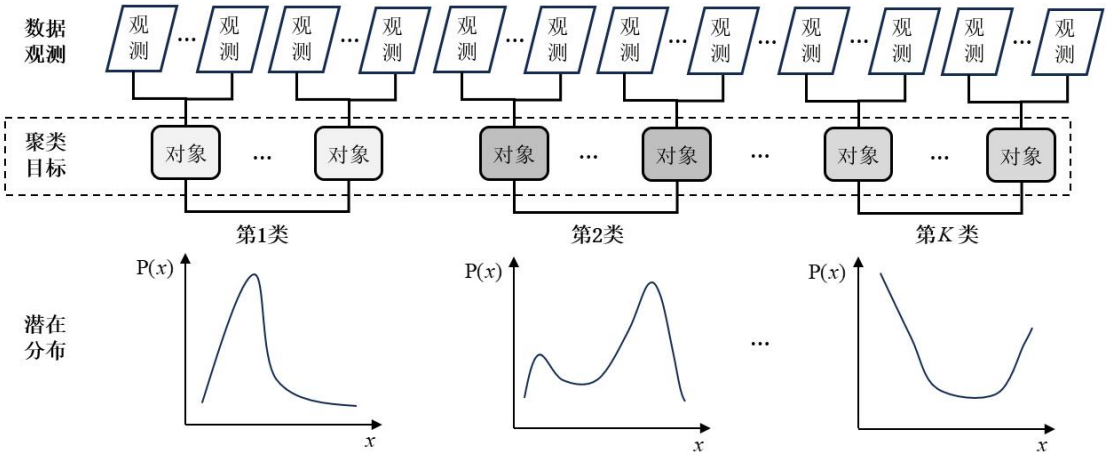
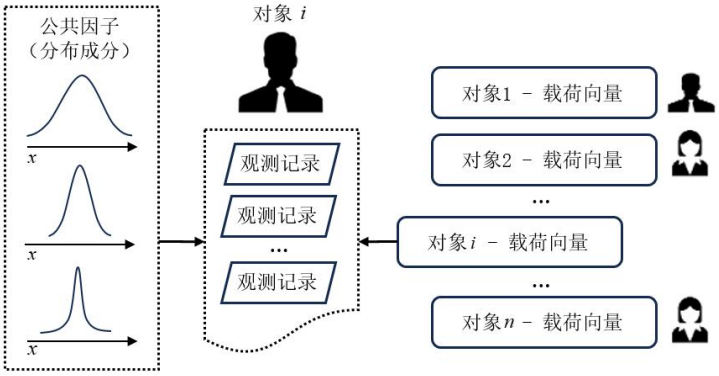
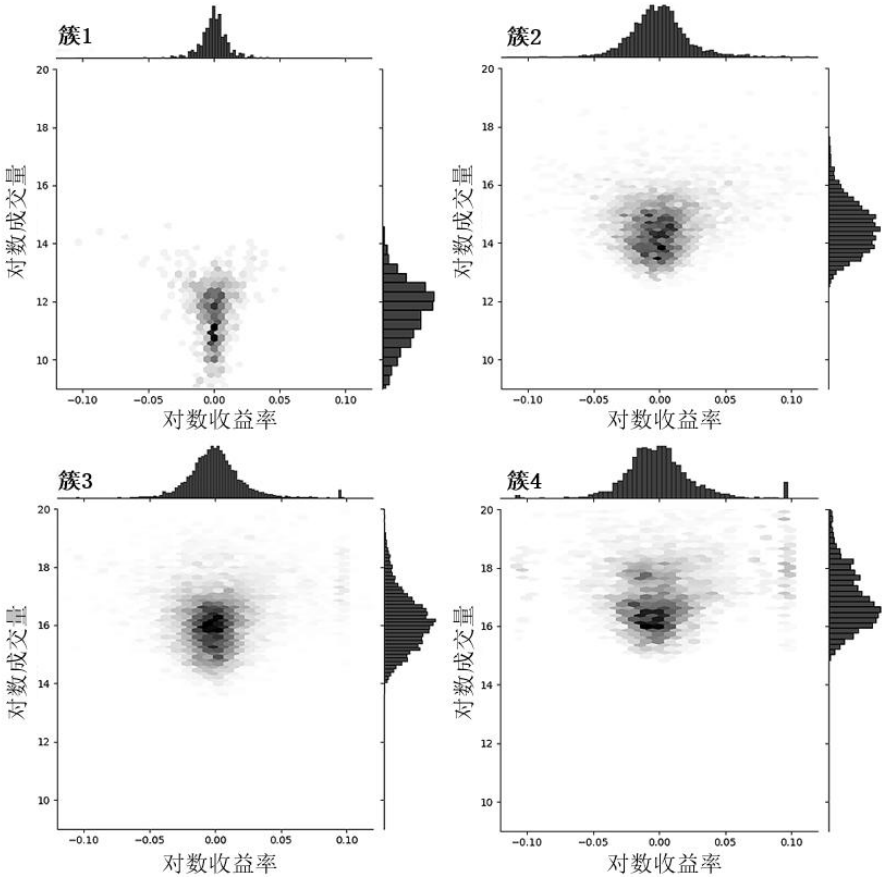
基于高斯混合模型的分布因子聚类方法—附录



