系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
28(2) (2008, 2), 243–250

基于样条变换的 PLS 回归的非线性结构分析*

孟 洁

(北京航空航天大学, 北京 100083; 中央财经大学, 北京 100081)

王 惠 文 黄 海 军 苏 建 宁

(北京航空航天大学, 北京 100083)

摘要 基于样条变换的 PLS 非线性回归模型既吸取了样条函数分段拟合以适应任意曲线连续变化的优点，又借鉴了偏最小二乘回归方法能够有效解决自变量集合高度相关的技术. 针对多元加法模型，从理论和仿真试验的角度分别验证了，对于多个独立自变量对单因变量为非线性关系的数据系统，基于样条变换的 PLS 回归方法不仅能够有效实现自变量对因变量的整体预测，而且能够提取各维自变量对因变量的单独非线性作用特征，从而确定数据系统内部的复杂非线性结构关系，增强了模型的可解释性.

关键词 样条函数，偏最小二乘回归，非线性，特征提取，结构分析.

MR(2000) 主题分类号 62-07, 62J02

1 引 言

1984 年 Wold 提出的偏最小二乘回归 (partial least-squares regression, 简称 PLS 回归) 为在变量集合多重相关条件下的回归建模提供了一个有效的工具. 为了进一步解决非线性建模问题， 1989 年 Wold 提出了二次多项式偏最小二乘回归 [1], 又于 1992 年提出了样条偏最小二乘回归[2];1992 年，Hoskuldsson 提出了另一种不同的二次多项式偏最小二乘回归 [3,4];1994 年， LarsAarhus 发表了《 Nonlinear empirical modeling using local PLS models 》，提出对解释变量局部分段使用 PLS 回归方法. 上述这些非线性 PLS 回归研究扩大了 PLS 回归的应用范畴. 然而，由于目前这些模型还仅局限于对模型的整体预测上，因此对变量间的关系探讨并不尽人意. 此外，有关模型的计算方法也比较复杂，使其应用受到局限. 基于此，就有必要对数据特征及其内部的相互关系作进一步研究和探讨，建立更直观、简洁并能够反映变量间非线性作用特征的结构模型.

1997 年，Jean-Francois Durand 提出基于多元加法样条变换的 PLS 回归模型 (multivariate spline)[5], 即利用样条基函数，将自变量与因变量之间的未知非线性关系按照各维自变量对

* 国家自然科学基金 (70371007) 和国家杰出青年科学基金 (70125003) 资助课题.
收稿日期：2004-05-31, 收到修改稿日期：2006-05-12.

因变量的拟线性关系相加展开，再进行偏最小二乘回归求参，从而得到自变量对因变量的整体函数解析式．该方法主要有两个方面的突出特点．第一，由于样条函数具有分段拟合、按需要裁剪以适应任意曲线连续变化的特点，因此可以有效地避免龙格 (Runge) 现象 [6,7]，同时保持函数的光滑和连续性，使拟合曲线对原始数据中的特异点不过分敏感，以排除噪声的影响．第二，在经样条基函数变换后，模型维数显著增加，极有可能导致样本点数少于变换后的自变量个数，或变换后数据多重相关的情况．例如 Eubank(1988) 和 Stone(1985) 在使用加法样条普通多元回归方法时 [8,9]，就曾遇到这一难题．而采用 PLS 回归方法，就可以有效地解决观察值数量少或自变量存在多重相关的问题 [10]．

在基于样条变换的 PLS 回归的基础上，本文将进一步研究模型内部的非线性结构特征．从理论上证明：对于多元加法模型 (additive model)，在各维自变量独立的情况下，使用基于样条变换的 PLS 非线性回归方法得到整体模型后，可以从中提取各维自变量的回归关系式，分别画出各维自变量对因变量的函数曲线，从而真实地反映各自变量对因变量的解释作用情况，分析其对因变量的贡献 [11]．通过仿真实验，这一结论也得到充分验证．

2 基于样条变换的非线性 PLS 回归建模

2.1 建模思路

设有因变量 $\{y\}$ 和 p 个自变量 $\{x_1, x_2, \cdots, x_p\}$，观测了 n 个样本点，由此构成了自变量与因变量的数据表 $X = [x_1, x_2, \cdots, x_p]_{n\times p}$ 和 $Y = [y]_{n\times 1}$．建模的目标是找到自变量 X 与因变量 Y 之间的非线性关系，更进一步地，求得每个 $x_j(j = 1, 2, \cdots, p)$ 对 y 的作用特征，用函数关系表示为

$$y = f(x_1, x_2, \cdots, x_p) + \varepsilon = \beta_0 + f_1(x_1) + \cdots + f_p(x_p) + \varepsilon, \tag{1}$$

