

Dependent Dirichlet-Multinomial Processes with Random Number of Components

Andrea Cremaschi¹ Beatrice Franzolini²

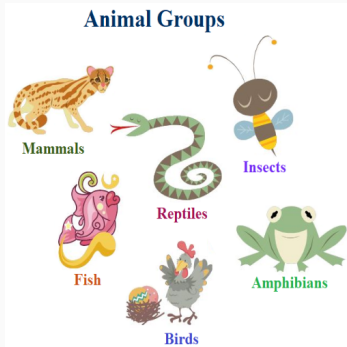
SISBAYES 2025

Department of Statistical Sciences, University of Padova

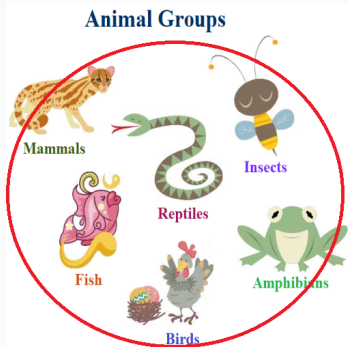
4-5 September 2025

¹School of Science and Technology, IE University, Madrid

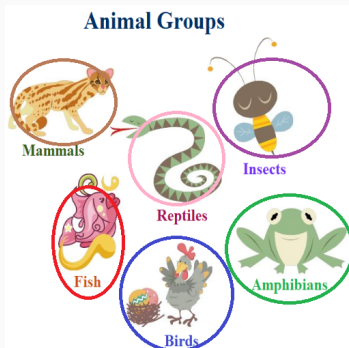
²Bocconi Institute for Data Science and Analytics, Bocconi University, Milan



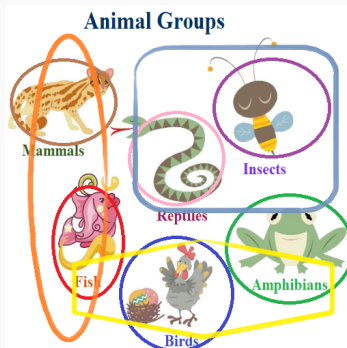
- Several settings present data with a grouped structure



- Several settings present data with a grouped structure
- ✗ Ignore grouping. Neglects heterogeneity across groups



- Several settings present data with a grouped structure
- ✗ Ignore grouping. Neglects heterogeneity across groups
- ✗ Assume independence across groups. Does not allow for sharing of information



- Several settings present data with a grouped structure
- ✗ Ignore grouping. Neglects heterogeneity across groups
- ✗ Assume independence across groups. Does not allow for sharing of information
- ✓ Allow dependence within and between groups

In Bayesian Statistics (and Nonparametrics) several modelling approaches exists
👉 different ways to induce dependence among groups

In Bayesian Statistics (and Nonparametrics) several modelling approaches exist
👉 different ways to induce dependence among groups

FOCUS on of this talk: mixture models with random number of components

They are less studied than the infinite-dimensional approaches, especially in the context of grouped data

In Bayesian Statistics (and Nonparametrics) several modelling approaches exist
👉 different ways to induce dependence among groups

FOCUS on of this talk: mixture models with random number of components

They are less studied than the infinite-dimensional approaches, especially in the context of grouped data

Most recent and related work:

👉 Colombi et al. (2024) “Hierarchical Mixture of Finite Mixtures”. *Bayesian Analysis*.

We exploit the well-known Dirichlet-Multinomial construction.

We exploit the well-known Dirichlet-Multinomial construction.

Let G groups, each equipped with a mixture of M components.

For each group $g = 1, \dots, G$:

$$P_g(\cdot) \stackrel{\text{a.s.}}{=} \sum_{m=1}^M W_{gm} \delta_{\theta_m}(\cdot) \quad (1)$$

$$\theta_1, \dots, \theta_M \mid M \stackrel{\text{i.i.d.}}{\sim} P_0 \quad (2)$$

$$\mathbf{W}_g = (W_{g1}, \dots, W_{gM}) \mid M \sim \text{Dir}(\alpha, \dots, \alpha) \quad (3)$$

where:

(1) are the *mixing measures*

(2) are the *atoms* (shared across groups)

(3) are the mixture *weights* with Dirichlet marginals

- within each group  Dirichlet-Multinomial process $P_g \sim \text{DMP}(\alpha, P_0)$

IDEA  Can we directly model the matrix of weights $\mathbf{W} = \{\mathbf{W}_g, g = 1, \dots, G\}$ to induce dependence across groups?

IDEA  Can we directly model the matrix of weights $\mathbf{W} = \{\mathbf{W}_g, g = 1, \dots, G\}$ to induce dependence across groups?

Properties we require:

- $W_{gm} \in (0, 1)$, for $g = 1, \dots, G$, $m = 1, \dots, M$
- $\sum_{m=1}^M W_{gm} = 1$, for $g = 1, \dots, G$
- $\mathbf{W}_g = (W_{g1}, \dots, W_{gM}) \sim \text{Dir}(\alpha, \dots, \alpha)$

Let $P_g, P_\ell \sim \text{DMP}(\alpha, P_0)$; $X \mid P_g \sim P_g$ and $Y \mid P_\ell \sim P_\ell$.

For any compatible joint distribution of the weights the following holds:

Proposition (Lower bound correlations)

If $\text{Corr}(W_{gm}, W_{\ell m}) \geq 0$ for any m , then:

- (i) $\text{Corr}[P_g(A), P_\ell(A)] \geq \frac{M\alpha + 1}{M\alpha + M}$
- (ii) $\text{Corr}(X, Y) \geq \frac{1}{M}$
- (iii) *The lower bounds obtained with independent sequences of weights $\mathbf{W}_g \perp \mathbf{W}_\ell$*

Let $P_g, P_\ell \sim \text{DMP}(\alpha, P_0)$; $X \mid P_g \sim P_g$ and $Y \mid P_\ell \sim P_\ell$.

