混合高斯参数估计的两种 EM 算法比较

刘旺锁[1,2]，王平波[1]，顾雪峰[1]
(1. 海军工程大学，湖北武汉 430033; 2. 广州大学，广东广州 510006)

摘要：混合高斯模型是一种典型的非高斯概率密度模型，获得广泛应用。其参数的优效估计可以通过最大似然方法获得，但最大似然估计往往因其非线性而难以实现，故期望最大化(Expectation-Maximization, EM)迭代算法成为一种常用的替代方法。常规 EM 算法性能受迭代初值设置影响大，且不能对模型阶数做出估计。一种名为贪婪 EM 的改进算法可以克服这两个缺点，获得更为准确的模型参数估计，但其运算量一般会远大于前者。本文对这两种 EM 算法进行综合研究，深入挖掘两者之间的关系，并基于相同的数值仿真实例，直观地演示比较两者的性能差异。

关键词：混合高斯；最大似然估计；期望最大化；贪婪期望最大化
中图分类号：TN911.7 **文献标识码**：A **文章编号**：1000-3630(2014)-06-0539-05
DOI 编码：10.3969/j.issn1000-3630.2014.06.012

Comparison of two EM algorithms for Gaussian mixture parameter estimation

LIU Wang-suo[1,2], WANG Ping-bo[1], GU Xue-feng[1]
(1. *Naval University of Engineering*, *Wuhan* 430033, *Hubei*, *China*; 2. *University of Guangzhou*, *Guangzhou* 510006, *Guangdong*, *China*)

Abstract: Gaussian mixture is a typical and widely-used non-Gaussian probability density distribution model. The expectation-maximization algorithm is a usual iterative realization for the maximum likelihood estimation of its parameters. However, its performance depends highly on the initial values. And it can not estimate the order of Gaussian mixture. The greedy expectation-maximization algorithm can solve these problems by incrementally adding Gaussian components to the mixture. But its operation quantity is often much larger than the former. The relationship between these two algorithms is discussed, and their concrete realization methods are given comparatively. With the same numerical instance, their performance differences are illustrated and studied.

Key words: Gaussian mixture; Maximum Likelihood Estimation(MLE); Expectation-Maximization(EM); Greedy Expectation-Maximization(GEM)

0 引 言

混合高斯(Gaussian Mixture, GM)模型是一种比较优秀的非高斯概率密度(Probability Density Function, PDF)模型，它具有参数少、结构简单、物理意义明显直观、拟合性能好等一系列优点，为雷达、声呐、通信、语音、图像等信号处理领域所广泛采用。使用 GM 模型对数据进行非高斯 PDF 建模的关键是如何快速、准确地得到 GM 模型参数估计。众所周知，对于非随机未知确定量，若其优效估计存在，则必然是最大似然估计(Maximum Likelihood Estimation, MLE)[1]。所以，首选是寻求 GM 参数的 MLE。

但是，对于多参量同时估计问题，MLE 一般难以严格实现。此时，往往代之以一种名为期望最大化(Expectation-Maximization, EM)的迭代算法[2]。文献[3]中提出了 GM 参数估计的 EM 迭代算法。这种 EM 算法存在参数初始化问题，如果初始化不恰当，迭代可能会错误地收敛于局部极值，不能得到正确的参数估计。不幸的是，对于 GM 参数估计，理想的 EM 迭代初始化方案尚未建立，这是 EM 算法应用的主要局限性所在。EM 算法的第二个局限性是，它不能对 GM 模型阶数做出估计，只能在固定的 GM 阶数下进行。这就是说，使用 EM 算法，必须对 GM 模型阶数做出预先假定，而对于某些极端非高斯数据，实际中很难事先确定其 GM 阶数。

为了克服 EM 算法的这两个缺陷，文献[4]提出了一种名为贪婪 EM(Greedy EM, GEM)迭代的改进算法(为以示区别，下文把传统 EM 算法简记为 CEM)。理论上，GEM 算法不依赖于初始值，且可自适应估计 GM 阶数。

