Dimension mixtures of finite-dimensional models

Jeffrey W. Miller   Brown University, Division of Applied Math

1. Introduction

Summary

Many of the commonly-used nonparametric models (such as the Dirichlet process mixture (DPM), hierarchical Dirichlet process (HDP), and Indian buffet process (IBP)) can be interpreted as an infinite-dimensional limit of finite-dimensional models. A less common approach is simply to put a prior on the dimension — that is, to take a “dimension mixture” of finite-dimensional models. This is very natural from the Bayesian perspective, however, it has been believed that inference in such models is difficult and requires techniques such as reversible jump MCMC.

To the contrary, we have found that this approach gives rise to combinatorial stochastic processes that closely parallel those of the DPM, HDP, and IBP. Consequently, efficient approximate inference for such dimension mixture models can be done in much the same way as for these standard nonparametric models.

Although more experimentation needs to be done, the method of dimension mixtures appears to be an attractive and widely-applicable approach to constructing nonparametric Bayesian models.

Similarities with standard nonparametric models

1. Approximate inference techniques (e.g. Gibbs sampling) are nearly identical
2. Interpretation in terms of “restaurant processes”
3. Exchangeability properties
4. Posterior predictive performance appears to be nearly identical

Advantages

1. Interpretability and conceptual simplicity
2. Natural Bayesian approach (if something is unknown, put a prior on it)
3. Cleaner clusters/topics/features (no tendency to make tiny superfluous groups)
4. Complete control over the distribution on the number of clusters/topics/features
5. Consistency typically holds automatically (assuming identifiability)

Disadvantages

1. MCMC mixing time may be longer
2. Slightly more complicated formulas
3. More parameters (due to item 4 above)

2. Mixture of finite mixtures (MFM)

Generative model description

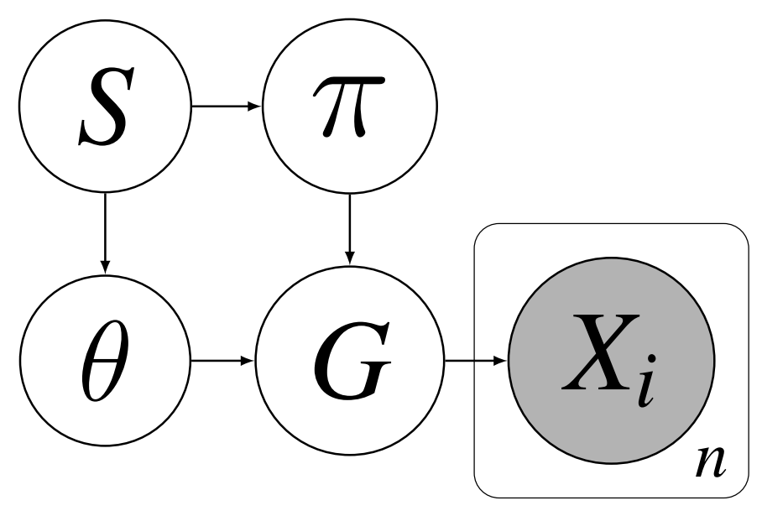
Many researchers have considered the following natural alternative to DPMs.

(e.g. Nobile (1994, 2007), Richardson & Green (1997, 2001), Stephens (2000), etc.)

Instead of $G \sim \mathrm{DP}(\alpha, H)$, choose the mixing measure G to put mass on a randomly-selected number S of (randomly-selected) parameter values $\theta_1, \ldots, \theta_S$:

Mixture of finitely-supported measures (MF)

$S \sim q(s)$, a p.m.f. on $\{1, 2, \ldots\}$
$\pi \sim \mathrm{Dirichlet}_s(\gamma, \ldots, \gamma)$ (given $S = s$)
$\theta_1, \ldots, \theta_s \overset{\text{iid}}{\sim} H$ (given $S = s$)
$G = \sum_{i=1}^{S} \pi_i \delta_{\theta_i} \quad \Longrightarrow \quad G \sim \mathrm{MF}(\gamma, H, q)$.



