Mixture of Poisson PCA for joint clustering and dimension reduction of count-data

Nicolas Jouvin, Julien Chiquet & others (Disclaimer: work-in-progress)

Statlearn - Montpellier, jeudi 06 avril 2023

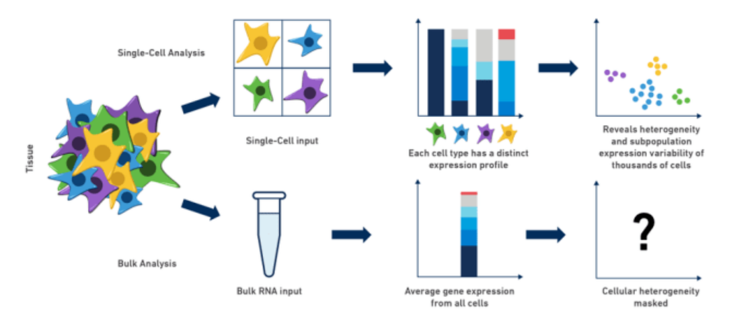
MIA
PARIS-SACLAY
SOLSTIS

Motivations: clustering & visualization of count-data

Count-data arise in many modern scientific field

Example 1: biology & single-cell RNAseq

Group similar cells based on their *gene* expression profile



Source: 10x Genomics

Count-data arise in many modern scientific field

Example 2: ecology

Group ecological sites based on their *species* abundance (thx @ Eleni Matechou)

Site	Psy	Hym	Ath	Cea	...	Set
1	27	2	0	0	...	4
2	220	15	0	0	...	0
3	1173	0	1	2	...	71
4	2671	12	1	3	...	49
⋮						

Count-data arise in many modern scientific field

Example 3: document clustering

Group similar texts based on their *words* profile

MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

Les prélèvements examinés correspondent à des fragments de tissu mammaire remanié par une prolifération tumorale dont les caractères morphologiques sont ceux d'un adénocarcinome canalaire infiltrant. Cette lésion est peu différenciée, d'architecture essentiellement trabéculaire. Les cellules néoplasiques comportent des atypies nucléaires marquées. L'index mitotique est élevé (22 mitoses sur 10 champs au grandissement 400). Deux fragments de 8 et 15 mm. Adénocarcinome mammaire de type canalaire infiltrant peu différencié. Grade histo-pronostique (EE) : III Index mitotique élevé.

MICROBIOPSIE SOUS ECHOGRAPHIE DU SEIN DROIT

MACROSCOPIE

Cinq fragments de 5 à 15 mm

MICROSCOPIE

L'examen histologique met en évidence des lésions tumorales dont les caractères morphologiques sont ceux d'un carcinome canalaire infiltrant moyennement différencié. La lésion est d'architecture trabéculaire et glanduliforme. Les cellules sont caractérisées par des atypies cytonucléaires modérées. L'activité mitotique est faible : deux mitoses ont été dénombrées sur dix champs au grandissement 400. Ces lésions sont associées à un stroma dense fibreux. Elles infiltrent le tissu adipeux. Deux séries de prélèvements ont été confiées : A - 1er tour : onze cylindres biopsiques mesurant 10 à 30 mm de long. B - 2ème tour : onze cylindres biopsiques mesurant 5 à 30 m de long. Adénocarcinome mammaire de type canalaire infiltrant. Grade histopronostique (EE) I. Index mitotique faible.

. . .

MACROBIOPSIE DU SEIN GAUCHE

MACROSCOPIE

3 fragments de 7 à 15 mm

MICROSCOPIE

Tous les prélèvements ont un aspect histologique similaire. Ils correspondent à des fragments de tissu mammaire remanié par des lésions de mastose fibreuse commune. Présence d'un discret infiltrat inflammatoire. On retrouve également quelques microcalcifications. L'un des prélèvements cryo-préservés sera analysé histologiquement et un compte rendu complémentaire adressé ultérieurement. Trois fragments de 7 à 15 mm. Lésions de mastose fibreuse. Le prélèvement paraît peu significatif. Une analyse complémentaire sur le prélèvement cryo-préservé sera réalisée.

