Modèles d'imputation pour des comptages d'oiseaux d'eau en Afrique du Nord

Barbara BRICOUT

Encadrants: L. Dami, P. Defos du Rau, S. Donnet, T. Galewski, S. Robin

Sorbonne Université

March 25, 2024

Les oiseaux d'eau

Ensemble des oiseaux aquatiques sauvages qui vivent dans l'eau et les zones humides.

But : Estimer l'abondance des oiseaux d'eau en Afrique du Nord à partir des données DIOE.

Matrice de comptages

	X1990	X1991	X1992	X1993	X1994	X1995	X1996
Site 1	_	_	_	2	_	_	_
Site 2	_	0	0	0	35	_	0
Site 3	_	_	_	46	_	_	_
Site 4	_	_	_	_	0	_	_
Site 5	_	_	_	2	_	_	_

Dimensions : 300 sites, 30 années, 60% de données manquantes

Covariables

1. Covariables **spatiales** (en ligne) : Distance aux villes, à la côte, superficie des surface d'eau, écosystème, altitude
2. Covariables **temporelles** (en colonne) : Températures et précipitations en Europe au printemps et en hiver, oscillation Nord - Atlantique
3. Covariables **spatio-temporelles** (en ligne-colonne): Pourcentage de terres agricoles, densité de population, croissance économique, précipitations

Les covariables sont **complètes**.

Notations : Comptages

1. Y : Comptages sur n sites et p années
2. Y_{ij} est le nombre d'oiseaux sur le site i l'année j
3. Ω : Matrice de présence
4. $\Omega_{ij} = 1$ si Y_{ij} est observé et 0 si Y_{ij} est manquant
5. $\mathcal{O} = \{Y_{ij} : \Omega_{ij} = 1\}$ ensemble des sites $\times$ dates observés
6. $Y^{\mathcal{O}} = \{Y_{ij} : (i,j) \in \mathcal{O}\}$ ensemble des comptages observés

Notations : Covariables

1. $\underset{n \times d_1}{R}$: Covariables sur les sites (lignes)
2. $\underset{p \times d_2}{C}$: Covariables sur les années (colonnes)
3. $\underset{np \times d_3}{U}$: Covariables sur les intéractions site-année
4. $X = [R \otimes \mathbf{1}_p \quad \mathbf{1}_n \otimes C \quad U]$ matrice de covariables de dimension $np \times d$ avec $d = d_1 + d_2 + d_3$.

Ignorabilité

Hypothèses

1. Indépendance conditionnelle de l'observation et de l'abondance :

$$P(Y, \Omega|X) = P(\Omega|X)P(Y|X)$$

2. Séparabilité des paramètres

$$P_{\alpha,\theta}(Y, \Omega|X) = P_{\alpha}(\Omega|X)P_{\theta}(Y|X)$$

Donc

$$P_{\alpha,\theta}(Y^O, \Omega|X) = P_{\alpha}(\Omega|X)P_{\theta}(Y^O|X)$$

Objectif: inférence de θ par maximisation de $P_{\theta}(Y^O|X)$

Modèle

Modèle saturé :

$$Y_{ij} \sim \mathcal{P}(\exp(X_{ij}^T \gamma + \zeta_{ij}))$$

$\underset{n \times p}{\zeta}$ = Interaction sites $\times$ années

Modèle

Modèle saturé :

$$Y_{ij} \sim \mathcal{P}(\exp(X_{ij}^T \gamma + \zeta_{ij}))$$

$\underset{n \times p}{\zeta}$ = Interaction sites $\times$ années

Données incomplètes $\Rightarrow$ Hypothèse de faible rang :

$$\zeta = \underset{n \times q}{W} \underset{p \times q}{B}^T$$

Modèle

Modèle saturé :

$$Y_{ij} \sim \mathcal{P}(\exp(X_{ij}^T \gamma + \zeta_{ij}))$$

$\underset{n \times p}{\zeta}$ = Interaction sites $\times$ années

Données incomplètes $\Rightarrow$ Hypothèse de faible rang :

$$\zeta = \underset{n \times q}{W} \underset{p \times q}{B}^T$$

Choix de modélisation : W aléatoire

$$(W_i)_{i,\dots,n} \text{ i.i.d., } W_i \sim \mathcal{N}(0_q, I_q)$$

Modèle

Modèle saturé :