其中，$\varepsilon \sim N(0, \sigma^2)$ 为随机误差项，ε 与自变量独立，β_0 为常数项，$f_j(x_j)(j = 1, 2, \cdots, p)$ 为 x_j 对 y 的非线性作用函数．

对于 $f_j(x_j)(j = 1, 2, \cdots, p)$，本文采用 3 次 B 样条函数进行拟合，具体展开为

$$\widehat{f}_j(x_j) = \sum_{l=0}^{M_j+2} \beta_{j,l}\Omega_3\Big(\frac{x_j - \xi_{j,l-1}}{h_j}\Big), \tag{2}$$

式中，$\beta_{j,l}$ 为模型的待定参数．而 $\Omega_3\big(\frac{x_j-\xi_{j,l-1}}{h_j}\big) = \frac{1}{3!h_j^3}\sum\limits_{p=0}^{4}(-1)^p \begin{pmatrix} 4 \\ p \end{pmatrix} (x_j - \xi_{j,l-3+p})_+^3$ 为 3 次 B 样条基函数 [12]；在计算过程中，定义 $\xi_{j,l-1}$，h_j，M_j 分别为 x_j 上划分的区间分点、分段长度以及分段个数，

$$\xi_{j,l-1} = \min(x_j) + (l-1)h_j, \tag{3}$$

其中

$$h_j = \frac{\max(x_j) - \min(x_j)}{M_j}, \quad l = 0, 1, \cdots, M_j + 2.$$

结合 (1) 和 (2) 式，得到全体自变量与因变量的非线性拟合函数为

$$\widehat{y} = \beta_0 + \sum_{j=1}^{p}\widehat{f}_j(x_j) = \beta_0 + \sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}\Omega_3\Big(\frac{x_j - \xi_{j,l-1}}{h_j}\Big), \tag{4}$$

(4) 式是关于 $z_{j,l}=\Omega_3(\frac{x_j-\xi_{j,l-1}}{h_j})$ 的线性函数，采用 PLS 回归方法进行参数求解.

2.2　模型的建立

根据上述建模思路，下面给出模型建立的具体算法步骤.

第 1 步　对自变量空间的每一维 $j(1\le j\le p)$ 进行 3 次 B 样条变换 $x_j\to Z_j$.

1) 确定 M_j, 并根据 (3) 式求得分点 $\xi_{j,l-1}$;

2) 对 $[x_j]_{n\times 1}$ 作如下 3 次 B 样条变换

$$z_{j,0}=\Omega_3\Big(\frac{x_j-\xi_{j,-1}}{h_j}\Big), z_{j,1}=\Omega_3\Big(\frac{x_j-\xi_{j,0}}{h_j}\Big),\cdots, z_{j,M_j+2}=\Omega_3\Big(\frac{x_j-\xi_{j,M_j+1}}{h_j}\Big), \tag{5}$$

则 $Z_j=\big\{z_{j,0},z_{j,1},\cdots,z_{j,M_j+2}\big\}=\big\{\Omega_3\big(\frac{x_j-\xi_{j,l-1}}{h_j}\big),\ l=0,1,\cdots,M_j+2\big\}$.

第 2 步　对因变量及新的自变量进行中心标准化处理

$$\widetilde{z}^i_{j,l}=\frac{z^i_{j,l}-\overline{z}_{j,l}}{s_{j,l}},\quad \widetilde{y}_i=\frac{y_i-\overline{y}}{s_y},\quad l=0,1,\cdots,M_j+2;\ j=1,2,\cdots,p\ ;\ i=1,2,\cdots,n. \tag{6}$$

其中，$\overline{z}_{j,l},\overline{y}$ 分别是 $z_{j,l}$, y 的样本均值，$s_{j,l}$, s_y 分别是 $z_{j,l}$, y 的样本方差.