For any compatible joint distribution of the weights the following holds:

Proposition (Lower bound correlations)

If $\text{Corr}(W_{gm}, W_{\ell m}) \geq 0$ for any m , then:

- (i) $\text{Corr}[P_g(A), P_\ell(A)] \geq \frac{M\alpha + 1}{M\alpha + M}$
- (ii) $\text{Corr}(X, Y) \geq \frac{1}{M}$
- (iii) *The lower bounds obtained with independent sequences of weights $\mathbf{W}_g \perp \mathbf{W}_\ell$*

Colombi et al. (2024) with Dirichlet weights 📄 lowest possible correlation

Let $P_g, P_\ell \sim \text{DMP}(\alpha, P_0)$; $X \mid P_g \sim P_g$ and $Y \mid P_\ell \sim P_\ell$.

For any compatible joint distribution of the weights the following holds:

Proposition (Lower bound correlations)

If $\text{Corr}(W_{gm}, W_{\ell m}) \geq 0$ for any m , then:

- (i) $\text{Corr}[P_g(A), P_\ell(A)] \geq \frac{M\alpha + 1}{M\alpha + M}$
- (ii) $\text{Corr}(X, Y) \geq \frac{1}{M}$
- (iii) *The lower bounds obtained with independent sequences of weights $\mathbf{W}_g \perp \mathbf{W}_\ell$*

Colombi et al. (2024) with Dirichlet weights 📄 lowest possible correlation

Different approach: joint distribution of the weights via matrix-variate random variables

Let $\mathbf{U} = (U_1, \dots, U_M)$ be an M -dimensional array of $G \times G$ matrices

$\mathbf{U}$ has a (symmetric) *matrix-variate Dirichlet distribution* with parameter α , and we write $\mathbf{U} \mid \alpha \sim \text{MDir}_{M,G}(\mathbf{U} \mid \alpha)$, if it has a density:

$$p(\mathbf{U} \mid \alpha) = \frac{1}{\beta_G^\alpha} \prod_{m=1}^M |U_m|^{\alpha - (G+1)/2} \mathbb{1}_{\mathbf{U} \in S_{M,G}}$$

where:

- $\alpha > G/2$
- β_G^α is the multivariate beta function (Gupta & Nagar, 2018)
- support:

$$S_{M,G} = \left\{ (U_1, \dots, U_M) : U_m \text{ is positive definite for each } m \text{ and } U_M = \mathbb{I}_G - \sum_{m=1}^{M-1} U_m \right\}$$

- Olkin and Rubin (1964)  study of the distributional and independence properties
-  Extensively studied in Gupta and Nagar (2018)
-  Admits constructive definition via normalisation of Wishart matrices

- Olkin and Rubin (1964)  study of the distributional and independence properties
-  Extensively studied in Gupta and Nagar (2018)
-  Admits constructive definition via normalisation of Wishart matrices

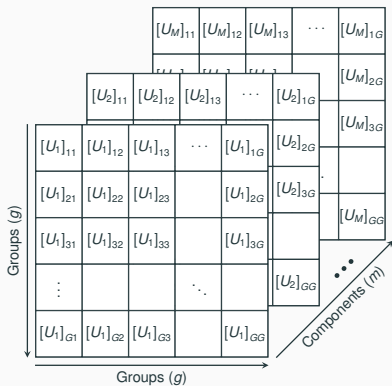
Why is it of interest to us?

- Olkin and Rubin (1964) 📖 study of the distributional and independence properties
- 📖 Extensively studied in Gupta and Nagar (2018)
- 📖 Admits constructive definition via normalisation of Wishart matrices

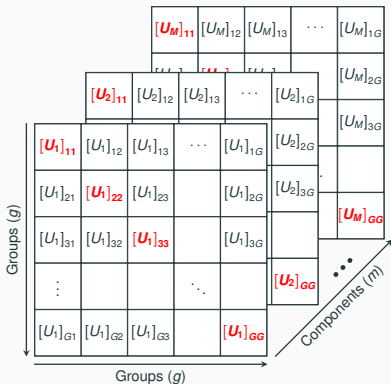
Why is it of interest to us?

Let $W_{gm} \stackrel{d}{=} [U_m]_{gg}$, then:

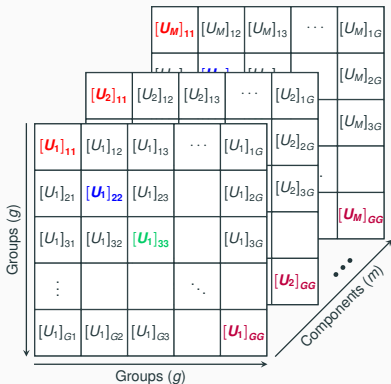
- ✓ $W_{gm} \in (0, 1)$, for $g = 1, \dots, G$ and $m = 1, \dots, M$
- ✓ $\sum_{m=1}^M W_{gm} = 1$, for $g = 1, \dots, G$
- ✓ $\mathbf{W}_g = (W_{g1}, \dots, W_{gM}) \sim \text{Dir}(\alpha, \dots, \alpha)$



- Let $\mathbf{U} \mid \alpha \sim \text{MDir}_{M,G}(\mathbf{U} \mid \alpha)$



- Let $\mathbf{U} \mid \alpha \sim \text{MDir}_{M,G}(\mathbf{U} \mid \alpha)$
- The diagonal slice of this array has the required properties

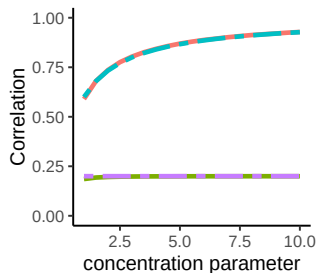


- Let $\mathbf{U} \mid \alpha \sim \text{MDir}_{M,G}(\mathbf{U} \mid \alpha)$
- The diagonal slice of this array has the required properties
- We set $W_{gm} = [U_m]_{gg}$
for $g = 1, \dots, G$ and $m = 1, \dots, M$

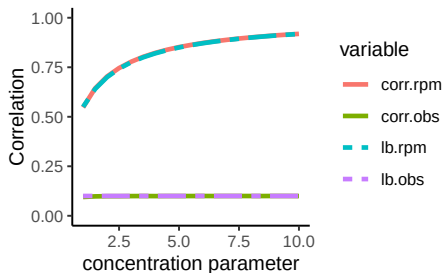
What is the dependence among groups when using $\mathbf{U} \mid \alpha \sim \text{MDir}_{M,G}(\mathbf{U} \mid \alpha)$?