$$Y_{ij} \sim \mathcal{P}(\exp(X_{ij}^T \gamma + \zeta_{ij}))$$

$\underset{n \times p}{\zeta}$ = Interaction sites $\times$ années

Données incomplètes $\Rightarrow$ Hypothèse de faible rang :

$$\zeta = \underset{n \times q}{W} \underset{p \times q}{B}^T$$

Choix de modélisation : W aléatoire

$$(W_i)_{i,\dots,n} \text{ i.i.d., } W_i \sim \mathcal{N}(0_q, I_q)$$

Remarque : Si $X = R$ et $Y^O = Y$ alors il s'agit du modèle PLN-PCA (Chiquet et al. 2018 [1])

Propriétés du modèle

1. Sur-dispersion (aléa gaussien) [2]
2. Sites indépendants
3. Dépendances entre les années :

$$V(Z_i) = \Sigma = BB^T$$

Algorithme EM

Vraisemblance observée : $\log p_\theta(Y^O)$

Décomposition pour l'algorithme EM :

$$\log p_\theta(Y^O) = E(\log p_\theta(Y^O, W) \mid Y^O) + \mathcal{H}(p_\theta(W \mid Y^O))$$

Problème : $p_\theta(W \mid Y^O)$ pas calculable
$\Rightarrow$ Approximation "variationnelle"

Version variationnelle de l'EM

Pour toute loi de probabillité $\tilde{p}$ sur W.

$$\begin{aligned} J_q(Y^O; \tilde{p}, \theta) &= \log p_\theta(Y^O) - KL(\tilde{p}(W) \| p_\theta(W \mid Y^O)) \\ &= E_{\tilde{p}}(\log p(Y^O, W)) + H(\tilde{p}(W)) \\ &\leq \log p_\theta(Y^O) \end{aligned}$$

But : Maximiser la borne inférieure $J_q(\tilde{p}; \theta)$ en $\tilde{p}$ et θ.

Remarque: Revient à minimiser $KL(\tilde{p}(W) \| p_\theta(W \mid Y^O))$
$\Rightarrow \tilde{p} =$ **approximation variationnelle de** $p_\theta(W \mid Y^O)$

Choix de $\tilde{p}$

Indépendance des sites :

$$p_\theta(W \mid Y^O) = \prod_{i=1}^{n} p_\theta(W_i \mid Y^O)$$

donc

$$\tilde{p}(W) = \prod_{i=1}^{n} \tilde{p}_i(W_i)$$

Choix :

$$\tilde{p}_i = \mathcal{N}(m_i, S_i), \qquad S_i = diag(s_i \odot s_i)$$

$(m_i, s_i) \in \mathbb{R}^q \times \mathbb{R}^p$.

Optimisation

Prise en compte des données manquantes

$$
\begin{aligned}
J_q(Y^O; \tilde{p}, \theta) =& \mathbf{1}_n^\top \left(\Omega \odot \left(Y \odot \left(X\gamma + MB^\top \right) - A \right) \right) \mathbf{1}_p \\
& - \frac{1}{2} \mathbf{1}_n^\top \left(M \odot M + S \odot S - 2 \log(S) \right) \mathbf{1}_q + c
\end{aligned}
$$

avec

$$
A = \exp\left(X\gamma + MB^\top + \frac{1}{2}(S \odot S)(B \odot B)^\top \right),
$$

(Idem gradients)
Optimisation : Montée de gradient

Optimisation

Prise en compte des données manquantes

$$
\begin{aligned}
J_q(Y^O; \tilde{p}, \theta) =& \mathbf{1}_n^\top \left(\Omega \odot (Y \odot (X\gamma + MB^\top) - A) \right) \mathbf{1}_p \\
& - \frac{1}{2} \mathbf{1}_n^\top \left(M \odot M + S \odot S - 2 \log(S) \right) \mathbf{1}_q + c
\end{aligned}
$$

