Feature Partitions and Multi-View Clustering

Yarin Gal, Zoubin Ghahramani

University of Cambridge, UK

In Short:

- We define **a new combinatorial structure that unifies Kingman [1978]'s random partitions and Broderick, Pitman, and Jordan [2013]'s feature frequency models**.
- **This structure underlies non-parametric multi-view clustering models**, where data points are simultaneously clustered into different possible clusterings.
- The de Finetti measure is a **product of paintbox constructions**, tying together Kingman [1978]'s paintbox and Broderick et al. [2013]'s feature paintbox.
- Characterising the properties of feature partitions allows us to **understand the relations between the models they underlie and to share algorithmic insights between the models**.

Motivation

Many non-parametric multi-view clustering models and applications exist:

- Multi-view clustering model developed for **identity and pose identification** in portrait photos, and machines identification in sound data [Guan et al., 2010],
- Another suggested for **text prediction** [Niu et al., 2012],
- **Network modelling** based on multi-view clustering, predicting links in **protein interaction networks** of the yeast *S. cerevisiae* [Palla et al., 2012],
- **Cognitive models** using multi-view clustering, capturing human reasoning about high dimensional data [Shafto et al., 2006].

Surprisingly, the models above have a common underlying structure that unites them all

Preliminaries – Partitions & Kingman's Paintbox

- A *partition of* $\mathbb{N}$ is a mutually exclusive and exhaustive set of subsets of $\mathbb{N}$.
- An *exchangeable random partition* is a random element in the set of partitions, invariant to permutations of the naturals.
- *Kingman's paintbox* is the directing measure underlying exchangeable random partitions. It describes a sampling procedure that allows us, conditioned on some measure, to generate iid samples from an exchangeable random partition (Fig. 1).

FIGURE 1: **A Kingman paintbox** and the corresponding partition generated from it: $\mathcal{R} = \{\{1\}, \{2,3\}, \{4\}\}$.

Feature Partitions

Definition 1. *A* **feature partition** $\mathcal{F} = \{A_1, A_2, ...\}$ *over the data points* $[N] = \{1, ..., N\}$ *with* K *possible features* $[K] = \{1, ..., K\}$ *is defined as a partition of* $[N] \times [K]$*, pairs of natural numbers where the first element denotes the data point label and the second element denotes the feature label. We require all subsets* A_k *to have the property that:*

$$\text{if } (i,j) \in A_k \text{ and } (i',j') \in A_k, \text{ then it must be that } j = j'.$$

- A random feature partition F is a random element in the set of feature partitions,
- F is said to be *exchangeable* if

$$\sigma_1 \times \sigma_2(F) \stackrel{d}{=} F$$

for all σ_1 permutations of $[N]$ and σ_2 permutations of $[K]$.

$$\mathcal{F} = \{\{(1,1),\ (4,1)\},\ \{(2,1)\},\ \{(3,1)\},\ \{(1,2),\ (2,2)\},\ \{(3,2),\ (4,2)\},\ \{(1,3),\ (2,3),\ (3,3),\ (4,3)\},\ \dots\ \}$$

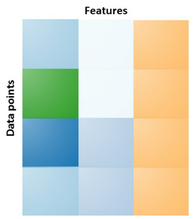

FIGURE 2: **A feature partition** (left) depicted as a matrix with categorical entries corresponding to the subsets A_i (right). Rows in the matrix correspond to data points n, columns correspond to features k; colours denote categorical value assignments for features.

Factorial Paintbox Construction

The following construction extends Kingman's paintbox to feature partitions:

Definition 2. *Given a sequence of probability measures* (μ_j) *each defined over the interval* $[0,1]$ *with disjoint support sets, for every* j *generate a sequence of random variables* $X_{i,j} \sim \mu_j$ *iid for* $i \in \mathbb{N}$*. The sequence* $(X_{i,j})$ *defines a random feature partition* F *exchangeable in data points by* $F_{x,y} = \{\omega | X_x(\omega) = X_y(\omega)\}$*, the event that* x *and* y *belong to the same block for* $x = (i,j), y = (i',j')$*. If for all* j *we have in addition that if* $\mu_j \sim \mu$ *then* F *is exchangeable in features as well.*

The factorial paintbox construction (Fig. 3) ties together:

- Kingman [1978]'s paintbox construction (by restricting it to a single feature), and
- Broderick et al. [2013]'s feature paintbox construction for feature frequency models (through a restriction to two values per feature).

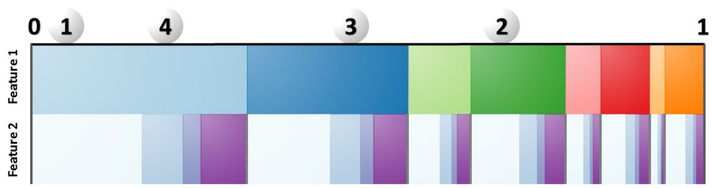

FIGURE 3: **Factorial paintbox corresponding to the first 2 features of $\mathcal{F}$ in Fig. 2.** The nested paintboxes are identical.

de Finetti's theorem for exchangeable random feature partitions:

Theorem 3. *The underlying directing measure for a feature partition is a factorial paintbox construction with some random measure* α *over random measures* (α_j)*.*

Exchangeable Probability Function

The probability function for the feature partition:

Theorem 4. *Let* $((P_{i,j})_{i=1}^{\infty})_{j=1}^{\infty}$ *be a sequence of independent sequences of random variables with the constraints that* $P_{i,j} \geq 0$ *and* $\sum_{i=1}^{\infty} P_{i,j} \leq 1$ *for all* j*. Let* $p(\cdot)$ *be defined as*

$$p\left((n_{i,j})_{i=1}^{k(j)})_{j=1}^{K}\right) = E_{((P_{i,j})_{i=1}^{k(j)})_{j=1}^{K}}\left[\prod_{j=1}^{K}\left(\prod_{i=1}^{k(j)} P_{i,j}^{n_{i,j}-1} \cdot \prod_{i=1}^{k(j)-1}\left(1 - \sum_{k=1}^{i} P_{k,j}\right)\right)\right]$$

for $((n_{i,j})_{i=1}^{k(j)})_{j=1}^{K}$ *a sequence of sequences of natural numbers.*
There exists a random feature partition F *exchangeable w.r.t. the data points with asymptotic frequencies in order of appearance given by the distribution* $((P_{i,j})_{i=1}^{\infty})_{j=1}^{\infty}$ *iff* $p(\cdot)$ *is symmetric w.r.t. re-orderings of the counts within each sequence* j*.* $p(\cdot)$ *is the probability function of* F *then.*

This function generalises the exchangeable random partition probability functions and feature allocation probability function.

Impact

Many models and applications use multi-view clustering...

- ... we **identified** various multi-view clustering **models as equivalent**,
- ... we **collapsed** many multi-view clustering **applications into the same class**,
- ... we can **share algorithmic insights** between the various models **using the feature partition as their underlying model**:
  - Explain away differences between the models,
  - Unify inference for the different models (Gibbs sampling used for some and variational inference for others)
  - A clear way of generalisation (using various distributions over the partitions, introducing new dependencies, etc.).

Future Research

Introduce dependencies to the factorial paintbox construction:

- Could be depicted as each block having its own paintbox sampled from a distribution conditioned on the block itself,
- Could be used to model the underlying structure of correlated multi-clustering models such as in Doshi-Velez and Ghahramani [2009],
- Corresponds to a special case of the fragmentation chain [Bertoin, 2006].