Doc 1	"Lésions cancéreuses (. . .) carcinome canalaire"
Doc 2	"Lésions cancéreuses (. . .) carcinome lobulaire"
. . .	. . .
Doc n	"Lésions bénignes (. . .) métaplasie"

Statistical context & problematics

Multivariate data $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$

- discrete: $\boldsymbol{y}_i \in \mathbb{N}^p$
- possibly highly-dimensional ($p >> n$) or with small sample size.
- No Gaussianity, sparse data, over-dispersion

Statistical context & problematics

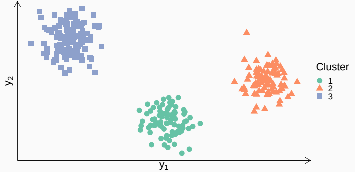
Multivariate data $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$

- discrete: $\boldsymbol{y}_i \in \mathbb{N}^p$
- possibly highly-dimensional ($p >> n$) or with small sample size.
- No Gaussianity, sparse data, over-dispersion

Our goal: unsupervised data analysis

(1) Clustering $\rightsquigarrow$ partition

(2) Dimension reduction $\rightsquigarrow$ visualization



Idealized version

How ? design a statistical model on $\boldsymbol{Y}$ integrating (1) & (2)

Overview

Count data modeling: the Poisson Log Normal (PLN) family

A mixture of PLN-PCA for joint clustering and dimension reduction

Inference

Conclusion

Count data modeling: the Poisson Log Normal (PLN) family


The world if everything
was Gaussian

What If ? Everything was Gaussian

In a Gaussian world, we would love to use the GLM framework

$$\boldsymbol{y}_i = \underbrace{\boldsymbol{x}_i^\top \boldsymbol{B}}_{\text{covariates}} + \underbrace{\boldsymbol{o}_i}_{\text{offset}} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$$

What If ? Everything was Gaussian

In a Gaussian world, we would love to use the GLM framework

$$\boldsymbol{y}_i = \underbrace{\boldsymbol{x}_i^\top \boldsymbol{B}}_{\text{covariates}} + \underbrace{\boldsymbol{o}_i}_{\text{offset}} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Sigma})$$

Pros:

- account for offset $\boldsymbol{o}_i$ and covariates $\boldsymbol{x}_i$ when available
- $\boldsymbol{\Sigma}$ capture all the remaining covariance
- flexible, generalizes

What If ? Everything was Gaussian

In a Gaussian world, we would love to use the GLM framework

$$\boldsymbol{y}_i = \underbrace{\boldsymbol{x}_i^\top \boldsymbol{B}}_{\text{covariates}} + \underbrace{\boldsymbol{o}_i}_{\text{offset}} + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{\Sigma})$$

Pros:

- account for offset $\boldsymbol{o}_i$ and covariates $\boldsymbol{x}_i$ when available
- $\boldsymbol{\Sigma}$ capture all the remaining covariance
- flexible, generalizes

However... for multivariate counts

- Data transformation (log, normalization): quick and dirty
- Non-Gaussian multivariate distributions: do not scale to data dimension yet

The vanilla PLN model: a Gaussian love story

Gaussian latent layer encoding for Poisson (log-)intensities (Aitchison et al. 1989)

$$\begin{aligned} \boldsymbol{\eta}_i &\sim \mathcal{N}_p(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}), && \text{(param)} \\ \boldsymbol{y}_i \mid \boldsymbol{\eta}_i &\sim \otimes_j \mathcal{P}(\exp(\eta_{ij})) && \text{(emission)} \end{aligned} \tag{PLN}$$

The vanilla PLN model: a Gaussian love story

Gaussian latent layer encoding for Poisson (log-)intensities (Aitchison et al. 1989)

$$\begin{aligned} \boldsymbol{\eta}_i &\sim \mathcal{N}_p(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}), && \text{(param)} \\ \boldsymbol{y}_i \mid \boldsymbol{\eta}_i &\sim \otimes_j \mathcal{P}(\exp(\eta_{ij})) && \text{(emission)} \end{aligned} \tag{PLN}$$