avec

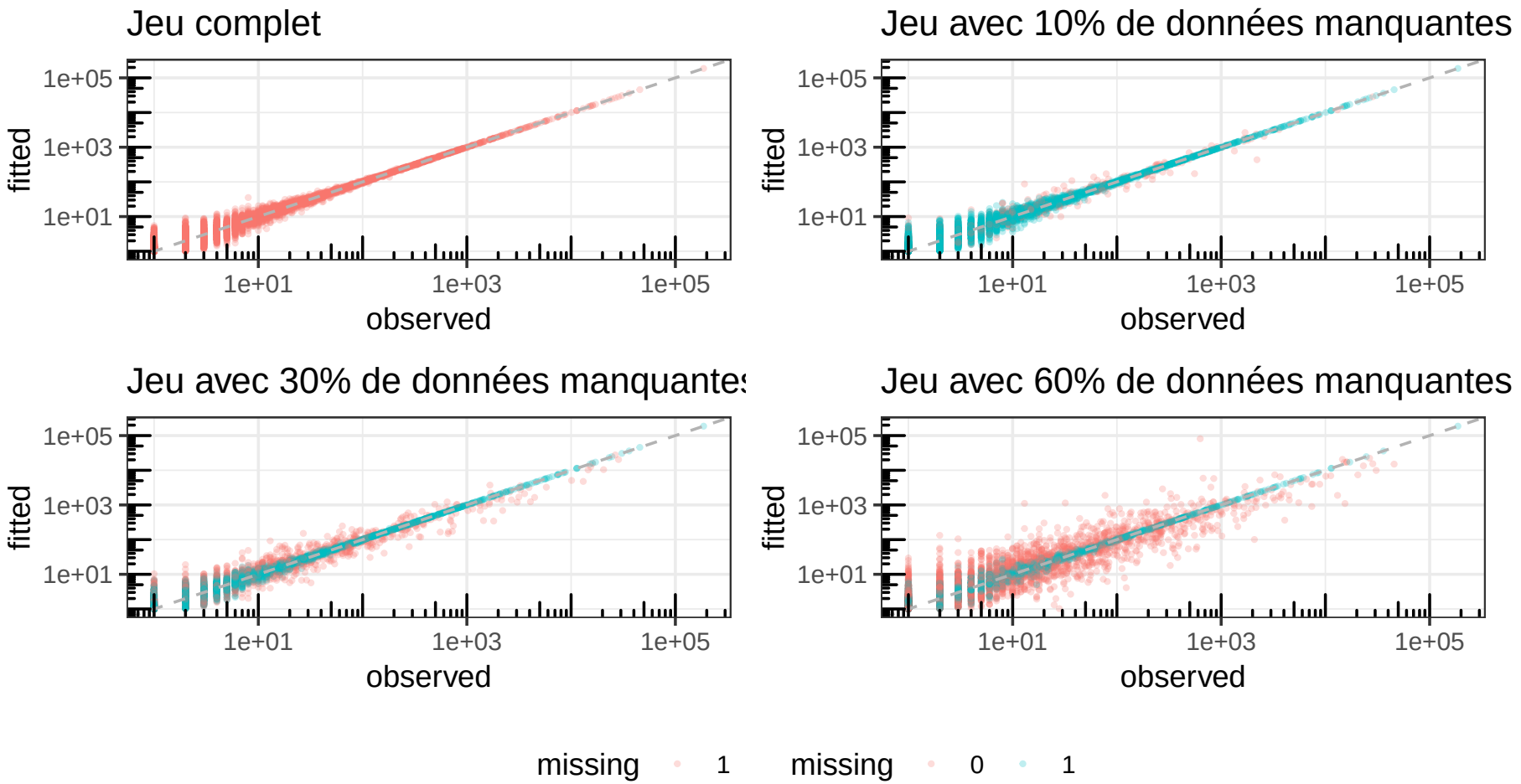
$$
A = \exp\left(X\gamma + MB^\top + \frac{1}{2}(S \odot S)(B \odot B)^\top \right),
$$

(Idem gradients)
Optimisation : Montée de gradient **Prédictions** :