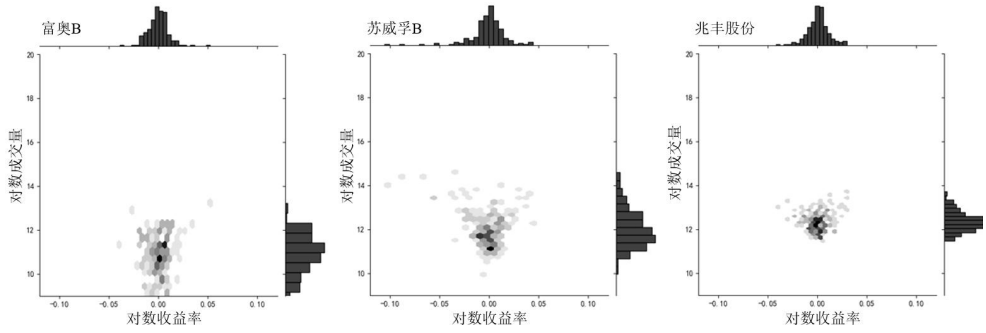
附图 1 分布函数聚类数据结构



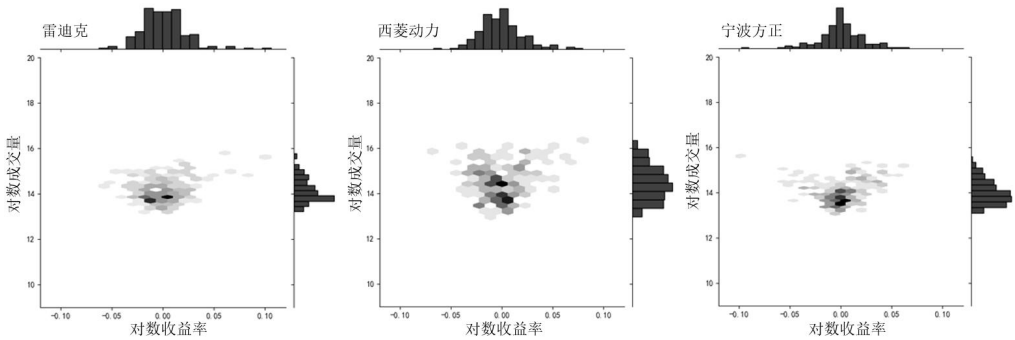
附图 2 分布因子模型示意图



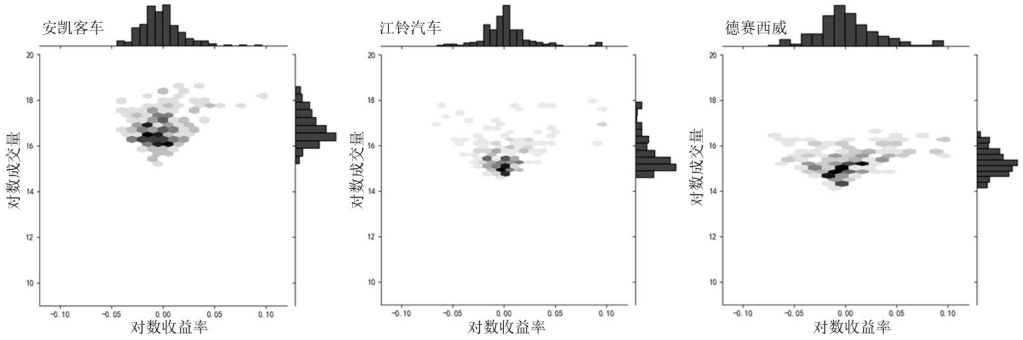
附图 3 股票数据中 DFM 方法输出各个簇的联合分布



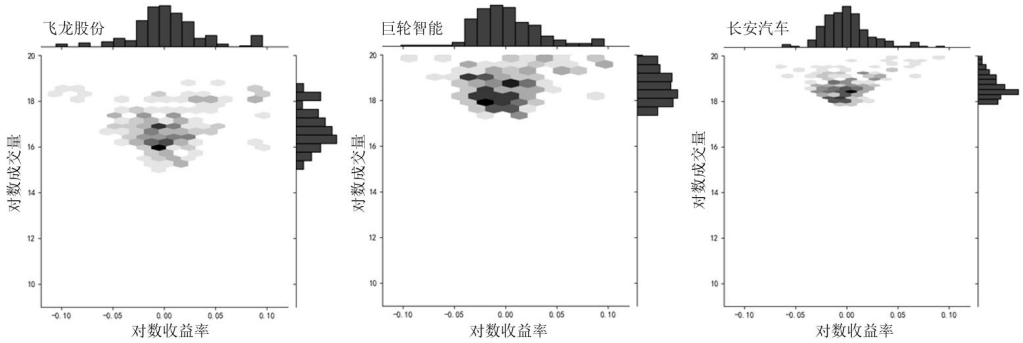
附图 4 簇 1 中典型股票的分布



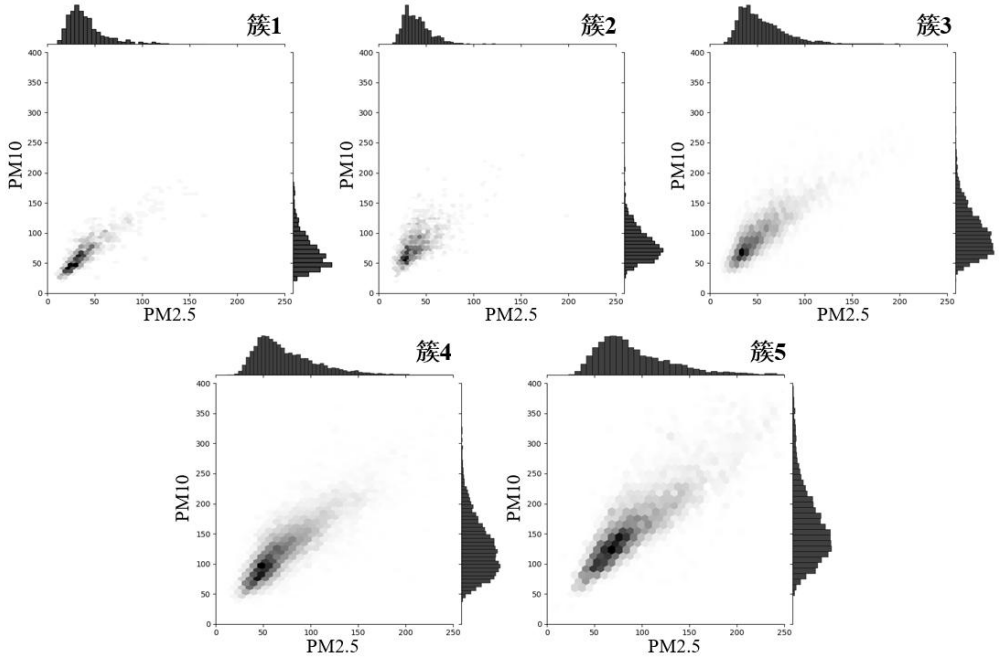
附图 5 簇 2 中典型股票的分布



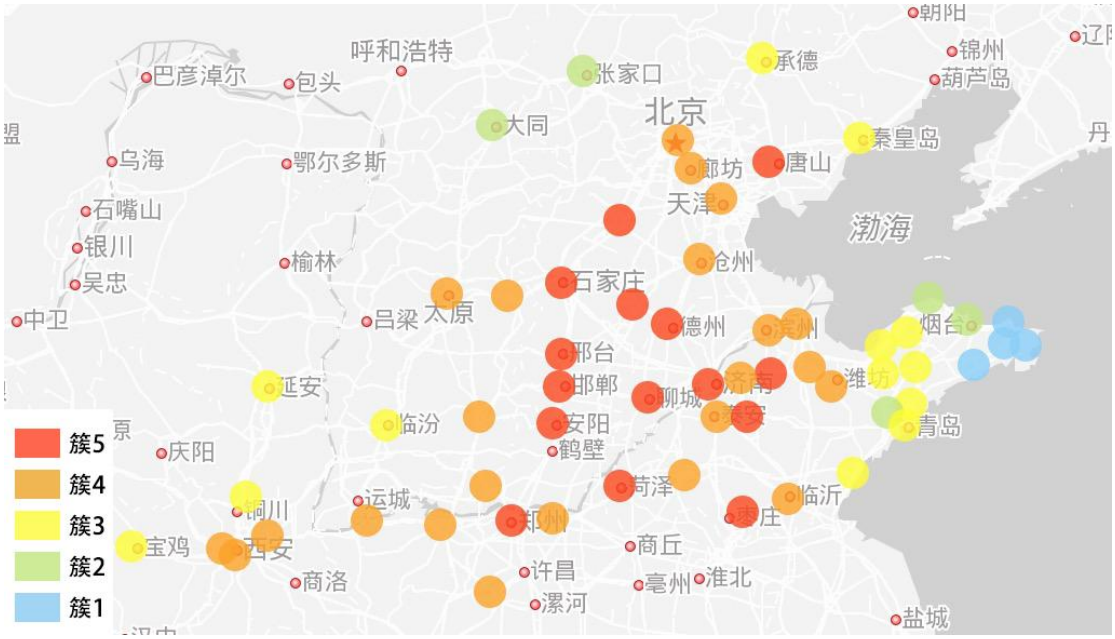
附图 6 簇 3 中典型股票的分布



附图 7 簇 4 中典型股票的分布



附图 8 大气污染数据中 DFM 方法输出各个簇的联合分布



附图 9 不同簇中城市的地理位置分布

（一）附录 1：DFM 方法具体算法流程

算法 1：DFM 聚类初始化步骤

输入：n 个载荷向量：$\hat{\boldsymbol{\varphi}}_i, i=1,\cdots,n$；聚类个数：$K$。