收稿日期：2014-04-29；**修回日期**：2014-08-07
基金项目：国家自然科学基金资助项目(51109218)。
作者简介：刘旺锁(1965－)，男，江苏金坛人，硕士生导师，研究方向为声纳装备保障与效能评估、水声信号处理。
通讯作者：王平波，E-mail: blackberet@126.com

文献[5]、[6]分别把 CEM 算法和 GEM 算法引入到了水声信号处理中。而且，文献[5]中提出了一种多初值初始化方案，可以部分地避免错误收敛问题。受当时数值仿真方法的限制，文献[6]直接使用一段海试数据简单演示了 GEM 算法的性能。本文将对 CEM 算法和 GEM 算法进行综合研究，深入发掘两者之间的关系，并基于相同的数值仿真实例，直观而详细地比较两者的性能差异。

1 混合高斯模型

一般地，GM 模型的 PDF 可表述为如式(1)所示的加分布形式：

$$f(x_n)=\sum_{m=1}^{M}\varepsilon_m\cdot f_m(x_n) \tag{1}$$

式中：x_n 为非高斯序列 $\boldsymbol{x}=[x_1, x_2, \cdots, x_N]^{\mathrm{T}}$ 的第 n 个样本值；M 为模型阶数；ε_m 为混合参数，满足式(2)所示的关系；f_m 为高斯分布 $\mathbb{N}(\mu_m, \sigma_m^2)$ 的 PDF，如式(3)所示。

$$\sum_{m=1}^{M}\varepsilon_m=1 \tag{2}$$

$$f_m(x_n)=\frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\left[-\frac{(x_n-\mu_m)^2}{2\sigma_m^2}\right] \tag{3}$$

其中：μ_m 为均值，σ_m^2 为方差。可以看到，GM 序列的 PDF 是一种纯粹的数学级数表达式，但它仍有自己明确的物理解释。事实上，从式(1)所示的 GM 模型的 PDF 中，恰恰能看清混合高斯信号形成的物理机理，即混合高斯信号是由多个高斯源信号依据一定的概率分布组成的。具体而言，M 代表了组成非高斯信号的高斯源(亦称高斯分量)数量；f_m 就是第 m 个高斯源 $G_m\sim\mathbb{N}(\mu_m, \sigma_m^2)$ 的 PDF；ε_m 为当前样本 u_n 来自第 m 个高斯源的概率。

显然，通过简单地调整 GM 参数 $\boldsymbol{\theta}=[\varepsilon_1,\cdots,\varepsilon_M, \mu_1,\cdots,\mu_M,\sigma_1^2,\cdots,\sigma_M^2]^{\mathrm{T}}$ 就可以拟合几乎任意的单钟型或多钟型 PDF。这就是 GM 模型在非高斯信号处理中倍受青睐的原因。

可以用图 1 所示的“完全数据”的概念来解释混合高斯数据 $\boldsymbol{x}$ 的形成过程：

设有 M 组高斯数据 $\{G_m\sim\mathbb{N}(\mu_m, \sigma_m^2)\,|m=1, 2, \cdots, M\}$，对于任一时刻 n，以 $\{\varepsilon_m\,|m=1, 2, \cdots, M\}$ 的概率分布从 $\{G_m\}$ 中抽取一点作为当前样本值 x_n，如此即可得到 N 点的混合高斯序列 $\boldsymbol{x}=\{x_n|n=1, 2, \cdots, N\}$。这里不难看到，仅从数据 $\boldsymbol{x}$ 本身，并不能分辨每一个样本值 x_n 来自哪个高斯分布序列，从这个意义上看，$\boldsymbol{x}$ 并不包含数据的全部信息，是“非完全数据”。但是，如图 1 所示，若在每一个样本值 x_n 之后再缀以一个可以指示样本值来源的 M 维向量(比如，若 x_n 来自于 $\boldsymbol{g}_m$，则 $\boldsymbol{z}_n$ 除第 m 个元素为 1 外其余所有元素皆为 0)，则数据 $\boldsymbol{u}=\{u_n=[x_n, \boldsymbol{z}_n]|n=1, 2, \cdots, N\}$ 就包含了包括样本来源在内的全部数据信息，在这个意义上，数据 $\boldsymbol{u}$ 称为“完全数据”，向量 $\boldsymbol{z}_n$ 则称为指示向量。

