基于正交 GF 系统的散乱数据拟合及分析*

蔡占川，陈 伟
（澳门科技大学资讯科技学院，澳门）

摘 要：提出了一种散乱数据的正交表示方法，该方法利用正交 GF 系统来逼近或插值给定的散乱数据点集。k（k 为非负整数）次 GF 系统是一类正交样条函数系，Haar 函数及 Franklin 正交函数恰好分别是 $k=0$ 及 $k=1$ 时的特殊情形。基于 GF 系统，提出了求解散乱数据问题的新的能量模型，根据该能量模型的频谱，可以对散乱数据进行不同层次的曲面重构。实验结果表明该方法高效且效果良好。

关键词：散乱数据；最佳平方逼近；GF 系统
中图分类号：TP399 **文献标志码**：A **文章编号**：0529－6579（2013）05－0073－05

Least Square Approximation and Analysis for Scattered Data Based on Orthogonal GF System

CAI Zhanchuan, *CHEN Wei*
(Faculty of Information Technology, Macau University of Science and Technology, Macau, China)

Abstract: Base on GF system, an orthogonal representation algorithm for scattered data is proposed. When $k=0$ and $k=1$, GF system are Haar functions and Franklin functions respectively. A new energy model is proposed to solve this problem based on GF system. According to GF spectrum, different hierarchical surfaces could be reconstructed for scattered data. The experiments show that the method proposed is efficient and can produce pleasing results.

Key words: scattered data; least square approximation; GF system

散乱数据是指在二维平面上或三维空间中，无规则的、随机分布的抽样数据点[1]。自然界中的很多观测信号都是散乱的，诸如卫星的观测数据、海平面船载测量数据等。在对这些数据进行分析时，除了抽样的数值之外，也需要知道任意位置的数值，在这种情况下，就需要对所得到的观测数据进行插值或逼近（工程上通常称为拟合）[2-4]。正如 Roussell 所言“描述自然的问题最后一般都归结为逼近问题。”事实上，求解逼近问题的指导思想有二：其一，要选择简单的函数空间作为描述对象的逼近空间；其二，这个空间要有好的逼近度。为了得到好的逼近效果，通常选择最佳平方逼近。接下来的问题是，选择什么样的函数空间？对于散乱数据而言，传统的散乱数据的曲面造型插值方法是 Shepard 方法，它是一个与距离成反比的加权方法，有在插值点附近会出现平台、精度较差、函数重建质量较差和计算量大等缺点。后来陆续又有了各种各样的对多元散乱数据的插值方法，都有一定的效果，如薄板样条、光滑余因子方法、Box 样条[5]、Bézier 曲面[6]、顶点样条[7]、多项式自然样条和三角网最优插值[8]. 这些方法具有能量极小的性质，避免插值点附近出现平台现象，重建质量较好；近年来，细分算法成为一个热点，还有层次 B 样条方法[9]、径向基函数方法，但是它们都没有选择正交基函数。本文试图选择一类正交基函数对散乱数据做逼近。利用正交基逼近后，可以对散乱数据

* **收稿日期**：2013－04－12
基金项目：国家重点基础研究发展计划“973”资助项目（2011CB302400）；澳门科技发展基金资助项目（084/2012/A3，004/2011/A1，006/2011/A1，015/2010/A）；国家自然科学基金面上资助项目（61170320，61272364）；浙江大学 CAD & CG 国家重点实验室开放课题资助项目（A1310）；广东省自然科学基金资助项目（S2011040002981）
作者简介：蔡占川（1974 年生），男；**研究方向**：计算机图形学、数据处理与分析；E-mail：zccai@ must. edu. mo

进行更加深入的分析和挖掘，诸如频谱分析等。

最近，文献［10］提出了一类新的正交样条函数系——GF 系统，作为样条函数空间的一组标准正交基，它能较为精确表达连续样条函数。GF 系统是通过对一组线性无关的函数组施行正交化过程得到，零次 GF 系统恰是 Haar 函数。

本文首先简要介绍 GF 系统的构造过程，然后详细介绍了基于 GF 系统的散乱数据的最佳逼近算法，接着是实验结果，最后是结论与展望。

1 k 次 GF 系统简介

k 次 GF 系统是由一组线性无关的函数组经正交化过程得到。首先定义 k 次 GF 系统正交化之前的线性无关函数组，记为 $G_{k,n}^{j}(t)$。令

$$G_{k,n}^{j}(t)=\begin{cases}0, & t\in[0,\dfrac{2j-1}{2^{n-1}}];\\ (t-\dfrac{2j-1}{2^{n-1}})^{k}, & t\in(\dfrac{2j-1}{2^{n-1}},1];\end{cases}=(t-\frac{2j-1}{2^{n-1}})_{+}^{k},t\in[0,1]$$