输出：一组初始簇中心：$\mathrm{C}^{(0)}=\{C_1^{(0)},\cdots,C_K^{(0)}\}$。

1. 随机选择 1 个对象的载荷作为第一个初始簇中心 $C_1^{(0)}$，$\mathrm{C}^{(0)} \leftarrow \{C_1^{(0)}\}$；
2. **for** $k=2,\cdots,K$ **do**
3. 对 $\{\hat{\boldsymbol{\varphi}}_i : i=1,\cdots,n\}$，根据概率 p_i 选取一个 $\mathrm{C}^{(0)}$ 外的载荷向量 $\hat{\boldsymbol{\varphi}}_i$，其中

$$p_i=\frac{D_m(i,\mathrm{C}^{(0)})}{\sum_{i\notin C^{(0)}} D_m(i,\mathrm{C}^{(0)})};$$

4. 将 $\hat{\boldsymbol{\varphi}}_i$ 作为新的初始簇中心 $C_k^{(0)}$ 添加到 $\mathrm{C}^{(0)}$，$\mathrm{C}^{(0)} \leftarrow \mathrm{C}^{(0)} \cup \{C_k^{(0)}\}$；
5. **end for**

算法 2：DFM 聚类算法流程

输入：n 个对象的观测：$\{\mathrm{M}_i : i=1,\cdots,n\}$；聚类个数：$K$；高斯成分个数：$G$；DFM 参数估计最大迭代次数：$T_0$；聚类划分最大迭代次数：$T_1$。

输出：聚类划分结果：$\mathrm{S}_1,\cdots,\mathrm{S}_K$，其中 S_k 为划分到第 k 个簇的对象序号组成的集合。

1. 对 $g=1,\cdots,G$，产生随机的 $\boldsymbol{\mu}_g^{(0)},\boldsymbol{\Sigma}_g^{(0)}$。对 $i=1,\cdots,n$，产生随机的 $\boldsymbol{\varphi}_i^{(0)}$；
2. $t \leftarrow 0$；
3. **repeat**
4. 根据式（3）得到 $\{\hat{\psi}_{ijg} : i=1,\cdots,n; j=1,\cdots,V_i; g=1,\cdots,G\}$；
5. **for** $g=1,\cdots,G$ **do**
6. 根据式（4）、式（5）更新得到 $\boldsymbol{\mu}_g^{(t+1)}$ 和 $\boldsymbol{\Sigma}_g^{(t+1)}$；