Parameters θ

- $\boldsymbol{B}$, the regression parameters
- $\boldsymbol{\Sigma}$, the variance-covariance matrix

Observations

- $\boldsymbol{Y}$: the count-data matrix $n \times p$
- $\boldsymbol{X}$: the covariates matrix $n \times d$
- $\boldsymbol{O}$: the offsets matrix $n \times p$

Some properties of PLN models

Over-dispersion

- mean: $\mathbb{E}(Y_{ij}) = \exp\left(o_{ij} + \mathbf{x}_i^\top \mathbf{B}_{\cdot j} + \sigma_{jj}/2\right) > 0$
- variance: $\mathbb{V}(Y_{ij}) = \mathbb{E}(Y_{ij}) + \mathbb{E}(Y_{ij})^2 \left(e^{\sigma_{jj}} - 1\right) > \mathbb{E}(Y_{ij})$
- covariance: $\mathrm{Cov}(Y_{ij}, Y_{ik}) = \mathbb{E}(Y_{ij})\mathbb{E}(Y_{ik}) \left(e^{\sigma_{jk}} - 1\right)$

Underlying assumption: correlations are captured in the latent layer

Flexible, with many extensions: **clustering**, **dimension reduction**, network inference, ...
(Chiquet et al. 2021) & an **R** package PLNmodels

Extension to clustering: mixture modeling

Goal find a partition $\boldsymbol{Z}$ into K groups

How ? Use a Gaussian mixture in the latent layer

$$\boldsymbol{\eta}_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}_p(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{M}_K(1, \boldsymbol{\pi}), && \text{(membership)} \\ \boldsymbol{\eta}_i \mid \{\boldsymbol{z}_i = k\} &\sim \mathcal{N}_p(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), && \text{(param)} \\ \boldsymbol{y}_i \mid \boldsymbol{\eta_i} &\sim \otimes_j \mathcal{P}(\exp(\eta_{ij})) && \text{(emission)} \end{aligned} \qquad \text{(PLNmixture)}$$

$$\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \boldsymbol{B}\}$$

Clustering ? Via the *posterior* distribution

$$p(\boldsymbol{Z} \mid \boldsymbol{Y}, \theta)$$

Problem : p^2 parameters in $\mathbf{\Sigma}$, what if p is ”large” ?

Extension to dimension reduction with the PLN-PCA (Chiquet et al. 2018)

Problem : p^2 parameters in $\boldsymbol{\Sigma}$, what if p is "large" ?

Regularization Low-rank factorization of $\boldsymbol{\Sigma} = \boldsymbol{C}\boldsymbol{C}^\top$, with $\boldsymbol{C}$ a $p \times q$ matrix

Extension to dimension reduction with the PLN-PCA (Chiquet et al. 2018)

Problem : p^2 parameters in $\mathbf{\Sigma}$, what if p is "large" ?

Regularization Low-rank factorization of $\mathbf{\Sigma} = \boldsymbol{C}\boldsymbol{C}^\top$, with $\boldsymbol{C}$ a $p \times q$ matrix

Probabilistic PCA formulation (Tipping et al. 1999b)

$$\begin{aligned} \boldsymbol{w}_i &\sim \mathcal{N}_p(\mathbf{0}_q, \boldsymbol{I}_q) && \text{(low-dimensional subspace)} \\ \boldsymbol{\eta}_i &= \boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{C}\boldsymbol{w}_i && \text{(linear transformation)} \\ \boldsymbol{y}_i \mid \boldsymbol{\eta}_i &\sim \otimes_j \mathcal{P}(\exp(\eta_{ij})) && \text{(emission)} \end{aligned} \qquad \text{(PLN-PCA)}$$

$\theta = \{\boldsymbol{B}, \boldsymbol{C}\}$

Dimension reduction ? $p(\boldsymbol{W} \mid \boldsymbol{Y}, \theta)$, $\boldsymbol{W} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_n\}$