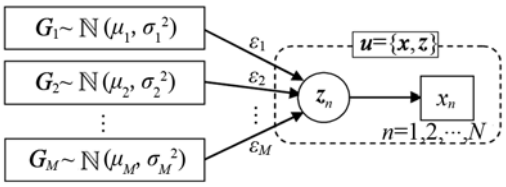


图 1 完全数据形成图解

Fig.1 Composition of the complete data in EM

2 CEM 算法

GM 参数的 MLE 问题可以描述为：给定样本数据 $\boldsymbol{x}=[x_1, x_2, \cdots, x_N]^{\mathrm{T}}$，寻求令式(4)所示函数值取得最大的参数值 $\boldsymbol{\theta}=[\varepsilon_1,\cdots,\varepsilon_M;\mu_1,\cdots,\mu_M;\sigma_1^2,\cdots,\sigma_M^2]^{\mathrm{T}}$：

$$l_M(\theta)=\sum_{n=1}^{N}\ln f_M(x_n) \tag{4}$$

基于对图 1 所示完全数据概念的理解，Aaron[3] 导出了 GM 参数 MLE 的 CEM 算法，省略中间步骤，主要迭代公式如(5)所示。

$$p(m|x_n)=\frac{\varepsilon_m\varphi_m(x_n)}{f_M(x_n)} \tag{5-a}$$

$$\varepsilon_m=\frac{1}{N}\sum_{n=1}^{N}p(m|x_n) \tag{5-b}$$

$$\mu_m=\frac{\sum_{n=1}^{N}p(m|x_n)x_n}{\sum_{n=1}^{N}p(m|x_n)} \tag{5-c}$$

$$\sigma_m^2=\frac{\sum_{n=1}^{N}p(m|x_n)(x_n-\mu_m)^2}{\sum_{n=1}^{N}p(m|x_n)} \tag{5-d}$$

这里，$p(m|x_n)$ 为样本 x_n 来自第 m 个高斯源 G_m 的先验概率，这实质上是一个关于 $\boldsymbol{z}_n$ 的期望数值。

CEM 迭代算法流程如图 2 所示。

可以看到，在假定模型阶数和初始化模型参数后，即可使用式(5)更新参数估计，直到满足令式(4)所示似然函数积分值取得最大。这就是 CEM 迭代算法的基本思想[3]。