7. **for** $i=1,\cdots,n$ **do**
8. 根据式（6）更新得到 $\varphi_{ig}^{(t+1)}$；
9. **end for**
10. **end for**
11. $t \leftarrow t+1$；
12. **until** 划分不发生变化 **or** $t > T_0$；
13. 以估计结果 $\{\boldsymbol{\varphi}_i^{(T_0)} : i=1,\cdots,n\}$ 作为对载荷向量的估计 $\{\widehat{\boldsymbol{\varphi}}_i : i=1,\cdots,n\}$；
14. 根据算法 1 确定初始簇中心 $\mathrm{C}^{(0)} = \{C_1^{(0)},\cdots,C_K^{(0)}\}$；
15. $t \leftarrow 0$；
16. **repeat**
17. **for** $k=1,\cdots,K$ **do**
18. $\mathrm{S}_k^{(t)} \leftarrow \varnothing$；
19. **end for**
20. **for** $i=1,\cdots,n$ **do**
21. $k_i \leftarrow \arg\min_k \left\| \widehat{\boldsymbol{\varphi}}_i - C_k^{(t)} \right\|$；
22. 将对象 i 分配至 $\mathrm{S}_{k_i}^{(t)}$，$\mathrm{S}_{k_i}^{(t)} \leftarrow \mathrm{S}_{k_i}^{(t)} \cup \{i\}$；
23. **end for**
24. **for** $k=1,\cdots,K$ **do**
25. $C_k^{(t+1)} \leftarrow \left|\mathrm{S}_k^{(t)}\right|^{-1} \sum_{i\in \mathrm{S}_k^{(t)}} \widehat{\boldsymbol{\varphi}}_i$；
26. **end for**
27. $t \leftarrow t+1$；
28. **until** 划分不发生变化 **or** $t > T_1$；
29. 将最终聚类划分得到的 $\mathrm{S}_1^{(t)},\cdots,\mathrm{S}_K^{(t)}$ 作为结果 $\mathrm{S}_1,\cdots,\mathrm{S}_K$ 返回；

（二）附录 2：DFM 方法计算复杂度的证明

证明：根据算法 2，可参照循环的操作次数，直接得出算法各个环节的计算复杂度：

（1）对载荷矩阵的估计：迭代估计中 E-step 需要的操作次数为 $O(T_0 nBG)$，M-step 需要的操作次数为 $O(T_0 G)$。由于 T_0、B、G 均为给定常数，因此这一环节计算复杂度为 $O(n)$；

（2）聚类初始状态选择：根据算法 1，该环节需要的操作次数为 $O(Kn)$，给定聚类个数 K，其计算复杂度为 $O(n)$；

（3）聚类划分：迭代聚类的操作次数为 $O(T_1 nK)$ 和 $O(T_1 K)$，由于 T_1 为给定常数，该环节计算复杂度为 $O(n)$。

综上，DFM 聚类方法的计算复杂度为 $O(n)$。

（三）附录 3：引理 1 的证明

证明：由于欧氏距离满足三角不等式，可得

$$\left\|\boldsymbol{\varphi}_i-\boldsymbol{\varphi}_j\right\|\le\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_j\right\|+\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|$$

从而有

$$\begin{aligned}\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_j\right\|&\ge\left\|\boldsymbol{\varphi}_i-\boldsymbol{\varphi}_j\right\|-\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|\\&\ge\left\|\boldsymbol{\pi}_{\gamma_i}-\boldsymbol{\pi}_{\gamma_j}\right\|-\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|\\&\ge\pi_{\min}-2^{-1}\pi_{\min}\ge2^{-1}\pi_{\min}>\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|\end{aligned}$$

