以及乳成分的提高有利，也有利于降低体细胞数，增强个体乳房炎的抗性，鉴于此，本文研究了检测调控奶牛动态性状基因位点的检测方法[6]。

应用B-spline函数检测调控动态性状QTL的关键是结点个数、结点位置以及阶次的选择。对于样本容量较小，观测值较少的数据集仍然是用较简单的多项式作为子模型拟合效果更好。根据文献资料显示，

对于奶牛的身高、体重这些体尺性状选择5、6、7阶的，结点个数为3个的B-spline拟合效果更好。

2. 方法

以 BC 群体为例，动态性状表型值和 QTL 的遗传效应及环境效应的关系可由如下模型描述

$$y_i = \mu + z_j\alpha + \varepsilon_j \tag{1}$$

这里 μ 为群体均值， z_j 为 QTL 基因型指示变量，取值 1 和−1 分别为对应着 QTL 基因型 QQ 和 Qq. α 为加性效应， ε_j 为剩余误差，服从正态分布 $N(0,\sigma^2_{\varepsilon_j})$ 记 $X_j=(1,z_j)$， $B=(\mu,\alpha)^{\mathrm{T}}$，从而模型(1)记为矩阵形式如下：

$$y_i = X_jB + \varepsilon_j \tag{2}$$

2.1. Legendre 多项式嵌入随机回归模型中

用k阶Legendre多项式描述动态性状随时间的变化规律，“从而第i个体在第t个测定日”动态性状表型值的遗传模型可表示为

$$y_i(t) = \mu(t) + \sum_{j=1}^{q} x_{ij}\alpha_j(t) + \beta_i(t) + \varepsilon_i \tag{1}$$

其中， $i=1,2,\cdots,n$，n为个体数； $\mu(t)$ 为时刻t的群体均值， x_{ij} 为基因型指示变量，当基因型为QQ时，指示变量为1，当基因型为Qq时，指示变量为−1； $\alpha_j(t)$ （$j=1,2,\cdots,q$）为第j个QTL的遗传值，q为基因组上观测到的QTL的最大个数， $\beta_i(t)$ 为指定个体依赖于时间的环境效应，服从正态分布 $N\left[0,\sigma^2_\beta(t)\right]$； ε_i 为指定个体随时间独立的环境效应，服从正态分布 $N\left(0,\sigma^2_\varepsilon\right)$。可知该模型为固定效应模型， $\mu(t)$ 和 $\alpha_j(t)$ 为固定效应， $\beta_i(t)$ 为随机效应。

2.2. B-spline 函数嵌入随机回归模型中

模型中的所有参数，除了 σ^2_ε 都是时间的函数。而参数和时间之间的函数关系可用B样条(B-spline)来描述。1974年，Gordon和Riesenfeld用B样条基线函数代替了Bernstein基线函数，构造了B-spline样条曲线。B-spline样条曲线分段组成。每一段的参数t的区间为[0，1]。这样就克服了Bezier曲线的缺点：改变Berier曲线任意一个控制点，曲线上的所有点都变换。B-Spline曲线的优点：修改某一控制点只引起与该控制点相邻的曲线形状发生变化，远处的曲线形状不受影响。