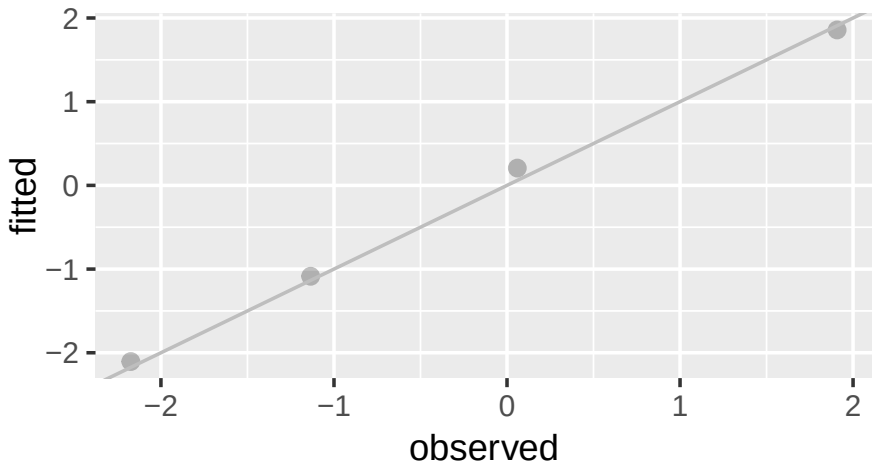
$$
\widehat{Y} = \widehat{E}(e^Z) = \exp(X\widehat{\gamma} + \widehat{M}\widehat{B}^T + \frac{1}{2}(\widehat{S} \odot \widehat{S})(\widehat{B} \odot \widehat{B})^T)
$$

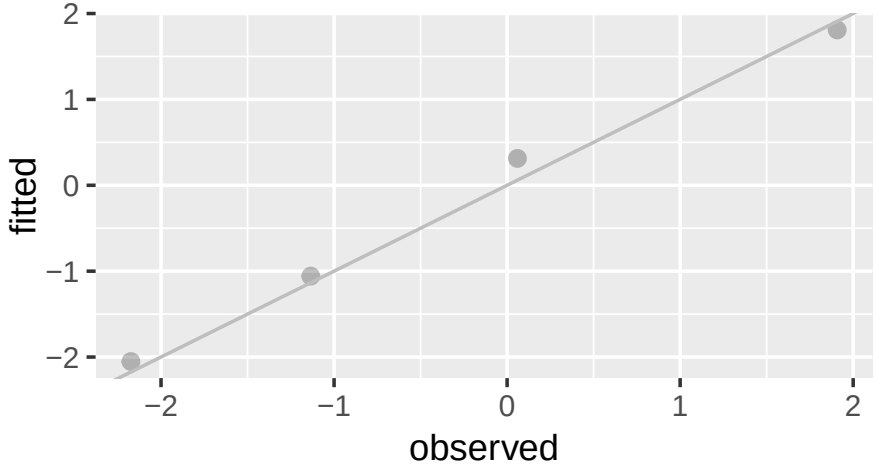
Simulations

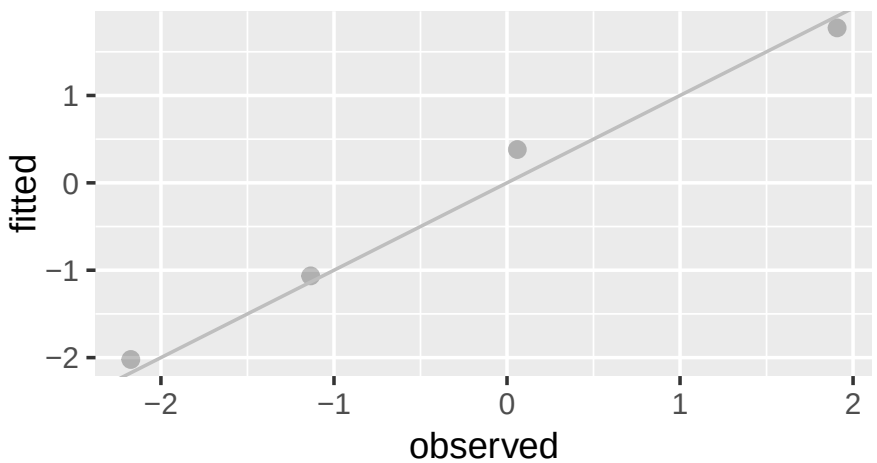
Données simulées pour $n = 300$, $p = 30$, $d = 3$ et $q = 5$