What is the dependence among groups when using $\mathbf{U} \mid \alpha \sim \text{MDir}_{M,G}(\mathbf{U} \mid \alpha)$?

Numerical example with $G = 2, \alpha > 1$



$M = 5$



$M = 1 + \text{Poi}(10)$

ISSUES

1. matrix-variate Dirichlet construction does not yield a wide range of correlations
2. $\alpha > G/2$ limits span of correlations and flexibility of resulting distributions

ISSUES

1. matrix-variate Dirichlet construction does not yield a wide range of correlations
2. $\alpha > G/2$ limits span of correlations and flexibility of resulting distributions

SOLUTIONS

1. leverage the unnormalised construction of the matrix-variate Dirichlet distribution
2. start from *Generalised-Wishart distribution* to allow for $\alpha \leq G/2$ (Srivastava, 2003)

New construction:

Let $S_m \stackrel{\text{i.i.d.}}{\sim} \text{Gen-Wishart}_G(2\alpha, \Psi)$, then for $m = 1, \dots, M$

$$S_g = \sum_{m=1}^M [S_m]_{gg}$$

$$W_{gm} = [S_m]_{gg} / S_g$$

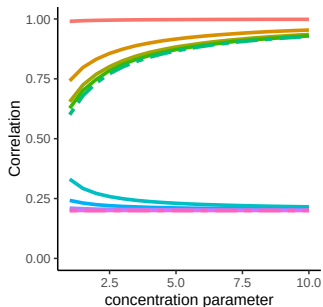
and we write $(P_1, \dots, P_G) \sim \text{W-DMP}(\alpha, P_0, \Psi)$

👉 this construction retains information from scale matrix Ψ

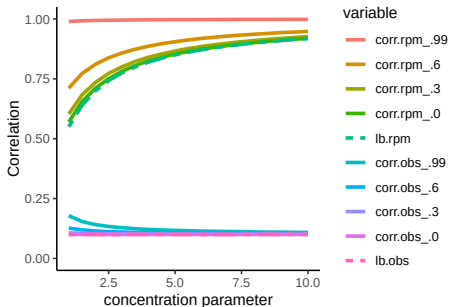
👉 no need for matrix inversion or Cholesky decomposition!

★ Correlation under new construction

Numerical example with $G = 2$, $\alpha > 1$, $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix}$



$M = 5$



$M = 1 + \text{Poi}(10)$

(i) **Closure under marginalisation:** $P_g \sim \text{DMP}(\alpha, P_0)$

☆ Ψ does not affect the marginal distribution of the processes

(ii) **Independence**

$\mathbf{W}_g \perp \mathbf{W}_\ell$ (Colombi et al. 2024, Dirichlet weights) recovered when $\Psi = \mathbb{I}_G$

(iii) **Full-dependence**

$\psi_{gl} \rightarrow 1$ or $\psi_{gl} \rightarrow -1 \implies \mathbf{W}_g \stackrel{\text{a.s.}}{=} \mathbf{W}_\ell, P_g \stackrel{\text{a.s.}}{=} P_\ell$

☆ Ψ act as correlation matrix btw rpms

Ψ can be chosen *a priori* to encode the dependence among rpms

(i) Distribution of the matrix of weights $[W]_{gm}$, for $G = 2$, conditionally on $M > 0$

Let $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix}$ and $\rho^2 = \frac{\psi_{12}^2}{\psi_{11}\psi_{22}} \in [0, 1)$, then:

$$p(\mathbf{W}) = \frac{(1 - \rho^2)^{M\alpha}}{\Gamma(\alpha)^M} \sum_{(k_1, \dots, k_M)} \rho^{2K} \Gamma(K + M\alpha)^2 \prod_{m=1}^M \frac{(W_{1m}W_{2m})^{k_m + \alpha - 1}}{k_m! \Gamma(k_m + \alpha)} \mathbb{1}_{\mathbf{W} \in \Delta_M^2}$$

where $k_m \in \mathbb{N} \cup 0$, $K = \sum_m k_m$, Δ_M^2 is an appropriate probability simplex.

(i) Distribution of the matrix of weights $[W]_{gm}$, for $G = 2$, conditionally on $M > 0$

Let $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix}$ and $\rho^2 = \frac{\psi_{12}^2}{\psi_{11}\psi_{22}} \in [0, 1)$, then:

$$p(\mathbf{W}) = \frac{(1 - \rho^2)^{M\alpha}}{\Gamma(\alpha)^M} \sum_{(k_1, \dots, k_M)} \rho^{2K} \Gamma(K + M\alpha)^2 \prod_{m=1}^M \frac{(W_{1m}W_{2m})^{k_m + \alpha - 1}}{k_m! \Gamma(k_m + \alpha)} \mathbb{1}_{\mathbf{W} \in \Delta_M^2}$$

where $k_m \in \mathbb{N} \cup 0$, $K = \sum_m k_m$, Δ_M^2 is an appropriate probability simplex.

