Bayesian Repulsive Gaussian Mixture Model

Fangzheng Xie and Yanxun Xu

Department of Applied Mathematics and Statistics, Johns Hopkins University
3400 North Charles Street Baltimore Maryland 21218 USA
fxie5@jhu.edu yanxun.xu@jhu.edu

1. Overview

We develop a general class of Bayesian repulsive Gaussian mixture models that encourage well-separated clusters, aiming at reducing potentially redundant components produced by independent priors for locations (such as the Dirichlet process). The asymptotic results for the posterior distribution of the proposed models are derived, including posterior consistency and posterior contraction rate in the context of nonparametric density estimation. More importantly, we show that compared to the independent prior on the component centers, the repulsive prior introduces additional shrinkage effect on the tail probability of the posterior number of components, which serves as a measurement of the model complexity. In addition, an efficient and easy-to-implement blocked-collapsed Gibbs sampler is developed based on the exchangeable partition distribution and the corresponding urn model. We evaluate the performance and demonstrate the advantages of the proposed model through extensive simulation studies and real data analysis.

2. Bayesian Repulsive Mixture Model

The hierarchical Bayesian mixture model:

$$
\begin{aligned}
(f(\boldsymbol{y}) \mid F) &= \int_{\mathbb{R}^p\times\mathcal{S}} \phi(\boldsymbol{y}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{d}F(\boldsymbol{\mu},\boldsymbol{\Sigma}),\\
(F\mid K, w_{1:K},\boldsymbol{\mu}_{1:K},\boldsymbol{\Sigma}_{1:K}) &= \sum_{k=1}^K w_k\delta_{(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)},\\
(\boldsymbol{\mu}_{1:K},\boldsymbol{\Sigma}_{1:K}\mid K) &\sim p(\boldsymbol{\mu}_{1:K},\boldsymbol{\Sigma}_{1:K}\mid K),\\
(w_{1:K}\mid K) &\sim \mathcal{D}_K(\beta),\\
K &\sim p_K(K).
\end{aligned}
$$

The joint prior on component-specific parameters:

$$
p(\boldsymbol{\mu}_{1:K},\boldsymbol{\Sigma}_{1:K}\mid K) = \frac{1}{Z_K}\left[\prod_{k=1}^K p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_k)p_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_k)\right]h_K(\boldsymbol{\mu}_{1:K}).
$$

Z_K is the normalizing constant,

$$
Z_K = \int_{\mathbb{R}^p}\cdots\int_{\mathbb{R}^p} h_K(\boldsymbol{\mu}_{1:K})\prod_{k=1}^K p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_k)\mathrm{d}\boldsymbol{\mu}_1\cdots\mathrm{d}\boldsymbol{\mu}_K,
$$

The function $h_K : (\mathbb{R}^p)^K \to [0,1]$ is permutation-invariant and satisfies the following repulsive condition: $h_K(\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_K) = 0$ if and only if $\boldsymbol{\mu}_k = \boldsymbol{\mu}_{k'}$ for some $k\neq k'$, $k,k'\in\{1,\cdots,K\}$

Two classes of repulsive priors:

$$
\begin{aligned}
h_K(\boldsymbol{\mu}_{1:K}) &= \min_{1\le k<k'\le K} g(\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k'}\|),\\
h_K(\boldsymbol{\mu}_{1:K}) &= \prod_{1\le k<k'\le K} g\left(\|\boldsymbol{\mu}_k-\boldsymbol{\mu}_{k'}\|\right)^{1/K}.
\end{aligned}
$$

The normalizing constants Z_K behaves:

Theorem 1. *Suppose the repulsive function h_K is of the form above. If $\iint_{\mathbb{R}^p\times\mathbb{R}^p}\left[\log g(\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|)\right]^2 p(\boldsymbol{\mu}_1)p(\boldsymbol{\mu}_2)\mathrm{d}\boldsymbol{\mu}_1\mathrm{d}\boldsymbol{\mu}_2 < \infty$, then $0\le -\log Z_K \le c_1 K$ for some constant $c_1>0$.*

3. Theoretical Properties: Convergence

Regularity conditions for the model:

A0 True distribution is of Gaussian mixture form:

$$
f_0(\boldsymbol{y}) = \int \phi(\boldsymbol{y}\mid\boldsymbol{\mu},\boldsymbol{\Sigma})\mathrm{d}F_0(\boldsymbol{\mu},\boldsymbol{\Sigma}),
$$

and $F_0(\|\boldsymbol{\mu}\|\ge t)\le B_1\exp(-b_1t^2)$ for some $B_1, b_1>0$.