记经过中心标准化处理的自变量为 $\widetilde{Z}$, 因变量为 $\widetilde{Y}$, 则原变量空间变换为

$$\begin{aligned}[X,Y]&=[x_1,x_2,\cdots,x_p,y]_{n\times(p+1)}\to\\ [\widetilde{Z},\widetilde{Y}]&=[(\widetilde{Z}_1)_{n\times(M_1+3)},\cdots,(\widetilde{Z}_p)_{n\times(M_p+3)},\widetilde{Y}]\\ &=[\widetilde{z}_{1,0},\cdots,\widetilde{z}_{1,M_1+2},\cdots,\widetilde{z}_{p,0},\cdots,\widetilde{z}_{p,M_p+2},\widetilde{y}]_{n\times\Big\{\Big(\sum\limits_{j=1}^{p}(M_j+3)\Big)+1\Big\}},\end{aligned}$$

从而得到新数据表符合线性关系

$$\widetilde{y}=\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\alpha_{j,l}\widetilde{z}_{j,l}. \tag{7}$$

第 3 步　对 (7) 式进行 PLS 回归求参 (PLS 回归具体算法详见文献 [10]), 提取最多的 PLS 成分数，求得回归系数 $\alpha_{j,l}(j=1,2,\cdots,p\ ;\ l=0,1,\cdots,M_j+2)$.

第 4 步　将 (6) 式代入 (7) 式，即

$$\frac{y-\overline{y}}{s_y}=\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\alpha_{j,l}\frac{z_{j,l}-\overline{z}_{j,l}}{s_{j,l}},$$

得到

$$y=\beta_0+\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}z_{j,l}, \tag{8}$$

其中

$$\beta_{j,l}=s_y\frac{\alpha_{j,l}}{s_{j,l}},\quad \beta_0=\overline{y}-\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}\overline{z}_{j,l}. \tag{9}$$

第 5 步 将回归系数及变换式 (5) 代入 (8) 式，得到 Y 关于 X 的非线性模型

$$\widehat{y}=\beta_0+\sum_{j=1}^{p}\widehat{f}_j(x_j)=\beta_0+\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}\Omega_3\Big(\frac{x_j-\xi_{j,l-1}}{h_j}\Big). \tag{10}$$

3 各维自变量对因变量的非线性解释能力

本节将在加法模型的条件下，讨论各维自变量对因变量的解释贡献能力，找出自变量对因变量的内在非线性结构特征，使模型更具有实际意义和应用价值．下面就从理论证明和仿真试验两方面讨论，在原始自变量之间满足独立的条件下，应用本模型能够有效辨识各维自变量对因变量的作用情况．

引理 3.1 随机变量 ξ, η, x, y 分别是 ξ, η 的容量为 n 的样本，样本均值 $\overline{x}$, $\overline{y}$ 为零，相关系数 $\mathrm{Cov}(x,y)=0$, 则有 $\|x+y\|^2=\|x\|^2+\|y\|^2$.

证 由 $\mathrm{Cov}(x,y)=\frac{1}{n}\langle x-\overline{x},y-\overline{y}\rangle=0$ 及 $\overline{x}=\overline{y}=0$, 得 $\langle x,y\rangle=0$, 所以

$$\|x+y\|^2=\langle x+y,x+y\rangle=\langle x,x\rangle+\langle x,y\rangle+\langle y,x\rangle+\langle y,y\rangle=\|x\|^2+\|y\|^2.$$

引理 3.2 若 x,y 是两个独立的随机变量， f, g 是使 $f(x)$, $g(y)$ 成为随机变量的函数，则 $f(x)$, $g(y)$ 为独立的随机变量．

具体证明过程见文献 [13] 第二章第 4 节推论 3.

根据引理 3.1 和引理 3.2, 可以证明，在自变量互相独立的条件下，使用本文的方法对数据系统建立非线性模型，按照整体拟合精度最高的原则求解模型参数；在利用所得到的模型进行非线性结构分析时，从整体模型中提取各维自变量的表达式，能够真实反映该自变量对因变量的解释作用和趋势特征．下面给出相应定理及证明．

定理 3.1 若原始数据表自变量各维 x_j $(j=1,2,\cdots,p)$ 互相独立，则从使用 “基于样条变换的 PLS 非线性结构分析 ” 得到的整体回归模型中，提取各维的拟合曲线，能够真实再现各维自变量对因变量非线性作用特征．

证 设真实函数为

$$y=\sum_{j=1}^{p}f_j(x_j)+\varepsilon=\sum_{j=1}^{p}(f_j(x_j)-\overline{f}_j(x_j))+\overline{y}+\varepsilon,$$