- $\boldsymbol{C}$: the *loadings* matrix, basis of the latent subspace
- $\boldsymbol{w}_i$: the *scores*, coordinates in the subspace

Related to exponential family pPCA (Collins et al. 2001)

A mixture of PLN-PCA for joint clustering and dimension reduction

Integrating both approaches: mixture of PLN-PCA

Gaussian mixture in the common latent q-dimensional subspace

$$
\begin{array}{ll}
\boldsymbol{z}_i \sim \mathcal{M}_K(1, \boldsymbol{\pi}) & \text{(clustering)} \\
\boldsymbol{w}_i \mid \boldsymbol{z}_{ik} = 1 \sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) & \text{(subspace)} \\
\boldsymbol{\eta}_i \mid \boldsymbol{w}_i = \boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{C}\boldsymbol{w}_i & \text{(linear transform)} \\
\boldsymbol{y}_i \mid \boldsymbol{\eta_i} \sim \otimes_j \mathcal{P}(\exp(\eta_{ij})) & \text{(emission)}
\end{array}
\quad \text{(mPLN-PCA)}
$$

With parameters $\theta = \{\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{C}, \boldsymbol{\pi}\}$

The latent layer could be summarized[1]

$$
\boldsymbol{\eta}_i \sim \sum_{k=1}^{K} \pi_k \mathcal{N}_p(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{C}\boldsymbol{\mu}_k, \boldsymbol{C}\Lambda_k \boldsymbol{C}^\top)
$$

[1] Analogy with mixture of factors models (Tipping et al. 1999a; McNicholas et al. 2008; McParland et al. 2019)

Properties & discussion

Identifiability with common loadings

- (Scale invariance) $\boldsymbol{C}^\top \boldsymbol{C} = \mathrm{Id}_q$
- (Rotational invariance): $\boldsymbol{C}$ and $\boldsymbol{\Lambda}_k$ only up to a rotation of the latent space

Let $\boldsymbol{R} \in \mathrm{Rot}(q)$, likelihood invariant under$(\boldsymbol{C}, \boldsymbol{\Lambda}_k) \mapsto (\boldsymbol{C}\boldsymbol{R}, \boldsymbol{R}^\top \boldsymbol{\Lambda}_k \boldsymbol{R})$

$\rightsquigarrow$ can always align the axis with the principal directions of one cluster

General model A more general model could be written with one $\boldsymbol{C}_k$ per cluster

- one subspace per cluster $\rightsquigarrow$ no common projection
- $\boldsymbol{C}_k \boldsymbol{\Lambda}_k \boldsymbol{C}_k{}^\top \rightsquigarrow \boldsymbol{\Lambda}_k$ should be diagonal

Inference

Intractable likelihood

Goal estimating $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{C}, B\}$ via $\arg\max_\theta p_\theta(\boldsymbol{Y})$

EM algorithm Standard for latent variable models, use decomposition

$$\log p_\theta(\boldsymbol{Y}) = \mathbb{E}_{p_\theta(\boldsymbol{W},\boldsymbol{Z}|\boldsymbol{Y})}\left[\log p_\theta(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z})\right] + \mathcal{H}(p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y}))$$

with $\mathcal{H}(p) = -\int p \log p$ the entropy of p.

Intractable likelihood

Goal estimating $\theta = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k, \boldsymbol{C}, B\}$ via $\arg\max_\theta p_\theta(\boldsymbol{Y})$

EM algorithm Standard for latent variable models, use decomposition

$$\log p_\theta(\boldsymbol{Y}) = \mathbb{E}_{p_\theta(\boldsymbol{W},\boldsymbol{Z}|\boldsymbol{Y})}\left[\log p_\theta(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z})\right] + \mathcal{H}(p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y}))$$

with $\mathcal{H}(p) = -\int p \log p$ the entropy of p.