在文献[5]中提出的多初值 CEM 方案如图 3 所示。依据图示方案实施估计，可以大大降低迭代收敛于局部极值点的错误概率。

3 GEM 算法

GEM 算法的理论依据是，一个混合分布的似然函数最大化过程可以通过一种所谓“贪婪”吸附

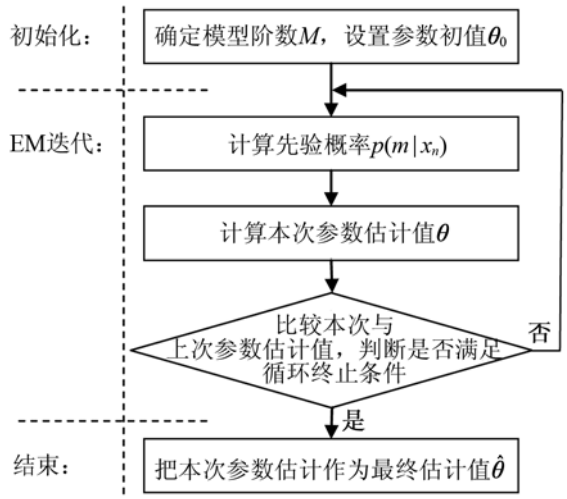


图 2　CEM 估计算法流程图

Fig.2　Flow chart of CEM algorithm

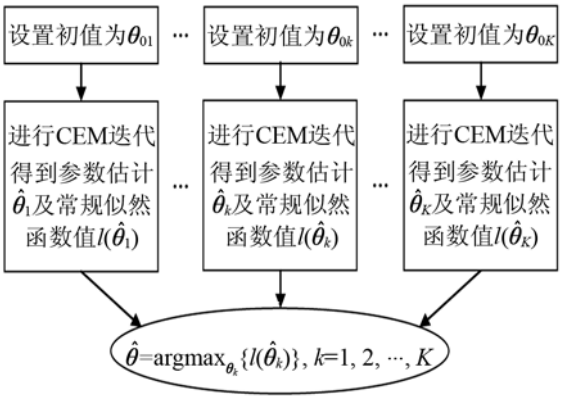


图 3　多初值的 CEM 算法

Fig.3　Block diagram of multi-initialization CEM algorithm

的方式完成，即依据一定的规则，相继向混合分布中增加新的高斯成员，直至最终完成数据的 PDF 拟合任务。

设现有 k 阶 GM 模型的 PDF 为 $f_k(x_n)$，以概率 α 新加入第 k+1 个 PDF 为 $\varphi(x_n,\beta_{k+1})$ 的高斯源 G_{k+1}，则新的 k+1 阶 GM 模型的 PDF 为

$$f_{k+1}(x_n)=(1-\alpha)f_k(x_n)+\alpha\varphi(x_n,\beta_{k+1}) \tag{6}$$

这里：α 即高斯源 G_{k+1} 的混合权系数；$\beta_{k+1}=(\mu_{k+1},\sigma_{k+1}^2)$ 为 G_{k+1} 的 PDF 参数。这样，对于每一个 k 和 $f_k(x_n)$，只要合理选择新加入源 G_{k+1} 的参数 α 和 β_{k+1}，确保能够使得下述新的对数似然函数积分值取得最大即可。

$$l_{k+1}(\theta)=\sum_{n=1}^{N}\ln f_{k+1}(x_n)=\sum_{n=1}^{N}\ln[(1-\alpha)f_k(x_n)+\alpha\varphi(x_n,\beta_{k+1})] \tag{7}$$

不难发现，不同于 CEM 算法在起始时即对 GM 所有可能源成份做出初始配置，GEM 算法是从一阶最佳 GM 拟合(即一个最佳高斯源)开始的，其初始化是根据数据通过计算完成的。接下来，即重复如下两步，直到满足循环停止条件(比如达到了预定模型阶数)：

(1) 插入一个新的源成份；

(2) 应用 EM 迭代直到收敛。

GEM 算法每次只对两个分量进行处理，这样就减少了对初始值的依赖。但对于每一次加入的高斯分量 $\varphi(x_n,\beta_{k+1})$，需要寻找使对数似然函数积分 $l_{k+1}(\theta)$ 最大的参数 α 和 β_{k+1}，以此作为该高斯分量的参数。

根据后验概率密度 $p(m|x_n)$，把给定的样本数据 $\boldsymbol{x}$ 分成 k 个互不相交的子集 $A_i(1\le i\le k)$，其中，$A_i=\{x_n\in\boldsymbol{x}: p(i|x_n)=\max\{p(m|x_n)|_{m=1}^{M}\}\}$。为了提高收敛精度，可以对各子集 A_i 再进行分块。方法是在 A_i 中均匀地选取两个数 x_a、x_b，把 A_i 分成互不相交的两个子块 A_{i1}、A_{i2}，其中，A_{i1} 中元素与 x_a 的距离更近，A_{i2} 中元素与 x_b 的距离更近。各子块的均值和方差作为新加入高斯分量的初始参数，权系数设为 $\alpha_i/2$。运用同样的方法，还可对 A_{i1}、A_{i2} 再进行分块，直至达到所需要求。本文将各 A_i 分成两子块，收敛精度较好，符合要求。

有了初始化高斯参数和权系数，用式(8)对参量进行更新(此即 partial EM 迭代算法，简记为 pEM)，选择使对数似然函数最大的一组参数作为新加入高斯分量的参数。

$$p(k+1|x_n)=\frac{\alpha_i\varphi(x_n,\beta_{k+1})}{(1-\alpha_i)f_k(x_n)+\alpha_i\varphi(x_n,\beta_{k+1})} \tag{8-a}$$

$$\alpha_{k+1}=\frac{\sum_{x_n\in A_{ij}}p(k+1|x_n)}{N_{ij}} \tag{8-b}$$

$$\mu_{k+1}=\frac{\sum_{x_n\in A_{ij}}p(k+1|x_n)x_n}{\sum_{x_n\in A_{ij}}p(k+1|x_n)} \tag{8-c}$$

$$\sigma_{k+1}^2=\frac{\sum_{x_n\in A_{ij}}p(k+1|x_n)(x_n-\mu_{k+1})^2}{\sum_{x_n\in A_{ij}}p(k+1|x_n)} \tag{8-d}$$