Then, draw $X_1, X_2, \ldots \overset{\text{iid}}{\sim} f_G(x) = \sum_{i=1}^{S} \pi_i p_{\theta_i}(x)$. We call this a **MFM**.

Partition distribution (DP vs MF)

A partition $\mathcal{C}$ of $\{1, \ldots, n\}$ into $t = |\mathcal{C}|$ parts has probability

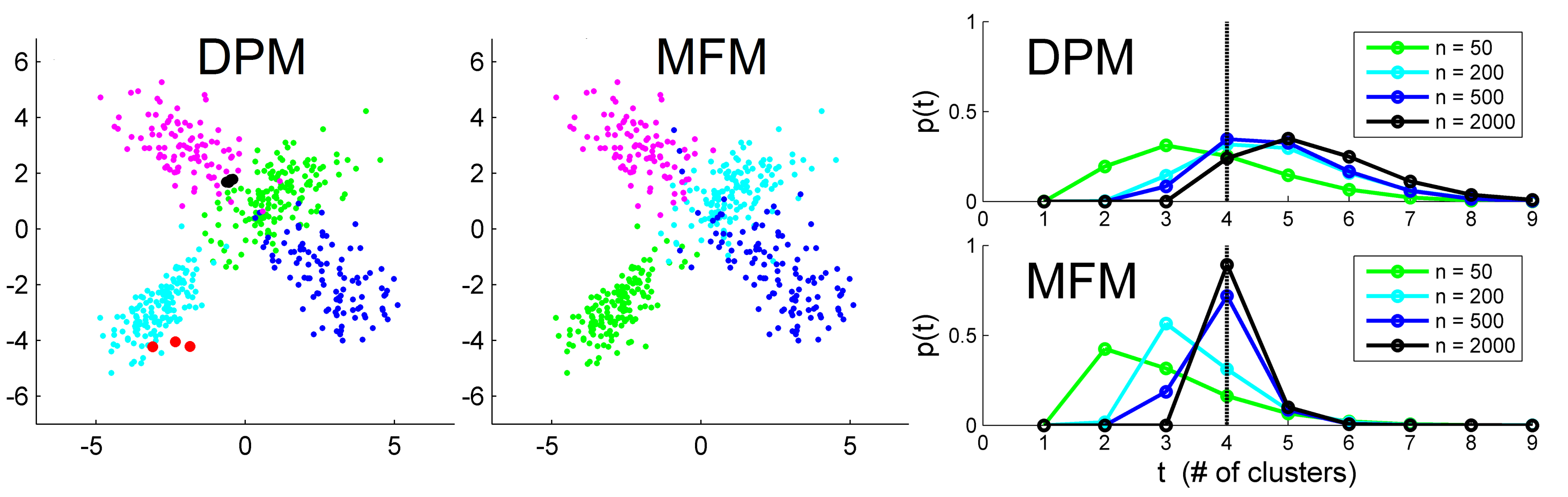
$$P_{\mathrm{DP}}^{(n)}(\mathcal{C}) = \frac{\alpha^t}{\alpha^{(n)}} \prod_{c \in \mathcal{C}} (|c| - 1)! \qquad\qquad P_{\mathrm{MF}}^{(n)}(\mathcal{C}) = v_n(t) \prod_{c \in \mathcal{C}} \gamma^{(|c|)}$$

where $x_{(k)} = x(x-1)\cdots(x-k+1)$, $x^{(k)} = x(x+1)\cdots(x+k-1)$, and $v_n(t) = \sum_{s=1}^{\infty} \frac{s_{(t)}}{(\gamma s)^{(n)}} q(s)$.

- Both P_{DP} and P_{MF} are “EPPFs” of Gibbs form (as in Pitman, 2006).
- The numbers $v_n(t)$ can be efficiently precomputed to arbitrary precision.
- This leads to a simple “restaurant process” closely resembling the CRP.
- Gibbs sampling for MFMs and DPMs is nearly identical.