A1 For some $\delta>0, c_2>0$, $g(x)\ge c_2\epsilon$ if $\epsilon\le x, \epsilon\in(0,\delta)$.

A2 Normalizing constant Z_K behaves:

$$
\iint [\log g(\|\boldsymbol{\mu}_1-\boldsymbol{\mu}_2\|)]^2 p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_1)p_{\boldsymbol{\mu}}(\boldsymbol{\mu}_2)\mathrm{d}\boldsymbol{\mu}_1\mathrm{d}\boldsymbol{\mu}_2 < \infty.
$$

A3 $\lambda(\boldsymbol{\Sigma})$'s are uniformly bounded away from 0 and ∞.

A4 $\boldsymbol{\Sigma}$'s are simultaneously diagonalizable with a fixed $\boldsymbol{U}$.

Regularity conditions for the prior:

B1 Weakly informative symmetric Dirichlet prior on weights: $(w_{1:K}\mid K)\sim\mathcal{D}_K(\beta)$ for some $\beta\in(0,1]$.

B2 Sub-Gaussian tail for $p_{\boldsymbol{\mu}}$:

$$
\int_{\{\|\boldsymbol{\mu}\|\ge t\}} p_{\boldsymbol{\mu}}(\boldsymbol{\mu})\mathrm{d}\boldsymbol{\mu}\le B_2\exp(-b_2t^2)
$$

for some $b_2, B_2>0$

B3 Power exponential lower bound for $p_{\boldsymbol{\mu}}$:

$$
p(\boldsymbol{\mu})\ge B_3\exp(-b_3\|\boldsymbol{\mu}\|^\alpha)
$$

for some $\alpha\ge 2, B_3, b_3>0$.

B4 Spectrum of covariance matrices: $\lambda_j(\boldsymbol{\Sigma})\sim p_\lambda$ independently with $\mathrm{supp}(p_\lambda)$ being compact.

B5 Tail condition for K: There exists some $B_4, b_4>0$ such that

$$
\begin{aligned}
p_K(K) &\ge \exp\left(-b_4K\log K\right),\\
\sum_{N=K}^\infty p_K(N) &\le \exp\left(-B_4K\log K\right).
\end{aligned}
$$

Convergence results of the posterior:

Theorem 2 (Consistency). *Assume conditions A0-A4 and B1-B5 hold. Then for all $\epsilon>0$,*

$$
\Pi(\|f-f_0\|_1>\epsilon\mid\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)\to 0
$$

in $\mathbb{P}_{f_0}$-probability.

Theorem 3 (Contraction Rate). *Assume conditions A0-A4 and B1-B5 hold. Then for any $t>p+(\alpha+2)/4$,*

$$
\Pi\left(\|f-f_0\|_1>\frac{(\log n)^t}{\sqrt{n}}\mid\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n\right)\to 0
$$

in $\mathbb{P}_{f_0}$-probability.

4. Theoretical Properties: Model Shrinkage

Model Shrinkage for location-mixture repulsive model:

C1 Assume the location-mixture only $(p_{\boldsymbol{\Sigma}}=\delta_{\boldsymbol{\Sigma}_0})$

C2 Take $g(x)=x/(g_0+x)$ where $g_0\in[0,\infty)$ controls the repulsion among $\boldsymbol{\mu}_k$'s, and $p_{\boldsymbol{\mu}}=\mathrm{N}(\boldsymbol{0},\tau^2\boldsymbol{I})$

C3 W.L.O.G. assume $\int\boldsymbol{\mu}F_0(\mathrm{d}\boldsymbol{\mu})=\boldsymbol{0}$.