式中，A_{ij} 表示对子集 A_i 进行分块产生的子块，N_{ij} 表示 A_{ij} 中元素的个数。

对给定的样本序列 $\boldsymbol{x}$，上述思路的 GM 模型参数 GEM 估计算法流程如图 4 所示。

4　仿真实例

使用复合抽样法[7]，产生一个 3 阶 GM 序列 $\boldsymbol{x}$，其参数为 $\boldsymbol{\theta}=[0.4,\ 0.2,\ 0.4;\ -5,\ 0,\ 5;\ 9,\ 1,\ 9]^{\mathrm{T}}$，样本点数 N=2000。样本波形和根据设置值 $\boldsymbol{\theta}$ 绘出的理

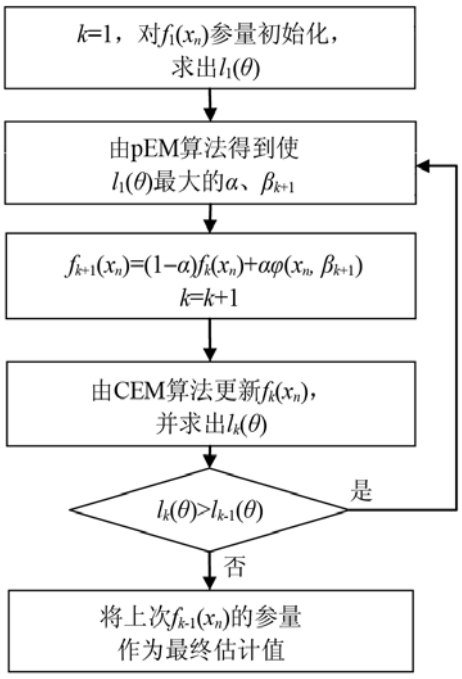
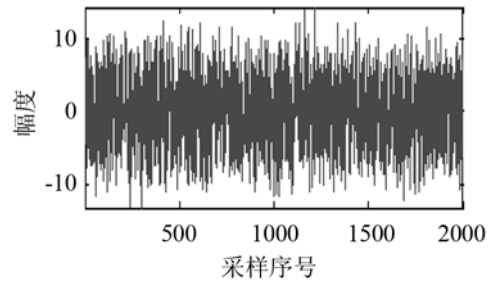