引理 1 得证。

（四）附录 4：定理 1 的证明

证明：根据式（6），对$1\le g\le G$，有$\widehat{\varphi}_{ig}=V_i^{-1}\sum_{j=1}^{V_i}\widehat{\psi}_{ijg}$，其中$\widehat{\psi}_{ijg}$为迭代结束时$\psi_{ijg}$的条件期望。对每个$\widehat{\psi}_{ijg}$，可得其期望为$E(\widehat{\psi}_{ijg})=\varphi_{ig}$，方差为：

$$\mathrm{var}(\widehat{\psi}_{ijg})=E(\widehat{\psi}_{ijg}^2)-\left[E(\widehat{\psi}_{ijg})\right]^2=E(\widehat{\psi}_{ijg})-\varphi_{ig}^2=\varphi_{ig}-\varphi_{ig}^2$$

由于对任意对象i，关于V_i个观测的$\{\widehat{\psi}_{ijg}:j=1,\cdots,V_i\}$独立同分布，因而$\widehat{\varphi}_{ig}$的期望为$E(\widehat{\varphi}_{ig})=\varphi_{ig}$,方差为$\mathrm{var}(\widehat{\varphi}\ )=V^{-2}\sum_{j=1}^{V_i}\mathrm{var}(\widehat{\psi}_{\ g})=V^{-1}(\varphi\ -\varphi^2)$。由于$\varphi_{ig}\in[0,1]$，可得$\varphi_{ig}-\varphi_{ig}^2\le4^{-1}$。并且$V_i^{-1}\le v_{\min}^{-1}$。从而对任意的$i$和$g$，有$\mathrm{var}(\widehat{\varphi}_{ig})\le(4v_{\min})^{-1}$。记$\mathrm{X}=\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|^2$，有

$$E(\mathrm{X})=E\left[\sum_{i=1}^{n}\sum_{g=1}^{G}(\widehat{\varphi}_{ig}-\varphi_{ig})^2\right]=\sum_{i=1}^{n}\sum_{g=1}^{G}\mathrm{var}(\widehat{\varphi}_{ig})\le\frac{nG}{4v_{\min}}$$

根据马尔可夫不等式，对$\varepsilon>0$，有

$$P(\mathrm{X}\ge\varepsilon)\le\frac{E(\mathrm{X})}{\varepsilon}\le\frac{nG}{4\varepsilon v_{\min}}$$

带入$\varepsilon=nG\ln n/(4v_{\min})$，可得

$$P\left(\mathrm{X}\ge\frac{nG\ln n}{4v_{\min}}\right)\le\frac{1}{\ln n}$$

该式等价于

$$P\left(\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|^2<\frac{nG\ln n}{4v_{\min}}\right)>1-\frac{1}{\ln n}$$

进而对 DFM 方法的聚类误差率，可得

$$\begin{aligned}P_e&=\frac{1}{n}\sum_{i=1}^{n}I\left(\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|^2>\frac{\pi_{\min}^2}{4}\right)\\&\le\frac{4}{n\pi_{\min}^2}\sum_{i=1}^{n}I\left(\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|^2>\frac{\pi_{\min}^2}{4}\right)\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|^2\\&\le\frac{4}{n\pi_{\min}^2}\sum_{i=1}^{n}\left\|\widehat{\boldsymbol{\varphi}}_i-\boldsymbol{\varphi}_i\right\|^2\le\frac{G\ln n}{v_{\min}\pi_{\min}^2}\end{aligned}$$

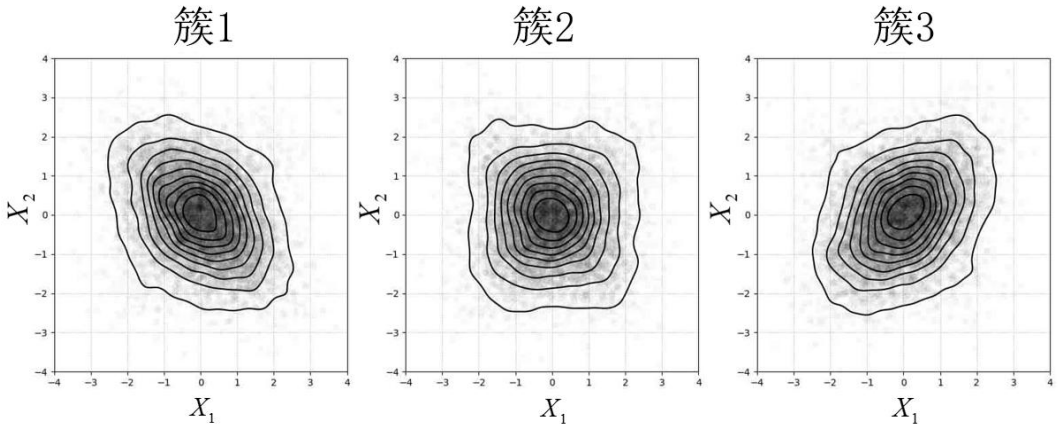
以不小于$1-\ln^{-1} n$的概率成立。同时根据假设 4，存在常数$c_0>0$和正整数$n_0>0$，对$n\geq n_0$，有$v_{\min}\geq c_0\ln^2 n$。因此，对$n\geq n_0$，不等式右侧可进一步化为