C4 Prior on K: proportional to Poisson modulus Z_K

$$
p(K)\propto Z_K\frac{\lambda^K}{K!},\quad K=1,2,\cdots
$$

Theorem 4. *Assume conditions A0-A3, B1-B3, C1-C4 hold with $\beta=1$. Then when $N\ge 3$, we have the following result:*

$$
\begin{aligned}
&\mathbb{E}_0\left[\Pi(K\ge N\mid\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n)\right]\\
&\quad\le C(\lambda)\chi(g_0;n,N)\exp\left[\frac{n\tau^2}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_0^{-1}\right)\right]\sum_{K=N+1}^\infty\frac{\lambda^K}{(\mathrm{e}^\lambda-1)K!},
\end{aligned}
$$

where $C(\lambda)$ are some constants depending on λ only, and $\chi(g_0;n,N)$ is a shrinkage constant such that $\chi(0;n,N)=1$ and $\lim_{g_0\to\infty}\chi(g_0;n,N)=0$ as long as τ is sufficiently large.

5. Posterior Inference

Posterior inference via a Gibbs sampler

Notations: $\mathcal{C}_n$: The partition induced by the mixture model; $(\boldsymbol{\gamma}_c^\star,\boldsymbol{\Gamma}_c^\star)$: the cluster-specific parameter for cluster c; $(\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star)$: auxiliary variables; $\mathcal{C}_\varnothing$: components without observations; $\boldsymbol{\theta}_i$: component-specific parameters for the ith observation.

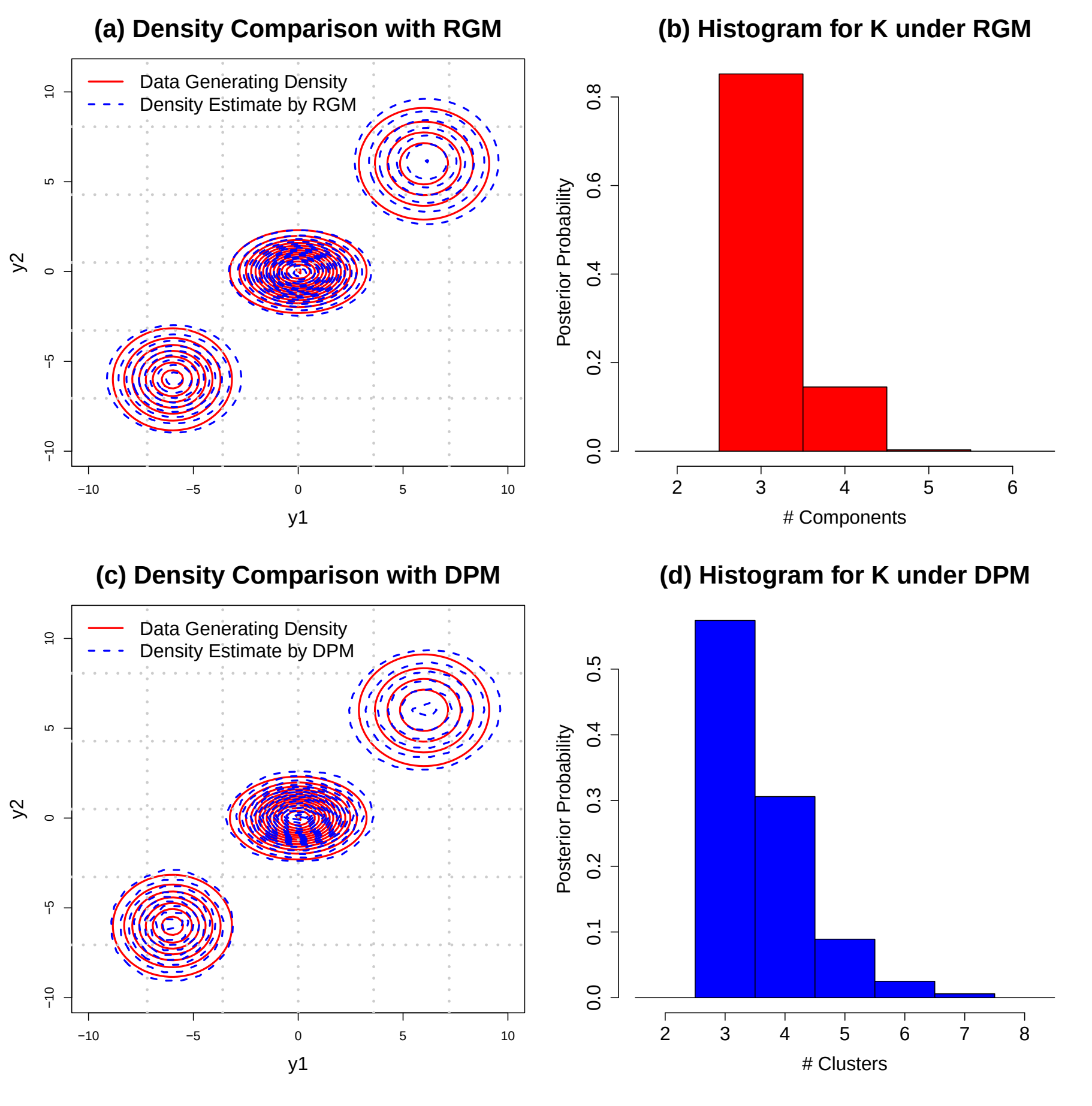
The Gibbs sampler: For each observation $i=1,\cdots,n$, do:

