

# DPM

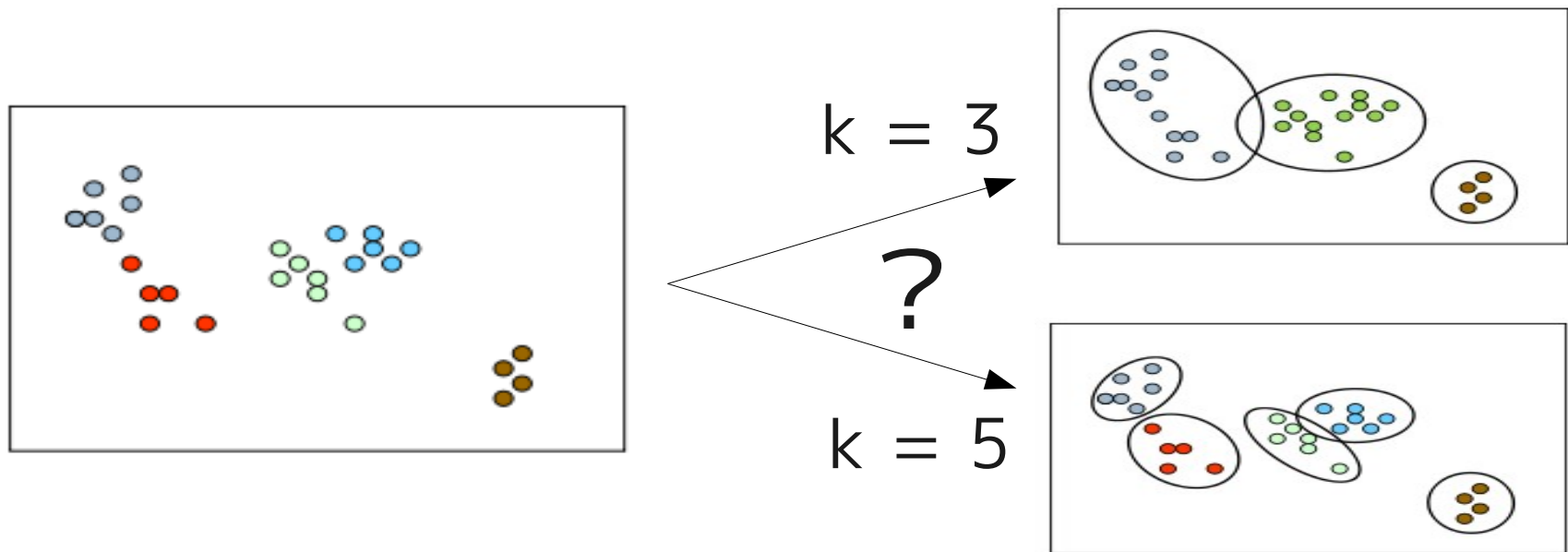
- DPM = Dirichlet Process Mixture  
= 混合ディリクレ過程
- ベイズ統計を用いたクラスタリング手法
- 発端となった研究は古い

Thomas S. Ferguson. "A Bayesian Analysis of Some Nonparametric Problems" 1973

→ IT の発展に伴い再注目

# Motivation

- ある分布をクラスタリングしたいが、k-means のような手法では  $k$  の値によって結果が大きく変わる。



DPM を用いれば、統計的により尤もらしい  $k$  を自動で判断してくれる。

# Dirichlet Process

- Dirichlet分布

- クラスタリングを表現する確率分布を値とする確率分布
- $k$  面サイコロ生成器とも言える。

サイコロは、

- 各面の出る確率は同等でない、歪んでいる。
  - 各面の値が
    - クラスタ ID および
    - そのクラスタを表現する確率分布のパラメータ
- に対応

→ Dirichlet Process とは任意面サイコロ生成器

# Dirichlet Process Mixture

- 実際にDPから無限面のサイコロが得られると仮定し、データを生成するモデル
- Chinese Restaurant Process

DPMの構成法の一つで、 $k$  の値は中華料理店の客が座っているテーブル数になぞらえて説明される。

- 客は着座数が多いテーブルに着座しやすい。

# 実装アルゴリズム

- 所与のデータ分布に対して、どのようなサイコロが尤もらしいか判断するために、何回もサイコロをサンプリングする。
  - 効率の良いサンプリング法が必要
- Markov Chain Monte Carlo

マルコフ過程(現在の事象は直前の事象にのみ影響を受ける)であるような確率過程のサンプリング法の総称

  - Gibbs sampler
    - CRPと相性がよい
    - 実装が容易
    - 他のアルゴリズム ( 変分ベイズ法 ) と比較して、同等かそれ以上の精度

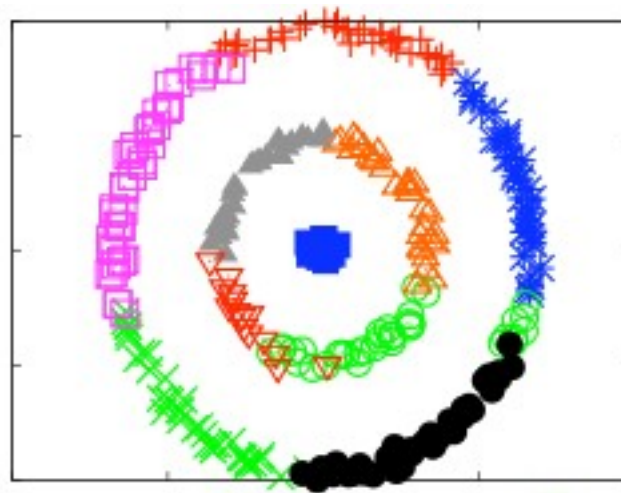
# DPMの問題点(1)

- 要素分布の選択

DPMを用いる前提として、データがどのような確率分布から生成されたか(サイコロの各面に対応する確率分布)について事前知識がないと、正しい  $K$  の推定は行えない。

- 要素分布が妥当でない例

同心円上のデータ点に対し、混合正規分布を選択。一つの円上の点を複数のクラスタに分割してしまっている。



# DPMの問題点(2)

- 事前分布が制限される

前述の問題故、事前分布は様々なものが適用できなければならないが、アルゴリズムの制約上、共役事前分布に限られる。

→ ディリクレ分布、指数型分布族など

- $K$  の値がばらつく

理論的には無限回 MCMC を繰り返せば収束するはずであるが、時間は限られている。

→ Dirichlet Process のパラメータから収束性を判断することが考えられる。