(ii) Hierarchical representation of $p(\mathbf{W})$

$$\begin{aligned} \mathbf{W}_{g\cdot} = (W_{g1}, \dots, W_{gM}) \mid \mathbf{k} = (k_1, \dots, k_M) &\stackrel{iid}{\sim} \text{Dirichlet}(k_1 + \alpha, \dots, k_M + \alpha), \quad g = 1, 2 \\ \mathbf{k} = (k_1, \dots, k_M) &\stackrel{iid}{\sim} \text{NegBin}(\alpha, 1 - \rho^2), \end{aligned}$$

Let $X_{gi} \mid (P_1, \dots, P_G) \stackrel{\text{ind}}{\sim} P_g$ and $(P_1, \dots, P_G) \sim \text{W-DMP}(\alpha, P_0, \Psi)$. Denote with $(X_m^*, m = 1, \dots, M^{(\text{obs})})$ the unique values in the data, and with $\mathbf{n}_g = (n_{g1}, \dots, n_{gM})$ the corresponding numerosities, then a posteriori (conditionally on $M > 0$)

$$P_g(\cdot) \stackrel{\text{a.s.}}{=} \sum_{m=1}^{M^{(\text{obs})}} W_{gm} \delta_{X_m^*}(\cdot) + \sum_{m=M^{(\text{obs})}+1}^M W_{gm} \delta_{\theta_m}(\cdot), \quad g = 1, 2$$

$$\mathbf{W}_g = (W_{g1}, \dots, W_{gM}) \mid \mathbf{k} \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{k} + \mathbf{n}_g + \alpha) \quad \text{for } g = 1, 2$$

$$p(\mathbf{k}) \propto \frac{\prod_{g=1}^2 \text{B}(\mathbf{k} + \mathbf{n}_g + \alpha)}{\text{B}(\mathbf{k} + \alpha)^2} \prod_{m=1}^M \left[\binom{k_m + \alpha - 1}{k_m} \rho^{2k_m} \right] \mathbb{1}_{k_m \in \mathbb{N} \cup 0}$$

Let $X_{gi} \mid (P_1, \dots, P_G) \stackrel{\text{ind}}{\sim} P_g$ and $(P_1, \dots, P_G) \sim \text{W-DMP}(\alpha, P_0, \Psi)$. Denote with $(X_m^*, m = 1, \dots, M^{(\text{obs})})$ the unique values in the data, and with $\mathbf{n}_g = (n_{g1}, \dots, n_{gM})$ the corresponding numerosities, then a posteriori (conditionally on $M > 0$)

$$P_g(\cdot) \stackrel{\text{a.s.}}{=} \sum_{m=1}^{M^{(\text{obs})}} W_{gm} \delta_{X_m^*}(\cdot) + \sum_{m=M^{(\text{obs})}+1}^M W_{gm} \delta_{\theta_m}(\cdot), \quad g = 1, 2$$

$$\mathbf{W}_g = (W_{g1}, \dots, W_{gM}) \mid \mathbf{k} \stackrel{\text{iid}}{\sim} \text{Dirichlet}(\mathbf{k} + \mathbf{n}_g + \alpha) \quad \text{for } g = 1, 2$$

$$p(\mathbf{k}) \propto \frac{\prod_{g=1}^2 B(\mathbf{k} + \mathbf{n}_g + \alpha)}{B(\mathbf{k} + \alpha)^2} \prod_{m=1}^M \left[\binom{k_m + \alpha - 1}{k_m} \rho^{2k_m} \right] \mathbb{1}_{k_m \in \mathbb{N} \cup 0}$$

Extensions to $G > 2$ available (more complicated)

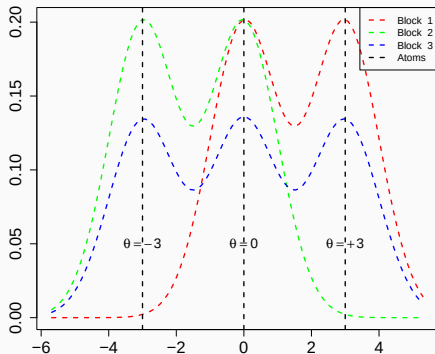
Simulation Study

Simulation setting:

- $G = 30$ split into 3 blocks of size 10
- Within each block, the true data-generating mixture has different weights but same atoms

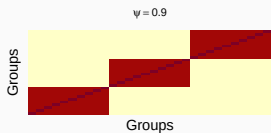
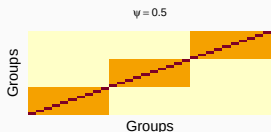
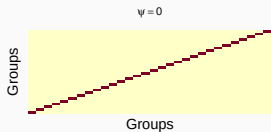
$$W_g^{true} = \begin{cases} (0, \frac{1}{2}, \frac{1}{2}), & g = 1, \dots, 10 \\ (\frac{1}{2}, \frac{1}{2}, 0), & g = 11, \dots, 20 \\ (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}), & g = 21, \dots, 30 \end{cases}$$

Data-generating mixture



We fit a univariate Normal-Inverse Gamma model

For each simulation setting, the matrix $\Psi = [\psi]_{g_1 g_2}$ has different off-diagonal elements for $g_1, g_2 = 1, \dots, G$