1. **Sample auxiliary variable $(\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star)$:** If $\mathcal{C}_n=\mathcal{C}_{-i}\cup\{\{i\}\}$, then set $(\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star)=\boldsymbol{\theta}_i$; Otherwise sample $(\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star)$ by:
   i) Sample $K\sim p(K\mid\mathcal{C}_n=\mathcal{C}_{-i}\cup\{\{i\}\})$, set $\ell=|\mathcal{C}_{-i}|$, compute $\mathcal{C}_\varnothing$ with $|\mathcal{C}_\varnothing|=K-\ell$, and set $\boldsymbol{\theta}_{-i}=(\boldsymbol{\theta}_1,\cdots,\boldsymbol{\theta}_n)\backslash\{\boldsymbol{\theta}_i\}$.
   ii) Sample $\boldsymbol{\Gamma}_{\underline{c}}^\star\sim p_{\boldsymbol{\Sigma}}(\boldsymbol{\Gamma}_{\underline{c}}^\star)$. Sample $(\boldsymbol{\gamma}_c^\star:c\in\mathcal{C}_\varnothing)$ by accept-reject sampling: Sample $(\boldsymbol{\gamma}_c^\star:c\in\mathcal{C}_\varnothing)$ independently from $p_{\boldsymbol{\mu}}$ and $U\sim\mathrm{Unif}(0,1)$, independent of $(\boldsymbol{\gamma}_c^\star:c\in\mathcal{C}_\varnothing)$; If $U<h_K(\boldsymbol{\gamma}_c^\star:c\in\mathcal{C}_{-i}\cup\mathcal{C}_\varnothing)$, then accept the new proposed samples; Otherwise resample $(\boldsymbol{\gamma}_c^\star:c\in\mathcal{C}_\varnothing)$ from $p_{\boldsymbol{\mu}}$ and U until $U<h_K(\boldsymbol{\gamma}_c^\star:c\in\mathcal{C}_{-i}\cup\mathcal{C}_\varnothing)$. Discard all $(\boldsymbol{\gamma}_c^\star,\boldsymbol{\Gamma}_c^\star:c\in\mathcal{C}_\varnothing\backslash\{\underline{c}\})$.
2. **Sample $\mathcal{C}_n$ from $p(\mathcal{C}_n|,\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star,\boldsymbol{y}_i,\mathcal{C}_{-i},\boldsymbol{\theta}_{-i})$:**