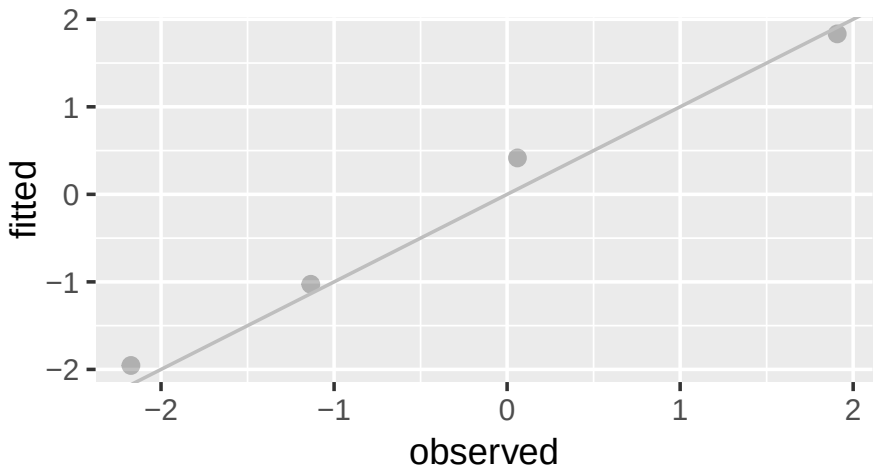


Estimation de γ



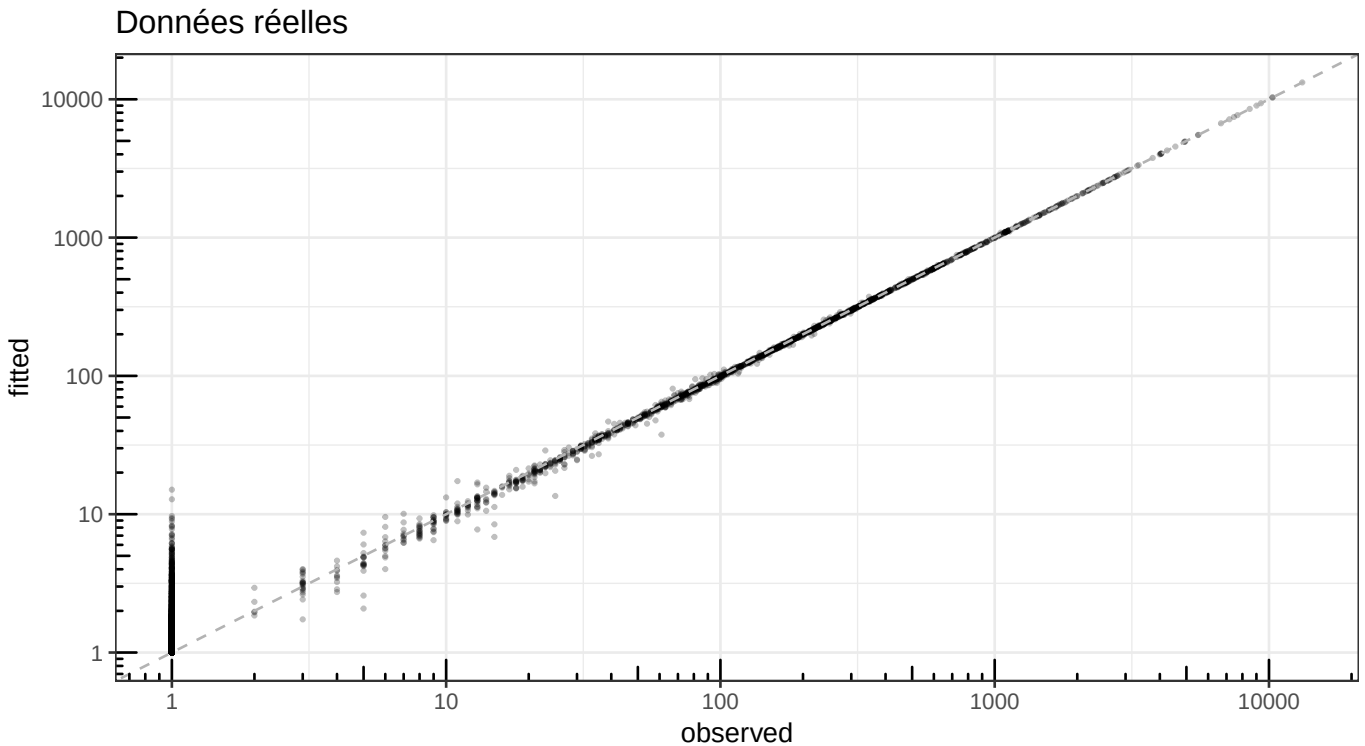






Données Tour du Valat

Comptages de Colvert en France complets



Incapacité du modèle à prédire les 0.