Problem(s) Even for vanilla PLN, intractable

1. likelihood
$$p_\theta(\boldsymbol{y}_i) = \sum_{\boldsymbol{z}_i} \int_{\mathbb{R}^q} p_\theta(\boldsymbol{y}_i, \boldsymbol{w}_i, \boldsymbol{z}_i)\, \mathrm{d}\boldsymbol{w}_i$$
2. posterior $p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})$ (or its first moments)

Solution use variational inference !

Variational inference: the Evidence Lower BOund

For any distribution q on $(\boldsymbol{W}, \boldsymbol{Z})$, the following inequality holds

$$\log p_\theta(\boldsymbol{Y}) \geq \mathcal{J}(\theta, q) := \mathbb{E}_q \left[\log p_\theta(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z})\right] + \mathcal{H}(q)$$

With a quantified gap

$$\log p_\theta(\boldsymbol{Y}) - \mathcal{J}(\theta, q) = \mathrm{KL}(q \,\|\, p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})) \geq 0$$

Variational inference: the Evidence Lower BOund

For any distribution q on $(\boldsymbol{W}, \boldsymbol{Z})$, the following inequality holds

$$\log p_\theta(\boldsymbol{Y}) \geq \mathcal{J}(\theta, q) := \mathbb{E}_q \left[\log p_\theta(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z})\right] + \mathcal{H}(q)$$

With a quantified gap

$$\log p_\theta(\boldsymbol{Y}) - \mathcal{J}(\theta, q) = \mathrm{KL}(q \,\|\, p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})) \geq 0$$

Fix θ Without constraints, minimization leads to $q = p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})$

$\rightsquigarrow$ we constrain $q = q_\psi$ in a parametric class $\mathcal{Q}$: the *variational family*

$$\arg\min_\psi \mathrm{KL}(q_\psi \,\|\, p_\theta(\cdot \mid \boldsymbol{Y})) = \arg\max_\psi \mathcal{J}(\theta, \psi)$$

Variational inference: the Evidence Lower BOund

For any distribution q on $(\boldsymbol{W}, \boldsymbol{Z})$, the following inequality holds

$$\log p_\theta(\boldsymbol{Y}) \geq \mathcal{J}(\theta, q) := \mathbb{E}_q \left[\log p_\theta(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{Z})\right] + \mathcal{H}(q)$$

With a quantified gap

$$\log p_\theta(\boldsymbol{Y}) - \mathcal{J}(\theta, q) = \mathrm{KL}(q \,\|\, p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})) \geq 0$$

Fix θ Without constraints, minimization leads to $q = p_\theta(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})$

$\rightsquigarrow$ we constrain $q = q_\psi$ in a parametric class $\mathcal{Q}$: the *variational family*

$$\arg\min_\psi \mathrm{KL}(q_\psi \,\|\, p_\theta(\cdot \mid \boldsymbol{Y})) = \arg\max_\psi \mathcal{J}(\theta, \psi)$$

Resulting VEM algorithm Iteratively solve

(VE-step) $\psi^{(t+1)} = \arg\max_\psi \mathcal{J}(\theta^{(t)}, \psi)$
(M-step) $\theta^{(t+1)} = \arg\max_\theta \mathcal{J}(\theta, \psi^{(t+1)})$

The variational family & the ELBO

Mean-field assumption & variational family [2]

$$\mathcal{Q} := \left\{ q_\psi(\boldsymbol{W}, \boldsymbol{Z}) = \prod_{i=1}^{n} q_i(\boldsymbol{w}_i, \boldsymbol{z}_i) = q_i(\boldsymbol{w}_i) q_i(\boldsymbol{z}_i) \;:\; \left(\begin{array}{ll} \bullet & q_i(\boldsymbol{w}_i) = \mathcal{N}_q(\boldsymbol{m}_i,\, \mathrm{diag}(\boldsymbol{s}_i \odot \boldsymbol{s}_i)) \\ \bullet & q_i(\boldsymbol{z}_i) = \mathcal{M}_K(1, \boldsymbol{\tau}_i) \end{array} \right. \right\}$$