定义 $\boldsymbol{\psi}(t)=\left[\psi_{0,p}(t)\quad\psi_{1,p}(t)\quad\cdots\quad\psi_{r,p}(t)\right]$ 为B样条的协变量，带有 k 个节点， p 阶多项式，$r=k-p-2$。定义 $\boldsymbol{\mu}=\left[\mu_0\quad\mu_1\quad\cdots\quad\mu_r\right]^{\mathrm{T}}$ 为随时间独立的群体均值做成的向量。在这里我们可以视B样条为权重，构造 $\boldsymbol{\mu}$ 的线性组合来描述随时间独立的群体均值 $\mu(t)$。其他的参数可以用同一个B样条来描述，具体为 $\alpha_j(t)=\boldsymbol{\psi}(t)\boldsymbol{\alpha}_j$， $\boldsymbol{\alpha}_j=\left[\alpha_{j0}\quad\alpha_{j1}\quad\cdots\quad\alpha_{jr}\right]^{\mathrm{T}}$， $j=1,2,\cdots,q$； $\beta_i(t)=\boldsymbol{\psi}(t)\boldsymbol{\beta}_i$，$\boldsymbol{\beta}_i=\left[\beta_{i0}\quad\beta_{i1}\quad\cdots\quad\beta_{ir}\right]^{\mathrm{T}}$， $i=1,2,\cdots,n$。由于 $\boldsymbol{\beta}_i$ 为随机回归效应，我们可以假设服从正态分布 $N\left(0,\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)$，$\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ 为指定个体随时间独立的随机回归效应做成的协方差矩阵，这样我们利用B样条重新参数化后，(1)改写成如下线性模型

$$y_i(t) = \boldsymbol{\psi}(t)\boldsymbol{\mu} + \sum_{j=1}^{q} x_{ij}\boldsymbol{\psi}(t)\boldsymbol{\alpha}_j + \boldsymbol{\psi}(t)\boldsymbol{\beta}_i + \varepsilon_i \tag{2}$$

在$t+1$个固定时间点测得每个个体的表型值， $t_0,t_1,\cdots,t_m$，从而这t+1个向量作成向量形式如下

$$\boldsymbol{y}_i=\left[y_i(t_0)\quad y_i(t_1)\quad\cdots\quad y_i(t_m)\right]^{\mathrm{T}}$$

定义$\boldsymbol{\psi}=\left[\boldsymbol{\psi}^{\mathrm{T}}\left(t_0\right)\ \ \boldsymbol{\psi}^{\mathrm{T}}\left(t_1\right)\ \ \cdots\ \ \boldsymbol{\psi}^{\mathrm{T}}\left(t_m\right)\right]$为$(r+1)\times(m+1)$矩阵，这样$y_i$的线性模型记为矩阵形式

$$\boldsymbol{y}_i=\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\mu}+\sum_{j=1}^{p}x_{ij}\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\alpha}_j+\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\beta}_i+\boldsymbol{\varepsilon}_i \tag{3}$$

这里，$\boldsymbol{\varepsilon}_i=\left[\varepsilon_{it_0}\ \ \cdots\ \ \varepsilon_{it_m}\right]^{\mathrm{T}}$为$(m+1)\times1$剩余效应向量，服从正态分布，$\boldsymbol{\varepsilon}_i\sim N\left(\boldsymbol{0},\boldsymbol{I}\sigma_\varepsilon^2\right)$。

2.3. 参数估计

模型(3)的数学期望为

$$\mathrm{E}\left(\boldsymbol{y}_i\left|\boldsymbol{\mu},\boldsymbol{\alpha}_j\right.\right)=\boldsymbol{U}_i=\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\mu}+\sum_{j=1}^{p}x_{ij}\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\alpha}_j$$

协方差矩阵为

$$\mathrm{Var}\left(\boldsymbol{y}_i\left|\boldsymbol{\mu},\boldsymbol{\alpha}_j\right.\right)=\boldsymbol{V}=\boldsymbol{\psi}^{\mathrm{T}}\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\boldsymbol{\psi}+\boldsymbol{I}\sigma_\varepsilon^2$$

似然函数为

$$p\left(\boldsymbol{y}\left|\boldsymbol{M},\boldsymbol{\theta}\right.\right)=\prod_{i=1}^{n}p\left(\boldsymbol{y}_i\left|\boldsymbol{M},\boldsymbol{\theta}\right.\right)\propto\left|\boldsymbol{V}\right|^{-n(m+1)/2}\exp\left[\sum_{i=1}^{n}\left(\boldsymbol{y}_i-\boldsymbol{U}_i\right)^{\mathrm{T}}\boldsymbol{V}^{-1}\left(\boldsymbol{y}_i-\boldsymbol{U}_i\right)\right] \tag{4}$$