Modèle Présence absence

$$\forall (i,j) \in \mathcal{O}$$

Modèle saturé :

$$V_{ij} \sim \mathcal{B}(\text{logit}^{-1}(X_{ij}^T \gamma + \zeta_{ij}))$$

$\underset{n \times p}{\zeta}$ = Interaction sites $\times$ années

Données incomplètes $\Rightarrow$ Hypothèse de faible rang :

$$\zeta = \underset{n \times q}{W} \underset{p \times q}{B}^T$$

Choix de modélisation : W aléatoire

$$(W_i)_{i,\dots,n} \text{ i.i.d., } W_i \sim \mathcal{N}(0_q, I_q)$$

Borne inférieure

Soit une distribution $\tilde{p} = \otimes_{i=1}^{n} \tilde{p}_i$ de T_i qui approche $p(T|V)$ et la borne inférieure :

$$J_q(\tilde{p};\theta) = \sum_{i=1}^{n}\sum_{j=1}^{p} V_{ij}(\mu + D_j^T m_i) + E_{\tilde{p}_i}[-\log(1+\exp(\mu + D_j^T T_i)]$$
$$-\frac{1}{2}(\|m_i\|_2^2 + \|s_i\|_2^2) - \frac{1}{2}\sum_{k=1}^{q} \log s_{ik}^2$$

Minoration de la borne inférieure

Proposition

$\forall a, x \in \mathbb{R}$, *avec* $a \neq 0$,

$$\log(1+e^x) \leq \log(1+e^a) + \frac{x-a}{1+e^{-a}} + \frac{(e^a-1)(x-a)^2}{4a(e^a+1)}$$

De plus, pour tout a, il s'agit de la plus petite borne supérieure de $\log(1+e^x)$ *avec égalité lorsque* $x = a$.

[3], [4]

Recherche du a qui maximise la borne

Soit x une variable aléatoire telle que $x \sim p = \mathcal{N}(\mu, \sigma)$. On pose :

$$\begin{aligned} f(a) =& E_p(\log(1+e^a) + \frac{x-a}{1+e^{-a}} + \frac{(e^a-1)(x-a)^2}{4a(e^a+1)}) \\ =& \log(1+e^a) + \frac{e^a(\mu-a)}{1+e^a} + \frac{e^a-1}{4a(e^a+1)}(\sigma^2 + (\mu-a)^2) \end{aligned}$$

On cherche ensuite a qui minimise f, en commençant par dériver f

$$f'(a) = -\frac{(e^{2a} - 2ae^a - 1)(s^2 + m^2 - a^2)}{4a^2(e^A+1)^2}$$

Proposition

f' est une fonction symétrique qui s'annule 3 fois et atteint son minimum en deux points : $\pm\sqrt{m^2+s^2}$.

Conclusion

On a un modèle qui permet d'imputer des données manquantes dans des données de comptages sur-dispersées.

Prochaines étapes :

1. Implémenter le modèle en présence absence
2. Modèle pour des comptages avec inflations de zéros

Mon animal mignon

Bibliographie

[1] Julien Chiquet, Mahendra Mariadassou **and** Stéphane Robin. "Variational inference for probabilistic Poisson PCA". **in***The Annals of Applied Statistics*: 12.4 (2018), **pages** 2674–2698.

[2] J. Aitchison **and** C.H Ho. "The multivariate Poisson-log normal distribution". **in***Biometrika*: 76.4 (1989), **pages** 643–653.

[3] Tommi S Jaakkola **and** Michael I Jordan. "Bayesian parameter estimation via variational methods". **in***Statistics and Computing*: 10 (2000), **pages** 25–37.

[4] Julyan Arbel, Olivier Marchal **and** Hien D Nguyen. "On strict sub-Gaussianity, optimal proxy variance and symmetry for bounded random variables". **in***ESAIM: Probability and Statistics*: 24 (2020), **pages** 39–55.