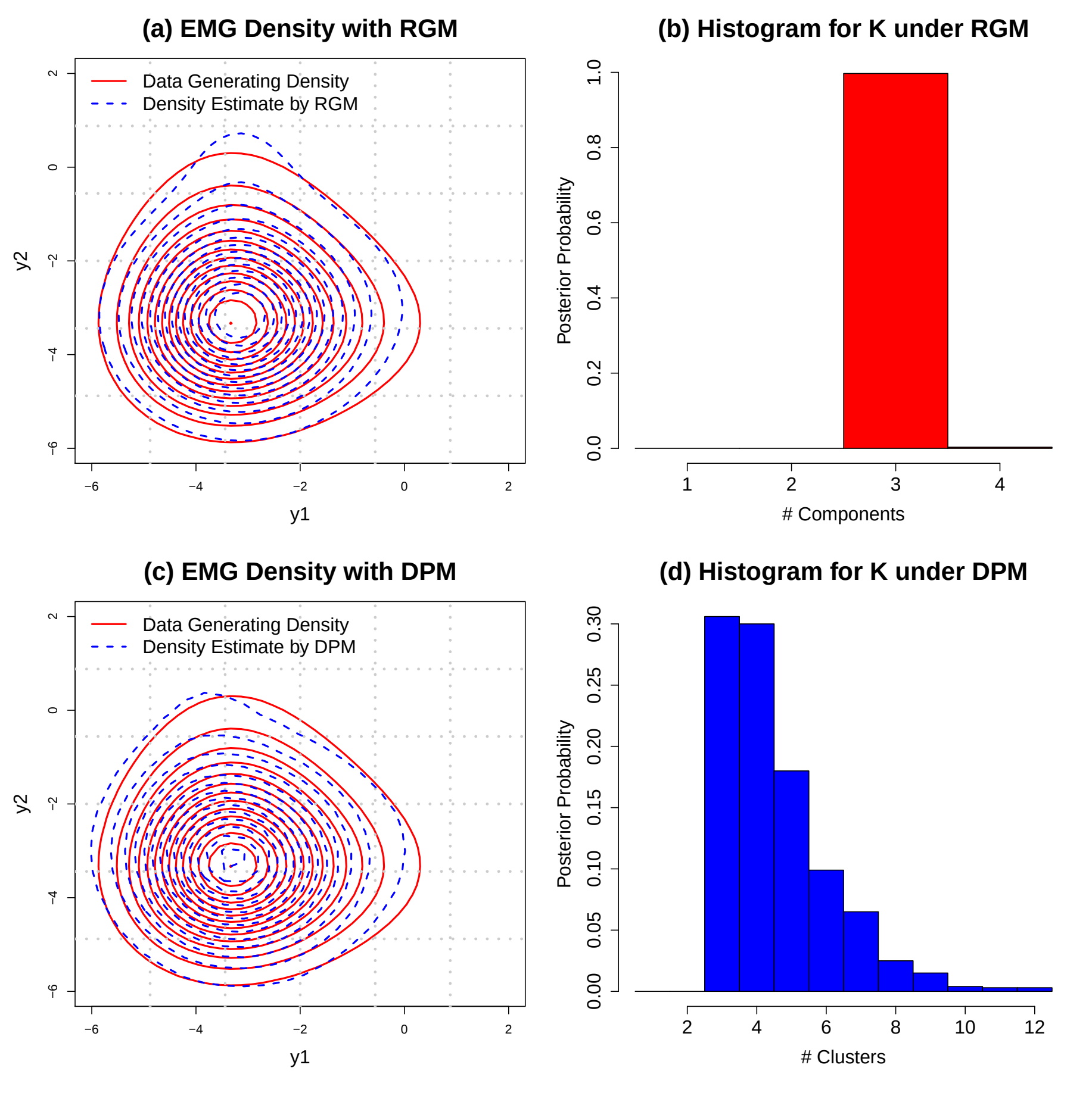
$$
\begin{aligned}
\Pi(i\text{ is a singleton}\mid -)&\propto\left[\frac{V_n(|\mathcal{C}_{-i}|+1)\beta}{V_n(|\mathcal{C}_{-i}|)}\right]\phi(\boldsymbol{y}_i\mid\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star),\\
\Pi(i\to c\mid -)&\propto(|c|+\beta)\,\phi(\boldsymbol{y}_i\mid\boldsymbol{\gamma}_c^\star,\boldsymbol{\Gamma}_c^\star).
\end{aligned}
$$

3. **Assign $\boldsymbol{\theta}_i$ value.** Set $\boldsymbol{\theta}_i=(\boldsymbol{\gamma}_{\underline{c}}^\star,\boldsymbol{\Gamma}_{\underline{c}}^\star)$ if $\mathcal{C}_n=\mathcal{C}_{-i}\cup\{\{i\}\}$, and set $\boldsymbol{\theta}_i=(\boldsymbol{\gamma}_c^\star,\boldsymbol{\Gamma}_c^\star)$ if $\mathcal{C}_n=(\mathcal{C}_{-i}\backslash\{c\})\cup(\{c\cup\{i\}\})$ for some $c\in\mathcal{C}_{-i}$.