$G_{k,n}^{j}(t)$ 表示 k 次第 n 组第 j 个函数。其中，$n=2,3,4,\cdots,j=1,2,\cdots,2^{n-2}$。

线性无关函数组 $G_{k,n}^{j}(t)$ 分组排列如下：

第 1 组 $G_{k,1}^{1}(t),G_{k,1}^{2}(t),\cdots,G_{k,1}^{k+1}(t)$；

第 2 组 $G_{k,2}^{1}(t)$；

第 3 组 $G_{k,3}^{1}(t),G_{k,3}^{2}(t)$；

第 4 组 $G_{k,4}^{1}(t),G_{k,4}^{2}(t),G_{k,4}^{3}(t),G_{k,4}^{4}(t)$；

M

第 n 组 $G_{k,n}^{1}(t),G_{k,n}^{2}(t),\cdots,G_{k,n}^{2^{n-2}}(t)$

其中，第 1 组表达式为：

$$\text{第 1 组}\begin{cases}G_{k,1}^{1}(t)=1, & t\in[0,1]\\ G_{k,1}^{2}(t)=t, & t\in[0,1]\\ M\\ G_{k,1}^{k+1}(t)=t^{k}, & t\in[0,1]\end{cases}$$

对 $G_{k,n}^{j}(t)$ 施行 Gram-Schmidt 正交化，即得到相应的 k 次 GF 系统，记为 $F_{k,n}^{j}(t)$。

图 1 是 $k=0,1,2,3$ 时，$G_{k,n}^{j}(t)$ 与 $F_{k,n}^{j}(t)$ 的部分函数图形。

2 基于 GF 系统的散乱数据的最佳平方逼近

2.1 基本思想

记 GF 系统前 m 项基函数 $j_j,j=1,2,\cdots,m$ 所张成的函数空间为 F，即 $F=\text{span}\{\varphi_j(x)\}_{j=1}^{m}$。如

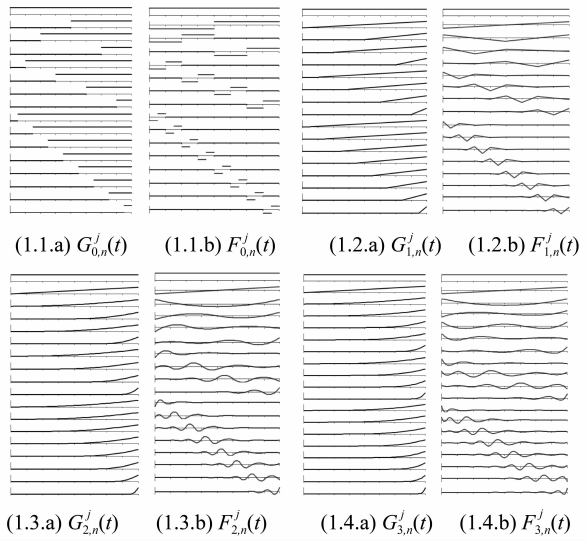


图 1 $G_{3,n}^{j}(t)$ 与正交化后函数组 $F_{k,n}^{j}(t)$（k 次 GF 系统）部分函数图形

Fig. 1 $G_{3,n}^{j}(t)$ and their orthogonalization function group $F_{k,n}^{j}(t)$ (GF system of degree k) graphics

果函数 $g(x)=\sum a_j\varphi_j(x)$ 与测量值数据 $\{f_j\}$ 有误差向量：

$$r^{\mathrm{T}}=(g(x_1)-f_1,\cdots,g(x_n)-f_n)$$

所谓的散乱数据最佳平方逼近就是在 F 中，寻找函数 $g(x)$，使得这个误差向量的平方和取最小，也就是通常意义下的最小二乘逼近。

2.2 算法描述

本文处理的散乱数据集限于双自变量的数据集：$P=\{(x,y,z)\mid z=f(x,y)\}$，算法可以用如下过程描述：已知二维平面域上散乱数据点集 $P=\{(x_k,y_k,z_k)\mid z_k=f(x_k,y_k),k=1,2,\cdots,N\}$，如图 2 所示，$W$ 为包含点集 P 的最小矩形域，$x_k\in[X_{\min},X_{\max}],y_k\in[Y_{\min},Y_{\max}]$。