其中，$\overline{f}_j(x_j),\overline{y}$ 分别为 $f_j(x_j),y$ 的样本均值，显然，$\overline{y}=\sum\limits_{j=1}^{p}\overline{f}_j(x_j)$, 记 $\widetilde{f}_j(x_j)=f_j(x_j)-\overline{f}_j(x_j)$, 则 $y=\sum\limits_{j=1}^{p}\widetilde{f}_j(x_j)+\overline{y}+\varepsilon$, 且易见 $\widetilde{f}_j(x_j)$ 的均值为零．

由 (1) 式，使用 “基于样条变换的 PLS 非线性结构分析 ” 得到的整体拟合函数为

$$\widehat{y}=\beta_0+\sum_{j=1}^{p}\widehat{f}_j(x_j)=\Big(\overline{y}-\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}\overline{z}_{j,l}\Big)+\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}z_{j,l}=\overline{y}+\sum_{j=1}^{p}\sum_{l=0}^{M_j+2}\beta_{j,l}(z_{j,l}-\overline{z}_{j,l}).$$

记 $\widetilde{\widehat{f}}_j(x_j)=\sum\limits_{l=0}^{M_j+2}\beta_{j,l}(z_{j,l}-\overline{z}_{j,l})$, 则 $\widehat{y}=\overline{y}+\sum\limits_{j=1}^{p}\widetilde{\widehat{f}}_j(x_j)$, 且易见 $\widetilde{\widehat{f}}_j(x_j)$ 的均值为零. 于是

$$\|y-\widehat{y}\|^2=\Big\|\Big(\sum_{j=1}^{p}\widetilde{f}_j(x_j)+\overline{y}+\varepsilon\Big)-\Big(\overline{y}+\sum_{j=1}^{p}\widetilde{\widehat{f}}_j(x_j)\Big)\Big\|^2=\Big\|\sum_{j=1}^{p}(\widetilde{f}_j(x_j)-\widetilde{\widehat{f}}_j(x_j))+\varepsilon\Big\|^2.$$

由 $x_j(j=1,2,\cdots,p)$, ε 互相独立及引理 2, 得 $g_j(x_j)=\widetilde{f}_j(x_j)-\widetilde{\widehat{f}}_j(x_j)(j=1,2,\cdots,p)$, ε 互相独立, 且易见 $g_j(x_j)$ 的均值为零.

又由引理 3.1, 有

$$\begin{aligned}\|y-\widehat{y}\|^2&=\Big\|\sum_{j=1}^{p}(\widetilde{f}_j(x_j)-\widetilde{\widehat{f}}_j(x_j))+\varepsilon\Big\|^2\\&=\Big\|\sum_{j=1}^{p}g_j(x_j)\Big\|^2+\|\varepsilon\|^2\\&=\sum_{j=1}^{p}\|g_j(x_j)\|^2+\sigma^2\\&=\sum_{j=1}^{p}\|\widetilde{f}_j(x_j)-\widetilde{\widehat{f}}_j(x_j)\|^2+\sigma^2.\end{aligned}$$

根据

$$\widetilde{f}_j(x_j)=f_j(x_j)-\overline{f}_j(x_j),\quad \widetilde{\widehat{f}}_j(x_j)=\sum_{l=0}^{M_j+2}\beta_{j,l}(z_{j,l}-\overline{z}_{j,l})=\widehat{f}_j(x_j)-\sum_{l=0}^{M_j+2}\beta_{j,l}\overline{z}_{j,l}$$

得

$$\|y-\widehat{y}\|^2=\sum_{j=1}^{p}\Big\|(\widehat{f}_j(x_j)-f_j(x_j))+\Big(\overline{f}_j(x_j)-\sum_{l=0}^{M_j+2}\beta_{j,l}\overline{z}_{j,l}\Big)\Big\|+\sigma^2.$$

由于模型按照整体拟合精度最高的原则求解参数, 即 $\|y-\widehat{y}\|$ 达到最高精度, 因此,

$$\Big\|(\widehat{f}_j(x_j)-f_j(x_j))+\Big(\overline{f}_j(x_j)-\sum_{l=0}^{M_j+2}\beta_{j,l}\overline{z}_{j,l}\Big)\Big\|\ (j=1,2,\cdots,p),$$

亦达到最高精度, 即各维自变量的拟合函数与真实函数之间, 除可能相差一个常数外, 其所反映的非线性趋势特征是真实可靠的.