$\boldsymbol{\mu}$的先验分布为常数，即为均匀先验。

所有参数的联合先验分布函数为

$$p\left(\boldsymbol{\theta}\right)=p\left(\boldsymbol{\mu}\right)p\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right)p\left(\sigma_\varepsilon^2\right)\prod_{j=1}^{q}p\left(\boldsymbol{\alpha}_j\left|\boldsymbol{A}_j\right.\right)p\left(\boldsymbol{A}_j\right) \tag{5}$$

结合数据的条件概率密度和参数的先验分布得到数据和参数的联合分布为

$$p\left(\boldsymbol{\theta}\left|\boldsymbol{y},\boldsymbol{M}\right.\right)\propto p\left(\boldsymbol{y},\boldsymbol{M}\left|\boldsymbol{\theta}\right.\right)p\left(\boldsymbol{\theta}\right) \tag{6}$$

与参数的联合后验分布成比例。此联合后验分布即为目标分布函数。

对似然函数取对数，求偏导，即可得到每个参数的极大似然估计的对数形式，由于每个参数都是待估计参数的函数，因此需要迭代算出。

3. 实例分析

动物资源群体来源于大庆本地牧场-红骥牧场。奶牛品种为中国荷斯坦奶牛，重要经济性状为体重、体长动态性状，测定时间从 2010 年 1 月至 2013 年 13 月。具体为 2010 年出生的 60 头奶牛、2011 年出生的 70 头奶牛及 2012 年出生的 80 头奶牛。由于奶牛早期生长发育状况与日后的产奶量、繁殖问题和健康问题有着密切关系，因此对奶牛生长性状的研究主要针对的是从出生到第一个泌乳期结束这段时间，分为初生、断奶、周岁和头胎分娩这四个时间点。在实际牛场管理过程中，这几个时间点正是牛只转群的时间，饲料成分、管理方式、饲养环境都发生了明显变化，这样的划分方式也与生产实际相契合。所以特别关注初生、断奶、周岁和头胎分娩这四个时间点的体重记录，将它们视为一组。共 4 组数据。处理数据，去掉异常值后剩余 143 个个体的 4 组数据。

分别用3，4，5，6阶Legendre多项式和5、6、7阶的B-spline插值函数(结点个数为3)拟合。对体重、体高的具体分析结果如表1~表3所示。

表中易见，阶数越高，拟合效果越好，而且 B-spline 函数的拟合效果要比 Legendre 多项式的好。

由拟合结果可见，一般而言，阶数越高，拟合结果与真值越接近，标准差也越来越小。通过将Legendre多项式和B-spline函数视为自模型，比较可见，前者的5阶和6阶的估计结果要比后者的估计结果有较大的标准差，且后者的7阶的估计结果更接近于真值(表2)。

由于荷斯坦奶牛的产奶量主要在体高的三个时期最大，所以我们只关注了三个阶段的体高，为156~167 cm，此时产奶量最高，120~135 cm 时比 136~155 cm 产奶量高。我们只拟合了这三个阶段的体高的生长轨迹。由于 3 阶 Legendre 多项式的拟合效果不好，所以我们略去了该结果(表 3)。

表中易见，阶数越高，拟合效果越好，而且B-spline函数的拟合效果要比Legendre多项式的好。

Table 1. Fitting accuracy of body weight growth trajectory for individuals under different models
表 1. 不同模型下个体的体重增长轨迹的拟合精度

	最小值	最大值	平均值	标准差
3阶Legendre多项式	0.7460	0.8380	0.7630	0.0345
4阶Legendre多项式	0.8190	0.8900	0.8342	0.0217
5阶Legendre多项式	0.8850	0.9210	0.8990	0.0097
6阶Legendre多项式	0.9120	0.9350	0.9243	0.0761
5阶B-spline函数	0.8960	0.9210	0.8970	0.0218
6阶B-spline函数	0.9140	0.9300	0.9160	0.0196
7阶B-spline函数	0.9340	0.9680	0.9516	0.0432