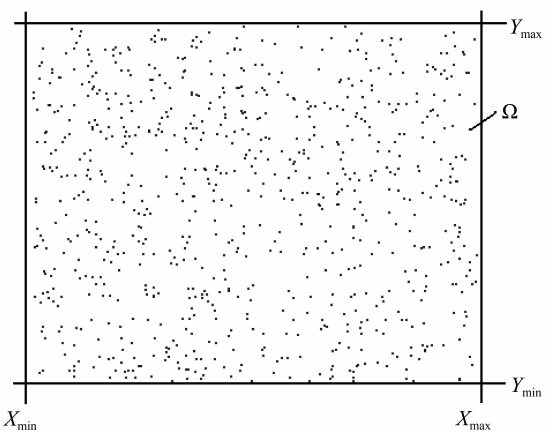


图 2 二维平面域上散乱数据点集 P 示意图

Fig. 2 Scattered data point set P in two-dimensional plane domain

由于 GF 系统函数的定义域为区间 [0,1]，首先将 Ω 映射到 [0,1] × [0,1] 矩形域上。规则如下：设 $l_x = X_{\max} - X_{\min}, l_y = Y_{\max} - Y_{\min}$，变换后点集 $P' = \{(x'_k, y'_k, z'_k), k = 1,2,\cdots,N\}$，其中

$$x'_k = \frac{x_k - X_{\min}}{l_x},$$

$$y'_k = \frac{y_k - Y_{\min}}{l_y},$$

$$z'_k = z_k$$

此时，$x'_k \in [0,1], y'_k \in [0,1]$。用 GF 系统作最小二乘拟合，兼顾光滑性与计算量，我们一般选用 3 次 GF 系统作为拟合基函数，假设拟合后曲面为

$$g(x,y) = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}\varphi_i(x)\varphi_j(y)$$

各采样点残差表达式为

$$r_k = g(x_k, y_k) - z_k = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}\varphi_i(x_k)\varphi_j(y_k) - z_k$$

残差平方和

$$F(C_{ij}; i = 1,2,\cdots,m, j = 1,2,\cdots,n) = \sum_{k=1}^{N} r_k^2$$

依据最小二乘算法，问题归结为确定 $m \times n$ 个系数 C_{ij} 使残差平方和最小。对 $F(C_{ij})$ 求偏导数并令其为零，即得到一个 $m \times n$ 阶线性方程组，求解此线性方程组就可以得到拟合系数 C_{ij}。

2.3　能量模型

由于 GF 系统是标准正交样条函数系，拟合系数 C_{ij} 即为频谱。令 $E = \sum_{i=1}^{m}\sum_{j=1}^{n} C_{ij}^2$，则称 E 为散乱点的能量。相应地，称 E 及 $C_{ij}; i = 1,2,\cdots,m, j = 1,2,\cdots,n$ 为散乱点的能量模型。因此，可以通过对能量模型中 E 及 C_{ij} 实现对散乱数据的分析，从而可以深度挖掘散乱数据的信息。

3　实　验

下面通过几个实验来测试 GF 系统对散乱数据的逼近效果，并利用得到的能量模型进行了散乱数据的分析。

实验一　选定一个原始模型，如图 3 所示。然后分别在模型上随机采样 600 ~ 2 600 个不等的散乱点，应用上述算法分别对各个散乱点集进行拟合，得到拟合曲面，并计算各拟合曲面与原始曲面的均方误差 MSE。如表 1 所示。图 4 表明了采样点数与均方误差的变化关系，表明了本文方法对散乱数据拟合具有较高的拟合度。

实验二　选择采用一个球面参数曲面模型，分别在曲面上采样 128 × 128 个点，并加入随机噪声扰动以模拟真实的数据。将基于 GF 系统的拟合结果与其他几种拟合方法的结果进行比较。