4　仿真试验

4.1　仿真试验的基本思路

试验的基本思路是：生成自变量 $X=[x_1,x_2,\cdots,x_p]_{n\times p}$ 的仿真数据, 选取有代表性的非线性函数 $f_j(x_j)$ $(j=1,2,\cdots,p)$, 据此生成相应的因变量 $Y=\Big[y=\sum\limits_{j=1}^{p}f_j(x_j)\Big]_{n\times 1}$ 的数据

样本；按照第 2 节中的基于样条变换的 PLS 非线性结构分析进行建模求解；对结果首先检验模型整体 $\widehat{y}=\widehat{f}(x_1,x_2,\cdots,x_p)$ 的拟合效果，再进一步从整体回归模型中提取各维上的非线性关系式 $\widehat{f_j}(x_j)$ $(j=1,2,\cdots,p)$, 验证其与原始选取的函数的真实非线性特征的一致性.

4.2 仿真试验

1) 仿真数据

随机生成各维自变量在 $[-1,1]$ 上的独立均匀分布；取样本点数为 $n=100$; 为了验证本方法的可行性，选取 3 个有代表性的函数 $f_1(x_1)=\sin(\frac{\pi x_1}{2})$(在 $[-1,1]$ 单调), $f_2(x_2)=x_2^2$(不一致单调), $f_3(x_3)=-x_3$(线性); 取随机误差项 $\varepsilon\sim N(0,0.01)$ 于是得到因变量的相应取值为 $y=f_1(x_1)+f_2(x_2)+f_3(x_3)+\varepsilon$. 至此, 构造出原始数据表 $X=[x_1,x_2,x_3]_{100\times 3}$ 和 $Y=[y]_{100\times 1}$.

2) 模型求解

取 $M_j=3$ $(j=1,2,3)$, 进行自变量空间的 3 次 B 样条变换，易见，经过变换后，自变量空间由 $p=3$ 维增至 $p\sum\limits_{j=1}^{3}(M_j+3)=18$ 维. PLS 回归方法提取全部成份进行回归求参. 事实上，经变换后，新变量间存在线性相关性，可以提取的 PLS 最大成分数为 $15(<18)$, 这也说明了比较普通多元回归，在本模型中使用 PLS 回归方法的必要性.

3) 试验结果

首先，总体拟合效果如图 1, 图 2 所示.

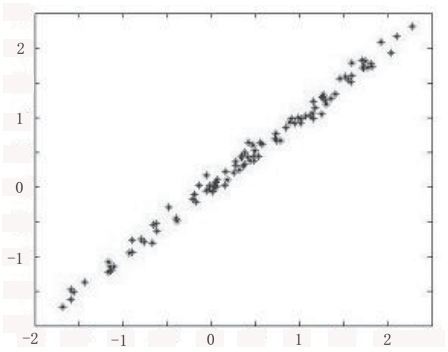

图 1

图 1 中，“$*$” 是以因变量的真实值为横坐标，以拟合值为纵坐标取得，这些点分布在对角线上，表明了整体回归模型对因变量的拟合效果很好.

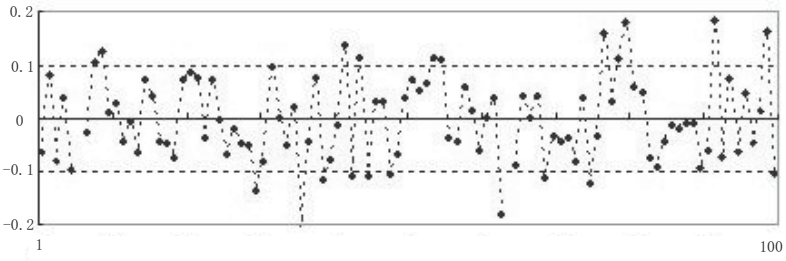

图 2

图 2 给出了 100 个样本点的误差值 (拟合值 - 真实值), 易见，拟合点误差主要在 ±0.1 范围内波动，平均相对误差为 0.098.

在此基础上，从求得的整体非线性回归关系式中，提取各维上的非线性成分，分别绘制其非线性曲线，如图 3 所示.