[2] Alternative choice of variational family, *e.g.* $q_\psi(\boldsymbol{w}_i, \boldsymbol{z}_i) = q_\psi(\boldsymbol{w}_i \mid \boldsymbol{z}_i) q_\psi(\boldsymbol{z}_i)$

The variational family & the ELBO

Mean-field assumption & variational family [2]

$$\mathcal{Q} := \left\{ q_\psi(\boldsymbol{W}, \boldsymbol{Z}) = \prod_{i=1}^{n} q_i(\boldsymbol{w}_i, \boldsymbol{z}_i) = q_i(\boldsymbol{w}_i) q_i(\boldsymbol{z}_i) \ : \ \begin{pmatrix} \bullet & q_i(\boldsymbol{w}_i) = \mathcal{N}_q(\boldsymbol{m}_i, \, \mathrm{diag}(\boldsymbol{s}_i \odot \boldsymbol{s}_i)) \\ \bullet & q_i(\boldsymbol{z}_i) = \mathcal{M}_K(1, \boldsymbol{\tau}_i) \end{pmatrix} \right\}$$

$$\mathcal{J}(\theta, \psi) = \sum_{i=1}^{n} \mathcal{J}_i(\theta, \psi_i), \quad \psi_i = (\boldsymbol{m}_i, \boldsymbol{s}_i, \boldsymbol{\tau}_i) \in \mathbb{R}^q \times \mathbb{R}^q \times \Delta_K \qquad \text{(ELBO)}$$

Let $\boldsymbol{A}_i := \mathbb{E}_{\boldsymbol{w}_i \sim q_i}[\exp(\boldsymbol{\eta}_i)] = \exp\left(\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{C}\boldsymbol{m}_i + \frac{1}{2}\boldsymbol{C}^2 \boldsymbol{s}_i^2\right)$

$$\begin{aligned} \mathcal{J}_i(\theta, \psi_i) =& \boldsymbol{y}_i^\top (\boldsymbol{o}_i + \boldsymbol{x}_i^\top \boldsymbol{B} + \boldsymbol{C}\boldsymbol{m}_i) - \boldsymbol{A}_i^\top \mathbf{1}_p + \log(\boldsymbol{s}_i^2)^\top \mathbf{1}_q + \text{cst} \\ & - \frac{1}{2} \sum_{k=1}^{K} \tau_{ik} \left(2 \log \frac{\tau_{ik}}{\pi_k} - \log|\boldsymbol{\Lambda}_k| + \mathrm{Tr}\left[\boldsymbol{\Lambda}_k^{-1} (\boldsymbol{m}_i \boldsymbol{m}_i^\top + \mathrm{diag}(\boldsymbol{s}_i^2)) \right] \right) \end{aligned}$$

[2] Alternative choice of variational family, *e.g.* $q_\psi(\boldsymbol{w}_i, \boldsymbol{z}_i) = q_\psi(\boldsymbol{w}_i \mid \boldsymbol{z}_i) q_\psi(\boldsymbol{z}_i)$

Properties of the ELBO

The ELBO $\mathcal{J}(\theta, \psi)$ is

- bi-concave wrt $\boldsymbol{\tau}$ and $(\boldsymbol{M}, \boldsymbol{S})$
- concave wrt $(\boldsymbol{C}, \boldsymbol{B})$
- bi-concave wrt to $\boldsymbol{\mu}_k$ and $\boldsymbol{\Lambda}_k$

but not jointly concave.

Closed-form M-step when ψ is fixed

- GMM part

$$\hat{\pi}_k = \frac{\tilde{n}_k}{n}, \quad \hat{\boldsymbol{\mu}}_k = \frac{1}{\tilde{n}_k} \boldsymbol{\tau}_{.k}^\top \boldsymbol{M}, \quad \hat{\boldsymbol{\Lambda}}_k = \frac{1}{\tilde{n}_k} (\boldsymbol{M} - \mathbf{1}_n \hat{\boldsymbol{\mu}}_k^\top)^\top \operatorname{diag}(\boldsymbol{\tau}_{.k})(\boldsymbol{M} - \mathbf{1}_n \hat{\boldsymbol{\mu}}_k^\top) + \operatorname{diag}(\boldsymbol{\tau}_{.k}^\top \boldsymbol{S})$$

- Covariables: $\hat{\boldsymbol{B}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X} \boldsymbol{M} \boldsymbol{C}^\top$

But not for $\boldsymbol{C}$ or VE-step

Inference: two possible strategies

(1) Standard variational EM: alternate between $\max_{\psi} \mathcal{J}$ & $\max_{\theta} \mathcal{J}$.