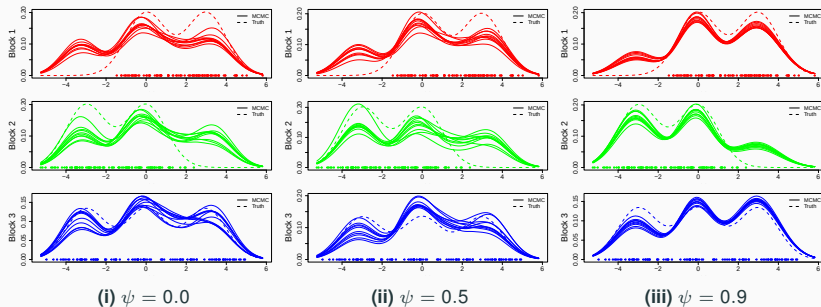


(i) $\psi = 0.0$

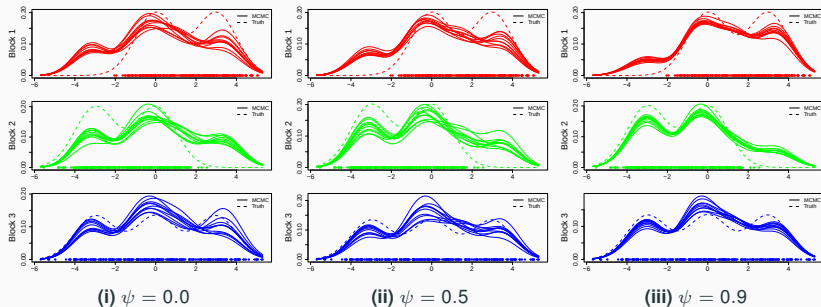
(ii) $\psi = 0.5$

(iii) $\psi = 0.9$

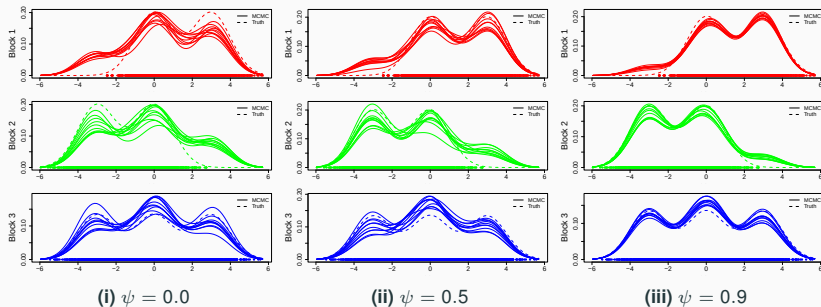
$n_g = 10$, for $g = 1, \dots, G$



$n_g = 25$, for $g = 1, \dots, G$

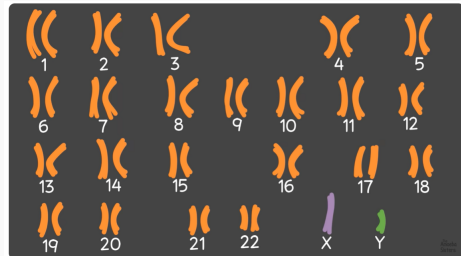


$n_g = 100$, for $g = 1, \dots, G$

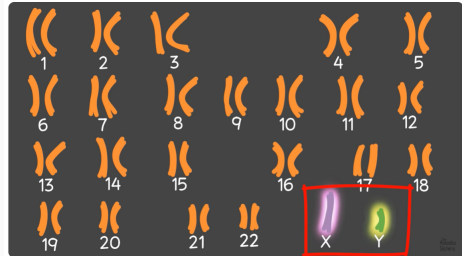


Sex differences in gene expressions from the human brain

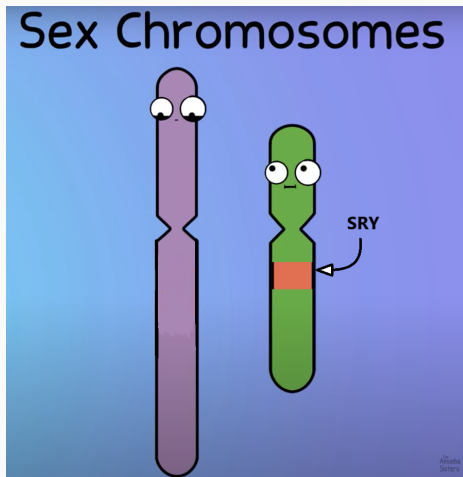
- Humans carry 23 pairs of chromosomes as their genetic material (*karyotype*)



- Humans carry 23 pairs of chromosomes as their genetic material (*karyotype*)
- 22 pairs are called autosome
1 special pair contains the XY chromosomes



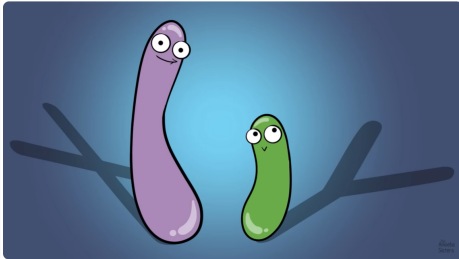
- Humans carry 23 pairs of chromosomes as their genetic material (*karyotype*)
- 22 pairs are called autosome
1 special pair contains the XY chromosomes
- *virilization process* (SRY gene present in the Y chromosome) 🚫 determines the sex of the offspring (Female XX / Male XY)

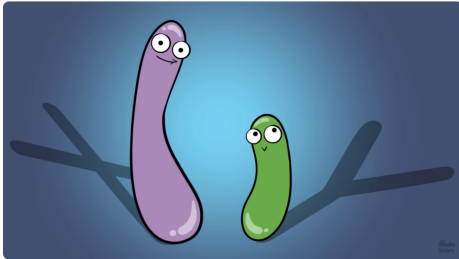


- Humans carry 23 pairs of chromosomes as their genetic material (*karyotype*)
- 22 pairs are called autosome
1 special pair contains the XY chromosomes
- *virilization process* (SRY gene present in the Y chromosome) 🐣 determines the sex of the offspring (Female XX / Male XY)
- Karyotype changes across species



- Some traits are sex-specific and the corresponding genes carried by X or Y are called X-linked and Y-linked genes

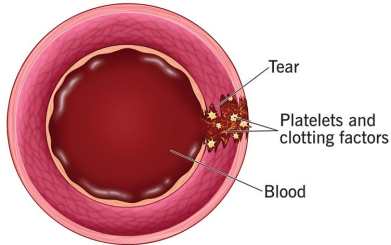




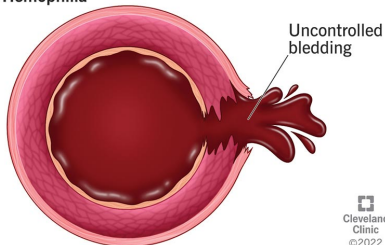
- Some traits are sex-specific and the corresponding genes carried by X or Y are called X-linked and Y-linked genes
- Some examples:

Hemophilia

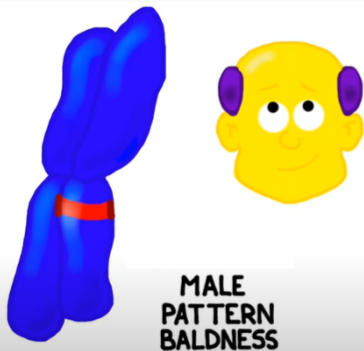
Normal blood vessel



Hemophilia



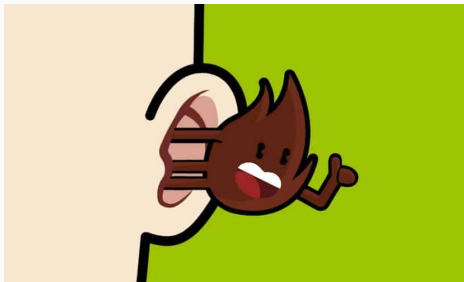
- Some traits are sex-specific and the corresponding genes carried by X or Y are called X-linked and Y-linked genes
- Some examples:
 - haemophilia 🩸 X-linked



- Some traits are sex-specific and the corresponding genes carried by X or Y are called X-linked and Y-linked genes
- Some examples:
 - haemophilia ➡ X-linked
 - baldness ➡ X-linked

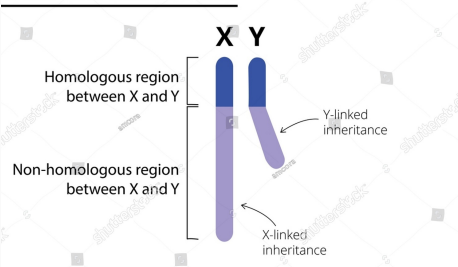


- Some traits are sex-specific and the corresponding genes carried by X or Y are called X-linked and Y-linked genes
- Some examples:
 - **haemophilia** 🐸 X-linked
 - **baldness** 🐸 X-linked
 - **webbed toes** 🐸 Y-linked

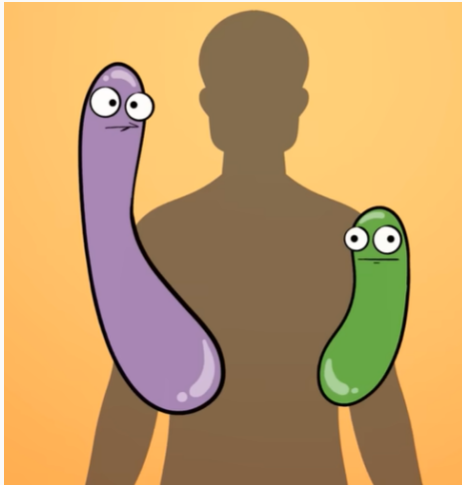


- Some traits are sex-specific and the corresponding genes carried by X or Y are called X-linked and Y-linked genes
- Some examples:
 - **haemophilia** 🚑 X-linked
 - **baldness** 🚑 X-linked
 - **webbed toes** 🚑 Y-linked
 - **hairy ears** 🚑 thought to be Y-linked but not yet confirmed (Lee et al. 2004)

Sex Chromosome



- Some traits are sex-specific and the corresponding genes carried by X or Y are called **X-linked** and **Y-linked** genes
- Some examples:
 - **haemophilia** 🚑 **X-linked**
 - **baldness** 🚑 **X-linked**
 - **webbed toes** 🚑 **Y-linked**
 - **hairy ears** 🚑 thought to be **Y-linked** but not yet confirmed (Lee et al. 2004)
- **X-Y homologues** are genes that are present in the X and Y chromosomes. They are very similar (but not identical)...



- Some traits are sex-specific and the corresponding genes carried by X or Y are called **X-linked** and **Y-linked** genes
- Some examples:
 - **haemophilia** ➡ **X-linked**
 - **baldness** ➡ **X-linked**
 - **webbed toes** ➡ **Y-linked**
 - **hairy ears** ➡ thought to be **Y-linked** but not yet confirmed (Lee et al. 2004)
- **X-Y homologues** are genes that are present in the X and Y chromosomes. They are very similar (but not identical)...
- ...also found in non-gonadal tissues

There is evidence that X- and Y-linked genes contribute to the development of non-gonadal tissues...what about the brain?

There is evidence that X- and Y-linked genes contribute to the development of non-gonadal tissues...what about the brain?

- Gene expressions measured from post-mortem brain samples Vawter et al. (2004)
 - three brain regions
 - three laboratories (UC Irvine, UC Davis, UMichigan Ann Arbor)
 - 10 subjects (5 male, 5 females)
- We select a subset of genes:
 - From Vawter et al. (2024): **Y-linked** 🚺 UTY, USP9Y, SMCY, DBY, RPS4Y
 - From Vawter et al. (2024): **X-linked** 🚺 XIST
 - + top 14 genes with highest variability across subjects/labs/brain regions ($H = 20$)

Each subject represents a group, $g = 1, \dots, G = 10$

👉 model the genes within each group via regression:

$$Y_{gi} \stackrel{\text{i.i.d.}}{\sim} N(\theta_{c_{gh_i}} + \beta \mathbf{x}_i, \sigma_{c_{gh_i}}^2), \quad \mathbf{x}_i = [\text{Lab}_i, \text{Brain Region}_i]$$

$$\mathbb{P}(c_{gh} = m \mid M, \mathbf{W}) = W_{gm}, \quad m = 1, \dots, M, \quad h = 1, \dots, H$$

$$(\theta_1, \sigma_1^2), \dots, (\theta_M, \sigma_M^2) \mid M \stackrel{\text{i.i.d.}}{\sim} \text{N-inv-}\Gamma(\theta, \sigma^2 \mid m_0, k_0, a_0, b_0)$$

$$\mathbf{W} \sim \text{W-DMP}(\alpha = 1, \Psi), \quad M \sim \text{Poi}_1(\Lambda)$$

where, within each group, we indicate by h_i the gene measured in the i -th observation.