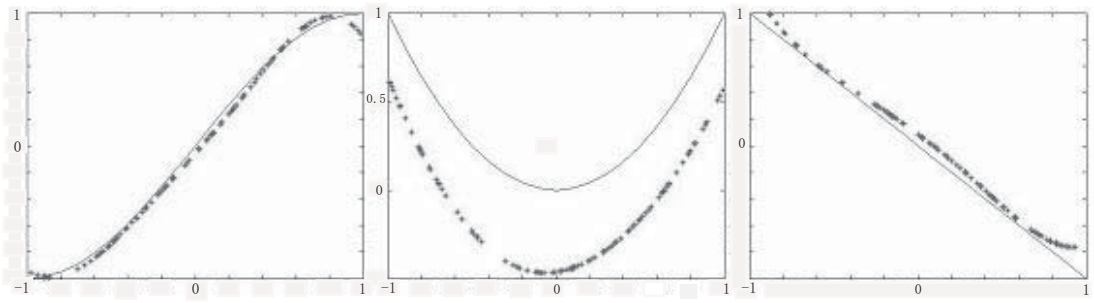

图 3

图 3 中，实线表示各维上真实非线性曲线，“$*$”为拟合曲线，显然，在每一维上，拟合得到的非线性关系反映了真实的非线性结构特征；x_1, x_2, x_3 上的非线性走势都得到了真实地再现.

5　小　结

本文主要对基于样条变换的 PLS 回归方法进行了非线性结构分析，研究结果表明，“基于样条变换的 PLS 回归的非线性结构分析”方法借鉴了逼近理论中的样条拟合，将非线性问题转化为拟线性问题，在此基础上，利用求解线性问题的 PLS 回归方法进行模型的系数求解，其结果，不仅整体模型能够达到较高的精度，更为重要的是，在原始自变量独立的情况下，从整体模型中能够有效提取出各维自变量对因变量的非线性作用特征，这在数据分析领域，研究复杂系统内部各自变量对因变量的解释贡献程度以及预测上都具有重要的意义和应用价值.

参　考　文　献

[1] Wold S, Kettaneh-Wold N and Skagerberg B. Nonlinear PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 1989, **7**(7): 53–65.

[2] Wold S. Nonlinear partial least squares modelling II spline inner function. *Chemometrics and Intelligent Laboratory Systems*, 1992, **14**(14): 71–84.

[3] Hoskuldsson A. The H-principle in modeling with applications to chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 1992, **14**(2): 139–154.
[4] Hoskuldsson A. Quadratic PLS regression. *Journal of Chemometrics*, 1992, **6**: 307–334.
[5] Durand Jean-Francois. Local polynomial additive regression through PLS and splines: PLSS. *Chemometrics and Intelligent Laboratory Systems*, 2001, **58**(2): 235–246.
[6] 冯康等. 数值计算方法. 北京：国防工业出版社， 1978, 1–40.
[7] 李岳生. 数值逼近. 北京：人民教育出版社， 1978, 65–68, 76–132.
[8] Eubank R L. Spline Smoothing and Nonparametric Regression. New York, Marcel Dekker, 1988.
[9] Stone C J. Additive regression and other nonparametric models. *The Annals of Statistics*, 1985, **13**(2): 689–705.
[10] 王惠文. 偏最小二乘回归方法及其应用. 北京：国防工业出版社， 1999.
[11] Hastie T and Tibshirani R. Generalized Additive Models. London, Chapman and Hall, 1990.
[12] 颜庆津. 数值分析. 北京：北京航空航天大学出版社. 2000.
[13] 严士健，刘秀芳. 概率论基础. 北京：科学出版社， 1982.
[14] 严华生，谢应齐，曹杰. 非线性统计预报方法及其应用. 昆明：云南科技出版社， 1998.

NONLINEAR STRUCTURE ANALYSIS WITH PARTIAL LEAST-SQUARES REGRESSION BASED ON SPLINE TRANSFORMATION

MENG Jie

(*Beijing University of Aeronautics and Astronautics, School of Economics and Management, Beijing* 100083; *Central University of Finance and Economics, School of Statistics, Beijing* 100081)

WANG Huiwen HUANG Haijun SU Jianning

(*Beijing University of Aeronautics and Astronautics, School of Economics and Management, Beijing* 100083)

Abstract Nonlinear Partial Least-Squares Regression Model based on Spline Transformation not only takes advantages of the characters of spline functions which can locally fit continuous curves properly, but also brings in Partial Least-Squares Regression Method which can effectively solve the problem of high correlations in the set of independent variables. In this paper, according to additive modeling methods both in theory and simulation, it is proven that Nonlinear Partial Least-Squares Regression Method based on Spline Transformation can not only get the exact whole forecasting model, but also successfully extract nonlinear features of each independent variable's effect on the dependent variable when dealing with nonlinear data systems with multi-absolute independent variables for one dependent variable. In this way, acquire the complex nonlinear structures of the data system and an explainable model can be acquired.

Key words Spline functions, partial least-squares regression, nonlinear, feature extraction, structure analysis.