(2) Joint optimization of J w.r.t. (θ, ψ)

- Scalability with n and p (work of **Bastien Batardière** on PLN-PCA)

WIP : currently working on torch implementation (**R** & **Python**)

- automatic differentiation framework to compute ∇_{θ} & ∇_{ψ}
- stochastic optimization (*e.g.* ADAM)
- Amortized VI: $q_{\psi} = q(\boldsymbol{w}_i, \boldsymbol{z}_i \mid g_{\psi}(\boldsymbol{y}_i))$, g_{ψ} neural net with weights ψ

Model selection

Choice of (K, q) is a model selection problem

Clustering context: integrated classification likelihood (ICL, Biernacki et al. 2000)

BIC-like: (very) temporarily adopt a Bayesian POV on parameters θ

$$\log p(\boldsymbol{Y}, \boldsymbol{Z}) = \int_{\theta} \int_{\boldsymbol{W}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}, \theta)$$

Model selection

Choice of (K, q) is a model selection problem

Clustering context: integrated classification likelihood (ICL, Biernacki et al. 2000)

BIC-like: (very) temporarily adopt a Bayesian POV on parameters θ

$$\log p(\boldsymbol{Y}, \boldsymbol{Z}) = \int_{\theta} \int_{\boldsymbol{W}} p(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W}, \theta)$$

Laplace + Stirling approximation + ELBO proxy leads to an approximate ICL

$$vICL(K, q) = \mathcal{J}(\hat{\theta}, \hat{\psi}) - \frac{1}{2}\left(pq - \frac{q(q+1)}{2} + K - 1 + Kq + K\frac{q(q+1)}{2}\right)\log(n) \tag{1}$$

$\mathcal{J}(\hat{\theta}, \hat{\psi})$ serves as a proxy for $\log p(\boldsymbol{Y}, \hat{\boldsymbol{Z}} \mid \hat{\theta})$

Conclusion

Summary of the methodology

Objectives

1. a partition (*clustering*)

$$\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$$

Summary of the methodology

Objectives

① a partition (*clustering*)

$$\boldsymbol{Z} = \{\boldsymbol{z}_1, \dots, \boldsymbol{z}_n\}$$

② low-dimensional representation

$$\boldsymbol{W} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_n\}$$

Summary of the methodology

Objectives

① a partition (*clustering*)

$$\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$$

② low-dimensional representation

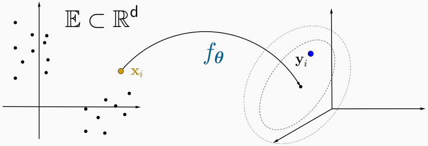
$$\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$$

Method

$$\boldsymbol{Z}, \boldsymbol{W} \sim p_\theta \qquad (latent)$$

$$\boldsymbol{\eta} = f_\theta(\boldsymbol{W}) \qquad (param)$$

$$\boldsymbol{Y} \mid \boldsymbol{\eta} \sim p(\cdot \mid \boldsymbol{\eta}) \qquad (obs)$$



Summary of the methodology

Objectives

① a partition (*clustering*)

$$\boldsymbol{Z} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n\}$$

② low-dimensional representation

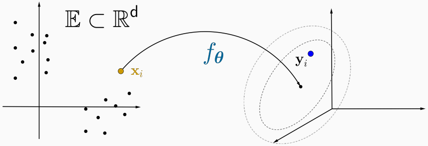
$$\boldsymbol{W} = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n\}$$