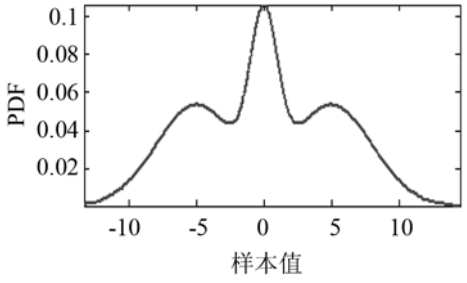
图 4 GEM 算法流程图

Fig.4 Flow chart of GEM algorithm

论 PDF 曲线如图 5 所示。可以看到，这是一种强非高斯情形。



(a) 样本波形



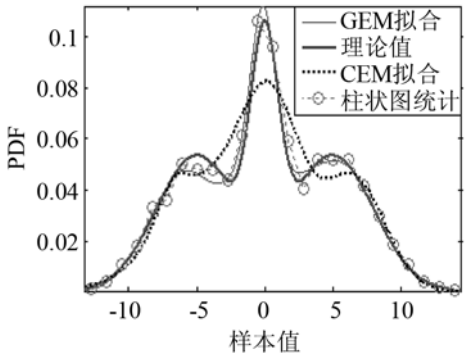
(b) 理论 PDF 曲线

图 5 数值仿真实例的波形和 PDF

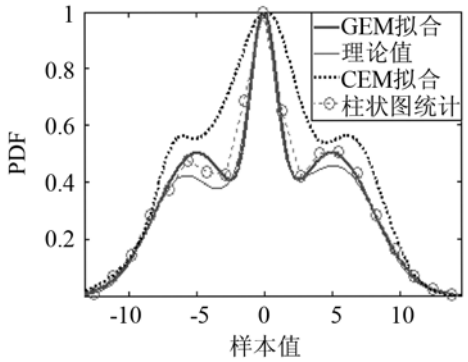
Fig.5 True description of the instance

分别使用 CEM 和 GEM 算法对此序列进行 GM 参数估计。此例中，预设的 GM 模型阶数为 3。但 GEM 算法估计的模型阶数为 6。在实际应用中，这种现象很常见，因为一条复杂形状的曲线可以有多种组成分解方式。为了与 GEM 估计结果比较，例中也设定 CEM 阶数为 6。应当指出，由于阶数估计值不同于预设值，所以难以从参数估计值 $\hat{\boldsymbol{\theta}}$ 和设定值 $\boldsymbol{\theta}$ 的直接对比中得到性能描述。但可以据此绘出 PDF 曲线，然后通过曲线比较进行性能直观比对。

图 6 给出了根据这两种方法得到的参数估计绘制出的 PDF 曲线(分别标记为 CEM、GEM)对比。为了便于比较，图中同时给出了理论 PDF 曲线和根据柱状图统计(这是一种常用的、适合于大样本量的分布统计分析方法：平分样本值域为若干连续区间，统计取值落于这些区间内的点数，以此可得到样本值的大致分布)得到的 PDF 曲线。于是，每幅图中共有 4 条 PDF 曲线。图 6(a)中显示 PDF 原值，为了方便观察各条曲线的区别，图 6(b)中给出了它们经过最大值归一化后的对比情形。



(a) 原始 PDF 值



(b) 最大值归一化的 PDF 值

图 6 PDF 曲线比较

Fig.6 Comparison of PDF curves for four methods

由图 6 可见，GEM 算法得到的 PDF 曲线与理论 PDF 曲线重合程度远高于 CEM 算法，这说明前者 PDF 拟合性能明显优于后者。在本例这种样本量较为充足的条件下，GEM 拟合与柱状图统计 PDF 拟合性能相当。

5 结 语

GM 模型是对数据进行非高斯 PDF 建模的有效

模型之一。CEM 算法是 GM 模型参数估计的常用方法，但 CEM 不能估计模型阶数，估计精度受初始值设置影响严重。而 GEM 算法则可以克服这些缺点，但运算量较大。

本文系统地梳理了 CEM 和 GEM 算法的关系及各自优缺点，并基于同一仿真实例对其性能进行了直观比较演示。比较结果表明：GEM 算法可以取得优于 CEM 算法的 PDF 拟合性能，且能自动估计 GM 模型阶数。

由于通过不断新增高斯分量的方式逼近最大似然估计，所以 GEM 算法不受初值设置影响(即使最初的初值设置不合理，它也可以通过后续新增高斯分量的方式予以弥补)，可以取得较为稳定的估计性能。但是每一次新增高斯分量，都会相应增加运算量，所以 GEM 算法运算量往往远大于 CEM 算法。下一步将重点研究如何把 GM 阶数限定在一个较小范围内应用 GEM 算法对水声混响数据进行高效的 PDF 建模，并对 CEM 和 GEM 算法估计精度和速度等性能做出统计性比较，以最终判定究竟哪种方法更适用于水声混响数据的统计建模。

参 考 文 献

[1] Steven M Kay. Fundamentals of statistical signal processing: Estimation theory[M]. New Jersey, USA: Prentice Hall, 1998: 326.

[2] Redner R A, Walker H F. Mixture densities, maximum likelihood, and the EM algorithm[J]. SIAM Review, 1984, **26**(2): 195-202.

[3] Aaron A D. Using EM to estimate a probability density with a mixture of Gaussians[DB/OL]. http://citeseer.ist.psu.edu, 2000.

[4] Verbeek J J, Vlassis N, Krose B. Efficient Greedy Learning of Gaussian Mixture Models[R]. Netherlands: Reports of Computer Science Institute of Amsterdam Univ, 2001: 153.