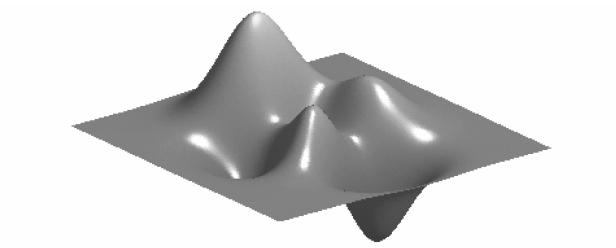


图 3　原始数据模型

Fig. 3　The original model

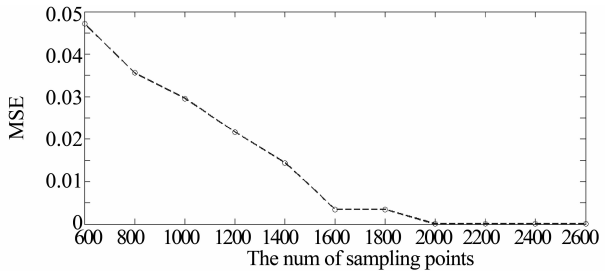


图 4　采样点与均方误差示意图

Fig. 4　Sampling points and the mean square error diagram

注：拟合曲面与理想标准曲面的差别用相对中误差 RRMSE 评价，设 i 点的真值为 256 × 256，拟合值为 z_i，误差 $\varepsilon_i = |Z_i - z_i|$，计算公式如下

$$\mathrm{RRMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\frac{\varepsilon_i}{Z_i}\right)^2}{N}}$$

PRMSE 值越小说明拟合效果越好。拟合对象是一个标准球体，其参数方程为

$$\begin{cases} x = \cos\theta\cos\varphi \\ y = \cos\theta\sin\varphi \quad 0 \leqslant \varphi \leqslant 2\pi, -\dfrac{\pi}{2} \leqslant \theta \leqslant \dfrac{\pi}{2} \\ z = \sin\theta \end{cases}$$

实验三　本例选取三组散乱数据，每组散乱数据又分别含有 1 000、1 200、1 500、2 000 个不同的采样点，如表 3 所示。

对上述散乱点采用本文算法进行拟合，并根据得到的拟合系数计算拟合曲面的能量。拟合曲面及其能量（E）如表 4 所示。

从表 4 可以看出，同一组内曲面的“能量”比较接近甚至相同，不同组间的曲面的“能量”差别较大。因此，利用 GF 系统进行散乱数据拟合，能够拟合出光滑的曲面，亦可以进行频谱分析。

表 1 采样本文方法拟合的曲面与原曲面的均方误差

Table 1 Mean square error of the original surface and the fitting surface by proposed method in this paper

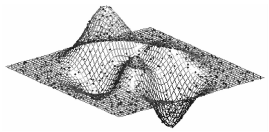
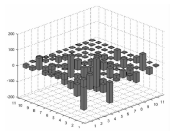
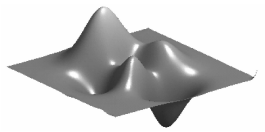
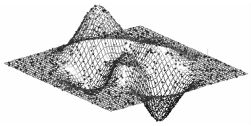
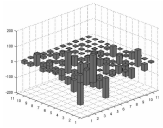
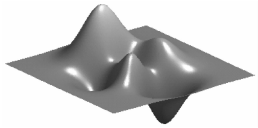
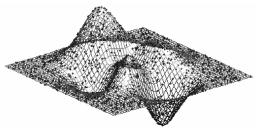
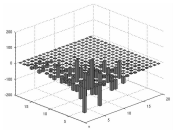
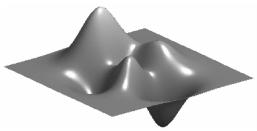
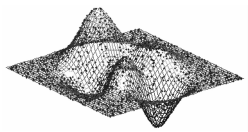
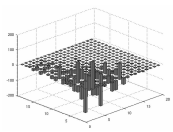
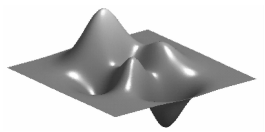
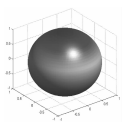
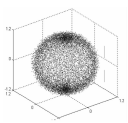
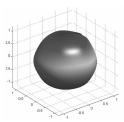
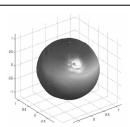
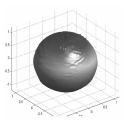
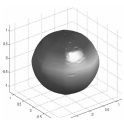
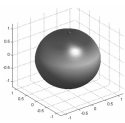
采样点数	实际采样图示 $\{x_k, y_k, z_k\}$	频谱	拟合曲面	与原曲面的均方误差
600				0.045 5
1 200				0.021 8
1 800				0.004 8
2 400				0.000 1

表 2 球体点云数据拟合

Table 2 Fitting of sphere point cloud data

目标曲面与采样点	 	相对中误差 RRMSE
DCT 拟合		0.994 3
db4 小波拟合		1.148 1
db8 小波拟合		0.996 3
sym4 小波拟合		0.848 9
3 次 GF 系统拟合		0.717 2