Method

$$\boldsymbol{Z}, \boldsymbol{W} \sim p_\theta \qquad (latent)$$

$$\boldsymbol{\eta} = f_\theta(\boldsymbol{W}) \qquad (param)$$

$$\boldsymbol{Y} \mid \boldsymbol{\eta} \sim p(\cdot \mid \boldsymbol{\eta}) \qquad (obs)$$



Inference to estimate $\hat{\theta}$ + *(variational) posterior* for $\hat{\boldsymbol{Z}}, \hat{\boldsymbol{W}} \approx \arg\max_{\boldsymbol{W},\boldsymbol{Z}} p_{\hat{\theta}}(\boldsymbol{W}, \boldsymbol{Z} \mid \boldsymbol{Y})$

Perspectives

Remaining things to do

- Finish the inference algorithm + integration in the PLNmodels package
- Application on sigle-cell RNAseq data
- Numerical investigation of the property of M-estimator (work of J. Chiquet et. al. on PLN using (Westling et al. 2015))

$$\hat{\theta}_n = \arg\max_{\theta} \left\{ \mathcal{J}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{J}_i\left(\theta, \hat{\psi}_i(\theta, \boldsymbol{y}_i)\right) \right\}, \quad \hat{\theta}_n \xrightarrow[n\to+\infty]{?} \arg\max_{\theta} \mathbb{E}_{\theta^\star}\left[\mathcal{J}\left(\theta, \hat{\psi}(\theta, \mathbf{Y})\right)\right]$$

- Zero-inflation $\rightsquigarrow$ introduce a binary *mask* variable $H_{ij} \sim \mathcal{B}(\text{logit}(\boldsymbol{x}_i^\top \boldsymbol{B}_j^0))$

Non-linear extensions: variational auto-encoders

- $\boldsymbol{\eta}_i = f_\theta(\boldsymbol{w}_i)$, f_θ neural net with weights θ (encoder)
- $q_\psi = q(\boldsymbol{w}_i, \boldsymbol{z}_i \mid g_\psi(\boldsymbol{y}_i))$, g_ψ neural net with weights ψ (decoder)

Thank you for your attention[3]

[3] And sorry for the lack of experiments today, though I'd like to share the blame with the French Government...

References

- Aitchison, J. and C.H. Ho (1989). “The multivariate Poisson-log normal distribution”. In: *Biometrika* 76.4, pp. 643–653.
- Biernacki, Christophe, Gilles Celeux, and Gérard Govaert (2000). “Assessing a mixture model for clustering with the integrated completed likelihood”. In: *IEEE transactions on pattern analysis and machine intelligence* 22.7, pp. 719–725.
- Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (2018). “Variational inference for probabilistic Poisson PCA”. In: *The Annals of Applied Statistics* 12.4, pp. 2674–2698.
- Chiquet, Julien, Mahendra Mariadassou, and Stéphane Robin (2021). “The Poisson-lognormal model as a versatile framework for the joint analysis of species abundances”. In: *Frontiers in Ecology and Evolution* 9, p. 588292.
- Collins, Michael, Sanjoy Dasgupta, and Robert E Schapire (2001). “A generalization of principal components analysis to the exponential family”. In: *Advances in neural information processing systems* 14.

McNicholas, Paul David and Thomas Brendan Murphy (2008). “Parsimonious Gaussian mixture models”. In: *Statistics and Computing* 18.3, pp. 285–296.

McParland, Damien and Thomas Brendan Murphy (2019). “Mixture modelling of high-dimensional data”. In: *Handbook of Mixture Analysis*. Chapman and Hall/CRC, pp. 239–270.

Tipping, Michael E and Christopher M Bishop (1999a). “Mixtures of probabilistic principal component analyzers”. In: *Neural computation* 11.2, pp. 443–482.

Tipping, Michael E and Christopher M Bishop (1999b). “Probabilistic principal component analysis”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61.3, pp. 611–622.

Westling, Ted and Tyler H. McCormick (2015). *Beyond prediction: A framework for inference with variational approximations in mixture models*.

Questions