Demonstration: bivariate Gaussian mixture

Typical clustering   Posterior on # of clusters



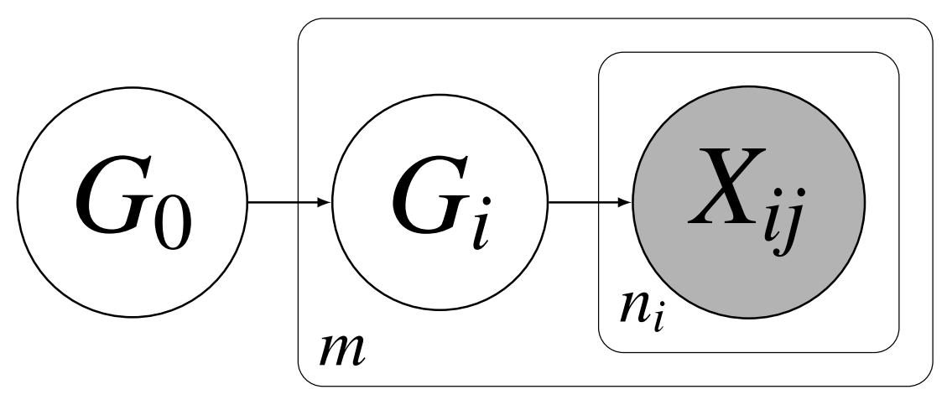
3. Hierarchical MFM (HMFM)

Generative model description

Similarly, we can construct an alternative to the Hierarchical DP (HDP) of Teh et al. (2006) by drawing G_0 in this way, and drawing a “lower level” of mixing measures $G_1, \ldots, G_m$ using G_0 as their base measure:

Hierarchical MF (HMF)

$G_0 \sim \mathrm{MF}(\gamma, H, q)$
$G_1, \ldots, G_m \overset{\text{iid}}{\sim} \mathrm{MF}(\gamma, G_0, q)$ (given G_0)



Then, draw $X_{ij} \sim f_{G_i}(x)$ indep. for $j \in \{1, \ldots, n_i\}$, $i \in \{1, \ldots, m\}$. We call this a **HMFM**.

Hierarchical partition distribution (HDP vs HMF)

For $i = 1, \ldots, m$, let $\mathcal{C}_i$ be a partition of $\{1, \ldots, n_i\}$, and let $t_i = |\mathcal{C}_i|$. Let $\mathcal{C}_0$ be a partition of $\{1, \ldots, N\}$ where $N = \sum t_i$, and let $t_0 = |\mathcal{C}_0|$. Then letting $\mathcal{C} = (\mathcal{C}_0, \mathcal{C}_1, \ldots, \mathcal{C}_m)$,