$$P_e\leq\frac{G\ln n}{v_{\min}\pi_{\min}^2}\leq\frac{G}{c_0\pi_{\min}^2\ln n}=\frac{1}{c_0\bar{\pi}_{\min}^2\ln n}$$

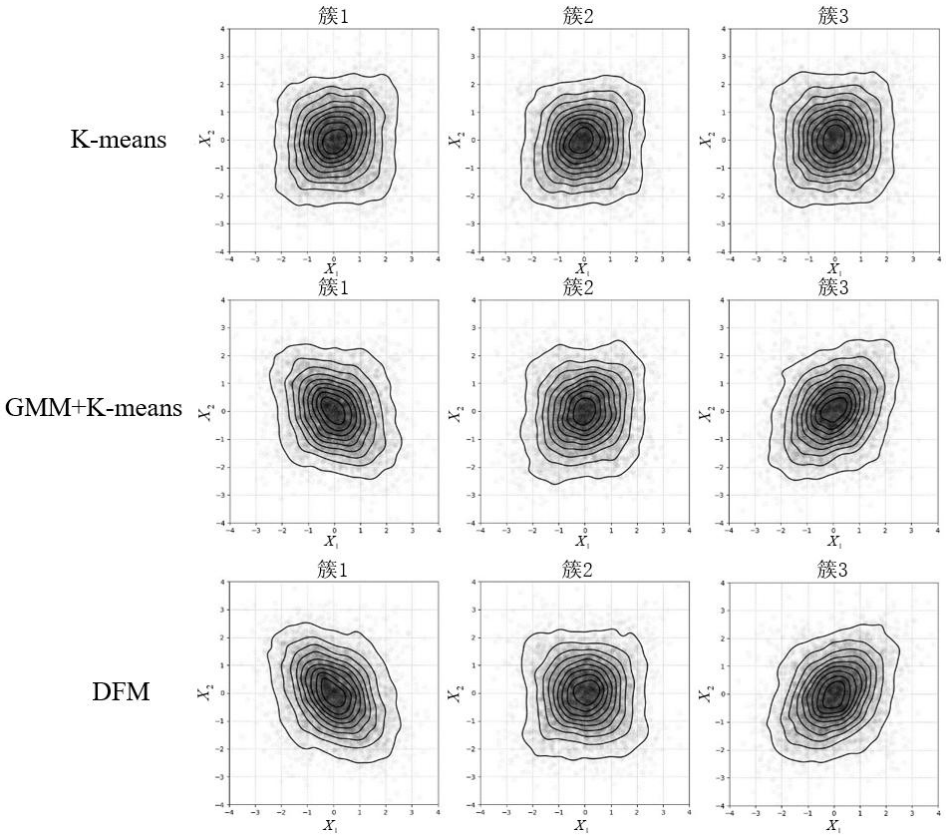
定理 1 得证。

（五）附录 5：模拟实验中簇分布的可视化展示

附图 3 以样本规模设置 1，$\beta=100$为例，展示了模拟生成数据的真实分布，可作为聚类结果的参照标准。附图 4 以样本规模设置 2，$\beta=200$为例，可视化展示各聚类方法输出的簇在经验分布上的对比。对比模拟设置的真实分布，DFM 方法在实验的三个方法中表现最好，得到的簇则区分度较高，和真实模拟分布非常接近。而 K-means 方法难以区分不同的分布模式，得到的簇在经验分布上相互难以辨认。



附录 5-图 1 模拟设置的数据分布



附录 5-图 2 不同方法在模拟数据上聚类结果的经验分布