4 结论与展望

本文利用 GF 系统对给定的散乱数据点进行最佳平方逼近，得到拟合曲面。实验结果表明，该方法不仅可以拟合出光滑的曲面，而且因 GF 系统是正交样条函数系，还可以引入频谱分析手段进行后续分析工作。本文利用拟合系数（即频谱），计算了拟合曲面的能量，从而给出了散乱数据的一种新度量。下一步我们将研究快速算法，从而能应用于更大规模的散乱数据拟合。利用散乱数据的能量模型可以深度挖掘散乱数据所隐藏的内部信息。

参考文献：

[1] FRANKE R. Scattered data interpolation: tests of some methods[J]. Mathematics of Computation, 1982, 38: 181 – 200.

[2] BAJAJ C L, IHM I. Algebraic surface design using Hermite interpolation[J]. ACM Transactions on Graphics, 1992, 11(1): 61 – 91.

[3] SHEPARD D. A two dimensional interpolation function for irregularly spaced data[C]// Proc ACM 23rd Nat'l Conf, 1986, 517 – 524.

表 3　几组模型的不同采样点图示

Table 3　Different sampling points for several groups models

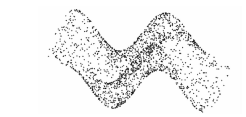
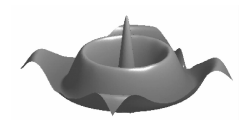
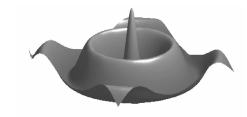
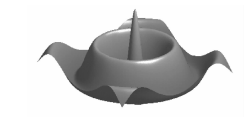
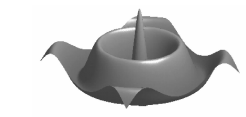
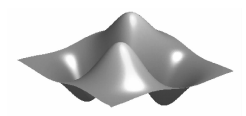
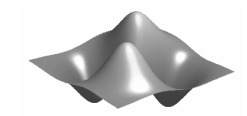
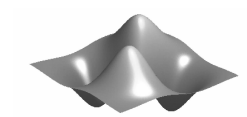
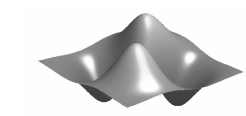
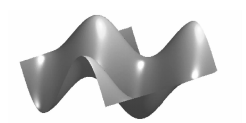
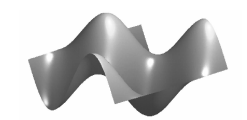
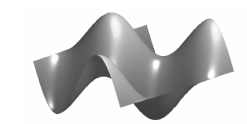
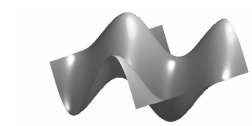
采样点数	1 000	1 200	1 500	2 000
第 1 组				
第 2 组				
第 3 组				

表 4　拟合曲面及其能量

Table 4　Fitting surfaces and their energies

第 1 组	 *E*=0.130 7	 *E*=0.113 8	 *E*=0.113 8	 *E*=0.113 7
第 2 组	 *E*=0.295 6	 *E*=0.295 6	 *E*=0.293 7	 *E*=0.295 5
第 3 组	 *E*=0.312 1	 *E*=0.312 9	 *E*=0.312 1	 *E*=0.312 1

[4] FRANKE R, NIELSON G M. Scattered data interpolation of large sets of scattered data[J]. Intl J Numerical Methods in Eng, 1980,15: 1691 – 1704.

[5] WANG R H, SHI X Q, LUO Z X, et al. Multivariate spline and its applications[M]. Beijing: Science Press, 1994.

[6] WANG N, MENG F Z, LI M. Interpolation of large scale scattered data base on Bezier surface[J]. Journal of Tianjin University of Technology, 2005, 21(2):40 – 43.

[7] WANG G F, SUN Y, ZHANG H X, et al. Large_scale interpolation of discrete data based on nodal point interpolation principle[J]. Journal of Harbin Engineering University, 2001, 22(1):45 – 48.

[8] CHUI CHARLES K, LAI M J. Filling polygonal holes using C1 cubic triangular spline patches[J]. Computer Aided Geometric Design, 2000, 17(4):297 – 307.

[9] ZHANG W Q, TANG Z S, LI J. Adaptive hierarchical b-spline surface approximation of large-scale scattered data [C]// Pacific Graphics'98, 1998.

[10] CAI Z C, CHEN W, QI D X, et al. A class of general Franklin functions and its application[J]. Chinese J Computers, 2009, 32(10): 2004 – 2013.