Each subject represents a group, $g = 1, \dots, G = 10$

👉 model the genes within each group via regression:

$$Y_{gi} \stackrel{\text{i.i.d.}}{\sim} N(\theta_{c_{gh_i}} + \beta \mathbf{x}_i, \sigma_{c_{gh_i}}^2), \quad \mathbf{x}_i = [\text{Lab}_i, \text{Brain Region}_i]$$

$$\mathbb{P}(c_{gh} = m \mid M, \mathbf{W}) = W_{gm}, \quad m = 1, \dots, M, \quad h = 1, \dots, H$$

$$(\theta_1, \sigma_1^2), \dots, (\theta_M, \sigma_M^2) \mid M \stackrel{\text{i.i.d.}}{\sim} \text{N-inv-}\Gamma(\theta, \sigma^2 \mid m_0, k_0, a_0, b_0)$$

$$\mathbf{W} \sim \text{W-DMP}(\alpha = 1, \Psi), \quad M \sim \text{Poi}_1(\Lambda)$$

where, within each group, we indicate by h_i the gene measured in the i -th observation.

👉 we estimate clustering of the genes across subjects and experimental conditions

Each subject represents a group, $g = 1, \dots, G = 10$

👉 model the genes within each group via regression:

$$Y_{gi} \stackrel{\text{ind.}}{\sim} N(\theta_{c_{gh_i}} + \beta \mathbf{x}_i, \sigma_{c_{gh_i}}^2), \quad \mathbf{x}_i = [\text{Lab}_i, \text{Brain Region}_i]$$

$$\mathbb{P}(c_{gh} = m \mid M, \mathbf{W}) = W_{gm}, \quad m = 1, \dots, M, \quad h = 1, \dots, H$$

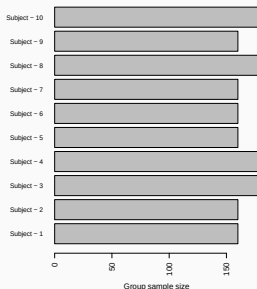
$$(\theta_1, \sigma_1^2), \dots, (\theta_M, \sigma_M^2) \mid M \stackrel{\text{i.i.d.}}{\sim} \text{N-inv-}\Gamma(\theta, \sigma^2 \mid m_0, k_0, a_0, b_0)$$

$$\mathbf{W} \sim \text{W-DMP}(\alpha = 1, \Psi), \quad M \sim \text{Poi}_1(\Lambda)$$

where, within each group, we indicate by h_i the gene measured in the i -th observation.

👉 we estimate clustering of the genes across subjects and experimental conditions

👉 Number of observations within each group (total $n = 1680$)



Each subject represents a group, $g = 1, \dots, G = 10$

👉 model the genes within each group via regression:

$$Y_{gi} \stackrel{\text{i.i.d.}}{\sim} N(\theta_{c_{gh_i}} + \beta \mathbf{x}_i, \sigma_{c_{gh_i}}^2), \quad \mathbf{x}_i = [\text{Lab}_i, \text{Brain Region}_i]$$

$$\mathbb{P}(c_{gh} = m \mid M, \mathbf{W}) = W_{gm}, \quad m = 1, \dots, M, \quad h = 1, \dots, H$$

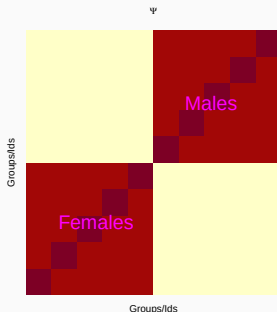
$$(\theta_1, \sigma_1^2), \dots, (\theta_M, \sigma_M^2) \mid M \stackrel{\text{i.i.d.}}{\sim} \text{N-inv-}\Gamma(\theta, \sigma^2 \mid m_0, k_0, a_0, b_0)$$

$$\mathbf{W} \sim \text{W-DMP}(\alpha = 1, \Psi), \quad M \sim \text{Poi}_1(\Lambda)$$

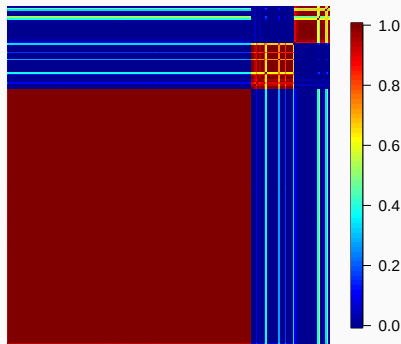
where, within each group, we indicate by h_i the gene measured in the i -th observation.

👉 we estimate clustering of the genes across subjects and experimental conditions

Scale matrix Ψ 👉 dependence
information
specified via the covariate Sex_i



- Posterior co-clustering probabilities for $H = 20$ unique genes
- We actually cluster $H \times G = 200$ different variables (dimension of the matrix *rightarrow*)
- Three main clusters are identified



- **Distinct clustering structures** found for Male and Female groups

- **Females:**

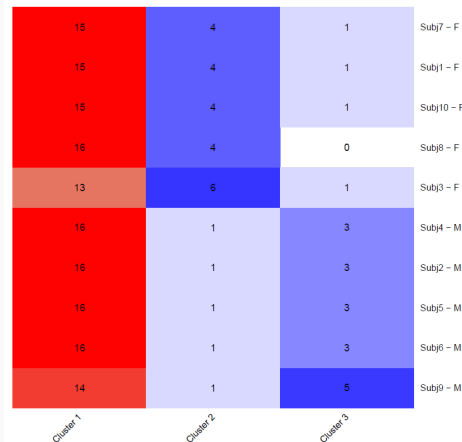
- Cluster 2: contains **Y-linked genes**
- Cluster 3: contains **XIST (X-linked) overexpressed**

- **Males:**

- Cluster 2: contains **XIST (X-linked)**
- Cluster 3: contains **Y-linked genes overexpressed**