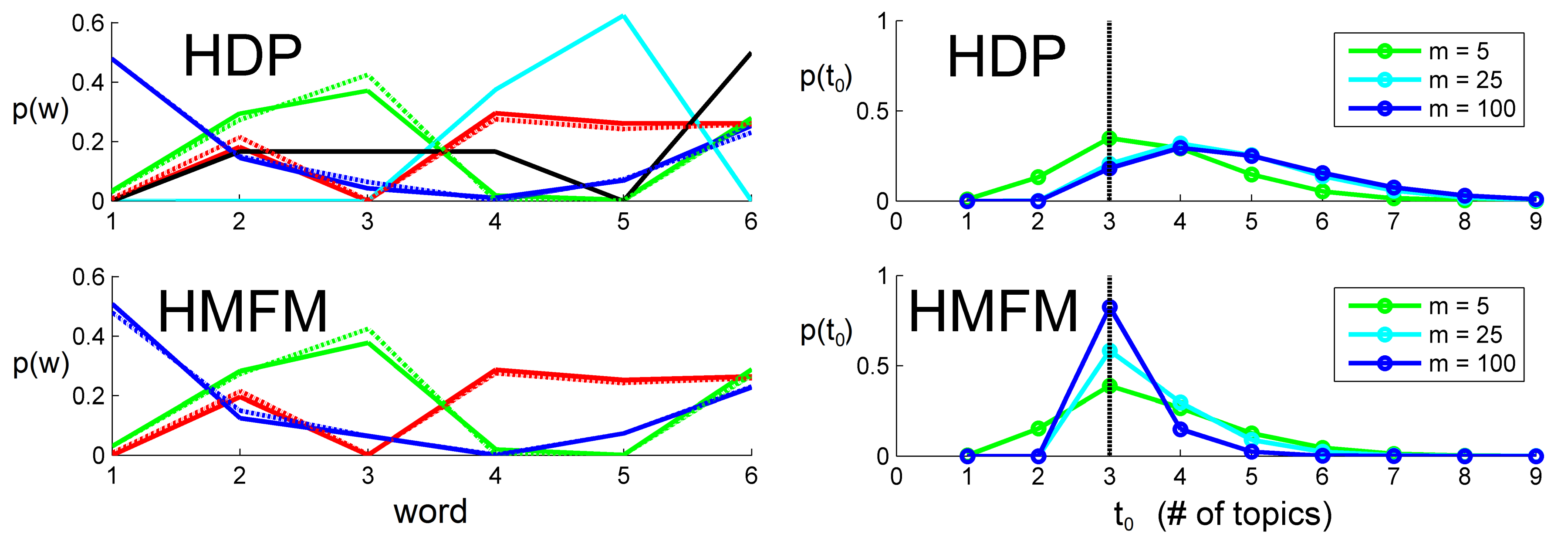
$$P_{\mathrm{HDP}}(\mathcal{C}) = P_{\mathrm{DP}}^{(N)}(\mathcal{C}_0) \prod_{i=1}^{m} P_{\mathrm{DP}}^{(n_i)}(\mathcal{C}_i) \qquad\qquad P_{\mathrm{HMF}}(\mathcal{C}) = P_{\mathrm{MF}}^{(N)}(\mathcal{C}_0) \prod_{i=1}^{m} P_{\mathrm{MF}}^{(n_i)}(\mathcal{C}_i)$$

with P_{DP} and P_{MF} as above. (Note that $\mathcal{C}_0$ depends on $\mathcal{C}_1, \ldots, \mathcal{C}_m$ through $N = \sum t_i$.)

- This leads to a simple “franchise process” closely resembling that of the HDP.
- Gibbs sampling for HMFMs and HDPs is nearly identical.
- Since N is not fixed, caching $v_N(t_0)$ is more memory-efficient than precomputing.

Demonstration: toy topic model

Typical topic distributions   Posterior on # of topics



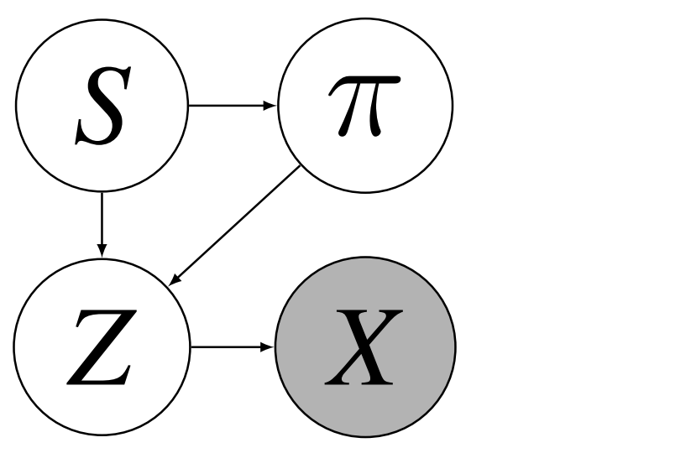
4. Mixture of finite feature models (MFFM)

Generative model description

In the same way, we can construct an alternative to the Indian Buffet Process (IBP) of Griffiths and Ghahramani (2005) by making the number of features S random.

A distribution on binary matrices

$S \sim q(s)$, a p.m.f. on $\{0, 1, 2, \ldots\}$
$\pi_1, \ldots, \pi_s \overset{\text{iid}}{\sim} \mathrm{Beta}(a, b)$ (given $S = s$)
For $j \in \{1, \ldots, s\}$ (given $S = s$ and π):
$Z_{1j}, \ldots, Z_{nj} \overset{\text{iid}}{\sim} \mathrm{Bernoulli}(\pi_j)$.



Then, draw $(X_1, \ldots, X_n)$ according to the feature matrix Z. We call this a **MFFM**.