[5] 王平波，蔡志明，刘旺锁. 混合高斯概率密度模型参数的 EM 估计[J]. 声学技术, 2007, **26**(3): 498-502.
WANG Pingbo, CAI Zhiming, LIU Wangsuo. EM Estimation of PDF Parameters for Gaussian Mixture Processes[J]. Technical Acoustics, 2007, **26**(3): 498-502.

[6] 卫红凯，王平波，蔡志明. 混响数据的混合高斯建模研究[J]. 声学技术, 2007, **26**(3): 514-518.
WEI Hongkai, WANG Pingbo, CAI Zhiming. Study of reverberation for gaussian mixture model[J]. Technical Acoustics, 2007, **26**(3): 514-518.

[7] 刘旺锁，王平波，顾雪峰，等. 一种非白非高斯数据的数值仿真方法[J]. 声学技术, 2013, **32**(3): 228-232.
LIU Wangsuo, WANG Pingbo, GU Xuefeng, et al. A simulation approach to colored non-Gaussian processes[J]. Technical Acoustics, 2013, **32**(3): 228-232.

上海章奎生声学工程顾问有限公司近日在沪揭牌成立

2014 年 11 月 19 日在上海大连路 1619 号骏丰国际财富广场 23 楼隆重举行了上海章奎生声学工程顾问有限公司揭牌开业聚会。公司新办公室阳光明媚，专家朋友济济一堂，热烈祝贺上海章奎生声学工程顾问有限公司开业。到场祝贺的有上海市声学学会秘书长龚农斌教授级高工、上海市演出行业协会副会长蔡健勇博士、上海交通大学博士生导师陈端石教授、上海第九设计研究院刘利民教授级高工、上海戏剧学院柳得安教授、安恒利扩声技术工程有限公司项珏总经理和陈永坚副总、上海电影制片厂国家一级录音师任大铭先生和吴国强先生、上海市舞台研究所周建国所长、上海歌舞团著名调音师及国家一级录音师史汇荣先生以及上海演艺设备行业、灯光音响界、舞台美术界的众多专家、企业界的朋友石敏先生、曹益群先生、马为民先生、严雷先生、谢咏冰先生、池文忠先生，还有沪上声学装备行业的许多专家和企业家如张明发先生、蒋彩荣先生也都到场祝贺上海章奎生声学工程顾问有限公司的成立和开业。

回顾 14 年前的 2000 年 7 月，在上海现代建筑设计集团有限公司有关领导的支持和指导下，年届 60 多岁的章奎生教授级高工领衔并主持成立了章奎生声学设计研究所。当时是集团单位内的第一个名人设计所。10 多年来既经历了市场的考验，也遇到了我国文化演艺建筑大发展的黄金年代。10 多年来章奎生声学设计研究所的业务稳步上涨，技术不断发展，效益逐年提高。10 余年间承担并完成了大量剧院和音乐厅、大会堂及体育馆、广电广播大楼、录音棚及电影院工程等，取得了显著的业绩。项目遍及我国 20 多个省、区及 40 多个大中城市，并改变了 90 年代境外声学工程师占领我国大多演艺建筑工程建声专业设计市场的不正常现象。章奎生声学设计研究所凭借先进的声学专业设计技术、众多的声学工程业绩、优良的音质效果和周到的配合服务，取得了业主的信任，才占领了我国演艺建筑声学设计的大半市场份额。如 2013 年国内评出十大优秀剧院中由章所承担建声专业设计或建声顾问的剧院就达六座之多。

2014 年 10 月起，年近八旬的章奎生教授再次牵头领衔与具有近 10 年建声设计和科研经验的宋拥民博士和冯善勇博士及已在章奎生声学设计研究所工作多年的余斌硕士联手创建成立上海章奎生声学工程顾问有限公司。这是一个新的起点，是一个由老中青专业技术人员组建的有业绩、有实力、有技术和知名度的小型高科技民营企业，并已通过工商管理局的注册批准。相信新公司一定会取得新的业绩和新的发展，创建出一片新的事业。

上海章奎生声学工程顾问有限公司　章奎生