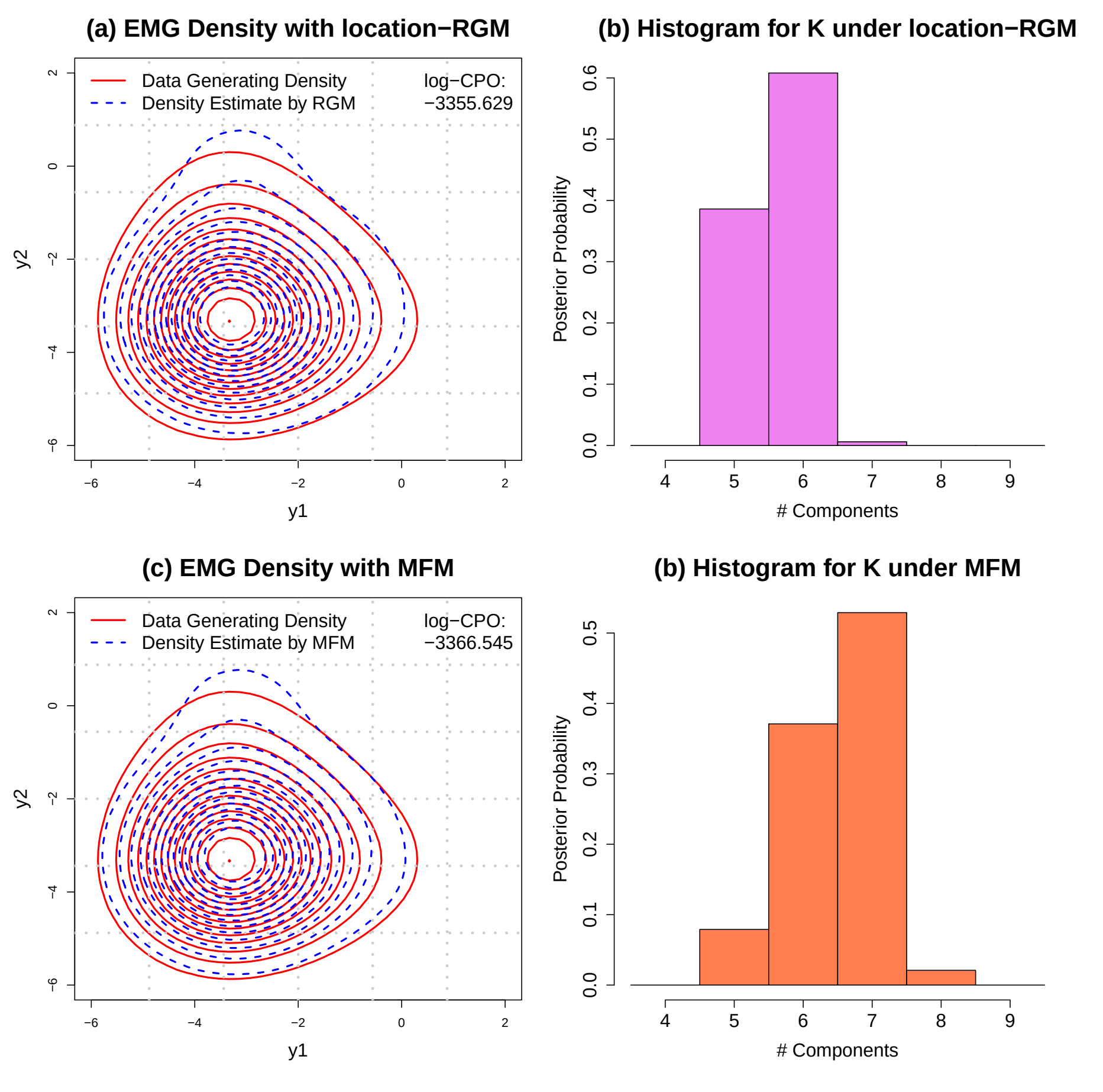
6. Numerical Results



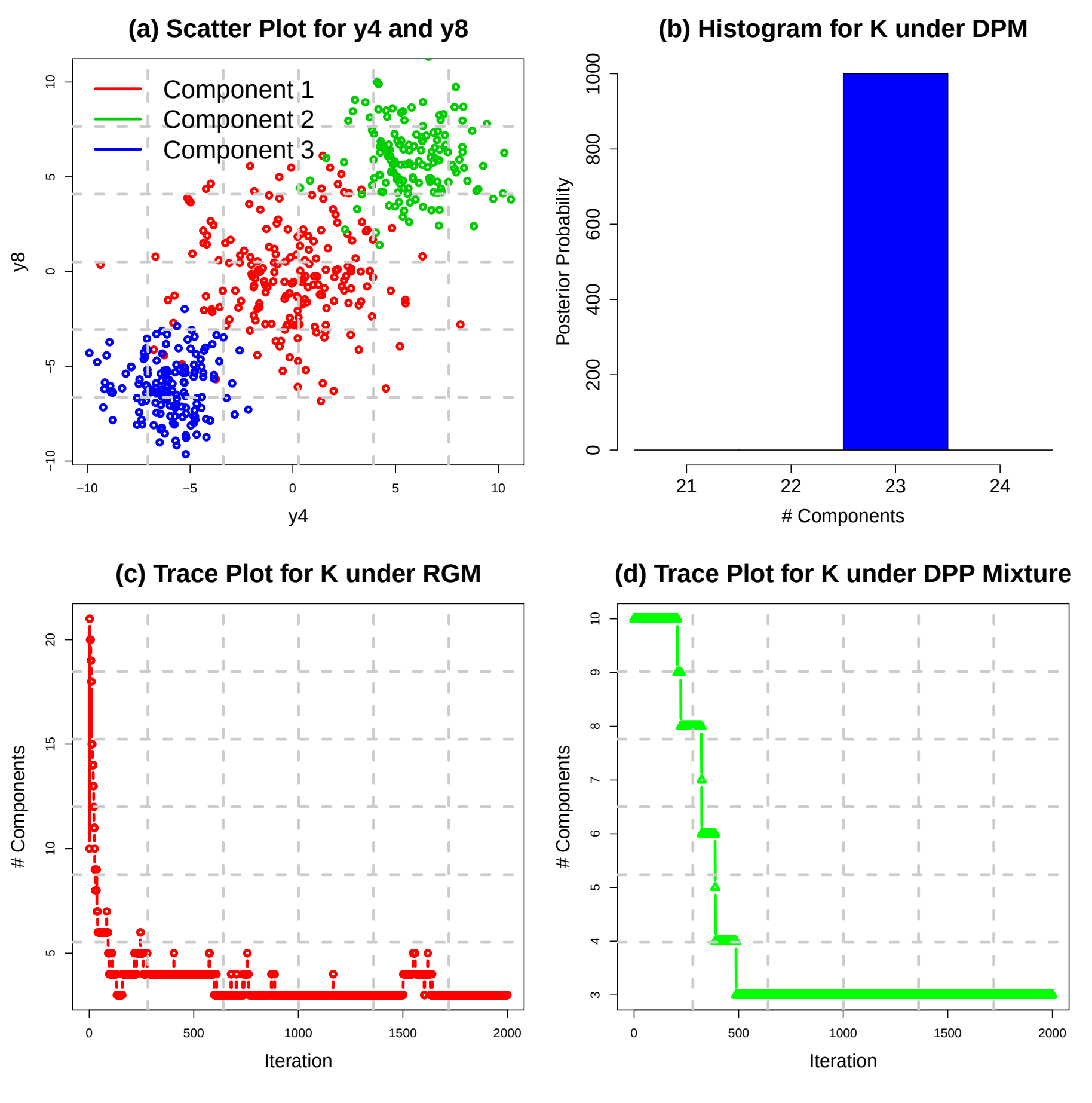
Fiting a discrete mixtures of Gaussians: Panels (a) and (c) are the contour plots for the posterior density estimation of the RGM model and the DPM model, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the RGM model and the posterior number of clusters under the DPM model, respectively, where the underlying true number of components is $K=3$.



Fitting a uni-modal density: Panels (a) and (c) are the contour plots for the posterior density estimation under the RGM model and the DPM model, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the RGM model and the posterior number of clusters under the DPM model, respectively.



Fitting a uni-modal density using location-mixtures only: Panels (a) and (c) are the contour plots for the posterior density estimation under the location-RGM and the MFM, respectively. Panels (b) and (d) are the histograms of the posterior number of components under the locatio-RGM and MFM, respectively.



A 10-dimensional model-based clustering: Panel (a) is the scatter plot of the 4th-versus-8th coordinate of the simulated data; Panel (b) is the histogram of the posterior number of clusters under the DPM model; Panels (c) and (d) are the trace plots for the posterior samples of K under the RGM model, and that of the number of clusters under the DPP mixture model, respectively.