Equivalence class distribution (IBP vs MFFM)

Consider two binary matrices equivalent if they are the same after removing any columns containing only zeros. The probability of obtaining $\bar{Z} \in \{0,1\}^{n \times t}$ with column sums $m_1, \ldots, m_t > 0$ after removing any zero columns from Z is

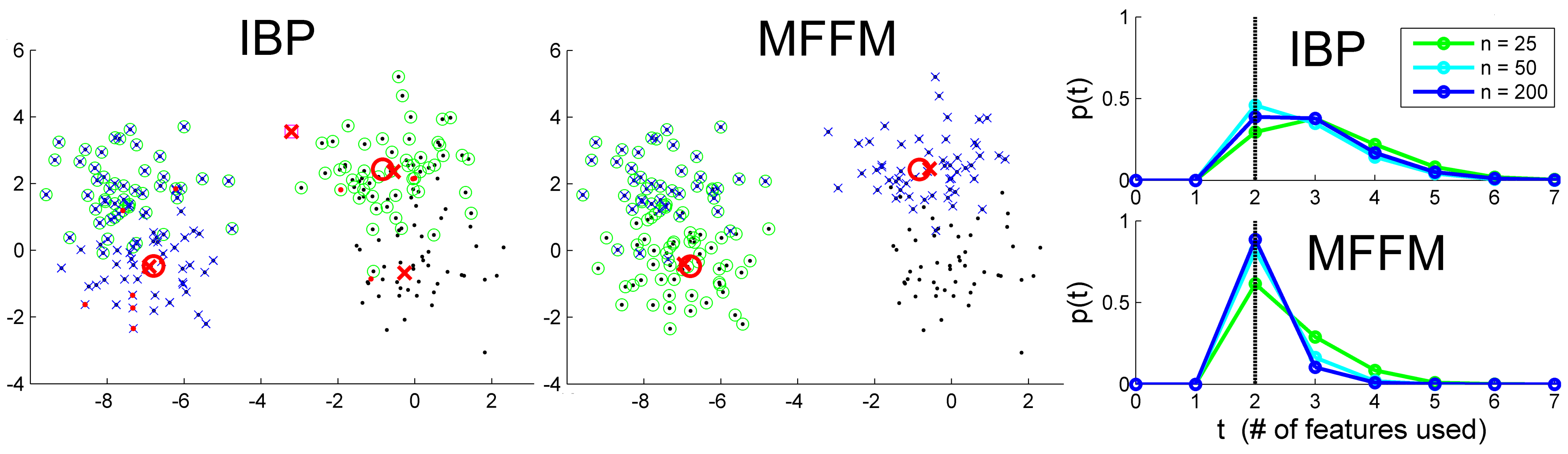
$$P_{\mathrm{IBP}}(\bar{Z}) = \frac{\alpha^t e^{-\alpha H_n}}{t!} \prod_{i=1}^{t} \frac{(m_i - 1)!\,(n - m_i)!}{n!} \qquad\qquad P_{\mathrm{MFFM}}(\bar{Z}) = v'_n(t) \prod_{i=1}^{t} \frac{a^{(m_i)}\, b^{(n-m_i)}}{(a+b)^{(n)}}$$

where $x^{(n)} = x(x+1)\cdots(x+n-1)$ and $v'_n(t) = \sum_{s=0}^{\infty} \binom{s}{t} c_n^{s-t} q(s)$, with $c_n = \frac{b^{(n)}}{(a+b)^{(n)}}$.

- If $q(s) = \mathrm{Poisson}(s \mid \lambda)$ then $v'_n(t) = e^{\lambda c_n} \mathrm{Poisson}(t \mid \lambda)$.
- In general, $v'_n(t)$ can be efficiently precomputed to arbitrary precision.
- This leads to a simple “restaurant process” closely resembling that of the IBP.
- Gibbs sampling for MFFMs and IBPs is nearly identical.

Demonstration: toy linear-Gaussian feature model

Typical feature assignments   Posterior on # of features


IBP
MFFM
IBP
MFFM
p(t)
n = 25
n = 50
n = 200
t (# of features used)