Table 2. Results under different models
表 2. 不同模型的体重检测结果

模型	QTL位置(cM)	μ	α	β	ε
3阶Legendre多项式	4 - 13.0	38.96 (19.36)	8.45 (2.11)	12.35 (2.58)	3.98
4阶Legendre多项式	7 - 55.0	37.33 (12.48)	9.63 (1.96)	15.85 (2.47)	3.65
5阶Legendre多项式	2 - 60	37.96 (13.68)	9.48 (1.57)	17.84 (2.39)	3.44
6阶Legendre多项式	3 - 28.7	40.68 (11.55)	10.25 (1.26)	19.64 (1.38)	2.48
5阶B-spline函数	4 - 105.0	36.58 (9.56)	10.87 (0.92)	19.64 (1.14)	3.96
6阶B-spline函数	6 - 19.4	39.87 (8.28)	8.23 (0.54)	19.28 (0.85)	2..38
7阶B-spline函数	4 - 67.8	34.33 (5.36)	9.35 (3.48)	18.25 (2.45)	3.78

Table 3. Fitting accuracy of body height growth trajectory for individuals under different models
表 3. 不同模型下体高的增长轨迹的拟合精度

	最小值	最大值	平均值	标准差
4阶Legendre多项式	0.4468	0.7233	0.5652	0.0460
5阶Legendre多项式	0.5344	0.8791	0.7836	0.0182
6阶Legendre多项式	0.8743	0.9261	0.8365	0.0562
5阶B-spline函数	0.8907	0.9397	0.8643	0.0325
6阶B-spline函数	0.9467	0.9823	0.9543	0.0205
7阶B-spline函数	0.8798	0.9560	0.476	0.0478

4. 总结

本文应用B-spline函数作为随机回归模型的子模型分析了控制动态性状QTL的检测。比较了B-spline函数与Legendre多项式在参数估计(包括均值和标准差)上的精确性，从对比结果可见，前者的5阶和6阶的估计结果要比后者的估计结果有较小的标准差，且前者的7阶的估计结果更接近于真值，说明用B-spline函数估计模型参数更准确。用所提出的方法分析了来源于当地农场奶牛的关于身高体重的实际数据集，对实际育种工作者具有一定的指导意义。

基金项目

2016 年度大庆市指导性科技计划项目：调控动态性状基因位点的贝叶斯定位方法的研究与实例分析(zd-2016-089)。

参考文献 (References)

[1] 杨运清, 李仁杰, 李淑玲. 动态性状遗传参数的估计方法[J]. 畜牧兽医学报, 1996, 27(5): 412-416.

[2] 杨润清, 高会江, 孙华, 等. 远交群体动态性状基因定位的似然分析 I. 理论方法[J]. 遗传学报, 2004, 31(1): 1116-1122.

[3] 黄少卿, 崔意旎, 杨润清. 基于 Legendre 多项式的动态性状功能定位[J]. 自然科学通报, 2005, 15(10): 1183-118.

[4] Rodriguez-Zas, S.L. (2002) Detection of Quantitative Trait Loci Influencing Dairy Traits Using a Model for Longitudinal Data. *Journal of Dairy Science*, **85**, 2681-2691. https://doi.org/10.3168/jds.S0022-0302(02)74354-3

[5] 高会江, 孙华, 等. F_2 群体动态性状基因定位的极大似然分析[J]. 东北林业大学学报, 2006, 34(1): 72-77.

[6] 毛杰, 王根林, 余盼. 上海地区荷斯坦奶牛体型性状——产奶性状和体细胞评分的遗传统计分析[J]. 南京农业大学学报, 2015, 38(4): 650-655.