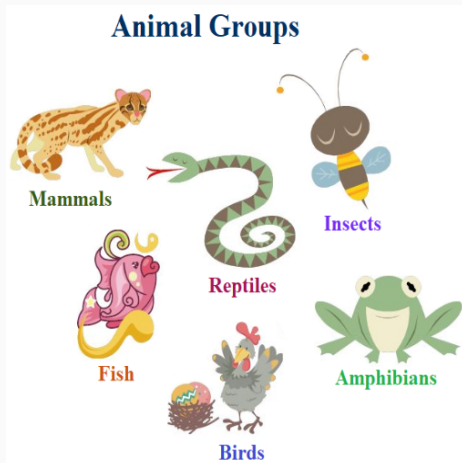
- **Cluster summary:**

- Cluster 2: *low-expressed genes*
- Cluster 3: *overexpressed genes*
- Cluster 1: *all the rest*

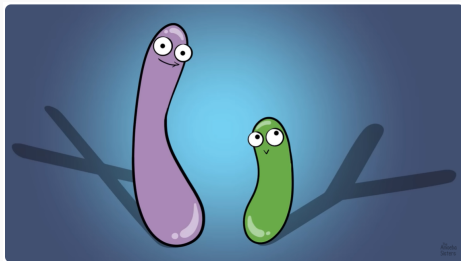


Estimated cluster sizes within each group (Binder estimate)

- New framework for grouped data via mixtures with random number of components
- Joint modelling of weights via matrix construction
- Prior correlation btw rpms encoded through matrix Ψ
- Improved estimation accuracy
- Applied to sex-specific gene expression; results align with literature and reveal gene clusters



- New framework for grouped data via mixtures with random number of components
- Joint modelling of weights via matrix construction
- Prior correlation btw rpms encoded through matrix Ψ
- Improved estimation accuracy
- Applied to sex-specific gene expression; results align with literature and reveal gene clusters



- Argiento & De Iorio (2022). "Is infinity that far? A Bayesian nonparametric perspective of finite mixture models". *The Annals of Statistics*, 50(5), 2641-2663.
- Colombi, Argiento, Camerlenghi and Paci (2024) "Hierarchical Mixture of Finite Mixtures" *Bayesian Analysis*, 1(1), pp.1-29.
- Olkin & Rubin (1964). "Multivariate beta distributions and independence properties of the Wishart distribution". *The Annals of Mathematical Statistics*, 261-269.
- Gupta & Nagar (2018). "Matrix variate distributions". Chapman and Hall/CRC.
- Vawter et al. (2004) "Gender-specific gene expression in post-mortem human brain: localization to sex chromosomes". *Neuropsychopharmacology*, 29(2), 373-384.
- Lee et al. (2004) "Molecular evidence for absence of Y-linkage of the Hairy Ears trait". *European journal of human genetics*, 12(12), 1077-1079.

We consider Wishart-type rvs for any integer α , a *Generalised Wishart* distribution.

We consider Wishart-type rvs for any integer α , a *Generalised Wishart* distribution.

Let $S \sim \text{Gen-Wishart}_G(2\alpha, \Psi)$, then:

$$f_{\text{Gen-Wishart}}(S) = \begin{cases} \frac{(\det S)^{\frac{\alpha-G-1}{2}} \exp\{-\frac{1}{2} \text{tr}(\Psi^{-1} S)\}}{2^{\frac{\alpha G}{2}} \Gamma_G(\frac{\alpha}{2}) (\det \Psi)^{\frac{\alpha}{2}}} & \alpha > G/2 \\ \frac{(\det S_{11})^{\frac{\alpha-G-1}{2}} \exp\{-\frac{1}{2} \text{tr}[\Psi^{-1} S]\}}{\pi^{\alpha(G-\alpha)/2} 2^{\frac{\alpha G}{2}} \Gamma_{\alpha}(\frac{\alpha}{2}) (\det \Psi)^{\frac{\alpha}{2}}} & \alpha \leq G/2 \end{cases}$$

where in the singular case $S = \begin{pmatrix} S_{11} & S_{12} \\ S'_{12} & S_{22} \end{pmatrix}$ and $S_{22} = S'_{12} S_{11}^{-1} S_{12}$.

1. Cluster Allocations and Shared Atoms ➤ simple full-conditionals

1. Cluster Allocations and Shared Atoms ➤ simple full-conditionals
2. Weights ➤ $s_{gm} \mid \cdot \sim$ mixtures of Gammas depending on n_{gm} , for $g = 1, \dots, 2\alpha$
We obtain a single Gamma rv when $n_{gm} = 0$ or $\Psi = \mathbb{I}_G$

1. Cluster Allocations and Shared Atoms ➤ simple full-conditionals
2. Weights ➤ $s_{gm} \mid \cdot \sim$ mixtures of Gammas depending on n_{gm} , for $g = 1, \dots, 2\alpha$
We obtain a single Gamma rv when $n_{gm} = 0$ or $\Psi = \mathbb{I}_G$
3. Number of components M ➤ unnormalised weights construction.
Let $U = \text{diag}(u_1, \dots, u_G)$ auxiliary variables, $M^{(\text{empty})} :=$ number of empty components (across groups):

$$q_{M^{(\text{empty})}}(m) = \frac{(m + M^{(\text{obs})})!}{m!} q_M(m + M^{(\text{obs})}) \psi(\mathbf{u})^m, \quad \psi(\mathbf{u}) = \det \left(\mathbb{I}_G + \frac{1}{2} \Psi U \right)^{-\alpha}$$

When $q_M = \text{Poi}_1(\Lambda)$:

$$q_{M^{(\text{empty})}}(m) = \pi_0 \text{Poi}_0(\Lambda \psi(\mathbf{u})) + \pi_1 \text{Poi}_1(\Lambda \psi(\mathbf{u}))$$

$$\pi_0 = \frac{M^{(\text{obs})}}{M^{(\text{obs})} + \Lambda \psi(\mathbf{u})}, \pi_1 = 1 - \pi_0$$

analogous to Argiento and De Iorio (2022) and Colombi et al. (2024) constructions.